An optimization perspective on log-concave sampling and beyond

by

Sinho Chewi

B.S., University of California, Berkeley, 2018

Submitted to the Department of Mathematics
in partial fulfillment of the requirements for the degree of Doctor of Philosophy
at the Massachusetts Institute of Technology.

May 2023

© Sinho Chewi. This work is licensed under a **CC BY-SA 2.0**. The author hereby grants to MIT a nonexclusive, worldwide, irrevocable, royalty-free license to exercise any and all rights under copyright, including to reproduce, preserve, distribute and publicly display copies of the thesis, or release the thesis under an open-access license.

Author: _____

Department of Mathematics
May 1, 2023

Certified by: _____

Philippe Rigollet
Professor of Mathematics
Thesis Supervisor

Accepted by: _____

Jonathan A. Kelner
Professor of Mathematics
Chair, Department Committee on Graduate Theses

An optimization perspective on log-concave sampling and beyond

by

Sinho Chewi

Submitted to the MIT Mathematics Department in May 2023,
in partial fulfillment of the requirements for the Ph.D. degree.

Abstract

The primary contribution of this thesis is to advance the theory of complexity for sampling from a continuous probability density over $\mathbb{R}^d$. Some highlights include: a new analysis of the proximal sampler, taking inspiration from the proximal point algorithm in optimization; an improved and sharp analysis of the Metropolis-adjusted Langevin algorithm, yielding new state-of-the-art guarantees for high-accuracy log-concave sampling; the first lower bounds for the complexity of log-concave sampling; an analysis of mirror Langevin Monte Carlo for constrained sampling; and the development of a theory of approximate first-order stationarity in non-log-concave sampling.

We further illustrate the main tools in this work—diffusions and Wasserstein gradient flows—through applications to functional inequalities, the entropic barrier, Wasserstein barycenters, variational inference, and diffusion models.

Thesis Supervisor: Philippe Rigollet
Title: Professor of Mathematics

In memory of Matthew Brennan (1994–2021).

Contents

Acknowledgements

When I look back at any aspect of my journey as a graduate student, I see the profound influence of my advisor, Philippe Rigollet. I see it in the way I pursue research directions and approach new topics with curiosity and enjoyment. I see it in my collaborators, all of whom I met through the opportunities he generously provided. Looking back further, I see it even at the very beginning: he was the one who convinced me that MIT was the perfect place for me (and he couldn't be more right!). Thank you, Philippe, for your endless support, and for shaping me into who I am today.

I thank my other committee members, Jonathan Kelner and Ankur Moitra, who have always been sources of inspiration for me. Ankur acted as a mentor for me during my first year of graduate school, and I owe him for generously offering his advice and his time.

I am incredibly fortunate to have collaborated with many brilliant people, who were instrumental to the completion of the work in this thesis: Kwangjun Ahn, Jason M. Altschuler, Francis Bach, Krishnakumar Balasubramanian, Silvère Bonnabel, Sébastien Bubeck, Hong-Bin Chen, Sitan Chen, Yongxin Chen, Xiang Cheng, Julien Clancy, Michael Diao, Jaume de Dios Pont, Murat A. Erdogdu, Patrik R. Gerber, Marc Lambert, Thibaut Le Gouic, Holden Lee, Yin Tat Lee, Jerry Li, Mufan (Bill) Li, Yuanzhi Li, Chen Lu, Tyler Maunu, Shyam Narayanan, Jonathan Niles-Weed, Aram-Alexandre Pooladian, Philippe Rigollet, Adil Salim, Ruoqi Shen, George Stepaniants, Austin J. Stromme, Felipe Suarez, Paxton Turner, Andre Wibisono, Anru R. Zhang, Matthew Zhang, Yi Zhang. Thank you also to all of the other wonderful people I have interacted with during my graduate studies (too many to list!). I look forward to meeting again.

I want to especially thank some people in the list above. Krishna, Sébastien, Murat, Thibaut, Tyler, Jon, Adil, and Andre have all acted as advocates and mentors for me—I deeply appreciate their unwavering belief in me, and I only hope I can repay them someday. And although I'm grateful to everyone listed above, allow me to give a shout out to Austin, who started his PhD at the same time I

did and was my closest collaborator during the scary early years. It was truly fun learning math together, and I learned a lot from his inspirational attitude.

I continue to think about the amazing life of Matthew Brennan. He was an incredible person, a steadfast friend, and my roommate. I miss him.

I met many of my collaborators for the first time at the Geometric Methods in Optimization and Sampling (GMOS) program at the Simons Institute for the Theory of Computing in Fall 2021. It was an incredibly formative experience, and I owe it to Philippe for his part in making it happen. I also had a great time visiting Jon at New York University, and the ML foundations group at Microsoft Research in Redmond as a summer intern.

I want to thank all of my friends from before graduate school who supported me throughout—you made everything so bearable. I thank Daniel Raban and Forest Yang in particular, for listening to me patiently whenever I unloaded my graduate school struggles unto them.

Finally, I thank my parents for everything I cannot express in words.

Chapter 1

Introduction

The primary aim of this thesis is to study the complexity of the task of *sampling*: given a target probability density $\pi \propto \exp(-V)$ on $\mathbb{R}^d$, how expensive is it to generate random variables whose law is close to π in suitable metrics? Since the dawn of the Markov chain Monte Carlo (MCMC) revolution [GS90], sampling has been the algorithmic cornerstone of Bayesian inference and scientific computing [RC04; Liu08; Gel+14]. How do we design fast samplers, and how can we develop a theory of complexity for this task?

The key to both of these questions lies in the remarkable connections between sampling and the mature field of optimization. Towards the question of algorithm design, there is a striking parallel between the *gradient flow*

$$\dot{X}_t = -\nabla V(X_t)\,,$$

the canonical continuous-time dynamics for obtaining a minimizer of V, and the *Langevin diffusion*

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \tag{1.1}$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion, which converges in law to its stationary distribution $\pi \propto \exp(-V)$. This suggests a first connection between the two fields in which sampling can be viewed as the "probabilistic counterpart" to optimization. Whereas in optimization we seek global minimizers of V, in sampling we must sample from $\pi \propto \exp(-V)$, thereby exploring regions in which V is small, and perhaps unsurprisingly the dynamics (1.1) for sampling is a noisy version of the gradient flow for optimization.

There is, however, a more profound link, due to the seminal work of [JKO98]. In this perspective, if we do not track the noisy evolution of the stochastic process $(X_t)_{t\geq 0}$ but instead focus our attention on the evolution of the marginal law $\mu_t := \mathrm{law}(X_t)$, then we obtain dynamics on the space $\mathcal{P}(\mathbb{R}^d)$ of probability measures over $\mathbb{R}^d$. Developing a calculus for understanding dynamics on this space introduces many new technical difficulties, but the price we pay for the increased level

of abstraction is richly compensated by a deep and newfound intuition for the Langevin diffusion. Namely, [JKO98] observed that once the space $\mathcal{P}(\mathbb{R}^d)$ is equipped with an appropriate geometric structure—the geometry arising from the theory of optimal transport [Vil03]—the marginal law $(\mu_t)_{t\geq 0}$ of the Langevin diffusion becomes a gradient flow for the Kullback–Leibler divergence $\mathsf{KL}(\cdot \,\|\, \pi)$. Thus, the Langevin diffusion is not merely a noisy variant of a gradient flow, but is in fact *exactly* a gradient flow from the right perspective!

The motto of this viewpoint can be succinctly summarized as saying that "sampling is optimization in the space of measures" [Wib18]. Besides its aesthetic appeal, it has inspired novel analyses of the Langevin diffusion [DMM19] and has given rise to a flurry of new samplers inspired by algorithms from convex optimization; for example, in this thesis we study sampling counterparts of the proximal point method (§4), Nesterov's accelerated gradient method (§6), and mirror descent (§8 and §9).

The second question we asked above was the problem of developing a theory of complexity for sampling. Here too, we draw inspiration from optimization through the celebrated oracle model of [NY83]. This model, adapted to the context of sampling, measures the work exerted by an algorithm in terms of the number of *queries* made to a first-order oracle for π. Given a query point $x \in \mathbb{R}^d$, the oracle returns $V(x) - V(0)$ and $\nabla V(x)$. Note that this query model accommodates applications such as Bayesian inference in which the normalization constant of π is unknown, since the oracle outputs can be simulated without this knowledge. Within this framework, the complexity of sampling of becomes an information-theoretic question, although it usually carries practical implications for algorithm design since for most samplers, the computational complexity and the oracle complexity are tightly related.

Once we adopt the oracle model, it is now possible to ask rather fine-grained complexity questions for sampling. One question of particular interest in this work is the following canonical one. Consider the following class of distributions: $\pi \propto \exp(-V)$, where V is α-convex and β smooth with $0 < \alpha < \beta < \infty$, and V is minimized at 0. What is the minimal number of queries to the first-order oracle necessary to output a sample which is ε-close to π in total variation distance?

Despite the extensive literature on MCMC methods, this flavor of complexity question which aims at truly understanding the intrinsic and non-asymptotic difficulty of sampling has only been studied in earnest relatively recently with early works such as [DT12; Dal17b]. This is the starting point of this thesis. In short, we use inspiration from optimization to design and analyze new samplers and make progress towards understanding the fundamental complexity problem.

We give an overview of relevant background in §2. In the rest of this introductory chapter, we summarize the contributions of the thesis.

Sampling under log-concavity and isoperimetry

§3: Analysis of Langevin Monte Carlo. We begin by studying the basic Langevin diffusion (1.1), which once discretized becomes the standard Langevin Monte Carlo (LMC) algorithm

$$X_{(k+1)h} = X_{kh} - h\,\nabla V(X_{kh}) + \sqrt{2}\,(B_{(k+1)h} - B_{kh})\,.$$

Here, the Brownian motion increment $B_{(k+1)h} - B_{kh} \sim \mathsf{normal}(0, hI_d)$ is easy to simulate and hence LMC is easily implemented.

Although LMC has been extensively studied in a non-asymptotic context since [DT12], still several questions about LMC remained unresolved.

- Is it possible to provide guarantees for LMC that hold in more stringent performance metrics? In particular, the family of *Rényi divergences* $\mathcal{R}_q$ (for $q \geq 1$) is particularly strong as it controls many other common divergences (§2.2.3). Rényi divergences have recently played an important role in the application of sampling to differential privacy [GT20], and they are also crucial for obtaining warm starts for high-accuracy samplers in §6.

- Can we obtain guarantees under weaker assumptions? Although [VW19] obtained a result under a log-Sobolev inequality (LSI), it was unknown how obtain a guarantee under a Poincaré inequality (PI). Moreover, most analyses of LMC assume that ∇V is Lipschitz, which is too restrictive when moving to the PI setting.

In this chapter, we address these questions by providing a suite of Rényi divergence guarantees under various assumptions. We prove the first Rényi guarantees under an LSI by extending the technique of [VW19]. We also develop an argument based on Girsanov's theorem that allows for a *Latała–Oleszkiewicz* inequality (LOI), which interpolates between PI and LSI, as well as Hölder continuity of ∇V (rather than only Lipschitz continuity). Altogether, our results paint a fuller understanding of the behavior of LMC in various settings.

Besides the results themselves, some of the techniques developed in this chapter are reused later. Namely, we find Lemma 3.6.3 to be particularly useful, and the Girsanov argument is extended to the underdamped Langevin diffusion in [Zha+23] and §6.

§4: Analysis of the proximal sampler. The results for LMC, however, suffer from some notable disadvantages. First, LMC is biased (for any positive $h > 0$, the stationary distribution of LMC is not equal to π); to control the size of the bias, we must take h polynomially small in the desired accuracy ε, which leads to a

low-accuracy guarantee, that is, the guarantee scales polynomially in $1/\varepsilon$. Second, the proofs for LMC are lengthy and tedious, and it is unclear if the guarantees we obtain are sharp. For instance, our complexity guarantee for LMC under a PI reads (with some simplifications) as $\widetilde{O}(\kappa^2 d^3/\varepsilon^2)$, where κ is an appropriate notion of "condition number" for this setting.

Many of these issues are resolved by instead considering the *proximal sampler* algorithm of [TP18; LST21c]. In this algorithm, we augment the target distribution π to a joint distribution $\boldsymbol{\pi}$ over $\mathbb{R}^d \times \mathbb{R}^d$ via

$$\boldsymbol{\pi}(x, y) \propto \exp\Big(-V(x) - \frac{1}{2h}\|x - y\|^2\Big)\,.$$

We then apply Gibbs sampling to $\boldsymbol{\pi}$, yielding the iterates

$$\begin{aligned} Y_k &\sim \pi^{Y|X}(\cdot \mid X_k) = \mathsf{normal}(X_k, hI_d)\,, \\ X_{k+1} &\sim \pi^{X|Y}(\cdot \mid Y_k) \propto \exp\Big(-V(\cdot) - \frac{1}{2h}\|\cdot - Y_k\|^2\Big)\,. \end{aligned}$$

This algorithm can be understood as a proximal discretization of the Wasserstein gradient flow of the KL divergence. Just as the proximal point method is well-known within optimization to be a more stable discretization of the gradient flow, we shall see that the proximal sampler affords substantial benefits over LMC. For example, since the proximal sampler is an asymptotically unbiased Markov chain, we generally it to be geometrically ergodic, leading to a *high-accuracy* sampler whose complexity scales as $\mathrm{polylog}(1/\varepsilon)$ w.r.t. the target accuracy ε.

In the second step, we must sample from the distribution $\pi^{X|Y}$, known as the *restricted Gaussian oracle* (RGO). This introduces a trade-off for the step size $h > 0$: if h is large, then the proximal sampler converges faster; however, if h is small, then the RGO is easier to implement (because it more closely resembles a Gaussian distribution). We explore different extremes of this trade-off:

- In §4, we consider an extremely small step size $h = \Theta(\frac{1}{\beta d})$, where β is the Lipschitz constant of ∇V, for which the RGO is extremely easy to implement via rejection sampling.[1]
- In §6, we consider a large step size $h = \Theta(\frac{1}{\beta})$, for which implementation of the RGO is non-trivial and requires the use of an auxiliary sampler.
- In §17, diffusion models can morally be considered instantiations of the proximal sampler with an extremely large step size, for which the proximal

[1]The later work of [FYC23] shows that with *approximate* rejection sampling, one can take a much larger step size of $h = \Theta(\frac{1}{\beta\sqrt{d}})$.

sampler converges in one iteration but the "RGO" is implemented with the use of deep learning.

Previously, [LST21c] established convergence guarantees for the proximal sampler under strong log-concavity. In §4, we introduce a new interpretation of the two steps of the proximal sampler as running a Brownian motion forwards and backwards in time respectively. Through this interpretation, we are able to prove new convergence results for the proximal sampler under weaker assumptions: under weak log-concavity, and under functional inequalities such as PI, LSI, or more generally, LOI. More broadly, the high-level message of our analysis is that, similarly to the relationship between the proximal point method and the gradient flow in optimization, the proximal sampler inherits any favorable convergence rates enjoyed by the continuous-time Langevin diffusion.

Already with the naïve rejection sampling implementation of the RGO, it yields surprising improvements over LMC; for instance, in the PI setting, the complexity guarantee for the proximal sampler reads $O(\kappa d^2 \log(1/\varepsilon))$ which is a major improvement w.r.t. every problem parameter. For strongly log-concave targets, the complexity is $O(\kappa d \log(1/\varepsilon))$.

In the step size regime $h = \Theta(\frac{1}{\beta d})$ used for these results, the proximal sampler is indeed directly comparable to LMC (which also uses step size scaling as $1/d$ w.r.t. the dimension d), and its interpretation as a proximal discretization is satisfying. However, even beyond this regime, the proximal sampler is a strikingly powerful algorithmic framework for designing faster samplers. Taking $h = \Theta(\frac{1}{\beta})$, implementation of the RGO amounts to sampling from a certain log-concave distribution with $O(1)$ condition number[2] to high accuracy (the latter requirement arises to prevent accumulation of errors from inexact implementation of the RGO). Crucially, this is true assuming only that V is β-smooth. The results of §4 therefore provide a general reduction of the task of sampling under various assumptions (e.g., PI and LSI) to the task of *high-accuracy well-conditioned log-concave sampling*, which will be profitably exploited in §6.

§5: Analysis of MALA from a warm start. As discussed above, through the proximal sampler reduction, the problem of high-accuracy log-concave sampling takes on special importance, and the next two chapters are dedicated to this problem.

A standard method for obtaining a high-accuracy sampler is to start with an proposal kernel Q and to accept or reject proposed moves according to a Metropolis–Hastings filter [Met+53; Has70]. When we apply this recipe with the proposal kernel taken to be one step of LMC, we arrive at the Metropolis-adjusted Langevin algorithm (MALA) [Bes+95], which remains quite popular in practice:

[2]The condition number of a distribution $\pi \propto \exp(-V)$ is the ratio between the smoothness and strong convexity parameters of V.

1. Propose $Y_{k+1} \sim Q(X_k, \cdot) = \mathsf{normal}(X_k - h\,\nabla V(X_k), 2h\,I_d)$.
2. Accept the proposal (i.e., set $X_{k+1} = Y_{k+1}$) with prob. $1 \wedge \frac{\pi(Y_{k+1})\,Q(Y_{k+1}, X_k)}{\pi(X_k)\,Q(X_k, Y_{k+1})}$; otherwise, set $X_{k+1} = X_k$.

The non-asymptotic analysis of MALA was carried out in [Dwi+19; Che+20a; LST20], yielding a complexity of $\widetilde{O}(\kappa d\,\mathrm{polylog}(1/\varepsilon))$ in the well-conditioned log-concave setting.[3] Suppressing the dependence on ε (which is always $\mathrm{polylog}(1/\varepsilon)$ for a high-accuracy sampler, by definition of "high accuracy"), the complexity of $\widetilde{O}(\kappa d)$ has also constituted a barrier for other high-accuracy samplers, such as the Metropolized random walk (MRW) and Metropolized Hamiltonian Monte Carlo (MHMC) [Che+20a]. In this chapter, we break this barrier for the first time by improving the complexity of MALA to $\widetilde{O}(\kappa\sqrt{d})$ (in the regime of small κ), under the additional assumption of a *warm start*: an initialization for the algorithm with $O(1)$ Rényi divergence from the target π.

To achieve this result, we introduce a new analysis technique for Metropolis-adjusted chains based on a projection characterization of the Metropolis–Hastings filter [BD01], which reduces the computation of the acceptance probability to a Girsanov discretization argument similarly to the one carried out in §3. We complement our results with a lower bound (later refined in [WSC22]) showing that our complexity bound under a warm start is tight.

The main drawback of this result is the need for a warm start. As shown in [LST21a], this issue is fundamental rather than merely technical because the complexity of MALA is $\widetilde{\Omega}(\kappa d)$ without this warm start. Hence, in the next chapter, we focus on the question of algorithmically obtaining a warm start for MALA.

§6: Algorithmic warm starts for MALA. A natural approach to obtaining the warm start is to use a *low-accuracy* sampler. For instance, we consider the underdamped Langevin diffusion, which is thought to be the sampling analogue of Nesterov's accelerated gradient flow (although the acceleration phenomenon $\kappa \mapsto \sqrt{\kappa}$ has not yet been established for log-concave sampling):

$$\begin{aligned} \mathrm{d}X_t &= P_t\,\mathrm{d}t\,, \\ \mathrm{d}P_t &= -\nabla V(X_t)\,\mathrm{d}t - \gamma P_t\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t\,. \end{aligned}$$

The use of this diffusion is well-motivated: it was shown in [DR20] that once discretized, the underdamped Langevin Monte Carlo (ULMC) algorithm enjoys a complexity guarantee of $\widetilde{O}(\kappa^{3/2}d^{1/2}/\varepsilon)$, in the *Wasserstein* metric. For the

[3] Note that the proximal sampler with rejection sampling implementation of the RGO already matches this guarantee.

purposes of a warm start, as we discuss in §6 it is critical to obtain a guarantee in the stronger *Rényi* divergence. Nevertheless, this result provides hope that if the convergence guarantee of ULMC can be "upgraded" to hold in Rényi, then the overall complexity of MALA with a warm start procured via ULMC would be $\widetilde{O}(\kappa^{3/2}d^{1/2} + \kappa d^{1/2}\operatorname{polylog}(1/\varepsilon))$

This distinction between Wasserstein and Rényi may at first appear innocuous, but is in fact deeper than it seems. Arguments in the Wasserstein metric are are considerably simplified through the use of coupling methods, and once we move to Rényi, we quickly run into long-standing challenges in the analysis of hypocoercive partial differential equations [Vil09a]. Here, our main innovation is the extension and judicious use of the shifted divergence method, a technique which originated in the literature on differential privacy [Fel+18] and recently applied to sampling in [AT22b]. Together with a suitable adaptation of the Girsanov argument of §3, we establish the desired Rényi divergence guarantees for ULMC.

Hence, ULMC provides a warm start for MALA, and together it yields a faster high-accuracy log-concave sampler than was known before. When we further feed this into the proximal sampler reduction, it sharpens the dependence on κ, leading to the current state-of-the-art complexity of $\widetilde{O}(\kappa\sqrt{d}\operatorname{polylog}(1/\varepsilon))$ for high-accuracy log-concave sampling.[4] As discussed above, the proximal sampler reduction also furnishes state-of-the-art results under more general assumptions, such as for targets satisfying a PI or LSI.

§7: **Lower bound in one dimension.** Thus far, we have focused on improved algorithmic guarantees for sampling, which provide upper bounds on the complexity of this task. As put forth in [NY83], however, a true understanding of this complexity also requires matching lower bounds which chart fundamental limitations shared by all potential algorithms. The problem of establishing such sampling lower bounds is extremely nascent, and we in fact found no prior work which directly address this question for log-concave sampling (although there have been several related approaches, see §7 for a discussion).

In this chapter, we establish the first lower bound for log-concave sampling. Our main result shows that the query complexity of sampling from densities $\pi \propto \exp(-V)$, where $V : \mathbb{R} \to \mathbb{R}$ is a univariate potential minimized at 0 and satisfying $1 \le V'' \le \kappa$, is $\Theta(\log\log\kappa)$. Despite being restricted to univariate distributions, and therefore falling short of capturing the dimension dependence of sampling which is of central interest, our work provides important insights towards further progress on the lower bound problem. In particular, our lower bound construction demonstrates the effectiveness of information-theoretic techniques

[4] The same complexity was arrived at concurrently and independently in [FYC23] via an approximate rejection sampling implementation of the RGO.

for this question. The upper bound is achieved via a tailored rejection sampling algorithm and, similarly to Nesterov's accelerated gradient method, was only found due to the presence of the lower bound. It cannot be achieved by existing MCMC samplers and serves as a reminder that the optimality of our existing algorithms remains ever in question without a theory of lower bounds.

In subsequent work [Che+23b], we make further progress by settling the complexity of log-concave sampling in any fixed dimension, and for the subclass of Gaussian distributions, although we omit these results from the thesis.

Constrained sampling and Brascamp–Lieb

In this next part of the thesis, we study *mirror Langevin Monte Carlo* (MLMC), which is the sampling analogue of the *mirror descent* algorithm for optimization, and which can be used for sampling from distributions with compact support (i.e., *constrained* sampling), as well as poorly conditioned distributions.

§8: **Continuous-time analysis of the mirror Langevin diffusion.** We begin with a study of the mirror Langevin diffusion in continuous time. The mirror Langevin diffusion is determined by the potential V of the target distribution $\pi \propto \exp(-V)$, as well as the choice of a *mirror map* $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, which is a convex function that determines the geometry of the algorithm. The diffusion is the solution $(X_t)_{t \geq 0}$ to

$$Y_t \coloneqq \nabla\phi(X_t)\,, \qquad \mathrm{d}Y_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,[\nabla^2\phi(X_t)]^{1/2}\,\mathrm{d}B_t\,. \tag{1.2}$$

Our main observation is that provided that V is *relatively convex* w.r.t. ϕ, that is, $\nabla^2 V \succeq \alpha\,\nabla^2\phi$ for some $\alpha > 0$, then a well-known functional inequality, the *Brascamp–Lieb inequality*, furnishes a spectral gap for the mirror Langevin diffusion, and hence the diffusion converges rapidly to its stationary distribution π. The notion of relative convexity is well-motivated from the convex optimization literature [BBT17; LFN18].

In particular, when V is strictly convex and we choose $\phi = V$, we arrive at the sampling analogue of *Newton's method* from optimization; we refer to the specialization of (1.2) to this case as the *Newton–Langevin diffusion*. Here, the relative convexity condition trivially holds with $\alpha = 1$, and consequently, the Newton–Langevin diffusion converges to stationarity exponentially fast with a rate that is independent of the conditioning of the problem and the dimension. This is reminiscent of the affine invariance of Newton's method.

In this chapter, we perform numerical experiments to demonstrate the potential applicability of this diffusion, and we leave the question of obtaining discretization bounds to the next chapter.

§9: **Discretization analysis of mirror Langevin Monte Carlo.** In this chapter, we take up the question deferred from the preceding one: how do we obtain non-asymptotic convergence guarantees for MLMC, in the spirit of part I of the thesis? This question is considerably more difficult than the corresponding one for LMC, stemming from the use of a non-isotropic and spatially dependent diffusion matrix in (1.2). Consequently, the prior work [Zha+20] did not obtain a satisfactory discretization result for the mirror Langevin diffusion, since their discretization bounds do not vanish even as the step size h of the discretization is taken to zero—hence, it is not possible to achieve any desired target accuracy ε from their results.

In [Zha+20], the authors considered the standard Euler–Maruyama discretization of (1.2). In our work, we take a different approach and propose a modified discretization in which we assume that the "null" mirror diffusion (that is, the diffusion (1.2) with the potential V set to zero) can be exactly simulated. This is a stringent assumption that limits the applicability of our results in practice, but it is natural within the oracle model because the "null" mirror diffusion can be simulated without making additional queries to the potential V. We can view the situation as follows: even for mirror descent algorithm in optimization, one must assume that the mirror map is simple enough so that basic operations (e.g., computing $\nabla\phi$ and $\nabla\phi^\star$) can be carried out. In the sampling setting, we require another algorithmic primitive involving the mirror map, namely, the simulation of the "null" mirror diffusion.

Once this assumption is made, however, we show that a clean analysis of MLMC can be carried out following the proof technique of [DMM19]. Appealingly, our analysis only requires assumptions on ϕ which are natural from the standpoint of convex analysis: relative convexity and smoothness of V w.r.t. ϕ, relative Lipschitzness of the gradient of V w.r.t. ϕ, and self-concordance of ϕ. We obtain discretization guarantees which recover state-of-the-art guarantees for LMC as a special case, and which avoids the aforementioned issue of carrying a bias term which does not vanish as $h \searrow 0$.

§10: **Interlude: two applications of Brascamp–Lieb inequalities.** We pause our discussion of sampling in order to explore two interesting consequences of the Brascamp–Lieb inequality which drives the convergence of mirror Langevin.

In the first application, we resolve an open question of [BE19] by showing that the entropic barrier for a convex body, which is known to be a self-concordant barrier for that body, in fact attains the optimal barrier parameter of d, where d is the ambient dimension. Self-concordant barriers are the cornerstone of interior-point methods in structured optimization [NN94], and the question of obtaining optimal and universal self-concordant barriers has been a long-standing one in that field. Our proof shows that the optimality of the entropic barrier is a direct

consequence of a certain dimensional refinement of the Brascamp–Lieb inequality.

Next, we give a new proof of Caffarelli's celebrated theorem on contractive properties of the optimal transport map [Caf00]; namely, the optimal transport map from a β-log-smooth distribution to an α-strongly log-concave distribution is $\sqrt{\beta/\alpha}$-Lipschitz. Our proof in fact provides an extension of this result to the *entropic* optimal transport map, thereby recovering Caffarelli's original result as the entropic regularization tends to zero. Key to our proof is the representation of the Hessians of the entropic optimal dual potentials as covariance matrices, to which we can apply a dual pair of covariance inequalities: the Brascamp–Lieb inequality and the Cramér–Rao inequality.

Optimization and sampling without convexity

§11: Dimension-free log-Sobolev inequalities for mixtures. We next aim to study sampling from non-log-concave distributions, which arise commonly in difficult but practical inference problems. A standard approach to obtaining sampling guarantees in this setting is to assume that the target distribution satisfies a functional inequality such as an LSI, as was done in §3, §4, and §6. The LSI is a flexible assumption which covers a wide range of non-log-concave distributions.

One barrier to pursuing this approach is that, surprisingly, the LSI constant is not tightly characterized even for the canonical non-log-concave example of a Gaussian mixture. In particular, it was an open question of [Bar+18] to show that the convolution of a measure with compact support and a Gaussian satisfies an LSI with a dimension-free constant (depending only on the radius of the support and the variance of the Gaussian). In this chapter, we resolve this question by proving a rather general result on the LSI constant of a mixture. Since the LSI arises as a property of keen interest throughout high-dimensional probability, we believe this result will be broadly useful.

§12: Lower bounds for stationary points in optimization. Although the LSI yields guarantees for non-log-concave sampling, they are (unavoidably) poor because the LSI constant typically scales exponentially in important problem parameters. This is a manifestation of the fact that non-log-concave sampling is, in the worst case, computationally hard. The same situation arises in the analogous field of non-convex optimization, but in that setting there is a general and well-developed theory on polynomial complexity bounds for obtaining approximate first-order *stationary points*, which is the best goal to which we can strive in such generality.

However, even in the mature setting of optimization in which the optimal complexity of finding stationary points is well-understood in the high-dimensional regime [Car+20] (and moreover attained by gradient descent), there remain im-

portant unresolved questions about the *low-dimensional* complexity of finding stationary points. This question is motivated by the question of whether there exists a cutting-plane method for optimization, the resolution to which would deepen our understanding of the limitations of non-convex optimization.

In this chapter, we make a contribution in this direction by tightly characterizing this complexity in dimension one, in four settings based on whether we consider deterministic or randomized algorithms, and whether or not the oracle returns zeroth-order information. One of the surprises of our findings is that gradient descent is already optimal in dimension one among deterministic algorithms using strictly first-order information, whereas this was previously only known in dimension $\widetilde{\Omega}(1/\varepsilon^2)$.

§13: Sampling upper bounds in the Fisher information metric. Motivated by the theory of stationary points in non-convex optimization, in this chapter we develop a theory of approximate first-order stationarity for non-log-concave sampling. Taking inspiration from the interpretation of sampling as minimizing the KL divergence functional over the space of probability measures endowed with the Wasserstein geometry, we take as our definition of an ε-stationary a point for which the norm of the Wasserstein gradient of the KL divergence is at most ε. This corresponds to a relative Fisher information bound $\mathsf{FI}(\mu \,\|\, \pi) \le \varepsilon^2$.

We provide an interpretation of this criterion in terms of the classical notion of metastability of diffusions. Moreover, mirroring the corresponding result for non-convex optimization, we show that under the sole assumption of log-smoothness of π, averaged LMC attains an ε-stationary point in polynomially many queries. As an interesting corollary, it implies an $\widetilde{O}(d^2/\varepsilon^4)$ iteration complexity bound for averaged LMC to reach ε total variation error when π satisfies a PI, which can be compared to the results in §3. Overall, our definition of approximate first-order stationarity for sampling is the foundation for a novel framework for studying the complexity of non-log-concave sampling which allows for quantitative comparisons between algorithms, as done in the next chapter.

§14: Sampling lower bounds in the Fisher information metric. Complementing the results of the previous chapter, here we obtain *lower bounds* on the complexity of reaching ε-stationarity in sampling. As discussed above, the theory of lower bounds for sampling is underdeveloped at present, and we view this as a promising step in this direction.

Among the results of this chapter, we highlight a surprising reduction of non-log-concave sampling to finding stationary points in non-convex optimization in a certain regime for the Fisher information. In this regime, it implies that the upper bound obtained for averaged LMC in the preceding chapter is *optimal*, whereas optimality of LMC was not previously known in any setting.

Other applications of Wasserstein gradient flows

In the last part of the thesis, we adopt a broader outlook and develop further applications of Wasserstein gradient flows to Wasserstein barycenters, variational inference, and diffusion models.

§15: **Bures–Wasserstein barycenters.** In this chapter, we consider the algorithmic problem of averaging Gaussian distributions in the optimal transport metric. Since this is an optimization problem with an intrinsic geometric structure, namely the Wasserstein geometry over the space of Gaussians (called the Bures–Wasserstein geometry), it is natural to consider Riemannian gradient algorithms for its solution. In this chapter, we provide the first non-asymptotic convergence guarantees for such algorithms. In doing so, we develop machinery for general Wasserstein barycenters (namely, a stability result in Theorem 15.3.3) as well as for optimization more generally over the Bures–Wasserstein space (which will be fruitfully employed in the next chapter).

§16: **Gaussian variational inference.** Next, we consider *variational Bayes*, which has recently emerged as a tractable alternative to MCMC sampling. In this approach, rather than using MCMC algorithms to sample from the posterior distribution $\pi \propto \exp(-V)$, we instead seek the best variational approximation of π from within a simpler class of distributions, hoping that this variational approximation is accurate enough to yield information about useful summary statistics of π, such as its mean and covariance. Here, we focus on *Gaussian variational inference*, in which the simpler class of distributions is taken to be the class of Gaussians, and the objective is to find a Gaussian p minimizing the KL divergence $\mathsf{KL}(p \,\|\, \pi)$.

Gaussian variational inference is naturally formulated as an optimization problem over the Bures–Wasserstein space, and in doing so we identify the Wasserstein geometry over Gaussians as a canonical one for this problem. This is justified because, as we show in §16, the objective of Gaussian variational inference is convex as soon as V is. Consequently, by discretizing the Bures–Wasserstein gradient flow of $\mathsf{KL}(\cdot \,\|\, \pi)$, we arrive at a principled algorithm for variational inference for which we can establish non-asymptotic convergence guarantees.

We can also extend our methodology to variational inference with the more flexible class of *Gaussian mixtures*, using the geometry introduced in [CGT19; DD20], albeit with a corresponding loss of theoretical guarantees. Nevertheless, our algorithm for mixtures of Gaussians (and a more flexible variant thereof which allows for time-varying weights via the Wasserstein–Fisher–Rao geometry) yields encouraging results in experiments, providing a proof of concept in favor of our geometrically motivated approach.

§17: **Theory for diffusion models.** We end this thesis with a theoretical study of *diffusion models* or *score-based generative models* (SGMs), which have achieved state-of-the-art performance for generative modelling. Similarly to the proximal sampler studied in §4, diffusion models are based on the idea of running a stochastic process forwards and backwards in time. However, here we take the forward process to be the Ornstein–Uhlenbeck (OU) process, which is known to converge exponentially rapidly to the standard Gaussian measure, and we run the process for a long time such that the resulting algorithm converges to the distribution of interest in one iteration. The challenge, then, is to find a tractable implementation of the backwards diffusion.

Unlike the results of the previous chapters, we depart from the oracle model and instead assume that we have access to L^2-accurate estimates of the *score functions* (the gradients of the log densities) along the forward process. In practice, such estimates are obtained by training a deep neural network with a score matching objective [Hyv05] using samples from the target distribution (e.g., a database of natural images). Under this assumption on score estimation error, the main result of this chapter is that diffusion models can sample from essentially arbitrary distributions (including highly non-log-concave distributions or distributions supported on lower-dimensional subsets) with polynomial complexity.

The catch, of course, is that it is unclear when the assumption of L^2-accurate score estimation is verified in practice, since this requires an understanding of the generalization performance of neural network training that is currently out of reach. Nevertheless, our result provides a principled justification for the use of diffusion models and points towards the importance of going beyond the oracle model in order to fully tackle the difficult task of non-log-concave sampling.

Chapter 2

Background

2.1 Background on optimal transport

We recall here basic background and notation on optimal transport and refer the reader to [Vil03; AGS08; Vil09b; San15] for more details.

2.1.1 Optimal transport costs

Wasserstein distance. Given a Polish space (E, d), we denote by $\mathcal{P}_2(E)$ the collection of all (Borel) probability measures μ on E such that $\mathbb{E}_{X\sim\mu}[\mathsf{d}(X,y)^2] < \infty$ for some $y \in E$. For two measures $\mu, \nu \in \mathcal{P}_2(E)$, let $\mathcal{C}(\mu,\nu)$ be set of couplings between μ and ν, that is, the collection of probability measures γ on $E \times E$ such that if $(X,Y) \sim \gamma$, then $X \sim \mu$ and $Y \sim \nu$.

Definition 2.1.1. *Given two probability measures $\mu, \nu \in \mathcal{P}_2(E)$, the* 2-Wasserstein distance *between μ and ν is*

$$W_2^2(\mu,\nu) \coloneqq \inf_{\gamma\in\mathcal{C}(\mu,\nu)} \int \mathsf{d}(x,y)^2 \,\mathrm{d}\gamma(x,y)\,. \tag{2.1}$$

We are primarily interested in the case when $E = \mathbb{R}^d$ equipped with the standard Euclidean metric. Thus, $\mathcal{P}_2(\mathbb{R}^d)$ denotes the space of probability measures on $\mathbb{R}^d$ with finite second moment, and $\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ denotes the space of measures P on $\mathcal{P}_2(\mathbb{R}^d)$ such that $\mathbb{E}_{\nu\sim P}\, W_2^2(\mu_0,\nu) < \infty$ for some, and therefore any, $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. If $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is absolutely continuous w.r.t. the Lebesgue measure, we write $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, and we similarly define the space $\mathcal{P}_2(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d))$.

Transport map. Given a measure μ and a map $T\colon \mathbb{R}^d \to \mathbb{R}^d$, the pushforward $T_{\#}\mu$ is the law of $T(X)$ when $X \sim \mu$.

Theorem 2.1.2 (Fundamental theorem of optimal transport)**.** *Suppose that $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. Then, the unique optimal transport plan $\gamma^\star$ (i.e., the minimizer in (2.1)) is induced by a transport map $T_{\mu\to\nu}$, in the sense that if $X \sim \mu$, then $(X, T_{\mu\to\nu}(X)) \sim \gamma^\star$ (this is known as* Brenier's theorem*).*

Moreover, $T_{\mu\to\nu}$ is characterized as the (μ-a.e. unique) gradient of a convex proper lower semicontinuous function $\phi_{\mu\to\nu} : \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ such that $(\nabla\phi_{\mu\to\nu})_{\#}\mu = \nu$. We refer to $\phi_{\mu\to\nu}$ as the Kantorovich potential*, and it is the solution to the dual optimal transport problem*

$$\frac{1}{2}W_2^2(\mu,\nu) = \sup_{\substack{\phi:\mathbb{R}^d\to\mathbb{R}\cup\{\infty\}\\ \text{convex proper LSC}}} \Big\{\int \big(\frac{\|\cdot\|^2}{2} - \phi\big)\,\mathrm{d}\mu + \int \big(\frac{\|\cdot\|^2}{2} - \phi^*\big)\,\mathrm{d}\nu\Big\}.$$

For $\alpha,\beta > 0$, if $\phi_{\mu\to\nu}$ is α-strongly convex and β-smooth, in the sense that for all $x, y \in \mathbb{R}^d$,

$$\frac{\alpha}{2}\|y-x\|^2 \le \phi_{\mu\to\nu}(y) - \phi_{\mu\to\nu}(x) - \langle\nabla\phi_{\mu\to\nu}(x), y-x\rangle \le \frac{\beta}{2}\|y-x\|^2, \qquad (2.2)$$

then we say that the potential $\varphi_{\mu\to\nu}$ is (α,β)-*regular*.

Metric and topological properties. The space $\mathcal{P}_2(\mathbb{R}^d)$ endowed with the 2-Wasserstein distance is a complete separable metric space, i.e., a Polish space. Convergence in the W_2 metric ($W_2(\mu_n,\mu)\to 0$) is equivalent to weak convergence and convergence of the second moment (i.e., $\int f\,\mathrm{d}\mu_n \to \int f\,\mathrm{d}\mu$ for all bounded continuous functions $f:\mathbb{R}^d\to\mathbb{R}$, and $\int\|\cdot\|^2\,\mathrm{d}\mu_n \to \int\|\cdot\|^2\,\mathrm{d}\mu$).

The 2-Wasserstein metric is useful because it lifts the geometry of $\mathbb{R}^d$ to the space $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures over $\mathbb{R}^d$; for example, the mapping $x\mapsto\delta_x$ is an isometric embedding of $\mathbb{R}^d$ into $\mathcal{P}_2(\mathbb{R}^d)$. As we discuss shortly, this geometry is particularly important because it admits a calculus (known as *Otto calculus*) which allows for a geometric study of dynamics on $\mathcal{P}_2(\mathbb{R}^d)$.

Extension to other costs. More generally, the theory of optimal transport can be fruitfully developed in the following abstract setting: E_1, E_2 are Polish spaces and $c : E_1\times E_2 \to \mathbb{R}\cup\{\infty\}$ is a lower semicontinuous cost function; the optimal transport cost is $\inf_{\gamma\in\mathcal{C}(\mu,\nu)}\int c(x,y)\,\mathrm{d}\gamma(x,y)$. The infimum is always realized by an optimal transport plan. The corresponding dual problem is to maximize the objective $\int f\,\mathrm{d}\mu + \int g\,\mathrm{d}\nu$ over pairs $(f,g)\in L^1(\mu)\times L^1(\nu)$ such that $f(x)+g(y)\le c(x,y)$ (for $\mu\otimes\nu$-a.e. $x,y\in E_1\times E_2$). Strong duality holds (the optimal transport cost equals the value of the dual problem), and the maximizers in the dual problem also characterized by the notion of *c-concavity* which generalizes the usual notion of convexity. Some key examples include:

- When $E_1 = E_2 = \mathbb{R}^d$ and $c(x,y) = \|x-y\|^p$ for some $p\ge 1$, then the corresponding optimal transport cost is the p-th power of the *p-Wasserstein distance* $W_p(\mu,\nu)$.

- When $E_1 = E_2$ and $c(x, y) = \mathbb{1}\{x \neq y\}$ then the corresponding optimal transport cost is the *total variation distance* $\|\mu - \nu\|_{\mathsf{TV}}$.
- In §9, we also make use of *Bregman coupling costs*.

2.1.2 Riemannian geometry

In this section we give a brief exposition to Riemannian geometry. We refer readers to [Car92] for a standard introduction.

An n-dimensional manifold M is a topological space which is Hausdorff, second countable, and locally homeomorphic to $\mathbb{R}^n$. A smooth atlas is a collection of smooth charts $\{\psi_\alpha\}_{\alpha \in \mathcal{A}}$ so that each $\psi_\alpha \colon U_\alpha \subset M \to \mathbb{R}^n$ is a homeomorphism from an open set U_α in M, $M = \bigcup_{\alpha \in \mathcal{A}} U_\alpha$, and such that for all $\alpha, \alpha' \in \mathcal{A}$, $\psi_\alpha \circ \psi_{\alpha'}^{-1}$ is smooth wherever defined. For a fixed choice of smooth atlas, we declare a function $f \colon M \to \mathbb{R}$ to be smooth if $f \circ \psi_\alpha^{-1}$ is for each $\alpha \in \mathcal{A}$. The manifold together with a smooth atlas defines a smooth n-dimensional manifold, and we shall always suppress mention of the atlas. A map $f \colon M \to N$ between two smooth manifolds is said to be smooth if its composition with smooth charts is.

Given a smooth n-dimensional manifold M and a point $p \in M$, the tangent space T_pM is the equivalence class of all smooth curves $\gamma \colon (-\varepsilon, \varepsilon) \to M$ such that $\gamma(0) = p$, where two such curves γ_0, γ_1 are equivalent if, with respect to every coordinate chart ψ defined in a neighborhood of p, $(\psi \circ \gamma_0)'(0) = (\psi \circ \gamma_1)'(0)$. As such, T_pM is a real n-dimensional vector space for each $p \in M$. The cotangent space at $p \in M$ is then the dual to T_pM, which we shall denote T_p^*M. The tangent bundle is the disjoint union $TM \coloneqq \bigsqcup_{p \in M} T_pM$, and the cotangent bundle is similarly the disjoint union $T^*M \coloneqq \bigsqcup_{p \in M} T_p^*M$. The smooth structure on M induces a smooth structure on TM and T^*M, so each is then a $2n$-dimensional smooth manifold in its own right.

A smooth vector field $X \colon M \to TM$ is then a smooth map $p \mapsto X_p$ such that $X_p \in T_pM$ for all $p \in M$, and similarly for a smooth covector field $\alpha \colon M \to T^*M$. Higher-order tensors are defined similarly: a (p, q)-tensor field is a smooth mapping $T \colon M \to (TM)^p \otimes (T^*M)^q$. The differential $df \colon M \to T^*M$ of a smooth function f on M is the smooth covector field such that $df_p \colon T_pM \to \mathbb{R}$ obeys $df_p(v) \coloneqq (f \circ \gamma)'(0)$, where γ is any curve with tangent vector $v \in T_pM$ at $\gamma(0) = p$.

A Riemannian manifold (M, g) is a smooth n-dimensional manifold M with a smooth metric tensor $g \colon M \to T^*M \otimes T^*M$; at each point of M, this is a positive definite bilinear form. The metric tensor therefore defines a smoothly varying choice of inner product on the tangent spaces of M. In addition to giving rise to notions of length and geodesics, the metric tensor provides a canonical isomorphism (the Riesz isomorphism) between the tangent space and cotangent

space: for a vector $v \in T_pM$ the covector $\alpha_v \in T_p^*M$ is defined by $\alpha_v(w) = g_p(v, w)$. For a covector $\alpha \in T_p^*M$ the vector $v_\alpha \in T_pM$ is defined as the unique solution of $\alpha(w) = g_p(v_\alpha, w)$ for all $w \in T_pM$. A smooth vector field X can be accordingly transformed into a smooth covector field denoted $X^\flat$, and a smooth covector field ω can be transformed into a smooth vector field $\omega^\#$. The gradient of a function $f\colon M \to \mathbb{R}$ is defined then as $\nabla f \coloneqq (df)^\#$: in other words, for all $p \in M$ and $v \in T_pM$, $df_p(v) = g_p(\nabla f(p), v)$.

We typically write $\langle \cdot, \cdot \rangle_p$ instead of $g_p(\cdot, \cdot)$, and we write $\|\cdot\|_p$ for the norm induced by the metric tensor, i.e., $\|v\|_p \coloneqq \sqrt{\langle v, v \rangle_p}$. In this notation, the distance between points $p, q \in M$ is defined as

$$d_M(p, q) \coloneqq \inf_{\gamma \in \Gamma(p,q)} \int_0^1 \|\gamma'(t)\|_{\gamma(t)} \, \mathrm{d}t \,,$$

where $\Gamma(p, q)$ is the collection of all smooth (or piecewise continuous) curves $\gamma\colon [0, 1] \to M$ such that $\gamma(0) = p$ and $\gamma(1) = q$. If M is connected, then the distance d_M is indeed a metric. If we additionally assume that (M, d_M) is complete as a metric space then by the Hopf–Rinow theorem the value of the above minimization problem is attained by at least one curve $\gamma\colon [0, 1] \to M$ such that $t \mapsto \|\gamma'(t)\|_{\gamma(t)}$ is constant, which is said to be a constant-speed geodesic.

For any $p \in M$, there always exists an $\varepsilon > 0$ such that for any vector $v \in T_pM$ with $\|v\|_p < \varepsilon$, there is a unique constant-speed geodesic $\gamma_v\colon [0, 1] \to M$ obeying $\gamma_v(0) = p$ and $\gamma_v'(0) = v$.[1] On the ball $B_\varepsilon(0)$ with radius ε and center $0 \in T_pM$ (with respect to the norm $\|\cdot\|_p$), we can now define the exponential map $\exp_p\colon B_\varepsilon(0) \to M$ by $v \in V_p \mapsto \gamma_v(1)$. The exponential map is a diffeomorphism onto its image, so we can define the inverse mapping $\log_p\colon \exp_p(B_\varepsilon(0)) \to T_pM$. If M is complete, the domain of definition of any constant-speed geodesic $\gamma\colon [0, 1] \to \mathbb{R}$ can be extended to all of $\mathbb{R}$ such that at each time γ is locally a constant-speed minimizing geodesic; in this case, the exponential mapping can be extended to a mapping $\exp_p\colon T_pM \to M$. Note, however, that the mapping $\log_p$ is not necessarily defined everywhere.

We lastly recall that for fixed $q \in M$ and p which does not belong to the cut locus of q (the set of points for which there exists more than one constant-speed minimizing geodesic from p),

$$[\nabla d_M^2(\cdot, q)](p) = -2 \log_p(q) \,. \tag{2.3}$$

[1] In fact, a stronger result holds: there exists a neighborhood U of p such that for any two points $q, q' \in U$, there is a unique constant-speed minimizing geodesic $\gamma\colon [0, 1] \to U$ joining q to q'. Such a neighborhood is called a totally normal neighborhood of p.

This statement has an intuitive meaning: it says that outside of the cut locus of q, the gradient of the squared distance points in the direction of maximum increase.[2]

2.1.3 Riemannian interpretation of Wasserstein space

In this section, we briefly explain the interpretation of the Wasserstein space of probability measures as a Riemannian manifold. This interpretation is motivated by the connection between dissipative evolution equations and the theory of gradient flows on the Wasserstein space, first discovered in [Ott98]; subsequently, this link was further developed and strengthened in the seminal works [JKO98; Ott01]. Although the paper [JKO98] chronologically precedes [Ott01], the intuition of the former is based heavily on the work of Otto in the latter paper, in which he develops the formal[3] rules governing the calculus which now bears his name. For more introductory expositions of this subject, we refer to [Vil03, §8] and [San15, §5]. The task of putting this formal discussion on rigorous footing is undertaken in [AGS08, §8]. We also note that the Wasserstein space is a length space within the framework of metric geometry; see [BBI01] for an introduction to this approach.

Otto calculus endows the space $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ with a formal Riemannian structure inspired by fluid dynamics. To describe the idea, suppose that $(\mu_t)_{t\geq 0}$ is a curve of probability measures, with μ_t representing the fluid density at time t. Also, let $(v_t)_{t\geq 0}$ denote the velocity vector fields governing the dynamics of the particles; this means that the trajectory $t \mapsto X_t$ of an individual particle evolves according to the ODE

$$\dot{X}_t = v_t(X_t)\,. \tag{2.4}$$

In probabilistic language, if X_0 is a random variable drawn from the density μ_0 and it evolves according to (2.4), then $X_t \sim \mu_t$ for all $t \geq 0$. From this, we can derive a partial differential equation (PDE) governing the evolution of $(\mu_t)_{t\geq 0}$ as follows: fix a test function $\psi : \mathbb{R}^d \to \mathbb{R}$ (which is bounded, smooth, etc.). Formally, if the integration by parts is justified, then

$$\begin{aligned}\partial_t \int \psi \,\mathrm{d}\mu_t = \partial_t\, \mathbb{E}\, \psi(X_t) = \mathbb{E}\, \partial_t \psi(X_t) = \mathbb{E}\langle \nabla\psi(X_t), v_t(X_t)\rangle \\ = \int \langle \nabla\psi, v_t\rangle \,\mathrm{d}\mu_t = -\int \psi \,\mathrm{div}(v_t\mu_t)\,,\end{aligned}$$

[2] When there are multiple constant-speed minimizing geodesics joining p to q, then the following fact is still true: the squared distance function $d_M^2(\cdot, q)$ is superdifferentiable at p. Moreover, for any constant-speed minimizing geodesic $\gamma\colon [0,1] \to M$ joining p to q, the vector $-2\gamma'(0) \in T_pM$ is a supergradient of $d_M^2(\cdot, q)$ at p.

[3] Here, "formal" is not a synonym for "rigorous".

from which we deduce the *continuity equation* of fluid dynamics:

$$\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0\,. \tag{2.5}$$

This PDE can be interpreted in a suitable weak sense: for any smooth test function ψ with compact support, the mapping $t \mapsto \int \psi \,\mathrm{d}\mu_t$ should be absolutely continuous and thus differentiable at almost every $t \in [0,1]$, and its derivative should satisfy $\partial_t \int \psi \,\mathrm{d}\mu_t = \int \langle \nabla\psi, v_t\rangle \,\mathrm{d}\mu_t$.

Conversely, if $(\mu_t)_{t\geq 0}$ is a sufficiently nice curve, then it is always possible to find a family of vector fields $(v_t)_{t\geq 0}$ such that the equation (2.5) holds, i.e., we can interpret $(\mu_t)_{t\geq 0}$ as the evolution of a fluid density. Since the vector fields $(v_t)_{t\in[0,1]}$ fully govern the evolution of the curve of measures $(\mu_t)_{t\in[0,1]} \subseteq \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, we would like to equip $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ with the structure of a Riemannian manifold such that $(v_t)_{t\in[0,1]}$ is interpreted as the tangent vectors to the curve $(\mu_t)_{t\in[0,1]}$. However, a problem arises: given a curve $(\mu_t)_{t\in[0,1]}$ in Wasserstein space, there are many choices for the vector fields $(v_t)_{t\in[0,1]}$ which solve (2.5) together with $(\mu_t)_{t\in[0,1]}$. Indeed, if we fix any pair $(\mu_t)_{t\in[0,1]}$, $(v_t)_{t\in[0,1]}$ solving (2.5), then we obtain another solution by replacing v_t with $v_t + w_t$, where w_t is any vector field satisfying $\operatorname{div}(w_t\mu_t) = 0$. This motivates the search for a *distinguished* family of vector fields solving (2.5).

To do so, we pick v_t to minimize the *kinetic energy*,

$$v_t = \arg\min \left\{ \int \|w_t\|^2 \,\mathrm{d}\mu_t \;\middle|\; w_t : \mathbb{R}^d \to \mathbb{R} \;\text{ satisfies }\; \operatorname{div}(\mu_t w_t) = -\partial_t \mu_t \right\}.$$

If μ_t is regular (admits a density w.r.t. Lebesgue measure), then the minimum is attained at a gradient vector field: $v_t = \nabla\psi_t$ for a function $\psi_t : \mathbb{R}^d \to \mathbb{R}$. We are led to define the tangent space

$$T_\mu \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) = \overline{\{\nabla\psi \mid \psi \in \mathcal{C}_c^\infty(\mathbb{R}^d)\}}^{L^2(\mu)}$$

and endow it with the inner product

$$\langle v, w\rangle_\mu = \int \langle v, w\rangle \,\mathrm{d}\mu\,.$$

This yields a formal Riemannian structure on $\mathcal{P}_2(\mathbb{R}^d)$. Moreover, the choice of picking the vector field with minimal kinetic energy is closely related to the idea of optimal transport of mass. Indeed, Brenier's theorem asserts that in the optimal transport problem of transporting a measure ν_0 to another measure ν_1, the optimal transport plan is induced by a transport map, which is the gradient of a convex function ϕ. In other words, if we interpret ν_0 as a collection of particles, then

each particle initially moves along the vector field $\nabla\phi - \mathrm{id}$. In particular, taking $\nu_0 = \mu_0$ and $\nu_1 = \mu_\varepsilon$ for a small $\varepsilon > 0$, we expect the tangent vector of $(\mu_t)_{t\in[0,1]}$ at time 0 to be of the form $\nabla\phi - \mathrm{id}$ for a convex function ϕ. Therefore, it is equivalent to write [see AGS08, §8] that

$$T_\mu \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) := \overline{\{\lambda\,(\nabla\phi - \mathrm{id})\colon\ \lambda > 0,\ \phi \in \mathcal{C}_c^\infty(\mathbb{R}^d),\ \phi \text{ convex}\}}^{L^2(\mu)}\,.$$

To complete the story, [BB99] proved that

$$W_2(\mu_0,\mu_1) = \inf\Big\{ \int_0^1 \|v_t\|_{\mu_t}\,\mathrm{d}t \;\Big|\; (\mu_t)_{t\in[0,1]}, (v_t)_{t\in[0,1]} \text{ solve (2.5)} \Big\}\,. \tag{2.6}$$

From the lens of Riemannian geometry, this says that the notion of distance induced by the Riemannian structure is precisely the quadratic Wasserstein distance, and hence we refer to the space $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ equipped with this Riemannian structure as the *Wasserstein space*.

Geodesics and generalized geodesics. Given two measures $\mu_0, \mu_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, there is a unique constant-speed minimizing geodesic joining μ_0 to μ_1.

Definition 2.1.3. *Given $\mu_0, \mu_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, the* (constant-speed) geodesic *joining μ_0 to μ_1 is the curve*

$$t \mapsto [(1-t)\,\mathrm{id} + t\,T]_\# \mu_0\,, \qquad t \in [0,1]\,,$$

where T is the optimal transport mapping from μ_0 to μ_1. This is also known as displacement interpolation *or* McCann's interpolation.

This geodesic has the following interpretation: draw a "particle" $X_0 \sim \mu_0$, and move X_0 to $T(X_0)$ with constant speed for one unit of time along the Euclidean geodesic (i.e., straight line) joining these endpoints; thus, at time t, the particle is at position $X_t = (1-t)\,X_0 + t\,T(X_0)$. Then, μ_t is simply the law of X_t.

Let $T_t := (1-t)\,\mathrm{id} + t\,T$. Since $\dot{X}_t = T(X_0) - X_0 = (T - \mathrm{id}) \circ T_t^{-1}(X_t)$, then along the geodesic we see that $(\mu_t, v_t)_{t\in[0,1]}$ solves the continuity equation (2.5), where the vector field is $v_t = (T - \mathrm{id}) \circ T_t^{-1}$. This solution achieves the minimum in the variational problem (2.6).

The geodesic satisfies

$$W_2(\mu_0,\mu_t) = t\,W_2(\mu_0,\mu_1) \qquad \forall t \in [0,1]\,. \tag{2.7}$$

Moreover, it can be shown that any constant-speed geodesic in $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, that is, any curve $(\mu_t)_{t\in[0,1]} \subseteq \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ satisfying (2.7), is necessarily of the form $\mu_t = [(1-t)\,\mathrm{id} + t\,T]_\#\mu_0$. The tangent vector to $(\mu_t)_{t\in[0,1]}$ at time 0 is the vector

field $T - \mathrm{id}$. With this notion of geodesics, the space $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ becomes a geodesic space [BBI01].

We also define the notion of a generalized geodesic, which has been used to prove existence for the minimizing movements scheme for Wasserstein gradient flows [AGS08]. This notion also turns out to be quite useful for applications; see, e.g., §9 and §15.

Definition 2.1.4. *For any $\nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, we define the* generalized geodesic *with base ν and connecting μ_0 to μ_1 to be the curve $(\mu_s^\nu)_{s\in[0,1]}$ where*

$$\mu_s^\nu := [(1-s)\, T_{\nu\to\mu_0} + s\, T_{\nu\to\mu_1}]_{\#}\nu\,.$$

Observe that the notion of generalized geodesic reduces to that of a geodesic when $\nu = \mu_0$, so that convexity along generalized geodesic is a stronger notion than convexity along geodesics. We say that a set $\mathcal{C} \subset \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ is convex along geodesics (resp. generalized geodesics) if its indicator function $\iota_{\mathcal{C}}$ is convex along geodesics (resp. generalized geodesics). Note that $\mathcal{C}$ is convex along generalized geodesics with base b if and only if the set $\log_b(\mathcal{C})$ is convex in the usual sense.

Exponential and logarithmic maps. For any $b, b' \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, define the map $\log_b : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to T_b\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ by $\log_b(b') := T_{b\to b'} - \mathrm{id}$. Reciprocally, we define the map $\exp_b : U \to \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ in some neighborhood U of the origin of $T_b\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ by $\exp_b(v) = (\mathrm{id} + v)_{\#}b$.

In Riemannian geometry, it is common to localize the argument around a measure μ, which loosely means replacing a measure ν with its image $\log_\mu \nu$ in the tangent space at μ. This is convenient because the tangent space at μ is embedded in the Hilbert space $L^2(\mu)$, and we can leverage Hilbert space arguments (e.g., computing inner products). In order to do this one must quantify the distortion introduced by the map $\log_\mu$, which is morally related to curvature.

Convexity. We are now in a position to define two notions of convexity in Wasserstein space. Consider any functional $\mathcal{F} : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to (-\infty, \infty]$ defined over the Wasserstein space.

Definition 2.1.5. *Let $\alpha \in \mathbb{R}$. We say that $\mathcal{F}$ is* α-geodesically convex *if for all $\mu_0, \mu_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, the constant-speed geodesic $(\mu_s)_{s\in[0,1]}$ from μ_0 to μ_1 satisfies*

$$\mathcal{F}(\mu_s) \le (1-s)\,\mathcal{F}(\mu_0) + s\,\mathcal{F}(\mu_1) - \frac{\alpha\, s\,(1-s)}{2}\, W_2^2(\mu_0,\mu_1)\,, \qquad \text{for all } s \in [0,1]\,.$$

We say that $\mathcal{F}$ is α-convex along generalized geodesics *if for all choices $\nu, \mu_0, \mu_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, it holds that*

$$\mathcal{F}(\mu_s^\nu) \le (1-s)\,\mathcal{F}(\mu_0) + s\,\mathcal{F}(\mu_1) - \frac{\alpha\, s\,(1-s)}{2}\, W_2^2(\mu_0,\mu_1)\,, \qquad \text{for all } s \in [0,1]\,.$$

If we omit mention of the parameter α, then we refer to the case $\alpha = 0$.

Note that convexity along generalized geodesics is stronger than geodesic convexity since it requires $\mathcal{F}$ to be convex along a larger set of curves.

The interpretation of generalized geodesics is that we linearize $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ on the tangent space $T_\nu \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. This means that we replace μ_0 with its image $\log_\nu \mu_0 = T_{\nu\to\mu_0} - \mathrm{id}$ in the tangent space, and similarly for μ_1. Since the tangent space is a subset of a Hilbert space, geodesics in the tangent space are described by straight lines, i.e.,

$$t \mapsto (1-t)\, T_{\nu\to\mu_0} + t\, T_{\nu\to\mu_1} - \mathrm{id}\,.$$

If we translate back to $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, we end up with the curve

$$t \mapsto \exp_\nu\big((1-t)\, T_{\nu\to\mu_0} + t\, T_{\nu\to\mu_1} - \mathrm{id}\big) = [(1-t)\, T_{\nu\to\mu_0} + t\, T_{\nu\to\mu_1}]_{\#}\nu = \mu_t^\nu\,.$$

Thus, the property of being convex along generalized geodesics can be reformulated as requiring that

$$\mathcal{F} \circ \exp_\nu : T_\nu \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathbb{R} \qquad \text{is convex for every } \nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)\,. \tag{2.8}$$

In Euclidean space, convexity of a function $f : \mathbb{R}^d \to \mathbb{R}$ is equivalent, via Jensen's inequality, to the following statement: for every probability measure P on $\mathbb{R}^d$, it holds that $f(\int x \,\mathrm{d}P(x)) \le \int f(x)\,\mathrm{d}P(x)$. Since the Wasserstein barycenter (see §15) is the Wasserstein analogue of the mean, we can write a similar definition on Wasserstein space. Given a probability measure P on $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, let b_P denote its Wasserstein barycenter. We say that $\mathcal{F} : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathbb{R}$ is *convex along barycenters* if

$$\mathcal{F}(b_P) \le \int \mathcal{F}(\mu)\,\mathrm{d}P(\mu)\,, \qquad \text{for all } P \in \mathcal{P}_2\big(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)\big)\,.$$

Similarly, via (2.8), we can define $\mathcal{F} : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathbb{R}$ to be convex along generalized barycenters if

$$\mathcal{F} \circ \exp_\nu\Big(\int v\,\mathrm{d}P(v)\Big) \le \int \mathcal{F} \circ \exp_\nu(v)\,\mathrm{d}P(v) \qquad \text{for all } \nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \text{ and } P \in \mathcal{P}_2\big(T_\nu \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)\big)\,. \tag{2.9}$$

However, since the tangent space is embedded in a Hilbert space, there is no difference between (2.8) and (2.9).

To summarize the relationship between these four concepts:

$$\begin{aligned}\text{convex along generalized barycenters} &\iff \text{convex along generalized geodesics}\\ &\implies \text{convex along barycenters}\\ &\implies \text{geodesically convex}\,.\end{aligned}$$

For a justification of these facts and further discussion, see [AC11].

Curvature. We often use the fact that $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ is non-negatively curved in the sense of Alexandrov. More specifically, we use the fact that for $\mu_0, \mu_1, \nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, if $(\mu_s)_{s\in[0,1]}$ denotes the constant-speed geodesic connecting μ_0 to μ_1, then for all $s \in [0,1]$,

$$W_2^2(\mu_s,\nu) \geq (1-s)\,W_2^2(\mu_0,\nu) + s\,W_2^2(\mu_1,\nu) - s\,(1-s)\,W_2^2(\mu_0,\mu_1)\,. \tag{2.10}$$

Moreover, for any $\mu, \nu, b \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ it holds that

$$\begin{aligned}W_2(\mu,\nu) \leq \|T_{b\to\nu}\circ T_{\mu\to b} - \mathrm{id}\|_{L^2(\mu)} &= \|T_{b\to\nu} - T_{b\to\mu}\|_{L^2(b)}\\ &= \|\log_b(\mu) - \log_b(\nu)\|_b\,.\end{aligned} \tag{2.11}$$

We note that the use of terminology from Riemannian geometry can be justified when the measures are regular, see [AGS08]. For our purposes these analogies are merely employed for better readability and intuition.

JKO scheme. This formal picture already allows one to compute gradients of functionals defined over $\mathcal{P}_2(\mathbb{R}^d)$ and hence to consider gradient flows, as well as to derive criteria which imply quantitative rates of convergence for these flows. However, it is a considerable technical undertaking to make the preceding formal considerations fully rigorous, and this was only accomplished later in the comprehensive monograph [AGS08]. Instead, in [JKO98], the authors sidestep this difficulty by considering an implicit time-discretization scheme which only requires the metric structure of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. For a step size $h > 0$, define the updates

$$\mu_{h,k+1} := \operatorname*{arg\,min}_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \left\{\mathcal{F}(\mu) + \frac{1}{2h}\,W_2^2(\mu,\mu_{h,k})\right\}, \tag{2.12}$$

where $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}\cup\{\infty\}$ is the functional of interest defined over the Wasserstein space. Note that in optimization, this is known as the "proximal point method" for minimizing $\mathcal{F}$.

As $h \searrow 0$, one hopes that we have convergence $\mu_{h,\lfloor t/h\rfloor} \to \mu_t$ in a suitable sense, and then the limiting curve $(\mu_t)_{t\geq 0}$ can be interpreted as the Wasserstein gradient flow of $\mathcal{F}$. This is indeed what [JKO98] showed in a particular, but important case.

Namely, if $\pi \propto \exp(-V)$ is a density on $\mathbb{R}^d$ obeying mild regularity conditions, and we take the functional to be the KL divergence, $\mathcal{F} = \mathsf{KL}(\cdot \,\|\, \pi)$, then the sequence of discrete approximations converges to the solution of the Fokker–Planck equation

$$\partial_t \mu_t = \operatorname{div}\big(\mu_t \, \nabla \ln \frac{\mu_t}{\pi}\big) \,. \tag{2.13}$$

As discussed further in §2.2, the Fokker–Planck equation governs the evolution of the marginal law of the Langevin diffusion

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \,,$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion on $\mathbb{R}^d$. Hence, this celebrated result says that the Langevin diffusion can be interpreted as the Wasserstein gradient flow of the KL divergence. The implicit discretization (2.12) is now commonly known as the "JKO scheme" after the authors Jordan, Kinderlehrer, and Otto.

Although the Wasserstein space is not truly a Riemannian manifold, many of the formal calculations of [Ott01] can now be justified rigorously, under appropriate technical conditions, due to the extensive theory developed in [AGS08; Vil09b].

2.1.4 Optimization over the Wasserstein space

Gradients and gradient flows. Given a functional $\mathcal{F} : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathbb{R}$, we can define its Wasserstein gradient formally as follows. The gradient of $\mathcal{F}$ at μ_0 is the element $\nabla_{W_2}\mathcal{F}(\mu_0) \in T_{\mu_0}\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ satisfying

$$\partial_t|_{t=0}\mathcal{F}(\mu_t) = \langle \nabla_{W_2}\mathcal{F}(\mu_0), v_0\rangle_{\mu_0} \tag{2.14}$$

for any curve $(\mu_t)_{t\in\mathbb{R}}$ with Wasserstein tangent vector v_0 at time 0; that is, $(\mu_t)_{t\in\mathbb{R}}$ and $(v_t)_{t\in\mathbb{R}}$ solve the continuity equation (2.5) with $v_t \in T_{\mu_t}\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ for a.e. $t \in \mathbb{R}$. To compute the Wasserstein gradient, suppose that $\mathcal{F}$ admits a *first variation* $\delta\mathcal{F}(\mu_0) : \mathbb{R}^d \to \mathbb{R}$, that is, for any such curve $(\mu_t)_{t\in\mathbb{R}}$, we have $\partial_t|_{t=0}\mathcal{F}(\mu_t) = \int \delta\mathcal{F}(\mu_0)\,\partial_t|_{t=0}\mu_t$.[4] By the continuity equation (2.5), we have $\partial_t|_{t=0}\mu_0 = -\operatorname{div}(\mu_0 v_0)$. Integrating by parts, we see that (2.14) equals $\langle \nabla\delta\mathcal{F}(\mu_0), v_0\rangle_{\mu_0}$. Moreover, since $\nabla\delta\mathcal{F}(\mu_0)$ is a gradient vector field, it belongs to $T_{\mu_0}\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. We conclude that

$$\nabla_{W_2}\mathcal{F}(\mu) = \nabla\delta\mathcal{F}(\mu) \,.$$

Definition 2.1.6. *A curve $(\mu_t)_{t\geq 0} \subseteq \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ is a* Wasserstein gradient flow *of the functional $\mathcal{F} : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathbb{R}\cup\{\infty\}$ if for a.e. $t \in \mathbb{R}$, the Wasserstein tangent*

[4] The first variation is only defined up to an additive constant.

vector to the curve at time t is $-\nabla_{W_2}\mathcal{F}(\mu_t)$. In light of (2.5), it means that the following PDE holds:

$$\partial_t \mu_t = \operatorname{div}\big(\mu_t \, \nabla_{W_2}\mathcal{F}(\mu_t)\big) \, .$$

The *Riemannian gradient descent* update for $\mathcal{F}$ with step size η starting at μ is given by

$$\mu^+ := \exp_\mu\big(-\eta \, \nabla_{W_2}\mathcal{F}(\mu)\big) = [\mathrm{id} - \eta \, \nabla_{W_2}\mathcal{F}(\mu)]_{\#}\mu \, .$$

Note that the step size η should be chosen small enough that $-\eta \, \nabla_{W_2}\mathcal{F}(\mu)$ lies in the domain of the exponential map. From the general description of the tangent space of Wasserstein space, $\nabla_{W_2}\mathcal{F}(\mu)$ is the gradient of a mapping $\psi : \mathbb{R}^d \to \mathbb{R}$; then, $-\eta \, \nabla_{W_2}\mathcal{F}(\mu)$ belongs to the domain of the exponential map if $\|\cdot\|^2/2 - \eta\psi$ is convex.

Convexity and smoothness. We say that $\mathcal{F}$ is α-convex if

$$\mathcal{F}(\mu_1) \geq \mathcal{F}(\mu_0) + \langle \nabla_{W_2}\mathcal{F}(\mu_0), \log_{\mu_0}\mu_1 \rangle_{\mu_0} + \frac{\alpha}{2} \, W_2^2(\mu_0, \mu_1) \, , \quad \forall \mu_0, \mu_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \, , \tag{2.15}$$

and β-smooth if

$$\mathcal{F}(\mu_1) \leq \mathcal{F}(\mu_0) + \langle \nabla_{W_2}\mathcal{F}(\mu_0), \log_{\mu_0}\mu_1 \rangle_{\mu_0} + \frac{\beta}{2} \, W_2^2(\mu_0, \mu_1) \, , \quad \forall \mu_0, \mu_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \, .$$

These two properties are formally equivalent to the following statements: for any constant-speed geodesic $(\mu_t)_{t\in[0,1]}$, one has

$$\partial_t^2|_{t=0}\mathcal{F}(\mu_t) \geq \alpha \, W_2^2(\mu_0, \mu_1) \qquad \text{or} \qquad \partial_t^2|_{t=0}\mathcal{F}(\mu_t) \leq \beta \, W_2^2(\mu_0, \mu_1),$$

respectively. Also, (2.15) is equivalent to $\mathcal{F}$ being α-geodesically convex in the sense of Definition 2.1.5.

2.2 Background on diffusions

We assume familiarity with basic notions from stochastic calculus, see [Le 16; Str18]. We refer to [BGL14] for further background.

2.2.1 Markov semigroup theory and functional inequalities

Diffusions. Diffusion processes play a predominant role in the study of sampling, and in this work we shall be particularly interested in the *Langevin diffusion*.

Definition 2.2.1. *Let $V : \mathbb{R}^d \to \mathbb{R}$ be a $\mathcal{C}^1$ function such that $\int \exp(-V) < \infty$, called the* potential. *The* Langevin diffusion *associated with V is the solution $(X_t)_{t\geq 0}$ to the stochastic differential equation (SDE)*

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,. \tag{2.16}$$

Here, $(B_t)_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$.

If we assume, as we do in most of this work for discretization purposes, that ∇V is Lipschitz, then according to the standard theory of SDEs there is a unique strong solution to (2.16). That is, given a filtered probability space $(\Omega, \mathbb{P}, \mathscr{F}, (\mathscr{F}_t)_{t\geq 0})$ supporting a standard Brownian motion $(B_t)_{t\geq 0}$, there is a unique adapted process $(X_t)_{t\geq 0}$ with continuous sample paths such that (2.16) holds, which is to be interpreted via stochastic integrals: for all $t \geq 0$, $X_t = X_0 - \int_0^t \nabla V(X_s)\,\mathrm{d}s + \sqrt{2}\,B_t$. This process satisfies the strong Markov property.

As we discuss below, the stationary distribution of the Langevin diffusion is $\pi \propto \exp(-V)$. In the special case when $V = \frac{\|\cdot\|^2}{2}$, then the stationary distribution is standard Gaussian and the diffusion is known as the *Ornstein–Uhlenbeck* (OU) process. In this case, the SDE is linear and can be solved in closed form.

Besides the Langevin diffusion, in §6 we will also study the *underdamped Langevin diffusion*, which describes motion in a particle well with friction.

Markov semigroups. In order to develop a useful calculus for working with diffusions, it is helpful to abstractly represent them through their actions on test functions, as captured via the following definition.

Definition 2.2.2. *A* Markov semigroup *is a semigroup of linear operators $(P_t)_{t\geq 0}$ acting on a suitable space of functions (containing constant functions) such that:*

1. *For all $t \geq 0$, $P_t 1 = 1$.*
2. *For all $t \geq 0$, if $f \geq 0$, then $P_t f \geq 0$.*
3. *We have $P_0 = \mathrm{id}$, and for all $s, t \geq 0$, it holds that $P_{s+t} = P_s \circ P_t = P_t \circ P_s$ (semigroup property).*

Given a Markov process $(X_t)_{t\geq 0}$, the corresponding Markov semigroup is given by $P_t f(x) \coloneqq \mathbb{E}[f(X_t) \mid X_0 = x]$. Conversely, there are many results which provide conditions under which there exists a corresponding Markov process for a given Markov semigroup $(P_t)_{t\geq 0}$.

Calculus enters the picture once we consider the "time derivative of the semigroup". The semigroup property ensures that it suffices to consider this derivative at time 0, as in the following definition.

Definition 2.2.3. *Given a Markov semigroup* $(P_t)_{t\geq 0}$*, the* infinitesimal generator $\mathscr{L}$ *is defined via*

$$\mathscr{L} f := \lim_{t \searrow 0} \frac{P_t f - f}{t} \,. \tag{2.17}$$

In Definitions 2.2.2 and 2.2.3, we are purposefully vague regarding the class of functions on which the semigroup acts, as well as the sense in which the limit (2.17) is taken. This is a rather subtle issue. In order to develop a satisfactory spectral theory, we would like to consider the space of functions $L^2(\pi)$, where π is the stationary distribution of the corresponding Markov process. However, not all functions $f \in L^2(\pi)$ have sufficient regularity for the limit in (2.17) to exist (e.g., in $L^2(\pi)$). To address this, one can consider $\mathscr{L}$ to be an *unbounded* operator on $L^2(\pi)$, meaning that it comes together with a corresponding *domain* of definition which is a strict subspace of $L^2(\pi)$. The choice of domain is not obvious, and it must be done carefully in order to properly define notions such as self-adjointness. These issues are also handled in [BGL14] through the formalism of a *Markov triple* which specifies an algebra of test functions (e.g., compactly supported and smooth functions). For the sake of this informal introduction, we ignore these issues and focus on the calculus itself.

For the Langevin diffusion (2.16), the infinitesimal generator is given by

$$\mathscr{L} f = \Delta f - \langle \nabla V, \nabla f \rangle \,.$$

For standard Brownian motion, the generator is $\mathscr{L} = \frac{1}{2}\Delta$.

The Markov semigroup encodes the dynamics of the Markov process, as shown by *Kolmogorov's equations*: for any test function f, it holds that $\partial_t P_t f = \mathscr{L} P_t f$. In other words, if we set $u_t := P_t f$ for all $t \geq 0$, then u solves the *heat equation* $\partial_t u_t = \mathscr{L} u_t$, which coincides with the usual heat equation (up to a factor $\frac{1}{2}$) for Brownian motion.

Dually, we can let the semigroup act on probability densities by setting $P_t^* \mu$ to denote the law of the diffusion at time t when initialized at μ. This notation is justified because for any test function f, $\mathbb{E}\, f(X_t) = \mathbb{E}\, P_t f(X_0) = \int P_t f \,\mathrm{d}\mu = \int f \, P_t^* \mu$, where P_t^* denotes the adjoint of P_t w.r.t. Lebesgue measure. Then, the dual to the heat equation is $\partial_t P_t^* \mu = \mathscr{L}^* P_t^* \mu$, where $\mathscr{L}^*$ denotes the Lebesgue adjoint of $\mathscr{L}$. In the case of Langevin diffusion, we have $\mathscr{L}^* \mu = \Delta \mu + \operatorname{div}(\mu \nabla V)$. Hence, if μ_t is the marginal of the Langevin diffusion at time t, we have the PDE for the evolution of the probability density, known as the *Fokker–Planck equation*:

$$\partial_t \mu_t = \Delta \mu_t + \operatorname{div}(\mu_t \nabla V) \,. \tag{2.18}$$

The Fokker–Planck equation readily implies that $\pi \propto \exp(-V)$ is stationary for the Langevin diffusion, because $\mathscr{L}^* \pi = 0$. Dually, $\mathbb{E}_\pi \mathscr{L} f = 0$ for f.

Reversibility and integration by parts. In the above discussion, we had to introduce P_t^* and $\mathscr{L}^*$ as new operators because the semigroup and generator are typically not self-adjoint w.r.t. Lebesgue measure. Instead, they are typically self-adjoint w.r.t. another measure π, in which case the Markov semigroup is called reversible.

Definition 2.2.4. *A Markov semigroup $(P_t)_{t\geq 0}$ is* reversible *w.r.t. a probability measure π if P_t defines a self-adjoint operator on $L^2(\pi)$ for all $t \geq 0$.*

If $(P_t)_{t\geq 0}$ is reversible w.r.t. π, then it implies that π is the stationary distribution of the corresponding Markov process. Also, reversibility implies (and is equivalent to) the generator $\mathscr{L}$ being self-adjoint on $L^2(\pi)$, once "self-adjoint" is defined appropriately for unbounded operators. As the name suggests, the property of reversibility indeed implies that the Markov process (started at the stationary distribution π) has the same law backwards and forwards in time. The Langevin diffusion is a key example of a reversible diffusion.

For a reversible diffusion, if we instead consider the relative density $\rho_t \coloneqq \mu_t/\pi$ w.r.t. the stationary distribution π, then the semigroup coincides with its $L^2(\pi)$ adjoint and hence the Fokker–Planck equation can be written simply as $\partial_t \rho_t = \mathscr{L}\rho_t$ or $\rho_t = P_t\rho_0$.

Observe that Kolmogorov's equation shows that the dynamics of the diffusion are encoded via a *linear* PDE involving the generator $\mathscr{L}$, and under the condition of reversibility, the operator $\mathscr{L}$ is self-adjoint. Hence, we expect to obtain a spectral theory for $\mathscr{L}$ in which $\mathscr{L}$ has real spectrum, and moreover this spectrum should govern the rate of convergence to stationarity for the diffusion. This is indeed the case, and the first step is to identify the quadratic form associated with $\mathscr{L}$. Actually, it is convenient to consider the form associated with $-\mathscr{L}$ instead.

Definition 2.2.5. *Suppose that $(P_t)_{t\geq 0}$ is a reversible Markov semigroup with generator $\mathscr{L}$ and stationary distribution π. The* Dirichlet energy *associated with $\mathscr{L}$ is the bilinear form*

$$\mathscr{E}(f,g) \coloneqq \langle f, (-\mathscr{L})\, g\rangle_\pi = \langle (-\mathscr{L})\, f, g\rangle_\pi \,.$$

One can show that the Dirichlet energy can be written as $\mathscr{E}(f,g) = \int \Gamma(f,g)\,\mathrm{d}\pi$, where Γ is a bilinear operator called the *carré du champ*; moreover, $\Gamma(f,f) \geq 0$ for any function f. In particular, $\mathscr{E}(f,f) \geq 0$. For example, for the Langevin diffusion, we have $\Gamma(f,g) = \langle \nabla f, \nabla g\rangle$, which does not depend on the potential V.

The inequality $\mathscr{E}(f,f) \geq 0$ for all f shows that $-\mathscr{L}$ is a positive operator. Constant functions always lie in the kernel of $-\mathscr{L}$, and the infimum of the spectrum restricted to the orthogonal complement of constant functions (that is, functions f with $\mathbb{E}_\pi f = 0$) is called the *spectral gap* of the diffusion.

Functional inequalities. A *functional inequality* is an inequality that holds true for all elements of a suitable function class. Functional inequalities can encode a wealth of information, with implications ranging from concentration of measure to rapid mixing of Markov processes.

In the context of sampling, the most well-studied functional inequalities are the *Poincaré inequality* (PI) and the *log-Sobolev inequality* (LSI).

Definition 2.2.6. *A probability distribution π on $\mathbb{R}^d$ satisfies a* Poincaré inequality *(PI) with constant C_{PI} if for all smooth and compactly supported functions $\phi : \mathbb{R}^d \to \mathbb{R}$,*

$$\operatorname{var}_\pi(\phi) \le C_{\mathsf{PI}}\, \mathbb{E}_\pi[\|\nabla\phi\|^2]\,. \tag{2.19}$$

Definition 2.2.7. *A probability distribution π on $\mathbb{R}^d$ satisfies a* log-Sobolev inequality *(LSI) with constant C_{LSI} if for all smooth and compactly supported functions $\phi : \mathbb{R}^d \to \mathbb{R}$,*

$$\operatorname{ent}_\pi(\phi^2) := \mathbb{E}_\pi\Big[\phi^2 \log \frac{\phi^2}{\mathbb{E}_\pi[\phi^2]}\Big] \le 2C_{\mathsf{LSI}}\, \mathbb{E}_\pi[\|\nabla\phi\|^2]\,. \tag{2.20}$$

By taking $f = \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}}$, the LSI can also be rewritten in the equivalent form

$$\mathsf{KL}(\mu \,\|\, \pi) \le 2C_{\mathsf{LSI}}\, \mathbb{E}_\pi\big[\big\|\nabla\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}}\big\|^2\big] = \frac{C_{\mathsf{LSI}}}{2}\, \mathbb{E}_\mu\big[\big\|\nabla \ln \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\big\|^2\big] =: \frac{C_{\mathsf{LSI}}}{2}\, \mathsf{FI}(\mu \,\|\, \pi)\,. \tag{2.21}$$

These functional inequalities are classically related to the ergodicity properties of the Langevin diffusion (3.1). Indeed, if π_t denotes the law of the diffusion at time t, then a PI is equivalent to

$$\chi^2(\pi_t \,\|\, \pi) \le \exp\Big(-\frac{2t}{C_{\mathsf{PI}}}\Big)\, \chi^2(\pi_0 \,\|\, \pi)\,, \qquad \text{for all } t \ge 0\,, \tag{2.22}$$

whereas an LSI is equivalent to

$$\mathsf{KL}(\pi_t \,\|\, \pi) \le \exp\Big(-\frac{2t}{C_{\mathsf{LSI}}}\Big)\, \mathsf{KL}(\pi_0 \,\|\, \pi)\,, \qquad \text{for all } t \ge 0\,. \tag{2.23}$$

We review information divergences such as KL and χ^2 in §2.2.3 below.

The inequality (2.22) can be understood from the spectral perspective: the right-hand side of the Poincaré inequality (2.19) is precisely the Dirichlet energy $\mathscr{E}(\phi, \phi)$ for the Langevin diffusion, and hence (2.19) is equivalent to a lower

bound of C_{PI}^{-1} on the spectral gap for the Langevin diffusion. On the other hand, the chi-squared divergence $\chi^2(\pi_t \,\|\, \pi)$ is just the squared $L^2(\pi)$ norm of the projection of the relative density π_t/π orthogonal to constant functions, i.e., $\chi^2(\pi_t \,\|\, \pi) = \|\pi_t/\pi - 1\|^2_{L^2(\pi)}$. Therefore, the equivalence between (2.19) and (2.22) follows from the Fokker–Planck equation and spectral theory. The equivalence between (2.20) and (2.23) does not admit such a spectral interpretation, but it follows via a quick calculation using Markov semigroup theory.

Improving upon the prior result of [CLL19], [VW19] showed that the functional inequalities (2.19) and (2.20) also imply convergence for the Langevin diffusion in Rényi divergence; see Theorem 3.2.1 in §3.

We next collect together key facts about these functional inequalities. The following results show that the class of distributions satisfying these inequalities is larger than the class of strongly log-concave distributions; see [BGL14, Proposition 5.1.3 and Corollary 5.7.2].

Lemma 2.2.8 (Strong log-concavity implies LSI implies PI)**.** *Let π be a distribution on $\mathbb{R}^d$.*

1. *(*Bakry–Émery theorem*) If π is α-strongly log-concave, then it satisfies an LSI with constant at most $1/\alpha$.*
2. *If π satisfies an LSI with constant C_{LSI}, then it also satisfies a PI with constant at most C_{LSI}.*

The second part of the lemma is standard and follows from linearizing the LSI. The first part of the lemma (the Bakry–Émery theorem) is deeper and we discuss it further below.

A useful consequence of the LSI is the following sub-Gaussian concentration inequality for Lipschitz functions, typically established via the Herbst argument; see [BGL14, Proposition 5.4.1].

Lemma 2.2.9 (LSI implies sub-Gaussian concentration)**.** *Suppose that π is a distribution on $\mathbb{R}^d$ satisfying an LSI with constant C_{LSI}. Then, for any L-Lipschitz function $\phi : \mathbb{R}^d \to \mathbb{R}$ and any $\lambda \in \mathbb{R}$,*

$$\mathbb{E}_\pi \exp\big(\lambda\,(\phi - \mathbb{E}_\pi\,\phi)\big) \le \exp\Big(\frac{\lambda^2 C_{\mathsf{LSI}} L^2}{2}\Big)\,.$$

Consequently, for all $\eta \ge 0$,

$$\pi\{\phi - \mathbb{E}_\pi\,\phi \ge \eta\} \le \exp\Big(-\frac{\eta^2}{2C_{\mathsf{LSI}}L^2}\Big)\,.$$

Similarly, the PI implies subexponential concentration, see [BGL14, §4.4.3].

Lemma 2.2.10 (PI implies subexponential concentration)**.** *Suppose that π is a distribution on $\mathbb{R}^d$ satisfying a PI with constant C_{PI}. Then, for any L-Lipschitz function $\phi : \mathbb{R}^d \to \mathbb{R}$ and any $\eta \geq 0$,*

$$\pi\{\phi - \mathbb{E}_\pi \phi \geq \eta\} \leq 3 \exp\Big(-\frac{\eta}{\sqrt{C_{\mathsf{PI}}}\, L}\Big) .$$

Next we recall two comparison inequalities which enable proving sampling guarantees in Wasserstein distance as an immediate corollary of proving sampling guarantees in other metrics—namely KL divergence in the LSI setting, and chi-squared divergence in the PI setting. Such comparison inequalities are often called transport inequalities. Specifically, the first result, attributed to Otto and Villani [OV00], shows that under an LSI, a transportation inequality between Wasserstein and KL divergence holds (this inequality is often referred to as Talagrand's T_2 inequality).

Lemma 2.2.11 (Otto–Villani theorem)**.** *Suppose that π is a distribution on $\mathbb{R}^d$ satisfying an LSI with constant C_{LSI}. Then, for all distributions $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$W_2^2(\mu, \pi) \leq 2 C_{\mathsf{LSI}}\, \mathsf{KL}(\mu \,\|\, \pi) .$$

The second result shows a similar transport inequality in the PI setting [Liu20]. Under a PI, Talagrand's T_2 inequality does not necessarily hold anymore. Nevertheless, a useful transport inequality still holds if one replaces the KL divergence by the chi-squared divergence.

Lemma 2.2.12 (Quadratic transport-variance inequality)**.** *Suppose that π is a distribution on $\mathbb{R}^d$ satisfying a PI with constant C_{PI}. Then, for all distributions $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$W_2^2(\mu, \pi) \leq 2 C_{\mathsf{PI}}\, \chi^2(\mu \,\|\, \pi) .$$

Finally, we record the following standard second-moment-type bound for strongly log-concave measures; see, e.g., [DKR22, Proposition 2]. We give a short proof sketch for the convenience of the reader.

Lemma 2.2.13 (Second moment bound)**.** *Suppose that $\pi \propto \exp(-V)$ is α-strongly log-concave, with mode at $x^\star$. Then, it holds that $\int \|\cdot - x^\star\|^2 \,\mathrm{d}\pi \leq d/\alpha$.*

Proof. Integration by parts shows that for any smooth function $\phi : \mathbb{R}^d \to \mathbb{R}$ of controlled growth, it holds that $\mathbb{E}_\pi \mathscr{L}\phi = \mathbb{E}_\pi[\Delta\phi - \langle \nabla V, \nabla\phi \rangle] = 0$, where $\mathscr{L}$ is the generator of the Langevin diffusion. We apply this to $\phi(x) := \frac{1}{2}\|x - x^\star\|^2$, for which $\nabla\phi(x) = x - x^\star$ and $\Delta\phi = d$. By strong convexity of V, $\langle \nabla V(x), x - x^\star \rangle \geq \alpha \|x - x^\star\|^2$, and the result follows. □

Functional inequalities are particularly useful for high-dimensional non-log-concave sampling because they tensorize (if two measures satisfy the same functional inequality, then their product also satisfies the functional inequality with the same constant) and they are stable under common operations such as bounded perturbation (replacing the potential V with $\widetilde{V}$, with $\sup|V-\widetilde{V}|<\infty$), Lipschitz mapping (replacing π with $T_{\#}\pi$ where $T:\mathbb{R}^d\to\mathbb{R}^d$ is Lipschitz and the pushforward $T_{\#}\pi$ is the distribution of $T(X)$ when $X\sim\pi$), or taking mixtures (see §11). We refer to [BGL14] for a comprehensive treatment.

Curvature-dimension condition. The Bakry–Émery theorem in Lemma 2.2.8 is a celebrated result due to its geometric interpretation, which we briefly describe. We have already mentioned that associated to a reversible Markov semigroup, we have a *carré du champ operator* Γ, which is a bilinear operator mapping pairs of functions to functions. One can also define another operator in the same spirit, called the *iterated carré du champ operator* and denoted Γ_2. Then, we say that the Markov diffusion satisfies the *curvature-dimension condition* $\mathsf{CD}(\alpha,d)$ if $\Gamma_2(f,f)\geq\alpha\Gamma(f,f)+(\mathscr{L}f)^2/d$. Although we have not defined the operators Γ, Γ_2, as they will not be used in the sequel, the salient point is that the curvature-dimension condition can be defined solely in terms of the Markov semigroup.

The relevance of the curvature-dimension condition is that the semigroup associated to the standard Brownian motion on a Riemannian manifold satisfies $\mathsf{CD}(\alpha,d)$ if and only if the Ricci curvature of the manifold is at least α and the dimension of the manifold is at most d. Thus, as the name indicates, in this context the curvature-dimension condition encodes curvature and dimension information in Riemannian geometry. However, since the curvature-dimension condition is purely written in terms of the Markov semigroup, we can also ask if it holds for diffusions outside of this Riemannian context; for instance, we can ask if the Langevin diffusion satisfies this condition. If so, we can interpret it as encoding abstract geometric properties *intrinsic* to the Markov process.

It turns out that the Langevin diffusion satisfies $\mathsf{CD}(\alpha,\infty)$ if and only if $\nabla^2 V\succeq\alpha I_d$. Hence, the curvature of the potential V acts as a substitute for the Ricci curvature of the ambient space. In this abstract context, the Bakry–Émery theorem asserts that $\mathsf{CD}(\alpha,\infty)$ with $\alpha>0$ implies the validity of an LSI with constant $1/\alpha$.

2.2.2 The Langevin diffusion as a Wasserstein gradient flow

We now briefly review the interpretation in [JKO98] of the Langevin diffusion as a Wasserstein gradient flow.

Let $V:\mathbb{R}^d\to\mathbb{R}$ denote the potential, and let $\pi\propto\exp(-V)$. Consider the KL

divergence $\mathsf{KL}(\mu \,\|\, \pi)$, which can be decomposed as a sum

$$\mathcal{F}(\mu) \coloneqq \mathsf{KL}(\mu \,\|\, \pi) = \int V \,\mathrm{d}\mu + \int \mu \ln \mu + \text{constant}\,.$$

The first term can be interpreted as the potential energy, and the second term is the (negative) entropy. Recall from §2.1 that the Wasserstein gradient of $\mathcal{F}$ is given by $\nabla \delta \mathcal{F}$. One can show via direct calculation that $\delta \mathcal{F}(\mu) = V + \ln \mu$ up to an additive constant, and we deduce that the Wasserstein gradient of the KL divergence is

$$\nabla_{W_2} \mathcal{F}(\mu) = \nabla V + \nabla \ln \mu = \nabla \ln \frac{\mu}{\pi}\,.$$

The Wasserstein gradient flow for the KL divergence is the curve $(\mu_t)_{t \geq 0}$ with

$$\partial_t \mu_t = \operatorname{div}\big(\mu_t \,\nabla \ln \frac{\mu_t}{\pi}\big)\,.$$

By comparison with (2.18), we can deduce that the marginal law of the Langevin diffusion traces out the Wasserstein gradient flow of $\mathsf{KL}(\cdot \,\|\, \pi)$. This celebrated result endows the Langevin diffusion with a geometric interpretation with close connections to optimization, which is a central theme explored in this thesis. Namely, since the Langevin diffusion is a gradient flow, we can use the theory of gradient flows to study its convergence.

The starting point is to investigate the convexity of the objective functional. Using Otto calculus, one can show that the Hessian of the KL divergence, viewed as a quadratic form on $T_\mu \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, formally takes the form

$$\nabla^2_{W_2} \mathcal{F}(\mu)[v, v] = \int \langle \nabla^2 V \, v, v \rangle \,\mathrm{d}\mu + \int \|\nabla v - I_d\|^2_{\mathrm{HS}} \,\mathrm{d}\mu\,.$$

In particular, if $\nabla^2 V \succeq \alpha I_d$, then $\nabla^2_{W_2} \mathcal{F}(\mu)[v, v] \geq \alpha \,\|v\|^2_\mu$ and hence $\mathcal{F}$ is α-convex along Wasserstein geodesics. Via general principles for gradient flows (see §16.8), it implies the following result: if $(\mu_t)_{t \geq 0}$, $(\nu_t)_{t \geq 0}$ are the marginal laws of two copies of the Langevin diffusion (but with possibly different initializations) corresponding to an α-convex potential V, then

$$W_2(\mu_t, \nu_t) \leq \exp(-\alpha t)\, W_2(\mu_0, \nu_0)\,.$$

This result could also be deduced by a synchronous coupling of the diffusions, together with Itô's formula.

Since $\nabla_{W_2}\mathcal{F}(\mu) = \nabla \ln(\mu/\pi)$, then the squared norm of the Wasserstein gradient is given by $\|\nabla \ln(\mu/\pi)\|_\mu^2$, which is also called the *relative Fisher information* $\mathsf{FI}(\mu \,\|\, \pi)$. By general calculation rules for gradient flows,

$$\partial_t \,\mathsf{KL}(\mu_t \,\|\, \pi) = -\Big\|\nabla \ln \frac{\mu_t}{\pi}\Big\|_{\mu_t}^2 = -\mathsf{FI}(\mu_t \,\|\, \pi)\,.$$

From this equality, we see that exponential decay of the KL divergence follows from the condition $\mathsf{FI}(\mu \,\|\, \pi) \geq 2\alpha\, \mathsf{KL}(\mu \,\|\, \pi)$ for all μ. Recalling (2.21), we see that this precisely amounts to a LSI with constant $1/\alpha$, and thus we obtain (2.23). In this context, however, the LSI takes on an operational meaning: namely, it is seen to be the *gradient domination* or *Polyak–Łojasiewicz* (PL) inequality from optimization (see, e.g., [KNS16]). In general, α-convexity implies a PL inequality, which therefore recovers the Bakry–Émery theorem (Lemma 2.2.8) for the Langevin diffusion. This perspective was first laid out in [OV00].

Finally, we mention that the Otto–Villani theorem (Lemma 2.2.11), which asserts that an LSI implies a transport inequality, is also an instantiation of a general fact from optimization, namely, that a PL inequality implies a quadratic growth inequality [KNS16].

■ 2.2.3 Comparisons between divergences

In this section, we collect together common divergences between probability measures as well as the relationships between them.

Definition 2.2.14. *The* total variation (TV) distance *between* μ *and* π *is*

$$\|\mu - \pi\|_{\mathsf{TV}} \coloneqq \sup_{A \subseteq \mathbb{R}^d \text{ measurable}} |\mu(A) - \pi(A)|\,.$$

Definition 2.2.15. *The* KL divergence *of* μ *from* π *is*

$$\mathsf{KL}(\mu \,\|\, \pi) \coloneqq \int \ln \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\,\mathrm{d}\mu = \int \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \ln \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\,\mathrm{d}\pi\,,$$

where $\mathsf{KL}(\mu \,\|\, \pi)$ *is understood to be* $+\infty$ *if* $\mu \not\ll \pi$.

Definition 2.2.16. *The* chi-squared divergence *of* μ *from* π *is*

$$\chi^2(\mu \,\|\, \pi) \coloneqq \int \big(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\big)^2\,\mathrm{d}\pi = \int \big(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\big)^2\,\mathrm{d}\pi - 1\,,$$

where $\chi^2(\mu \,\|\, \pi)$ *is understood to be* $+\infty$ *if* $\mu \not\ll \pi$.

We also introduce the family of *Rényi divergences*, which includes both the KL divergence and the chi-squared divergence as special cases.

Definition 2.2.17. *The* Rényi divergence *of order* $q \in (1, \infty)$ *of* μ *from* π *is*

$$\mathcal{R}_q(\mu \,\|\, \pi) \coloneqq \frac{1}{q-1} \ln \left\| \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right\|_{L^q(\pi)}^q ,$$

where $\mathcal{R}_q(\mu \,\|\, \pi)$ *is understood to be* $+\infty$ *if* $\mu \not\ll \pi$.

Remark 2.2.18 (Special cases of Rényi divergence)**.** *The Rényi divergence of order* $q = 1$ *coincides with the KL divergence, i.e.,*

$$\mathcal{R}_1(\mu \,\|\, \nu) = \mathsf{KL}(\mu \,\|\, \nu) .$$

The Rényi divergence of order $q = 2$ *is related to the* χ^2 *divergence via the formula*

$$\mathcal{R}_2(\mu \,\|\, \nu) = \ln\bigl(1 + \chi^2(\mu \,\|\, \nu)\bigr) .$$

The Rényi divergence of order $q = \infty$ *is given by*

$$\mathcal{R}_\infty(\mu \,\|\, \nu) = \ln \left\| \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \right\|_{L^\infty(\nu)} .$$

We repeatedly use the following elementary properties of the Rényi divergence. Further details about these properties and their proofs can be found in, e.g., the Rényi divergence survey [EH14] as Theorem 1, Theorem 3, Equation 10, and Remark 1, respectively.

Lemma 2.2.19 (Data-processing inequality for Rényi divergences)**.** *For any Rényi order* $q \geq 1$*, any Markov transition kernel* P*, and any probability distributions* μ*,* ν*, it holds that*

$$\mathcal{R}_q(\mu P \,\|\, \nu P) \leq \mathcal{R}_q(\mu \,\|\, \nu) .$$

Lemma 2.2.20 (Monotonicity of Rényi divergences)**.** *For any Rényi orders* $q' \geq q \geq 1$*, and any probability distributions* μ, ν*,*

$$\mathcal{R}_q(\mu \,\|\, \nu) \leq \mathcal{R}_{q'}(\mu \,\|\, \nu) .$$

Lemma 2.2.21 (Rényi divergence between isotropic Gaussians)**.** *For any Rényi order* $q \geq 1$*, any variance* $\sigma^2 > 0$*, and any means* $x, y \in \mathbb{R}^d$*,*

$$\mathcal{R}_q\bigl(\mathsf{normal}(x, \sigma^2 I_d) \,\big\|\, \mathsf{normal}(y, \sigma^2 I_d)\bigr) = \frac{q \,\|x - y\|^2}{2\sigma^2} .$$

Lemma 2.2.22 (Relation to f-divergences)**.** *For any Rényi order* $q \in (1, \infty)$*, the corresponding function* $\exp((q-1)\,\mathcal{R}_q(\cdot \,\|\, \cdot))$ *is an* f*-divergence, and thus in particular is jointly convex in its arguments.*

We end this section with one last property of Rényi divergences: the weak triangle inequality. The name of this property arises from the fact that although Rényi divergences do not satisfy the triangle inequality, they do satisfy a modified version of it in which the Rényi order is increased and the bound is weakened by a multiplicative factor. Since this property does not appear in the aforementioned survey [EH14] on Rényi divergences, we provide a brief proof for completeness. It can also be found in, e.g., [Mir17, Proposition 11].

Lemma 2.2.23 (Weak triangle inequality for Rényi divergence)**.** *For any Rényi order $q > 1$, any $\lambda \in (0,1)$, and any probability distributions μ, ν, π,*

$$\mathcal{R}_q(\mu \,\|\, \pi) \le \frac{q-\lambda}{q-1}\, \mathcal{R}_{q/\lambda}(\mu \,\|\, \nu) + \mathcal{R}_{(q-\lambda)/(1-\lambda)}(\nu \,\|\, \pi)\,.$$

Proof. Expand $\mathcal{R}_q(\mu\|\nu) = \frac{1}{q-1} \ln \int fg$ where $f = \mu^q/\nu^{q-\lambda}$ and $g = \nu^{q-\lambda}/\pi^{q-1}$, and then apply Hölder's inequality $\int fg \le (\int f^a)^{1/a} (\int g^b)^{1/b}$ using Hölder exponents $a = 1/\lambda$ and $b = 1/(1-\lambda)$. □

Under a Poincaré inequality (2.19), the quadratic transport-variance inequality of Lemma 2.2.12 together with standard comparison inequalities such as Pinsker's inequality (see [Tsy09]) imply the comparisons

$$\max\Big\{2\,\|\mu-\pi\|_{\mathsf{TV}}^2,\ \ln\big(1+\frac{1}{2C_{\mathsf{PI}}}\,W_2^2(\mu,\pi)\big),\ \mathsf{KL}(\mu \,\|\, \pi)\Big\} \le \mathcal{R}_2(\mu \,\|\, \pi)\,.$$

This makes Rényi divergences a convenient family of divergences for proving sampling guarantees, since they imply guarantees in many other common divergences.

2.3 Background on the Bures–Wasserstein space

The material from this section is used in §15 and §16.

2.3.1 Geometry

We now specialize concepts from §2.1 to the Bures–Wasserstein manifold of centered non-degenerate Gaussian measures (identified with their covariance matrices), equipped with the Wasserstein metric. Thus, the Bures–Wasserstein manifold is the space $\mathbf{S}^d_{++}$ of positive-definite symmetric matrices equipped with a certain Riemannian metric.

The optimal transport problem between Gaussians is discussed in many places, e.g., [BJL19]. Given two covariance matrices $\Sigma, \Sigma' \in \mathbf{S}^d_{++}$, the optimal transport map between the corresponding centered Gaussians is the linear map

$$T_{\Sigma\to\Sigma'} = \Sigma^{-1/2} \left(\Sigma^{1/2}\Sigma'\Sigma^{1/2}\right)^{1/2} \Sigma^{-1/2}\,. \tag{2.24}$$

Note that this is a symmetric matrix. Since $AX \sim \mathsf{normal}(0, A\Sigma A^{\mathsf{T}})$ for $X \sim \mathsf{normal}(0, \Sigma)$, the fact that $T_{\Sigma\to\Sigma'}X \sim \mathsf{normal}(0, \Sigma')$ reduces to the matrix identity $T_{\Sigma\to\Sigma'}\Sigma T_{\Sigma\to\Sigma'} = \Sigma'$, which can be verified by hand. The above formula yields

$$\begin{aligned} W_2^2(\Sigma, \Sigma') &= \mathbb{E}[\|X - T_{\Sigma\to\Sigma'}X\|^2] = \mathbb{E}[\|X\|^2 + \|T_{\Sigma\to\Sigma'}X\|^2 - 2\langle X, T_{\Sigma\to\Sigma'}X\rangle] \\ &= \operatorname{tr}(\Sigma + \Sigma' - 2\,\Sigma\,T_{\Sigma\to\Sigma'})\,. \end{aligned} \tag{2.25}$$

From the general description of Wasserstein geodesics, the constant-speed geodesic $(\Sigma_t)_{t\in[0,1]}$ joining Σ to Σ' is given by

$$\Sigma_t = \big((1-t)\,I_d + t\,T_{\Sigma\to\Sigma'}\big)\,\Sigma\,\big((1-t)\,I_d + t\,T_{\Sigma\to\Sigma'}\big)\,, \qquad t \in [0,1]\,. \tag{2.26}$$

The tangent space $T_\Sigma \mathbf{S}^d_{++}$ is identified with the space $\mathbf{S}^d$ of symmetric $d \times d$ matrices. Given $S \in T_\Sigma \mathbf{S}^d_{++}$, the tangent space norm of S is given by $\|S\|_{L^2(\mathcal{N}(0,\Sigma))} = \sqrt{\mathbb{E}[\|SX\|^2]} = \sqrt{\langle S^2, \Sigma\rangle}$, which we simply denote as $\|S\|_\Sigma$. More generally, given matrices A, B, we write $\langle A, B\rangle_\Sigma := \operatorname{tr}(A^{\mathsf{T}}\Sigma B)$. The exponential map[5] is $\exp_\Sigma S = (I_d + S)\,\Sigma\,(I_d + S)$, so that $\exp_\Sigma(T_{\Sigma\to\Sigma'} - I_d) = \Sigma'$. The inverse of the exponential map is then $\log_\Sigma \Sigma' = T_{\Sigma\to\Sigma'} - I_d$.

The description of the Bures–Wasserstein tangent space is in accordance with the general Riemannian structure of Wasserstein space (see [AGS08]). We now elaborate on other possible conventions, in order to dispel possible confusion.

The space $\mathbf{S}^d_{++}$ is often studied as a manifold in other contexts, and the tangent space at any point is usually identified with $\mathbf{S}^d$. It is crucial to realize, however, that a tangent space is not simply a vector space (or inner product space); a tangent space also has the interpretation of describing velocities of curves. In other words, for each tangent vector S, we also need to prescribe which curves have velocity S. In the usual way of describing the manifold structure of $\mathbf{S}^d_{++}$, this prescription is given as follows. Given a curve $(\Sigma_t)_{t\in\mathbb{R}} \subseteq \mathbf{S}^d_{++}$, if $\dot\Sigma_0$ denotes the ordinary time derivative of this curve at time 0, then we declare $\dot\Sigma_0$ to be the tangent vector of the curve at time 0. Although this prescription is natural, observe that it conflicts with our description of the tangent space structure of the Bures–Wasserstein manifold; in particular, for the curve in (2.26), we have described the tangent vector to this curve (at time 0) to be $T_{\Sigma\to\Sigma'} - I_d$, but the ordinary time derivative of this curve is $(T_{\Sigma\to\Sigma'} - I_d)\,\Sigma + \Sigma\,(T_{\Sigma\to\Sigma'} - I_d)$.

To summarize the discussion in the preceding paragraph: although the usual description of the tangent space of $\mathbf{S}^d_{++}$ at Σ and our description of the tangent space are formally the same, in that they are both identified with $\mathbf{S}^d$, they differ in

[5] Technically the exponential map is only defined if $S + I_d \succeq 0$; this is because if $S + I_d$ is not positive semidefinite, then $S + I_d$ is not an optimal transport map due to Brenier's theorem.

that tangent vectors from the two descriptions give rise to different curves. Note that if we were to adopt the usual description of the tangent space of $\mathbf{S}^d_{++}$, then we would have to define the tangent space norm $\|\cdot\|_\Sigma$ differently from above. In this thesis, we adopt the convention described earlier in this section in order to preserve the connection with the general setting of optimal transport.

2.3.2 Additional useful facts

Here we collect various facts about the Wasserstein metric for easy reference.

1. **Euclidean gradient vs. Bures–Wasserstein gradient**.

   Let $F : \mathbf{S}^d_{++} \to \mathbb{R}$ be a function. Temporarily denote by $\mathrm{D}\, F$ the usual Euclidean gradient of F, and we reserve ∇F for the gradient with respect to the Bures–Wasserstein geometry. In fact, under our tangent space convention, these two quantities are related as follows: let $(\Sigma_t)_{t\in\mathbb{R}}$ denote a curve in $\mathbf{S}^d_{++}$. We temporarily denote the Euclidean tangent vector (i.e., ordinary time derivative) to this curve via $\dot\Sigma^{\mathrm{E}}$, and the Bures–Wasserstein tangent vector via $\dot\Sigma^{\mathrm{BW}}$, which are related via $\dot\Sigma^{\mathrm{E}} = \dot\Sigma^{\mathrm{BW}}\Sigma + \Sigma\dot\Sigma^{\mathrm{BW}}$ (see the discussion above). We can compute the time derivative of F in two ways:

   $$\begin{aligned}\langle \nabla F(\Sigma_0), \dot\Sigma_0^{\mathrm{BW}}\rangle_{\Sigma_0} &= \partial_t|_{t=0} F(\Sigma_t) = \langle \mathrm{D}\, F(\Sigma_0), \dot\Sigma_0^{\mathrm{E}}\rangle \\ &= \langle \mathrm{D}\, F(\Sigma_0), \dot\Sigma_0^{\mathrm{BW}}\Sigma_0 + \Sigma_0\dot\Sigma_0^{\mathrm{BW}}\rangle = 2\,\langle \mathrm{D}\, F(\Sigma_0), \dot\Sigma_0^{\mathrm{BW}}\rangle_{\Sigma_0}\,.\end{aligned}$$

   From this we can conclude that

   $$\nabla F(\Sigma_0) = 2\ \mathrm{D}\, F(\Sigma_0)\,.$$

2. **Gradient of the squared Wasserstein distance**.

   For any $\nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, the gradient of the functional $W_2^2(\cdot,\nu)$ at μ is

   $$\nabla W_2^2(\cdot,\nu)(\mu) = -2\,(T_{\mu\to\nu} - \mathrm{id}) = -2\log_\mu \nu\,.$$

   This is derived in, e.g., [ZP19]; see also (2.3). In the Bures–Wasserstein setting, it can be proven via matrix calculus.

3. **Inverse of the transport map**.

   If $\Sigma, \Sigma' \in \mathbf{S}^d_{++}$, then the transport map $T_{\Sigma\to\Sigma'}$ is the inverse matrix for the transport map $T_{\Sigma'\to\Sigma}$. This can be verified from the formula (2.24) using the symmetry of the geometric mean. More generally, it is a special case of the convex conjugacy relation between optimal Kantorovich potentials.

4. **Diagonal case**.

   If $\Sigma_0, \Sigma_1 \in \mathbf{S}^d_{++}$ are *diagonal matrices*, then $W_2^2(\Sigma_0, \Sigma_1) = \|\Sigma_0^{1/2} - \Sigma_1^{1/2}\|^2_{\mathrm{HS}}$ is the squared Hilbert–Schmidt norm between the square roots. This can be verified, e.g., from the explicit formula (15.2) using the fact that Σ_0 and Σ_1 commute. Note that in one dimension, all matrices are diagonal. More generally, these observations extend to when Σ_0 and Σ_1 commute.

   Similarly, it can be seen from (2.26) that the geodesic is given by

$$\Sigma_t^{1/2} = (1-t)\,\Sigma_0^{1/2} + t\,\Sigma_1^{1/2}\,, \qquad t \in [0,1]\,,$$

   which says that the Bures–Wasserstein geodesic between diagonal (or commuting matrices) is simply the Euclidean geodesic after applying the square root map.

5. **The case of non-zero means**.

   For any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, suppose that the means of these distributions are m_μ and m_ν, respectively. Let $\bar{\mu}$, $\bar{\nu}$ denote the centered versions of these distributions. Then, it holds that

$$W_2^2(\mu, \nu) = \|m_\mu - m_\nu\|^2 + W_2^2(\bar{\mu}, \bar{\nu})\,.$$

   This can be proven directly from the definition (2.1).

6. **A lower bound on the Wasserstein distance**.

   Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. If $\tilde{\mu}$ and $\tilde{\nu}$ are *Gaussian* measures with the same moments up to order two as μ and ν, respectively, then $W_2(\mu, \nu) \geq W_2(\tilde{\mu}, \tilde{\nu})$ [CMT96]. This fact also follows from the dual formulation of optimal transport.

Part I

Sampling under log-concavity and isoperimetry

Chapter 3

Analysis of Langevin Monte Carlo

We begin our study of sampling with perhaps the most canonical algorithm for this task, namely, Langevin Monte Carlo (LMC). Although LMC has been intensively studied for more than a decade, still the following fundamental question remained open: what convergence guarantees can be obtained when the target distribution π satisfies a Poincaré inequality?

Classically, a Poincaré inequality implies exponential convergence for the continuous-time Langevin diffusion in the chi-squared divergence.[1] Using this fact to provide guarantees for the discrete-time LMC algorithm, however, is considerably more challenging due to the need for working with chi-squared or Rényi divergences, and prior works have largely focused on strongly log-concave targets. In this chapter, we provide the first convergence guarantees for LMC assuming that π satisfies either a Latała–Oleszkiewicz or modified log-Sobolev inequality, which interpolates between the Poincaré and log-Sobolev settings. Unlike prior works, our results allow for weak smoothness and do not require convexity or dissipativity conditions. The techniques we develop for Rényi discretization analysis also play a key role for obtaining warm starts in §6.

This chapter is based on [Che+21a], joint with Murat A. Erdogdu, Mufan (Bill) Li, Ruoqi Shen, and Matthew Zhang.[2]

3.1 Introduction

The task of sampling from a target distribution $\pi \propto \exp(-V)$ on $\mathbb{R}^d$, known only up to a normalizing constant, is fundamental in many areas of scientific computing [Mac03; RC04; Liu08; Gel+14]. As such, there has been a considerable amount of research dedicated to this task, yielding precise and non-asymptotic algorithmic guarantees when the potential V is strongly convex; see, e.g., [Dal17a; DMM19; Dwi+19; SL19; HBE20; LST20; CLW21]. Many distributions encountered in

[1] This fact will be revisited in the context of the mirror Langevin diffusion in §8.

[2] This work also appeared as an extended abstract at COLT 2022 [Che+22c].

practice, however, are non-log-concave, and it is therefore of central importance to provide sampling guarantees for such distributions. In this work, we address this problem by working under the assumption that π satisfies a suitable functional inequality, which we now motivate.

The canonical sampling algorithm, Langevin Monte Carlo (LMC), is based on a discretization of the continuous-time Langevin diffusion, which is the solution to the stochastic differential equation

$$\mathrm{d}Z_t = -\nabla V(Z_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,. \tag{3.1}$$

Here, $(B_t)_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$. Classically, if π satisfies a functional inequality such as a Poincaré inequality or a log-Sobolev inequality, then the law of the Langevin diffusion (3.1) converges exponentially fast to the target distribution π (see §2.2 for background). Namely, a Poincaré inequality implies exponential convergence in chi-squared divergence, whereas a log-Sobolev inequality (which is stronger than a Poincaré inequality) implies exponential convergence in KL divergence.

The class of measures satisfying a Poincaré inequality is quite large, including all strongly log-concave measures (due the Bakry–Émery criterion) and, more generally, all log-concave measures [KLS95; Bob99; Che21a]. It also includes many examples of non-log-concave distributions such as Gaussian convolutions of measures with bounded support (see §11), and it is closed under bounded perturbations of the log-density. Owing to its broad applicability and its favorable continuous-time convergence properties, this class of measures is thus a natural goal for providing quantitative guarantees for non-log-concave sampling.

Sampling guarantees under functional inequalities. Our work is inspired by [VW19], which advocated the use of a functional inequality paired with a smoothness condition as a minimal set of assumptions for obtaining sampling guarantees; in their work, Vempala and Wibisono prove convergence of LMC under a log-Sobolev inequality. This result was then improved using the proximal Langevin algorithm under higher-order smoothness in [Wib19] and subsequently extended to Riemannian manifolds in [LE23].

Despite the appeal of this program, however, the majority of works on non-log-concave sampling instead make an additional assumption on the growth of the potential known as a dissipativity condition, see, e.g., [RRT17; EMS18; EH21; NDC21; EHZ22; Mou+22]. A representative example of such a condition is $\langle \nabla V(x), x\rangle \geq a\,\|x\| - b$ for some constants $a, b > 0$. Although useful for discretization proofs, dissipativity conditions are arguably less natural from the standpoint of the quantitative theory of Markov processes [BGL14], and ultimately redundant in the presence of an appropriate functional inequality. Other drawbacks include

the fact that b is typically dimension-dependent, and that dissipativity conditions are not as stable under perturbations (see §3.4 for an example). Hence, we avoid such conditions in our work.

In our first main result (Theorem 3.3.4), we assume that the target π satisfies a Latała–Oleszkiewicz inequality (LOI) with parameter $\alpha \in [1, 2]$. The LOI is a well-studied functional inequality that elegantly interpolates between Poincaré and log-Sobolev inequalities [LO00]. Notably, the $\alpha = 1$ case reduces to the Poincaré inequality, while the $\alpha = 2$ case reduces to the log-Sobolev inequality; intermediate values of α enable capturing potentials with growth $V(x) \approx \|x\|^\alpha$ (see §3.2.2). We also complement our result by proving a sampling guarantee (Theorem 3.3.6) under the modified log-Sobolev inequalities considered in [EH21], which is useful for treating examples in which the LOI constant is dimension-dependent.

Towards weaker notions of smoothness. Since the assumption of a Poincaré inequality allows for a variety of non-convex potentials with at least linear growth, it is restrictive to pair this assumption with the gradient Lipschitz assumption which is usually invoked in the sampling literature. Hence, following [DGN14; Nes15; Cha+20; EH21], we instead assume that ∇V is Hölder-continuous with some exponent $s \in (0, 1]$.

An analysis in Rényi divergence. We now describe the main technical challenge of this work. Recall that a log-Sobolev inequality (LSI) implies exponential ergodicity of the diffusion (3.1) in KL divergence, and consequently the analysis of LMC under a LSI naturally proceeds with the KL divergence as the performance metric [VW19; Wib19; LE23]. Similarly, a Poincaré inequality implies exponential ergodicity of (3.1) in chi-squared divergence, and accordingly we analyze LMC in chi-squared divergence, or equivalently, in Rényi divergence. In turn, the techniques we develop for the analysis may be useful for other situations in which only a Poincaré-type inequality is available, such as the state-of-the-art convergence rate for the underdamped Langevin diffusion [CLW20] or for the mirror Langevin diffusion (which we discuss further in §8).

Via standard comparison inequalities, a convergence guarantee in Rényi divergence implies convergence for other common divergences (e.g., total variation distance, 2-Wasserstein distance, or KL divergence), and is therefore more desirable. Of particular interest in this regard is the role of Rényi divergence guarantees for providing "warm starts" for high-accuracy samplers such as the Metropolis-adjusted Langevin algorithm (MALA), see §5 and §6.

Unfortunately, working with Rényi divergences introduces substantial new technical hurdles as it prevents the use of standard coupling-based discretization arguments; as such, there are not many prior works to draw upon. The convergence of the diffusion (3.1) in Rényi divergence was first proven in [CLL19;

VW19]. The paper [VW19] also takes a first step towards discretization by introducing a technique based on differential inequalities for the Rényi divergence for a continuous-time interpolation of LMC. Although this strategy succeeds for obtaining KL convergence under an LSI, it falls short for Rényi divergence; indeed, the analysis of [VW19] only holds under the (currently unverifiable) assumption that the *biased stationary distribution* of the LMC algorithm satisfies a Poincaré inequality. Moreover, their result only establishes quantitative convergence of LMC to its biased limit; to recover a convergence guarantee to π, this also requires an estimate of the "Rényi bias" (the Rényi divergence between the biased stationary distribution and π), which was unresolved. Instead, [GT20] provided the first Rényi guarantee for LMC by using the adaptive composition theorem from differential privacy to control the discretization error, albeit suboptimally. Subsequently, their result was sharpened in [EHZ22] via a two-stage analysis combining the two papers [VW19; GT20].

In this paper, we first show how to modify the interpolation method of [VW19] to yield a genuine Rényi convergence guarantee for LMC under an LSI, thereby yielding a stronger result than [GT20; EHZ22] with a shorter and more elegant proof. We further extend this to the case when π is log-concave, but this technique is unable to cover the setting of a weaker functional inequality and smoothness condition. For this, we instead draw inspiration from the stochastic calculus-based analysis of [DT12] (see also the similar argument in §5). At the heart of our proofs is the introduction of new change-of-measure inequalities which intriguingly rely on the very fact that the analysis is carried out in Rényi divergence (and not any weaker metric). Thus, although the use of Rényi divergences introduces new technical obstructions, it also provides the key tool for overcoming them.

3.1.1 Contributions

Convergence of the diffusion under functional inequalities. Our first contribution is to establish quantitative Rényi convergence bounds for the Langevin diffusion (3.1) under the following functional inequalities: (1) the Latała–Oleszkiewicz inequalities (LOI) [LO00], which interpolate between the Poincaré and log-Sobolev inequalities (Theorem 3.2.2), and the modified log-Sobolev inequality (MLSI) used in [EH21] (Theorem 3.2.3). LOI and MLSI have relative merits, and they capture the tail behavior of the potential, providing an accurate characterization of the speed of convergence for both the diffusion as well as the LMC algorithm.

Improved guarantees for LMC under an LSI or log-concavity. As our second principal contribution (Theorem 3.3.1), we provide an elegant proof that under an LSI, the LMC algorithm (with appropriate step size) achieves ε^2 error in Rényi divergence in $\widetilde{O}(d/\varepsilon^2)$ iterations. This improves upon past works in several respects. First,

in the LSI case, a Rényi convergence guarantee for LMC was previously unknown; thus, our work strengthens [VW19] by proving convergence in a stronger metric (Rényi divergence rather than KL divergence). Second, even when the target π is strongly log-concave, our proof is both sharper and significantly shorter than the prior works [GT20; EHZ22] on Rényi convergence; moreover, our guarantee for fixed step size LMC does not degrade if the number of iterations is taken too large. As a corollary, we resolve an open question of [VW19] on the size of the "Rényi bias" in this setting (see Corollary 3.3.2).

With additional effort, we are able to extend the techniques to the case when π is (weakly) log-concave, and we obtain a guarantee with explicit dependence on the Poincaré constant of π (however, our guarantee is no longer stable); see Theorem 3.3.3. Our result is the state-of-the-art guarantee for LMC for sampling from isotropic log-concave targets.

Convergence of LMC under a functional inequality and weak smoothness. Our main contribution is to provide sampling guarantees assuming that the potential has a Hölder-continuous gradient of exponent $s \in (0, 1]$ and that π either satisfies LOI (Theorem 3.3.4) or MLSI (Theorem 3.3.6). As noted previously, these assumptions are considerably more general than what are usually considered in the sampling literature and do not require dissipativity. In particular, Theorem 3.3.4 completes the program of [VW19] by establishing the first sampling guarantees for LMC under a Poincaré inequality and a weak smoothness condition.

Generically, our final rate is $\widetilde{O}(d^{(2/\alpha)\,(1+1/s)-1/s}/\varepsilon^{2/s})$, where s is the Hölder continuity exponent of ∇V and α captures the growth of the potential at infinity. We give a number of illustrative examples in §3.4 and show that our results improve upon the rates given in [EH21].

3.1.2 Notation and organization

Throughout the chapter, $\pi \propto \exp(-V)$ denotes the target distribution on $\mathbb{R}^d$; the function $V : \mathbb{R}^d \to \mathbb{R}$ is referred to as the "potential". We abuse notation by identifying a measure with its density (w.r.t. Lebesgue measure on $\mathbb{R}^d$). We write $a \lesssim b$ and $a = O(b)$ to indicate that $a \le Cb$ for a universal constant $C > 0$; also, we use $\widetilde{O}(\cdot)$ as a shorthand for $O(\cdot)\log^{O(1)}(\cdot)$. Similar remarks apply to the notations $\gtrsim$, Ω, $\widetilde{\Omega}$, and $\asymp$, Θ, $\widetilde{\Theta}$.

The chapter is organized as follows. In §3.2, we begin by reviewing functional inequalities and their implications for the continuous-time convergence of the diffusion (3.1) in Rényi divergence. We then state our main theorems on the LMC algorithm in Section 3.3, and illustrate them with examples in §3.4. We give a technical exposition of our proof techniques in §3.5 and fill in the details in §3.6.

We conclude in §3.7 with a discussion of future directions of research.

3.2 Functional inequalities and continuous-time convergence

Our focus in this section is the convergence of the continuous-time Langevin diffusion (3.1) under various functional inequalities. Throughout the paper, we use Rényi divergences as measures of distance between two probability laws. The Rényi divergence of order $q \in (1, \infty)$ of μ from π is defined to be

$$\mathcal{R}_q(\mu \,\|\, \pi) := \frac{1}{q-1} \ln \left\| \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right\|_{L^q(\pi)}^q ,$$

where $\mathcal{R}_q(\mu \,\|\, \pi)$ is understood to be $+\infty$ if $\mu \not\ll \pi$. See §2.2.3 for background on these divergences.

3.2.1 Poincaré and log-Sobolev inequalities

In the context of sampling, the most well-studied functional inequalities are the *Poincaré inequality* (PI) and the *log-Sobolev inequality* (LSI); see §2.2 for background. We recall the basic definitions here. We say that π satisfies a PI with constant C_{PI} if, for all smooth functions $f : \mathbb{R}^d \to \mathbb{R}$, it holds that

$$\mathrm{var}_\pi(f) \le C_{\mathsf{PI}} \, \mathbb{E}_\pi[\|\nabla f\|^2] \,. \tag{PI}$$

Similarly, we say that π satisfies an LSI with constant C_{LSI} if for all smooth $f : \mathbb{R}^d \to \mathbb{R}$,

$$\mathrm{ent}_\pi(f^2) \le 2C_{\mathsf{LSI}} \, \mathbb{E}_\pi[\|\nabla f\|^2] \,, \tag{LSI}$$

where $\mathrm{ent}_\pi(f^2) := \mathbb{E}_\pi[f^2 \ln(f^2 / \mathbb{E}_\pi(f^2))]$. By a linearization argument, an LSI implies a PI with the same constant.

These functional inequalities are classically related to the ergodicity properties of the Langevin diffusion (3.1). Indeed, if π_t denotes the law of the diffusion at time t, then a PI is equivalent to

$$\chi^2(\pi_t \,\|\, \pi) \le \exp\Big(-\frac{2t}{C_{\mathsf{PI}}}\Big) \, \chi^2(\pi_0 \,\|\, \pi) \,, \qquad \text{for all } t \ge 0 \,,$$

whereas an LSI is equivalent to

$$\mathsf{KL}(\pi_t \,\|\, \pi) \le \exp\Big(-\frac{2t}{C_{\mathsf{LSI}}}\Big) \, \mathsf{KL}(\pi_0 \,\|\, \pi) \,, \qquad \text{for all } t \ge 0 \,.$$

Functional inequalities are particularly useful for high-dimensional non-log-concave sampling because they are preserved under a variety of common operations (see §2.2). This flexibility is key for capturing a wide variety of non-log-concave settings encountered in practice.

Before stating the convergence results, we recall from §2.2.3 that

$$\max\Big\{2\,\|\mu-\pi\|_{\mathsf{TV}}^2,\ \ln\big(1+\frac{1}{2C_{\mathsf{PI}}}\,W_2^2(\mu,\pi)\big),\ \mathsf{KL}(\mu\parallel\pi)\Big\}\le\mathcal{R}_2(\mu\parallel\pi)\,.$$

Note that in the Poincaré case, a T_2 transportation inequality does not necessarily hold, so a KL guarantee does not imply a matching W_2 guarantee; by working with Rényi divergences, we are able to provide a unified guarantee for all of these metrics simultaneously.

Improving upon the prior result of [CLL19], [VW19] showed that these inequalities also imply Rényi convergence for the diffusion.

Theorem 3.2.1 ([VW19, Theorems 3 and 5])**.** *Let $q\ge 2$, and let π_t denote the law of the continuous-time Langevin diffusion (3.1) at time t.*

1. *If π satisfies (LSI), then*

$$\partial_t\mathcal{R}_q(\pi_t\parallel\pi)\le-\frac{2}{qC_{\mathsf{LSI}}}\,\mathcal{R}_q(\pi_t\parallel\pi)\,.$$

2. *If π satisfies (PI), then*

$$\partial_t\mathcal{R}_q(\pi_t\parallel\pi)\le-\frac{2}{qC_{\mathsf{PI}}}\times\begin{cases}1\,, & \text{if } \mathcal{R}_q(\pi_t\parallel\pi)\ge 1\,,\\ \mathcal{R}_q(\pi_t\parallel\pi)\,, & \text{if } \mathcal{R}_q(\pi_t\parallel\pi)\le 1\,.\end{cases}$$

The above result states that under LSI, the Rényi divergence decays exponentially fast whereas under PI, dissipation can be explained in two phases; an initial phase of *slow* decay followed by exponential convergence. Thus, to obtain $\mathcal{R}_q(\pi_T\parallel\pi)\le\varepsilon^2$, it suffices to have

1. $T\ge\Omega\Big(qC_{\mathsf{LSI}}\ln\dfrac{\mathcal{R}_q(\pi_0\parallel\pi)}{\varepsilon^2}\Big)$ and 2. $T\ge\Omega\Big(qC_{\mathsf{PI}}\Big(\mathcal{R}_q(\pi_0\parallel\pi)+\ln\dfrac{1}{\varepsilon}\Big)\Big)$

under LSI and PI respectively.

■ 3.2.2 Latała–Oleszkiewicz inequalities

In order to interpolate between the Poincaré and log-Sobolev cases, we consider a family of functional inequalities known as Latała–Oleszkiewicz inequalities

(LOI) [LO00]. We say that π satisfies an LOI of order $\alpha \in [1,2]$ and constant $C_{\mathsf{LOI}(\alpha)}$ if for all smooth $f : \mathbb{R}^d \to \mathbb{R}$,

$$\sup_{p \in (1,2)} \frac{\mathbb{E}_\pi(f^2) - \mathbb{E}_\pi(f^p)^{2/p}}{(2-p)^{2\,(1-1/\alpha)}} \le C_{\mathsf{LOI}(\alpha)}\, \mathbb{E}_\pi[\|\nabla f\|^2]\,. \tag{LOI}$$

An LOI of order 1 is equivalent to a PI, and an LOI of order 2 is equivalent to an LSI. More generally, an LOI of order α captures measures whose potentials "have tail growth α"; indeed, two notable examples of distributions satisfying the LOI of order α are $\pi(x) \propto \exp(-\sum_{i=1}^d |x_i|^\alpha)$ and $\pi(x) \propto \exp(-\|x\|^\alpha)$ [LO00; Bar01]. The LOI is well-studied because it captures intermediate forms of concentration and is related to a number of other important inequalities, such as Sobolev inequalities; we refer readers to [LO00; Bar01; BR03; Cha04; Bou+05; Wan05; BCR06; BCR07; CGG07; Goz10].

As our first result in this section, we extend Theorem 3.2.1 to cover the LOI case. Our proof, which uses as an intermediary the super Poincaré inequality introduced in [Wan00], is deferred to §3.6.1.

Theorem 3.2.2. *Let $q \ge 2$, and let π_t denote the law of the continuous-time Langevin diffusion (3.1) at time t. Suppose that π satisfies (LOI) with order α. Then, it holds that*

$$\partial_t \mathcal{R}_q(\pi_t \,\|\, \pi) \le -\frac{1}{68 q C_{\mathsf{LOI}(\alpha)}} \times \begin{cases} \mathcal{R}_q(\pi_t \,\|\, \pi)^{2-2/\alpha}\,, & \text{if } \mathcal{R}_q(\pi_t \,\|\, \pi) \ge 1\,, \\ \mathcal{R}_q(\pi_t \,\|\, \pi)\,, & \text{if } \mathcal{R}_q(\pi_t \,\|\, \pi) \le 1\,. \end{cases}$$

The above theorem can be used to obtain $\mathcal{R}_q(\pi_T \,\|\, \pi) \le \varepsilon^2$ whenever

$$T \ge \Omega\Big(q C_{\mathsf{LOI}(\alpha)} \Big(\frac{\mathcal{R}_q(\mu_0 \,\|\, \pi)^{2/\alpha - 1} - 1}{2/\alpha - 1} + \ln \frac{1}{\varepsilon}\Big)\Big)\,;$$

we refer to Lemma 3.6.16 for details. We also remark that Theorem 3.2.2 reduces to Theorem 3.2.1 in the edge cases $\alpha = 2$ (LSI) and $\alpha = 1$ (PI) up to an absolute constant. For $\alpha \in (1,2)$, the initial phase of convergence interpolates between the *slow* decay induced by PI and the exponential decay under LSI.

■ 3.2.3 Modified log-Sobolev inequalities

In addition, we also consider the modified log-Sobolev inequality (MLSI) used in [EH21]. The MLSI of order $\alpha_0 \in [-1,2]$ states that for all $f : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}_\pi(f^2) = 1$,

$$\mathrm{ent}_\pi(f^2) \le 2 C_{\mathsf{MLSI}} \inf_{p \ge 2} \big\{ \mathbb{E}_\pi[\|\nabla f\|^2]^{1-\delta(p)}\, \widetilde{\mathfrak{m}}_p\big((1+f^2)\,\pi\big)^{\delta(p)} \big\}\,, \tag{MLSI}$$

where $\delta(p)$ and $\widetilde{\mathfrak{m}}_p(\mu)$ for a measure μ (not necessarily a probability measure) are given as

$$\delta(p) := \frac{2-\alpha_0}{p+2-2\alpha_0}\,, \qquad \widetilde{\mathfrak{m}}_p(\mu) := \int (1+\|\cdot\|^2)^{p/2}\,\mathrm{d}\mu\,.$$

The inequality (MLSI) is a careful refinement of [TV00], and provides convergence guarantees for both the Langevin diffusion and LMC under various tail growth conditions [EH21]. It is similar to log-Nash inequalities [BZ99; Zeg01], yet the main focus of the latter is infinite-dimensional semigroups. We consider (MLSI) as used in [EH21] since other MLSI-type results are stated by absorbing various dimension-dependent constants into C_{MLSI}, and thus they cannot provide sharp rates for LMC.

For technical reasons, we also pair this assumption with a concentration property of the target: for some $\mathfrak{m} \geq 0$ and $\alpha \in [0,1]$,

$$\pi\{\|\cdot\| \geq \mathfrak{m} + \lambda\} \leq 2\exp\Big\{-\Big(\frac{\lambda}{C_{\mathsf{tail}}}\Big)^{\alpha_1}\Big\}\,, \qquad \text{for all } \lambda \geq 0\,. \tag{α_1-tail}$$

The parameters α_0 and α_1 are analogous to the parameter α in the LOI; we refer to [EH21] and the examples in §3.4 for further discussion.

Similarly to Theorem 3.2.2, we can prove a quantitative continuous-time convergence rate for the Langevin diffusion (3.1) under (MLSI) and (α_1-tail). The proof is deferred to §3.6.5.

Theorem 3.2.3. *Suppose that π satisfies the conditions (MLSI) and (α_1-tail), and assume that $\varepsilon^{-1}, \mathfrak{m}, C_{\mathsf{MLSI}} \geq 1$ and that $\mathfrak{m}, C_{\mathsf{tail}}, \mathcal{R}_{2q}(\pi_0 \,\|\, \pi) \leq d^{O(1)}$. Let $(\pi_t)_{t\geq 0}$ denote the law of the continuous-time Langevin diffusion (3.1). Then, it holds that $\mathcal{R}_q(\pi_T \,\|\, \pi) \leq \varepsilon^2$ for*

$$T \geq \Omega\Big(qC_{\mathsf{MLSI}}^2\,\big(\mathfrak{m} + qC_{\mathsf{tail}}\,\mathcal{R}_{2q}(\pi_0 \,\|\, \pi)^{1/\alpha_1}\big)^{2-\alpha_0}\,\mathrm{polylog}\,\frac{d\,\mathcal{R}_q(\pi_0 \,\|\, \pi)}{\varepsilon}\Big)\,.$$

We remark that when $\alpha_0 = \alpha_1 = \alpha$, the dependence on the Rényi divergence at initialization in Theorems 3.2.2 and 3.2.3 match up to a logarithmic factor, and hence LOI and MSLI provide similar results in continuous time. However, as we discuss in Section 3.4, MLSI is useful for treating certain examples in which the LOI constant $C_{\mathsf{LOI}(\alpha)}$ may be dimension-dependent whereas C_{MLSI} is not.

■ 3.3 Main results on Langevin Monte Carlo

In this section, we present our main results on the Rényi convergence of LMC. Denoting the step size with $h > 0$, the LMC algorithm is defined by the iteration

$$X_{(k+1)h} = X_{kh} - h\,\nabla V(X_{kh}) + \sqrt{2h}\,\xi_k\,, \qquad k \in \mathbb{N}\,, \tag{LMC}$$

where $(\xi_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. standard Gaussian variables. Here, the indexing of the LMC iterates is chosen so that the iterate X_{kh} is comparable to the continuous-time diffusion (3.1) at time kh. We let μ_{kh} denote the law of X_{kh}.

Our first result deals with the LSI and gradient Lipschitz case.

Theorem 3.3.1. *Assume that π satisfies (LSI) and that ∇V is L-Lipschitz; assume for simplicity that $C_{\mathsf{LSI}}, L \geq 1$ and $q \geq 3$. Let μ_{Nh} denote the N-th iterate of LMC with step size h satisfying $0 < h < 1/(192q^2 C_{\mathsf{LSI}} L^2)$. Then, for all $N \geq N_0$, it holds that*

$$\mathcal{R}_q(\mu_{Nh} \,\|\, \pi) \leq \exp\Big(-\frac{(N-N_0)h}{4C_{\mathsf{LSI}}}\Big)\, \mathcal{R}_2(\mu_0 \,\|\, \pi) + \widetilde{O}(dhqC_{\mathsf{LSI}}L^2)\,,$$

where $N_0 = \lceil \frac{2C_{\mathsf{LSI}}}{h} \ln(q-1) \rceil$. In particular, for $h = \widetilde{\Theta}(\frac{\varepsilon^2}{dqC_{\mathsf{LSI}}L^2} \min(1, \frac{d}{q\varepsilon^2}))$,

$$\mathcal{R}_q(\mu_{Nh} \,\|\, \pi) \leq \varepsilon^2\,, \qquad \text{for all } N \geq \widetilde{\Omega}\Big(\frac{dqC_{\mathsf{LSI}}^2 L^2 \log \mathcal{R}_2(\mu_0 \,\|\, \pi)}{\varepsilon^2} \max\big\{1, \frac{q\varepsilon^2}{d}\big\}\Big)\,.$$

The comparison of Theorem 3.3.1 with [VW19; GT20; EHZ22] is summarized as Table 3.1. Since our guarantee is stable with respect to the number of iterations N, we can let $N \to \infty$ and obtain an estimate on the asymptotic bias of (LMC) in Rényi divergence; this answers an open question of [VW19].

Corollary 3.3.2. *Assume that π satisfies (LSI) and that ∇V is L-Lipschitz; assume for simplicity that $C_{\mathsf{LSI}}, L \geq 1$. Let $\mu_\infty^{(h)}$ denote the stationary distribution of LMC with step size h satisfying $0 < h < 1/(192q^2 C_{\mathsf{LSI}} L^2)$. Then,*

$$\mathcal{R}_q(\mu_\infty^{(h)} \,\|\, \pi) \leq \widetilde{O}(dhqC_{\mathsf{LSI}}L^2)\,.$$

Source	Assumption	Metric	Complexity	Stable?
[VW19]	(LSI)	KL ($q = 1$)	$dC_{\mathsf{LSI}}^2 L^2/\varepsilon^2$	✓
[GT20]	C_{SC}^{-1}-SLC	Rényi	$dq^2 C_{\mathsf{SC}}^4 L^4/\varepsilon^4$	✗
[EHZ22]	C_{SC}^{-1}-SLC	Rényi	$dq^4 C_{\mathsf{SC}}^4 L^4/\varepsilon^2$	✗
Theorem 3.3.1	(LSI)	Rényi	$dqC_{\mathsf{LSI}}^2 L^2/\varepsilon^2$	✓

Table 3.1: We compare the guarantee of Theorem 3.3.1 with prior results, omitting polylogarithmic factors. "SLC" refers to "strongly log-concave", and the last column refers to whether the bound is stable as the number of iterations of LMC tends to infinity. The complexity bound in the last row is stated for moderate values of q; when $q \gg d/\varepsilon^2$, then the dependence on q becomes $\widetilde{O}(q^2)$.

Extending the techniques of Theorem 3.3.1, we next give a result for the log-concave (which implies (PI)) and gradient Lipschitz case.

Theorem 3.3.3. *Assume that π is log-concave (and hence satisfies (PI)) and that ∇V is L-Lipschitz. Assume that V is minimized at 0. Let μ_{Nh} denote the N-th iterate of LMC with step size h satisfying $h = \widetilde{\Theta}(\frac{\varepsilon^2}{dq^2 C_{\mathsf{PI}} L^2} \min\{1, \frac{1}{q\varepsilon^2}, \frac{dC_{\mathsf{PI}}}{\varepsilon^2 L}\})$ and initialized at $\mu_0 = \mathsf{normal}(0, L^{-1} I_d)$. Then,*

$$\mathcal{R}_q(\mu_{Nh} \,\|\, \pi) \le \varepsilon^2 \qquad \text{after } N = \widetilde{\Theta}\Big(\frac{d^2 q^3 C_{\mathsf{PI}}^2 L^2}{\varepsilon^2} \max\big\{1, q\varepsilon^2, \frac{\varepsilon^2 L}{dC_{\mathsf{PI}}}\big\}\Big) \text{ iterations}.$$

In Table 3.2, we compare Theorem 3.3.3 with the prior works [DMM19; Dwi+19; Che+20a; DKR22]. Compared to these results, Theorem 3.3.3 is only beaten by the result for modified MALA (for which our result reads $\widetilde{O}(d^2/\varepsilon^2)$ whereas the result for modified MALA is $\widetilde{O}(d^2/\varepsilon^{3/2})$). Moreover, our result is given in the strongest metric (Rényi divergence). However, better results for the log-concave case will be obtained in §4 and §6.

Source	Algorithm	Metric	Complexity
[DMM19]	averaged LMC	$\sqrt{\mathrm{KL}}$	d^2/ε^4
[Dwi+19; Che+20a]	modified MALA	TV	$d^2/\varepsilon^{3/2}$
[DKR22]	modified LMC	W_1	d^2/ε^4
[DKR22]	modified LMC	W_2	d^2/ε^6
[DKR22]	modified ULMC	W_1	d^2/ε^3
[DKR22]	modified ULMC	W_2	d^2/ε^5
Theorem 3.3.3	LMC	$\sqrt{\text{Rényi}}$	d^2/ε^2

Table 3.2: We compare guarantees for sampling from an isotropic log-concave distribution with $C_{\mathsf{PI}}, L = O(1)$. MALA refers to the Metropolis-adjusted Langevin algorithm, whereas ULMC refers to the underdamped Langevin Monte Carlo algorithm.

Subsequently, we consider the general case of an LOI. We also assume weak smoothness for some $s \in (0, 1]$ and $L > 0$:

$$\|\nabla V(x) - \nabla V(y)\| \le L \|x - y\|^s \qquad \text{for all } x, y \in \mathbb{R}^d. \tag{s-Hölder}$$

We note that the LO order α and the Hölder exponent s need to satisfy $s + 1 \ge \alpha$.

Theorem 3.3.4. *Assume that the potential satisfies $\nabla V(0) = 0$, (LOI) of order α, and (s-Hölder). For simplicity, assume that $\varepsilon^{-1}, \mathfrak{m}, C_{\mathsf{LOI}(\alpha)}, L, \mathcal{R}_2(\mu_0 \,\|\, \hat{\pi}) \ge 1$ and $q \ge 2$; here, $\mathfrak{m} := \int \|\cdot\| \,\mathrm{d}\pi$ and $\hat{\pi}$ is a slightly modified version of π which is introduced in the analysis (§3.6.4). Then, LMC with an appropriate step size (given in (3.16)) satisfies $\mathcal{R}_q(\mu_{Nh} \,\|\, \pi) \le \varepsilon^2$ after*

$$N = \widetilde{\Theta}_s\Big(\frac{dq^{1+2/s} C_{\mathsf{LOI}(\alpha)}^{1+1/s} L^{2/s} \, \mathcal{R}_{2q-1}(\mu_0 \,\|\, \pi)^{(2/\alpha - 1)(1+1/s)}}{\varepsilon^{2/s}}$$

$$\times \max\Big\{1, q^{1/s}\varepsilon^{2/s}, \frac{\mathfrak{m}^s}{d}, \frac{\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})^{s/2}}{d}\Big\}\Big)$$

iterations. Here, $\widetilde{\Theta}_s(\cdot)$ hides polylogarithmic factors and constants depending only on s.

We now make a few remarks to simplify the rate. First, although initialization is more subtle in the non-log-concave case, it is reasonable to suppose that the quantities $\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})$, $\mathcal{R}_{2q-1}(\mu_0 \,\|\, \pi)$ are $\widetilde{O}(d)$; we defer a detailed discussion of initialization to §3.6.6. Next, it is also reasonable to assume[3] $\mathfrak{m} = O(d)$, in which case the third term in the maximum will never dominate. Focusing on the dependence on the dimension and target accuracy, we therefore obtain the simplified rate $\widetilde{O}(d^{(2/\alpha)\,(1+1/s)-1/s}/\varepsilon^{2/s})$; in particular, in the smooth ($s = 1$) case, the rate is $\widetilde{O}(d^{4/\alpha-1}/\varepsilon^2)$. Regarding prior works which handle a wide variety of growth rates and smoothness conditions for the potential, the closest to the present work is [EH21], which obtains a rate of $\widetilde{O}(d^{(2/\alpha+\mathbb{1}\{\alpha=1\})\,(1+1/s)-1}/\varepsilon^{2/s})$ for potentials of tail growth α satisfying (s-Hölder); note that our rate is strictly better as soon as $s < 1$ and avoids the jump in the rate at $\alpha = 1$. We emphasize, however, that despite the superficial similarity with [EH21], our result is the first one proven under a purely functional analytic condition on the target (together with weak smoothness).

Remark 3.3.5. *The case $\alpha = 1$ yields the bound $\widetilde{O}(d^{2+1/s}q^{1+2/s}C_{\mathsf{PI}}^{1+1/s}L^{2/s}/\varepsilon^{2/s})$ for LMC under the Poincaré inequality and weak smoothness. In the case $\alpha = 2$ and $s = 1$ (LSI and smooth case), the rate reduces to $\widetilde{O}(dq^3C_{\mathsf{LSI}}^2L^2/\varepsilon)$, which recovers the guarantee of Theorem 3.3.1 up to the dependence on q.*

When the LOI constant $C_{\mathsf{LOI}(\alpha)}$ is dimension-dependent, Theorem 3.3.4 may not give the sharpest rates. We therefore complement Theorem 3.3.4 with a result assuming (MLSI).

Theorem 3.3.6. *Assume that the potential satisfies $\nabla V(0) = 0$, (MLSI) of order α_0, (α_1-tail), and (s-Hölder). For simplicity, we also consider the regime $\varepsilon^{-1}, \mathfrak{m}, C_{\mathsf{MLSI}}, C_{\mathsf{tail}}, L, \mathcal{R}_2(\mu_0 \,\|\, \hat{\pi}) \geq 1$, $q \geq 2$, and $\mathfrak{m}, C_{\mathsf{tail}}, \mathcal{R}_2(\pi_0 \,\|\, \pi) \leq d^{O(1)}$; here, $\hat{\pi}$ is a slightly modified version of π which is introduced in the analysis (§3.6.4). Then, LMC with appropriately chosen step size (given in (3.18)) satisfies $\mathcal{R}_q(\mu_{Nh} \,\|\, \pi) \leq \varepsilon^2$ after*

$$N = \widetilde{\Theta}\Big(\frac{d\,\mathcal{R}_{2q}(\mu_0 \,\|\, \pi)^{(2-\alpha_0)\,(1+1/s)/\alpha_1}}{\varepsilon^{2/s}}$$

[3] This holds for, e.g., the potentials $V(x) = \|x\|^\alpha$ for all $\alpha \in [1, 2]$.

$$\times \max\Big\{1, \varepsilon^{2/s}, \frac{\mathfrak{m}^s}{d}, \frac{\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})^{s/2}}{d}, \Big(\frac{\mathfrak{m}}{\mathcal{R}_{2q}(\mu_0 \,\|\, \pi)^{1/\alpha_1}}\Big)^{(2-\alpha_0)/s}\Big\}\Big)$$

iterations. Here, the $\widetilde{\Theta}(\cdot)$ notation hides polylogarithmic factors as well as constants depending on α_0, α_1, q, s, C_{MLSI}, C_{tail}, and L; a more precise statement is given in §3.6.5.

For potentials of tail growth $\alpha \in (1, 2]$, we can suppose that (MLSI) and (α_1-tail) are satisfied with $\alpha_0 = \alpha_1 = \alpha$, where we take $\mathfrak{m} = O(d^{1/\alpha})$. Also, assuming $\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi}), \mathcal{R}_{2q}(\mu_0 \,\|\, \pi) = O(d)$, the rate is then $\widetilde{O}(d^{(2/\alpha)(1+1/s)-1/s}/\varepsilon^{2/s})$ as before. As discussed in the next section, the case $\alpha = 1$ is special and (MLSI) may not hold with $\alpha_0 = \alpha$.

Remark 3.3.7. *A number of recent works [DMM19; Cha+20; NDC21; LC22; Leh22] consider non-smooth and mixed-smooth potentials. By incorporating Gaussian smoothing, it seems possible to extend our techniques to cover these settings, but we do not pursue this direction here.*

■ 3.4 Illustrative examples

In this section, we illustrate our results on simple examples and compare our guarantees with prior work.

Example 3.4.1 (tail growth $\alpha \in (1, 2]$)**.** Consider the target $\pi_\alpha(x) \propto \exp(-\|x\|^\alpha)$ for $\alpha \in (1, 2]$, which satisfies (LOI) of order α and (s-Hölder) with $s = \alpha - 1$. Since π_α satisfies (PI) with $C_{\mathsf{PI}} = \Theta(d^{2/\alpha-1})$ [Bob03], then Theorem 3.3.4 does not yield a good result. Previously, [EH21] showed that π_α satisfies (MLSI) of order α, obtaining the complexity $\widetilde{O}(d^{(3-\alpha)/(\alpha-1)}/\varepsilon^{2/(\alpha-1)})$ to achieve ε^2-accuracy in KL divergence for this target. From Theorem 3.3.6, we have improved this rate to $\widetilde{O}((d/\varepsilon^2)^{1/(\alpha-1)})$ in Rényi divergence. Since (MLSI) is stable under bounded perturbations, the same rate holds for appropriately perturbed potentials such as $V(x) = \|x\|^\alpha + \cos\|x\|$.

Due to the use of the weighted CKP inequality [BV05], their KL bound yields $\widetilde{O}(d^{(5-\alpha)/(\alpha-1)}/\varepsilon^{2\alpha/(\alpha-1)})$ complexity to reach ε accuracy in the W_α metric. On the other hand, Theorem 3.3.6 together with the Poincaré inequality yields the complexity $\widetilde{O}(d^{2/(\alpha(\alpha-1))}/\varepsilon^{2/(\alpha-1)})$ to obtain ε accuracy in the W_2 metric. Hence, we have both improved the rate in W_α and proven a new guarantee in W_2 which previously could not be reached at all. ♢

Example 3.4.2 (tail growth $\alpha \in (1, 2]$ for smoothed potential)**.** Consider the target $\pi_\alpha(x) \propto \exp(-(1 + \|x\|^2)^{\alpha/2})$, which satisfies (LOI) of order α and (s-Hölder)

with $s = 1$ (i.e., ∇V is Lipschitz). Previously, [EH21] obtained the complexity $\widetilde{O}(d^{(4-\alpha)/\alpha}/\varepsilon^2)$ in KL divergence and $\widetilde{O}(d^{(4+\alpha)/\alpha}/\varepsilon^{2\alpha})$ in W_α. From Theorem 3.3.6, we have obtained the rate $\widetilde{O}(d^{(4-\alpha)/\alpha}/\varepsilon^2)$ in Rényi divergence and $\widetilde{O}(d^{(6-2\alpha)/\alpha}/\varepsilon^2)$ in W_2. As before, this rate is stable under suitable perturbations of the potential. ♢

Example 3.4.3 (tail growth $\alpha = 1$ for smoothed potential)**.** The case of $\alpha = 1$ is worth considering separately for comparison purposes. Consider the target $\pi_1(x) \propto \exp(-\sqrt{1 + \|x\|^2})$, which satisfies ($s$-Hölder) with $s = 1$ (i.e., ∇V is Lipschitz). Previously, [EH21] showed that π_1 satisfies (MLSI) with $\alpha_0 = -O(\frac{1}{\log d})$ and $C_{\mathsf{MLSI}} = O(\log d)$; also, π_1 satisfies (α_1-tail) with $\alpha_1 = 1$. Using this, they obtained the complexity $\widetilde{O}(d^5/\varepsilon^2)$ in KL divergence, whereas Theorem 3.3.6 implies the same rate in Rényi divergence. We also remark that their rate only holds for sufficiently small perturbations (e.g., their analysis does not cover the potential $V(x) = \|x\| + \cos\|x\|$) due to the need to preserve a dissipativity assumption, whereas our result has no such requirement. This highlights a benefit of working without dissipativity conditions.

Here, Theorem 3.3.3 applies to π_1 with $C_{\mathsf{PI}} = O(d)$ [Bob03] and yields a rate of $\widetilde{O}(d^4/\varepsilon^2)$ in Rényi divergence; in contrast, [DMM19] yields a rate of $\widetilde{O}(d^3/\varepsilon^4)$ in KL divergence (started from a distribution with $W_2^2(\mu_0, \pi_1) = O(d^2)$) for averaged LMC, and [Dwi+19; Che+20a] yields a rate of $\widetilde{O}(d^{3.5}/\varepsilon^{1.5})$ in $\|\cdot\|_{\mathsf{TV}}$ for modified MALA, although none of these rates is stable under perturbation. ♢

Example 3.4.4 (tail growth $\alpha \in [1, 2]$ for smoothed product potential)**.** For $x \in \mathbb{R}^d$, let $\langle x \rangle_i := \sqrt{1 + x_i^2}$. Consider the target $\pi_\alpha(x) \propto \exp(-\|\langle x \rangle\|_\alpha^\alpha)$, which satisfies (LOI) of order α [see LO00] and (s-Hölder) with $s = 1$ (i.e., ∇V is Lipschitz). The result of [EH21] implies a complexity of $\widetilde{O}(d^{(4-\alpha)/\alpha}/\varepsilon^2)$ in KL divergence and $\widetilde{O}(d^{(4+\alpha)/\alpha}/\varepsilon^{2\alpha})$ in W_α for $\alpha \in (1, 2]$, and $\widetilde{O}(d^5/\varepsilon^2)$ in KL divergence when $\alpha = 1$. From Theorem 3.3.4, we have obtained the rate $\widetilde{O}(d^{(4-\alpha)/\alpha}/\varepsilon^2)$ in Rényi divergence and hence also W_2 for all $\alpha \in [1, 2]$; in particular, there is no jump in the rate at $\alpha = 1$. ♢

Example 3.4.5 (LSI case with weakly smooth potential)**.** We also compare the results when $\alpha = 2$ and $s \in (0, 1]$. In this case, [Cha+20] obtained the rate $\widetilde{O}(d^{(2+s)/s}/\varepsilon^{2/s})$ in $\|\cdot\|_{\mathsf{TV}}$ for strongly log-concave distributions, whereas [EH21] obtained the rate $\widetilde{O}((d/\varepsilon^2)^{1/s})$ in KL divergence for perturbations of strongly log-concave distributions. In contrast, Theorem 3.3.4 yields the rate $\widetilde{O}(d/\varepsilon^{2/s})$ in Rényi divergence under (LSI). An example of such a potential is given by $V(x) = \frac{1}{2}\|x\|^2 + \cos(\|x\|^{1+s})$. ♢

3.5 Technical overview

3.5.1 Adapting the interpolation method to Rényi divergences

In the proof of Theorem 3.3.1, we follow the interpolation method of [VW19]. Namely, we introduce the following interpolation of (LMC): for $t \in [kh, (k+1)h]$,

$$X_t = X_{kh} - (t - kh)\,\nabla V(X_{kh}) + \sqrt{2}\,(B_t - B_{kh})\,, \tag{3.2}$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion, and let μ_t denote the law of X_t. Then, [VW19] derives the following differential inequality for the KL divergence:

$$\partial_t \mathsf{KL}(\mu_t \,\|\, \pi) \leq -\frac{3}{4} \times \underbrace{4\,\mathbb{E}_\pi[\|\nabla\sqrt{\rho_t}\|^2]}_{\text{Fisher information}} + \underbrace{\mathbb{E}[\|\nabla V(X_t) - \nabla V(X_{kh})\|^2]}_{\text{discretization error}}, \tag{3.3}$$

where we write $\rho_t \coloneqq \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}$. This inequality is an analogue of the celebrated de Bruijn identity from information theory for the interpolated process. Assuming that π satisfies (LSI) and that ∇V is L-Lipschitz, the Fisher information upper bounds the KL divergence and the discretization error is shown to be of order $O(dh^2L^2)$; this then yields a convergence guarantee in KL divergence.

The analogous differential inequality for the Rényi divergence is

$$\partial_t \mathcal{R}_q(\mu_t \,\|\, \pi) \leq -\underbrace{\frac{3}{q}\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)}}_{\text{Rényi Fisher information}} + \underbrace{q\,\frac{\mathbb{E}[\rho_t^{q-1}(X_t)\,\|\nabla V(X_t) - \nabla V(X_{kh})\|^2]}{\mathbb{E}_\pi(\rho_t^q)}}_{\text{discretization error}}. \tag{3.4}$$

(See the proof of [EHZ22, Lemma 6]; to make the chapter more self-contained, we also provide a derivation in Proposition 3.6.2.) Note that the $q = 1$ case of the above inequality formally corresponds to (3.3). Next, as shown in [VW19, Lemma 5], the Rényi Fisher information indeed upper bounds the Rényi divergence under an LSI. However, the discretization term is now far trickier to control.

Write $\psi_t \coloneqq \rho_t^{q-1} / \mathbb{E}_\pi(\rho_t^q)$. Observing that $\mathbb{E}\,\psi_t(X_t) = 1$, the discretization term can be written as an expectation under a *change of measure*:

$$\text{discretization error} = q\,\widetilde{\mathbb{E}}[\|\nabla V(X_t) - \nabla V(X_{kh})\|^2]\,,$$

where $\widetilde{\mathbb{E}}$ is the expectation under the measure $\widetilde{\mathbb{P}}$ defined via $\frac{\mathrm{d}\widetilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}} = \psi_t(X_t)$. Also, using the Lipschitzness of ∇V, we obtain

$$\|\nabla V(X_t) - \nabla V(X_{kh})\|^2 \leq 2h^2L^2\,\|\nabla V(X_{kh})\|^2 + 4L^2\,\|B_t - B_{kh}\|^2\,.$$

Hence, our task is to bound the expectation of these two terms under a complicated change of measure.

Towards that end, consider first the Brownian motion term. Using the Donsker–Varadhan variational principle, for any random variable X,

$$\widetilde{\mathbb{E}} X \le \mathsf{KL}(\widetilde{\mathbb{P}} \,\|\, \mathbb{P}) + \ln \mathbb{E} \exp X \,.$$

Applying this to $X = c\,(\|B_t - B_{kh}\| - \mathbb{E}\|B_t - B_{kh}\|)^2$ for a constant $c > 0$ to be chosen later, we can bound

$$\begin{aligned}\widetilde{\mathbb{E}}[\|B_t - B_{kh}\|^2] &\le 2\,\mathbb{E}[\|B_t - B_{kh}\|^2] + \frac{2}{c}\,\widetilde{\mathbb{E}} X \\ &\le 2\,\mathbb{E}[\|B_t - B_{kh}\|^2] \\ &\quad + \frac{2}{c}\left\{\mathsf{KL}(\widetilde{\mathbb{P}} \,\|\, \mathbb{P}) + \ln \mathbb{E} \exp\bigl(c\,(\|B_t - B_{kh}\| - \mathbb{E}\|B_t - B_{kh}\|)^2\bigr)\right\}. \qquad (3.5)\end{aligned}$$

Note that the first and third terms in the right-hand side of the above expression are expectations under the original measure $\mathbb{P}$, and can therefore be controlled; to ensure that the third term is bounded, we can take $c \asymp 1/h$. For the second term, a surprising calculation involving a judicious application of the LSI for π (see (3.8), (3.9), and (3.10)) shows that it is bounded by h times the Rényi Fisher information, and can therefore be absorbed into the first term of the differential inequality (3.4) for h sufficiently small.

The expectation of the drift term $\|\nabla V(X_{kh})\|^2$ under the change of measure can also be handled via similar methods, but this can be bypassed via a duality principle for the Fisher information; see Lemma 3.6.3. We also remark that naïvely, this proof incurs a cubic dependence on q, but this can be sharpened via an argument based on hypercontractivity (Proposition 3.6.4).

In the above proof outline, the LSI for π plays a crucial role in the arguments. In Theorem 3.3.3, we show that the method can be somewhat extended to the case when π does not satisfy an LSI, but is instead assumed to be (weakly) log-concave. In this case, we show that with an appropriate Gaussian initialization, the law μ_{kh} of the *iterate* X_{kh} of (LMC) satisfies an LSI, albeit with a constant which grows with the number of iterations (Lemma 3.6.5). In turn, this fact together with a suitable modification of the preceding proof strategy also allows us to obtain a convergence guarantee in this case (see §3.6.3 for details).

■ 3.5.2 Controlling discretization error via Girsanov's theorem

In the general case of a weaker functional inequality and smoothness condition, the preceding arguments do not apply. Instead, we start with the weak triangle

inequality for the Rényi divergence (when $q \geq 2$; see Lemma 2.2.23):

$$\mathcal{R}_q(\mu_T \,\|\, \pi) \lesssim \mathcal{R}_{2q}(\mu_T \,\|\, \pi_T) + \mathcal{R}_{2q-1}(\pi_T \,\|\, \pi)\,.$$

Here, $(\mu_t)_{t\geq 0}$ is the law of the interpolated process (3.2), whereas $(\pi_t)_{t\geq 0}$ is the law of the continuous-time Langevin diffusion (3.1) initialized at a draw from μ_0. The second term is handled via the continuous-time convergence results, either under the LOI (Theorem 3.2.2) or under the MLSI (Theorem 3.2.3), and the crux of the proof is to control the first term (the discretization error).

The discretization error $\mathcal{R}_{2q}(\mu_T \,\|\, \pi_T)$ was controlled in the prior works [GT20; EHZ22] via the adaptive composition theorem, albeit under stronger assumptions (strong convexity/dissipativity). Briefly, this theorem controls the Rényi divergence between the paths of the interpolated and original (continuous-time) processes by summing up the contribution to the Rényi divergence in each infinitesimal time step. In turn, due to the Brownian motion driving the SDEs, this reduces to a computation of the Rényi divergence between Gaussians. Making this approach rigorous, however, requires first applying it to the discrete-time algorithm and then performing a cumbersome limiting argument. Here, we streamline this technique by instead invoking Girsanov's theorem from stochastic calculus.

First, the data-processing inequality (Lemma 2.2.19) implies that $\mathcal{R}_{2q}(\mu_T \,\|\, \pi_T) \leq \mathcal{R}_{2q}(P_T \,\|\, Q_T)$, where P_T and Q_T are measures on path space representing the laws of the trajectories (on the interval $[0, T]$) of the interpolated and diffusion processes respectively. Next, Girsanov's theorem provides a closed-form formula for the Radon–Nikodym derivative $\frac{\mathrm{d}P_T}{\mathrm{d}Q_T}$, which leads to the inequality

$$\mathcal{R}_{2q}(P_T \,\|\, Q_T) \leq \frac{1}{2\,(2q-1)} \ln \mathbb{E} \exp\Big(4q^2 \int_0^T \|\nabla V(Z_t) - \nabla V(Z_{\lfloor t/h \rfloor\, h})\|^2 \,\mathrm{d}t\Big)\,,$$

where $(Z_t)_{t\geq 0}$ is the continuous-time Langevin diffusion (3.1). The use of Girsanov's theorem for deriving quantitative estimates on the discretization error in this manner was likely first introduced in [DT12] for the KL divergence. However, to the best our knowledge, this work is the first to adapt the Girsanov technique to provide a complete Rényi convergence result for LMC.

Controlling the discretization error over an interval $[0, h]$ corresponding to a single iteration of LMC is straightforward using the tools of stochastic calculus (see also the calculation in §5). Extending this to the full time interval $[0, T]$ is more challenging; indeed, if we bound the discretization error on $[h, 2h]$ conditional on $(Z_t)_{t\in[0,h]}$, then the resulting bound depends on $\|Z_h\|^2$, which prevents us from straightforwardly iterating the one-step discretization bound. To address this, we instead control intermediate error terms conditioned on the event $\mathcal{E}_{\delta,T} := \max_{k\in\mathbb{N},\, kh\leq T} \|Z_{kh}\| \leq R_{\delta,T}$, and $R_{\delta,T}$ is chosen so that $\mathbb{P}(\mathcal{E}_{\delta,T}) \geq 1 - \delta$.

Subsequently, we can use Lemma 3.6.10 to remove the conditioning, and hence providing a bound on $\mathcal{R}_{2q}(P_T \,\|\, Q_T)$ if $R_{\delta,T}$ does not grow too fast in $1/\delta$; in particular, it is required that $R_{\delta,T} \lesssim \sqrt{\log(1/\delta)}$.[4]

The requirement on $R_{\delta,T}$ is equivalent to requiring that for each $t \in [0,T]$, the random variable Z_t has sub-Gaussian tails. Observe however that the stationary distribution π may not have sub-Gaussian tails under our assumption of an LOI (indeed, in the Poincaré case, π may only have subexponential tails). Nevertheless, if the initialization μ_0 has sub-Gaussian tails, then for each $t \in [0,T]$ it may still be the case that π_t has sub-Gaussian tails. This turns out to be true, but it is quite non-trivial to prove without any dissipativity conditions on the potential V, and therefore constitutes our primary technical challenge.

To overcome this challenge, we introduce a novel technique based on comparison of the diffusion (3.1) with an auxiliary Langevin diffusion $(\hat{\pi}_t)_{t\geq 0}$ corresponding to a modified stationary distribution $\hat{\pi}$. The distribution $\hat{\pi}$ is constructed to have sub-Gaussian tails. To transfer the sub-Gaussianity of $\hat{\pi}$ to π_t, we apply the following change of measure inequality: for probability measures μ and ν, and any event $E \subseteq \mathbb{R}^d$,

$$\mu(E) = \nu(E) + \int \mathbb{1}_E \left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\right) \mathrm{d}\nu \leq \nu(E) + \sqrt{\chi^2(\mu \,\|\, \nu)\, \nu(E)}\,,$$

where the last inequality is the Cauchy–Schwarz inequality. This simple inequality states that in order to control the probability of an event E under a measure μ in terms of its probability under ν, it suffices to control the chi-squared divergence between μ and ν. Applying this to our context, we can establish sub-Gaussian tail bounds for π_t if we can control the Rényi divergences $\mathcal{R}_2(\pi_t \,\|\, \hat{\pi}_t)$ and $\mathcal{R}_2(\hat{\pi}_t \,\|\, \hat{\pi})$; the former is again controlled via Girsanov's theorem. We stress that the auxiliary process $(\hat{\pi}_t)_{t\geq 0}$ is introduced only for analysis purposes and does not affect the implementation of the algorithm.

The details of this strategy are carried out in Section 3.6.4.

3.6 Proofs

3.6.1 Proof of Theorem 3.2.2

In this section, we prove Theorem 3.2.2 on the Rényi convergence of the continuous-time Langevin diffusion (3.1) under an LOI. Using capacity inequalities as an intermediary, [BCR06; Goz10] established the equivalence of LOI with other functional inequalities such as modified Sobolev inequalities. For our purposes,

[4] See, however, §6 for an alternative approach.

it is instead convenient to work with *super Poincaré inequalities*, which were introduced in [Wan00].

We say that π satisfies a super Poincaré inequality with function $\beta : \mathbb{R}_+ \to \mathbb{R}_+$ if for all smooth $f : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E}_\pi(f^2) \le \beta(s)\, \mathbb{E}_\pi[\|\nabla f\|^2] + s\, (\mathbb{E}_\pi|f|)^2 \qquad \text{for all } s \ge 1\,. \tag{3.6}$$

For $\alpha \in [1, 2]$, define the function $\beta_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ via

$$\beta_\alpha(s) := \frac{96 C_{\mathsf{LOI}(\alpha)}}{\ln(\mathrm{e} + s)^{2-2/\alpha}}\,.$$

Then, it is known that (LOI) with order α implies a super Poincaré inequality with function β_α [see Goz10, Remark 5.16]. The following proof is inspired by the proof of [VW19, Theorem 5].

Proof of Theorem 3.2.2. From [VW19, Lemma 6], we have

$$\partial_t \mathcal{R}_q(\pi_t \,\|\, \pi) = -\frac{4}{q}\, \frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)}\,,$$

where $\rho_t := \frac{\mathrm{d}\pi_t}{\mathrm{d}\pi}$. Applying the super Poincaré inequality (3.6) with $f = \rho_t^{q/2}$ and $\beta = \beta_\alpha$ yields

$$\begin{aligned}
\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2] &\ge \frac{1}{\beta_\alpha(s)}\, \mathbb{E}_\pi(\rho_t^q) - \frac{s}{\beta_\alpha(s)}\, \{\mathbb{E}_\pi(\rho_t^{q/2})\}^2 \\
&= \frac{1}{\beta_\alpha(s)} \exp\{(q-1)\, \mathcal{R}_q(\pi_t \,\|\, \pi)\} - \frac{s}{\beta_\alpha(s)} \exp\{(q-2)\, \mathcal{R}_{q/2}(\pi_t \,\|\, \pi)\}\,.
\end{aligned}$$

Using the fact that $\mathcal{R}_{q/2} \le \mathcal{R}_q$, we can further lower bound this by

$$\begin{aligned}
\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2] &\ge \frac{\exp\{(q-1)\, \mathcal{R}_q(\pi_t \,\|\, \pi)\}}{\beta_\alpha(s)} \left(1 - s \exp\{-\mathcal{R}_q(\pi_t \,\|\, \pi)\}\right) \\
&= \frac{\mathbb{E}_\pi(\rho_t^q)}{\beta_\alpha(s)} \left(1 - s \exp\{-\mathcal{R}_q(\pi_t \,\|\, \pi)\}\right).
\end{aligned}$$

We now distinguish two cases. If $\mathcal{R}_q(\pi_t \,\|\, \pi) \ge 1$, then we choose $s = \frac{1}{2} \exp\{\mathcal{R}_q(\pi_t \,\|\, \pi)\}$, yielding

$$\partial_t \mathcal{R}_q(\pi_t \,\|\, \pi) \le -\frac{2}{q\beta_\alpha(s)} = -\frac{\ln(\mathrm{e} + \frac{1}{2} \exp \mathcal{R}_q(\pi_t \,\|\, \pi))^{2-2/\alpha}}{48 q C_{\mathsf{LOI}(\alpha)}}$$

$$\leq -\frac{1}{68qC_{\mathsf{LOI}(\alpha)}}\,\mathcal{R}_q(\pi_t \,\|\, \pi)^{2-2/\alpha}\,.$$

Otherwise, if $\mathcal{R}_q(\pi_t \,\|\, \pi) \leq 1$, then we choose $s = 1$, yielding

$$\begin{aligned}\partial_t \mathcal{R}_q(\pi_t \,\|\, \pi) &\leq -\frac{4}{q\beta_\alpha(1)}\left(1 - \exp\{-\mathcal{R}_q(\pi_t \,\|\, \pi)\}\right) \leq -\frac{2}{q\beta_\alpha(1)}\,\mathcal{R}_q(\pi_t \,\|\, \pi)\\ &\leq -\frac{1}{68qC_{\mathsf{LOI}(\alpha)}}\,\mathcal{R}_q(\pi_t \,\|\, \pi)\,,\end{aligned}$$

where we used the elementary inequality $1 - \exp(-x) \geq x/2$ for $x \in [0, 1]$. □

■ 3.6.2 Proof of Theorem 3.3.1

Throughout this section, recall the notation $\rho_t := \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}$ and $\psi_t := \rho_t^{q-1} / \mathbb{E}_\pi(\rho_t^q)$.

We begin by proving the differential inequality (3.4). Although this has appeared in the previous works [VW19; EHZ22], we include the proofs for the sake of completeness.

Proposition 3.6.1. *Let $(\mu_t)_{t\geq 0}$ denote the law of the interpolation (3.2) of LMC. Then, for $t \in [kh, (k+1)h]$,*

$$\partial_t \mu_t = \operatorname{div}\left(\left\{\nabla \ln \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} + \mathbb{E}[\nabla V(X_{kh}) - \nabla V(X_t) \mid X_t = \cdot]\right\} \mu_t\right).$$

Proof. For $s, t \in \mathbb{R}_+$, let $\mu_{t|s}(\cdot \mid X_s)$ denote the conditional law of X_t given X_s, and let $\mu_{s,t}$ denote the joint law of (X_s, X_t). Conditioned on X_{kh}, the Fokker–Planck equation for the interpolation (3.2) takes the form

$$\partial_t \mu_{t|kh}(\cdot \mid X_{kh}) = \Delta \mu_{t|kh}(\cdot \mid X_{kh}) + \operatorname{div}\left(\nabla V(X_{kh})\, \mu_{t|kh}(\cdot \mid X_{kh})\right).$$

Taking the expectation over X_{kh} yields

$$\begin{aligned}\partial_t \mu_t &= \Delta \mu_t + \operatorname{div}(\nabla V\, \mu_t) + \int \operatorname{div}\left(\{\nabla V(x_{kh}) - \nabla V(\cdot)\}\, \mu_{t|kh}(\cdot \mid x_{kh})\right) \mathrm{d}\mu_{kh}(x_{kh})\\ &= \operatorname{div}\left(\nabla \ln \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\, \mu_t\right) + \operatorname{div}\left(\left(\int \{\nabla V(x_{kh}) - \nabla V(\cdot)\}\, \mathrm{d}\mu_{kh|t}(x_{kh} \mid \cdot)\right) \mu_t(\cdot)\right)\\ &= \operatorname{div}\left(\nabla \ln \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\, \mu_t\right) + \operatorname{div}\left(\{\mathbb{E}[\nabla V(X_{kh}) \mid X_t = \cdot] - \nabla V\}\, \mu_t\right).\end{aligned}$$

Combining the two terms yields the result. □

Proposition 3.6.2. *Let $(\mu_t)_{t\geq 0}$ denote the law of the interpolation (3.2) of LMC. Also, let $\rho_t \coloneqq \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}$ and $\psi_t \coloneqq \rho_t^{q-1} / \mathbb{E}_\pi(\rho_t^q)$. Then, for $t \in [kh, (k+1)h]$,*

$$\partial_t \mathcal{R}_q(\mu_t \,\|\, \pi) \leq -\frac{3}{q} \frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + q\,\mathbb{E}[\psi_t(X_t)\,\|\nabla V(X_t) - \nabla V(X_{kh})\|^2]\,.$$

Proof. For brevity, in this proof we write $\Delta_t \coloneqq \mathbb{E}[\nabla V(X_{kh}) \mid X_t = \cdot] - \nabla V$. Elementary calculus together with Proposition 3.6.1 yields

$$\begin{aligned}
\partial_t \mathcal{R}_q(\mu_t \,\|\, \pi) &= \frac{q}{(q-1)\,\mathbb{E}_\pi(\rho_t^q)} \int \big(\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\big)^{q-1} \partial_t \mu_t \\
&= \frac{q}{(q-1)\,\mathbb{E}_\pi(\rho_t^q)} \int \rho_t^{q-1} \operatorname{div}(\{\nabla \ln \rho_t + \Delta_t\}\,\mu_t) \\
&= -\frac{q}{(q-1)\,\mathbb{E}_\pi(\rho_t^q)} \int \langle \nabla(\rho_t^{q-1}), \nabla \ln \rho_t + \Delta_t \rangle \,\mathrm{d}\mu_t \\
&= -\frac{1}{\mathbb{E}_\pi(\rho_t^q)} \Big\{ \frac{4}{q}\,\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2] + 2\,\mathbb{E}_{\mu_t}[\rho_t^{q/2-1} \langle \nabla(\rho_t^{q/2}), \Delta_t \rangle] \Big\}\,.
\end{aligned}$$

For the second term, Young's inequality implies

$$\begin{aligned}
&-\mathbb{E}_{\mu_t}[\rho_t^{q/2-1} \langle \nabla(\rho_t^{q/2}), \Delta_t \rangle] \\
&\quad = -\iint \rho_t^{q/2-1}(x_t) \langle \nabla(\rho_t^{q/2})(x_t), \nabla V(x_{kh}) - \nabla V(x_t) \rangle \,\mu_{kh|t}(\mathrm{d}x_{kh} \mid x_t)\,\mu_t(\mathrm{d}x_t) \\
&\quad = -\iint \rho_t^{q/2-1}(x_t) \langle \nabla(\rho_t^{q/2})(x_t), \nabla V(x_{kh}) - \nabla V(x_t) \rangle \,\mu_{kh,t}(\mathrm{d}x_{kh}, \mathrm{d}x_t) \\
&\quad = -\mathbb{E}[\rho_t^{q/2-1}(X_t) \langle \nabla(\rho_t^{q/2})(X_t), \nabla V(X_{kh}) - \nabla V(X_t) \rangle] \\
&\quad \leq \frac{1}{2q}\,\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2] + \frac{q}{2}\,\mathbb{E}[\rho_t^{q-1}(X_t)\,\|\nabla V(X_{kh}) - \nabla V(X_t)\|^2]\,.
\end{aligned}$$

Substituting this into the previous expression completes the proof. □

Next, we formulate a lemma to control the expectation of $\|\nabla V\|^2$ under a change of measure. Although this is not strictly necessary for the proof, it streamlines the argument.

Lemma 3.6.3. *Assume that ∇V is L-Lipschitz. For any probability measure μ, we have*

$$\mathbb{E}_\mu[\|\nabla V\|^2] \leq 4\,\mathbb{E}_\pi\big[\big\|\nabla \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}}\big\|^2\big] + 2dL = \mathbb{E}_\mu\big[\big\|\nabla \ln \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\big\|^2\big] + 2dL\,.$$

Proof. Let $\mathscr{L}$ denote the infinitesimal generator of the Langevin diffusion (3.1), i.e., $\mathscr{L}f = \Delta f - \langle \nabla V, \nabla f \rangle$. Observe that $\mathscr{L}V = \Delta V - \|\nabla V\|^2$. Applying integration by parts and recalling that $\mathbb{E}_\pi \mathscr{L}f = 0$ for any f,

$$\begin{aligned}
\mathbb{E}_\mu[\|\nabla V\|^2] &= -\mathbb{E}_\mu \mathscr{L}V + \mathbb{E}_\mu \Delta V \le -\int \mathscr{L}V \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\right) \mathrm{d}\pi + dL \\
&= \int \langle \nabla V, \nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \rangle \, \mathrm{d}\pi + dL \\
&= 2 \int \langle \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}} \nabla V, \nabla \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}} \rangle \, \mathrm{d}\pi + dL \\
&\le \frac{1}{2} \mathbb{E}_\mu[\|\nabla V\|^2] + 2 \mathbb{E}_\pi\big[\|\nabla \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}}\|^2\big] + dL \,.
\end{aligned}$$

Rearrange this inequality to obtain the desired result. □

We are now ready to give the proof of Theorem 3.3.1. In order to emphasize the main ideas, we first present a proof which incurs a suboptimal dependence on q and explain how to sharpen the argument afterwards.

Proof of Theorem 3.3.1. As encapsulated in the differential inequality of Proposition 3.6.2, the crux of the proof of Theorem 3.3.1 is to control the discretization error term $\mathbb{E}[\psi_t(X_t) \|\nabla V(X_t) - \nabla V(X_{kh})\|^2]$ for $t \in [kh, (k+1)h]$. Since ∇V is L-Lipschitz, we have $\|\nabla V(X_t) - \nabla V(X_{kh})\|^2 \le 2L^2 (t - kh)^2 \|\nabla V(X_{kh})\|^2 + 4L^2 \|B_t - B_{kh}\|^2$. However, it is more convenient to have a bound in terms of $\|\nabla V(X_t)\|$ rather than $\|\nabla V(X_{kh})\|$, so we use

$$\begin{aligned}
\|\nabla V(X_{kh})\| &\le \|\nabla V(X_t)\| + L \|X_t - X_{kh}\| \\
&\le \|\nabla V(X_t)\| + hL \|\nabla V(X_{kh})\| + \sqrt{2} L \|B_t - B_{kh}\| \,.
\end{aligned}$$

If $h \le 1/(3L)$, we can rearrange this inequality to obtain

$$\|\nabla V(X_{kh})\| \le \frac{3}{2} \|\nabla V(X_t)\| + \frac{3L}{\sqrt{2}} \|B_t - B_{kh}\| \,,$$

so

$$\begin{aligned}
&\|\nabla V(X_t) - \nabla V(X_{kh})\|^2 \\
&\quad \le 9L^2 (t - kh)^2 \|\nabla V(X_t)\|^2 + (18 h^2 L^4 + 4L^2) \|B_t - B_{kh}\|^2 \\
&\quad \le 9L^2 (t - kh)^2 \|\nabla V(X_t)\|^2 + 6L^2 \|B_t - B_{kh}\|^2 \,.
\end{aligned}$$

We will control the two error terms in turn.

For the first error term, applying Lemma 3.6.3 to the measure $\psi_t\mu_t$ yields

$$\mathbb{E}_{\psi_t\mu_t}[\|\nabla V\|^2] \le \mathbb{E}_{\mu_t}\big[\psi_t \big\|\nabla \ln\big(\psi_t \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\big)\big\|^2\big] + 2dL = \frac{\mathbb{E}_\pi[\rho_t^q \,\|\nabla \ln(\rho_t^q)\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + 2dL$$
$$= \frac{4\,\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + 2dL\,.$$

Note the calculation

$$\mathbb{E}_{\mu_t}\big[\psi_t \big\|\nabla \ln\big(\psi_t \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\big)\big\|^2\big] = \frac{4\,\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)}\,, \tag{3.7}$$

which will be used below as well.

For the second error term, we apply the Donsker–Varadhan variational principle as in (3.5).

$$\begin{aligned}
&\mathbb{E}[\psi_t(X_t)\,\|B_t - B_{kh}\|^2] \\
&\quad\le 2\,\mathbb{E}[\|B_t - B_{kh}\|^2] \\
&\qquad + \frac{2}{c}\left\{\mathsf{KL}(\widetilde{\mathbb{P}} \,\|\, \mathbb{P}) + \ln \mathbb{E}\exp\big(c\,(\|B_t - B_{kh}\| - \mathbb{E}\|B_t - B_{kh}\|)^2\big)\right\} \\
&\quad\le 2d\,(t-kh) + \frac{2}{c}\left\{\mathsf{KL}(\widetilde{\mathbb{P}} \,\|\, \mathbb{P}) + \ln \mathbb{E}\exp\big(c\,(\|B_t - B_{kh}\| - \mathbb{E}\|B_t - B_{kh}\|)^2\big)\right\},
\end{aligned}$$

where $\frac{\mathrm{d}\widetilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}} = \psi_t(X_t)$. Due to Gaussian concentration, if we set $c = \frac{1}{8\,(t-kh)}$, then

$$\mathbb{E}\exp\frac{(\|B_t - B_{kh}\| - \mathbb{E}\|B_t - B_{kh}\|)^2}{8\,(t-kh)} \le 2\,,$$

c.f. [BLM13, §2.3, Theorem 5.5]. Next, using the LSI for π, we compute

$$\mathsf{KL}(\widetilde{\mathbb{P}} \,\|\, \mathbb{P}) = \mathbb{E}_{\psi_t\mu_t} \ln \psi_t = \mathbb{E}_{\psi_t\mu_t} \ln \frac{\rho_t^{q-1}}{\mathbb{E}_{\mu_t}(\rho_t^{q-1})} = \frac{q-1}{q}\,\mathbb{E}_{\psi_t\mu_t} \ln \frac{\rho_t^q}{\mathbb{E}_{\mu_t}(\rho_t^{q-1})^{q/(q-1)}} \tag{3.8}$$

$$= \frac{q-1}{q}\left\{\mathbb{E}_{\psi_t\mu_t} \ln \frac{\rho_t^q}{\mathbb{E}_{\mu_t}(\rho_t^{q-1})} - \underbrace{\frac{1}{q-1}\ln \mathbb{E}_{\mu_t}(\rho_t^{q-1})}_{\ge 0}\right\}$$

$$\le \frac{q-1}{q}\,\mathsf{KL}(\psi_t\mu_t \,\|\, \pi) \tag{3.9}$$

$$\le \frac{(q-1)\,C_{\mathsf{LSI}}}{2q}\,\mathbb{E}_{\psi_t\mu_t}\big[\big\|\nabla \ln(\psi_t \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi})\big\|^2\big]$$

$$= \frac{2\,(q-1)\,C_{\mathsf{LSI}}}{q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)}\,, \tag{3.10}$$

where the last equality is (3.7). We have proved

$$\begin{aligned}
&\mathbb{E}[\psi_t(X_t)\,\|B_t - B_{kh}\|^2] \\
&\quad\le 2d\,(t-kh) + \frac{32h\,(q-1)\,C_{\mathsf{LSI}}}{q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + (16\ln 2)\,(t-kh) \\
&\quad\le 14d\,(t-kh) + 32hC_{\mathsf{LSI}}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)}\,.
\end{aligned}$$

Finally, collecting together the error terms and applying Proposition 3.6.2, we see that

$$\begin{aligned}
\partial_t \mathcal{R}_q(\mu_t \,\|\, \pi) \le -\frac{3}{q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + 9qL^2\,(t-kh)^2\left\{\frac{4\,\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + 2dL\right\} \\
+ 6qL^2\left\{14d\,(t-kh) + 32hC_{\mathsf{LSI}}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)}\right\}.
\end{aligned}$$

Assuming for simplicity that $C_{\mathsf{LSI}}, L \ge 1$, then $h \le 1/(192q^2C_{\mathsf{LSI}}L^2)$ implies

$$\begin{aligned}
\partial_t \mathcal{R}_q(\mu_t \,\|\, \pi) &\le -\frac{1}{q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + 18dqL^3\,(t-kh)^2 + 84dqL^2\,(t-kh) \\
&\le -\frac{1}{2qC_{\mathsf{LSI}}}\,\mathcal{R}_q(\mu_t \,\|\, \pi) + 18dqL^3\,(t-kh)^2 + 84dqL^2\,(t-kh)\,,
\end{aligned}$$

where the last line uses the fact that π satisfies LSI [see VW19, Lemma 5]. This then implies the differential inequality

$$\begin{aligned}
\partial_t\Big\{\exp\Big(\frac{t-kh}{2qC_{\mathsf{LSI}}}\Big)\,\mathcal{R}_q(\mu_t \,\|\, \pi)\Big\} &\le \exp\Big(\frac{t-kh}{2qC_{\mathsf{LSI}}}\Big)\,\{18dqL^3\,(t-kh)^2 + 84dqL^2\,(t-kh)\} \\
&\le 19dqL^3\,(t-kh)^2 + 85dqL^2\,(t-kh)\,.
\end{aligned}$$

Integrating this inequality over $t \in [kh, (k+1)h]$ yields the recursion

$$\begin{aligned}
\mathcal{R}_q(\mu_{(k+1)h} \,\|\, \pi) &\le \exp\Big(-\frac{h}{2qC_{\mathsf{LSI}}}\Big)\,\mathcal{R}_q(\mu_{kh} \,\|\, \pi) + \frac{19}{3}\,dh^3qL^3 + \frac{85}{2}\,dh^2qL^2 \\
&\le \exp\Big(-\frac{h}{2qC_{\mathsf{LSI}}}\Big)\,\mathcal{R}_q(\mu_{kh} \,\|\, \pi) + 43dh^2qL^2\,.
\end{aligned}$$

Iterating this yields

$$\mathcal{R}_q(\mu_{Nh} \,\|\, \pi) \le \exp\Big(-\frac{Nh}{2qC_{\mathsf{LSI}}}\Big)\,\mathcal{R}_q(\mu_0 \,\|\, \pi) + 86dhq^2C_{\mathsf{LSI}}L^2,$$

which completes the proof. □

We now outline the hypercontractivity argument to improve the dependence on q.

Proposition 3.6.4 (Hypercontractivity)**.** *Let $(\mu_t)_{t\geq 0}$ denote the law of the interpolation (3.2) of LMC. Also, let $q(t) \coloneqq 1 + (q_0 - 1)\exp\frac{t}{2C_{\mathsf{LSI}}}$ for $t \geq 0$, and write $\rho_t \coloneqq \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}$, $\psi_t \coloneqq \rho_t^{q(t)-1}/\mathbb{E}_\pi(\rho_t^{q(t)})$. Then, for $t \in [kh, (k+1)h]$,*

$$\partial_t\Big(\frac{1}{q(t)}\ln\int \rho_t^{q(t)}\,\mathrm{d}\pi\Big) \leq -\frac{2\,(q(t)-1)}{q(t)^2}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q(t)/2})\|^2]}{\mathbb{E}_\pi(\rho_t^{q(t)})} + (q(t)-1)\,\mathbb{E}[\psi_t(X_t)\,\|\nabla V(X_t) - \nabla V(X_{kh})\|^2]\,.$$

Proof. Using calculus together with Proposition 3.6.1, we compute the derivative in time as in Proposition 3.6.2, only now taking into account the additional time-dependent function q. Since the calculation is very similar to Proposition 3.6.2, we only record the final result:

$$\begin{aligned}
&\partial_t\Big(\frac{1}{q(t)}\ln\int \rho_t^{q(t)}\,\mathrm{d}\pi\Big)\\
&\quad= -\frac{1}{\mathbb{E}_\pi(\rho_t^{q(t)})}\int\langle\nabla(\rho_t^{q(t)-1}), \nabla\ln\rho_t + \Delta_t\rangle\,\mathrm{d}\mu_t + \frac{\dot q(t)\,\mathrm{ent}_\pi(\rho_t^{q(t)})}{q(t)^2\,\mathbb{E}_\pi(\rho_t^{q(t)})}\\
&\quad\leq -\frac{3\,(q(t)-1)}{q(t)^2}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q(t)/2})\|^2]}{\mathbb{E}_\pi(\rho_t^{q(t)})}\\
&\qquad\qquad + (q(t)-1)\,\mathbb{E}[\psi_t(X_t)\,\|\nabla V(X_t) - \nabla V(X_{kh})\|^2] + \frac{\dot q(t)\,\mathrm{ent}_\pi(\rho_t^{q(t)})}{q(t)^2\,\mathbb{E}_\pi(\rho_t^{q(t)})}\,,
\end{aligned}$$

where $\dot q$ is the derivative of q, we write $\Delta_t \coloneqq \mathbb{E}[\nabla V(X_{kh}) \mid X_t = \cdot] - \nabla V$, and the entropy functional is defined in §3.2. Applying (LSI),

$$\frac{\dot q(t)\,\mathrm{ent}_\pi(\rho_t^{q(t)})}{q(t)^2\,\mathbb{E}_\pi(\rho_t^{q(t)})} \leq \frac{2\dot q(t) C_{\mathsf{LSI}}\,\mathbb{E}_\pi[\|\nabla(\rho_t^{q(t)/2})\|^2]}{q(t)^2\,\mathbb{E}_\pi(\rho_t^{q(t)})} = \frac{q(t)-1}{q(t)^2}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q(t)/2})\|^2]}{\mathbb{E}_\pi(\rho_t^{q(t)})}$$

where the last equality follows from our choice of q. □

Proof of Theorem 3.3.1. **Initial waiting phase**. Let $\bar q \geq 3$. We apply Proposition 3.6.4 with $q_0 = 2$ and for $t \leq N_0 h$, where $N_0 = \lceil \frac{2C_{\mathsf{LSI}}}{h}\ln(\bar q - 1)\rceil$. As in the earlier proof of Theorem 3.3.1, we take $h \leq 1/(192 q^2 C_{\mathsf{LSI}} L^2)$; note that, $\bar q \leq q(N_0 h) \leq 2\bar q$. The bound on the error term from the previous proof implies

$$\partial_t\Big(\frac{1}{q(t)}\ln\int \rho_t^{q(t)}\,\mathrm{d}\pi\Big) \leq 18 d q(t) L^3\,(t-kh)^2 + 84 d q(t) L^2\,(t-kh)\,.$$

Integrating this over $t \in [kh, (k+1)h]$ yields

$$\frac{1}{q((k+1)h)} \ln \int \rho_{(k+1)h}^{q((k+1)h)} \,\mathrm{d}\pi - \frac{1}{q(kh)} \ln \int \rho_{kh}^{q(kh)} \,\mathrm{d}\pi \le 12dh^3\bar{q}L^3 + 84dh^2\bar{q}L^2 \le 85dh^2\bar{q}L^2 \,.$$

Iterating this yields

$$\frac{1}{q(N_0h)} \ln \int \rho_{N_0h}^{q(N_0h)} \,\mathrm{d}\pi - \frac{1}{2} \ln \int \rho_0^2 \,\mathrm{d}\pi \le 85dh^2\bar{q}L^2N_0 \le 170dh\bar{q}C_{\mathsf{LSI}}L^2 \ln \bar{q} \,.$$

Remainder of the convergence analysis. After shifting the time indices and applying the preceding proof of Theorem 3.3.1 with $q = 2$,

$$\begin{aligned}\mathcal{R}_{\bar{q}}(\mu_{(N+N_0)h} \,\|\, \pi) &\le \frac{3}{2\bar{q}} \ln \int \rho_{(N+N_0)h}^{\bar{q}} \,\mathrm{d}\pi \le \frac{3}{4} \mathcal{R}_2(\mu_{Nh} \,\|\, \pi) + 255dh\bar{q}C_{\mathsf{LSI}}L^2 \ln \bar{q} \\ &\le \frac{3}{4} \exp\Big(-\frac{Nh}{4C_{\mathsf{LSI}}}\Big) \mathcal{R}_2(\mu_0 \,\|\, \pi) + 258dhC_{\mathsf{LSI}}L^2 + 255dh\bar{q}C_{\mathsf{LSI}}L^2 \ln \bar{q} \\ &\le \exp\Big(-\frac{Nh}{4C_{\mathsf{LSI}}}\Big) \mathcal{R}_2(\mu_0 \,\|\, \pi) + 513dh\bar{q}C_{\mathsf{LSI}}L^2 \ln \bar{q} \,.\end{aligned}$$

This completes the proof. □

■ 3.6.3 Proof of Theorem 3.3.3

To prove Theorem 3.3.3, we show that the iterates of LMC satisfy (LSI) with a growing constant.

Lemma 3.6.5. *Assume that V is convex and ∇V is L-Lipschitz. Let $(\mu_{kh})_{k\in\mathbb{N}}$ denote the law of the iterates of LMC initialized at $\mu_0 = \mathsf{normal}(0, L^{-1}I_d)$ and run with step size $h \le 1/L$. Then, the LSI constant $C_{\mathsf{LSI}}(\mu_{kh})$ of μ_{kh} satisfies $C_{\mathsf{LSI}}(\mu_{kh}) \le L + 2kh$.*

Proof. With the condition on the step size, $\mathrm{id} - h\,\nabla V$ is a contraction. Using standard facts about the behavior of the log-Sobolev constant under contractions [BGL14, Proposition 5.4.3] and convolutions [see, e.g., Cha04], we obtain

$$C_{\mathsf{LSI}}(\mu_{(k+1)h}) \le C_{\mathsf{LSI}}\big((\mathrm{id} - h\nabla V)_{\#}\mu_{kh}\big) + 2h \le C_{\mathsf{LSI}}(\mu_{kh}) + 2h \,.$$

The result follows via iteration. □

We are now ready to prove Theorem 3.3.3, which builds upon the proof of Theorem 3.3.1.

Proof of Theorem 3.3.3. Using the differential inequality of Proposition 3.6.2, assuming $h \le 1/(3L)$, we want to control the error term

$$\begin{aligned}&\mathbb{E}[\psi_t(X_t)\,\|\nabla V(X_t) - \nabla V(X_{kh})\|^2] \\ &\quad \le 9L^2\,(t-kh)^2\,\mathbb{E}[\psi_t(X_t)\,\|\nabla V(X_t)\|^2] + 6L^2\,\mathbb{E}[\psi_t(X_t)\,\|B_t - B_{kh}\|^2]\,,\end{aligned}$$

see the first proof of Theorem 3.3.1. For the first term, an application of Lemma 3.6.3 again yields

$$\mathbb{E}[\psi_t(X_t)\,\|\nabla V(X_t)\|^2] \le \frac{4\,\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + 2dL\,.$$

For the second term, the Donsker–Varadhan variational principle (3.5) implies

$$\mathbb{E}[\psi_t(X_t)\,\|B_t - B_{kh}\|^2] \le 2d\,(t-kh) + 16\,(t-kh)\,\{\mathsf{KL}(\widetilde{\mathbb{P}}\,\|\,\mathbb{P}) + \ln 2\}\,.$$

Now comes a key difference in the proof: in Theorem 3.3.1, we bounded $\mathsf{KL}(\widetilde{\mathbb{P}}\|\mathbb{P}) \le \frac{q-1}{q}\,\mathsf{KL}(\psi_t\mu_t\,\|\,\pi)$ and applied the LSI for π. Here, we instead use $\mathsf{KL}(\widetilde{\mathbb{P}}\,\|\,\mathbb{P}) = \mathsf{KL}(\psi_t\mu_t\,\|\,\mu_t)$ and apply the LSI from Lemma 3.6.5 which worsens over time. We thus obtain

$$\mathsf{KL}(\widetilde{\mathbb{P}}\,\|\,\mathbb{P}) \le 2C_{\mathsf{LSI}}(\mu_t)\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} \le 2\,(L + 2\,(k+1)\,h)\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)}\,.$$

Let N denote the total number of iterations that we run LMC. Collecting together all of the error terms and using Proposition 3.6.2, we see that

$$\begin{aligned}\partial_t\mathcal{R}_q(\mu_t\,\|\,\pi) \le -\frac{3}{q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + 9qL^2\,(t-kh)^2\left\{\frac{4\,\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + 2dL\right\} \\ + 6qL^2\left\{14d\,(t-kh) + 32h\,(L+2Nh)\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)}\right\}.\end{aligned}$$

Assuming $h \le \frac{1}{384qL\sqrt{N}}\min\{1, \frac{\sqrt{N}}{qL^2}\}$, it yields

$$\begin{aligned}\partial_t\mathcal{R}_q(\mu_t\,\|\,\pi) &\le -\frac{1}{q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + 18dqL^3\,(t-kh)^2 + 84dqL^2\,(t-kh) \\ &\le -\frac{1}{qC_{\mathsf{PI}}}\,\{1 - \exp(-\mathcal{R}_q(\mu_t\,\|\,\pi))\} + 18dqL^3\,(t-kh)^2 + 84dqL^2\,(t-kh)\,,\end{aligned}$$

where the last inequality follows from [VW19, Lemma 17].

We now split the analysis into two phases. In the first phase, we consider $t \le N_0 h$, where N_0 is the largest integer such that $\mathcal{R}_q(\mu_{N_0 h} \,\|\, \pi) \ge 1$. Then,

$$\partial_t \mathcal{R}_q(\mu_t \,\|\, \pi) \le -\frac{1}{2qC_{\mathsf{PI}}} + 18dqL^3\,(t-kh)^2 + 84dqL^2\,(t-kh)\,.$$

Integration yields

$$\begin{aligned}\mathcal{R}_q(\mu_{(k+1)h} \,\|\, \pi) - \mathcal{R}_q(\mu_{kh} \,\|\, \pi) &\le -\frac{h}{2qC_{\mathsf{PI}}} + 6dh^3qL^3 + 42dh^2qL^2 \\ &\le -\frac{h}{2qC_{\mathsf{PI}}} + 43dh^2qL^2\,.\end{aligned}$$

If $h \le \frac{1}{172dq^2C_{\mathsf{PI}}L^2}$, then we deduce that $\mathcal{R}_q(\mu_{kh} \,\|\, \pi) \le \mathcal{R}_q(\mu_0 \,\|\, \pi) - \frac{kh}{4qC_{\mathsf{PI}}}$, and hence that the first phase ends after at most $N_0 \le 4qC_{\mathsf{PI}}\mathcal{R}_q(\mu_0 \,\|\, \pi)/h$ iterations.

In the second phase, we consider t such that $\mathcal{R}_q(\mu_t \,\|\, \pi) \le 1$. Using $1 - \exp(-x) \ge x/2$ for $x \in [0,1]$, in this phase we have the inequality

$$\partial_t \mathcal{R}_q(\mu_t \,\|\, \pi) \le -\frac{1}{2qC_{\mathsf{PI}}}\,\mathcal{R}_q(\mu_t \,\|\, \pi) + 18dqL^3\,(t-kh)^2 + 84dqL^2\,(t-kh)\,.$$

As in the proof of Theorem 3.3.1, it implies

$$\begin{aligned}\mathcal{R}_q(\mu_{Nh} \,\|\, \pi) &\le \exp\Big(-\frac{(N-N_0-1)h}{2qC_{\mathsf{PI}}}\Big)\,\mathcal{R}_q(\mu_{(N_0+1)h} \,\|\, \pi) + 88dhq^2C_{\mathsf{PI}}L^2 \\ &\le \exp\Big(-\frac{(N-N_0-1)h}{2qC_{\mathsf{PI}}}\Big) + 88dhq^2C_{\mathsf{PI}}L^2\,.\end{aligned}$$

To make this at most ε, we take $h \le \frac{\varepsilon}{176dq^2C_{\mathsf{PI}}L^2}$ and $N \ge N_0 + 1 + \frac{2qC_{\mathsf{PI}}}{h}\ln(2/\varepsilon)$.

From Lemma 3.6.17, we see that $\mathcal{R}_q(\mu_0 \,\|\, \pi) = \widetilde{O}(d)$, so that $N = \widetilde{\Theta}(\frac{dqC_{\mathsf{PI}}}{h})$. Substituting this into our earlier constraints on h, we see that if we take

$$h = \widetilde{\Theta}\Big(\frac{\varepsilon}{dq^2C_{\mathsf{PI}}L^2}\min\big\{1, \frac{1}{q\varepsilon}, \frac{dC_{\mathsf{PI}}}{\varepsilon L}\big\}\Big)\,,$$

then the iteration complexity is

$$N = \widetilde{\Theta}\Big(\frac{d^2q^3C_{\mathsf{PI}}^2L^2}{\varepsilon}\max\big\{1, q\varepsilon, \frac{\varepsilon L}{dC_{\mathsf{PI}}}\big\}\Big)\,.$$

This completes the proof. □

3.6.4 Proof of Theorem 3.3.4

3.6.4.1 Girsanov's theorem and change of measure

As discussed in Section 3.5.2, we will use the Girsanov's theorem, stated below in a form which is convenient for our purposes.

Theorem 3.6.6 (Girsanov's theorem, [Øks03, Theorem 8.6.8]). *Consider stochastic processes $(x_t)_{t\geq 0}$, $(b_t^P)_{t\geq 0}$, $(b_t^Q)_{t\geq 0}$ adapted to the same filtration. Let P_T and Q_T be probability measures on the path space $\mathcal{C}([0,T];\mathbb{R}^d)$ and $(X_t)_{t\geq 0}$ evolves according to*

$$\begin{aligned} \mathrm{d}X_t &= b_t^P\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t^P \qquad \text{under } P_T\,,\\ \mathrm{d}X_t &= b_t^Q\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t^Q \qquad \text{under } Q_T\,, \end{aligned}$$

where B^P is a P_T-Brownian motion and B^Q is a Q_T-Brownian motion. Assume that Novikov's condition

$$\mathbb{E}^{Q_T}\exp\Big(\frac{1}{4}\int_0^T \|b_t^P - b_t^Q\|^2\,\mathrm{d}t\Big) < \infty$$

holds. Then,

$$\frac{\mathrm{d}P_T}{\mathrm{d}Q_T} = \exp\Big(\frac{1}{\sqrt{2}}\int_0^T \langle b_t^P - b_t^Q, \mathrm{d}B_t^Q\rangle - \frac{1}{4}\int_0^T \|b_t^P - b_t^Q\|^2\,\mathrm{d}t\Big)\,.$$

Remark 3.6.7. *In our applications of Girsanov's theorem, although we do not check Novikov's condition explicitly, the validity of Novikov's condition follows from the proof.*

Actually, we only need the following corollary.

Corollary 3.6.8. *For any event $\mathcal{E}$ and $q \geq 1$,*

$$\mathbb{E}^{Q_T}\big[(\frac{\mathrm{d}P_T}{\mathrm{d}Q_T})^q\,\mathbb{1}_{\mathcal{E}}\big] \leq \sqrt{\mathbb{E}\Big[\exp\Big(q^2\int_0^T \|b_t^P - b_t^Q\|^2\,\mathrm{d}t\Big)\,\mathbb{1}_{\mathcal{E}}\Big]}\,,$$

provided that Novikov's condition holds:

$$\mathbb{E}^{Q_T}\exp\Big(q^2\int_0^T \|b_t^P - b_t^Q\|^2\,\mathrm{d}t\Big) < \infty\,.$$

Proof. Applying the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\mathbb{E}^{Q_T}\big[(\frac{\mathrm{d}P_T}{\mathrm{d}Q_T})^q \mathbb{1}_{\mathcal{E}}\big] \\
&\quad = \mathbb{E}^{Q_T}\Big[\exp\Big(\frac{q}{\sqrt{2}}\int_0^T \langle b_t^P - b_t^Q, \mathrm{d}B_t^Q\rangle - \frac{q}{4}\int_0^T \|b_t^P - b_t^Q\|^2\,\mathrm{d}t\Big)\mathbb{1}_{\mathcal{E}}\Big] \\
&\quad \le \sqrt{\mathbb{E}^{Q_T}\Big[\exp\Big((q^2 - \frac{q}{2})\int_0^T \|b_t^P - b_t^Q\|^2\,\mathrm{d}t\Big)\mathbb{1}_{\mathcal{E}}\Big]} \\
&\qquad \times \underbrace{\sqrt{\mathbb{E}^{Q_T}\exp\Big(\sqrt{2}q\int_0^T \langle b_t^P - b_t^Q, \mathrm{d}B_t^Q\rangle - q^2\int_0^T \|b_t^P - b_t^Q\|^2\,\mathrm{d}t\Big)}}_{=1} \\
&\quad \le \sqrt{\mathbb{E}^{Q_T}\Big[\exp\Big(q^2\int_0^T \|b_t^P - b_t^Q\|^2\,\mathrm{d}t\Big)\mathbb{1}_{\mathcal{E}}\Big]}\,,
\end{aligned}$$

where we used Itô's lemma to show that the underlined term equals 1; this step requires checking that the exponential local martingale is a bona fide martingale, which is implied by Novikov's condition. □

Next, we state and prove the change of measure principle described in §3.5.2. This lemma will be invoked repeatedly in the main arguments.

Lemma 3.6.9 (Change of measure)**.** *Let μ, ν be probability measures and let E be any event. Then,*

$$\mu(E) \le \nu(E) + \sqrt{\chi^2(\mu \,\|\, \nu)\,\nu(E)}\,.$$

In particular, if μ and ν are probability measures on $\mathbb{R}^d$ and

$$\nu\{\|\cdot\| \ge R_0 + \eta\} \le C\exp(-c\eta^2) \qquad \textit{for all } \eta \ge 0\,,$$

where $C \ge 1$, then

$$\mu\Big\{\|\cdot\| \ge R_0 + \sqrt{\frac{1}{c}\,\mathcal{R}_2(\mu \,\|\, \nu)} + \eta\Big\} \le 2C\exp\big(-\frac{c\eta^2}{2}\big) \qquad \textit{for all } \eta \ge 0\,.$$

Proof.

$$\mu(E) = \nu(E) + \int \mathbb{1}_E\,\big(\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\big)\,\mathrm{d}\nu \le \nu(E) + \sqrt{\chi^2(\mu \,\|\, \nu)\,\nu(E)}\,,$$

where the last inequality is the Cauchy–Schwarz inequality.

For the second statement, applying the change of measure principle to $E = \{\|\cdot\| \geq R_0 + \bar{\eta}\}$ yields

$$\mu\{\|\cdot\| \geq R_0 + \bar{\eta}\} \leq C \exp(-c\bar{\eta}^2) + \sqrt{C \exp\{-\big(c\bar{\eta}^2 - \mathcal{R}_2(\mu \,\|\, \nu)\big)\}}\,.$$

Now take $\bar{\eta} = \sqrt{\frac{1}{c}\,\mathcal{R}_2(\mu \,\|\, \nu)} + \eta$. □

Finally, we use the following lemma used to remove the conditioning on events.

Lemma 3.6.10 ([GT20, Lemma 14])**.** *Let $Y > 0$ be a random variable. Assume that for all $0 < \delta < 1/2$ there exists an event $\mathcal{E}_\delta$ with probability at least $1 - \delta$ such that $\mathbb{E}[Y^2 \mid \mathcal{E}_\delta] \leq \frac{v}{\delta^\xi}$ for some $\xi < 1$. Then, $\mathbb{E}\, Y \leq 4\sqrt{v}$.*

■ 3.6.4.2 Sub-Gaussianity of the Langevin diffusion

In this section, we introduce a modified distribution: for $\gamma, R > 0$,

$$\hat{\pi} \propto \exp(-\hat{V})\,, \qquad \hat{V}(x) \coloneqq V(x) + \frac{\gamma}{2}\,(\|x\| - R)_+^2\,. \tag{3.11}$$

Here, $(\|x\| - R)_+^2$ is interpreted as $\max\{\|x\| - R, 0\}^2$. Although $\hat{\pi}$ and $\hat{V}$ depend on the parameters γ and R, we will suppress this in the notation for simplicity. Note that by construction, $V = \hat{V}$ on the ball $B(0, R)$ of radius R centered at the origin. Also, the probability measure $\hat{\pi}$ has sub-Gaussian tails. We record this and other useful facts below.

Lemma 3.6.11 (properties of the modified potential)**.** *Let $\hat{\pi}$ and $\hat{V}$ be defined as in (3.11). Assume that $\nabla V(0) = 0$ and that ∇V satisfies (s-Hölder). Then, the following assertions hold.*

1. *(sub-Gaussian tail bound) Assume that R is chosen so that $\pi(B(0, R)) \geq 1/2$. Then, for all $\eta \geq 0$,*

$$\hat{\pi}\{\|\cdot\| \geq R + \eta\} \leq 2 \exp\big(-\frac{\gamma\eta^2}{2}\big)\,.$$

2. *(gradient growth) The gradient $\nabla\hat{V}$ satisfies*

$$\|\nabla\hat{V}(x)\| \leq L + (L + \gamma)\,\|x\|\,.$$

Proof. 1. We can write

$$\int \exp\big(\frac{\gamma}{2}\,(\|\cdot\| - R)_+^2\big)\,\mathrm{d}\hat{\pi} = \frac{\int \exp(-V)}{\int \exp(-\hat{V})}\,.$$

Next, we bound

$$\frac{\int \exp(-\hat{V})}{\int \exp(-V)} = \int \exp\bigl(-\frac{\gamma}{2}\,(\|\cdot\| - R)_+^2\bigr)\,\mathrm{d}\pi \geq \pi\bigl(B(0,R)\bigr) \geq \frac{1}{2}$$

by our assumption on R. The sub-Gaussian tail bound follows from Markov's inequality via

$$\hat{\pi}\{\|\cdot\| - R \geq \eta\} \leq \hat{\pi}\Bigl\{\exp\bigl(\frac{\gamma}{2}\,(\|\cdot\| - R)_+^2\bigr) \geq \exp\frac{\gamma\eta^2}{2}\Bigr\} \leq 2\exp\bigl(-\frac{\gamma\eta^2}{2}\bigr)\,.$$

2. First, note that $\|\nabla V(x)\| \leq L\,\|x\|^s \leq L\,(1 + \|x\|)$, using $\nabla V(0) = 0$ and (s-Hölder). Then,

$$\|\nabla\hat{V}(x)\| \leq \|\nabla V(x)\| + \gamma\,(\|x\| - R)_+ \leq L + (L+\gamma)\,\|x\|\,.$$

□

Throughout this section, we will assume that $R \geq \max\{1, 2\mathfrak{m}\}$, where $\mathfrak{m} \coloneqq \int\|\cdot\|\,\mathrm{d}\pi$, so that the sub-Gaussian tail bound in Lemma 3.6.11 is valid.

We now begin transferring the sub-Gaussianity of $\hat{\pi}$ to π_t. First, we establish sub-Gaussian tail bounds for $\hat{\pi}_t$, where $(\hat{\pi}_t)_{t\geq 0}$ is the law of the continuous-time Langevin diffusion

$$\mathrm{d}\hat{Z}_t = -\nabla\hat{V}(\hat{Z}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \tag{3.12}$$

with potential $\hat{V}$, initialized at $\hat{X}_0 \sim \mu_0$.

Lemma 3.6.12. *Let $(\hat{z}_t)_{t\geq 0}$ denote the modified diffusion (3.12) with potential $\hat{V}$. Assume that $h \leq 1/(2\,(L+\gamma))$ and $R \geq \max\{1, 2\mathfrak{m}\}$. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{t\in[0,Nh]} \|\hat{Z}_t\| \leq R + 4h\,(L+\gamma)\,R + \sqrt{\frac{8}{\gamma}\,\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})} + \sqrt{\bigl(96dh + \frac{32}{\gamma}\bigr)\ln\frac{8N}{\delta}}\,.$$

Proof. Apply the change of measure principle (Lemma 3.6.9) together with the sub-Gaussian tail bound in Lemma 3.6.11 to see that with probability at least $1-\delta$,

$$\|\hat{Z}_{kh}\| \leq R + \sqrt{\frac{2}{\gamma}\,\mathcal{R}_2(\hat{\pi}_t \,\|\, \hat{\pi})} + \sqrt{\frac{4}{\gamma}\ln\frac{4}{\delta}}\,.$$

Since the Rényi divergence is decreasing along the diffusion (3.12), then $\mathcal{R}_2(\hat{\pi}_t \,\|\, \hat{\pi}) \le \mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})$. Therefore, a union bound implies that with probability at least $1-\delta$,

$$\max_{k=0,1,\dots,N-1} \|\hat{Z}_{kh}\| \le R + \sqrt{\frac{2}{\gamma}\,\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})} + \sqrt{\frac{4}{\gamma}\ln\frac{4N}{\delta}}\,. \tag{3.13}$$

Next, for $t \le h$,

$$\begin{aligned}
\|\hat{Z}_{kh+t} - \hat{Z}_{kh}\| &\le \int_0^t \|\nabla \hat{V}(\hat{Z}_{kh+r})\|\,\mathrm{d}r + \sqrt{2}\,\|B_{kh+t} - B_{kh}\| \\
&\le hL + (L+\gamma)\int_0^t \|\hat{Z}_{kh+r}\|\,\mathrm{d}r + \sqrt{2}\,\|B_{kh+t} - B_{kh}\| \\
&\le hL + (L+\gamma)\Big(h\,\|\hat{Z}_{kh}\| + \int_0^t \|\hat{Z}_{kh+r} - \hat{Z}_{kh}\|\,\mathrm{d}r\Big) + \sqrt{2}\,\|B_{kh+t} - B_{kh}\|\,,
\end{aligned}$$

where we used Lemma 3.6.11. Grönwall's inequality implies

$$\begin{aligned}
&\sup_{t\in[0,h]} \|\hat{Z}_{kh+t} - \hat{Z}_{kh}\| \\
&\quad\le \big(hL + h\,(L+\gamma)\,\|\hat{Z}_{kh}\| + \sqrt{2}\sup_{t\in[0,h]}\|B_{kh+t} - B_{kh}\|\big)\exp\big(h\,(L+\gamma)\big) \\
&\quad\le 2hL + 2h\,(L+\gamma)\,\|\hat{Z}_{kh}\| + \sqrt{8}\sup_{t\in[0,h]}\|B_{kh+t} - B_{kh}\|
\end{aligned}$$

provided $h \le 1/(2\,(L+\gamma))$. Now, a union bound shows that

$$\begin{aligned}
&\mathbb{P}\Big\{\sup_{t\in[0,Nh]} \|\hat{Z}_t\| \ge \eta\Big\} \\
&\quad\le \mathbb{P}\Big\{\max_{k=0,1,\dots,N-1} \|\hat{Z}_{kh}\| \ge R'\Big\} \\
&\qquad + \sum_{k=0}^{N-1} \mathbb{P}\Big\{\sup_{t\in[0,h]} \|\hat{Z}_{kh+t} - \hat{Z}_{kh}\| \ge \eta - R',\ \max_{k=0,1,\dots,N-1} \|\hat{Z}_{kh}\| \le R'\Big\} \\
&\quad\le \mathbb{P}\Big\{\max_{k=0,1,\dots,N-1} \|\hat{Z}_{kh}\| \ge R'\Big\} \\
&\qquad + \sum_{k=0}^{N-1} \mathbb{P}\Big\{\sqrt{8}\sup_{t\in[0,h]} \|B_{kh+t} - B_{kh}\| \ge \eta - R' - 2hL - 2h\,(L+\gamma)\,R'\Big\}\,.
\end{aligned}$$

Taking $R' = R + \sqrt{\frac{2}{\gamma}\,\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})} + \sqrt{\frac{4}{\gamma}\ln\frac{8N}{\delta}}$ and applying a standard bound on the tail probability of Brownian motion (Lemma 3.6.21) shows that with probability

at least $1-\delta$, if $R \geq 1$,

$$\begin{aligned}\sup_{t\in[0,Nh]} \|\hat{Z}_t\| &\leq R' + 2hL + 2h\,(L+\gamma)\,R' + \sqrt{48dh\ln\frac{6N}{\delta}} \\ &\leq R + 4h\,(L+\gamma)\,R + \sqrt{\frac{8}{\gamma}\,\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})} + \sqrt{\big(96dh + \frac{32}{\gamma}\big)\ln\frac{8N}{\delta}}\end{aligned}$$

after simplifying some terms. □

Next, we control the Rényi divergence between π_t and $\hat{\pi}_t$, which ultimately allows us to transfer the sub-Gaussianity to π_t.

Proposition 3.6.13. *Let $T \coloneqq Nh$. Let Q_T, $\hat{Q}_T$ be the measures on path space corresponding to the original diffusion (3.1) and the modified diffusion (3.12) respectively, both initialized at μ_0. Assume that $h \leq \frac{1}{3}\min\{\frac{1}{L+\gamma}, \frac{T}{d}\}$ and $\gamma \leq \frac{1}{3072T}$. Also, suppose that $R \geq \max\{1, 2\mathfrak{m}\}$ and $\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi}) \geq 1$. Then,*

$$\mathcal{R}_2(Q_T \,\|\, \hat{Q}_T) \leq \frac{h\,(L+\gamma)^2\,R^2}{d} + 5\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})\ln(8N)\,.$$

Proof. For all $0 < \delta < 1/2$, let $\mathcal{E}_\delta$ denote the event that the conclusion of Lemma 3.6.12 holds, i.e.,

$$\mathcal{E}_\delta \coloneqq \Big\{\sup_{t\in[0,Nh]} \|\hat{Z}_t\| \leq R_\delta\Big\}$$

with

$$R_\delta \coloneqq R + 4h\,(L+\gamma)\,R + \sqrt{\frac{8}{\gamma}\,\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})} + \sqrt{\big(96dh + \frac{32}{\gamma}\big)\ln\frac{8N}{\delta}}\,.$$

Then, we know that $\mathbb{P}(\mathcal{E}_\delta) \geq 1-\delta$. Applying Girsanov's theorem in the form of Corollary 3.6.8,

$$\begin{aligned}\ln\mathbb{E}\big[\big(\frac{\mathrm{d}Q_T}{\mathrm{d}\hat{Q}_T}\big)^4\,\mathbb{1}_{\mathcal{E}_\delta}\big] &\leq \frac{1}{2}\ln\mathbb{E}\Big[\exp\Big(16\int_0^T \|\nabla V(\hat{z}_t) - \nabla\hat{V}(\hat{z}_t)\|^2\,\mathrm{d}t\Big)\,\mathbb{1}_{\mathcal{E}_\delta}\Big] \\ &= \frac{1}{2}\ln\mathbb{E}\Big[\exp\Big(16\gamma^2\int_0^T (\|\hat{z}_t\| - R)_+^2\,\mathrm{d}t\Big)\,\mathbb{1}_{\mathcal{E}_\delta}\Big] \\ &\leq \Big(384\gamma^2h^2\,(L+\gamma)^2\,R^2 + 192\gamma\,\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi}) + (2304\gamma^2dh + 768\gamma)\ln\frac{8N}{\delta}\Big)\,T\,.\end{aligned}$$

In order to apply Lemma 3.6.10 and remove the conditioning, we require the condition $2304\gamma^2 dhT + 768\gamma T < 1$. This can be achieved by taking $\gamma \le \frac{1}{3072T}$ and $h \le \frac{T}{3d}$. Then, Lemma 3.6.10 implies

$$\begin{aligned}
\mathcal{R}_2(Q_T \,\|\, \hat{Q}_T) &= \ln \mathbb{E}\big[\big(\frac{\mathrm{d}Q_T}{\mathrm{d}\hat{Q}_T}\big)^2\big] \\
&\le \ln 8 + \big(192\gamma^2 h^2 \left(L+\gamma\right)^2 R^2 + 96\gamma\, \mathcal{R}_2(\mu_0 \,\|\, \hat{\pi}) \\
&\qquad\qquad + (1152\gamma^2 dh + 384\gamma)\ln(8N)\big)\, T \\
&\le \ln 8 + \frac{h^2 \left(L+\gamma\right)^2 R^2}{T} + \mathcal{R}_2(\mu_0 \,\|\, \hat{\pi}) + \frac{dh \ln(8N)}{T} + \ln(8N) \\
&\le \frac{h \left(L+\gamma\right)^2 R^2}{d} + 5\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi}) \ln(8N)\,,
\end{aligned}$$

where we have combined terms using $\mathcal{R}_2(\mu_0 \| \hat{\pi}) \ge 1$ to simplify the final bound. □

Proposition 3.6.14. *Let $(Z_t)_{t\ge 0}$ denote the continuous-time diffusion (3.1) initialized at μ_0. Assume that $h \le \frac{1}{3}\min\{\frac{1}{L+T^{-1}}, \frac{T}{d}\}$ and $\mathfrak{m}, \mathcal{R}_2(\mu_0 \,\|\, \hat{\pi}) \ge 1$. Then, for all $\delta \in (0, 1/2)$, with probability at least $1-\delta$,*

$$\begin{aligned}
\max_{k=0,1,\ldots,N-1} \|Z_{kh}\| &\le 2\mathfrak{m} + 490\sqrt{T\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi}) \ln(8N)} \\
&\qquad + \frac{230 h^{1/2} \mathfrak{m} \left(L + T^{-1}\right) T^{1/2}}{d^{1/2}} + 160\sqrt{T \ln \frac{1}{\delta}}\,,
\end{aligned}$$

where we write $T \coloneqq Nh$.

Proof. Recall from the proof of Lemma 3.6.12 that with probability at least $1-\delta$,

$$\max_{k=0,1,\ldots,N-1} \|\hat{Z}_{kh}\| \le R + \sqrt{\frac{2}{\gamma}\,\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})} + \sqrt{\frac{4}{\gamma} \ln \frac{4N}{\delta}}$$

(see (3.13)). Equivalently,

$$\mathbb{P}\Big\{\max_{k=0,1,\ldots,N-1} \|\hat{Z}_{kh}\| \ge R + \sqrt{\frac{2}{\gamma}\,\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})} + \eta\Big\} \le 4N \exp\big(-\frac{\gamma\eta^2}{4}\big)\,.$$

Applying the change of measure principle (Lemma 3.6.9) again to Q_T and $\hat{Q}_T$ with the choice $\gamma = \frac{1}{3072T}$ and $R = 2\mathfrak{m}$ reveals that for all $\delta \in (0, 1/2)$, with probability at least $1-\delta$,

$$\max_{k=0,1,\ldots,N-1} \|Z_{kh}\| \le R + \sqrt{\frac{2}{\gamma}\,\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})} + \sqrt{\frac{4}{\gamma}\,\mathcal{R}_2(Q_T \,\|\, \hat{Q}_T)} + \sqrt{\frac{8}{\gamma} \ln \frac{8N}{\delta}}$$

$$\leq 2\mathfrak{m} + 490\sqrt{T\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})\ln(8N)} + \frac{230h^{1/2}\mathfrak{m}\,(L+T^{-1})\,T^{1/2}}{d^{1/2}} + 160\sqrt{T\ln\frac{1}{\delta}}\,,$$

after simplifying the bound. □

■ 3.6.4.3 Bounding the discretization error

In this section, we prove our main bound on the discretization error.

Proposition 3.6.15. *Let $(\mu_t)_{t\geq 0}$ denote the law of the interpolated process (3.2) and let $(\pi_t)_{t\geq 0}$ denote the law of the continuous-time Langevin diffusion (3.1), both initialized at μ_0. Assume that ∇V satisfies $\nabla V(0) = 0$ and (s-Hölder). For simplicity, assume that $\varepsilon^{-1}, \mathfrak{m}, L, T, \mathcal{R}_2(\mu_0 \,\|\, \hat{\pi}) \geq 1$ and $q \geq 2$. If the step size h satisfies*

$$h \leq \widetilde{O}_s\Big(\frac{\varepsilon^{1/s}}{dq^{1/s}L^{2/s}T^{1/s}}\min\Big\{1, \frac{1}{q^{1/s}\varepsilon^{1/s}}, \frac{d}{\mathfrak{m}^s}, \frac{d}{\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})^{s/2}}\Big\}\Big)\,,$$

where the notation $\widetilde{O}_s$ hides constants depending on s as well as polylogarithmic factors, then for $T \coloneqq Nh$,

$$\mathcal{R}_q(\mu_T \,\|\, \pi_T) \leq \varepsilon\,.$$

Proof. Let P, Q denote the measures on path space corresponding to the interpolated process (3.2) and the continuous-time diffusion (3.1) respectively, both initialized at μ_0. Also, let

$$G_t \coloneqq \frac{1}{\sqrt{2}}\int_0^r \langle \nabla V(Z_r) - \nabla V(Z_{\lfloor r/h\rfloor h}), \mathrm{d}B_r\rangle - \frac{1}{4}\int_0^r \|\nabla V(Z_r) - \nabla V(Z_{\lfloor r/h\rfloor h})\|^2\,\mathrm{d}r\,,$$

where $(Z_t)_{t\geq 0}$ is the continuous-time diffusion (3.1). By applying Girsanov's theorem (Theorem 3.6.6) and Itô's formula, we obtain

$$\begin{aligned}\mathbb{E}^{Q_T}\big[\big(\frac{\mathrm{d}P_T}{\mathrm{d}Q_T}\big)^q\big] - 1 &= \mathbb{E}\exp(qG_T) - 1\\ &= \frac{q\,(q-1)}{4}\,\mathbb{E}\int_0^T \exp(qG_t)\,\|\nabla V(Z_t) - \nabla V(Z_{\lfloor t/h\rfloor h})\|^2\,\mathrm{d}t\\ &\leq \frac{q^2}{4}\int_0^T \sqrt{\mathbb{E}[\exp(2qG_t)]\,\mathbb{E}[\|\nabla V(Z_t) - \nabla V(Z_{\lfloor t/h\rfloor h})\|^4]}\,\mathrm{d}t\,.\end{aligned} \tag{3.14}$$

We bound the two expectations in turn. From Corollary 3.6.8 and (s-Hölder),

$$\begin{aligned}\mathbb{E}\exp(2qG_t) &\le \sqrt{\mathbb{E}\exp\Big(4q^2\int_0^t \|\nabla V(Z_r) - \nabla V(Z_{\lfloor r/h\rfloor h})\|^2\,\mathrm{d}r\Big)} \\ &\le \sqrt{\mathbb{E}\exp\Big(4q^2L^2\int_0^t \|Z_r - Z_{\lfloor r/h\rfloor h}\|^{2s}\,\mathrm{d}r\Big)}\end{aligned}$$

and we control this term by conditioning on the event

$$\mathcal{E}_{\delta,kh} \coloneqq \Big\{\max_{j=0,1,\dots,k-1}\|Z_{jh}\| \le R_\delta\Big\},$$

where

$$R_\delta \coloneqq 2\mathfrak{m} + 490\sqrt{T\mathcal{R}_2(\mu_0\,\|\,\hat\pi)\ln(8N)} + \frac{230h^{1/2}\mathfrak{m}\,(L+T^{-1})\,T^{1/2}}{d^{1/2}} + 160\sqrt{T\ln\frac{1}{\delta}}\,.$$

By Proposition 3.6.14, we know that $\mathbb{P}(\mathcal{E}_{\delta,kh}) \ge 1-\delta$.

One step error. We first consider the error over an interval $[0,h]$ conditionally on Z_0, corresponding to a single step of the LMC algorithm. This step requires bounding the exponential moment of $\sup_{t\in[0,h]}\|Z_t - Z_0\|^{2s}$, which is a slightly tedious exercise in stochastic calculus; hence, we postpone the calculation to §3.6.7. We quote the final result here: assuming that $h \lesssim 1/(d^s q^2 L^2)^{1/(1+s)}$, Lemma 3.6.22 implies

$$\begin{aligned}\ln\mathbb{E}\exp\Big(8q^2L^2\int_0^h\|Z_t - Z_0\|^{2s}\,\mathrm{d}t\Big) &\le \ln\mathbb{E}\exp\big(8hq^2L^2\sup_{t\in[0,h]}\|Z_t - Z_0\|^{2s}\big) \\ &\lesssim h^{2s+1}q^2L^{2s+2}\,(1+\|Z_0\|^{2s^2}) + d^s h^{s+1}q^2L^2\,.\end{aligned}$$

Iterating the bound. Let $(\mathscr{F}_t)_{t\ge0}$ denote the filtration and introduce the shorthand notation $H_t \coloneqq \int_0^t\|x_r - x_{\lfloor r/h\rfloor h}\|^{2s}\,\mathrm{d}r$. By conditioning on $\mathscr{F}_{(N-1)h}$, we can apply the one step bound to derive the bound

$$\begin{aligned}&\ln\mathbb{E}[\exp\{8q^2L^2H_{Nh}\}\,\mathbb{1}_{\mathcal{E}_{\delta,Nh}}] \\ &\quad\le \ln\mathbb{E}\big[\exp\big\{8q^2L^2H_{(N-1)h} \\ &\qquad\qquad + O\big(h^{2s+1}q^2L^{2s+2}\,(1+\|Z_{(N-1)h}\|^{2s^2}) + d^s h^{s+1}q^2L^2\big)\big\}\,\mathbb{1}_{\mathcal{E}_{\delta,Nh}}\big] \\ &\quad\le \ln\mathbb{E}[\exp\{8q^2L^2H_{(N-1)h}\}\,\mathbb{1}_{\mathcal{E}_{\delta,(N-1)h}}] \\ &\qquad + O\big(h^{2s+1}q^2L^{2s+2}\,(1+R_\delta^{2s^2}) + d^s h^{s+1}q^2L^2\big)\,.\end{aligned}$$

Iterating this recursion yields

$$\ln \mathbb{E}[\exp\{8q^2L^2H_{Nh}\}\,\mathbb{1}_{\mathcal{E}_{\delta,Nh}}] \lesssim h^{2s}q^2L^{2s+2}R_\delta^{2s^2}T + d^s h^s q^2 L^2 T\,.$$

where we recall $T \coloneqq Nh$. In order to apply Lemma 3.6.10 to remove the conditioning, we require the step size to satisfy $h \lesssim_s 1/(q^{1/s}L^{(s+1)/s}T^{(s^2+1)/(2s)})$, where the notation $\lesssim_s$ hides a constant depending only on s. Applying the lemma,

$$\begin{aligned}\ln \mathbb{E}\exp\{4q^2L^2H_{Nh}\} &\lesssim 1 + d^s h^s q^2 L^2 T \\ &\quad + h^{2s}q^2L^{2s+2}T \\ &\qquad \times \Big(\mathfrak{m} + \sqrt{T\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})\ln(8N)} + \frac{h^{1/2}\mathfrak{m}\,(L + T^{-1})\,T^{1/2}}{d^{1/2}}\Big)^{2s^2}\,.\end{aligned}$$

We pause here to give a remark which may clarify the proof. The $+1$ term above arises for two reasons. First, Lemma 3.6.10 requires a bound on the conditional expectation $\mathbb{E}[\exp\{8q^2L^2H_{Nh}\} \mid \mathcal{E}_{\delta,Nh}]$ whereas we have bounded $\mathbb{E}[\exp\{8q^2L^2H_{Nh}\}\,\mathbb{1}_{\mathcal{E}_{\delta,Nh}}]$; passing from the latter to the former incurs a factor of 2 (for $\delta \le 1/2$). Second, the conclusion of Lemma 3.6.10 also contributes a factor of 4. This shows that the application of Lemma 3.6.10 inherently adds a constant to the bound on the logarithm of the expectation. This also explains why, at the beginning of this proof in (3.14), we first applied Itô's formula to $\exp(qG_T)$ rather than applying Lemma 3.6.10 to $\mathbb{E}\exp(qG_T)$ directly. If we had done the latter, then it would not be possible to make the Rényi divergence $\mathcal{R}_q(P_T \,\|\, Q_T)$ arbitrarily small with an appropriate choice of h.

We now choose h in order to make $\mathbb{E}\exp\{4q^2L^2H_{Nh}\} \lesssim 1$. This is accomplished by taking

$$h \le \widetilde{O}_s\Big(\frac{1}{dq^{2/s}L^{2/s}T^{1/s}}\min\Big\{1, \frac{d}{\mathfrak{m}^s}, \frac{d}{\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})^{s/2}}, \frac{d^{(2s+2)/(s+2)}}{\mathfrak{m}^{2s/(s+2)}}\Big\}\Big)\,. \tag{3.15}$$

The last term in the minimum can also be eliminated from consideration; indeed, if $d^{(2s+2)/(s+2)}/\mathfrak{m}^{2s/(s+2)} \ge 1$, then it is not active in the minimum. Otherwise, raising this expression to the power $(s+2)/2 \ge 1$,

$$\frac{d^{(2s+2)/(s+2)}}{\mathfrak{m}^{2s/(s+2)}} \ge \frac{d^{s+1}}{\mathfrak{m}^s} \ge \frac{d}{\mathfrak{m}^s}\,.$$

Controlling the remaining term. Next, we must bound the difference $\mathbb{E}[\|\nabla V(Z_t) - \nabla V(Z_{kh})\|^4]$ for $t \in [kh, (k+1)h]$. Although this can also be handled directly via stochastic calculus, we will deduce the bound from Lemma 3.6.22 to avoid repeating work. This yields

$$\mathbb{E}[\exp(\lambda\,\|Z_t - Z_{kh}\|^{2s}) \mid Z_{kh}] \lesssim 1\,,$$

provided that λ is chosen as

$$\lambda \asymp \frac{1}{d^s h^s} \wedge \frac{1}{h^{2s} L^{2s} \left(1 + \|Z_{kh}\|^{2s^2}\right)} .$$

In turn, it implies the tail bound

$$\mathbb{P}\{\|Z_t - Z_{kh}\|^{4s} \geq \eta \mid Z_{kh}\} \lesssim \exp(-\lambda\sqrt{\eta})$$

which is integrated to yield

$$\begin{aligned}\sqrt{\mathbb{E}[\|\nabla V(Z_t) - \nabla V(Z_{kh})\|^4]} &\leq L^2 \sqrt{\mathbb{E}[\|Z_t - Z_{kh}\|^{4s}]} \lesssim L^2 \sqrt{\mathbb{E}\,\frac{1}{\lambda^2}} \\ &\lesssim d^s h^s L^2 + h^{2s} L^{2s+2} \sqrt{1 + \mathbb{E}[\|Z_{kh}\|^{4s^2}]} .\end{aligned}$$

Integrate the sub-Gaussian tail bound from Proposition 3.6.14 to obtain

$$\sqrt{1 + \mathbb{E}[\|Z_{kh}\|^{4s^2}]} \leq \widetilde{O}\Big(\mathfrak{m}^{2s^2} + T^{s^2} \mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})^{s^2} + \frac{h^{s^2} \mathfrak{m}^{2s^2} L^{2s^2} T^{s^2}}{d^{s^2}}\Big) .$$

Finishing the proof. Combining together the previous steps, we have proven

$$\begin{aligned}&\mathbb{E}^{Q_T}\big[\big(\frac{\mathrm{d}P_T}{\mathrm{d}Q_T}\big)^q\big] - 1 \\ &\quad \leq \widetilde{O}\Big(d^s h^s q^2 L^2 T \\ &\qquad + h^{2s} q^2 L^{2s+2} T \left(\mathfrak{m}^{2s^2} + T^{s^2} \mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})^{s^2} + \frac{h^{s^2} \mathfrak{m}^{2s^2} L^{2s^2} T^{s^2}}{d^{s^2}}\right)\Big) .\end{aligned}$$

The step size condition from (3.15) makes the right-hand side of the above expression $\lesssim 1$. Taking logarithms,

$$\begin{aligned}&\mathcal{R}_q(P_T \,\|\, Q_T) \\ &\quad \leq \widetilde{O}\Big(d^s h^s q L^2 T + h^{2s} q L^{2s+2} T \left(\mathfrak{m}^{2s^2} + T^{s^2} \mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})^{s^2} + \frac{h^{s^2} \mathfrak{m}^{2s^2} L^{2s^2} T^{s^2}}{d^{s^2}}\right)\Big) .\end{aligned}$$

We now choose h to make the Rényi divergence at most ε^2. By similar reasoning as before, it suffices to take

$$h \leq \widetilde{O}_s\Big(\frac{\varepsilon^{2/s}}{d q^{1/s} L^{2/s} T^{1/s}} \min\Big\{1, \frac{d}{\mathfrak{m}^s}, \frac{d}{\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})^{s/2}}\Big\}\Big) .$$

This completes the proof. □

■ 3.6.4.4 Finishing the proof

Finally, we use Theorem 3.2.2 on the continuous-time convergence of the Langevin diffusion (3.1) in Rényi divergence under an LOI. Together with our discretization bound, it will imply Theorem 3.3.4.

Lemma 3.6.16. *Let $(\pi_t)_{t\geq 0}$ denote the law of the continuous-time diffusion (3.1) initialized at μ_0, and assume that π satisfies (LOI) with order α. If*

$$T \geq 68qC_{\mathsf{LOI}(\alpha)}\left(\frac{\mathcal{R}_q(\mu_0 \,\|\, \pi)^{2/\alpha-1}-1}{2/\alpha-1} + \ln\frac{1}{\varepsilon^2}\right),$$

we obtain $\mathcal{R}_q(\pi_T \,\|\, \pi) \leq \varepsilon^2$.

Proof. Recall from Theorem 3.2.2 that

$$\partial_t \mathcal{R}_q(\pi_t \,\|\, \pi) \leq -\frac{1}{68qC_{\mathsf{LOI}(\alpha)}} \times \begin{cases} \mathcal{R}_q(\pi_t \,\|\, \pi)^{2-2/\alpha}\,, & \text{if } \mathcal{R}_q(\pi_t \,\|\, \pi) \geq 1\,, \\ \mathcal{R}_q(\pi_t \,\|\, \pi)\,, & \text{if } \mathcal{R}_q(\pi_t \,\|\, \pi) \leq 1\,. \end{cases}$$

In general, if $R : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies the ODE $R' = -CR^\beta$ for some $\beta \in (0,1)$, then a calculation shows that

$$R(t) = \left\{R(0)^{1-\beta} - C\,(1-\beta)\,t\right\}^{1/(1-\beta)}.$$

Thus, if $\alpha < 2$, we obtain $\mathcal{R}_q(\pi_{T_0} \,\|\, \pi) \leq 1$ at time

$$T_0 = \frac{68qC_{\mathsf{LOI}(\alpha)}}{2/\alpha-1}\left\{\mathcal{R}_q(\mu_0 \,\|\, \pi)^{2/\alpha-1} - 1\right\}.$$

Observe that as $\alpha \to 2$, then $T_0 \to 68qC_{\mathsf{LOI}(2)} \ln \mathcal{R}_q(\mu_0 \,\|\, \pi)$ which recovers the continuous-time convergence under (LSI). Then, at time

$$T = T_0 + 68qC_{\mathsf{LOI}(\alpha)} \ln\frac{1}{\varepsilon^2}$$

we obtain $\mathcal{R}_q(\pi_T \,\|\, \pi) \leq \varepsilon^2$. □

Proof of Theorem 3.3.4. Let $(\mu_t)_{t\geq 0}$ denote the marginal law of the interpolated process (3.2) and let $(\pi_t)_{t\geq 0}$ denote the law of the continuous-time Langevin diffusion (3.1), both initialized at μ_0. By the weak triangle inequality (when $q \geq 2$; Lemma 2.2.23), we can bound

$$\mathcal{R}_q(\mu_{Nh} \,\|\, \pi) \lesssim \mathcal{R}_{2q}(\mu_{Nh} \,\|\, \pi_{Nh}) + \mathcal{R}_{2q-1}(\pi_{Nh} \,\|\, \pi)\,.$$

For $T \coloneqq Nh$, we can make the second term at most $\varepsilon^2/2$ if we choose

$$T = \widetilde{\Theta}\big(qC_{\mathsf{LOI}(\alpha)}\,\mathcal{R}_{2q-1}(\mu_0 \parallel \pi)^{2/\alpha-1}\big)$$

by Lemma 3.6.16. Then, by Proposition 3.6.15, we can make the first term at most $\varepsilon^2/2$ taking

$$h = \widetilde{\Theta}_s\Big(\frac{\varepsilon^{2/s}}{dq^{2/s}C_{\mathsf{LOI}(\alpha)}^{1/s}L^{2/s}\,\mathcal{R}_{2q-1}(\mu_0 \parallel \pi)^{(2/\alpha-1)/s}} \tag{3.16}$$

$$\times \min\Big\{1, \frac{1}{q^{1/s}\varepsilon^{2/s}}, \frac{d}{\mathfrak{m}^s}, \frac{d}{\mathcal{R}_2(\mu_0 \parallel \hat{\pi})^{s/2}}\Big\}\Big)\,. \tag{3.17}$$

Then, the total number of iterations of LMC is

$$N = \frac{T}{h} = \widetilde{\Theta}_s\Big(\frac{dq^{1+2/s}C_{\mathsf{LOI}(\alpha)}^{1+1/s}L^{2/s}\,\mathcal{R}_{2q-1}(\mu_0 \parallel \pi)^{(2/\alpha-1)\,(1+1/s)}}{\varepsilon^{2/s}}$$

$$\times \max\Big\{1, q^{1/s}\varepsilon^{2/s}, \frac{\mathfrak{m}^s}{d}, \frac{\mathcal{R}_2(\mu_0 \parallel \hat{\pi})^{s/2}}{d}\Big\}\Big)\,.$$

This completes the proof. □

■ 3.6.5 Proof of Theorems 3.2.3 and 3.3.6

We first prove the continuous-time convergence for the Langevin diffusion (3.1) under (MLSI) and (α_1-tail).

Proof of Theorem 3.2.3. From [VW19, Lemma 6], we have

$$\partial_t \mathcal{R}_q(\pi_t \parallel \pi) = -\frac{4}{q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)}\,,$$

where $\rho_t \coloneqq \frac{\mathrm{d}\pi_t}{\mathrm{d}\pi}$. Following the calculations of [VW19, Lemma 5] and applying (MLSI) to $f^2 = \rho_t^q / \mathbb{E}_\pi(\rho_t^q)$,

$$\begin{aligned}
\frac{4}{q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} &\geq \frac{4}{q}\Big(\frac{\mathrm{ent}_\pi(\rho_t^q)}{2C_{\mathsf{MLSI}}\,\mathbb{E}_\pi(\rho_t^q)\,\widetilde{\mathfrak{m}}_p((1+\rho_t^q/\,\mathbb{E}_\pi(\rho_t^q))\,\pi)^{\delta(p)}}\Big)^{1/(1-\delta(p))} \\
&\geq \frac{1}{qC_{\mathsf{MLSI}}^2\,\widetilde{\mathfrak{m}}_p((1+\rho_t^q)\,\pi)^{\delta(p)/(1-\delta(p))}}\,\Big(\frac{\mathrm{ent}_\pi(\rho_t^q)}{\mathbb{E}_\pi(\rho_t^q)}\Big)^{1/(1-\delta(p))} \\
&\geq \frac{1}{qC_{\mathsf{MLSI}}^2\,\widetilde{\mathfrak{m}}_p((1+\rho_t^q)\,\pi)^{\delta(p)/(1-\delta(p))}}\,\mathcal{R}_q(\pi_t \parallel \pi)^{1/(1-\delta(p))}
\end{aligned}$$

$$\geq \frac{\varepsilon^{2\delta(p)/(1-\delta(p))}}{qC_{\mathsf{MLSI}}^2\,\widetilde{\mathfrak{m}}_p((1+\rho_t^q)\,\pi)^{\delta(p)/(1-\delta(p))}}\,\mathcal{R}_q(\pi_t \parallel \pi)$$

as long as $\mathcal{R}_q(\pi_t \parallel \pi) \geq \varepsilon^2$. Next, we bound the moments. It is a standard exercise [see Ver18, Exercise 2.7.3] to show that (α_1-tail) implies $\widetilde{\mathfrak{m}}_p(\pi)^{1/p} \lesssim \mathfrak{m} + C_{\mathsf{tail}}\,p^{1/\alpha_1}$. Also, by a slight modification of the change of measure principle (Lemma 3.6.9), we can show that $\widetilde{\mathfrak{m}}_p(\rho_t^q\pi)^{1/p} \lesssim \mathfrak{m} + C_{\mathsf{tail}}\,\mathcal{R}_2(\rho_t^q\pi \parallel \pi)^{1/\alpha_1} + C_{\mathsf{tail}}\,p^{1/\alpha_1}$, and that $\mathcal{R}_2(\rho_t^q\pi \parallel \pi) \lesssim q\mathcal{R}_{2q}(\pi_t \parallel \pi) \leq q\mathcal{R}_{2q}(\pi_0 \parallel \pi)$. Therefore,

$$\begin{aligned}\widetilde{\mathfrak{m}}_p\big((1+\rho_t^q)\,\pi\big)^{\delta(p)/(1-\delta(p))} &\leq \widetilde{\mathfrak{m}}_p(\pi)^{\delta(p)/(1-\delta(p))} + \widetilde{\mathfrak{m}}_p(\rho_t^q\pi)^{\delta(p)/(1-\delta(p))} \\ &\lesssim \{\mathfrak{m} + qC_{\mathsf{tail}}\,\mathcal{R}_{2q}(\pi_0 \parallel \pi)^{1/\alpha_1} + C_{\mathsf{tail}}\,p^{1/\alpha_1}\}^{(2-\alpha_0)\,(1+\alpha_0/(p-\alpha_0))}\,.\end{aligned}$$

Using the assumption that $\mathfrak{m}, C_{\mathsf{tail}}, \mathcal{R}_{2q}(\pi_0 \parallel \pi) \leq d^{O(1)}$, for $p \gtrsim \log d$,

$$\widetilde{\mathfrak{m}}_p\big((1+\rho_t^q)\,\pi\big)^{\delta(p)/(1-\delta(p))} \lesssim \{\mathfrak{m} + qC_{\mathsf{tail}}\,\mathcal{R}_{2q}(\pi_0 \parallel \pi)^{1/\alpha_1} + C_{\mathsf{tail}}\,p^{1/\alpha_1}\}^{2-\alpha_0}\,.$$

Together, it implies that $\mathcal{R}_q(\pi_T \parallel \pi) \leq \varepsilon^2$ whenever

$$T \geq \Omega\Big(\frac{qC_{\mathsf{MLSI}}^2}{\varepsilon^{4\delta(p)}}\,\{\mathfrak{m} + qC_{\mathsf{tail}}\,\mathcal{R}_{2q}(\pi_0 \parallel \pi)^{1/\alpha_1} + C_{\mathsf{tail}}\,p^{1/\alpha_1}\}^{2-\alpha_0}\,\ln\frac{\mathcal{R}_q(\pi_0 \parallel \pi)}{\varepsilon^2}\Big)\,.$$

Next, choosing $p \asymp \ln(d/\varepsilon^2)$, we obtain $\varepsilon^{4\delta(p)} \gtrsim 1$, so that

$$T \geq \Omega\Big(qC_{\mathsf{MLSI}}^2\,\{\mathfrak{m} + qC_{\mathsf{tail}}\,\mathcal{R}_{2q}(\pi_0 \parallel \pi)^{1/\alpha_1} + C_{\mathsf{tail}}\ln(d/\varepsilon^2)^{1/\alpha_1}\}^{2-\alpha_0}\,\ln\frac{\mathcal{R}_q(\pi_0 \parallel \pi)}{\varepsilon^2}\Big)\,,$$

completing the proof. □

With the continuous-time result in hand, it is now straightforward to combine it with the discretization result (Proposition 3.6.15) from the previous section.

Proof of Theorem 3.3.6. Let $(\mu_t)_{t\geq 0}$ denote the marginal law of the interpolated process (3.2) and let $(\pi_t)_{t\geq 0}$ denote the law of the continuous-time Langevin diffusion (3.1), both initialized at μ_0. By the weak triangle inequality (when $q \geq 2$; Lemma 2.2.23), we can bound

$$\mathcal{R}_q(\mu_{Nh} \parallel \pi) \lesssim \mathcal{R}_{2q}(\mu_{Nh} \parallel \pi_{Nh}) + \mathcal{R}_{2q-1}(\pi_{Nh} \parallel \pi)\,.$$

For $T \coloneqq Nh$, we can make the second term at most $\varepsilon^2/2$ if we choose

$$T = \widetilde{\Theta}(qC_{\mathsf{MLSI}}^2\,\{\mathfrak{m} + qC_{\mathsf{tail}}\,\mathcal{R}_{2q}(\mu_0 \parallel \pi)^{1/\alpha_1}\}^{2-\alpha_0})$$

by Theorem 3.2.3. Then, by Proposition 3.6.15, we can make the first term at most $\varepsilon^2/2$ taking

$$h = \widetilde{\Theta}_s\Big(\frac{\varepsilon^{2/s}}{dq^{(4-\alpha_0)/s} C_{\mathsf{MLSI}}^{2/s} C_{\mathsf{tail}}^{(2-\alpha_0)/s} L^{2/s}\, \mathcal{R}_{2q}(\mu_0 \,\|\, \pi)^{(2-\alpha_0)/(\alpha_1 s)}} \times \min\Big\{1, \frac{1}{q^{1/s}\varepsilon^{2/s}}, \frac{d}{\mathfrak{m}^s}, \frac{d}{\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})^{s/2}}, \Big(\frac{\mathcal{R}_{2q}(\mu_0 \,\|\, \pi)^{1/\alpha_1}}{\mathfrak{m}}\Big)^{(2-\alpha_0)/s}\Big\}\Big)\,. \tag{3.18}$$

Then, the total number of iterations of LMC is

$$\begin{aligned} N &= \frac{T}{h} \\ &= \widetilde{\Theta}_s\Big(\frac{dq^{(1+(3-\alpha_0)\,(1+s))/s} C_{\mathsf{MLSI}}^{2\,(1+1/s)} C_{\mathsf{tail}}^{(2-\alpha_0)\,(1+1/s)} L^{2/s}\, \mathcal{R}_{2q}(\mu_0 \,\|\, \pi)^{(2-\alpha_0)\,(1+1/s)/\alpha_1}}{\varepsilon^{2/s}} \\ &\qquad \times \max\Big\{1, q^{1/s}\varepsilon^{2/s}, \frac{\mathfrak{m}^s}{d}, \frac{\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})^{s/2}}{d}, \Big(\frac{\mathfrak{m}}{\mathcal{R}_{2q}(\mu_0 \,\|\, \pi)^{1/\alpha_1}}\Big)^{(2-\alpha_0)/s}\Big\}\Big)\,. \end{aligned}$$

This completes the proof. □

■ 3.6.6 Initialization

In this section, we give bounds on the Rényi divergence at initialization. We begin with the convex case.

Lemma 3.6.17. *Suppose that V is convex with $V(0) = 0$ and $\nabla V(0) = 0$, and assume that ∇V is L-Lipschitz. Let $\mathfrak{m} \coloneqq \int \|\cdot\| \,\mathrm{d}\pi$. Then, for $\mu_0 = \mathsf{normal}(0, L^{-1} I_d)$,*

$$\mathcal{R}_\infty(\mu_0 \,\|\, \pi) \le 2 + \frac{d}{2} \ln(2\mathfrak{m}^2 L)\,.$$

Proof. We can write

$$\sup \frac{\mu_0}{\pi} = \sup_{x \in \mathbb{R}^d} \exp\big\{V(x) - \frac{L}{2}\|x\|^2\big\} \frac{\int \exp(-V)}{\int \exp(-V - \delta\|\cdot\|^2)} \frac{\int \exp(-V - \delta\|\cdot\|^2)}{(2\pi/L)^{d/2}} \tag{3.19}$$

for some $\delta > 0$ to be chosen later. We bound the three ratios in turn. First,

$$\exp\big\{V(x) - \frac{L}{2}\|x\|^2\big\} \le 1$$

using $V(x) \le L\,\|x\|^2/2$. Next,

$$\frac{\int \exp(-V - \delta\,\|\cdot\|^2)}{\int \exp(-V)} = \int \exp(-\delta\,\|\cdot\|^2)\,\mathrm{d}\pi \ge \exp(-4\delta\mathfrak{m}^2)\,\pi\{\|\cdot\| \le 2\mathfrak{m}\}$$
$$\ge \frac{1}{2}\exp(-4\delta\mathfrak{m}^2)$$

by Markov's inequality. Finally, since $V \ge 0$,

$$\frac{\int \exp(-V - \delta\,\|\cdot\|^2)}{(2\pi/L)^{d/2}} \le \frac{\int \exp(-\delta\,\|\cdot\|^2)}{(2\pi/L)^{d/2}} = \big(\frac{L}{2\delta}\big)^{d/2}.$$

Taking $\delta = 1/(4\mathfrak{m}^2)$, we obtain

$$\mathcal{R}_\infty(\mu_0 \,\|\, \pi) = \ln \sup \frac{\mu_0}{\pi} \le 2 + \frac{d}{2}\ln(2\mathfrak{m}^2 L)\,,$$

which is $O(d)$, up to a logarithmic factor. □

We next extend this result to the general case.

Lemma 3.6.18. *Suppose that $\nabla V(0) = 0$ and that ∇V satisfies (s-Hölder) with constant $L > 0$. Let $\mathfrak{m} := \int \|\cdot\|\,\mathrm{d}\pi$. Then, for $\mu_0 = \mathsf{normal}(0, (2L)^{-1} I_d)$,*

$$\mathcal{R}_\infty(\mu_0 \,\|\, \pi) \le 2 + L + V(0) - \min V + \frac{d}{2}\ln(4\mathfrak{m}^2 L)\,.$$

Proof. We consider the same decomposition as in (3.19). First, for some $\lambda \in [0,1]$, we have

$$|V(x) - V(0)| = |\langle \nabla V(\lambda x), x\rangle| \le \|\nabla V(\lambda x) - \nabla V(0)\|\,\|x\| \le L\,\|x\|^{1+s}.$$

Therefore,

$$\exp\{V(x) - L\,\|x\|^2\} \le \exp\{V(x) - V(0) + V(0) - L\,\|x\|^2\}$$
$$\le \exp\{V(0) + L\,\|x\|^{1+s} - L\,\|x\|^2\} \le \exp\{V(0) + L\}$$

using $t^{1+s} \le 1 + t^2$ for all $t \ge 0$. Next,

$$\frac{\int \exp(-V - \delta\,\|\cdot\|^2)}{\int \exp(-V)} \ge \frac{1}{2}\exp(-4\delta\mathfrak{m}^2)$$

as before. Lastly,

$$\frac{\int \exp(-V - \delta\,\|\cdot\|^2)}{(\pi/L)^{d/2}} \le \frac{\exp(-\min V)\int \exp(-\delta\,\|\cdot\|^2)}{(\pi/L)^{d/2}} = \exp(-\min V)\,\big(\frac{L}{\delta}\big)^{d/2}.$$

This yields

$$\mathcal{R}_\infty(\mu_0 \,\|\, \pi) = \ln \sup \frac{\mu_0}{\pi} \le 2 + L + V(0) - \min V + \frac{d}{2} \ln(4\mathfrak{m}^2 L)\,,$$

with the choice $\delta = 1/(4\mathfrak{m}^2)$. □

In order to obtain an initialization with $\mathcal{R}_\infty(\mu_0 \,\|\, \pi) = \widetilde{O}(d)$, the lemma requires finding a stationary point $x \in \mathbb{R}^d$ such that the optimality gap $V(x) - \min V$ is not too large, i.e., of order $O(d)$. Since ∇V satisfies (s-Hölder), it suffices to find a stationary point which lies in a ball of radius $O(d^{1/(1+s)})$ centered at the minimizer of V. Based on this result, it seems reasonable to assume that the initialization typically satisfies $\mathcal{R}_\infty(\mu_0 \,\|\, \pi) = \widetilde{O}(d)$.

Actually, in the setting of Theorem 3.3.4, we also need a bound on the Rényi divergence $\mathcal{R}_2(\mu_0 \,\|\, \hat{\pi})$, where $\hat{\pi}$ is a slight modification of π (see Section 3.6.4). The following lemma is proven just as in Lemma 3.6.18, so the proof is omitted.

Lemma 3.6.19. *Suppose that $\nabla V(0) = 0$ and that ∇V satisfies (s-Hölder) with constant $L > 0$. For some $\gamma > 0$, let $\hat{V}(x) := V(x) + \frac{\gamma}{2}(\|x\| - R)_+^2$, and let $\hat{\pi} \propto \exp(-\hat{V})$. Also, let $\hat{\mathfrak{m}} := \int \|\cdot\| \,\mathrm{d}\hat{\pi}$. Then, for $\mu_0 = \mathsf{normal}(0, (2L+\gamma)^{-1} I_d)$,*

$$\mathcal{R}_\infty(\mu_0 \,\|\, \hat{\pi}) \le 2 + L + \frac{\gamma}{2} + V(0) - \min V + \frac{d}{2} \ln(4\hat{\mathfrak{m}}^2 L)\,.$$

From the tail bound in Lemma 3.6.11, we can deduce an upper bound for $\hat{\mathfrak{m}}$ as follows

$$\begin{aligned}
\hat{\mathfrak{m}} &= \int_0^\infty \hat{\pi}(\|\cdot\| \ge t)\,\mathrm{d}t \\
&= \int_0^R \hat{\pi}(\|\cdot\| \ge t)\,\mathrm{d}t + \int_0^\infty \hat{\pi}(\|\cdot\| \ge R + \eta)\,\mathrm{d}\eta \\
&\le R + \int_0^\infty 2\exp\Big(-\frac{\gamma\eta^2}{2}\Big)\,\mathrm{d}\eta \\
&\lesssim R + \sqrt{\frac{1}{\gamma}}\,.
\end{aligned}$$

In Proposition 3.6.14, we eventually take γ roughly of order $1/d \lesssim \gamma \lesssim 1$, and $R \lesssim \mathfrak{m}$. Hence, if $L + V(0) - \min V = \widetilde{O}(d)$ and $\mathfrak{m} \le d^{O(1)}$, then $\mathcal{R}_\infty(\mu_0 \| \hat{\pi}) = \widetilde{O}(d)$.

■ 3.6.7 Additional technical lemmas

In this section, we collect together technical lemmas which appear in the proofs of §3.6.4. The proofs rely on standard arguments from stochastic calculus.

We first present a bound on the moment generating function of the supremum of a one-dimensional Brownian motion using the reflection principle.

Lemma 3.6.20. *Let $(B_s)_{s\geq 0}$ be a standard one-dimensional Brownian motion. For $h, \lambda > 0$, such that $\lambda < \frac{1}{2h}$ the following holds:*

$$\mathbb{E}\exp\Big(\lambda \sup_{s\in[0,h]} |B_s|^2\Big) \leq \frac{1+2h\lambda}{1-2h\lambda}\,.$$

Proof. The reflection principle [KS91, Proposition 6.19, 2.2.6] states that for every $t > 0$, it holds that

$$\mathbb{P}\big(\sup_{s\in[0,h]} B_s > t\big) = 2\,\mathbb{P}(B_h > t)\,.$$

As a result, we have that

$$\begin{aligned}\mathbb{P}\big(\sup_{s\in[0,h]} |B_s|^2 > t\big) &= \mathbb{P}\big(\sup_{s\in[0,h]} |B_s| > \sqrt{t}\big)\\ &\leq \mathbb{P}\big(\sup_{s\in[0,h]} B_s > \sqrt{t}\big) + \mathbb{P}\big(\inf_{s\in[0,h]} B_s < -\sqrt{t}\big)\\ &= 4\,\mathbb{P}(B_h > \sqrt{t}) \leq 2\exp\big(-\frac{t}{2h}\big)\,.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}\exp\Big(\lambda \sup_{s\in[0,h]} |B_s|^2\Big) &= 1 + \lambda\int_0^\infty \exp(\lambda t)\,\mathbb{P}\big(\sup_{s\in[0,h]} |B_s|^2 > t\big)\,\mathrm{d}t\\ &\leq 1 + 2\lambda\int_0^\infty \exp\big(-\frac{1-2h\lambda}{2h}\,t\big)\,\mathrm{d}t = 1 + \frac{4h\lambda}{1-2h\lambda}\,.\end{aligned}$$

□

Lemma 3.6.21. *Let $(B_t)_{t\geq 0}$ be a standard Brownian motion in $\mathbb{R}^d$. Then, if $\lambda \geq 0$ and $h \leq 1/(4\lambda)$,*

$$\mathbb{E}\exp\big(\lambda \sup_{t\in[0,h]} \|B_t\|^2\big) \leq \exp(6dh\lambda)\,.$$

In particular, for all $\eta \geq 0$,

$$\mathbb{P}\big\{\sup_{t\in[0,h]} \|B_t\| \geq \eta\big\} \leq 3\exp\big(-\frac{\eta^2}{6dh}\big)\,.$$

Next, for $s \in (0,1)$ and $0 \leq \lambda < 1/(12dh)^s$,

$$\mathbb{E}\exp\big(\lambda \sup_{t\in[0,h]} \|B_t\|^{2s}\big) \leq \exp(144d^s h^s \lambda)\,.$$

Proof. The first statement follows from Lemma 3.6.20, and the second follows from the first by taking $\lambda = 1/(6dh)$ and applying Markov's inequality.

We now turn towards the proof of the third statement. Using the tail bound

$$\mathbb{P}\Big\{ \sup_{t\in[0,h]} \|B_t\|^{2s} \geq \eta \Big\} \leq 3\exp\big(-\frac{\eta^{1/s}}{6dh}\big)$$

we now bound $\mathbb{E}\exp(\lambda \sup_{t\in[0,h]} \|B_t\|^{2s})$.

$$\begin{aligned}\mathbb{E}\exp\big(\lambda \sup_{t\in[0,h]} \|B_t\|^{2s}\big) &= 1 + \lambda\int_0^\infty \exp(\lambda\eta)\,\mathbb{P}\Big\{ \sup_{t\in[0,h]} \|B_t\|^{2s} \geq \eta\Big\}\,\mathrm{d}\eta \\ &\leq 1 + 3\lambda\int_0^\infty \exp\big(\lambda\eta - \frac{\eta^{1/s}}{6dh}\big)\,\mathrm{d}\eta\,.\end{aligned}$$

Split the integral into whether or not $\eta \geq (12dh\lambda)^{s/(1-s)}$. For the first part,

$$\begin{aligned}\lambda\int_0^{(12dh\lambda)^{s/(1-s)}} \exp(\lambda\eta)\,\mathrm{d}\eta &\leq (12dh)^{s/(1-s)}\lambda^{1/(1-s)}\exp\{(12dh)^{s/(1-s)}\lambda^{1/(1-s)}\} \\ &\leq 3\,(12dh)^{s/(1-s)}\lambda^{1/(1-s)}\end{aligned}$$

provided that $\lambda \leq 1/(12dh)^s$. For the second part, using the change of variables $\tau = \eta^{1/s}/(12dh)$,

$$\begin{aligned}\lambda\int_{(12dh\lambda)^{s/(1-s)}}^\infty \exp\big(\lambda\eta - \frac{\eta^{1/s}}{6dh}\big)\,\mathrm{d}\eta &\leq \lambda\int_{(12dh\lambda)^{s/(1-s)}}^\infty \exp\big(-\frac{\eta^{1/s}}{12dh}\big)\,\mathrm{d}\eta \\ &\leq (12dh)^s s\lambda\int_0^\infty \frac{\exp(-\tau)}{\tau^{1-s}}\,\mathrm{d}\tau \\ &= (12dh)^s s\lambda\Gamma(s) = (12dh)^s\lambda\Gamma(1+s) \\ &\leq (12dh)^s\lambda\,,\end{aligned}$$

where we used Gautschi's inequality to obtain $\Gamma(1+s) \leq 1$. We have proven

$$\mathbb{E}\exp\big(\lambda \sup_{t\in[0,h]} \|B_t\|^{2s}\big) \leq 1 + 9\,(12dh)^{s/(1-s)}\lambda^{1/(1-s)} + 3\,(12dh)^s\lambda \leq 1 + 144d^sh^s\lambda\,,$$

which implies the result. □

Lemma 3.6.22. *Let $(Z_t)_{t\geq 0}$ denote the continuous-time Langevin diffusion (3.1) started at Z_0, and assume that the gradient ∇V of the potential satisfies $\nabla V(0) = 0$ and (s-Hölder). Also, assume that $h \leq 1/(6L)$ and $\lambda \leq 1/(96d^sh^s)$. Then,*

$$\mathbb{E}\exp\big(\lambda \sup_{t\in[0,h]} \|Z_t - Z_0\|^{2s}\big) \leq \exp\{8h^{2s}L^{2s}\,(1 + \|Z_0\|^{2s^2})\,\lambda + 1152d^sh^s\lambda\}\,.$$

Proof. Let $f(t) \coloneqq \sup_{r\in[0,t]} \|Z_r - Z_0\|^2$. Then, for $0 \le t \le h$, since $\|\nabla V(x)\| \le L\|x\|^s$,

$$\begin{aligned}
\|Z_t - Z_0\|^2 &= \left\| -\int_0^t \nabla V(Z_r)\,\mathrm{d}r + \sqrt{2}\,B_t \right\|^2 \le 2t\int_0^t \|\nabla V(Z_r)\|^2\,\mathrm{d}r + 4\|B_t\|^2 \\
&\le 4t\int_0^t \|\nabla V(Z_r) - \nabla V(Z_0)\|^2\,\mathrm{d}r + 4t^2\|\nabla V(Z_0)\|^2 + 4\|B_t\|^2 \\
&\le 4tL^2\int_0^t \|Z_r - Z_0\|^{2s}\,\mathrm{d}r + 4t^2L^2\|Z_0\|^{2s} + 4\|B_t\|^2 \\
&\le 4tL^2\int_0^t \|Z_r - Z_0\|^2\,\mathrm{d}r + 4t^2L^2\left(1 + \|Z_0\|^{2s}\right) + 4\|B_t\|^2\,,
\end{aligned}$$

which yields

$$f(t) \le 4t^2L^2\left(1 + \|Z_0\|^{2s}\right) + 4\sup_{r\in[0,t]}\|B_r\|^2 + 4tL^2\int_0^t f(r)\,\mathrm{d}r\,.$$

Grönwall's inequality yields

$$\begin{aligned}
f(h) &\le \left(4h^2L^2\left(1 + \|Z_0\|^{2s}\right) + 4\sup_{r\in[0,h]}\|B_r\|^2\right)\exp(2h^2L^2) \\
&\le 8h^2L^2\left(1 + \|Z_0\|^{2s}\right) + 8\sup_{r\in[0,h]}\|B_r\|^2
\end{aligned}$$

using $h \le 1/(6L)$. It also yields

$$\sup_{t\in[0,h]}\|Z_t - Z_0\|^{2s} \le 8h^{2s}L^{2s}\left(1 + \|Z_0\|^{2s^2}\right) + 8\sup_{r\in[0,h]}\|B_r\|^{2s}\,.$$

The result now follows from Lemma 3.6.21. □

■ 3.7 Conclusion

In this work, we have given a suite of sampling guarantees for the LMC algorithm which assume only that a functional inequality and a smoothness condition hold. In particular, no such guarantees were previously known beyond the LSI case considered in [VW19]. Consequently, we have resolved the open questions of estimating the Rényi bias of LMC (Corollary 3.3.2) and establishing quantitative convergence guarantees for LMC under a Poincaré inequality. Our results and techniques are also of interest because they work with a stronger metric (namely, Rényi divergence) than what is usually considered in the sampling literature.

To conclude, we list a few directions for future research.

- It is not clear how sharp our bounds are, and it is worth investigating whether our techniques can be improved.
- As discussed in the introduction, obtaining guarantees in Rényi divergence is useful for applications to differential privacy, as well as for obtaining warm starts for high-accuracy algorithms. Hence, we ask whether Rényi convergence guarantees can be proved for more sophisticated algorithms, such as randomized midpoint discretizations [SL19; HBE20].

In follow-up work [Zha+23] with Matthew Zhang, Mufan (Bill) Li, Krishnakumar Balasubramanian, and Murat A. Erdogdu, I have also extended the techniques in this chapter to study the underdamped Langevin Monte Carlo (ULMC) algorithm. The strongly log-concave case of these results will be presented in §6, where it will be applied to provide a warm start for the Metropolis-adjusted Langevin algorithm (MALA).

Chapter 4

Analysis of the proximal sampler

In the previous chapter, we obtained new sampling guarantees for LMC under isoperimetry, albeit at the expense of somewhat involved calculations. Moreover, there are notable weaknesses of the LMC algorithm itself. For instance, it is *biased*: its stationary distribution for positive step size $h > 0$ does not equal the desired target π, and consequently the step size must be chosen appropriately small to control the size of this bias.

To overcome these issues, we ask the following question: since LMC can be interpreted as a discretization of the Wasserstein gradient flow for the KL divergence, are there better methods of implementing this flow? In this chapter, we study the proximal sampler algorithm of [TP18; LST21c] based on a novel interpretation as an alternating iteration of Brownian motion forward and backward in time. It leads to new convergence bounds which then translate into improved sampling guarantees under isoperimetry, with simpler proofs than §3.

This chapter is based on [Che+22b], joint with Yongxin Chen, Adil Salim, and Andre Wibisono.

4.1 Introduction

We again study the problem of sampling from a target density $\pi^X \propto \exp(-V)$ on $\mathbb{R}^d$, which enjoys surprising and deep connections with the field of optimization. Indeed, the standard Langevin algorithm can be viewed as a gradient flow of the Kullback–Leibler (KL) divergence on the space of probability measures equipped with the geometry of optimal transport (see §2.2), a perspective which has led to new analyses [DMM19; SR20] and algorithms [Per16; Zha+20; DL21; Ma+21] inspired by the theory of convex optimization.

Among the algorithms in the optimization toolkit, we focus on *proximal* methods. Classically, proximal methods are used to minimize composite objectives of the form $f + g$, where g is smooth and convex and f is non-smooth but simple enough to allow for evaluation of the proximal map $\mathrm{prox}_f : y \mapsto$

$\arg\min_{x\in\mathbb{R}^d}\{f(x)+\frac{1}{2h}\|x-y\|^2\}$. However, the setting of our investigation is more closely related to the minimization of a non-composite objective f, for which the proximal method is known as the *proximal point algorithm* [Mar70; Roc76].

As a natural first step towards developing a proximal point algorithm for sampling, one can combine the proximal map with the standard Langevin algorithm, leading to the *proximal Langevin algorithm*. This algorithm was introduced in [Per16] and analyzed in the papers [Ber18; Wib19; SR20]. Although these results are encouraging, the analogy between optimization methods and Langevin-based algorithms is imperfect because the discretization of the latter leads to asymptotic *bias*, a feature which is typically not present in optimization (see [Wib18] for a thorough discussion).

Remarkably, a new proximal algorithm for sampling was proposed recently in [LST21c] which overcomes this issue via a novel Gibbs sampling approach. Briefly, the *proximal sampler* is a sampling algorithm which assumes access to samples from an oracle distribution, known as the *restricted Gaussian oracle* (RGO); the RGO is a sampling analogue of the proximal map from optimization. Under this assumption, as well as the additional assumption that the target π^X is strongly log-concave, [LST21c] proved that the proximal sampler converges exponentially fast to π^X in total variation distance. In their paper, the proximal sampler was used as a *reduction framework* to improve the condition number dependence of other sampling algorithms. Indeed, the RGO is a better conditioned distribution than the target distribution, so that implementing the RGO is easier than solving the original sampling task. In turn, the reduction framework allowed them to establish improved complexity results for a variety of structured log-concave sampling problems. We review the proximal sampler and its implementability in §4.2.2.

Our contributions. Prior to our work, the convergence of the proximal sampler was only known in the case when $\pi^X \propto \exp(-V)$ is strongly log-concave. In this chapter, we greatly expand the classes of targets to which the proximal sampler is applicable by providing new convergence guarantees.

First, we consider the case when V is weakly convex. We show that after k iterations, the proximal sampler outputs a distribution whose KL divergence to the target is $O(1/k)$. Our proof is analogous to, and is inspired by, the corresponding guarantee for minimizing a weakly convex function (in particular, the $O(1/k)$ rate matches the optimization result).

Next, we assume that π^X satisfies a *functional inequality*, e.g., a Poincaré inequality or a log-Sobolev inequality. Such functional inequalities have been employed in the sampling literature as tractable settings for non-log-concave sampling; see [VW19] and §3. For these distributions, we show that the proximal

sampler converges to the target in Rényi divergence (or any other weaker metric, such as KL divergence) with a rate that matches the known convergence rates for the continuous-time Langevin diffusion under the same assumptions.

In each of these settings, if we additionally assume that ∇V is Lipschitz, then the RGO is implementable, as it becomes a smooth strongly log-concave distribution. Hence, we obtain new sampling guarantees for gradient Lipschitz potentials when the target is weakly log-concave or satisfies a functional inequality. In all cases, our results are *stronger* than known results in the literature. Subsequent works have also considered implementability of the RGO under weaker smoothness conditions [GLL22; LC22; LC23].

Finally, we clarify the connection between the proximal sampler and the proximal point algorithm in optimization in the following ways: (1) We show that convergence proofs for the proximal sampler can be translated to yield convergence proofs for the proximal point algorithm. As a consequence, we obtain a new convergence guarantee for the proximal point method under a gradient domination condition with optimal rate, which is (to the best of our knowledge) a new result. (2) We show that the RGO can be interpreted as a proximal mapping on the Wasserstein space.

Other related work. Sampling algorithms which are conceptually similar or directly related to the proximal sampler have been previously proposed in the literature [GC11; Mar+16; TP18; VPD22]. The RGO has also been considered as an adjoint of the heat semigroup in [KP21], which was then used in the recent breakthrough on the KLS conjecture in [KL22]. After the first version of our work appeared online, our result under LSI (Theorem 4.3.3) was recovered via the framework of localization schemes in [CE22].

Organization. The rest of the chapter is organized as follows. We begin with background on the proximal sampler in §4.2. We then give our main results in §4.3. In particular, we state our new convergence guarantees for the proximal sampler in §4.3.1, and we give applications of our results in §4.3.2. We then describe the connections between the proximal sampler and the proximal point method in §4.3.3. All proofs are given in §4.4.

Finally, we conclude and list open directions in §4.6.

4.2 Background and notation

4.2.1 Divergences between probability measures

Throughout the paper, we abuse notation by identifying a probability measure with its density w.r.t. Lebesgue measure.

We recall the following divergences between probability measures, see §2.2.3 for further background. For a probability measure $\rho \ll \pi$, we define the *KL divergence*, the *chi-squared divergence*, and the *Rényi divergence* of order $q \geq 1$ respectively via

$$\mathsf{KL}(\rho \,\|\, \pi) \coloneqq \int \rho \log \frac{\rho}{\pi}\,, \;\; \chi^2(\rho \,\|\, \pi) \coloneqq \int \frac{\rho^2}{\pi} - 1\,, \;\; \mathcal{R}_q(\rho \,\|\, \pi) \coloneqq \frac{1}{q-1} \log \int \frac{\rho^q}{\pi^{q-1}}\,,$$

with $\mathcal{R}_1 = \mathsf{KL}$. We recall that for $1 \leq q \leq q' < \infty$, we have the monotonicity property $\mathcal{R}_q \leq R_{q'}$, and that $\mathcal{R}_2 = \ln(1 + \chi^2)$.

We also define the 2-Wasserstein distance between ρ and π to be

$$W_2^2(\rho, \pi) \coloneqq \inf_{\gamma \in \mathcal{C}(\rho,\pi)} \int \|x - y\|^2 \,\mathrm{d}\gamma(x, y)\,,$$

where $\mathcal{C}(\rho, \pi)$ is the set of *couplings* of ρ and π, i.e., joint distributions on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are ρ and π.

■ 4.2.2 The proximal sampler

Our goal is to sample from a target probability distribution π^X on $\mathbb{R}^d$ with density $\pi^X \propto \exp(-V)$ and finite second moment, where $V \colon \mathbb{R}^d \to \mathbb{R}$ is the *potential*.

Following [LST21c], we define the joint target distribution $\boldsymbol{\pi}$ on $\mathbb{R}^d \times \mathbb{R}^d$ with (Lebesgue) density

$$\boldsymbol{\pi}(x, y) \propto \exp\Big(-V(x) - \frac{1}{2h} \|x - y\|^2\Big)\,,$$

where $h > 0$ is the *step size* of the algorithm.

Observe that the X-marginal of $\boldsymbol{\pi}$ is equal to the original target distribution π^X, whereas the conditional distribution of Y given X is Gaussian: $\pi^{Y|X}(\cdot \mid x) = \mathsf{normal}(x, hI)$. Therefore, the Y-marginal is the convolution of π^X with a Gaussian, $\pi^Y = \pi^X * \mathsf{normal}(0, hI)$. The perspective that we adopt in our proofs is that π^Y is obtained by evolving π^X along the heat flow for time h.

The conditional distribution of X given Y is the "regularized" distribution

$$\pi^{X|Y}(x \mid y) \propto_x \exp\Big(-V(x) - \frac{1}{2h} \|x - y\|^2\Big)\,.$$

The *restricted Gaussian oracle* (RGO) is defined as an oracle that, given $y \in \mathbb{R}^d$, outputs a random variable distributed according to $\pi^{X|Y}(\cdot \mid y)$. We also write $\pi^{X|Y}(\cdot \mid y) = \pi^{X|Y=y}$.

Proximal sampler: The proximal sampler is initialized at a point $X_0 \in \mathbb{R}^d$ and performs Gibbs sampling on the joint target $\boldsymbol{\pi}$. That is, the proximal sampler iterates the following two steps:

1. From X_k, sample $Y_k \mid X_k \sim \pi^{Y|X}(\cdot \mid X_k) = \mathsf{normal}(X_k, hI)$.
2. From X_k, sample $X_{k+1} \mid Y_k \sim \pi^{X|Y}(\cdot \mid Y_k)$.

The first step consists in sampling a Gaussian random variable centered at X_k, and is therefore easy to implement. The second step calls the RGO at the point Y_k and requires a suitable implementation, as we discuss below.

As is well-known from the theory of Gibbs sampling, the iterates $(X_k, Y_k)_{k\in\mathbb{N}}$ form a reversible Markov chain with stationary distribution $\boldsymbol{\pi}$. That is, the proximal sampler is an *unbiased* sampling algorithm, unlike algorithms based on discretizations of stochastic processes such as the unadjusted Langevin algorithm. This is because the proximal sampler is an idealized algorithm in which we assume *exact* access to the RGO. For our applications, we implement the RGO via rejection sampling; see §4.3.2 for details and §4.3.4 for an explicit example in the Gaussian case. We develop faster implementations for the RGO in §6.

4.3 Results for the proximal sampler

4.3.1 New convergence results for the proximal sampler

In this section, we describe our new convergence results for the proximal sampler under various assumptions, beginning with the strongly log-concave and weakly log-concave cases, and then proceeding to targets satisfying functional inequalities which allow for non-log-concavity.

4.3.1.1 Strong log-concavity

We start by recalling the W_2 contraction result from [LST21c][1] for the proximal sampler under strong log-concavity.

Theorem 4.3.1 ([LST21b, Lemma 2])**.** *Assume that $\pi^X \propto \exp(-V)$ is α-strongly log-concave (i.e., V is α-strongly convex), where $\alpha \geq 0$. For any $h > 0$ and for any two initial distributions ρ_0^X, $\bar{\rho}_0^X$, after k iterations of the proximal sampler with step size h, the respective distributions ρ_k^X, $\bar{\rho}_k^X$ satisfy the bound*

$$W_2(\rho_k^X, \bar{\rho}_k^X) \leq \frac{W_2(\rho_0^X, \bar{\rho}_0^X)}{(1+\alpha h)^k}\,. \tag{4.1}$$

[1] See the arXiv version [LST21b] for the proof.

Although this result was stated in [LST21b] as a convergence result rather than a contraction, the latter is implicit in the proof. From the proof of [LST21b], one can also read off a convergence guarantee in KL divergence, although this will be a corollary of our result in §4.3.1.3.

We revisit Theorem 4.3.1 in §4.4.2 and provide a proof which more closely resembles a classical convergence proof of the proximal point algorithm. We use Wasserstein subdifferential calculus.

We note that this is the sampling analogue of the classical fact that the proximal map for an α-strongly convex function with step size h is a $\frac{1}{1+\alpha h}$-contraction. In §4.5.1, we give a new proof of this fact about the proximal point method by translating the proof of [LST21b] into optimization.

■ 4.3.1.2 Log-concavity

The preceding result does not yield convergence when $\alpha = 0$. We provide a new convergence guarantee for the weakly convex case which mirrors a Lyapunov analysis of gradient flows for convex functions.

Theorem 4.3.2. *Assume that $\pi^X \propto \exp(-V)$ is log-concave (i.e., V is convex). For the k-th iterate ρ_k^X of the proximal sampler,*

$$\mathsf{KL}(\rho_k^X \parallel \pi^X) \le \frac{W_2^2(\rho_0^X, \pi^X)}{kh}\,.$$

Proof. §4.4.3. □

■ 4.3.1.3 Log-Sobolev inequality

Recall from §2.2 that a probability distribution π satisfies the log-Sobolev inequality (LSI) with constant $1/\alpha > 0$ ($1/\alpha$-LSI) if for any probability distribution ρ, the following inequality holds:

$$\mathsf{KL}(\rho \parallel \pi) \le \frac{1}{2\alpha}\,\mathsf{FI}(\rho \parallel \pi)\,. \tag{4.2}$$

Here $\mathsf{FI}(\rho \parallel \pi)$ is the Fisher information of ρ w.r.t. π. Recall that strong log-concavity implies LSI, and that LSI is equivalent to the gradient domination condition for relative entropy $\mathsf{KL}(\cdot \parallel \pi)$; see also §4.3.3.1.

Theorem 4.3.3. *Assume that $\pi^X \propto \exp(-V)$ satisfies $1/\alpha$-LSI. For any $h > 0$ and any initial distribution ρ_0^X, the k-th iterate ρ_k^X of the proximal sampler with step size h satisfies*

$$\mathsf{KL}(\rho_k^X \parallel \pi^X) \le \frac{\mathsf{KL}(\rho_0^X \parallel \pi^X)}{(1+\alpha h)^{2k}}\,. \tag{4.3}$$

Furthermore, for all $q \geq 1$:

$$\mathcal{R}_q(\rho_k^X \parallel \pi^X) \leq \frac{\mathcal{R}_q(\rho_0^X \parallel \pi^X)}{(1+\alpha h)^{2k/q}}\,. \tag{4.4}$$

Proof. §4.4.4. □

■ 4.3.1.4 Poincaré inequality

Recall from §2.2 that a probability distribution π satisfies the Poincaré inequality (PI) with constant $1/\alpha > 0$ ($1/\alpha$-PI) if for any smooth compactly supported function $\psi : \mathbb{R}^d \to \mathbb{R}$, the following inequality holds:

$$\mathrm{var}_\pi(\psi) \leq \frac{1}{\alpha}\,\mathbb{E}_\pi[\|\nabla\psi\|^2]\,. \tag{4.5}$$

Recall also that $1/\alpha$-LSI implies $1/\alpha$-PI.

Theorem 4.3.4. *Assume $\pi^X \propto \exp(-V)$ satisfies $1/\alpha$-PI. For any $h > 0$ and any initial distribution ρ_0^X, the k-th iterate ρ_k^X of the proximal sampler with step size h satisfies*

$$\chi^2(\rho_k^X \parallel \pi^X) \leq \frac{\chi^2(\rho_0^X \parallel \pi^X)}{(1+\alpha h)^{2k}}\,. \tag{4.6}$$

Furthermore, for all $q \geq 2$, if we set

$$c_0 \coloneqq \frac{q}{2\ln(1+\alpha h)}\left(\mathcal{R}_q(\rho_0^X \parallel \pi^X) - 1\right),$$

then

$$\mathcal{R}_q(\rho_k^X \parallel \pi^X) \leq \begin{cases} \mathcal{R}_q(\rho_0^X \parallel \pi^X) - \frac{2k\ln(1+\alpha h)}{q}\,, & \text{if } k \leq c_0\,, \\ 1/(1+\alpha h)^{2(k-k_0)/q}\,, & \text{if } k \geq \lceil c_0 \rceil\,. \end{cases} \tag{4.7}$$

Proof. §4.4.5. □

■ 4.3.1.5 Latała–Oleszkiewicz inequality

We next consider a family of functional inequalities which interpolate between PI and LSI. A probability distribution π satisfies the Latała–Oleszkiewicz inequality (LOI) of order $r \in [1,2]$ and constant $1/\alpha > 0$ ($(r, 1/\alpha)$-LOI) if for any smooth bounded function $\psi : \mathbb{R}^d \to \mathbb{R}_+$, the following inequality holds:

$$\sup_{p\in(1,2)} \frac{\mathrm{var}_{p,\pi}(\psi)}{(2-p)^{2(1-1/r)}} \coloneqq \sup_{p\in(1,2)} \frac{\mathbb{E}_\pi[\psi^2] - \mathbb{E}_\pi[\psi^p]^{2/p}}{(2-p)^{2(1-1/r)}} \leq \frac{1}{\alpha}\,\mathbb{E}_\pi[\|\nabla\psi\|^2]\,.$$

This inequality was introduced in [LO00], and sampling guarantees for the Langevin algorithm under LOI were given in §3. The LOI for $r = 1$ is equivalent to PI and the LOI for $r = 2$ is equivalent to LSI, up to absolute constants. Generally speaking, (r, α)-LOI captures targets $\pi \propto \exp(-V)$ such that the tails of V grow as $\|\cdot\|^r$ at infinity.

Theorem 4.3.5. *Assume $\pi^X \propto \exp(-V)$ satisfies $(r, 1/\alpha)$-LOI with $r \in [1, 2)$. For any $h > 0$, $q \geq 2$, and any initial distribution ρ_0^X, the k-th iterate ρ_k^X of the proximal sampler with step size h satisfies*

$$\mathcal{R}_q(\rho_k^X \parallel \pi^X) \leq \begin{cases} \left(\mathcal{R}_q(\rho_0^X \parallel \pi^X)^{2/r-1} - \frac{(2/r-1)\, k \ln(1+\alpha h)}{68q}\right)^{r/(2-r)}, & \text{if } k \leq c_0\,, \\ 1/(1+\alpha h)^{(k-\lceil c_0 \rceil)/(68q)}\,, & \text{if } k \geq \lceil c_0 \rceil\,, \end{cases} \tag{4.8}$$

where

$$c_0 \coloneqq \frac{68q}{(2/r-1)\ln(1+\alpha h)} \left(\mathcal{R}_q(\rho_0^X \parallel \pi^X)^{2/r-1} - 1\right).$$

(For $r = 2$, we can instead use Theorem 4.3.3.)

Proof. §4.4.6. □

To interpret the result, suppose that $\mathcal{R}_q(\rho_0^X \parallel \pi^X) = O(d)$ at initialization and that $h \ll 1/\alpha$. Then, the theorem states that after an initial waiting period of $\lceil c_0 \rceil = O(d^{2/r-1}/h)$ iterations, in which the Rényi divergence decays to $O(1)$, the Rényi divergence decays exponentially thereafter. This interpolates between a waiting time of $O(d/h)$ under PI ($r = 1$; Theorem 4.3.4) and a waiting time of $O((\log d)/h)$ under LSI ($r = 2$; Theorem 4.3.3).

■ 4.3.2 Applications of the convergence results

We start with a corollary of Theorem 4.3.2. Suppose that f is β-smooth, i.e., ∇f is β-Lipschitz. Then, provided $\frac{1}{h} \geq \beta$, the RGO $\pi^{X|Y}$ is strongly-log-concave, with condition number $(1 + \beta h)/(1 - \beta h) \leq O(1)$. We can implement the RGO via rejection sampling.

Rejection sampling: Given a target distribution $\tilde{\pi} \propto \exp(-\tilde{V})$, where $\tilde{V}$ is $\tilde{\alpha}$-strongly convex, perform the following steps.

1. Compute the minimizer $x^\star$ of $\tilde{V}$.
2. Repeat until acceptance: draw a random variable $Z \sim \mathcal{N}(x^\star, \tilde{\alpha}^{-1} I)$ and accept it with probability $\exp(-\tilde{V}(Z) + \tilde{V}(x^\star) + \frac{\tilde{\alpha}}{2}\|Z - x^\star\|^2)$.

The resulting sample is distributed according to $\tilde{\pi}$, and one can show that the expected number of iterations of the algorithm is bounded by $\tilde{\kappa}^{d/2}$ with $\tilde{\kappa} \coloneqq \tilde{\beta}/\tilde{\alpha}$ and $\tilde{\beta}$ is the smoothness of $\tilde{V}$; see, e.g., §4.4.7.

We apply this to $\tilde{V}$ given by $\tilde{V}(x) = V(x) + \frac{1}{2h}\|x-y\|^2$. The algorithm above requires exact minimization of $\tilde{V}$, which we assume for simplicity (since it is well-known how to efficiently minimize a strongly convex and smooth function). With the choice $h \asymp \frac{1}{\beta d}$, the expected number of iterations is $O(1)$. Combining this implementation of the RGO with Theorem 4.3.2, we obtain:

Corollary 4.3.6. *Suppose $\pi^X \propto \exp(-V)$ where V is convex and β-smooth. Take $h \asymp \frac{1}{\beta d}$ and implement the RGO with rejection sampling as described above. Then, the proximal sampler outputs ρ_k^X with $\mathsf{KL}(\rho_k^X \parallel \pi^X) \le \varepsilon^2$ and the expected number of calls to an oracle for V is $O(\beta d\, W_2^2(\rho_0^X, \pi^X)/\varepsilon^2)$.*

More precisely, our algorithm requires access to an oracle of V which can evaluate V and compute the proximity operator for V.

We now compare this rate with others in the literature. Let $\mathfrak{m}_2$ denote the second moment of π^X. For example, $\mathfrak{m}_2 = O(d)$ for a product measure, and $\mathfrak{m}_2 = O(d^2)$ when $V(x) = \sqrt{1+\|x\|^2}$. It is reasonable to assume that the Poincaré constant α of π^X is $O(\mathfrak{m}_2/d)$ and that $W_2^2(\rho_0^X, \pi^X) = O(\mathfrak{m}_2)$. With these simplifications, our complexity is $O(\beta d \mathfrak{m}_2/\varepsilon^2)$; averaged LMC achieves $\widetilde{O}(\beta d \mathfrak{m}_2/\varepsilon^4)$ [DMM19]; MALA achieves $\widetilde{O}(\beta^{3/2} d^{1/2} \mathfrak{m}_2^{3/2}/\varepsilon^{3/2})$ albeit in the TV distance [Dwi+19; Che+20a]; and LMC achieves $\widetilde{O}(\beta^2 \mathfrak{m}_2^2/\varepsilon^2)$ in the stronger Rényi metric (§3). Since all these complexity results also hold in terms of the total variation distance, our result is arguably the best one for this setting (at least, if dimension dependence is the primary consideration).

Similarly, implementing the RGO with rejection sampling in Theorem 4.3.5 yields the following corollary:

Corollary 4.3.7. *Suppose $\pi^X \propto \exp(-V)$ where V is β-smooth and π^X satisfies $(r, 1/\alpha)$-LOI. Take $h \asymp \frac{1}{\beta d}$ and implement the RGO with rejection sampling as described above. Then, the proximal sampler outputs ρ_k^X with $\mathcal{R}_q(\rho_k^X \parallel \pi^X) \le \varepsilon^2$ and the expected number of calls to an oracle for V is*

$$\widetilde{O}\Big(\frac{\beta d q}{\alpha}\left(\mathcal{R}_q(\rho_0^X \parallel \pi^X)^{2/r-1} \vee \log\frac{1}{\varepsilon}\right)\Big).$$

Even for the special case of a Poincaré inequality and smoothness, the first sampling guarantee under these assumptions is the one in §3. Let us write $\hat{\kappa} \coloneqq \beta/\alpha$ for the "condition number" and assume $\mathcal{R}_q(\rho_0^X \parallel \pi^X) = O(d)$ (see, e.g., §3.6.6). Then, our complexity is $\widetilde{O}(\hat{\kappa} d q\,(d^{2/r-1} \vee \log(1/\varepsilon)))$, whereas Theorem 3.3.4 gives

a complexity bound for LMC of order $\widetilde{O}(\hat{\kappa}^2 d^{4/r-1} q^3/\varepsilon^2)$. We note that our result is the *first* high-accuracy guarantee for this setting (i.e., the complexity depends polylogarithmically on ε). Moreover, even in the low-accuracy regime $\varepsilon \asymp 1$, our complexity of $\widetilde{O}(\hat{\kappa} d^{2/r} q)$ is always better (e.g., in the Poincaré case $r = 1$, our rate is $\widetilde{O}(\hat{\kappa} d^2 q)$ whereas Theorem 3.3.4 yields $\widetilde{O}(\hat{\kappa}^2 d^3 q^3)$), although we note that Theorem 3.3.4 handles the more general weakly smooth case.

Surprisingly, the same strategy of rejection sampling also applies to non-smooth potentials. In [LC22], it was shown that when the above rejection sampling is applied to $\tilde{V}(x) = V(x) + \frac{1}{2h}\|x - y\|^2$ with V being a convex and M-Lipschitz function, if $h \le 1/(16M^2 d)$, the expected number of iterations of the algorithm is bounded above by 2. Moreover, the result is insensitive to the inexactness of the minimizer of $\tilde{V}$ [LC22]. Combining it with Theorem 4.3.2 and Theorem 4.3.4 we establish another corollary:

Corollary 4.3.8. *Suppose $\pi^X \propto \exp(-V)$ where V is convex and M-Lipschitz. Take $h \asymp \frac{1}{M^2 d}$ and implement the RGO with rejection sampling as described above.*

1. *Applying Theorem 4.3.2, we deduce that the proximal sampler outputs ρ_k^X with $\mathsf{KL}(\rho_k^X \,\|\, \pi^X) \le \varepsilon^2$ and the expected number of calls to an oracle for V is $O(M^2 d\, W_2^2(\rho_0^X, \pi^X)/\varepsilon^2)$.*

2. *Applying Theorem 4.3.4 (using the fact that log-concave measures satisfy $1/\alpha$-PI for some $\alpha > 0$), we deduce that the proximal sampler outputs ρ_k^X with $\mathcal{R}_q(\rho_k^X \,\|\, \pi^X) \le \varepsilon^2$ and the expected number of calls to an oracle for V is $O(\frac{M^2 dq}{\alpha}(\mathcal{R}_q(\rho_0^X \,\|\, \pi^X) \vee \log(1/\varepsilon)))$.*

We make the same simplifications as above to compare the rates. Our complexity (from the second part of Corollary 4.3.8) is $O(M^2 \mathfrak{m}_2 (d \vee \log(1/\varepsilon)))$, whereas [DMM19] achieves $O(M^2 \mathfrak{m}_2/\varepsilon^4)$ in KL divergence and [LC22] achieves $\widetilde{O}(M^2 d \mathfrak{m}_2/\varepsilon)$ in total variation distance. In particular, when $\mathfrak{m}_2 = O(d)$, our result is substantially better.

We summarize the ways in which the proximal sampler improves upon the standard discretized Langevin algorithm.

1. Under weaker assumptions on the target π^X, such as a Poincaré inequality, the analysis of the Langevin algorithm is affected in two ways: first, the continuous-time convergence of the diffusion is slower; and second, the discretization analysis becomes much more challenging. In contrast, although the ideal proximal sampler also converges more slowly under weaker assumptions, the second issue is no longer present. In particular, regardless of the isoperimetric assumption on π^X, as soon as ∇V is Lipschitz we can implement the RGO via rejection sampling, yielding a simple analysis with strong convergence guarantees.

2. Related to the first point, it is currently not known how to perform a discretization analysis of the Langevin algorithm with linear dependence on the condition number $\hat{\kappa} = \frac{\beta}{\alpha}$ under $1/\alpha$-LSI or $1/\alpha$-PI. Our results therefore constitute the first $O(\hat{\kappa})$ guarantees for such distributions.

3. When implemented via rejection sampling, the proximal sampler provides a new approach to obtaining high-accuracy guarantees for sampling (i.e., complexity guarantees with dependence polylog$(1/\varepsilon)$ on the accuracy ε). The simplicity of the analysis makes it an attractive alternative to Metropolis–Hastings algorithms, whose analysis is often involved.

4. Finally, we mention that when the RGO is implemented via the Metropolized random walk [Dwi+19], the resulting algorithm only uses *zeroth-order* queries to f, which is crucial for certain applications (e.g., Bayesian inverse problems).

4.3.3 On the relation between the proximal sampler and the proximal point algorithm

The proximal sampler is motivated by the proximal point method in optimization. Recall that in optimization, the proximal point method for minimizing f is the iteration of the proximal mapping

$$\mathrm{prox}_{hf}(y) \coloneqq \arg\min_{x \in \mathbb{R}^d} \left\{ f(x) + \frac{1}{2h} \|x - y\|^2 \right\} \tag{4.9}$$

with some step size $h > 0$. Formally, using the analogy between optimization and sampling (in optimization, wish to minimize f; in sampling we wish to sample from $\exp(-V)$); the RGO can be viewed as the sampling analogue of the proximal mapping in which we sample from a regularized version of the target π.

In this section, we establish a more precise correspondence between the proximal sampler algorithm (for sampling from $\exp(-V)$) and the proximal point method (for minimizing f).

4.3.3.1 Convergence under LSI/PL

We recall that LSI for $\pi \propto \exp(-V)$ is equivalent to the statement that the relative entropy $\mathsf{KL}(\cdot \parallel \pi)$ satisfies the gradient domination condition (or the Polyak–Łojasiewicz (PL) inequality) in the Wasserstein metric [OV00]. Thus, in the optimization setting, the analogous assumption to LSI is that f satisfies PL.

We recall f satisfies the PL inequality with constant $1/\alpha > 0$ ($1/\alpha$-PL) if for all $x \in \mathbb{R}^d$,

$$\|\nabla f(x)\|^2 \geq 2\alpha \left(f(x) - f^\star \right),$$

where $f^* = \inf f$. The PL inequality allows for mild non-convexity of f, yet still implies exponential convergence of gradient flow or proximal point method for minimizing f; see for example [KNS16].

In light of our convergence guarantee for the proximal sampler under LSI in Theorem 4.3.3, it is natural to ask whether there is an analogous result for the proximal point method under PL. We answer this affirmatively via the following theorem. We note that a less careful proof of the argument gives the suboptimal contraction factor $\frac{1}{1+\alpha h}$; to the best of our knowledge, we are not aware of another reference which obtains the optimal contraction factor under PL [AB09].[2]

Theorem 4.3.9. *Suppose that $f : \mathbb{R}^d \to (-\infty, +\infty]$ is differentiable and satisfies $1/\alpha$-PL and let $x' \in \mathrm{prox}_{hf}(x)$. Also, write $f^\star = \inf f$. Then, it holds that*

$$f(x') - f^\star \leq \frac{1}{(1+\alpha h)^2} \left\{ f(x) - f^\star \right\}.$$

Proof. §4.5.2. □

■ 4.3.3.2 RGO as a proximal operator on Wasserstein space

Consider $y \in \mathbb{R}^d$. Noting that $\pi^{X|Y=y}(\mathrm{d}x) \propto_x \exp(-\frac{1}{2h}\|x-y\|^2)\,\pi^X(\mathrm{d}x)$ and using [AGS08, Remark 9.4.2] we have

$$\mathsf{KL}(\rho^X \parallel \pi^X) = \mathsf{KL}(\rho^X \parallel \pi^{X|Y=y}) - \int \frac{1}{2h}\|x-y\|^2 \,\mathrm{d}\rho^X(x) + C(y)\,,$$

where $C(y)$ is a constant depending only on y. Using $\arg\min \mathsf{KL}(\cdot \parallel \pi^{X|Y=y}) = \pi^{X|Y=y}$, the RGO can be expressed as

$$\begin{aligned} \pi^{X|Y=y} &= \operatorname*{arg\,min}_{\rho^X \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathsf{KL}(\rho^X \parallel \pi^X) + \frac{1}{2h}\int \|x-y\|^2 \,\mathrm{d}\rho^X(x) \right\} \\ &= \operatorname*{arg\,min}_{\rho^X \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathsf{KL}(\rho^X \parallel \pi^X) + \frac{1}{2h} W_2^2(\rho^X, \delta_y) \right\}. \end{aligned} \tag{4.10}$$

Thus, by replacing the Euclidean distance by the Wasserstein distance, $\pi^{X|Y=y} = \mathrm{prox}_{h\,\mathsf{KL}(\cdot\|\pi^X)}(\delta_y)$. We use this fact in §4.4.2 to provide a new proof of the contraction of the proximal sampler under strong log-concavity (Theorem 4.3.1). The proximal operator over the Wasserstein space is also known as the JKO scheme [JKO98], and hence the proximal sampler can be viewed as a method of implementing the proximal discretization of the Wasserstein gradient flow while

[2] The optimality of our bound can be obtained by considering $f(x) = \frac{\alpha}{2}\|x\|^2$.

having access only to a "restricted" proximal operator (we are only able to evaluate the proximal operator on Dirac measures δ_y for $y \in \mathbb{R}^d$).

See [Che+22b] for an alternative interpretation of the proximal sampler as an entropically-regularized JKO scheme.

4.3.4 Example: Gaussian case

Suppose that the target distribution is $\mathsf{normal}(0, \Sigma)$, i.e., $V(x) = \frac{1}{2}\langle x, \Sigma^{-1} x\rangle$. In this case we can compute the iterations of the proximal sampler explicitly.

If we initialize the proximal sampler at

$$\rho_0^X = \mathsf{normal}(m_0, \Sigma_0)\,,$$

then some calculations show that

$$\begin{aligned}\rho_k^Y &= \mathsf{normal}(m_k, \Sigma_k + hI)\,,\\ \rho_{k+1}^X &= \mathsf{normal}(m_{k+1}, \Sigma_{k+1})\,,\end{aligned}$$

where[3]

$$\begin{aligned}m_{k+1} &\coloneqq \Sigma\,(\Sigma + hI)^{-1}\, m_k\,,\\ \Sigma_{k+1} &\coloneqq \Sigma\,(\Sigma + hI)^{-1}\,(\Sigma_k + hI)\,(\Sigma + hI)^{-1}\,\Sigma + h\,\Sigma\,(\Sigma + hI)^{-1}\,.\end{aligned}$$

Specializing to the case where $\Sigma = I$, $h = 1$, and we initialize at $\mathsf{normal}(0, \sigma_0^2 I)$, we obtain

$$|\sigma_k^2 - 1| = \frac{|\sigma_0^2 - 1|}{4^k}\,. \tag{4.11}$$

In particular, the contraction factor $\frac{1}{(1+\alpha h)^2}$ in Theorem 4.3.3 is sharp.

4.4 Proofs for the proximal sampler

4.4.1 Techniques

At a high level, our proofs proceed by considering the change in KL divergence or Rényi divergence when we apply the following two operations to the law ρ_k^X of the iterate and the target π^X: (1) we simultaneously evolve the two measures

[3] We can also notice that $m_{k+1} = \mathrm{prox}_{hV}(m_k)$, i.e., the means of the distributions follow the proximal point algorithm for V. Moreover, $m_k \to 0 = \arg\min V$ which is the mean of the target distribution.

along the heat flow for time h, and then (2) we apply the RGO to the resulting measures.

For the first step, we formulate a remarkably general lemma in §4.4.1.1 which shows that the computation of the time derivative of *any* ϕ-divergence along the simultaneous heat flow is similar (in a precise sense) to the analogous computation when studying the continuous-time Langevin diffusion. It is this property that allows us to apply functional inequalities which are usually used for the Langevin diffusion, such as the Poincaré and log-Sobolev inequalities, in order to study the convergence of the proximal sampler.

In the second step, we are applying the same operation (of sampling from the RGO) to each measure, so the data-processing inequality implies that the KL divergence or Rényi divergence can only decrease. Combined with the previous step, it is sufficient to prove a convergence guarantee for the proximal sampler; however, the rate turns out to be suboptimal. In order to recover the optimal rate, we introduce an argument based on the *Doob h-transform* (described in §4.4.1.2) to obtain contraction in the second step as well, using the backward version of our general lemma (see §4.4.1.3). We summarize our technique in §4.4.1.4.

■ 4.4.1.1 Lemma on the simultaneous heat flow

Let Φ_π be a ϕ-divergence for some convex function ϕ, i.e.

$$\Phi_\pi(\rho) \coloneqq \mathbb{E}_\pi\big[\phi\big(\frac{\rho}{\pi}\big)\big]\,.$$

We assume that ϕ is regular enough to justify the interchange of differentiation and integration and to perform integration by parts; this is satisfied for all of our applications.

We will use the following result in each forward step of the proximal sampler. This is a generalization of [VW19, Lemma 16].

Lemma 4.4.1. *Let $(\mu_t^X)_{t\ge0}$ be the law of the continuous-time Langevin diffusion with target distribution π^X, and define the* dissipation functional *D_{π^X} via the time derivative of Φ_{π^X} along the diffusion:*

$$D_{\pi^X}(\mu_t^X) \coloneqq -\partial_t \Phi_{\pi^X}(\mu_t^X) = \mathbb{E}_{\mu_t^X}\Big\langle \nabla\big(\phi' \circ \frac{\mu_t^X}{\pi^X}\big), \nabla \log \frac{\mu_t^X}{\pi^X}\Big\rangle\,.$$

If $(\rho^X Q_t)_{t\ge0}$ and $(\pi^X Q_t)_{t\ge0}$ evolve according to the simultaneous heat flow,

$$\partial_t \rho^X Q_t = \frac{1}{2}\,\Delta(\rho^X Q_t)\,, \qquad \partial_t \pi^X Q_t = \frac{1}{2}\,\Delta(\pi^X Q_t)\,,$$

then

$$\partial_t \Phi_{\pi^X Q_t}(\rho^X Q_t) = -\frac{1}{2}\,D_{\pi^X Q_t}(\rho^X Q_t)\,.$$

Proof. On one hand, we know that $(\mu_t^X)_{t\geq 0}$ satisfies the Fokker–Planck equation

$$\partial_t \mu_t^X = \operatorname{div}\big(\mu_t^X \nabla \ln \frac{\mu_t^X}{\pi^X}\big)$$

so that

$$\begin{aligned}\partial_t \Phi_{\pi^X}(\mu_t^X) &= \int \phi'\big(\frac{\mu_t^X}{\pi^X}\big)\, \partial_t \mu_t^X = \int \phi'\big(\frac{\mu_t^X}{\pi^X}\big) \operatorname{div}\big(\mu_t^X \nabla \ln \frac{\mu_t^X}{\pi^X}\big) \\ &= -\int \Big\langle \nabla\big[\phi'\big(\frac{\mu_t^X}{\pi^X}\big)\big], \nabla \ln \frac{\mu_t^X}{\pi^X} \Big\rangle \mu_t^X \,.\end{aligned}$$

On the other hand, writing $\rho_t^X \coloneqq \rho^X Q_t$ and $\pi_t^X \coloneqq \pi^X Q_t$ for brevity, along the simultaneous heat flow we compute

$$\begin{aligned}2\,\partial_t \Phi_{\pi_t^X}(\rho_t^X) &= 2\int \phi'\big(\frac{\rho_t^X}{\pi_t^X}\big)\Big(\partial_t \rho_t^X - \frac{\rho_t^X}{\pi_t^X}\,\partial_t \pi_t^X\Big) + 2\int \phi\big(\frac{\rho_t^X}{\pi_t^X}\big)\,\partial_t \pi_t^X \\ &= \int \phi'\big(\frac{\rho_t^X}{\pi_t^X}\big)\Big(\operatorname{div}(\rho_t^X \nabla \ln \rho_t^X) - \frac{\rho_t^X}{\pi_t^X}\operatorname{div}(\pi_t^X \nabla \ln \pi_t^X)\Big) \\ &\qquad + \int \phi\big(\frac{\rho_t^X}{\pi_t^X}\big)\operatorname{div}(\pi_t^X \nabla \ln \pi_t^X) \\ &= -\int \Big\langle \nabla\big[\phi'\big(\frac{\rho_t^X}{\pi_t^X}\big)\big], \nabla \ln \rho_t^X \Big\rangle \rho_t^X \\ &\qquad + \int \Big\langle \nabla\big[\phi'\big(\frac{\rho_t^X}{\pi_t^X}\big)\,\frac{\rho_t^X}{\pi_t^X}\big], \nabla \ln \pi_t^X \Big\rangle \pi_t^X \\ &\qquad - \int \Big\langle \nabla\big[\phi\big(\frac{\rho_t^X}{\pi_t^X}\big)\big], \nabla \ln \pi_t^X \Big\rangle \pi_t^X \\ &= -\int \Big\langle \nabla\big[\phi'\big(\frac{\rho_t^X}{\pi_t^X}\big)\big], \nabla \ln \frac{\rho_t^X}{\pi_t^X} \Big\rangle \rho_t^X + \int \Big\langle \nabla \frac{\rho_t^X}{\pi_t^X}, \nabla \ln \pi_t^X \Big\rangle \phi'\big(\frac{\rho_t^X}{\pi_t^X}\big)\,\pi_t^X \\ &\qquad - \int \Big\langle \nabla \frac{\rho_t^X}{\pi_t^X}, \nabla \ln \pi_t^X \Big\rangle \phi'\big(\frac{\rho_t^X}{\pi_t^X}\big)\,\pi_t^X \\ &= -D_{\pi_t^X}(\rho_t^X)\,.\end{aligned}$$

□

Remark 4.4.2. *A similar statement holds if we replace the ϕ-divergence Φ_π with any function $\psi \circ \Phi_\pi$ of the ϕ-divergence. This allows us to cover the Rényi divergence introduced in §4.2.*

■ 4.4.1.2 Doob's h-transform

Doob's h-transform is a useful method to analyze the properties of a diffusion process conditioned on its value at some terminal time point. Consider a general

diffusion process modeled by the stochastic differential equation (SDE)

$$\mathrm{d}Z_t = b(t, Z_t)\,\mathrm{d}t + \sigma(t, Z_t)\,\mathrm{d}B_t\,, \qquad Z_0 \sim \mu_0\,, \tag{4.12}$$

where $(B_t)_{t\geq 0}$ denotes a standard Wiener process. Assume that $b(t,z)$ and $\sigma(t,z)$ are piecewise continuous with respect to t and Lipschitz continuous with respect to z so that the above SDE (4.12) has a unique solution. The Doob h-transform characterizes the process conditional on its terminal value Z_T, summarized in the following lemma [SS19].

Lemma 4.4.3. *Let $(\hat{Z}_t)_{0\leq t\leq T}$ be the process (4.12) conditioned to satisfy $Z_T = z$. Then, the process satisfies the following SDE backwards in time:*

$$\mathrm{d}\hat{Z}_t = [b(t,\hat{Z}_t) - \sigma(t,\hat{Z}_t)\,\sigma(t,\hat{Z}_t)^{\mathsf{T}}\,\nabla \ln \mu_t(\hat{Z}_t)]\,\mathrm{d}t + \sigma(t,\hat{Z}_t)\,\mathrm{d}B_t\,,$$

where μ_t is the marginal distribution of Z_t in (4.12) and the SDE is started with $\hat{Z}_T = z$.

Equivalently, if we define the SDE

$$\begin{aligned}\mathrm{d}\hat{Z}_t^{\leftarrow} = [-b(T-t,\hat{Z}_t^{\leftarrow}) + \sigma(T-t,\hat{Z}_t^{\leftarrow})\,\sigma(T-t,\hat{Z}_t^{\leftarrow})^{\mathsf{T}}\,\nabla \ln \mu_{T-t}(\hat{Z}_t^{\leftarrow})]\,\mathrm{d}t \\ + \sigma(T-t,\hat{Z}_t^{\leftarrow})\,\mathrm{d}B_t\,,\end{aligned} \tag{4.13}$$

started at $\hat{Z}_0^{\leftarrow} = z$, then at time T the law of $\hat{Z}_T^{\leftarrow}$ is the conditional distribution of Z_0 given $Z_T = z$.

■ 4.4.1.3 Lemma on the simultaneous backward heat flow

We present the following backward version of Lemma 4.4.1, which we use in each backward step of the proximal sampler. We assume the same set up as in Lemma 4.4.1. Namely, let $\Phi_\pi(\rho) = \mathbb{E}_\pi[\phi(\frac{\rho}{\pi})]$ be a ϕ-divergence for some convex function ϕ, i.e.,

$$\Phi_\pi(\rho) \coloneqq \mathbb{E}_\pi\big[\phi\big(\frac{\rho}{\pi}\big)\big]$$

and let

$$D_\pi(\rho) = \mathbb{E}_\rho\Big\langle \nabla\big(\phi' \circ \frac{\rho}{\pi}\big), \nabla \ln \frac{\rho}{\pi}\Big\rangle$$

so that D_π is the dissipation of Φ_π along the Langevin dynamics with target π.

Lemma 4.4.4. *Let π^X be a probability distribution and let $\boldsymbol{\pi}$ be a joint density for (X,Y) with Y obtained from X by running the heat flow for time h. Let $\pi^{X|Y}$ be*

the conditional distribution of X given Y under π, and let π^Y denote the marginal distribution of Y. Then, for each $t \in [0,h]$, there exists a channel $Q_t^{\leftarrow}$ that maps probability measures to probability measures, with the following properties: (1) $Q_0^{\leftarrow}$ is the identity channel; (2) $Q_h^{\leftarrow}$ maps a probability measure ρ^Y to the the measure $\rho^Y Q_h^{\leftarrow}(x) = \int \pi^{X|Y}(x \mid y)\, \rho^Y(\mathrm{d}y)$; (3) for every t, $\pi^Y Q_t^{\leftarrow} = \pi * \mathsf{normal}(0, (h-t)I)$; and (4) for every ρ^Y,

$$\partial_t \Phi_{\pi^Y Q_t^{\leftarrow}}(\rho^Y Q_t^{\leftarrow}) = -\frac{1}{2} D_{\pi^Y Q_t^{\leftarrow}}(\rho^Y Q_t^{\leftarrow}) \,.$$

The channel is obtained from the Doob h-transform. To give intuition for the construction, consider the process $\mathrm{d}Z_t = \mathrm{d}B_t$ started at $Z_0 \sim \pi^X$, i.e., Brownian motion initialized from π^X. Then, the joint target distribution $\boldsymbol{\pi}$ of the proximal sampler can be expressed as $\boldsymbol{\pi} = \mathrm{law}(Z_0, Z_h)$, and consequently we have $\pi^{X|Y=y} = \mathrm{law}(Z_0 \mid Z_h = y)$. If we define the time reversal $Z_t^{\leftarrow} = Z_{h-t}$, then we can also express this as $\pi^{X|Y=y} = \mathrm{law}(Z_h^{\leftarrow} \mid Z_0^{\leftarrow} = y)$; moreover, the reversed process $(Z_t^{\leftarrow})_{t\in[0,h]}$ satisfies the SDE given in Lemma 4.4.3. Hence, we can take $\mu Q_t^{\leftarrow} \coloneqq \mathrm{law}(Z_h^{\leftarrow} \mid Z_0^{\leftarrow} \sim \mu)$ and use calculus in order to prove the result.

Proof. Let $\pi_t \coloneqq \pi^X * \mathsf{normal}(0, tI)$. We define $Q_t^{\leftarrow}$ as follows: given ρ^Y, we set $\rho^Y Q_t^{\leftarrow}$ to be the law at time t of the SDE

$$\mathrm{d}\hat{Z}_t^{\leftarrow} = \nabla \ln \pi_{h-t}(\hat{Z}_t^{\leftarrow})\, \mathrm{d}t + \mathrm{d}B_t \,, \tag{4.14}$$

started at $\hat{Z}_0^{\leftarrow} \sim \rho^Y$. According to Lemma 4.4.3 applied to the Brownian motion process (started at π^X), the channels $(Q_t^{\leftarrow})_{0\le t\le h}$ satisfy properties (1), (2), and (3). It remains to verify (4). In the proof, we write $\pi_t^{\leftarrow} \coloneqq \pi^Y Q_t^{\leftarrow}$ and $\rho_t^{\leftarrow} \coloneqq \rho^Y Q_t^{\leftarrow}$ for brevity. Note that $\pi_{h-t} = \pi_t^{\leftarrow}$ by construction, and we have the Fokker–Planck equations:

$$\begin{aligned}
\partial_t \pi_t^{\leftarrow} &= -\operatorname{div}(\pi_t^{\leftarrow} \nabla \ln \pi_t^{\leftarrow}) + \frac{1}{2}\Delta \pi_t^{\leftarrow} = -\frac{1}{2}\Delta \pi_t^{\leftarrow} \,, \\
\partial_t \rho_t^{\leftarrow} &= -\operatorname{div}(\rho_t^{\leftarrow} \nabla \ln \pi_t^{\leftarrow}) + \frac{1}{2}\Delta \rho_t^{\leftarrow} = \operatorname{div}\big(\rho_t^{\leftarrow} \nabla \ln \frac{\rho_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big) - \frac{1}{2}\Delta \rho_t^{\leftarrow} \,.
\end{aligned}$$

Hence,

$$\begin{aligned}
2\,\partial_t \Phi_{\pi_t^{\leftarrow}}(\rho_t^{\leftarrow}) &= 2\int \phi'\big(\frac{\rho_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big)\Big(\partial_t \rho_t^{\leftarrow} - \frac{\rho_t^{\leftarrow}}{\pi_t^{\leftarrow}}\,\partial_t \pi_t^{\leftarrow}\Big) + 2\int \phi\big(\frac{\rho_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big)\,\partial_t \pi_t^{\leftarrow} \\
&= \int \phi'\big(\frac{\rho_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big)\Big(2\operatorname{div}\big(\rho_t^{\leftarrow}\nabla \ln \frac{\rho_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big) - \Delta \rho_t^{\leftarrow} + \frac{\rho_t^{\leftarrow}}{\pi_t^{\leftarrow}}\Delta \pi_t^{\leftarrow}\Big) \\
&\quad - \int \phi\big(\frac{\rho_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big)\Delta \pi_t^{\leftarrow}
\end{aligned}$$

$$
\begin{aligned}
&= 2 \int \phi'\big(\frac{\rho_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big)\, \mathrm{div}\big(\rho_t^{\leftarrow} \nabla \ln \frac{\rho_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big) \\
&\quad \underbrace{- \int \phi'\big(\frac{\rho_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big) \Big(\Delta \rho_t^{\leftarrow} - \frac{\rho_t^{\leftarrow}}{\pi_t^{\leftarrow}} \Delta \pi_t^{\leftarrow}\Big) + \int \phi\big(\frac{\rho_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big)\, \Delta \pi_t^{\leftarrow}}_{= -D_{\pi_t^{\leftarrow}}(\rho_t^{\leftarrow}) \text{ by Lemma 4.4.1}} \\
&= -2 \int \Big\langle \nabla\big[\phi'\big(\frac{\rho_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big)\big], \nabla \ln \frac{\rho_t^{\leftarrow}}{\pi_t^{\leftarrow}} \Big\rangle \rho_t^{\leftarrow} + D_{\pi_t^{\leftarrow}}(\rho_t^{\leftarrow}) \\
&= -2 D_{\pi_t^{\leftarrow}}(\rho_t^{\leftarrow}) + D_{\pi_t^{\leftarrow}}(\rho_t^{\leftarrow}) = -D_{\pi_t^{\leftarrow}}(\rho_t^{\leftarrow})\,. \qquad \square
\end{aligned}
$$

■ 4.4.1.4 General strategy of the proofs

Suppose that we want to understand the change in the ϕ-divergence $\Phi_{\pi^X}(\rho_1^X)$ after one iteration of the proximal sampler, compared to the ϕ divergence $\Phi_{\pi^X}(\rho_0^X)$ at initialization. We split the analysis into two steps.

1. **Forward step:** In the first step, we draw $Y_0 \mid X_0 \sim \mathsf{normal}(X_0, hI)$.

   This creates a joint distribution $\boldsymbol{\rho}_0$ with the correct conditionals: $\rho_0^{Y|X} = \pi^{Y|X}$. Therefore, the ϕ-divergence of the joint distribution is equal to the initial ϕ-divergence of the X-marginal: $\Phi_{\boldsymbol{\pi}}(\boldsymbol{\rho}_0) = \Phi_{\pi^X}(\rho_0^X)$.

   Consider the Y-marginal $Y_0 \sim \rho_0^Y$. Observe that $\rho_0^Y = \rho_0^X * \mathsf{normal}(0, hI)$ is the output $\rho_0^Y = \tilde{\rho}_h$ of the heat flow $\partial_t \tilde{\rho}_t = \frac{1}{2}\Delta \tilde{\rho}_t$ at time $t = h$ starting from $\tilde{\rho}_0 = \rho_0^X$. We denote this by $\rho_0^Y = \rho_0^X Q_h$, where $(Q_t)_{t \geq 0}$ denotes the heat semigroup. Similarly, we write the Y-marginal of the target as $\pi^Y = \pi^X * \mathsf{normal}(0, hI) = \pi^X Q_h$.

   In particular, $(\rho_0^X Q_t)_{t \geq 0}$ and $(\pi^X Q_t)_{t \geq 0}$ evolve following the simultaneous heat flow.

   By Lemma 4.4.1, along the simultaneous heat flow,

   $$\partial_t \Phi_{\pi^X Q_t}(\rho_0^X Q_t) = -\frac{1}{2} D_{\pi^X Q_t}(\rho_0^X Q_t)$$

   where $D_\cdot(\cdot)$ denotes the dissipation functional for the ϕ-divergence along the Langevin dynamics. Hence, a lower bound on $D_{\pi^X Q_t}(\rho_0^X Q_t)$ leads to an upper bound on

   $$\Phi_{\pi^Y}(\rho_0^Y) - \Phi_{\pi^X}(\rho_0^X) = \Phi_{\pi^X Q_h}(\rho_0^X Q_h) - \Phi_{\pi^X}(\rho_0^X)\,.$$

2. **Backward step:** In the second step, we draw $X_1 \mid Y_0 \sim \pi^{X|Y=Y_0}$.

This time, we consider the backward heat flow and apply Lemma 4.4.4, which yields the Doob channels $(Q_t^{\leftarrow})_{0\le t\le h}$ with $\rho_1^X = \rho_0^Y Q_h^{\leftarrow}$ and $\pi^X = \pi^Y Q_h^{\leftarrow}$. Lemma 4.4.4 implies that

$$\partial_t \Phi_{\pi^Y Q_t^{\leftarrow}}(\rho_0^Y Q_t^{\leftarrow}) = -\frac{1}{2} D_{\pi^Y Q_t^{\leftarrow}}(\rho_0^Y Q_t^{\leftarrow}) .$$

Observe that this is almost symmetric with the forward step! In particular, a lower bound on $D_{\pi^Y Q_t^{\leftarrow}}(\rho_0^Y Q_t^{\leftarrow})$ leads to an upper bound on

$$\Phi_{\pi^X}(\rho_1^X) - \Phi_{\pi^Y}(\rho_0^Y) = \Phi_{\pi^Y Q_h^{\leftarrow}}(\rho_0^Y Q_h^{\leftarrow}) - \Phi_{\pi^Y}(\rho_0^Y) .$$

Combining the two steps allows us to understand each iteration of the proximal sampler algorithm.

■ 4.4.2 Convergence under strong log-concavity

Suppose that A is a set-valued mapping on $\mathbb{R}^d$ which is strongly monotone, in the sense that

$$\langle A(x) - A(y), x - y\rangle \ge \alpha \|x - y\|^2 \qquad \text{for all } x, y \in \mathbb{R}^d .$$

Suppose that $x' \in x - hA(x')$ and $y' \in y - hA(y')$. Then, by expanding out the square, one can easily show that $\|x' - y'\|^2 \le \frac{1}{(1+\alpha h)^2}\|x - y\|^2$. In particular, by applying this to the subdifferential $A = \partial f$, where f is α-strongly convex, one immediately obtains the fact that the proximal point algorithm is a $\frac{1}{1+\alpha h}$-contraction. In this section, we translate this proof to the sampling setting.

Recall from (4.10) that $\pi^{X|Y=y} = \mathrm{prox}_{hF}(\delta_y)$, where $F = \mathsf{KL}(\cdot \parallel \pi^X)$ is α-geodesically strongly strongly convex [AGS08, Equation 10.1.8]. Then, from the first-order optimality conditions on Wasserstein space [see AGS08, Lemma 10.1.2], we have

$$0 \in \partial F(\pi^{X|Y=y}) + \frac{1}{h}(\mathrm{id} - y) , \qquad \pi^{X|Y=y}\text{-a.s.}, \tag{4.15}$$

where ∂F denotes the Wasserstein subdifferential of F.

Proof of Theorem 4.3.1. First, let $y, \bar{y} \in \mathbb{R}^d$. Then, from (4.15):

$$\mathrm{id} \in y - h\,\partial F(\pi^{X|Y=y}) , \qquad \pi^{X|Y=y}\text{-a.s.} \tag{4.16}$$

$$\mathrm{id} \in \bar{y} - h\,\partial F(\pi^{X|Y=\bar{y}}) , \qquad \pi^{X|Y=\bar{y}}\text{-a.s.} \tag{4.17}$$

Let T be the optimal transport map from $\pi^{X|Y=y}$ to $\pi^{X|Y=\bar{y}}$. We rewrite (4.17) as

$$T \in \bar{y} - h\,\partial F(\pi^{X|Y=\bar{y}}) \circ T , \qquad \pi^{X|Y=y}\text{-a.s.} \tag{4.18}$$

We now abuse notation and write $\partial F(\pi^{X|Y=y})$ for an element of the subdifferential. Then, using (4.16) and (4.18), $\pi^{X|Y=y}$-a.s.,

$$\begin{aligned}\|T - \mathrm{id}\|^2 = \|\bar{y} - y\|^2 - 2h \left\langle \partial F(\pi^{X|Y=\bar{y}}) \circ T - \partial F(\pi^{X|Y=y}), T - \mathrm{id} \right\rangle \\ - h^2 \left\| \partial F(\pi^{X|Y=\bar{y}}) \circ T - \partial F(\pi^{X|Y=y}) \right\|^2 .\end{aligned}$$

Integrating with respect to $\pi^{X|Y=y}$, and using the geodesic strong convexity of F [AGS08, Equation 10.1.8],

$$\begin{aligned}&W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}) \\ &\quad \le \|y - \bar{y}\|^2 - 2\alpha h \, W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}) - \alpha^2 h^2 \, W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}) .\end{aligned}$$

Therefore,

$$W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}) \le \frac{1}{(1+\alpha h)^2} \|y - \bar{y}\|^2 .$$

The rest of the argument is concluded as in [LST21b, Lemma 2]. We provide the details here for completeness. First, along the proximal sampler, we have $W_2(\rho_0^Y, \bar{\rho}_0^Y) \le W_2(\rho_0^X, \bar{\rho}_0^X)$ because the heat flow is a Wasserstein contraction (see §4.4.1.4 for the notation). Next, let γ denote an optimal coupling of ρ_0^Y and $\bar{\rho}_0^Y$, and for all $y, y \in \mathbb{R}^d$ let $\gamma_{y,\bar{y}}$ denote an optimal coupling of $\pi^{X|Y=y}$ and $\pi^{X|Y=\bar{y}}$. We check that the measure $\hat{\gamma}(\mathrm{d}x, \mathrm{d}\bar{x}) := \gamma(\mathrm{d}y, \mathrm{d}\bar{y}) \, \gamma_{y,\bar{y}}(\mathrm{d}x, \mathrm{d}\bar{x})$ is a valid coupling of ρ_1^X and $\bar{\rho}_1^X$. To check that, for instance, the first marginal of $\hat{\gamma}$ is ρ_1^X, we take a bounded measurable function $\psi : \mathbb{R}^d \to \mathbb{R}$ and calculate

$$\begin{aligned}\int \psi(x) \, \hat{\gamma}(\mathrm{d}x, \mathrm{d}\bar{x}) &= \iint \psi(x) \, \gamma(\mathrm{d}y, \mathrm{d}\bar{y}) \, \gamma_{y,\bar{y}}(\mathrm{d}x, \mathrm{d}\bar{x}) \\ &= \iint \psi(x) \, \gamma(\mathrm{d}y, \mathrm{d}\bar{y}) \, \pi^{X|Y=y}(\mathrm{d}x) \\ &= \iint \psi(x) \, \rho_0^Y(\mathrm{d}y) \, \pi^{X|Y=y}(\mathrm{d}x) = \int \psi(x) \, \rho_1^X(\mathrm{d}x) ,\end{aligned}$$

and similarly the second marginal of $\hat{\gamma}$ is $\bar{\rho}_1^X$. Therefore,

$$\begin{aligned}W_2^2(\rho_1^X, \bar{\rho}_1^X) &\le \int \|x - \bar{x}\|^2 \, \hat{\gamma}(\mathrm{d}x, \mathrm{d}\bar{x}) = \iint \|x - \bar{x}\|^2 \, \gamma(\mathrm{d}y, \mathrm{d}\bar{y}) \, \gamma_{y,\bar{y}}(\mathrm{d}x, \mathrm{d}\bar{x}) \\ &= \int W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}) \, \gamma(\mathrm{d}y, \mathrm{d}\bar{y}) \\ &\le \frac{1}{(1+\alpha h)^2} \int \|y - \bar{y}\|^2 \, \gamma(\mathrm{d}y, \mathrm{d}\bar{y}) = \frac{1}{(1+\alpha h)^2} \, W_2^2(\rho_0^Y, \bar{\rho}_0^Y) ,\end{aligned}$$

which completes the proof. □

4.4.3 Convergence under log-concavity

For a probability distribution ρ with smooth relative density $\frac{\rho}{\pi}$, the *Fisher information* of ρ with respect to π is

$$\mathsf{FI}(\rho \parallel \pi) := \int \rho \left\| \nabla \log \frac{\rho}{\pi} \right\|^2 = \mathbb{E}_\pi \left[\frac{\pi}{\rho} \left\| \nabla \frac{\rho}{\pi} \right\|^2 \right]. \tag{4.19}$$

Recall that Fisher information is the dissipation of KL divergence along the Langevin dynamics.

Proof of Theorem 4.3.2. We follow the strategy and notation of §4.4.1.4.

1. **Forward step:** By log-concavity of $\pi^X Q_t$ (since log-concavity is preserved by convolution [SW14]), the convexity of $\mathsf{KL}(\cdot \parallel \pi^X Q_t)$ along Wasserstein geodesics [AGS08, Theorem 9.4.11] yields the inequality

$$\begin{aligned} 0 &= \mathsf{KL}(\pi^X Q_t \parallel \pi^X Q_t) \\ &\geq \mathsf{KL}(\rho_0^X Q_t \parallel \pi^X Q_t) + \mathbb{E}_{(X_t, Y_t) \sim \mathsf{OPT}(\rho_0^X Q_t, \pi^X Q_t)} \left\langle \nabla \log \frac{\rho_0^X Q_t}{\pi^X Q_t}(X_t), Y_t - X_t \right\rangle \end{aligned}$$

where $\mathsf{OPT}(\cdot, \cdot)$ is used to denote the optimal transport plan. Hence,

$$\underbrace{\mathbb{E}_{\rho_0^X Q_t}\left[\left\| \nabla \log \frac{\rho_0^X Q_t}{\pi^X Q_t} \right\|^2 \right]}_{= \mathsf{FI}(\rho_0^X Q_t \parallel \pi^X Q_t)} W_2^2(\rho_0^X Q_t, \pi^X Q_t) \geq \mathsf{KL}(\rho_0^X Q_t \parallel \pi^X Q_t)^2. \tag{4.20}$$

So, by Lemma 4.4.1 and (4.20),

$$\partial_t \, \mathsf{KL}(\rho_0^X Q_t \parallel \pi^X Q_t) = -\frac{1}{2} \mathsf{FI}(\rho_0^X Q_t \parallel \pi^X Q_t) \leq -\frac{1}{2} \frac{\mathsf{KL}(\rho_0^X Q_t \parallel \pi^X Q_t)^2}{W_2^2(\rho_0^X Q_t, \pi^X Q_t)}.$$

Also, observe that $t \mapsto W_2^2(\rho_0^X Q_t, \pi^X Q_t)$ is decreasing because the heat flow is a W_2 contraction (which can be proven directly quite easily). Solving this differential inequality yields

$$\frac{1}{\mathsf{KL}(\rho_0^Y \parallel \pi^Y)} = \frac{1}{\mathsf{KL}(\rho_0^X Q_h \parallel \pi^X Q_h)} \geq \frac{1}{\mathsf{KL}(\rho_0^X \parallel \pi^X)} + \frac{h}{2 W_2^2(\rho_0^X, \pi^X)}.$$

2. **Backward step:** By Lemma 4.4.4 and (4.20),

$$\partial_t \, \mathsf{KL}(\rho_0^Y Q_t^{\leftarrow} \parallel \pi^Y Q_t^{\leftarrow}) = -\frac{1}{2} \mathsf{FI}(\rho_0^Y Q_t^{\leftarrow} \parallel \pi^Y Q_t^{\leftarrow}) \leq -\frac{1}{2} \frac{\mathsf{KL}(\rho_0^Y Q_t^{\leftarrow} \parallel \pi^Y Q_t^{\leftarrow})^2}{W_2^2(\rho_0^Y Q_t^{\leftarrow}, \pi^Y Q_t^{\leftarrow})}.$$

By (4.13), the channels $(Q_t^{\leftarrow})_{t\geq 0}$ can be modeled by the diffusion

$$\mathrm{d}Z_t = \nabla \ln \pi_{h-t}(Z_t)\,\mathrm{d}t + \mathrm{d}B_t.$$

Since $\ln \pi_{h-t}$ is concave, with a standard coupling argument, one can show that $t \mapsto W_2(\rho_0^Y Q_t^{\leftarrow}, \pi^Y Q_t^{\leftarrow})$ is decreasing. Hence,

$$W_2(\rho_0^Y Q_t^{\leftarrow}, \pi^Y Q_t^{\leftarrow}) \leq W_2(\rho_0^Y Q_0^{\leftarrow}, \pi^Y Q_0^{\leftarrow}) = W_2(\rho_0^Y, \pi^Y) \leq W_2(\rho_0^X, \pi^X)\,.$$

Therefore, we deduce that

$$\frac{1}{\mathsf{KL}(\rho_1^X \parallel \pi^X)} = \frac{1}{\mathsf{KL}(\rho_0^Y Q_h^{\leftarrow} \parallel \pi^Y Q_h^{\leftarrow})} \geq \frac{1}{\mathsf{KL}(\rho_0^Y \parallel \pi^Y)} + \frac{h}{2W_2^2(\rho_0^X, \pi^X)}\,.$$

Finally, we iterate this inequality and recall that $W_2^2(\rho_k^X, \pi^X) \leq W_2^2(\rho_0^X, \pi^X)$ for all $k \in \mathbb{N}$ (see Theorem 4.3.1 for $\alpha = 0$). It quickly yields

$$\frac{1}{\mathsf{KL}(\rho_k^X \parallel \pi^X)} \geq \frac{1}{\mathsf{KL}(\rho_0^X \parallel \pi^X)} + \frac{kh}{W_2^2(\rho_0^X, \pi^X)}$$

or

$$\mathsf{KL}(\rho_k^X \parallel \pi^X) \leq \frac{\mathsf{KL}(\rho_0^X \parallel \pi^X)}{1 + kh\,\mathsf{KL}(\rho_0^X \parallel \pi^X)/W_2^2(\rho_0^X, \pi^X)} \leq \frac{W_2^2(\rho_0^X, \pi^X)}{kh}\,. \qquad \square$$

The above proof can be compared to the $O(1/t)$ convergence of the objective gap for the gradient flow $t \mapsto x_t$ of a convex function $f : \mathbb{R}^d \to \mathbb{R}$, which follows from differentiating the Lyapunov function $t \mapsto 2t\,\{f(x_t) - f(x^\star)\} + \|x_t - x^\star\|^2$, where $x^\star = \arg\min f$.

4.4.4 Convergence under LSI

We recall the following definitions. For a probability distribution ρ with smooth relative density $\frac{\rho}{\pi}$, the *Rényi information* of ρ with respect to π of order $q \geq 1$ is

$$\mathsf{FI}_q(\rho \parallel \pi) \coloneqq q\,\frac{\mathbb{E}_\pi\big[\big(\frac{\pi}{\rho}\big)^{q-2}\,\big\|\nabla \frac{\rho}{\pi}\big\|^2\big]}{\mathbb{E}_\pi\big[\big(\frac{\pi}{\rho}\big)^{q}\big]}\,.$$

Note that $\mathsf{FI}_1 = \mathsf{FI}$, where FI is the Fisher information (4.19). Recall that by definition, π satisfies $1/\alpha$-LSI if for all ρ, $\mathsf{FI}(\rho \parallel \pi) \geq 2\alpha\,\mathsf{KL}(\rho \parallel \pi)$. One can show this also implies for all $q \geq 1$:

$$\mathsf{FI}_q(\rho \parallel \pi) \geq \frac{2\alpha}{q}\,\mathcal{R}_q(\rho \parallel \pi)\,, \tag{4.21}$$

see [VW19, Lemma 5]. Just as Fisher information is the dissipation of KL divergence along the Langevin dynamics, Rényi information is the dissipation of Rényi divergence along the Langevin dynamics.

Proof of Theorem 4.3.3. We will prove the following one-step improvement lemma for Rényi divergence of order $q \geq 1$: For any initial distribution ρ_0^X, after one iteration of the proximal sampler with step size $h > 0$, the resulting distribution ρ_1^X satisfies

$$\mathcal{R}_q(\rho_1^X \parallel \pi^X) \leq \frac{\mathcal{R}_q(\rho_0^X \parallel \pi^X)}{(1+\alpha h)^{2/q}}\,. \tag{4.22}$$

Iterating this lemma for k iterations yields the desired convergence rate in the theorem. The result for KL divergence is the special case $q = 1$.

We follow the strategy and notation of §4.4.1.4.

1. **Forward step:** By Lemma 4.4.1, along the simultaneous heat flow,

$$\partial_t \mathcal{R}_q(\rho_0^X Q_t \parallel \pi^X Q_t) = -\frac{1}{2}\,\mathsf{FI}_q(\rho_0^X Q_t \parallel \pi^X Q_t) \leq -\frac{\alpha_t}{q}\,\mathcal{R}_q(\rho_0^X Q_t \parallel \pi^X Q_t)$$

where by (4.21), the last inequality holds if $\pi^X Q_t$ is $1/\alpha_t$-LSI. Since π^X satisfies $1/\alpha$-LSI by assumption, recall that $\pi^X Q_t = \pi^X * \mathsf{normal}(0, tI)$ satisfies $1/\alpha_t$-LSI with $\alpha_t = (\frac{1}{\alpha} + t)^{-1} = \frac{\alpha}{1+\alpha t}$. Integrating, we get

$$\mathcal{R}_q(\rho_0^X Q_t \parallel \pi^X Q_t) \leq \exp(-A_t)\,\mathcal{R}_q(\rho_0^X \parallel \pi^X)$$

where $A_t = \frac{1}{q}\int_0^t \alpha_s\,\mathrm{d}s = \frac{1}{q}\int_0^t \frac{\alpha}{1+\alpha s}\,\mathrm{d}s = \frac{1}{q}\ln(1+\alpha t)$. Therefore, after the forward step,

$$\mathcal{R}_q(\rho_0^Y \parallel \pi^Y) = \mathcal{R}_q(\rho_0^X Q_h \parallel \pi^X Q_h) \leq \frac{\mathcal{R}_q(\rho_0^X \parallel \pi^X)}{(1+\alpha h)^{1/q}}\,.$$

2. **Backward step:** By Lemma 4.4.4, along the simultaneous backwards heat flow, it holds that

$$\begin{aligned}\partial_t \mathcal{R}_q(\rho_0^Y Q_t^{\leftarrow} \parallel \pi^Y Q_t^{\leftarrow}) &= -\frac{1}{2}\,\mathsf{FI}_q(\rho_0^Y Q_t^{\leftarrow} \parallel \pi^Y Q_t^{\leftarrow}) \\ &\leq -\frac{\alpha_{h-t}}{q}\,\mathcal{R}_q(\rho_0^Y Q_t^{\leftarrow} \parallel \pi^Y Q_t^{\leftarrow})\end{aligned}$$

where the last inequality holds since $\pi^Y Q_t^{\leftarrow} = \pi * \mathsf{normal}(0, (h-t)I)$ is $1/\alpha_{h-t}$-LSI. Therefore, just as in the forward step, integration yields

$$\mathcal{R}_q(\rho_1^X \parallel \pi^X) = \mathcal{R}_q(\rho_0^Y Q_h^{\leftarrow} \parallel \pi^Y Q_h^{\leftarrow}) \leq \frac{\mathcal{R}_q(\rho_0^Y \parallel \pi^Y)}{(1+\alpha h)^{1/q}}\,.$$

Combining the two steps above yields the desired contraction rate in (4.22). □

4.4.5 Convergence under PI

The dissipation of the chi-squared divergence along the Langevin dynamics is

$$\mathsf{FI}_{\chi^2}(\rho \,\|\, \pi) \coloneqq 2\,\mathbb{E}_\pi\left[\left\|\nabla \frac{\rho}{\pi}\right\|^2\right].$$

Proof of Theorem 4.3.4. We follow the strategy and notation of §4.4.1.4.

1. **Forward step:** Along the simultaneous heat flow, Lemma 4.4.1 yields

$$\partial_t \chi^2(\rho_0^X Q_t \,\|\, \pi^X Q_t) = -\frac{1}{2}\,\mathsf{FI}_{\chi^2}(\rho_0^X Q_t \,\|\, \pi^X Q_t)\,,$$
$$\partial_t \mathcal{R}_q(\rho_0^X Q_t \,\|\, \pi^X Q_t) = -\frac{1}{2}\,\mathsf{FI}_q(\rho_0^X Q_t \,\|\, \pi^X Q_t)\,.$$

   Since π^X satisfies $1/\alpha$-PI, then $\pi^X Q_t$ satisfies $1/\alpha_t$-PI with $\alpha_t = \frac{\alpha}{1+\alpha t}$. Applying this yields

$$\partial_t \chi^2(\rho_0^X Q_t \,\|\, \pi^X Q_t) = -\frac{1}{2}\,\mathsf{FI}_{\chi^2}(\rho_0^X Q_t \,\|\, \pi^X Q_t) \le -\alpha_t\,\chi^2(\rho_0^X Q_t \,\|\, \pi^X Q_t)$$

   and therefore

$$\chi^2(\rho_0^Y \,\|\, \pi^Y) = \chi^2(\rho_0^X Q_h \,\|\, \pi^X Q_h) \le \frac{\chi^2(\rho_0^X \,\|\, \pi^X)}{1+\alpha h}$$

   upon integration.

   Next, from [VW19, Lemma 17], $1/\alpha_t$-PI implies

$$\begin{aligned}\partial_t \mathcal{R}_q(\rho_0^X Q_t \,\|\, \pi^X Q_t) &= -\frac{1}{2}\,\mathsf{FI}_q(\rho_0^X Q_t \,\|\, \pi^X Q_t) \\ &\le -\frac{2\alpha_t}{q}\left\{1 - \exp(-\mathcal{R}_q(\rho_0^X Q_t \,\|\, \pi^X Q_t))\right\}.\end{aligned}$$

   We split into two cases. If $\mathcal{R}_q(\rho_0^X \|\pi^X) \ge 1$, then as long as $\mathcal{R}_q(\rho_0^X Q_t \|\pi^X Q_t) \ge 1$ we can use the inequality $1 - \exp(-x) \ge \frac{1}{2}$ for $x \ge 1$, so that

$$\partial_t \mathcal{R}_q(\rho_0^X Q_t \,\|\, \pi^X Q_t) \le -\frac{\alpha_t}{q}\,.$$

   Integrating, we obtain

$$\mathcal{R}_q(\rho_0^Y \,\|\, \pi^Y) = \mathcal{R}_q(\rho_0^X Q_h \,\|\, \pi^X Q_h) \le \left(\mathcal{R}_q(\rho_0^X \,\|\, \pi^X) - \frac{\ln(1+\alpha h)}{q}\right) \vee 1\,.$$

In the second case, if $\mathcal{R}_q(\rho_0^X \,\|\, \pi^X) \le 1$, then we use $1 - \exp(-x) \ge \frac{x}{2}$ for $x \in [0,1]$ to obtain

$$\partial_t \mathcal{R}_q(\rho_0^X Q_t \,\|\, \pi^X Q_t) \le -\frac{\alpha_t}{q}\, \mathcal{R}_q(\rho_0^X Q_t \,\|\, \pi^X Q_t)\,.$$

Integrating,

$$\mathcal{R}_q(\rho_0^Y \,\|\, \pi^Y) = \mathcal{R}_q(\rho_0^X Q_h \,\|\, \pi^X Q_h) \le \frac{\mathcal{R}_q(\rho_0^X \,\|\, \pi^X)}{(1+\alpha h)^{1/q}}\,.$$

2. **Backward step:** Along the simultaneous backward heat flow, Lemma 4.4.4 next yields

$$\partial_t \chi^2(\rho_0^Y Q_t^{\leftarrow} \,\|\, \pi^Y Q_t^{\leftarrow}) = -\frac{1}{2}\, \mathsf{FI}_{\chi^2}(\rho_0^Y Q_t^{\leftarrow} \,\|\, \pi^Y Q_t^{\leftarrow})\,,$$
$$\partial_t \mathcal{R}_q(\rho_0^Y Q_t^{\leftarrow} \,\|\, \pi^Y Q_t^{\leftarrow}) = -\frac{1}{2}\, \mathsf{FI}_q(\rho_0^Y Q_t^{\leftarrow} \,\|\, \pi^Y Q_t^{\leftarrow})\,.$$

Using entirely analogous arguments as in the forward step, we obtain

$$\chi^2(\rho_1^X \,\|\, \pi^X) = \chi^2(\rho_0^Y Q_h^{\leftarrow} \,\|\, \pi^Y Q_h^{\leftarrow}) \le \frac{\chi^2(\rho_0^Y \,\|\, \pi^Y)}{1+\alpha h}$$

for the chi-squared divergence,

$$\mathcal{R}_q(\rho_1^X \,\|\, \pi^X) = \mathcal{R}_q(\rho_0^Y Q_h^{\leftarrow} \,\|\, \pi^Y Q_h^{\leftarrow}) \le \left(\mathcal{R}_q(\rho_0^Y \,\|\, \pi^Y) - \frac{\ln(1+\alpha h)}{q}\right) \vee 1$$

for the Rényi divergence if $\mathcal{R}_q(\rho_0^Y \,\|\, \pi^Y) \ge 1$, and

$$\mathcal{R}_q(\rho_1^X \,\|\, \pi^X) = \mathcal{R}_q(\rho_0^Y Q_h^{\leftarrow} \,\|\, \pi^Y Q_h^{\leftarrow}) \le \frac{\mathcal{R}_q(\rho_0^Y \,\|\, \pi^Y)}{(1+\alpha h)^{1/q}}$$

if $\mathcal{R}_q(\rho_0^Y \,\|\, \pi^Y) \le 1$.

□

■ 4.4.6 Convergence under LOI

Before giving the convergence proof under LOI, we recall the following property of the behavior of LOI under convolution.

Lemma 4.4.5. *Suppose that μ_0 satisfies $(r, 1/\alpha_0)$-LOI and μ_1 satisfies $(r, 1/\alpha_1)$-LOI. Then, $\mu_0 * \mu_1$ satisfies $(r, 1/\alpha_0 + 1/\alpha_1)$-LOI.*

Proof. Let $X_0 \sim \mu_0$ and $X_1 \sim \mu_1$ be independent. Then, we can write

$$\mathrm{var}_{p,\mu_0*\mu_1}(\psi) = \mathbb{E}[\Phi(\psi^p(X_0+X_1))] - \Phi(\mathbb{E}[\psi^p(X_0+X_1)])$$

where $\Phi(x) \coloneqq x^{2/p}$. One can then deduce the conclusion of the lemma easily from the subadditivity of the Φ-entropy [BLM13, Theorem 14.1]. □

Proof of Theorem 4.3.5. We follow the strategy and notation of §4.4.1.4.

1. **Forward step:** Along the simultaneous heat flow, Lemma 4.4.1 yields

$$\partial_t \mathcal{R}_q(\rho_0^X Q_t \parallel \pi^X Q_t) = -\frac{1}{2}\,\mathsf{FI}_q(\rho_0^X Q_t \parallel \pi^X Q_t)\,.$$

Since π^X satisfies $(r, 1/\alpha)$-LOI and $\mathcal{N}(0, tI)$ satisfies (r', t)-LOI for any $r' \in [1,2]$ [see LO00, Corollary 1], then by Lemma 4.4.5, $\pi^X Q_t$ satisfies $(r, 1/\alpha_t)$-LOI with $\alpha_t = \frac{\alpha}{1+\alpha t}$.

Next, from Theorem 3.2.2, $(r, 1/\alpha_t)$-LOI implies

$$\begin{aligned}\partial_t \mathcal{R}_q(\rho_0^X Q_t \parallel \pi^X Q_t) &= -\frac{1}{2}\,\mathsf{FI}_q(\rho_0^X Q_t \parallel \pi^X Q_t)\\ &\le -\frac{\alpha_t}{136q}\begin{cases}\mathcal{R}_q(\rho_0^X Q_t \parallel \pi^X Q_t)^{2-2/r}\,, & \mathcal{R}_q(\rho_0^X Q_t \parallel \pi^X Q_t) \ge 1\,,\\ \mathcal{R}_q(\rho_0^X Q_t \parallel \pi^X Q_t)\,, & \mathcal{R}_q(\rho_0^X Q_t \parallel \pi^X Q_t) \le 1\,.\end{cases}\end{aligned}$$

We split into two cases. If $\mathcal{R}_q(\rho_0^X \| \pi^X) \ge 1$, then as long as $\mathcal{R}_q(\rho_0^X Q_t \| \pi^X Q_t) \ge 1$, we have

$$\begin{aligned}\partial_t\, \mathcal{R}_q(\rho_0^X Q_t \parallel \pi^X Q_t)^{2/r-1} &= (\frac{2}{r}-1)\,\frac{\partial_t \mathcal{R}_q(\rho_0^X Q_t \parallel \pi^X Q_t)}{\mathcal{R}_q(\rho_0^X Q_t \parallel \pi^X Q_t)^{2-2/r}}\\ &\le -\frac{\alpha_t}{136q}\,(\frac{2}{r}-1)\end{aligned}$$

and therefore

$$\begin{aligned}\mathcal{R}_q(\rho_0^Y \parallel \pi^Y)^{2/r-1} &= \mathcal{R}_q(\rho_0^X Q_h \parallel \pi^X Q_h)^{2/r-1}\\ &\le \Big(\mathcal{R}_q(\rho_0^X \parallel \pi^X)^{2/r-1} - \frac{(2/r-1)\ln(1+\alpha h)}{136q}\Big) \vee 1\,.\end{aligned}$$

In the second case, if $\mathcal{R}_q(\rho_0^X \parallel \pi^X) \le 1$, then

$$\partial_t \mathcal{R}_q(\rho_0^X Q_t \parallel \pi^X Q_t) \le -\frac{\alpha_t}{136q}\,\mathcal{R}_q(\rho_0^X Q_t \parallel \pi^X Q_t)\,.$$

Integrating,

$$\mathcal{R}_q(\rho_0^Y \parallel \pi^Y) = \mathcal{R}_q(\rho_0^X Q_h \parallel \pi^X Q_h) \le \frac{\mathcal{R}_q(\rho_0^X \parallel \pi^X)}{(1+\alpha h)^{1/(136q)}}\,.$$

2. **Backward step:** Along the simultaneous backward heat flow, Lemma 4.4.4 next yields

$$\partial_t \mathcal{R}_q(\rho_0^Y Q_t^{\leftarrow} \,\|\, \pi^Y Q_t^{\leftarrow}) = -\frac{1}{2}\,\mathsf{FI}_q(\rho_0^Y Q_t^{\leftarrow} \,\|\, \pi^Y Q_t^{\leftarrow})\,.$$

Using entirely analogous arguments as in the forward step, we obtain

$$\begin{aligned}\mathcal{R}_q(\rho_1^X \,\|\, \pi^X)^{2/r-1} &= \mathcal{R}_q(\rho_0^Y Q_h^{\leftarrow} \,\|\, \pi^Y Q_h^{\leftarrow})^{2/r-1} \\ &\le \Big(\mathcal{R}_q(\rho_0^Y \,\|\, \pi^Y)^{2/r-1} - \frac{(2/r-1)\ln(1+\alpha h)}{136q}\Big) \vee 1\end{aligned}$$

if $\mathcal{R}_q(\rho_0^Y \,\|\, \pi^Y) \ge 1$, and

$$\mathcal{R}_q(\rho_1^X \mid \pi^X) = \mathcal{R}_q(\rho_0^Y Q_h^{\leftarrow} \,\|\, \pi^Y Q_h^{\leftarrow}) \le \frac{\mathcal{R}_q(\rho_0^Y \,\|\, \pi^Y)}{(1+\alpha h)^{1/(136q)}}$$

if $\mathcal{R}_q(\rho_0^Y \,\|\, \pi^Y) \le 1$.

□

■ 4.4.7 Rejection sampling implementation of the RGO

The following result on rejection sampling is standard, and we include it for the sake of completeness.

Theorem 4.4.6. *Suppose we have query access to the unnormalized target $\tilde{p} = pZ_p$ supported on $\mathscr{X}$, and that we have an upper envelope $\tilde{q} \ge \tilde{p}$. Let q denote the corresponding normalized probability distribution and write Z_q for the normalizing constant, i.e., $\tilde{q} = qZ_q$. Then, rejection sampling with acceptance probability $\tilde{p}/\tilde{q}$ outputs a point distributed according to p, and the number of samples drawn from q until a sample is accepted follows a geometric distribution with mean Z_q/Z_p.*

Proof. Since $\tilde{q}$ is an upper envelope for $\tilde{p}$, then $\tilde{p}(X)/\tilde{q}(X) \le 1$ is a valid acceptance probability. Clearly, the number of rejections follows a geometric distribution. The probability of accepting a sample is given by

$$\mathbb{P}(\text{accept}) = \int_{\mathscr{X}} \frac{\tilde{p}(x)}{\tilde{q}(x)}\, q(\mathrm{d}x) = \frac{Z_p}{Z_q}\int_{\mathscr{X}} p(\mathrm{d}x) = \frac{Z_p}{Z_q}\,.$$

Let $X_1, X_2, X_3, \ldots$ be a sequence of i.i.d. samples from q and let $U_1, U_2, U_3, \ldots$ be i.i.d. $\mathsf{uniform}[0,1]$. Let $A \subseteq \mathscr{X}$ be a measurable set, and let X be the output

of the rejection sampling algorithm. Partitioning by the number of rejections, we may write

$$\begin{aligned}
\mathbb{P}(X \in A) &= \sum_{n=0}^{\infty} \mathbb{P}\Big(X_{n+1} \in A,\ U_i > \frac{\tilde{p}(X_i)}{\tilde{q}(X_i)}\,\forall\, i \in [n],\ U_{n+1} \le \frac{\tilde{p}(X_{n+1})}{\tilde{q}(X_{n+1})}\Big) \\
&= \sum_{n=0}^{\infty} \mathbb{P}\Big(X_{n+1} \in A,\, U_{n+1} \le \frac{\tilde{p}(X_{n+1})}{\tilde{q}(X_{n+1})}\Big)\, \mathbb{P}\Big(U_1 > \frac{\tilde{p}(X_1)}{\tilde{q}(X_1)}\Big)^n \\
&= \sum_{n=0}^{\infty} \Big(\int_A \frac{\tilde{p}(x)}{\tilde{q}(x)}\, q(\mathrm{d}x)\Big)\Big(\int_{\mathscr{X}} \big(1 - \frac{\tilde{p}(x)}{\tilde{q}(x)}\big)\, q(\mathrm{d}x)\Big)^n \\
&= p(A)\, \frac{Z_p}{Z_q} \sum_{n=0}^{\infty} \big(1 - \frac{Z_p}{Z_q}\big)^n = p(A)\,.
\end{aligned}$$

□

■ 4.5 Optimization proofs inspired by the proximal sampler

■ 4.5.1 Alternative proof of the contractivity of the proximal map

The following theorem is well-known in optimization.

Theorem 4.5.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be α-strongly convex and differentiable. Then, the proximal mapping*

$$\mathrm{prox}_{hf}(y) \coloneqq \arg\min_{x \in \mathbb{R}^d} \Big\{ f(x) + \frac{1}{2h} \|x - y\|^2 \Big\}$$

is a $\frac{1}{1+\alpha h}$-contraction.

Here, we give a new proof of the theorem which translates the convergence proof of the proximal sampler in [LST21b] to optimization.

We recall that α-strong convexity implies the $1/\alpha$-PL inequality (or gradient domination inequality)

$$\|\nabla f(x)\|^2 \ge 2\alpha \{f(x) - \min f\} \qquad \text{for all } x \in \mathbb{R}^d\,,$$

which in turn implies the $1/\alpha$-quadratic growth inequality

$$f(x) - \min f \ge \frac{\alpha}{2} \|x - x^\star\|^2 \qquad \text{for all } x \in \mathbb{R}^d\,,$$

with $x^\star = \arg\min f$, see [OV00; BB18].

Proof of Theorem 4.5.1. Let $f_x(z) \coloneqq f(z) + \frac{1}{2h}\|x - z\|^2$, and define f_y similarly. Then, by definition,

$$x' \coloneqq \mathrm{prox}_{hf}(x) = \arg\min f_x\,,$$
$$y' \coloneqq \mathrm{prox}_{hf}(y) = \arg\min f_y\,.$$

Since f_x is $(\alpha + \frac{1}{h})$-strongly convex, then by applying the quadratic growth and PL inequalities,

$$\begin{aligned}\|x' - y'\|^2 &\le \frac{2}{\alpha + 1/h}\{f_x(y') - f_x(x')\} \le \frac{1}{(\alpha + 1/h)^2}\|\nabla f_x(y')\|^2 \\ &= \frac{1}{(\alpha + 1/h)^2}\left\|\nabla f(y') + \frac{1}{h}(y' - x)\right\|^2 \\ &= \frac{1}{(\alpha + 1/h)^2}\left\|-\frac{1}{h}(y' - y) + \frac{1}{h}(y' - x)\right\|^2 = \frac{1}{(1 + \alpha h)^2}\|x - y\|^2\end{aligned}$$

where the last line uses the optimality condition $\nabla f(y') + \frac{1}{h}(y' - y) = 0$ from the definition of y'. □

By comparing with the proof of [LST21b, Lemma 2], we see that f_y is analogous to $\mathsf{KL}(\cdot \,\|\, \pi^{X|Y=y})$ for the proximal sampler.

At first glance, it may appear that the proof above only requires a PL inequality, and not strong convexity. However, this is not the case, as it in fact requires that f_y satisfies $1/(\alpha + 1/h)$-PL for all $y \in \mathbb{R}^d$, which does not follow from (for example) the assumption that f satisfies $1/\alpha$-PL.

■ 4.5.2 Optimal contraction factor for the proximal point method under PL

Our proof uses the Hopf–Lax semigroup, guided by the following intuition. There is an analogy between the standard algebra $(+, \times)$ and the tropical algebra $(\inf, +)$; see, e.g., [Bac+92, Section 9.4] or [ABS21, Lecture 16]. The following table describes these analogies.

$(+, \times)$	$(\inf, +)$
convolution	inf-convolution
Fourier transform	convex conjugate
diffusion	gradient flow
heat equation	Hamilton–Jacobi equation
heat semigroup	Hopf–Lax semigroup

As described in §4.4.1.4, our proofs for the proximal sampler involve computing the time derivative of $t \mapsto \mathsf{KL}(\rho_0^X Q_t \,\|\, \pi^X Q_t)$ where $(\pi^X Q_t)_{t \ge 0}$, $(\rho_0^X Q_t)_{t \ge 0}$ are

simultaneously evolving according to the heat flow. In what follows, we will consider the time derivative of $t \mapsto f_t(x_t)$, where f_t is the Moreau envelope of f.

Proof of Theorem 4.3.9. Let us define, for $t > 0$,

$$f_{t,x}(z) \coloneqq f(z) + \frac{1}{2t} \|z - x\|^2 , \qquad x_t \coloneqq \arg\min f_{t,x} . \tag{4.23}$$

Then $x_t = \mathrm{prox}_{tf}(x)$ and $x \mapsto f_{t,x}(x)$ is the Moreau envelope of f. Recall the optimality condition

$$\nabla f(x_t) + \frac{1}{t} (x_t - x) = 0 .$$

The Moreau envelope satisfies the Hamilton–Jacobi equation

$$\partial_t f_{t,x}(x_t) = \langle \underbrace{\nabla f_{t,x}(x_t)}_{=0}, \dot{x}_t \rangle - \frac{1}{2t^2} \|x_t - x\|^2 .$$

Using the PL inequality,

$$\begin{aligned}
\partial_t f_{t,x}(x_t) &= -\frac{\alpha}{2t\,(1+\alpha t)} \|x_t - x\|^2 - \frac{1}{2t^2\,(1+\alpha t)} \|x_t - x\|^2 \\
&= -\frac{\alpha}{2t\,(1+\alpha t)} \|x_t - x\|^2 - \frac{1}{2\,(1+\alpha t)} \|\nabla f(x_t)\|^2 \\
&\le -\frac{\alpha}{2t\,(1+\alpha t)} \|x_t - x\|^2 - \frac{\alpha}{1+\alpha t} \{f(x_t) - f^\star\}
\end{aligned}$$

which yields

$$\partial_t \{f_{t,x}(x_t) - f^\star\} \le -\frac{\alpha}{1+\alpha t} \{f_{t,x}(x_t) - f^\star\} .$$

Integrating this yields[4]

$$f_{h,x}(x_h) - f^\star \le \{f(x) - f^\star\} \exp\Big(-\int_0^h \frac{\alpha}{1+\alpha t} \,\mathrm{d}t\Big) = \frac{1}{1+\alpha h} \{f(x) - f^\star\} .$$

Hence,

$$\begin{aligned}
\frac{1}{1+\alpha h} \{f(x) - f^\star\} &\ge f(x') - f^\star + \frac{1}{2h} \|x' - x\|^2 = f(x') - f^\star + \frac{h}{2} \|\nabla f(x')\|^2 \\
&\ge f(x') - f^\star + \alpha h \{f(x') - f^\star\} = (1+\alpha h) \{f(x') - f^\star\} .
\end{aligned}$$

This completes the proof. □

[4] Denote by $(Q_t^{\mathrm{HL}})_{t \ge 0}$ the Hopf–Lax semigroup defined by $Q_t^{\mathrm{HL}} f(x) = f_{t,x}(x_t)$. One can check that $Q_t^{\mathrm{HL}} f(x^\star) = f(x^\star)$ where $x^\star = \arg\min f$. So, we can rewrite this inequality as $Q_t^{\mathrm{HL}} f(x) - Q_t^{\mathrm{HL}} f(x^\star) \le \frac{1}{(1+\alpha t)} \{f(x) - f(x^\star)\}$.

4.6 Conclusion

In this chapter, we have studied in detail the proximal sampler of [TP18; LST21c]. In particular, we have given new convergence proofs under weaker assumptions than what were previously considered, allowing for a much wider class of distributions beyond log-concavity. In some cases, our proofs are inspired by convex optimization; in others, they show a remarkable parallel with the continuous-time theory of the Langevin diffusion under isoperimetry. Additionally, we have drawn more precise links between the proximal sampler and the proximal point method in optimization.

We conclude by listing a few directions for future study.

1. Is there an extension of the theory we have developed to the problem of sampling from composite potentials $\pi^X \propto \exp(-f-g)$?

2. Is there an accelerated version of the proximal sampler?

Additionally, since the complexity of the proximal sampler hinges on the complexity of the subroutine used to implement the RGO, this represents a potential avenue towards better sampling guarantees. In this chapter, we have only considered a simple rejection sampling implementation for the RGO. In the next two chapters, we shall develop a faster implementation based on the Metropolis-adjusted Langevin algorithm, which ultimately leads to improved complexities over the ones obtained in this chapter by a factor of $\sqrt{d}$. In the concurrent and independent work [FYC23], similar improvements were obtained via approximate rejection sampling implementations of the RGO.

Chapter 5

Analysis of MALA from a warm start

In §4, we studied the proximal sampler which is an *unbiased* sampling algorithm (with perfect implementation of the RGO) and hence leads to high-accuracy sampling guarantees. Another method for designing unbiased samplers is to add a Metropolis–Hastings filter step; when applied to LMC, it yields the Metropolis-adjusted Langevin algorithm (MALA). In the strongly log-concave case, the results of §4 match the existing complexity guarantees for MALA [Dwi+19; Che+20a; LST20]; in particular, the dimension dependence of both algorithms scale as $\widetilde{O}(d)$.

In this chapter, we break the $\widetilde{O}(d)$ barrier by showing that the dimension dependence of MALA improves to $\widetilde{O}(\sqrt{d})$ under a *warm start*; moreover, we show via lower bounds that for MALA, this rate is tight. The question of algorithmically obtaining a warm start for MALA to take advantage of this faster rate will be addressed in §6.

This chapter is based on [Che+21b], joint with Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet.

5.1 Introduction

The class of Metropolis–Hastings (MH) adjusted algorithms [Met+53; Has70], which includes the Metropolized random walk (MRW) algorithm, the Metropolis-adjusted Langevin algorithm (MALA), and Metropolized Hamiltonian Monte Carlo (MHMC), is particularly popular for sampling in practice. As such, their convergence properties are of central theoretical and practical interest. More specifically, with the ever-growing size of sample spaces, a precise characterization of how dimension affects convergence rates is a necessary step to develop a better understanding and, ultimately, practical guidelines for this suite of algorithms. In this chapter, we address this pressing question by characterizing the dimension dependence of MALA over a natural class of distributions from a warm start.

Formally, we consider the task of sampling from a target distribution π supported on $\mathbb{R}^d$, with density $\pi(\boldsymbol{x}) \propto \exp(-V(\boldsymbol{x}))$, where $V : \mathbb{R}^d \to \mathbb{R}$ is a strongly convex and smooth potential. Here, [RGG97] initiated the study of dimension dependence of MRW by means of an asymptotic framework: namely, when π is a product distribution, a scaling limit exists for MRW as the dimension tends to infinity with a dimension-dependent step size $h \approx d^{-1}$, thereby suggesting that the number of steps needed for MRW to reach stationarity is on the order of d. Subsequently, [RR98] [see also PST12] extended the scaling limit approach to MALA, suggesting that the dimension dependence for MALA is $d^{1/3}$ for sufficiently regular potentials and step size $h \approx d^{-1/3}$. Beyond its theoretical beauty, this result has had a tremendous practical impact by guiding the choice of step size for MALA even for distributions far beyond the scope of their seminal paper. Understanding the applicability of this result, and ultimately the optimal rate of convergence of MALA, requires a careful inspection of the framework laid out in [RR98]. It turns out that it is rather limited in several aspects. Perhaps most notably, it requires π to be a product distribution, which excludes distributions with complex dependence structures that are now routinely encountered in high-dimensional statistics. Moreover, it applies only to potentials V with higher-order regularity; this is not a mere technical artefact since the limit acceptance probability of MALA as $d \to \infty$ involves the third derivative of V. Finally, the asymptotic nature of the scaling limit result only suggests dimension dependence in the asymptotic limit as $d \to \infty$, so it potentially washes away important effects that may arise for finite d.

Thus it is natural to investigate the rate of convergence of MALA from a perspective that is now customary in the machine learning and optimization literature: by establishing non-asymptotic rates of convergence that hold uniformly over natural classes of target distributions which go beyond product distributions. We begin with the simplest and most natural setting and ask:

> What is the optimal dimension dependence of the mixing time of MALA uniformly over the class of α-strongly convex and β-smooth potentials?

Interestingly, and somewhat surprisingly, we show that while the rate $d^{1/3}$ originally established by [RR98] is indeed optimal for some product distributions such as the standard Gaussian, it is not optimal uniformly over the class of smooth and strongly convex potentials of interest in this work. In fact, for any choice of d, we exhibit a product distribution with infinitely differentiable potential on which MALA requires a stepsize much smaller than $d^{-1/3}$, thus resulting in a worse mixing time. This construction confirms the limitations of the scaling limit approach to establishing optimal dimension dependence.

Related work. The non-asymptotic performance of sampling algorithms uniformly over the class of smooth and strongly convex potentials has been the object

of intense research activity recently. For example, [Dwi+19; Che+20a] show that on this class of potentials, MRW can draw samples with at most ε^2 error in chi-squared divergence with $\widetilde{O}(d \log \frac{1}{\varepsilon})$ steps, thereby providing a non-asymptotic affirmation of the scaling limit of [RGG97]. However, far less is known about optimal rates for MALA. The current best result for MALA on the class of smooth and strongly convex potentials is the paper [LST20], which proves a complexity of $\widetilde{O}(d \log \frac{1}{\varepsilon})$ steps to achieve ε^2 error in chi-squared divergence. They also raise the question of whether there is a gap between the complexities of MRW and MALA.

The paper [MV19] took a direct aim at improving the dimension dependence of mixing time bounds for MALA. They succeeded in obtaining a bound of $\widetilde{O}(d^{2/3})$ albeit at the cost of stringent hypotheses. More specifically, they assume bounds on the third and fourth derivatives of the potential V; when these bounds are $O(1)$ (which is true for the standard Gaussian) then their mixing time is $\widetilde{O}(d^{2/3})$; see the discussion in [Che+20a].

Our contributions. In this work, we show that the mixing time in chi-squared divergence for MALA on the class of smooth and strongly convex potentials with a warm start is $\widetilde{\Theta}(d^{1/2})$. Our result consists of two parts: an upper bound on the mixing time which improves to optimality prior results such as [Dwi+19; Che+20a; LST20], as well as the construction of smooth and strongly convex potentials on which the mixing time of MALA is no better than $d^{1/2}$.

In order to prove our upper bound on the mixing time, we introduce new techniques based on the characterization of the Metropolis filter as a projection of the Markov transition kernel in expected L^1 distance [BD01]. Our techniques effectively reduce the problem of bounding the mixing time to controlling the discretization error between the continuous-time and discretized Langevin processes, which has been extensively studied in the sampling literature. We do not aim to give a comprehensive bibliography here, but we note that our discretization analysis is closest to the papers [DT12; Dal17b], as well as the Girsanov argument of §3. In this way, our upper bound has the potential to connect the vast literature on discretization of SDEs with the more difficult analysis of Metropolized algorithms, although it is likely that further innovations are necessary before the study of the latter is completely reduced to the former.

Notation. We use the symbol $\boldsymbol{x}$ to denote a d-dimensional vector, and the plain symbol x to denote a scalar variable. We abuse notation by identifying measures with their densities (w.r.t. Lebesgue measure); thus, for instance, π represents the stationary distribution (a measure), and the notation $\pi(\boldsymbol{x})$ refers to the corresponding density evaluated at $\boldsymbol{x}$.

5.2 Preliminaries

5.2.1 Assumptions

We consider the problem of sampling from a distribution π supported on $\mathbb{R}^d$. The density of the distribution is given by $\pi(\boldsymbol{x}) \propto \exp(-V(\boldsymbol{x}))$, and we refer to $V : \mathbb{R}^d \to \mathbb{R}$ as the *potential*. Throughout the paper, we will assume that V is twice continuously differentiable, α-strongly convex, and β-smooth, meaning

$$\alpha I_d \preceq \nabla^2 V(\boldsymbol{x}) \preceq \beta I_d \,, \qquad \forall\, \boldsymbol{x} \in \mathbb{R}^d \,.$$

We denote by $\kappa \coloneqq \beta/\alpha$ the *condition number*. For the sake of normalization, we assume that $V(\mathbf{0}) = \min V = 0$, so that $\nabla V(\mathbf{0}) = \mathbf{0}$.

5.2.2 Metropolis-adjusted Langevin algorithm (MALA)

Before stating our main results, we give some background on MALA and tools for establishing convergence rates of Markov chains.

Given a step size $h > 0$, MALA produces a sequence $(\boldsymbol{X}_n)_{n\geq 0}$ of random points in $\mathbb{R}^d$ as follows. First, MALA is initialized at $\boldsymbol{X}_0 \sim \mu_0$. Then, for $n \geq 0$, repeat the following two-step procedure:

1. Proposal step: sample $\boldsymbol{Y}_{n+1} \sim Q(\boldsymbol{X}_n, \cdot)$, where

$$Q(\boldsymbol{x}, \cdot) \coloneqq \frac{1}{(4\pi h)^{d/2}} \exp\Big(-\frac{\|\cdot - \boldsymbol{x} + h\,\nabla V(\boldsymbol{x})\|^2}{4h}\Big) \,.$$

   This proposal density corresponds to one step of the unadjusted Langevin algorithm.

2. Accept-reject step: set

$$\boldsymbol{X}_{n+1} = \begin{cases} \boldsymbol{Y}_{n+1} & \text{with probability } A(\boldsymbol{X}_n, \boldsymbol{Y}_{n+1}) \\ \boldsymbol{X}_n & \text{with probability } 1 - A(\boldsymbol{X}_n, \boldsymbol{Y}_{n+1}) \end{cases}$$

   where the acceptance probability is given by

$$A(\boldsymbol{x}, \boldsymbol{y}) \coloneqq 1 \wedge a(\boldsymbol{x}, \boldsymbol{y}) \,, \qquad a(\boldsymbol{x}, \boldsymbol{y}) \coloneqq \frac{\pi(\boldsymbol{y}) Q(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x}) Q(\boldsymbol{x}, \boldsymbol{y})} \,. \tag{5.1}$$

It is well-known that MALA outputs a sequence of random variables $(\boldsymbol{X}_n)_{n\geq 0}$ that forms a reversible Markov chain with stationary distribution π and Markov

transition kernel given by

$$\begin{aligned} T(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) &= [1 - A(\boldsymbol{x})]\, \delta_{\boldsymbol{x}}(\mathrm{d}\boldsymbol{y}) + Q(\boldsymbol{x}, \mathrm{d}\boldsymbol{y})\, A(\boldsymbol{x}, \boldsymbol{y})\,, \\ A(\boldsymbol{x}) &= \int Q(\boldsymbol{x}, \boldsymbol{y})\, A(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y} \geq 0\,. \end{aligned} \tag{5.2}$$

For the rest of the chapter, it is important to note that A, Q, etc. depend on the step size h.

There are many choices to measure proximity of the MALA output with the target distribution (see §2.2.3). In this work, we focus on the total variation distance (TV), the Kullback–Leibler divergence (KL), the chi-squared divergence (χ^2), and the 2-Wasserstein distance (W_2). Given a measure of discrepancy d between probability measures, we define the mixing time, with initial distribution μ_0, as follows:

$$\tau_{\mathrm{mix}}(\varepsilon, \mu_0; \mathsf{d}) \coloneqq \inf\{n \in \mathbb{N} \,:\, \boldsymbol{X}_0 \sim \mu_0,\ \mathsf{d}(\mu_n, \pi) \leq \varepsilon\}\,.$$

Extensions to other discrepancies, such as the p-Wasserstein distance for $p \leq 2$ or the Hellinger distance, are straightforward and omitted for brevity.

The mixing time of a Markov chain is governed by its spectral gap, which we now introduce. To that end, recall that the *Dirichlet form* associated with the MALA kernel T is the quadratic form

$$\mathcal{E}(f, g) = \mathbb{E}_\pi[f\,(\mathrm{id} - T)\, g]\,, \qquad f, g \in L^2(\pi)\,,$$

where $(Tg)(\boldsymbol{x}) \coloneqq \int g(\boldsymbol{y})\, T(\boldsymbol{x}, \mathrm{d}\boldsymbol{y})$. The *spectral gap* is defined as

$$\lambda \coloneqq \inf\left\{\frac{\mathcal{E}(f, f)}{\operatorname{var} f} \,:\, f \in L^2(\pi),\ \operatorname{var} f > 0\right\}. \tag{λ}$$

Since it is often difficult to control the spectral gap directly, it is also convenient to introduce the *conductance*, defined as

$$\mathsf{C} \coloneqq \inf\left\{\frac{\int_S T(\boldsymbol{x}, S^{\mathsf{c}})\, \pi(\mathrm{d}\boldsymbol{x})}{\pi(S)} \,:\, S \subseteq \mathbb{R}^d,\ \pi(S) \leq \frac{1}{2}\right\}. \tag{C}$$

By Cheeger's inequality [LS88], it holds that

$$\mathsf{C}^2 \lesssim \lambda \lesssim \mathsf{C}\,. \tag{5.3}$$

Actually, in order for the mixing time results we invoke to be valid, we must instead consider the $\frac{1}{2}$-lazy version of the chain, in which each proposal is discarded with probability $\frac{1}{2}$. Since this only affects the mixing time bounds by a factor of 2, we henceforth ignore this distinction.

The Metropolis-adjusted Langevin algorithm (MALA) has been studied for nearly three decades since [Bes+95], especially within an asymptotic framework; see, e.g., the influential work of [RR98].

5.3 The Gaussian case

As our work is motivated by the diffusion scaling limit of [RR98], which predicts a $d^{1/3}$ mixing time for MALA, it is natural to begin our investigations by asking whether this is indeed the correct order of the mixing time in the simplest possible setting: namely, when π is the standard Gaussian distribution. Our first contribution is to establish that it is indeed the case even for finite d. We formulate here an informal result and postpone a more detailed statement together with a proof to §5.8. Though it is expected, this result appears to be new.

Theorem 5.3.1 (Informal)**.** *If the target distribution π is the standard Gaussian distribution, then the mixing time of MALA under a warm start is $\Theta(d^{1/3})$, and is achieved with step size $h \approx d^{-1/3}$.*

The proof of this result is based on explicit calculations. While limited to the Gaussian case, its inspection is instructive for potential extensions to other distributions.

On the one hand, the upper bound on the mixing time relies on fine cancellations in the acceptance probability using the explicit form of the Gaussian distribution, which is unavailable for more general potentials. In general, it is difficult to control the acceptance probability directly, and this seems to be the main obstacle to sharpening the mixing time bound in [Dwi+19]. This observation motivates us to seek an indirect way of controlling the acceptance probability in the next section.

On the other hand, while the Gaussian target distribution readily yields a lower bound over the class of potentials with smooth and strongly convex potentials, it turns out to be too loose to address the optimality of MALA. In §5.5, we show that a tighter lower bound may be achieved using a carefully chosen perturbation of the Gaussian distribution.

See also [LST21a], which proved that the mixing time is $\widetilde{\Theta}(d^{1/2})$ if a warm start is not available.

5.4 Upper bound

In order to prove an upper bound on the mixing time of MALA, we assume that we have access to a *warm start*. This is a common assumption which has been employed in previous works on MALA, e.g., [Dwi+19; MV19; Che+20a].

Definition 5.4.1 (Warm start)**.** *We say that the initial distribution μ_0 is M_0-*warm *with respect to π if for any Borel set $E \subseteq \mathbb{R}^d$, it holds that $\mu_0(E) \leq M_0\pi(E)$. When clear from the context, we simply say that an algorithm has a M_0-warm start*

to indicate that it is initialized at an M_0-warm *distribution and omit reference to the target distribution.*

We now state our upper bound on the mixing time of MALA, which shows that under a warm start the mixing time of MALA is $\widetilde{O}(\sqrt{d})$.

Theorem 5.4.2. *Fix* $\varepsilon > 0$ *and consider a target distribution* π *satisfying the assumptions of §5.2.1. Then MALA with a* M_0*-warm start and step size*

$$h = \frac{c}{\beta}\left(\frac{1}{d^{1/2}\log(\kappa d M_0/\varepsilon)} \wedge \frac{1}{\kappa}\right)$$

for a sufficiently small absolute constant $c > 0$*, has mixing time given by*

$$\tau_{\mathrm{mix}}(\varepsilon, \mu_0; \mathsf{d}) \lesssim \left(\kappa d^{1/2} \log \frac{\kappa d M_0}{\varepsilon} + \kappa^2\right) \log\left(\frac{M_0}{\varepsilon}\right)$$

for each of the distances

$$\mathsf{d} \in \{\mathsf{TV},\ \sqrt{\mathsf{KL}},\ \sqrt{\chi^2},\ \sqrt{\alpha}\, W_2\}\,.$$

The main properties of strongly log-concave distributions that we use in the proof are summarized in Lemma 5.6.11. As long as π satisfies these properties, the upper bound technique may be applied under weaker assumptions, e.g., a log-Sobolev inequality. We do not pursue these extensions further in this paper.

We primarily work with the total variation distance to establish the above upper bound on the mixing time and translate this result to the chi-squared divergence by using M_0-warmness of all the iterates of the MALA chain. In turn, this result extends to the KL divergence using a standard comparison inequality [see, e.g., Tsy09, §2] and ultimately to the Wasserstein distance using Talagrand's transport inequality for strongly log-concave distributions; see §2.2.3.

The quantity $\log M_0$ is important because it can introduce additional dimensional factors under a feasible start [Dwi+19]. We address this issue in §6.

Since our upper bound proof may be of interest for analyzing other sampling algorithms based on Metropolis–Hastings filters, we now proceed to give a technical overview of the ideas involved in the upper bound. Throughout, we use the notation $Q_{\boldsymbol{x}}(\cdot)$, $T_{\boldsymbol{x}}(\cdot)$, etc. as a shorthand for the kernels $Q(\boldsymbol{x}, \cdot)$, $T(\boldsymbol{x}, \cdot)$, etc.

We begin by describing the approach of [Dwi+19], which will serve as a reference. The standard technique for bounding the conductance of geometric random walks is the following lemma [see, e.g., LV18a, Lemma 13].

Lemma 5.4.3. *Suppose that for all* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ *with* $\|\boldsymbol{x} - \boldsymbol{y}\| \le r$*, it holds that* $\|T_{\boldsymbol{x}} - T_{\boldsymbol{y}}\|_{\mathsf{TV}} \le 3/4$*. Then, the conductance of the MALA chain satisfies* $\mathsf{C} \gtrsim \sqrt{\alpha}\, r$*.*

In light of this lemma, [Dwi+19] considers the following decomposition:

$$\|T_{\boldsymbol{x}} - T_{\boldsymbol{y}}\|_{\mathsf{TV}} \le \|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} + \|Q_{\boldsymbol{x}} - Q_{\boldsymbol{y}}\|_{\mathsf{TV}} + \|T_{\boldsymbol{y}} - Q_{\boldsymbol{y}}\|_{\mathsf{TV}} \,. \tag{5.4}$$

The middle term is the TV distance between two Gaussian distributions, and using Pinsker's inequality it is straightforward to show that

$$\|Q_{\boldsymbol{x}} - Q_{\boldsymbol{y}}\|_{\mathsf{TV}} \le \frac{\|\boldsymbol{x} - \boldsymbol{y}\|}{\sqrt{2h}} \,, \qquad \text{provided } h \le \frac{2}{\beta} \,,$$

see [Dwi+19, Lemma 3]. On the other hand, bounding the first and third terms in the decomposition (5.4) requires carefully controlling the acceptance probability of MALA. In [Dwi+19], the authors show that these terms can be controlled when the step size is of order $h \approx 1/d$. An application of Lemma 5.4.3 with $r \approx \sqrt{h}$ yields a conductance bound of $\mathsf{C} = \Omega(1/\sqrt{d})$ and in turn, a spectral gap bound of $\lambda = \Omega(1/d)$ by Cheeger's inequality (5.3). Overall, this approach yields a mixing time bound of $O(d)$.

In order to prove a stronger mixing time bound of $\widetilde{O}(\sqrt{d})$, we must consider much larger step sizes (of order $h \approx 1/\sqrt{d}$), and in this regime, controlling the acceptance probabilities by hand requires a daunting computational effort. In fact, [RR98] already resort to a computer-aided proof to study the asymptotics of the acceptance probability. Our first main idea is to use the well-known fact [BD01] that for any proposal Q, the corresponding Metropolis-adjusted kernel T is the closest Markov kernel to Q, among all reversible Markov kernels with stationary distribution π.

Lemma 5.4.4. *Let Q be an atomless proposal kernel, and let T be the kernel obtained from Q by Metropolis adjustment (defined by (5.1) and (5.2)). Let $\bar{Q}$ be any kernel that is reversible with respect to π and has no atoms. Then, for $\boldsymbol{x} \sim \pi$, it holds that*

$$\mathbb{E}\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} \le 2\,\mathbb{E}\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} \,.$$

Proof. See §5.6.2. □

We apply this result by comparing the MALA kernel T with the transition kernel $\bar{Q}$ of the continuous-time Langevin diffusion run for time h. In other words, $\bar{Q}(\boldsymbol{x}, \cdot)$ is the law of $\bar{\boldsymbol{X}}_h$, where $(\bar{\boldsymbol{X}}_t)_{t \ge 0}$ evolves according to the stochastic differential equation

$$\mathrm{d}\bar{\boldsymbol{X}}_t = -\nabla V(\bar{\boldsymbol{X}}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t \,, \qquad \bar{\boldsymbol{X}}_0 = \boldsymbol{x} \,, \tag{5.5}$$

and $(\boldsymbol{B}_t)_{t\geq 0}$ is a standard Brownian motion. Using standard stochastic calculus arguments (see (5.11)), we show that $\mathbb{E}\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} = O(h\sqrt{d})$ (see (5.11)). This suggests that we can take the step size to be $h \asymp 1/\sqrt{d}$. However, since the lemma only controls the first and third terms of the decomposition (5.4) in expectation, it is not enough to yield a good lower bound on the conductance via Lemma 5.4.3. To remedy this, we prove a new pointwise version of the projection characterization of Metropolis adjustment.

Theorem 5.4.5. *Let Q be an atomless proposal kernel, and let T be the kernel obtained from Q by Metropolis adjustment (defined by (5.1) and (5.2)). Let $\bar{Q}$ be any kernel that is reversible with respect to π and has no atoms. Then, for every $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} \leq 2\,\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} + \int \frac{\pi(\boldsymbol{y})\,\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\,\Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|\,\mathrm{d}\boldsymbol{y}\,. \tag{5.6}$$

Consequently, for any convex increasing function $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ and $\boldsymbol{x} \sim \pi$, $\boldsymbol{y} \sim \bar{Q}(\boldsymbol{x},\cdot)$,

$$\mathbb{E}\,\Phi(\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}}) \leq \frac{1}{2}\,\mathbb{E}\,\Phi(4\,\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}}) + \frac{1}{2}\,\mathbb{E}\,\Phi\big(2\,\big|\frac{Q(\boldsymbol{x},\boldsymbol{y})}{\bar{Q}(\boldsymbol{x},\boldsymbol{y})} - 1\big|\big)\,. \tag{5.7}$$

Proof. See §5.6.2. □

Remark 5.4.6. *If we take the expectation of (5.6) when $\boldsymbol{x} \sim \pi$, we obtain*

$$\mathbb{E}\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} \leq 4\,\mathbb{E}\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}}\,,$$

which qualitatively recovers Lemma 5.4.4.

The second inequality in Theorem 5.4.5 can be used in the usual way to deduce concentration bounds for $\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}}$ when $\boldsymbol{x} \sim \pi$. A key feature of this approach is that both terms on the right-hand side of (5.7), in the case of MALA, involve only quantities which measure the discrepancy between the continuous-time Langevin kernel $\bar{Q}$ and the discretized Langevin proposal Q. Therefore, to control the quantity $\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}}$, it suffices to apply well-established techniques for studying the discretization of SDEs.

Once we show that $\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}}$ is controlled with high probability, we are then able to apply a conductance argument, similar to Lemma 5.4.3, in order to prove our mixing time bound. We give an in-depth overview of the proof and provide proofs of technical details in §5.6.

5.5 Lower bound

It is a standard fact that the mixing time is governed by the inverse of the spectral gap[1]. Hence, an upper bound on the spectral gap λ yields a lower bound on the mixing time. In addition, we know from Cheeger inequality (5.3) that $\lambda \lesssim \mathsf{C}$, where C denotes the conductance of the Markov chain. For these reasons, we identify a lower bound on the mixing time with an upper bound on either the conductance C or the spectral gap λ.

To complement our upper bound on the mixing time of MALA, we provide a nearly matching lower bound, thereby settling the question of the dimension dependence of MALA for log-smooth and strongly log-concave targets. To that end, we exhibit a target distribution (in fact a family of distributions) such that the MALA chain with step size h has exponentially small conductance whenever $h \gg d^{-1/2}$. More precisely, fix $\eta \in (0, 1/4)$ and define the adversarial target distribution π_η as a product distribution with potential V_η defined by

$$V_\eta(\boldsymbol{x}) = \frac{\|\boldsymbol{x}\|^2}{2} - \frac{1}{2d^{2\eta}} \sum_{i=1}^{d} \cos(d^\eta x_i)\,. \tag{5.8}$$

It is not hard to see that V_η is $1/2$-strongly convex and $3/2$-smooth. To motivate this choice, recall from [RR98, Theorem 1] that the acceptance probability of MALA tends to a positive constant as $d \to \infty$ whenever the second moment of the third derivative of the potential is finite and the step size is chosen as $h = \Theta(d^{-1/3})$. The choice V_η in (5.8) is an example of a smooth and strongly convex potential where this condition is violated asymptotically, therefore suggesting that $h = \Theta(d^{-1/3})$ is too large to prevent the acceptance probability to vanish for large d. Our first result below indicates that h should be taken significantly smaller than $d^{-1/3}$; in fact nearly as small as $d^{-1/2}$ when $\eta \approx 1/4$.

In the following theorem, we set $\eta = 1/4 - \delta$, for some small $\delta > 0$.

Theorem 5.5.1. *Fix $\delta \in (0, 1/18)$, let $\eta = 1/4 - \delta$, and let C denote the conductance of the MALA chain with target distribution π_η and step size h. Then, $\mathsf{C} \lesssim \exp[-\Omega(d^{4\delta})]$ for any $h \in [d^{-\frac{1}{2}+3\delta}, d^{-\frac{1}{3}}]$.*

Note that as $\delta \searrow 0$, the above theorem shows that MALA must take step sizes which are (essentially) at most of order $d^{-1/2}$.

[1]By definition, the spectral gap corresponds to the smallest eigenvalue of the Dirichlet form. Hence, for an initial distribution μ_0 that is correlated with the eigenfunction corresponding to λ, it follows that $\tau_{\mathrm{mix}}(\varepsilon, \mu_0; \sqrt{\chi^2}) = \widetilde{\Omega}(\lambda^{-1})$. See, e.g., [BGL14, §4] for a rigorous treatment of spectral theory.

The next result shows that the spectral gap of MALA is no better than h. Together with our upper bound, it implies in particular that the choice $h \approx d^{-1/2}$ is the optimal step size for MALA for a target distribution π_η and hence, cannot be improved uniformly over the class of distributions with smooth and strongly convex potentials.

Theorem 5.5.2. *The spectral gap λ of MALA with target distribution π_η and step size $0 < h \leq 1$ satisfies $\lambda \lesssim h$.*

We give the proofs of these theorems in §5.7.

Remark 5.5.3. *In fact, in our proof, we construct an event E with $\pi(E) \geq 1/2$, such that with step size h in the range $[d^{-\frac{1}{2}+3\delta}, d^{-\frac{1}{3}}]$ the acceptance probability starting at any point in E is $\exp[-\Omega(d^{4\delta})]$. Note then that the initialization $\mu_0 = \pi(\cdot \mid E)$ is M_0-warm w.r.t. π with $M_0 = 2$. Hence, our construction provides a lower bound on the mixing time of MALA from a warm start.*

5.6 Proof of the upper bound

This section presents the proof of Theorem 5.4.2.

5.6.1 High-level overview of the proof

The bulk of the proof controls the mixing time in total variation and we use results from §5.6.7 to extend it to the other distances.

For the proof, it is technically convenient to work with a refinement of the conductance known as the *s-conductance*: for $0 < s < 1/2$, define

$$\mathsf{C}_s := \inf\Big\{ \frac{\int_S T(\boldsymbol{x}, S^{\mathsf{c}})\, \pi(\mathrm{d}\boldsymbol{x})}{\pi(S) - s} \;\Big|\; S \subseteq \mathbb{R}^d,\ s < \pi(S) \leq \frac{1}{2} \Big\}. \tag{5.9}$$

A lower bound on the s-conductance translates into an upper bound on the mixing time in total variation distance, via the following lemma.

Lemma 5.6.1 ([LS93, Corollary 1.6])**.** *For any $n \in \mathbb{N}$ and $0 < s < 1/2$, the distribution of the n-th iterate μ_n of the MALA satisfies*

$$\|\mu_n - \pi\|_{\mathsf{TV}} \leq H_s + \frac{H_s}{s} \exp\Big(-\frac{\mathsf{C}_s^2 n}{2}\Big),$$

where $H_s := \sup\{|\mu_0(A) - \pi(A)| : \pi(A) \leq s\}$.

By Hölder's inequality, we have $H_s \leq M_0 s$. It yields the following corollary.

Corollary 5.6.2. *Taking $s = \varepsilon/(2M_0)$, it follows that*

$$\|\mu_n - \pi\|_{\mathsf{TV}} \le \varepsilon \qquad \textit{provided that} \qquad n \ge \frac{2}{\mathsf{C}_s^2} \ln \frac{2M_0}{\varepsilon}\,.$$

Motivated by the standard conductance lemma (Lemma 5.4.3) and the decomposition (5.4), in order to bound the s-conductance from below we will first bound $\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}}$, as in §5.4. The outline of the proof is as follows:

1. In §5.6.2, we prove the projection properties of MALA (Lemma 5.4.4 and Theorem 5.4.5).

2. In §5.6.3, we use the projection property (Lemma 5.4.4) along with stochastic calculus to bound the expectation $\mathbb{E}\,\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}}$ when $\boldsymbol{x} \sim \pi$.

3. In §5.6.4, we use the pointwise projection property, together with more stochastic calculus computations, in order to prove a concentration inequality for $\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}}$ when $\boldsymbol{x} \sim \pi$.

4. In §5.6.5, we use the concentration bound of §5.6.4, together with ideas from the proof of the standard conductance lemma (Lemma 5.4.3), in order to lower bound the s-conductance. Together with Corollary 5.6.2, it yields the mixing time bound of Theorem 5.4.2 in total variation distance.

5. Finally in §5.6.7, we explain how the mixing time bound in total variation distance implies mixing time bounds in other distances between probability measures.

■ 5.6.2 Proof of the projection properties

We start with a basic fact about MALA.

Proposition 5.6.3. *Let Q be the proposal kernel and let T be the MALA kernel with proposal Q. Then,*

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} = \int_{\mathbb{R}^d \setminus \{\boldsymbol{x}\}} |T(\boldsymbol{x}, \boldsymbol{y}) - Q(\boldsymbol{x}, \boldsymbol{y})|\, \mathrm{d}\boldsymbol{y} = 1 - \int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y})\, A(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y}.$$

Proof. First, since $T_{\boldsymbol{x}}$ has an atom at $\boldsymbol{x}$ and $Q_{\boldsymbol{x}}$ does not, we have

$$\|Q_{\boldsymbol{x}} - T_{\boldsymbol{x}}\|_{\mathsf{TV}} = \frac{1}{2}\left(T_{\boldsymbol{x}}(\{\boldsymbol{x}\}) + \int_{\mathbb{R}^d \setminus \{\boldsymbol{x}\}} |T(\boldsymbol{x}, \boldsymbol{y}) - Q(\boldsymbol{x}, \boldsymbol{y})|\, \mathrm{d}\boldsymbol{y}\right).$$

By the definition of the accept-reject step,

$$T_{\boldsymbol{x}}(\{\boldsymbol{x}\}) = 1 - \int_{\mathbb{R}^d \setminus \{\boldsymbol{x}\}} T(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} = 1 - \int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y}) \, A(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \,,$$

whereas

$$\int_{\mathbb{R}^d \setminus \{\boldsymbol{x}\}} |T(\boldsymbol{x}, \boldsymbol{y}) - Q(\boldsymbol{x}, \boldsymbol{y})| \, \mathrm{d}\boldsymbol{y} = 1 - \int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y}) \, A(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \,.$$

The result follows. □

We now prove the projection properties (Lemma 5.4.4 and Theorem 5.4.5).

Proof of Lemma 5.4.4. Since the kernel $\bar{Q}$ corresponding to the continuous-time Langevin diffusion is reversible with stationary distribution π, it follows from the result of [BD01] that

$$\iint_{(\mathbb{R}^d \times \mathbb{R}^d) \setminus \Delta} |T(\boldsymbol{x}, \boldsymbol{y}) - Q(\boldsymbol{x}, \boldsymbol{y})| \, \pi(\mathrm{d}\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} \leq \iint_{(\mathbb{R}^d \times \mathbb{R}^d) \setminus \Delta} |\bar{Q}(\boldsymbol{x}, \boldsymbol{y}) - Q(\boldsymbol{x}, \boldsymbol{y})| \, \pi(\mathrm{d}\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} \,,$$

where $\Delta = \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^d \times \mathbb{R}^d \,:\, \boldsymbol{x} = \boldsymbol{y}\}$. Since $Q_{\boldsymbol{x}}$ and $\bar{Q}_{\boldsymbol{x}}$ have no atoms, the right-hand side is equal to $2\,\mathbb{E}_{\boldsymbol{x} \sim \pi} \|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}}$. On the other hand, the left-hand side is equal to $\mathbb{E}_{\boldsymbol{x} \sim \pi} \|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}}$ due to Proposition 5.6.3. □

Proof of Theorem 5.4.5. For any $\boldsymbol{x}$, we have

$$\begin{aligned}
\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} &= \int \{1 - A(\boldsymbol{x}, \boldsymbol{y})\} \, Q(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \\
&= \int \Big[1 - \Big(1 \wedge \frac{\pi(\boldsymbol{y}) \, Q(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x}) \, Q(\boldsymbol{x}, \boldsymbol{y})}\Big)\Big] \, Q(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \\
&\leq \int \Big|1 - \frac{\pi(\boldsymbol{y}) \, Q(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x}) \, Q(\boldsymbol{x}, \boldsymbol{y})}\Big| \, Q(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \\
&\leq \int \Big|1 - \frac{\pi(\boldsymbol{y}) \, \bar{Q}(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x}) \, Q(\boldsymbol{x}, \boldsymbol{y})}\Big| \, Q(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} + \int \frac{\pi(\boldsymbol{y}) \, \bar{Q}(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x})} \Big|\frac{Q(\boldsymbol{y}, \boldsymbol{x})}{\bar{Q}(\boldsymbol{y}, \boldsymbol{x})} - 1\Big| \, \mathrm{d}\boldsymbol{y} \,.
\end{aligned}$$

Observe that the first term is given by

$$\begin{aligned}
\int \Big|1 - \frac{\pi(\boldsymbol{y}) \, \bar{Q}(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x}) \, Q(\boldsymbol{x}, \boldsymbol{y})}\Big| \, Q(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} &= \int \Big|Q(\boldsymbol{x}, \boldsymbol{y}) - \frac{\pi(\boldsymbol{y}) \, \bar{Q}(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x})}\Big| \, \mathrm{d}\boldsymbol{y} \\
&= 2 \, \|Q_{\boldsymbol{x}} - \bar{Q}_{\boldsymbol{x}}\|_{\mathsf{TV}} \,,
\end{aligned}$$

where in the second identity, we used the reversibility of $\bar{Q}$. This concludes the proof of the first inequality.

We now deduce the second inequality from the first. Using monotonicity and convexity of Φ respectively, we get,

$$\begin{aligned}\mathbb{E}\,\Phi(\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}}) &\le \mathbb{E}\,\Phi\Big(2\,\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} + \int \frac{\pi(\boldsymbol{y})\,\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\,\Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|\,\mathrm{d}\boldsymbol{y}\Big) \\ &\le \frac{1}{2}\,\mathbb{E}\,\Phi(4\,\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}}) + \frac{1}{2}\,\mathbb{E}\,\Phi\Big(2\int \frac{\pi(\boldsymbol{y})\,\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\,\Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|\,\mathrm{d}\boldsymbol{y}\Big)\,,\end{aligned}$$

where we take expectation with respect to $\boldsymbol{x} \sim \pi$. Next, noting from stationarity that $\int \pi(\boldsymbol{y})\,\bar{Q}(\boldsymbol{y},\boldsymbol{x})\,\mathrm{d}\boldsymbol{y} = \pi(\boldsymbol{x})$, we apply Jensen's inequality to yield

$$\begin{aligned}&\mathbb{E}\,\Phi\Big(2\int \frac{\pi(\boldsymbol{y})\,\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\,\Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|\,\mathrm{d}\boldsymbol{y}\Big) \\ &\quad= \int \Phi\Big(2\int \frac{\pi(\boldsymbol{y})\,\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\,\Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|\,\mathrm{d}\boldsymbol{y}\Big)\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\ &\quad\le \iint \Phi\Big(2\,\Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|\Big)\,\pi(\boldsymbol{y})\,\bar{Q}(\boldsymbol{y},\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} \\ &\quad= \iint \Phi\Big(2\,\Big|\frac{Q(\boldsymbol{x},\boldsymbol{y})}{\bar{Q}(\boldsymbol{x},\boldsymbol{y})} - 1\Big|\Big)\,\pi(\boldsymbol{x})\,\bar{Q}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}\,,\end{aligned}$$

where we switched $\boldsymbol{x}$ and $\boldsymbol{y}$ in the notation of the last line. □

■ 5.6.3 Expectation of the total variation

We now bound the expectation $\mathbb{E}\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}}$ when $\boldsymbol{x} \sim \pi$ using the projection property (Lemma 5.4.4). Akin to prior work such as [DT12], our primary tool to analyze the discretization of the Langevin diffusion is the Girsanov theorem from stochastic calculus [see, e.g., SV06; Le 16, for classical treatments].

Lemma 5.6.4 (Girsanov theorem)**.** *Let $\bar{\mathbf{Q}}_{\boldsymbol{x}}$ denote the probability measure on path space induced by the solution $(\bar{\boldsymbol{X}}_t)_{t\in[0,h]}$ of the continuous-Langevin diffusion SDE (5.5) started at $\boldsymbol{x}$ and run for time $h > 0$. Moreover, let $\mathbf{Q}_{\boldsymbol{x}}$ denote the probability measure on path space induced by the solution of the following SDE with constant drift*

$$\mathrm{d}\boldsymbol{X}_t = -\nabla V(\boldsymbol{x})\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t\,, \qquad \boldsymbol{X}_0 = \boldsymbol{x}\,.$$

Then, $\mathbf{Q}_{\boldsymbol{x}}$ is absolutely continuous with respect to $\bar{\mathbf{Q}}_{\boldsymbol{x}}$ and has density given by Radon–Nikodym derivative:

$$\frac{\mathrm{d}\mathbf{Q}_{\boldsymbol{x}}}{\mathrm{d}\bar{\mathbf{Q}}_{\boldsymbol{x}}}\big((\bar{\boldsymbol{X}}_t)_{t\in[0,h]}\big) = \exp\Big[\frac{1}{\sqrt{2}}\int_0^h \langle \nabla V(\bar{\boldsymbol{X}}_t) - \nabla V(\boldsymbol{x}), \mathrm{d}\boldsymbol{B}_t\rangle - \frac{1}{4}\int_0^h \|\nabla V(\bar{\boldsymbol{X}}_t) - \nabla V(\boldsymbol{x})\|^2 \,\mathrm{d}t\Big] .$$

Proof. See the proof of Proposition 2 in [DT12]. □

In the following lemma, we use Lemma 5.6.12.

Lemma 5.6.5. *Assume $h \lesssim 1/\beta$. For any $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} \lesssim \beta h \sqrt{d + \beta^2 h \,\|\boldsymbol{x}\|^2} \,.$$

Proof. Let end denote the function that maps a continuous curve $(y_t)_{t\in[0,h]}$ in $\mathbb{R}^d$ to its endpoint: $\mathsf{end}((y_t)_{t\in[0,h]}) \coloneqq y_h$. Then, it is clear that

$$Q_{\boldsymbol{x}} = \mathsf{end}_{\#}\,\mathbf{Q}_{\boldsymbol{x}} \qquad \text{and} \qquad \bar{Q}_{\boldsymbol{x}} = \mathsf{end}_{\#}\,\bar{\mathbf{Q}}_{\boldsymbol{x}} \,,$$

where the notation $f_{\#}\mu$ denotes the pushforward of a measure μ under the mapping f. On the one hand, it follows from the data processing inequality that

$$\mathsf{KL}(\bar{Q}_{\boldsymbol{x}} \,\|\, Q_{\boldsymbol{x}}) = \mathsf{KL}(\mathsf{end}_{\#}\,\bar{\mathbf{Q}}_{\boldsymbol{x}} \,\|\, \mathsf{end}_{\#}\,\mathbf{Q}_{\boldsymbol{x}}) \le \mathsf{KL}(\bar{\mathbf{Q}}_{\boldsymbol{x}} \,\|\, \mathbf{Q}_{\boldsymbol{x}}) \,.$$

On the other hand, the Girsanov theorem (in the form of Lemma 5.6.4) implies

$$\begin{aligned}\mathsf{KL}(\bar{\mathbf{Q}}_{\boldsymbol{x}} \,\|\, \mathbf{Q}_{\boldsymbol{x}}) &= -\,\mathbb{E}\ln\frac{\mathrm{d}\mathbf{Q}_{\boldsymbol{x}}}{\mathrm{d}\bar{\mathbf{Q}}_{\boldsymbol{x}}}(\bar{\boldsymbol{X}}_t) = \frac{1}{4}\int_0^h \mathbb{E}[\|\nabla V(\bar{\boldsymbol{X}}_t) - \nabla V(\boldsymbol{x})\|^2]\,\mathrm{d}t \\ &\le \frac{\beta^2}{4}\int_0^h \mathbb{E}[\|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2]\,\mathrm{d}t \lesssim \beta^2 h^2\,(d + \beta^2 h\,\|\boldsymbol{x}\|^2)\,,\end{aligned}$$

where we used the β-smoothness of V and Lemma 5.6.12. Now applying Pinsker's inequality, we obtain the desired inequality. □

It follows from Lemma 5.6.5 that when $\boldsymbol{x} \sim \pi$, we get

$$\begin{aligned}\mathbb{E}\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} &\lesssim \beta h\,\mathbb{E}\sqrt{d + \beta^2 h\,\|\boldsymbol{x}\|^2} \le \beta h\sqrt{d + \beta^2 h\,\mathbb{E}[\|\boldsymbol{x}\|^2]} \\ &\le \beta h\sqrt{(1+\beta\kappa h)\,d}\,,\end{aligned} \tag{5.10}$$

where we used the second moment bound of Lemma 2.2.13. Together with the projection property (Lemma 5.4.4), it yields

$$\mathbb{E}\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} \le 2\,\mathbb{E}\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} \lesssim \beta h\sqrt{(1+\beta\kappa h)\,d}\,. \tag{5.11}$$

We conclude this section with a concentration inequality which we use later in the argument.

Lemma 5.6.6. *Assume $h \lesssim 1/\beta$ and let $\boldsymbol{x} \sim \pi$. For any $\delta > 0$, with probability at least $1 - \delta$,*

$$\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} \lesssim \beta h \sqrt{d + \beta\kappa h \left(d + \log \frac{1}{\delta}\right)} \,.$$

Proof. Let $f(\boldsymbol{x}) \coloneqq \beta h \sqrt{d + \beta^2 h \,\|\boldsymbol{x}\|^2}$. Then,

$$\|\nabla f(\boldsymbol{x})\| = \frac{\beta^3 h^2 \,\|\boldsymbol{x}\|}{\sqrt{d + \beta^2 h \,\|\boldsymbol{x}\|^2}} \leq \beta^2 h^{3/2} \,.$$

Thus, $f(\boldsymbol{x})$ is $\beta^2 h^{3/2}$-Lipschitz, and it follows from sub-Gaussian concentration (Lemma 5.6.11) that with probability at least $1 - \delta$,

$$f(\boldsymbol{x}) \leq \mathbb{E}\, f(\boldsymbol{x}) + \beta^2 h^{3/2} \sqrt{\frac{2}{\alpha} \ln \frac{1}{\delta}} \,.$$

We have calculated $\mathbb{E}\, f(\boldsymbol{x}) \lesssim \beta h \sqrt{(1 + \beta\kappa h)\, d}$ in (5.10), and the result now follows from the pointwise bound in Lemma 5.6.5. □

■ 5.6.4 Concentration of the total variation

Equation (5.11) provides a control the total variation distance between the MALA kernel and the proposal *in expectation*. The main result of this section is an extension of this result to a control *with high probability* captured in the following proposition.

Proposition 5.6.7. *Fix $c_0 > 0$ and $0 < s < 1/2$. Then, there exists a constant $c_1 > 0$, depending only on c_0, such that with step size*

$$h = \frac{c_1}{\beta} \left(\frac{1}{d^{1/2} \log(d\kappa/s)} \wedge \frac{1}{\kappa} \right) , \tag{5.12}$$

the following holds with probability at least $1 - c_0 s \sqrt{\alpha h}$,

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} \leq \frac{1}{6} \,.$$

The idea of the proof is to use the pointwise projection of Theorem 5.4.5, and to obtain high probability bounds for each of the two terms in (5.6). An upper bound for the first term follows directly from Lemma 5.6.6. To control the second term, we will first obtain a bound on its moments.

Lemma 5.6.8. *Let $k \geq 1$ be any integer. Suppose that*

$$h \lesssim \frac{1}{\beta d^{1/2} k} \left(1 \wedge \frac{d^{1/6} k^{1/3}}{\kappa^{1/3}} \wedge \frac{d^{1/2} k}{\kappa}\right). \tag{5.13}$$

Then, it holds that

$$\left\{\mathbb{E}_{\boldsymbol{x}\sim\pi}\left[\left|\int \frac{\pi(\boldsymbol{y})\,\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\left|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\right| \mathrm{d}\boldsymbol{y}\right|^k\right]\right\}^{1/k} \lesssim \beta h \sqrt{k}\,(\sqrt{d} + \sqrt{k})\,.$$

The proof, given in §5.6.4.1, uses extensively tools from stochastic calculus. We remark that the quantity in Lemma 5.6.8 can be interpreted as a bound on the Rényi divergence between the discretized and continuous Langevin processes. A similar result has appeared as [GT20, Corollary 11]; see also the Girsanov argument of §3.

We are now in a position to prove Proposition 5.6.7.

Proof of Proposition 5.6.7. Assume that the step size h is small enough so that Lemmas 5.6.6 and 5.6.8 both hold. More specifically, since the requirement of Lemma 5.6.8 is more stringent than that of Lemma 5.6.6, so we can simply impose that (5.13) holds.

From Lemma 5.6.6 with $\delta = c_0 s\sqrt{\alpha h}/2$, there exists a constant $C_1 > 0$ such that with probability at least $1 - c_0 s\sqrt{\alpha h}/2$,

$$\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} \leq C_1 \beta h \sqrt{d + \beta\kappa h \ln \frac{2}{c_0 s\sqrt{\alpha h}}}\,.$$

From Lemma 5.6.8 and Markov's inequality, there exists a constant $C_2 > 0$ such that for any $\delta > 0$, with probability at least $1 - \delta$,

$$\int \frac{\pi(\boldsymbol{y})\,\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\left|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\right| \mathrm{d}\boldsymbol{y} \leq C_2 \beta h \sqrt{k}\,(\sqrt{d} + \sqrt{k})\,\delta^{-1/k}\,.$$

Taking $k \sim \ln \frac{2}{c_0 s\sqrt{\alpha h}}$ and $\delta = c_0 s\sqrt{\alpha h}/2$, we have $\delta^{-1/k} = \Theta(1)$ and hence

$$\int \frac{\pi(\boldsymbol{y})\,\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\left|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\right| \mathrm{d}\boldsymbol{y} \leq C_2 \beta h \sqrt{\ln \frac{2}{c_0 s\sqrt{\alpha h}}}\left(\sqrt{d} + \sqrt{\ln \frac{2}{c_0 s\sqrt{\alpha h}}}\right).$$

Combining these two inequalities with the pointwise projection property (Theorem 5.4.5), it follows that with probability at least $1 - c_0 s\sqrt{\alpha h}$,

$$\begin{aligned}\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} \lesssim\ & \beta h \sqrt{d + \beta\kappa h \ln \frac{2}{c_0 s\sqrt{\alpha h}}} \\ & + \beta h \sqrt{\ln \frac{2}{c_0 s\sqrt{\alpha h}}}\left(\sqrt{d} + \sqrt{\ln \frac{2}{c_0 s\sqrt{\alpha h}}}\right).\end{aligned} \tag{5.14}$$

If we choose the constant $c_1 > 0$ small enough, then choosing the step size as in (5.12) makes the both terms in the left-hand side of (5.14) less than $1/12$. This completes the proof of Proposition 5.6.7. □

■ 5.6.4.1 Proof of Lemma 5.6.8

We now prove the moment upper bound (Lemma 5.6.8). Since $\int \pi(\boldsymbol{y})\, \bar{Q}(\boldsymbol{y}, \boldsymbol{x})\, \mathrm{d}\boldsymbol{y} = \pi(\boldsymbol{x})$, we can apply Jensen's inequality to get

$$\begin{aligned}
&\int \pi(\boldsymbol{x}) \Big| \int \frac{\pi(\boldsymbol{y})\, \bar{Q}(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x})} \Big| \frac{Q(\boldsymbol{y}, \boldsymbol{x})}{\bar{Q}(\boldsymbol{y}, \boldsymbol{x})} - 1 \Big| \, \mathrm{d}\boldsymbol{y} \Big|^k \, \mathrm{d}\boldsymbol{x} \\
&\quad \le \iint \pi(\boldsymbol{y})\, \bar{Q}(\boldsymbol{y}, \boldsymbol{x}) \Big| \frac{Q(\boldsymbol{y}, \boldsymbol{x})}{\bar{Q}(\boldsymbol{y}, \boldsymbol{x})} - 1 \Big|^k \, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y} \\
&\quad = \int \Big(\int \Big| \frac{Q(\boldsymbol{x}, \boldsymbol{y})}{\bar{Q}(\boldsymbol{x}, \boldsymbol{y})} - 1 \Big|^k \, \bar{Q}(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) \Big) \, \pi(\mathrm{d}\boldsymbol{x}) \,,
\end{aligned}$$

where we switched $\boldsymbol{x}$ and $\boldsymbol{y}$ in the last line. The inner integral equals the f-divergence $D_f(Q_{\boldsymbol{x}} \,\|\, \bar{Q}_{\boldsymbol{x}})$, with $f(x) \coloneqq |x - 1|^k$. Recall the definitions of $\bar{\mathbf{Q}}_{\boldsymbol{x}}$ and $\mathbf{Q}_{\boldsymbol{x}}$ in Lemma 5.6.4. Hence we may apply the data processing inequality and bound the above by

$$F_k \coloneqq \int \Big(\int \Big| \frac{\mathrm{d}\mathbf{Q}_{\boldsymbol{x}}}{\mathrm{d}\bar{\mathbf{Q}}_{\boldsymbol{x}}} - 1 \Big|^k \, \mathrm{d}\bar{\mathbf{Q}}_{\boldsymbol{x}} \Big) \, \pi(\mathrm{d}\boldsymbol{x}) \,. \tag{5.15}$$

Recall from Lemma 5.6.4 that

$$\frac{\mathrm{d}\mathbf{Q}_{\boldsymbol{x}}}{\mathrm{d}\bar{\mathbf{Q}}_{\boldsymbol{x}}}(\bar{\boldsymbol{X}}) = \exp H_h \,,$$

where for $t \ge 0$,

$$H_t \coloneqq \frac{1}{\sqrt{2}} \int_0^t \langle \nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x}), \mathrm{d}\boldsymbol{B}_s \rangle - \frac{1}{4} \int_0^t \| \nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x}) \|^2 \, \mathrm{d}s \,.$$

Applying Itô's formula to $(H_t)_{t \ge 0}$ and the function exp, we deduce that

$$\exp H_h - 1 = \frac{1}{\sqrt{2}} \int_0^h (\exp H_t) \, \langle \nabla V(\bar{\boldsymbol{X}}_t) - \nabla V(\boldsymbol{x}), \mathrm{d}\boldsymbol{B}_t \rangle \,.$$

In what follows, $\bar{\mathbf{E}}_{\boldsymbol{x}}$ denotes the expectation under $\bar{\mathbf{Q}}_{\boldsymbol{x}}$ (the measure under which $\bar{\boldsymbol{X}}$ is a continuous-time Langevin diffusion). Also, we will use the letter C to

denote a numerical constant which may change from line to line. Based on the upper bound (5.15) on the k-th moment, we wish to estimate

$$\begin{aligned} F_k &= \bar{\mathbf{E}}_{\boldsymbol{x}}[|\exp H_h - 1|^k] = \frac{1}{2^{k/2}}\, \bar{\mathbf{E}}_{\boldsymbol{x}}\Big[\Big|\int_0^h (\exp H_t)\, \langle \nabla V(\bar{\boldsymbol{X}}_t) - \nabla V(\boldsymbol{x}), \mathrm{d}\boldsymbol{B}_t\rangle\Big|^k\Big] \\ &\le (Ck)^{k/2}\, \bar{\mathbf{E}}_{\boldsymbol{x}}\Big[\Big|\int_0^h \exp(2H_t)\, \|\nabla V(\bar{\boldsymbol{X}}_t) - \nabla V(\boldsymbol{x})\|^2\, \mathrm{d}t\Big|^{k/2}\Big] \end{aligned}$$

where the last line is the Burkholder–Davis–Gundy inequality with optimal constants [Bur73; Dav76]. Together with the Cauchy–Schwarz inequality and Hölder's inequality, it yields

$$\begin{aligned} F_k &\le (C\beta^2 k)^{k/2}\, \bar{\mathbf{E}}_{\boldsymbol{x}}\Big[\Big|\int_0^h \exp(4H_t)\, \mathrm{d}t\Big|^{k/4} \Big|\int_0^h \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^4\, \mathrm{d}t\Big|^{k/4}\Big] \\ &\le (C\beta^2 k)^{k/2} \sqrt{\bar{\mathbf{E}}_{\boldsymbol{x}}\Big[\Big|\int_0^h \exp(4H_t)\, \mathrm{d}t\Big|^{k/2}\Big]\, \bar{\mathbf{E}}_{\boldsymbol{x}}\Big[\Big|\int_0^h \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^4\, \mathrm{d}t\Big|^{k/2}\Big]} \\ &\le (C\beta^2 k)^{k/2}\, h^{k/2-1} \underbrace{\sqrt{\Big(\bar{\mathbf{E}}_{\boldsymbol{x}} \int_0^h \exp(2kH_t)\, \mathrm{d}t\Big)}}_{\text{(A)}} \underbrace{\sqrt{\Big(\bar{\mathbf{E}}_{\boldsymbol{x}} \int_0^h \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^{2k}\, \mathrm{d}t\Big)}}_{\text{(B)}}. \end{aligned}$$

We will control the two terms separately, starting with the first term (A).

Lemma 5.6.9. *Let $0 \le t \le h \lesssim \frac{1}{\beta k}$. Then,*

$$\ln \bar{\mathbf{E}}_{\boldsymbol{x}} \exp(2kH_t) \lesssim \beta^2 h^2 k^2\, (\beta^2 h\, \|\boldsymbol{x}\|^2 + d)\,.$$

Proof. Recall the following fact, which follows from Itô's lemma [Le 16, Theorem 5.10]: for any adapted process $(\boldsymbol{Z}_s)_{s\ge 0}$, we have

$$\bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Big(\int_0^t \langle \boldsymbol{Z}_s, \mathrm{d}\boldsymbol{B}_s\rangle - \frac{1}{2}\int_0^t \|\boldsymbol{Z}_s\|^2\, \mathrm{d}s\Big) = 1\,.$$

Together with the Cauchy–Schwarz inequality, it yields

$$\begin{aligned} &\bar{\mathbf{E}}_{\boldsymbol{x}} \exp(2kH_t) \\ &\quad = \bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Big[\sqrt{2}k \int_0^t \langle \nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x}), \mathrm{d}\boldsymbol{B}_s\rangle \\ &\qquad\qquad - \frac{k}{2}\int_0^t \|\nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x})\|^2\, \mathrm{d}s\Big] \end{aligned}$$

$$
\begin{aligned}
&= \bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Big[\sqrt{2}k \int_0^t \langle \nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x}), \mathrm{d}\boldsymbol{B}_s \rangle \\
&\qquad\qquad + \big(-4k^2 + 4k^2 - \frac{k}{2}\big) \int_0^t \|\nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x})\|^2 \,\mathrm{d}s\Big] \\
&\le \sqrt{\bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Big[8k^2 \int_0^t \|\nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x})\|^2 \,\mathrm{d}s\Big]} \\
&\le \sqrt{\bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Big[8\beta^2 k^2 \int_0^t \|\bar{\boldsymbol{X}}_s - \boldsymbol{x}\|^2 \,\mathrm{d}s\Big]} \le \sqrt{\bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Big[8\beta^2 h k^2 \sup_{s\in[0,h]} \|\bar{\boldsymbol{X}}_s - \boldsymbol{x}\|^2\Big]}\,.
\end{aligned}
$$

In order to upper bound the above quantity, we apply Lemma 5.6.12 with $\lambda \coloneqq 8\beta^2 h k^2$. In order to satisfy the preconditions of Lemma 5.6.12, we impose the restriction $h \lesssim \frac{1}{\beta k}$. Then, it follows that

$$
\ln \bar{\mathbf{E}}_{\boldsymbol{x}} \exp(2kH_t) \lesssim \beta^2 h^2 k^2 \left(\beta^2 h \|\boldsymbol{x}\|^2 + d\right).
$$

This is our desired bound. □

Hence, from Lemma 5.6.9, we obtain

$$
\text{(A)} \le \sqrt{h \exp\big(O(\beta^4 h^3 k^2 \|\boldsymbol{x}\|^2 + \beta^2 d h^2 k^2)\big)}\,.
$$

Next, we estimate (B). In fact, Lemma 5.6.12 together with standard moment bounds under sub-exponential concentration (e.g., [Ver18, Proposition 2.7.1]) gives

$$
\bar{\mathbf{E}}_{\boldsymbol{x}} \sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^{2k} \le C^k \left(\beta^{2k} h^{2k} \|\boldsymbol{x}\|^{2k} + d^k h^k + h^k k^k\right),
$$

where $C > 0$ is a numerical constant. See Corollary 5.6.13 in §5.6.6.2 for details. Hence, it holds that

$$
\text{(B)} = \sqrt{\int_0^h \bar{\mathbf{E}}_{\boldsymbol{x}}[\|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^{2k}] \,\mathrm{d}t} \le C^k h^{1/2} \left(\beta^k h^k \|\boldsymbol{x}\|^k + d^{k/2} h^{k/2} + h^{k/2} k^{k/2}\right).
$$

Hence,

$$
\begin{aligned}
(5.15) &\le (C\beta^2 k)^{k/2} h^{k/2-1} \times \text{(A)} \times \text{(B)} \\
&\le (C\beta^2 k)^{k/2} h^{k/2-1} \times h^{1/2} \exp\big(O(\beta^4 h^3 k^2 \|\boldsymbol{x}\|^2 + \beta^2 d h^2 k^2)\big) \\
&\qquad \times C^k h^{1/2} \left(\beta^k h^k \|\boldsymbol{x}\|^k + d^{k/2} h^{k/2} + h^{k/2} k^{k/2}\right) \\
&\le (C\beta^2 h k)^{k/2} \exp\big(O(\beta^2 d h^2 k^2)\big)
\end{aligned}
$$

$$\times \exp\bigl(O(\beta^4 h^3 k^2 \|\boldsymbol{x}\|^2)\bigr)\,(\beta^k h^k \|\boldsymbol{x}\|^k + d^{k/2}h^{k/2} + h^{k/2}k^{k/2})\,.$$

Next, we take the expectation w.r.t. $\boldsymbol{x} \sim \pi$ and use Cauchy–Schwarz:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\pi}\,\bar{\mathbf{E}}_{\boldsymbol{x}}[|\exp H_h - 1|^k] \\
&\quad \le (C\beta^2 hk)^{k/2}\exp\bigl(O(\beta^2 d h^2 k^2)\bigr) \\
&\qquad \times \sqrt{\mathbb{E}_{\boldsymbol{x}\sim\pi}\exp\bigl(O(\beta^4 h^3 k^2\|\boldsymbol{x}\|^2)\bigr)\,\mathbb{E}_{\boldsymbol{x}\sim\pi}[\beta^{2k}h^{2k}\|\boldsymbol{x}\|^{2k} + d^k h^k + h^k k^k]}\,.
\end{aligned}$$

For the two terms involving exponentials: the first will be bounded by a numerical constant provided that $h \le \frac{1}{C\beta k\sqrt{d}}$, and using concentration properties of π (see, e.g., Lemma 5.6.11), the second will be bounded provided $h \le \frac{\alpha^{1/3}}{C\beta^{4/3}d^{1/3}k^{2/3}}$. Taking this to be the case, the moment bounds in Lemma 5.6.11 now imply the bound

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\pi}\,\bar{\mathbf{E}}_{\boldsymbol{x}}[|\exp H_h - 1|^k] \\
&\quad \le (C\beta^2 hk)^{k/2} \times (\alpha^{-k/2}\beta^k d^{k/2}h^k + \alpha^{-k/2}\beta^k h^k k^{k/2} + d^{k/2}h^{k/2} + h^{k/2}k^{k/2})\,.
\end{aligned}$$

Taking k-th roots,

$$\begin{aligned}
&(\mathbb{E}_{\boldsymbol{x}\sim\pi}\,\bar{\mathbf{E}}_{\boldsymbol{x}}[|\exp H_h - 1|^k])^{1/k} \\
&\quad \lesssim \beta\sqrt{hk} \times (\alpha^{-1/2}\beta d^{1/2}h + \alpha^{-1/2}\beta h k^{1/2} + d^{1/2}h^{1/2} + h^{1/2}k^{1/2}) \\
&\quad \lesssim \beta h\sqrt{k}\,(\sqrt{d} + \sqrt{k})\,,
\end{aligned}$$

provided that $h \le \alpha/\beta^2$. This concludes the proof.

■ 5.6.5 Conductance argument

In this section, we use the results from the previous sections in order to prove a lower bound on the s-conductance. The argument is similar to the proof of the standard conductance lemma (Lemma 5.4.3).

Towards the goal of applying the bound on the mixing time via s-conductance given in Corollary 5.6.2, we take $s := \varepsilon/(2M_0)$, and we choose the step size as in Proposition 5.6.7. Then, Proposition 5.6.7 guarantees the existence of an event E with probability $\pi(E) \ge 1 - c_0 s\sqrt{\alpha h}$ such that

$$\boldsymbol{x} \in E \implies \|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} \le \frac{1}{6}\,.$$

Let S be a measurable subset of $\mathbb{R}^d$ with $s \le \pi(S) \le 1/2$. Define the following subsets of $\mathbb{R}^d$:

$$S_1 := \Bigl\{\boldsymbol{x} \in S \;\Big|\; T(\boldsymbol{x}, S^{\mathsf{c}}) \le \frac{1}{4}\Bigr\}\,, \qquad \text{bad set 1}$$

$$S_2 \coloneqq \left\{ \boldsymbol{x} \in S^{\mathsf{c}} \;\middle|\; T(\boldsymbol{x}, S) \le \frac{1}{4} \right\}, \qquad \text{bad set 2}$$
$$S_3 \coloneqq (S_1 \cup S_2)^{\mathsf{c}}. \qquad \text{good set}$$

If $\pi(S_1) < \pi(S)/2$ or $\pi(S_2) < \pi(S^{\mathsf{c}})/2$, then we may conclude from reversibility of the MALA kernel T that

$$\begin{aligned}\int_S T(\boldsymbol{x}, S^{\mathsf{c}})\, \pi(\mathrm{d}\boldsymbol{x}) &= \frac{1}{2}\left(\int_S T(\boldsymbol{x}, S^{\mathsf{c}})\, \pi(\mathrm{d}\boldsymbol{x}) + \int_{S^{\mathsf{c}}} T(\boldsymbol{x}, S)\, \pi(\mathrm{d}\boldsymbol{x}) \right) \\ &\ge \frac{1}{2} \cdot \frac{\pi(S)}{2} \cdot \frac{1}{4} = \frac{\pi(S)}{16}.\end{aligned}$$

Therefore, for the purpose of proving a lower bound on the s-conductance, we may assume that $\pi(S_1) \wedge \pi(S_2) \ge \pi(S)/2$.

Now we consider $\boldsymbol{x} \in E \cap S_1$ and $\boldsymbol{y} \in E \cap S_2$. From the definitions of S_1 and S_2, it follows that

$$\|T_{\boldsymbol{x}} - T_{\boldsymbol{y}}\|_{\mathsf{TV}} \ge \frac{1}{2}.$$

Since $\boldsymbol{x}, \boldsymbol{y} \in E$, we also have

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} \wedge \|T_{\boldsymbol{y}} - Q_{\boldsymbol{y}}\|_{\mathsf{TV}} \le \frac{1}{6}.$$

Thus, using the decomposition (5.4),

$$\begin{aligned}\frac{1}{2} &\le \|T_{\boldsymbol{x}} - T_{\boldsymbol{y}}\|_{\mathsf{TV}} \le \|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} + \|Q_{\boldsymbol{x}} - Q_{\boldsymbol{y}}\|_{\mathsf{TV}} + \|T_{\boldsymbol{y}} - Q_{\boldsymbol{y}}\|_{\mathsf{TV}} \\ &\le \frac{1}{6} + \frac{\|\boldsymbol{x} - \boldsymbol{y}\|}{\sqrt{2h}} + \frac{1}{6},\end{aligned}$$

where the middle term is controlled via

$$\|Q_{\boldsymbol{x}} - Q_{\boldsymbol{y}}\|_{\mathsf{TV}} \le \frac{\|\boldsymbol{x} - \boldsymbol{y}\|}{\sqrt{2h}}, \qquad \text{if } h \le \frac{2}{\beta},$$

see [Dwi+19, Lemma 3]. Hence, we obtain:

$$\frac{\sqrt{2h}}{6} \le \|\boldsymbol{x} - \boldsymbol{y}\|,$$

which implies that $\operatorname{dist}(E \cap S_1, E \cap S_2) \ge \sqrt{2h}/6 =: r$. By the isoperimetric inequality (see Lemma 5.6.11), there is an absolute constant $c > 0$ such that

$$\pi\big(((E \cap S_1)^r)^{\mathsf{c}} \setminus (E \cap S_1)\big) \ge \frac{c\sqrt{2}}{6} \sqrt{\alpha h}\, \pi(E \cap S_1).$$

Since S_1, S_2, and S_3 partition $\mathbb{R}^d$, we see that the set on the left-hand side is contained in $E \cap S_2$.

$$\begin{aligned}\pi(S_3) + c_0 s\sqrt{\alpha h} \geq \pi(E \cap S_3) &\geq \frac{c\sqrt{2}}{6}\sqrt{\alpha h}\,\pi(E \cap S_1) \\ &\geq \frac{c\sqrt{2}}{6}\sqrt{\alpha h}\,\{\pi(S_1) - \pi(E^{\mathsf{c}})\} \\ &\geq \frac{c\sqrt{2}}{6}\sqrt{\alpha h}\,\Big\{\frac{\pi(S)}{2} - \pi(E^{\mathsf{c}})\Big\} \\ &\geq \frac{c\sqrt{2}}{12}\sqrt{\alpha h}\,\pi(S)\,, \qquad (5.16)\end{aligned}$$

where (5.16) follows since $\pi(S)/2 \geq s/2 \geq 2c_0 s\sqrt{\alpha h} \geq 2\pi(E^{\mathsf{c}})$ provided that $4c_0^2 h \leq 1/\alpha$.

Since $\pi(S) \geq s$, it follows that, provided we choose c_0 small enough (and thus, the constant c_1 in the step size (5.12) small enough), we obtain

$$\pi(S_3) \geq \frac{c\sqrt{2}}{24}\sqrt{\alpha h}\,\pi(S)\,.$$

From this,

$$\begin{aligned}\int_S T(\boldsymbol{x}, S^{\mathsf{c}})\,\pi(\mathrm{d}\boldsymbol{x}) &= \frac{1}{2}\Big(\int_S T(\boldsymbol{x}, S^{\mathsf{c}})\,\pi(\mathrm{d}\boldsymbol{x}) + \int_{S^{\mathsf{c}}} T(\boldsymbol{x}, S)\,\pi(\mathrm{d}\boldsymbol{x})\Big) \\ &\geq \frac{1}{2}\cdot\frac{1}{4}\cdot\pi(S_3) \geq \frac{c\sqrt{2}}{192}\sqrt{\alpha h}\,\pi(S)\,.\end{aligned}$$

Collecting the arguments, we obtain a lower bound on the s-conductance.

Proposition 5.6.10. *If the step size h is chosen as (5.12) for a sufficiently small constant c_1, then the s-conductance of the MALA chain satisfies*

$$\mathsf{C}_s \gtrsim \sqrt{\alpha h}\,.$$

Together with the mixing time bound in Corollary 5.6.2, we have proven Theorem 5.4.2.

5.6.6 Auxiliary lemmas

5.6.6.1 Standard facts about strongly log-concave measures

The following properties of strongly log-concave measures are well-known.

Lemma 5.6.11. *The α-strong convexity of V implies the following properties:*

1. *(moment and tail bounds) For $\boldsymbol{x} \sim \pi$, it holds that $\mathbb{E}[\|\boldsymbol{x}\|^2] \le d/\alpha$.*

   In fact, for all $k \ge 2$,

   $$\mathbb{E}\|\boldsymbol{x}\|^k \le \frac{3^k \left(d^{k/2} + k^{k/2}\right)}{\alpha^{k/2}} .$$

   Consequently, $\mathbb{E}\exp(\lambda \|\boldsymbol{x}\|^2)$ is bounded above by a universal constant, provided that $0 \le \lambda \le \alpha/(40d)$.

2. *(isoperimetry) For any $S \subseteq \mathbb{R}^d$ with $\pi(A) \le 1/2$, it holds that $\pi(S^\varepsilon \setminus S) \gtrsim \varepsilon\sqrt{\alpha}\,\pi(S)$, where*

   $$S^\varepsilon \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d \mid \exists y \in S \text{ with } \|\boldsymbol{x} - \boldsymbol{y}\| \le \varepsilon\} .$$

3. *(sub-Gaussian concentration) For any 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ and $\delta > 0$, with probability at least $1 - \delta$ it holds that*

   $$f(\boldsymbol{x}) - \mathbb{E}_\pi f \le \sqrt{\frac{2}{\alpha} \ln \frac{1}{\delta}} ,$$

   when $\boldsymbol{x} \sim \pi$.

Proof. The first statement is a simplification of [DKR22, Lemma 2]. For the second statement, in fact strongly log-concave measures satisfy a stronger isoperimetric inequality (sometimes called a Gaussian isoperimetric inequality, or a log-isoperimetric inequality in [Che+20a]); we refer to [BGL14, §8.5.2] and the monograph [BH97] which explains the relationship between integral form of the isoperimetric inequality employed here and the more traditional differential version. Finally, for the third statement, see Lemma 2.2.9. □

■ 5.6.6.2 Stochastic calculus results

Below, we also collect together some inequalities proven via stochastic calculus. In what follows, $(\bar{\boldsymbol{X}}_t)_{t\ge 0}$ is the Langevin diffusion (5.5), started at $\boldsymbol{x}$.

We use the following lemma, which follows from the proof of Lemma 3.6.22.

Lemma 5.6.12. *If $(\bar{\boldsymbol{X}}_t)_{t\ge 0}$ denotes the continuous-time Langevin process (5.5) started at $\boldsymbol{x}$, then for all $\lambda \ge 0$ and $h \lesssim 1/(\beta \vee \lambda)$, we have*

$$\ln \mathbb{E}\exp\big(\lambda \sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2\big) \lesssim (\beta^2 h^2 \|\boldsymbol{x}\|^2 + dh)\,\lambda .$$

In particular, for $t \lesssim 1/\beta$,

$$\mathbb{E}[\|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2] \lesssim \beta^2 t^2 \|\boldsymbol{x}\|^2 + dt .$$

Corollary 5.6.13. *Assume $h \lesssim 1/\beta$. There exists a numerical constant $C > 0$ such that for all $k \geq 1$,*

$$\mathbb{E} \sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^{2k} \leq C^k \left(\beta^{2k} h^{2k} \|\boldsymbol{x}\|^{2k} + d^k h^k + h^k k^k\right).$$

Proof. In Lemma 5.6.12, take $\lambda \asymp 1/h$ to yield

$$\ln \mathbb{E} \exp\left(\lambda \sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2\right) \lesssim \beta^2 h \|\boldsymbol{x}\|^2 + d\,.$$

The result now follows from standard moment bounds under sub-exponential concentration [see, e.g., Ver18, Proposition 2.7.1]. □

Remark 5.6.14. *Bounds such as the one in Corollary 5.6.13 are standard and have appeared in the literature before, e.g., [Mou+22].*

■ 5.6.7 From total variation to other distances

In this section, we deduce the mixing time results of Theorem 5.4.2 for the KL divergence, the chi-squared divergence, and the 2-Wasserstein distance.

We begin with the following lemma which shows that the warmness parameter (defined in Definition 5.4.1) is preserved by the iterations of MALA. In fact, this is true for all reversible Markov chains and is a consequence of the data-processing inequality (Lemma 2.2.19). We give a direct proof for completeness.

Lemma 5.6.15. *Let $(\mu_n)_{n\in\mathbb{N}}$ denote the iterates of a Markov chain whose kernel T is reversible with respect ot π, and assume that μ_0 is M_0-warm with respect to π. Then, for all $n \in \mathbb{N}$, the iterate μ_n is also M_0-warm with respect to π.*

Proof. The proof is by induction. For any $\boldsymbol{y} \in \mathbb{R}^d$,

$$\begin{aligned}\frac{\mu_{n+1}(\boldsymbol{y})}{\pi(\boldsymbol{y})} &= \int \frac{\mu_n(\boldsymbol{x})}{\pi(\boldsymbol{y})}\, T(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x} = \int \frac{\mu_n(\boldsymbol{x})}{\pi(\boldsymbol{x})}\,\frac{\pi(\boldsymbol{x})\,T(\boldsymbol{x},\boldsymbol{y})}{\pi(\boldsymbol{y})}\,\mathrm{d}\boldsymbol{x} \\ &\leq M_0 \int T(\boldsymbol{y},\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = M_0\,,\end{aligned}$$

where we use the inductive assumption and the reversibility of T. □

Under a warmness condition, the total variation distance controls the chi-squared divergence.

Lemma 5.6.16. *Let μ be M_0-warm with respect to π. Then,*

$$\chi^2(\mu \,\|\, \pi) \leq 2M_0 \|\mu - \pi\|_{\mathsf{TV}}\,.$$

Proof. From the definition of the chi-squared divergence,

$$\chi^2(\mu \,\|\, \pi) = \int \left|\frac{\mu}{\pi} - 1\right|^2 \mathrm{d}\pi \le M_0 \int \left|\frac{\mu}{\pi} - 1\right| \mathrm{d}\pi = 2M_0 \,\|\mu - \pi\|_{\mathsf{TV}} \,.$$

Here we use the fact that pointwise, $|\mu/\pi - 1| \le \max\{1, M_0 - 1\} \le M_0$. □

It immediately implies the following result on mixing times.

Corollary 5.6.17. *Fix $\varepsilon > 0$. Then, MALA initialized with a distribution μ_0 which is M_0-warm with respect to π satisfies the following mixing time bounds:*

$$\tau_{\mathrm{mix}}(\varepsilon, \mu_0; \mathsf{d}) \le \tau_{\mathrm{mix}}\big(\frac{\varepsilon^2}{2M_0}, \mu_0; \mathsf{TV}\big)$$

for each of the distances

$$\mathsf{d} \in \big\{\sqrt{\mathsf{KL}},\ \sqrt{\chi^2},\ \sqrt{\frac{\alpha}{2}}\, W_2\big\} \,.$$

Proof. The mixing time in the chi-squared distance is a straightforward consequence of Lemmas 5.6.15 and 5.6.16. The result for the KL divergence now follows since $\mathsf{KL} \le \chi^2$ [Tsy09, Lemma 2.7]. Finally, for the result in 2-Wasserstein distance we can use Talagrand's transport inequality

$$\frac{\alpha}{2}\, W_2^2(\mu, \pi) \le \mathsf{KL}(\mu \,\|\, \pi) \,, \qquad \text{for all probability measures } \mu \ll \pi \,,$$

which is a consequence of the strong convexity of V [in fact it is a consequence of the weaker assumption of a log-Sobolev inequality, see BGL14, Theorem 9.6.1]. □

Corollary 5.6.17 implies the remaining mixing time results in Theorem 5.4.2.

■ 5.7 Proof of the lower bound

This section presents the proofs of Theorems 5.5.1 and 5.5.2. The majority of this section is devoted to the proof of the upper bound on the conductance when $h \gg d^{-1/2}$ (Theorem 5.5.1). The proof of the upper bound on the spectral gap (Theorem 5.5.2) is given in §5.7.3.

5.7.1 High-level overview of the proof

Recall that we take $\eta = 1/4 - \delta$, where $\delta > 0$ is fixed throughout. As mentioned in §5.5, we consider the potential

$$V(\boldsymbol{x}) = \frac{\|\boldsymbol{x}\|^2}{2} - \frac{1}{2d^{2\eta}} \sum_{i=1}^{d} \cos(d^\eta x_i) \tag{5.17}$$

$$=: V_{\mathsf{G}}(\boldsymbol{x}) + V_{\mathsf{pert}}(\boldsymbol{x})\,. \tag{5.18}$$

From the construction, it immediately follows that V is $1/2$-strongly convex and $3/2$-smooth.

We begin with some intuition for the above construction. At a high level, our construction can be seen as a "perturbed" Gaussian distribution; V_{G} is the potential corresponding to a standard Gaussian and V_{pert} corresponds to a perturbation. Having this interpretation, we are interested in constructing a distribution (i) that is significantly different from the standard Gaussian, yet (ii) the difference is not noticed by each step of MALA.

(i) A quick calculation (see Lemma 5.7.8) shows that $\mathsf{KL}(\mathsf{normal}(0,1) \parallel \pi) = O(d^{1-4\eta})$. So, we must take $\eta \le 1/4$ to ensure that π is significantly different from the standard Gaussian.

(ii) On the other hand, V_{pert} is an oscillatory perturbation. Hence, MALA would not see the contribution from V_{pert} as long as its movement due to the Langevin proposal is at least as long as the length scale of the fluctuations of V_{pert}.

With this in mind, note that the fluctuations of V_{pert} is of order $d^{-\eta}$, while the movement of a single coordinate under the Langevin proposal is of order $\sqrt{h}$ (due to the Gaussian part). Hence, MALA would essentially ignore V_{pert} as long as $h \gg d^{-2\eta}$.

We formalize the above heuristic in the rest of this section.

To prove the upper bound on the conductance in Theorem 5.5.1, we use the following proposition.

Proposition 5.7.1. *Let E be an event such that $\pi(E) \ge 1/2$. Then,*

$$\mathsf{C} \le 2 \sup_{\boldsymbol{x} \in E} \int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y})\, A(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y}\,.$$

Proof. Let E_0 be a subset of E with $\pi(E_0) = 1/2$. From the definition of the conductance (C),

$$\mathsf{C} = \inf_{\substack{S \subseteq \mathbb{R}^d \\ \pi(S) \le 1/2}} \frac{\int_S T(\boldsymbol{x}, S^{\mathsf{c}})\, \pi(\mathrm{d}\boldsymbol{x})}{\pi(S)} \le 2 \int_{E_0} T(\boldsymbol{x}, E_0^{\mathsf{c}})\, \pi(\mathrm{d}\boldsymbol{x})$$

$$\leq 2 \int_{E_0} \Big(\int_{E_0^{\mathsf{c}}} Q(\boldsymbol{x}, \boldsymbol{y})\, A(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y} \Big)\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$$
$$\leq 2 \int_{E_0} \Big(\int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y})\, A(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y} \Big)\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$$
$$\leq 2 \sup_{\boldsymbol{x} \in E_0} \int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y})\, A(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y} \leq 2 \sup_{\boldsymbol{x} \in E} \int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y})\, A(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y}\,. \qquad \square$$

From Proposition 5.7.1, it therefore suffices to show that there is an event $E \subseteq \mathbb{R}^d$ with probability $\pi(E) \geq 1/2$ such that

$$\sup_{\boldsymbol{x} \in E} \int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y})\, A(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y} \leq \exp[-\Omega(d^{4\delta})]$$

By definition of the Metropolis–Hasting accept-reject step (5.1), we have

$$\begin{aligned} Q(\boldsymbol{x}, \boldsymbol{y})\, A(\boldsymbol{x}, \boldsymbol{y}) &= Q(\boldsymbol{x}, \boldsymbol{y}) \min\Big\{1, \frac{\pi(\boldsymbol{y})\, Q(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x})\, Q(\boldsymbol{x}, \boldsymbol{y})}\Big\} \\ &\leq \frac{\pi(\boldsymbol{y})\, Q(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x})} \\ &= \frac{1}{(4\pi h)^{d/2}} \exp\Big[V(\boldsymbol{x}) - V(\boldsymbol{y}) - \frac{\|\boldsymbol{y} - \boldsymbol{x} - h\nabla V(\boldsymbol{y})\|^2}{4h}\Big]\,. \end{aligned} \tag{5.19}$$

We substitute in the definition of our potential (5.17) and expand out the terms in (5.19), grouping them according to whether they involve V_{pert} or not:

$$(5.19) = \frac{1}{(4\pi h)^{d/2}} \exp\Big[\frac{1}{2}\|\boldsymbol{x}\|^2 - \frac{1}{2}\|\boldsymbol{y}\|^2 - \frac{1}{4h}\|(1-h)\,\boldsymbol{y} - \boldsymbol{x}\|^2\Big] \tag{5.20}$$
$$\times \exp\Big[V_{\mathsf{pert}}(\boldsymbol{x}) - V_{\mathsf{pert}}(\boldsymbol{y}) + \frac{1}{2}\langle (1-h)\,\boldsymbol{y} - \boldsymbol{x}, \nabla V_{\mathsf{pert}}(\boldsymbol{y})\rangle - \frac{h}{4}\|\nabla V_{\mathsf{pert}}(\boldsymbol{y})\|^2\Big]\,. \tag{5.21}$$

Some algebra yields that (5.20) is equal to

$$\underbrace{\Big(\frac{1+h^2}{4\pi h}\Big)^{d/2} \exp\Big[-\frac{1+h^2}{4h}\Big\|\boldsymbol{y} - \frac{1-h}{1+h^2}\,\boldsymbol{x}\Big\|^2\Big]}_{=:\mu_{\boldsymbol{x}}(\boldsymbol{y})} \frac{1}{(1+h^2)^{d/2}} \exp\Big[\frac{h^2\|\boldsymbol{x}\|^2}{2\,(1+h^2)}\Big]\,.$$

The first term, which we denote by $\mu_{\boldsymbol{x}}(\boldsymbol{y})$, is the probability density function of the distribution $\mathsf{normal}(\frac{1-h}{1+h^2}\,\boldsymbol{x}, \frac{2h}{1+h^2}\,I_d)$ evaluated at $\boldsymbol{y}$. Using this observation,

the quantity $\int_{\mathbb{R}^d} Q(\boldsymbol{x},\boldsymbol{y})\,A(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}$ is upper bounded by

$$\underbrace{\frac{\exp\left[\frac{h^2\,\|\boldsymbol{x}\|^2}{2\,(1+h^2)}+V_{\mathsf{pert}}(\boldsymbol{x})\right]}{(1+h^2)^{d/2}}}_{①} \times \underbrace{\mathbb{E}_{\boldsymbol{y}\sim\mu_{\boldsymbol{x}}}\exp\Big[-V_{\mathsf{pert}}(\boldsymbol{y})+\frac{1}{2}\,\langle(1-h)\,\boldsymbol{y}-\boldsymbol{x},\nabla V_{\mathsf{pert}}(\boldsymbol{y})\rangle-\frac{h}{4}\,\|\nabla V_{\mathsf{pert}}(\boldsymbol{y})\|^2\Big]}_{②}\,.$$

Having this upper bound, we will prove that there is a set $E\subseteq\mathbb{R}^d$ with $\pi(E)\geq 1/2$ such that the following bounds hold for all $\boldsymbol{x}\in E$:

1. (Lemma 5.7.5)

$$① \leq \exp\big[-\frac{1}{8}\,d^{1-4\eta}+o(d^{1-4\eta})\big]\,.$$

2. (Lemma 5.7.6)

$$② \leq \exp\big[\frac{1}{16}\,d^{1-4\eta}+o(d^{1-4\eta})\big]\,.$$

From these bounds and the preceding calculations, we have

$$\sup_{\boldsymbol{x}\in E}\int Q(\boldsymbol{x},\boldsymbol{y})\,A(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \leq \exp\big[-\frac{1}{8}\,d^{1-4\eta}+o(d^{1-4\eta})\big]\,.$$

This completes the proof of Theorem 5.5.1.

The next section is devoted to proving the two main bounds (Lemmas 5.7.5 and 5.7.6).

5.7.2 Proofs of technical statements

5.7.2.1 Notation and technical lemmas

We use the following notation:

$$\begin{cases} V_1(x) \coloneqq \frac{1}{2}\,x^2-\frac{1}{2}\,d^{-2\eta}\cos(d^\eta x)\,,\\ V(\boldsymbol{x}) \coloneqq \sum_{i=1}^d V_1(x_i)=\frac{1}{2}\,\|\boldsymbol{x}\|^2-\frac{1}{2}\,d^{-2\eta}\sum_{i=1}^d\cos(d^\eta x_i)\,,\\ \pi_1(x)\propto\exp(-V_1(x))\,,\\ \pi(\boldsymbol{x})\propto\exp(-V(\boldsymbol{x}))\,. \end{cases} \tag{5.22}$$

Thus, π_1 is the marginal distribution of π. We first list useful technical lemmas for proving Lemmas 5.7.5 and 5.7.6. First, the following trigonometric inequality will be used several times.

Lemma 5.7.2. *Let $\xi \sim \mathsf{normal}(0,1)$, let p be a polynomial, and let $a, b \in \mathbb{R}$, $\gamma > 0$ be constants. Then, there exists $C > 0$ (depending on p, a, b, and γ) such that*

$$|\mathbb{E}[p(\xi)\sin(a + bd^\gamma \xi)]| \le \frac{C}{d}\,.$$

Proof. The key fact we use is that the characteristic function $\mathbb{E}[e^{\mathsf{i}t\xi}]$ of a Gaussian is equal to $\exp(-\frac{1}{2}t^2)$. First consider the case $p \equiv 1$. Let $\mathsf{im}(\cdot)$ denote the imaginary part. Then, we have

$$\begin{aligned}
\mathbb{E}\sin(a + bd^\gamma\xi) &= \mathbb{E}\,\mathsf{im}\exp\big(\mathsf{i}\,(a + bd^\gamma\xi)\big) \\
&= \mathsf{im}\big(\exp(\mathsf{i}a)\,\mathbb{E}\exp(\mathsf{i}bd^\gamma\xi)\big) \\
&= \mathsf{im}\Big(\exp\big(\mathsf{i}a - \frac{b^2d^{2\gamma}}{2}\big)\Big) \\
&= \sin(a)\exp\big(-\frac{b^2d^{2\gamma}}{2}\big)\,.
\end{aligned}$$

It is then clear that the result holds for $p = 1$. Next, when $p(x) = x^\ell$ for some $\ell \in \mathbb{N}^+$,

$$\begin{aligned}
\mathbb{E}[\xi^\ell \sin(a + bd^\gamma\xi)] &= \mathsf{im}\big(\exp(\mathsf{i}a)\,\mathbb{E}[\xi^\ell\exp(\mathsf{i}bd^\gamma\xi)]\big) \\
&= \mathsf{im}\Big(\exp(\mathsf{i}a)\,\mathsf{i}^{-\ell}\,\mathbb{E}\Big[\frac{\mathrm{d}^\ell}{\mathrm{d}t^\ell}\exp(\mathsf{i}t\xi)\Big|_{t=bd^\gamma}\Big]\Big) \\
&= \mathsf{im}\Big(\exp(\mathsf{i}a)\,\mathsf{i}^{-\ell}\,\frac{\mathrm{d}^\ell}{\mathrm{d}t^\ell}\exp\big(-\frac{t^2}{2}\big)\Big|_{t=bd^\gamma}\Big)\,.
\end{aligned}$$

Thus, it is clear that the lemma holds for this choice of p too. The case of a general polynomial follows from linearity. □

Clearly, the statement of the previous lemma can be substantially strengthened, but this will not be necessary for the MALA lower bound.

Now we list some useful facts about the adversarial target distribution.

Lemma 5.7.3. *Assume $\eta < 1/4$. The following hold for π_1 and π defined in (5.22):*

(a) Let $Z := \int_{\mathbb{R}} \exp(-V_1(\boldsymbol{x}))\,\mathrm{d}x$ be the one-dimensional normalizing constant. Then, we have $Z = \sqrt{2\pi} + O(d^{-4\eta})$.

(b) $\mathbb{E}_{x\sim\pi_1}[x^2] \le 1 + O(d^{-4\eta})$. *Consequently,* $\mathbb{E}_{\boldsymbol{x}\sim\pi}[\|\boldsymbol{x}\|^2] \le d + O(d^{1-4\eta})$.

(c) $\mathbb{E}_{x\sim\pi_1}[\cos(d^\eta x)] \le \frac{1}{4}\, d^{-2\eta} + O(d^{-6\eta})$.

Proof. (a) Letting $\xi \sim \mathsf{normal}(0,1)$, then

$$\begin{aligned} Z - \sqrt{2\pi} &= \int_{\mathbb{R}} \exp\Big(-\frac{1}{2}\, x^2 + \frac{1}{2d^{2\eta}} \cos(d^\eta x)\Big)\, \mathrm{d}x - \sqrt{2\pi} \\ &= \sqrt{2\pi} \int_{\mathbb{R}} \exp\Big(\frac{1}{2d^{2\eta}} \cos(d^\eta x)\Big)\, \frac{\exp(-\frac{1}{2}\, x^2)}{\sqrt{2\pi}}\, \mathrm{d}x - \sqrt{2\pi} \\ &= \sqrt{2\pi} \left(\mathbb{E} \exp\big(\frac{1}{2d^{2\eta}}\, \cos(d^\eta \xi)\big) - 1\right) \\ &= \frac{\sqrt{2\pi}}{2d^{2\eta}}\, \mathbb{E} \cos(d^\eta \xi) + O(d^{-4\eta})\,. \end{aligned}$$

By Lemma 5.7.2, we have $|\mathbb{E}\cos(d^\eta\xi)| = O(d^{-1}) = o(d^{-4\eta})$, since $\eta < 1/4$. The proof of (a) then follows.

(*b*) Similarly, letting $\xi \sim \mathsf{normal}(0,1)$,

$$\begin{aligned} \mathbb{E}_{x\sim\pi_1}[x^2] &= \int x^2\, \frac{\exp(-V_1(\boldsymbol{x}))}{Z}\, \mathrm{d}x \\ &= \frac{\sqrt{2\pi}}{Z}\, \mathbb{E}\big[\xi^2 \exp\big(\frac{1}{2d^{2\eta}} \cos(d^\eta\xi)\big)\big] \\ &= \big(1 + O(d^{-4\eta})\big)\, \mathbb{E}\big[\xi^2 \exp\big(\frac{1}{2d^{2\eta}} \cos(d^\eta\xi)\big)\big]\,. \end{aligned}$$

By Taylor expansion,

$$\mathbb{E}\big[\xi^2 \exp\big(\frac{1}{2d^{2\eta}} \cos(d^\eta\xi)\big)\big] = 1 + \frac{1}{2d^{2\eta}}\, \mathbb{E}[\xi^2 \cos(d^\eta\xi)] + O(d^{-4\eta})\,.$$

Again by Lemma 5.7.2, the second term is $O(d^{-(2\eta+1)}) = o(d^{-6\eta})$. Hence, the result follows.

(*c*) Similarly, it holds that

$$\begin{aligned} \mathbb{E}_{x\sim\pi_1} \cos(d^\eta x) &= \frac{\sqrt{2\pi}}{Z}\, \mathbb{E}\big[\cos(d^\eta\xi) \exp\big(\frac{1}{2d^{2\eta}} \cos(d^\eta\xi)\big)\big] \\ &= \big(1 + O(d^{-4\eta})\big) \Big[\mathbb{E}\cos(d^\eta\xi) + \frac{1}{2d^{2\eta}}\, \mathbb{E}\cos^2(d^\eta\xi) + O(d^{-4\eta})\Big]\,. \end{aligned}$$

By Lemma 5.7.2, the first term is $\mathbb{E}\cos(d^\eta \xi) = o(d^{-4\eta})$. Next, the second term can be written

$$\frac{1}{2d^{2\eta}}\,\mathbb{E}\cos^2(d^\eta\xi) = \frac{1}{4d^{2\eta}} + \frac{1}{4d^{2\eta}}\,\mathbb{E}\cos(2d^\eta\xi)\,.$$

From Lemma 5.7.2, $\mathbb{E}\cos(2d^\eta\xi) = o(d^{-4\eta})$. Therefore, the result follows. □

Lemma 5.7.4. *For $\boldsymbol{x} \sim \pi$, the following holds with probability at least $1 - 1/(4d)$:*

$$\|\boldsymbol{x}\|_\infty < 4\sqrt{\ln(8d)}\,.$$

Proof. By symmetry, we just need to show that with probability at least $1-1/(8d)$,

$$\max_{i\in[d]} x_i < 4\sqrt{\ln d}\,.$$

Since $V_1'' \geq 1/2$, each $|x_i|$ will be stochastically dominated by $|\xi|$, where $\xi \sim \mathsf{normal}(0, 2)$. Hence, if $\xi_1, \ldots, \xi_d$ are i.i.d. copies of ξ, we just need to show that

$$\max_{i\in[d]} \xi_i < 4\sqrt{\ln d}$$

with probability at least $1 - 1/d$. The standard argument based on the moment generating function (e.g., [Han16, Lemma 5.1]) tells us that $\mathbb{E}[\max_{i\in[d]} \xi_i] \leq 2\sqrt{\ln d}$, and Gaussian concentration (e.g., [Han16, Theorem 3.25]) implies

$$\mathbb{P}\big(\max_{i\in[d]} \xi_i > \mathbb{E}\max_{i\in[d]} \xi_i + t\big) \leq \exp\big(-\frac{t^2}{4}\big)\,.$$

Plug in $t = 2\sqrt{\ln(8d)}$ and we get the lemma as claimed. □

Now let us state and prove the technical statements in order.

■ 5.7.2.2 Proof of Lemma 5.7.5

Lemma 5.7.5. *Assume that $0 < h \leq d^{-1/3}$. Then there exists an event E_1 with $\pi(E_1) \geq 3/4$ such that for $\boldsymbol{x} \in E_1$,*

$$\frac{\exp\Big[\frac{h^2\,\|\boldsymbol{x}\|^2}{2\,(1+h^2)} + V_{\mathsf{pert}}(\boldsymbol{x})\Big]}{(1+h^2)^{d/2}} \leq \exp\big[-\frac{1}{8}\,d^{1-4\eta} + o(d^{1-4\eta})\big]\,.$$

Proof. We decompose the left-hand side as

$$\frac{\exp\left[\frac{h^2\,\|\boldsymbol{x}\|^2}{2\,(1+h^2)} + V_{\mathsf{pert}}(\boldsymbol{x})\right]}{(1+h^2)^{d/2}} = \frac{1}{(1+h^2)^{d/2}} \exp\left[\frac{h^2\,\|\boldsymbol{x}\|^2}{2\,(1+h^2)}\right] \times \exp[V_{\mathsf{pert}}(\boldsymbol{x})]$$

and bound each term separately.

We begin with the first term. By Lemma 5.7.3(b), we know that the second moment of π is $d + O(d^{1-4\eta})$. Since π is $1/2$-strongly log concave, a standard concentration argument (see, e.g., Lemma 5.6.11) shows that there exists a subset E_1' with $\pi(E_1') \geq 7/8$ such that for $\boldsymbol{x} \in E_1'$,

$$\|\boldsymbol{x}\|^2 \leq\ d + O(d^{1-4\eta}) + O(d^{1/2})\,.$$

Now, using the fact that $\ln(1+x) \geq x - x^2/2$ for $x \geq 0$,

$$\begin{aligned}
\frac{1}{(1+h^2)^{d/2}} \exp\left[\frac{h^2\,\|\boldsymbol{x}\|^2}{2\,(1+h^2)}\right] &\leq \exp\left[\frac{h^2\,(d + O(d^{1-4\eta}) + O(d^{1/2}))}{2\,(1+h^2)} - \frac{d}{2}\ln(1+h^2)\right] \\
&\leq \exp\left[\frac{h^2\,(d + O(d^{1-4\eta}) + O(d^{1/2}))}{2\,(1+h^2)} - \frac{dh^2}{2} + \frac{dh^4}{4}\right] \\
&= \exp\left[\frac{h^2\,(O(d^{1-4\eta}) + O(d^{1/2}))}{2\,(1+h^2)} - \frac{dh^4}{2\,(1+h^2)} + \frac{dh^4}{4}\right] \\
&= \exp\left[\frac{h^2\,(O(d^{1-4\eta}) + O(d^{1/2}))}{2\,(1+h^2)} + \frac{-dh^4 + 2dh^6}{4\,(1+h^2)}\right] \\
&\leq \exp[O(d^{1-4\eta}h^2) + O(d^{1/2}h^2)]\,,
\end{aligned}$$

where the last line follows since $h^2 \leq 1/2$. In order to show that the exponent of the above term is $o(d^{1-4\eta})$, we must check that $d^{1/2}h^2 = o(d^{1-4\eta})$, which holds if $h = o(d^{1/4-2\eta}) = o(d^{-1/4+2\delta})$. This indeed follows from our assumption that $h \leq d^{-1/3}$.

Next, we consider the second term. Recall from the calculation in the proof of Lemma 5.7.3(c) that $\mathbb{E}_{x\sim\pi_1}[\cos(d^\eta x)] \leq \frac{1}{4}\,d^{-2\eta} + O(d^{-6\eta})$. Hence, it follows that

$$\mathbb{E}_{\boldsymbol{x}\sim\pi}[V_{\mathsf{pert}}(\boldsymbol{x})] = -\frac{1}{2d^{2\eta}} \sum_{i=1}^{d} \mathbb{E}_{x_i\sim\pi_1} \cos(d^\eta x_i) = -\frac{1}{8}\,d^{1-4\eta} + O(d^{1-8\eta})\,.$$

Since π is $1/2$-strongly log-concave, another sub-Gaussian concentration argument (Lemma 5.6.11) shows that there exists a subset E_1'' with $\pi(E_1'') \geq 7/8$ such that for $\boldsymbol{x} \in E_1''$,

$$\exp[V_{\mathsf{pert}}(\boldsymbol{x})] \leq \exp\left[-\frac{1}{8}\,d^{1-4\eta} + O(d^{1-8\eta}) + O(d^{1/2-2\eta})\right]$$

$$\leq \exp\Big[-\frac{1}{8}\, d^{1-4\eta} + o(d^{1-4\eta})\Big]\,,$$

since $1 - 4\eta > 0$ by the hypothesis.

Now taking $E_1 \coloneqq E_1' \cap E_1''$, the above calculations show that for $\boldsymbol{x} \in E_1$,

$$\frac{\exp\Big[\frac{h^2\,\|\boldsymbol{x}\|^2}{2\,(1+h^2)} + V_{\mathsf{pert}}(\boldsymbol{x})\Big]}{(1+h^2)^{d/2}} \leq \exp\Big[-\frac{1}{8}\, d^{1-4\eta} + o(d^{1-4\eta})\Big]\,,$$

which completes the proof. □

■ 5.7.2.3 Proof of Lemma 5.7.6

Lemma 5.7.6. *Assume that $h \in [d^{-\frac{1}{2}+3\delta}, d^{-\frac{1}{3}}]$. Then there exists an event E_2 with $\pi(E_2) \geq 3/4$ such that for $\boldsymbol{x} \in E_2$,*

$$\mathbb{E}_{\boldsymbol{y}\sim\mu_{\boldsymbol{x}}} \exp\Big[-V_{\mathsf{pert}}(\boldsymbol{y}) + \frac{1}{2}\,\langle (1-h)\,\boldsymbol{y} - \boldsymbol{x}, \nabla V_{\mathsf{pert}}(\boldsymbol{y})\rangle - \frac{h}{4}\,\|\nabla V_{\mathsf{pert}}(\boldsymbol{y})\|^2\Big] \leq \exp\Big[\frac{1}{16}\, d^{1-4\eta} + o(d^{1-4\eta})\Big]\,.$$

Proof. Recall the definition $V_{\mathsf{pert}}(\boldsymbol{x}) = -\frac{1}{2}\, d^{-2\eta} \sum_{i=1}^d \cos(d^\eta x_i)$. Since V_{pert} is separable, it suffices to consider the following quantity: for $\mu_{x_i} \coloneqq \mathsf{normal}(\frac{1-h}{1+h^2}\, x_i, \frac{2h}{1+h^2})$,

$$\max_{i\in[d]} \mathbb{E}_{y_i\sim\mu_{x_i}} \exp\Big(\frac{\cos(d^\eta y_i)}{2d^{2\eta}} + \frac{((1-h)\,y_i - x_i)\sin(d^\eta y_i)}{4d^\eta} - \frac{h\sin^2(d^\eta y_i)}{16 d^{2\eta}}\Big)\,. \tag{5.23}$$

Indeed, the lemma is proved as soon as we show

$$(5.23) \leq \exp\Big[\frac{1}{16}\, d^{-4\eta} + o(d^{-4\eta})\Big]\,. \tag{5.24}$$

For the proof, we will therefore work with a single coordinate; for simplicity of notation, we will use the first coordinate.

To prove the inequality (5.24), let us first simplify the expression (5.23). Letting $\xi \sim \mathsf{normal}(0,1)$, we can equivalently write $y_1 = \frac{1-h}{1+h^2}\, x_1 + \sqrt{\frac{2h}{1+h^2}}\, \xi$. From this, we get

$$(1-h)\,y_1 - x_1 = -\frac{2h}{1+h^2}\, x_1 + (1-h)\sqrt{\frac{2h}{1+h^2}}\, \xi\,.$$

Since our regime of interest is $h = o(1)$, we simplify the notation by defining

$$\bar{h} \coloneqq \frac{h}{1+h^2} \qquad \text{and} \qquad \tilde{h} \coloneqq \frac{(1-h)^2}{1+h^2}\, h\,,$$

and treat them as being on the same order as h. Using these simplifying notations and rearranging, we are left to consider

$$\mathbb{E}\exp\Big(\underbrace{\frac{\cos(d^\eta y_1)}{2d^{2\eta}}}_{=:\Delta_1}-\underbrace{\frac{h\sin^2(d^\eta y_1)}{16d^{2\eta}}}_{=:\Delta_2}-\underbrace{\frac{2\bar{h}x_1\sin(d^\eta y_1)}{4d^\eta}}_{=:\Delta_3}+\underbrace{\frac{\sqrt{2\tilde{h}}\xi\sin(d^\eta y_1)}{4d^\eta}}_{=:\Delta_4}\Big)\,, \tag{5.25}$$

where $y_1=\frac{1-h}{1+h^2}\,x_1+\sqrt{\frac{2h}{1+h^2}}\,\xi$. Now we will estimate (5.25) by a Taylor expansion.

Throughout, we will assume $\|\boldsymbol{x}\|_\infty\le 4\sqrt{\ln(8d)}$. By Lemma 5.7.4, this holds on an event E_2 of probability $\pi(E_2)\ge 3/4$. From this, we note the bounds

$$|\Delta_1|=O(d^{-2\eta})\,,\quad |\Delta_2|=O(d^{-2\eta}h)\,,\quad |\Delta_3|=\widetilde{O}(d^{-\eta}h)\,,\quad |\Delta_4|=O_{\mathsf{p}}(d^{-\eta}\sqrt{h})\,.$$

Here, O_{p} denotes probabilistic big-O notation. Using $h=O(d^{-1/3})=o(d^{-4\eta/3})$, we have

$$\begin{aligned}|\Delta_1|&=O(d^{-2\eta})\,,\\ |\Delta_2|&=o(d^{-(3+1/3)\eta})\,,\\ |\Delta_3|&=o(d^{-(2+1/3)\eta})\,,\\ |\Delta_4|&=o_{\mathsf{p}}(d^{-(1+2/3)\eta})\,.\end{aligned} \tag{5.26}$$

From, this, we see that the third- or higher-order terms in the Taylor expansion, after taking the expectation, are $o(d^{-5\eta})$. Indeed, the dominant term is the term $\mathbb{E}[|\Delta_4|^3]=o(d^{-5\eta})$.

We also note that the common argument of the trigonometric terms is

$$d^\eta y_1=d^\eta\,\frac{1-h}{1+h^2}\,x_1+d^\eta\sqrt{\frac{2h}{1+h^2}}\,\xi\,,$$

so the coefficient in front of ξ is of order $d^\eta\sqrt{h}=\Omega(d^{\delta/2})$ by the assumption $h\ge d^{-\frac12+3\delta}$. Thus, the trigonometric terms precisely fit into the setting of Lemma 5.7.2, and we will apply Lemma 5.7.2 to estimate these terms.

Now let us estimate the terms of order one and two.

- *First- and lower-order terms.* We have

$$(\le 1\text{st order})=1+\mathbb{E}\,\Delta_1-\mathbb{E}\,\Delta_2-\mathbb{E}\,\Delta_3+\mathbb{E}\,\Delta_4\,.$$

By Lemma 5.7.2, we know $\mathbb{E}\,\Delta_1=O(d^{-1-2\eta})=o(d^{-6\eta})$. For $\mathbb{E}\,\Delta_2$, we have

$$-\mathbb{E}\,\Delta_2=-\frac{h}{32d^{2\eta}}+\frac{h}{32d^{2\eta}}\,\mathbb{E}\cos(2d^\eta y_1)=-\frac{h}{32d^{2\eta}}+o(d^{-6\eta})\,,$$

where we use Lemma 5.7.2 again. For $\mathbb{E}\,\Delta_3$, we have

$$-\mathbb{E}\,\Delta_3 = -\mathbb{E}\,\frac{2\bar{h}x_1\sin(d^\eta y_1)}{4d^\eta} = \widetilde{O}(d^{-(1+\eta)}h) = o(d^{-5\eta})\,,$$

where the last line is due to Lemmas 5.7.2 and 5.7.4. For $\mathbb{E}\,\Delta_4$, we have

$$\mathbb{E}\,\Delta_4 = \mathbb{E}\,\frac{\sqrt{2\tilde{h}}\xi\sin(d^\eta y_1)}{4d^\eta} = O(d^{-(1+\eta)}\sqrt{h}) = o(d^{-5\eta})\,,$$

where we use Lemma 5.7.2. Collecting together the terms, we have

$$(\le \text{1st order}) = 1 - \frac{h}{32d^{2\eta}} + o(d^{-5\eta})\,. \tag{5.27}$$

- *Second-order terms.* For the reader's convenience, we have organized the terms which appear in the second-order Taylor expansion as Table 5.1.

	$O(d^{-2\eta})$	$o(d^{-(3+1/3)\eta})$	$o(d^{-(2+1/3)\eta})$	$o_{\mathsf{p}}(d^{-(1+2/3)\eta})$
$O(d^{-2\eta})$	(5.28)	$o(d^{-4\eta})$	$o(d^{-4\eta})$	(5.29)
$o(d^{-(3+1/3)\eta})$		$o(d^{-4\eta})$	$o(d^{-4\eta})$	$o_{\mathsf{p}}(d^{-4\eta})$
$o(d^{-(2+1/3)\eta})$			$o(d^{-4\eta})$	$o_{\mathsf{p}}(d^{-4\eta})$
$o_{\mathsf{p}}(d^{-(1+2/3)\eta})$				(5.30)

Table 5.1: Terms which appear in the second-order Taylor expansion. The rows and columns are indexed by the terms Δ_1, Δ_2, Δ_3, Δ_4; refer to (5.26).

We now estimate the terms which are not covered by the table. Let us estimate the remaining terms one by one. First, by Lemma 5.7.2,

$$\frac{1}{2}\,\mathbb{E}[\Delta_1^2] = \mathbb{E}\,\frac{\cos^2(d^\eta y_1)}{8d^{4\eta}} = \frac{1}{16d^{4\eta}} + \mathbb{E}\,\frac{\cos(2d^\eta y_1)}{16d^{4\eta}} = \frac{1}{16d^{4\eta}} + o(d^{-8\eta})\,. \tag{5.28}$$

Next, by Lemma 5.7.2,

$$\mathbb{E}[\Delta_1\Delta_4] = \mathbb{E}\big[\frac{\sqrt{2\tilde{h}}\xi}{8d^{3\eta}}\cos(d^\eta y_1)\sin(d^\eta y_1)\big] = \frac{\sqrt{2\tilde{h}}}{16d^{3\eta}}\,\mathbb{E}[\xi\sin(2d^\eta y_1)] = o(d^{-7\eta})\,. \tag{5.29}$$

Lastly, invoking Lemma 5.7.2 yet again,

$$\frac{1}{2}\,\mathbb{E}[\Delta_4^2] = \mathbb{E}\,\frac{\tilde{h}\xi^2\sin^2(d^\eta y_1)}{16d^{2\eta}} = \mathbb{E}\,\frac{\tilde{h}\xi^2}{32d^{2\eta}} - \mathbb{E}\,\frac{\tilde{h}\xi^2\cos(2d^\eta y_1)}{32d^{2\eta}}$$

$$= \frac{\tilde{h}}{32d^{2\eta}} + o(d^{-6\eta})\,. \tag{5.30}$$

Combining all together, we obtain,

$$(\text{2nd order}) = \frac{1}{16d^{4\eta}} + \frac{\tilde{h}}{32d^{2\eta}} + o(d^{-4\eta})\,. \tag{5.31}$$

Therefore, we combine (5.27) and (5.31) to conclude

$$\begin{aligned}(5.25) &\le \exp\Big[\frac{1}{16}\,d^{-4\eta} - \frac{h}{32d^{2\eta}} + \frac{\tilde{h}}{32d^{2\eta}} + o(d^{-4\eta})\Big] \\ &= \exp\Big[\frac{1}{16}\,d^{-4\eta} + o(d^{-4\eta})\Big]\,,\end{aligned}$$

where the last line follows from $\tilde{h} - h = \frac{(1-h)^2}{1+h^2}\,h - h \le 0$. This implies (5.24), and hence the proof is complete. □

■ 5.7.3 Upper bound on the spectral gap

Note that when $\eta < 1/4$, the adversarial potential defined in (5.22) satisfies the assumptions of the following theorem, as a consequence of our computation in Lemma 5.7.3.

Theorem 5.7.7. *Consider a potential $V : \mathbb{R}^d \to \mathbb{R}$ which is separable: $V(\boldsymbol{x}) = \sum_{i=1}^d v(x_i)$ for a function $v : \mathbb{R} \to \mathbb{R}$. Assume that:*

- *V is symmetric about the origin, and $V(\mathbf{0}) = \min V$.*
- *V is $O(1)$-smooth.*
- *For the distribution $\pi_1 \propto \exp(-v)$, we have $\mathbb{E}_{x\sim\pi_1}[x^2] \asymp 1$.*

Then, spectral gap of MALA with target distribution $\pi \propto \exp(-V)$ and step size $h \le 1$ satisfies

$$\lambda \lesssim h\,.$$

Proof. Consider the function $f : \mathbb{R}^d \to \mathbb{R}$ given by $f(\boldsymbol{x}) \coloneqq x_1$. Since V is symmetric about the origin, we have $\mathbb{E}_\pi f = 0$.

From the definition the spectral gap (λ),

$$\lambda \le \frac{\mathbb{E}_\pi[f\,(\mathrm{id} - T)\,f]}{\mathbb{E}_\pi[f^2]} \lesssim \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}\sim\pi \\ \boldsymbol{y}\sim T(\boldsymbol{x},\cdot)}} [(x_1 - y_1)^2]\,.$$

Next, using the definition of the MALA kernel T, if ξ is a standard Gaussian random variable, then

$$\begin{aligned}\mathop{\mathbb{E}}_{\substack{\boldsymbol{x}\sim\pi\\ \boldsymbol{y}\sim T(\boldsymbol{x},\cdot)}}[(x_1-y_1)^2] &= \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}\sim\pi\\ \boldsymbol{y}\sim Q(\boldsymbol{x},\cdot)}}[(x_1-y_1)^2\,\mathbb{1}_{\text{proposal }\boldsymbol{x}\to\boldsymbol{y}\text{ is accepted}}]\\ &\le \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}\sim\pi\\ \boldsymbol{y}\sim Q(\boldsymbol{x},\cdot)}}[(x_1-y_1)^2] = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\pi}[\{hv'(x_1)-\sqrt{2h}\xi\}^2]\\ &\le 2h^2\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\pi}[v'(x_1)^2]+4h\,\mathbb{E}[\xi^2]\lesssim h^2\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\pi}[x_1^2]+h\lesssim h\,,\end{aligned}$$

by our assumptions. This completes the proof. □

5.7.4 Auxiliary lemmas

Lemma 5.7.8. *Let $\gamma \coloneqq \mathsf{normal}(0, I_d)$ and let π be the adversarial target distribution defined in (5.22). Then,*

$$\mathsf{KL}(\gamma \,\|\, \pi) \le O(d^{1-4\eta})\,.$$

Proof. From the definition of the KL divergence, if $\xi_1, \dots, \xi_d$ are i.i.d. random variables drawn according to γ, then

$$\begin{aligned}\mathsf{KL}(\gamma \,\|\, \pi) &= \int \gamma(\boldsymbol{x}) \ln\Big(\frac{Z^d}{(2\pi)^{d/2}}\exp V_{\mathsf{pert}}(\boldsymbol{x})\Big)\,\mathrm{d}\boldsymbol{x}\\ &= d\ln\frac{Z}{\sqrt{2\pi}} - \frac{1}{2d^{2\eta}}\sum_{i=1}^{d}\mathbb{E}\cos(d^\eta\xi_i)\,.\end{aligned}$$

From our estimate of the normalizing constant in Lemma 5.7.3,

$$d\ln\frac{Z}{\sqrt{2\pi}} = d\ln\big(1+O(d^{-4\eta})\big) = O(d^{1-4\eta})\,.$$

On the other hand, from the proof of Lemma 5.7.2,

$$-\frac{1}{2d^{2\eta}}\sum_{i=1}^{d}\mathbb{E}\cos(d^\eta\xi_i) = o(d^{1-4\eta})\,.$$

The result follows. □

5.8 Calculations for a Gaussian target distribution

In this section, we provide calculations for MALA when the target distribution π is the standard Gaussian. Since MALA applied to the Gaussian distribution has a scaling limit in the sense of [RR98], one would expect the mixing time of the Gaussian distribution to be of order $d^{1/3}$, and that is indeed what we show below.

5.8.1 Upper bound

First, we show that, under a warm start, the mixing time of MALA applied to the standard Gaussian mixes at $O(d^{1/3})$ rate.

Proposition 5.8.1. *Let $\varepsilon > 0$, and let the target distribution π be the standard Gaussian on $\mathbb{R}^d$. For a step size $h = cd^{-1/3}$, where $c > 0$ is a small constant, and an initial distribution μ_0 that is M_0-warm with respect to π such that $\log \frac{M_0}{\varepsilon h} = O(d^{1/3})$, the mixing time of MALA satisfies*

$$\tau_{\mathrm{mix}}(\varepsilon, \mu_0; \mathsf{TV}) \lesssim d^{1/3} \log\Big(\frac{M_0}{\varepsilon}\Big) .$$

Using the results of §5.6.7, the mixing time bounds can then be extended to the KL divergence, the chi-squared divergence, and the 2-Wasserstein distance.

The proof crucially relies on the fact that when $h \approx d^{-1/3}$, the acceptance probability $A(\boldsymbol{x})$ (see (5.2)) when $\boldsymbol{x} \sim \pi$ is of order $\Omega(1)$ with high probability, which is formalized below.

Lemma 5.8.2. *Let π be the standard Gaussian. For $h = c_0 d^{-1/3}$, where $c_0 > 0$ is sufficiently small, and $\boldsymbol{x} \sim \pi$, there exists $c_1 > 0$ such that with probability at least $1 - 2\exp(-c_1 d^{1/3})$, it holds that $A(\boldsymbol{x}) \geq 5/6$.*

Proof of Proposition 5.8.1. We sketch the proof, following the s-conductance mixing time strategy outlined in §5.6.1. Let $E := \{\boldsymbol{x} \in \mathbb{R}^d \mid A(\boldsymbol{x}) \geq 5/6\}$. Lemma 5.8.2 guarantees that $\pi(E) \geq 1 - 2\exp(-c_1 d^{1/3})$. By our assumption, we have $\log(\varepsilon h/M_0) = \Omega(d^{-1/3})$, so $\pi(E) \geq 1 - c'\sqrt{h}s$ for some constant $c' > 0$, where $s := \varepsilon/(2M_0)$. Moreover, on the event E we have (by Proposition 5.6.3)

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} = 1 - A(\boldsymbol{x}) \leq \frac{1}{6} .$$

The argument in the proof of Proposition 5.6.10 implies that the s-conductance, defined in (5.9), is lower bounded by $\mathsf{C}_s \gtrsim \sqrt{h}$, and Corollary 5.6.2 gives the desired mixing time bound. □

Proof of Lemma 5.8.2. Let $\boldsymbol{x} \sim \pi$ and $\boldsymbol{y} \sim Q(\boldsymbol{x}, \cdot)$. We will use c to denote universal constants, which can change from line to line. First note that by concentration of the norm [Ver18, Theorem 3.1.1], we have that for all $t > 0$,

$$\mathbb{P}\big(\big|\, \|\boldsymbol{x}\| - \sqrt{d}\,\big| > t\big) \leq 2\exp(-ct^2) .$$

As a result, the event

$$E_1 := \big\{\big|\, \|\boldsymbol{x}\| - \sqrt{d}\,\big| \leq t_1\big\}$$

holds with probability at least $1 - 2\exp(-ct_1^2)$.

By the radial symmetry of the standard Gaussian, we can assume that the only non-zero coordinate of $\boldsymbol{x}$ is the first coordinate: $\boldsymbol{x} = (x_1, 0, \dots, 0)$. Given $\boldsymbol{x}$, we draw $\boldsymbol{y}$ by:

$$\boldsymbol{y} = (1-h)\,\boldsymbol{x} + \sqrt{2h}\,\boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(0, I_d)\,.$$

We can write $\boldsymbol{\xi} = (\xi_1, \boldsymbol{\xi}_{-1})$, where $\xi_1 \sim \mathcal{N}(0,1)$, and $\boldsymbol{\xi}_{-1} \sim \mathcal{N}(0, I_{d-1})$. By Gaussian concentration, the event

$$E_2 \coloneqq \{|\xi_1| \le t_2\}$$

holds with probability at least $1 - 2\exp(-ct_2^2)$, and the event

$$E_3 \coloneqq \left\{ \left| \, \|\boldsymbol{\xi}_{-1}\| - \sqrt{d} \, \right| \le t_3 \right\}$$

hold with probability at least $1 - 2\exp(-ct_3^2)$. Define the quantities

$$\epsilon_1 \coloneqq \|\boldsymbol{x}\| - \sqrt{d}\,, \qquad \epsilon_2 \coloneqq \xi_1\,, \qquad \epsilon_3 \coloneqq \|\boldsymbol{\xi}_{-1}\| - \sqrt{d}\,.$$

Note that when π is the standard Gaussian, a brief calculation using the definition (5.1) shows that $a(\boldsymbol{x}, \boldsymbol{y}) = \exp(\frac{h}{4}(\|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2))$. Then, on the event $E_1 \cap E_2 \cap E_3$, we have that

$$\begin{aligned}
\frac{h}{4}\left| \, \|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2 \, \right| &= \frac{h}{4}\left| x_1^2 - \left[(1-h)x_1 + \sqrt{2h}\,\xi_1\right]^2 - 2h\,\|\boldsymbol{\xi}_{-1}\|^2 \right| \\
&= \frac{h}{4}\left| \left(\sqrt{d} + \epsilon_1\right)^2 - \left[(1-h)\left(\sqrt{d} + \epsilon_1\right) + \sqrt{2h}\,\epsilon_2\right]^2 - 2h\left(\sqrt{d} + \epsilon_3\right)^2 \right| \\
&= O(dh^3 + d^{1/2}h^2 t_1 + h^{3/2}d^{1/2}t_2 + d^{1/2}h^2 t_3)\,,
\end{aligned}$$

assuming that $t_1 = O(d^{1/2})$. In fact, we take $t_1, t_3 = d^{1/6}$. If we take t_2 to be a sufficiently large constant (and the dimension d is large), then we can ensure that the event $E_2 \cap E_3$ holds with probability at least $10/11$. With these choices,

$$\frac{h}{4}\left| \, \|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2 \, \right| = O(dh^3 + d^{2/3}h^2 + d^{1/2}h^{3/2})\,.$$

Taking $h \le c/d^{1/3}$ for a sufficiently small constant $c > 0$, we can ensure that $a(\boldsymbol{x}, \boldsymbol{y}) \ge 11/12$. Thus, on the event E_1, we have

$$A(\boldsymbol{x}) = \mathbb{E}[A(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{x}] \ge \mathbb{E}[A(\boldsymbol{x}, \boldsymbol{y})\,\mathbb{1}_{E_2 \cap E_2} \mid \boldsymbol{x}] \ge \frac{11}{12} \cdot \frac{10}{11} = \frac{5}{6}\,.$$

This completes the proof. □

5.8.2 Lower bound

We show that when the step size is chosen as $h \gg d^{-1/3}$, then the conductance of the MALA chain with Gaussian target is exponentially small.

Proposition 5.8.3. *For every $\theta < 1/3$, if we take step size $h = d^{-\theta}$, then the conductance of the MALA chain is exponentially small:*

$$\exists \delta > 0 \qquad \textit{such that} \qquad \mathsf{C} \lesssim \exp[-\Omega(d^{\delta})] \,.$$

Proof. We want to upper bound the conductance, defined in (C). It suffices to show that there exists an event $E \subseteq \mathbb{R}^d$ with $\pi(E) \geq 1/2$ such that

$$\sup_{\boldsymbol{x} \in E} \int Q(\boldsymbol{x}, \boldsymbol{y})\, A(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y} = \exp[-\Omega(d^{\delta})] \,,$$

see Proposition 5.7.1. Specifically, we take $E := \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\| \leq \sqrt{d}\}$; note that

$$\pi(E) = \frac{\Gamma(\frac{d}{2}, 0) - \Gamma(\frac{d}{2}, \frac{d}{2})}{\Gamma(\frac{d}{2})} > \frac{1}{2} \,.$$

From the definition (5.1), we have $A(\boldsymbol{x}, \boldsymbol{y}) = a(\boldsymbol{x}, \boldsymbol{y}) \wedge 1 \leq \sqrt{a(\boldsymbol{x}, \boldsymbol{y})}$.[2] Since $V(\boldsymbol{x}) = \frac{1}{2} \|\boldsymbol{x}\|^2$, a little algebra using the definition (5.1) shows that

$$a(\boldsymbol{x}, \boldsymbol{y}) = \exp\Big(\frac{h}{4} \big(\|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2\big)\Big) \,.$$

Further calculations show that

$$\begin{aligned}
\int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y})\, A(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y} &\leq \int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y})\, a(\boldsymbol{x}, \boldsymbol{y})^{1/2}\, \mathrm{d}\boldsymbol{y} \\
&= \int_{\mathbb{R}^d} \frac{1}{(4\pi h)^{d/2}} \exp\Big(-\frac{1}{4h} \|\boldsymbol{y} - (1-h)\,\boldsymbol{x}\|^2\Big) \exp\Big(\frac{h}{2} \big(\|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2\big)\Big)\, \mathrm{d}\boldsymbol{y} \\
&= \frac{1}{(4\pi h)^{d/2}} \int_{\mathbb{R}^d} \exp\Big(-\frac{1 + h^2/2}{4h} \Big\|\boldsymbol{y} - \frac{1-h}{1+h^2/2}\,\boldsymbol{x}\Big\|^2\Big)\, \mathrm{d}\boldsymbol{y} \\
&\qquad\qquad \times \exp\Big(\frac{h^2\,(1 - h/4)}{1 + h^2/2} \|\boldsymbol{x}\|^2\Big) \\
&= \exp\Big(\frac{h^2\,(1-h/4)}{4\,(1+h^2/2)} \|\boldsymbol{x}\|^2 - \frac{d}{2} \ln\Big(1 + \frac{h^2}{2}\Big)\Big) \,.
\end{aligned}$$

[2] One can check that the simple bound $A(\boldsymbol{x}, \boldsymbol{y}) \leq a(\boldsymbol{x}, \boldsymbol{y})$ is not enough for the proof to go through. A similar argument to upper bound the acceptance probability is made in [HSV14].

For $\boldsymbol{x} \in E$, we can bound this via

$$\begin{aligned}\int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y})\, A(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y} &\leq \exp\Big(\frac{h^2\,(1 - h/4)d}{4\,(1 + h^2/2)} - \frac{d}{2}\ln\big(1 + \frac{h^2}{2}\big)\Big) \\ &= \exp\Big(-\frac{h^3 d}{16}\,\big(1 + O(h)\big)\Big)\end{aligned}$$

which completes the proof. □

The next result shows that the spectral gap of the MALA chain is always upper bounded by the step size. Together with the preceding result, it implies that the mixing time of the MALA chain with Gaussian target is no better than the claimed $O(d^{1/3})$ rate.

Proposition 5.8.4. *The spectral gap of MALA with Gaussian target distribution and step size h satisfies*

$$\lambda \lesssim h\,.$$

Proof. This is a special case of Theorem 5.7.7. □

■ 5.9 Conclusion

By establishing the sharp dimension dependence of MALA for smooth and strongly convex potentials, our work parallels well-known trends in optimization [Bub15; Nes18] and high-dimensional statistics [Tsy09; Wai19] which seek to characterize the complexity of various learning tasks uniformly over a given function class. It is an interesting open question to extend our results on MALA to other natural function classes, such as smooth and weakly convex potentials, as well as to other sampling algorithms.

We mention two further works on the complexity of MALA which were published after the first version of this work appeared online. In [LST21a], Lee, Shen, and Tian proved that from certain initializations with warm start parameter $M_0 \sim \exp(d)$, the complexity of MALA is lower bounded by $\widetilde{\Omega}(d)$. This shows that the warm start dependence in our mixing time bound (Theorem 5.5.1) is actually *necessary*. In [WSC22], Wu, Schmidler, and Chen refined our lower and upper bounds, showing that the complexity of MALA from a warm start is $\widetilde{\Theta}(\kappa d)$.

To conclude, we list some specific directions that require further investigations.

Obtaining a warm start. The results mentioned above carry the message that in order to obtain further improvements for the complexity of MALA, it is not possible to further improve the upper bound analyses (e.g., by improving the

dependence on the warm start parameter). Instead, the natural next step is to algorithmically obtain a warm start for MALA in order to take advantage of the faster rates that the warm start unlocks. This program will be carried out in §6.

Analysis of other Metropolis–Hastings chains. An interesting feature of Theorem 5.4.2 is that the majority of the computations involve controlling the discretization error between the continuous-time and discretized Langevin processes, leading to the hope that the vast literature on discretization of SDEs can be leveraged to obtain mixing time bounds for the corresponding Metropolis–Hastings chains. However, a critical component of this program is the choice of a reversible Markov diffusion to which the MALA kernel can be compared via the projection property (Theorem 5.4.5). For example, consider the following two settings:

1. Under higher-order smoothness, the diffusion scaling limit of [RR98] suggests that the mixing time of MALA should scale as $d^{1/3}$, using step size $h \approx d^{-1/3}$. Indeed, our computations in §5.8 confirm this prediction for a Gaussian target distribution. However, in this regime, the discretized Langevin proposal is too far from the continuous-time Langevin diffusion for our upper bound strategy to succeed. Thus, in this example, the natural choice of reversible Markov diffusion fails to yield the correct mixing time for MALA.

2. The underdamped Langevin SDE [Che+18b] is an example of a Markov diffusion which is not reversible. We can consider adding a Metropolis adjustment after a proposal which consists of one step of the discretized underdamped Langevin process. It is not clear that our techniques apply to this example because there does not appear to be a natural *reversible* Markov diffusion with which to compare the resulting Metropolis-adjusted kernel.

Despite these obstacles, we believe that there is a wide variety of applications to which our upper bound technique applies, which we leave for future research.

Chapter 6

Algorithmic warm starts for MALA

In §5, we showed that the dimension dependence of MALA improves from $\widetilde{\Theta}(d)$ to $\widetilde{\Theta}(d^{1/2})$ under a warm start. In this chapter, we show how to obtain such a warm start in $\widetilde{O}(d^{1/2})$ queries through a Rényi divergence analysis of underdamped Langevin Monte Carlo (ULMC), thereby improving the dimension dependence of high-accuracy log-concave sampling to $\widetilde{O}(d^{1/2})$. In turn, when combined with the proximal sampler reduction of §4, it leads to new state-of-the-art sampling guarantees under isoperimetry.

This chapter is based on [AC23], joint with Jason M. Altschuler.

6.1 Introduction

We consider the problem of efficiently sampling from a high-dimensional probability distribution π on $\mathbb{R}^d$. Due to the many important applications of sampling throughout applied mathematics, engineering, and statistics, significant research effort has been devoted to designing fast sampling algorithms and analyzing their convergence rates. We refer to the book draft [Che23] for a recent exposition of the extensive literature and its history.

Yet, despite several decades of progress, many fundamental theoretical questions remain open about the complexity of sampling. Arguably one of the foremost questions in this field is:

What is the first-order query complexity for sampling from π?

Recall that a first-order query refers to accessing $V(x)$ and $\nabla V(x)$ at a query point $x \in \mathbb{R}^d$, where V denotes the negative log-density of $\pi \propto \exp(-V)$ (up to an additive normalization constant). This well-studied notion of a first-order query is inspired on one hand by the fact that such queries do not require knowledge of the normalization constant $\int \exp(-V)$ and thus are readily available in many practical applications, and on the other hand also inspired by the analogous and influential theory of complexity for convex optimization [NY83].

This problem of determining the query complexity for sampling has remained open even for the canonical and seemingly simple class of strongly log-concave and log-smooth (in brief, "well-conditioned") distributions π, let alone in more complicated settings. It is worth emphasizing that this state of affairs for *sampling* is in sharp contrast to that for *optimization*—indeed, the analogous query complexity questions for convex optimization were solved long ago in celebrated results from the 1980s [NY83; Nes18].

Within the literature, it is of central interest to understand this complexity question in the *high-accuracy regime*[1], since classical high-accuracy samplers such as the Metropolis-adjusted Langevin algorithm (MALA) and the Metropolized Hamiltonian Monte Carlo algorithm (MHMC) remain the de facto gold standard in practice. Yet the complexity for this high-accuracy setting has been particularly difficult to pin down, as we explain shortly.

The purpose of this chapter is to develop faster high-accuracy samplers, and in doing so move towards a better understanding of the first-order complexity of sampling. For simplicity of exposition, let us presently assume that π is well-conditioned, since by the proximal reduction framework [LST21c] (see §4), it is known that improvements to the complexity of well-conditioned sampling lead to improvements in more general settings such as when π is (non-strongly) log-concave, or even non-log-concave but satisfies standard isoperimetric assumptions such as the log-Sobolev or Poincaré inequality. (Indeed, our results improve upon the state-of-the-art for all these settings.)

The gap between low-accuracy and high-accuracy samplers. A central motivation of this chapter is the large gap between (our current understanding of) the complexity of low-accuracy samplers and high-accuracy samplers. To explain this gap, let us briefly provide relevant background on both classes of algorithms.

Low-accuracy samplers arise as discretizations of stochastic processes with stationary distribution π, such as the Langevin diffusion [the sampling analog of the gradient flow, see JKO98; Wib18] or the underdamped Langevin diffusion [the sampling analog of the accelerated gradient flow, see Ma+21]. Once discretized, however, the resulting discrete-time Markov chain is typically *biased*, i.e., its stationary distribution is no longer equal to π. In order to control the size of the bias, the step size of the algorithm is chosen to scale polynomially with ε, and hence the overall running time scales polynomially with $1/\varepsilon$. Despite this drawback, the discretization analysis is by now well-understood, with state-of-the-art results

[1]Throughout, we use the standard terminology *low accuracy* to refer to complexity results which scale polynomially in $1/\varepsilon$, and the term *high accuracy* for results which scale polylogarithmically in $1/\varepsilon$; here, ε is the desired target accuracy. These two regimes require different algorithms and analyses, as explained in the sequel.

achieving a complexity of $\widetilde{O}(d^{1/3}/\varepsilon^{2/3})$ [SL19; FLO21; BM22]; see [CLW21] for a discussion of tightness.

High-accuracy samplers, in contrast, are typically designed in such a way that there is no bias. This is achieved by, e.g., appending a Metropolis–Hastings filter to each step (see §5.2.2 for background). Common examples of these algorithms include MALA and MHMC, which are routinely deployed in large-scale applications and are the default implementations of sampling routines in many modern software packages [GLG15; Aba+16b]. However, the filter which debiases the algorithm also greatly complicates the analysis, and thus far the best complexity result for these algorithms[2] is $\widetilde{O}(d\log^{O(1)}(1/\varepsilon))$ [Dwi+19; Che+20a; LST20]. Note that the dimension dependence of this result is substantially worse than what is known in the low-accuracy regime and is at odds with the popularity of high-accuracy samplers in practice.

The mystery of warm starts. A promising first step towards resolving this gap is the result of §5, later refined in [WSC22]: when initialized from a *warm start* (i.e., a measure μ_0 with $\chi^2(\mu_0 \,\|\, \pi) \le O(1)$), the complexity of MALA improves to $\widetilde{O}(d^{1/2}\log^2(1/\varepsilon))$ since it can safely take much larger step sizes (of size $d^{-1/2}$ rather than d^{-1}). This raises the natural question: is the warm start condition merely an artefact of the analyses? Rather surprisingly, it was shown in [LST21a] that there exist bad initializations for MALA for which the dimension dependence is at least $\widetilde{\Omega}(d)$. Taken together, these results show that the complexity of MALA fundamentally hinges on the warmness of its initialization.

The key question is thus: can such a warm start be obtained algorithmically? Or more precisely:

Is there an algorithm which makes $\widetilde{O}(d^{1/2})$ queries to a first-order oracle for V and outputs a measure μ_0 with $\chi^2(\mu_0 \,\|\, \pi) \le O(1)$?

The requirement that the algorithm makes $\widetilde{O}(d^{1/2})$ queries is essential, else the cost of obtaining the warm start dominates the subsequent cost of running MALA. Yet this was the state of affairs—previously, the fastest algorithms took significantly longer to produce a warm start than to actually use it, defeating the purpose of the warm start. Resolving this discrepancy has been posed as an important question in many papers, e.g., [LST21a; LW22; WSC22].

The main challenge for answering this warm start question is that the chi-squared divergence is quite a strong performance metric. (We emphasize that it is essential to obtain the warm start in the chi-squared divergence, or more generally in a Rényi divergence $\mathcal{R}_q$ of order $q > 1$, rather than other common

[2] We discuss the result of [LW22] for the zigzag sampler further in §6.1.3.

metrics such as total variation, Wasserstein, or KL divergence; see §6.1.2 for an in-depth discussion.) The aforementioned results in the low-accuracy regime fall short of achieving this goal, since they only hold in the Wasserstein metric (for which standard coupling arguments are readily available). Despite significant effort, the best known guarantee for producing a warm start—achieved by the Langevin Monte Carlo (LMC) algorithm (see §3)—is far too costly as it requires $\widetilde{O}(d)$ queries, which defeats the purpose of the warm start.

Towards this hope of algorithmic warm starts, [WSC22] made the promising empirical observation that MALA mixes much faster if it is initialized at the output of the *underdamped* Langevin Monte Carlo (ULMC) algorithm. However, they left open the question of rigorously proving that this yields a warm start. While it is widely believed that ULMC is substantially faster than LMC, the previous best results for computing a warm start with ULMC had dimension dependence $\widetilde{O}(d^{5/2})$ (implicit from [GT20]) or very recently $\widetilde{O}(d^2)$ (implicit from [Zha+23]), see the prior work section §6.1.3 for details. We emphasize that this dimension dependence is not only a far cry from the elusive $\widetilde{O}(d^{1/2})$ goal, but moreover is even worse than known results for the simpler LMC algorithm. Unfortunately, any improvement to these ULMC warm start bounds appears to require overcoming fundamental difficulties with studying hypocoercive differential equations which remain unsolved today, despite being the focus of intensive research activity within the PDE community since the work of Kolmogorov [Kol34]. For a further discussion of these technical obstacles, see §6.1.2.

■ 6.1.1 Contributions

In this paper, we develop techniques which bypass longstanding challenges for analyzing hypocoercive dynamics, thereby establishing the first $\widetilde{O}(d^{1/2})$ Rényi mixing results for ULMC. This resolves the aforementioned warm start conjecture, which has been raised in a number of prior works, e.g., [LST21a; LW22; WSC22]. As discussed above, this enables us to design significantly faster high-accuracy samplers—both for the log-concave setting and far beyond. Finally, this also closes the long line of work devoted to understanding the complexity of MALA (see Table 6.1). We present our results in more detail below, and then discuss our new techniques in §6.1.2.

Result 1: Algorithmic warm starts via ULMC. Our first main result is an improvement of the state-of-the-art Rényi mixing bounds for ULMC from $\widetilde{O}(d^2)$ to $\widetilde{O}(d^{1/2})$. This resolves the warm start question in the affirmative. We remark that although the warm start problem was stated above for χ^2 convergence, our result actually holds more generally for Rényi divergences $\mathcal{R}_q$ of any order $q \geq 1$, and thus we state it as such. (For the purpose of warm starts, it suffices to take $q = 2$ since

Reference	Complexity	Algorithmically Achievable?
[Dwi+19]	$\kappa d + \kappa^{3/2} d^{1/2}$	No, requires a warm start
[Che+20a]	$\kappa d + \kappa^{3/2} d^{1/2}$	Yes
[LST20]	κd	Yes
Theorem 5.4.2	$\kappa d^{1/2} + \kappa^2$	No, requires a warm start
[WSC22]	$\kappa d^{1/2}$	No, requires a warm start
Theorem 6.5.1	$\kappa d^{1/2}$	**Yes (Theorem 6.4.1)**

Table 6.1: This table summarizes the community's progress towards non-asymptotic complexity bounds for MALA; the asymptotic study of MALA is much more classical, and dates back to at least [RR98]. The complexity bounds displayed are upper bounds; for brevity, we hide logarithmic factors as well as the dependence on ε since all results scale polylogarithmically in $1/\varepsilon$. As discussed in the main text, Theorem 6.5.1 completes our understanding of MALA due to matching lower bounds in [LST21a; WSC22] and Theorem 5.5.1 in §5.

$\chi^2 = \exp(\mathcal{R}_2) - 1$ is of constant size when $\mathcal{R}_2$ is.) Below, α and β denote the strong convexity and smoothness bounds, and $\kappa := \beta/\alpha$ is the condition number.

Theorem 6.1.1 (Rényi guarantees for ULMC; informal version of Theorem 6.4.1)**.** *Consider densities of the form $\pi \propto \exp(-V)$ on $\mathbb{R}^d$, where $\alpha I \preceq \nabla^2 V \preceq \beta I$ and $\kappa := \beta/\alpha < \infty$. The ULMC algorithm outputs a measure μ satisfying $\mathcal{R}_q(\mu \,\|\, \pi) \leq \varepsilon^2$ using $\widetilde{O}(\kappa^{3/2} d^{1/2} q^{1/2}/\varepsilon)$ first-order queries.*

As we detail in §6.1.2, the main barrier to obtaining this result is that the underdamped Langevin dynamics falls within a class of PDEs known as hypocoercive equations, for which fundamental questions remain unresolved.

Result 2: Faster high-accuracy log-concave sampling. Theorem 6.1.1 provides the first algorithm for computing warm starts that is not significantly slower than the use of the warm start. This enables us to exploit, for the first time, the results of §5 and [WSC22] which improve the complexity of MALA from $\widetilde{O}(d)$ to $\widetilde{O}(d^{1/2})$ from a warm start.[3] By combining this with additional algorithmic tools for improving the dependence on the condition number, we obtain our second main result, which substantially advances the state-of-the-art for high-accuracy log-concave sampling.

[3]We remark that all of our results could replace MALA with the zigzag algorithm [LW22]. Indeed, the zigzag sampler has the same key issue as MALA: it requires a warm start in chi-squared divergence for the known $d^{1/2}$ mixing result to apply. However, we focus on MALA because MALA's robust empirical performance has made it a central focus of study in the MCMC literature for nearly three decades [Bes+95].

Theorem 6.1.2 (High-accuracy log-concave sampling; informal version of Theorem 6.5.1)**.** *Consider the class of densities of the form $\pi \propto \exp(-V)$ on $\mathbb{R}^d$, where $\alpha I \preceq \nabla^2 V \preceq \beta I$ and $\kappa := \beta/\alpha$. There is an algorithm which outputs a sample with law μ satisfying $\mathsf{d}(\mu, \pi) \leq \varepsilon$, for any performance metric $\mathsf{d} \in \{\mathsf{TV}, \sqrt{\mathsf{KL}}, \sqrt{\chi^2}, \sqrt{\alpha}\, W_2\}$, after making $\kappa d^{1/2} \log^{O(1)}(\kappa d/\varepsilon)$ first-order queries.*

The algorithmic warm start result of Theorem 6.1.1 confirms the aforementioned empirical conjecture of [WSC22] and provides the final missing piece in our understanding of the complexity of MALA, closing the line of work developed in [RR98; Dwi+19; Che+20a; LST20; Che+21b; LST21a; WSC22] (Table 6.1). Indeed, due to matching lower bounds in [LST21a; WSC22], the complexities $\widetilde{O}(\kappa d)$ and $\widetilde{O}(\kappa d^{1/2})$ with or without a warm start are known to be tight.

The complexity in Theorem 6.1.2 constitutes a natural barrier for high-accuracy sampling. Indeed, regarding the dimension dependence, any further progress beyond $\widetilde{O}(d^{1/2})$ would seem to require completely different algorithms—both for obtaining a warm start and also for exploiting a warm start. For example, the $\widetilde{O}(d^{1/2})$ complexity of MALA is unimprovable even under arbitrarily warm starts (Theorem 5.5.1 and [WSC22]). And regarding the condition number dependence, any further progress beyond $\widetilde{O}(\kappa)$ in the high-dimensional regime[4] would constitute a major breakthrough in the complexity of sampling since it is currently unknown whether an acceleration phenomenon holds in the sampling context.

More broadly, our result provides evidence of the potential for designing faster high-accuracy samplers by combining low-accuracy samplers for computing a warm start, together with improved high-accuracy mixing from the warm start. We believe that this research program may be crucial for future progress in high-accuracy sampling, since faster mixing from a warm start seems likely to hold for other Metropolized algorithms. See §6.9 for further discussion in this direction.

Result 3: Faster high-accuracy sampling beyond log-concavity. We recall that high-accuracy log-concave sampling is the key to obtaining state-of-the-art complexity results for a wide class of distributions beyond log-concavity. This is achieved by using our faster log-concave sampler in Theorem 6.1.2 to improve the per-iteration complexity of the proximal sampler (see §4). This approach is overviewed in the techniques section §6.1.2, and leads to the following result.

Corollary 6.1.3 (Sampling from other classes of distributions; informal version of results in §6.5.2)**.** *For each of the following classes of distributions, we obtain complexity bounds which improve by a factor of $d^{1/2}$ over the results of §4:*

[4] Analogous to classical optimization results, there are sampling algorithms which achieve logarithmic dependence on κ at the expense of larger polynomial dependence on d. The open question mentioned here is really: can one improve the condition dependence beyond near-linear while also maintaining comparable dimension dependence?

- π *is log-smooth and weakly log-concave.*
- π *is log-smooth and satisfies a log-Sobolev inequality.*
- π *is log-smooth and satisfies a Poincaré inequality.*

The latter two assumptions of log-Sobolev and Poincaré—called *functional inequalities*—capture strictly richer classes of target distributions than strong-log-concavity. There are two major motivations for studying the complexity of sampling in this setting. First, functional inequalities are quite flexible, as they are preserved under common operations such as bounded perturbations and Lipschitz mappings (for background, see §2.2). Consequently, they often capture the breadth of settings encountered in practice, including non-log-concave settings. Second, these functional inequalities classically imply convergence of diffusions in continuous time, making them natural assumptions under which to study the corresponding discretizations.

Despite the appeal of this program, proving sampling guarantees under functional inequalities introduces a number of additional technical complications and was only accomplished recently, starting with [VW19] and continued in [Wib19; Ma+21; LE23] and §3. Our result continues this line of work, and in particular highlights the use of high-accuracy samplers for well-conditioned distributions as a powerful algorithmic tool for the broader problem of sampling under isoperimetry.

6.1.2 Challenges and techniques

6.1.2.1 Challenges for warm starts: Rényi divergence and hypocoercivity

Why Rényi? To explain what properties are needed for a warm start requires first explaining why a warm start helps. Briefly, the complexity of MALA is governed by the largest possible step size for which the algorithm still accepts a reasonable fraction of the proposals (see §5.2.2 for background on MALA). The basic reason why we might expect to improve the complexity of MALA from $\widetilde{O}(d)$ to $\widetilde{O}(d^{1/2})$ is that *at the stationary distribution* π, the step size can be increased significantly from d^{-1} to $d^{-1/2}$ while keeping the acceptance probability high. More precisely, with step size $d^{-1/2}$, the acceptance probability is large from a typical point from π; however, it can be exponentially small in regions that are atypical (i.e., exponentially rare under π). The existence of such regions implies that there are "bottlenecks" in the state space which take exponentially long to traverse. The role of a warm start initialization is to avoid such bottlenecks.

In other words, a key property that a warm start μ_0 must satisfy is that if π assigns exponentially small probability to an event, then so must μ_0. Crucially, this property does not hold if μ_0 is only known to be close to π in common

probability metrics such as total variation, Wasserstein, or KL divergence—but this property *does* hold if μ_0 is close to π in the chi-squared divergence, or more generally any Rényi divergence $\mathcal{R}_q$ of order $q > 1$.[5]

The key to warm starts: low-accuracy algorithms. In the preceding discussion, taking large step sizes from a non-warm initialization was problematic due to the rejections in the Metropolis–Hastings filter step. A natural idea, then, is to remove the filter for the initial stage of the algorithm and later reinstate it when the law of the iterate is closer to the target π. Since the proposal of MALA is just one step of the LMC algorithm, this amounts to using LMC to procure the warm start. More generally, we can consider using any low-accuracy sampler as our warm start algorithm, and indeed, as we discuss next, it will be crucial to consider ULMC instead of LMC in order to achieve the desired $\widetilde{O}(d^{1/2})$ dimension dependence.

At a high level, if we discretize a diffusion with step size h for continuous time T, then the total number of iterations is $N = T/h$. In order to understand the dimension dependence of the algorithm, one must therefore understand both h and T. These two terms reflect two distinct aspects of mixing analysis: the discretization bias and the convergence time.

The first part—the discretization bias—is now relatively well-understood (see the prior work discussions in §6.1.3), even for the chi-squared divergence and more general Rényi divergences. In particular, it is known that the Rényi bias of LMC is controlled for step sizes $h \lesssim 1/(dT)$, and the Rényi bias of ULMC is controlled for step sizes $h \lesssim 1/\sqrt{dT}$. (In fact, we streamline arguments in the literature in order to provide a shorter and simpler proof of this in §6.7.3.) Since the Langevin diffusion does not reach approximate stationarity until time $T \geq \Omega(\log d)$, it follows that LMC requires at least $N = T/h = \widetilde{\Omega}(d)$ iterations, which is too slow for our purposes.

ULMC is more promising, as the discretization bounds lead to iteration complexity bounds of $N = T/h = d^{1/2}T^{3/2}$. However, in order to reach our warm start goal of $N = \widetilde{O}(d^{1/2})$, this means that the convergence time T must be nearly dimension-free, i.e., of size $\widetilde{O}(1)$.

Why are nearly dimension-free convergence rates possible in continuous time? Since the Rényi divergence to π initially scales as $\widetilde{\Theta}(d)$, in order to obtain nearly dimension-free bounds on T, we require the diffusion to converge to stationarity in Rényi divergence with an *exponential* rate. This is a strong property of the diffusion, which we call *hyperequilibration*.

Hyperequilibration was not even known for the simpler (standard, overdamped)

[5] This is the same reason why differential privacy requires guarantees in Rényi divergences [Mir17].

Langevin diffusion (LD) until quite recently [CLL19; VW19]. While a spectral gap for LD (or equivalently, a Poincaré inequality for π) classically implies exponential decay of the chi-squared divergence, this is far weaker than hyperequilibration. Indeed, hyperequilibration requires exponential decay of $\mathcal{R}_2$, which amounts to *doubly* exponential decay of the chi-squared divergence, since $\mathcal{R}_2 = \log(\chi^2 + 1)$. Under the stronger assumption of a log-Sobolev inequality for π, it is well-known that the KL divergence decays exponentially fast, but it was unclear that the same holds for the *Rényi divergence* which, as discussed above, is crucial for warm starts. It was only through the inspired semigroup calculations of [CLL19; VW19] that we now know this to be true, namely, a log-Sobolev inequality implies hyperequilibration for LD.[6]

Recall, though, that the LD incurs too much discretization bias. To obtain sufficient control over both the discretization bias and the convergence time, we therefore need to establish hyperequilibration for the *underdamped* Langevin diffusion (ULD). However, this question brings us to longstanding challenges from the theory of hypocoercive PDEs.

Hypocoercivity: a fundamental barrier for underdamped analysis. To recap: for LD, we have exponential decay of the chi-squared divergence under a Poincaré inequality, exponential decay of the KL divergence under a log-Sobolev inequality, and finally hyperequilibration under a log-Sobolev inequality. What, then, are the analogous results for ULD? Since its introduction in the 1930s by Kolmogorov [Kol34], the regularity and convergence of ULD have been the focus of intensive research. It took nearly half a century to establish mixing [Tro77], and a further 30 years and Villani's "slightly miraculous-looking computations" [Vil09a, pg. 42] to prove exponential decay of the KL divergence under a log-Sobolev inequality. Establishing hyperequilibration for ULD remains out of reach for existing techniques.

The reason for this sudden jump in difficulty from the overdamped to the underdamped diffusions is due to a fundamental issue: the *degeneracy* of ULD. In brief, whereas LD is driven by a full-dimensional Brownian motion, ULD is driven by a degenerate one which is only added to a subset of the coordinates. For sampling purposes, this degeneracy is a desirable feature as it leads to smoother sample paths and smaller discretization error; however, this same degeneracy is also the source of deep questions in PDE theory which have motivated research in that field for nearly a century. The key challenge here is that the standard tools of Markov semigroup theory—which provide the backbone of the analysis for LD—completely break down for ULD. To address this difficulty, the theory

[6] This explains our choice of the terminology *hyperequilibration*: it is inspired by the analogy to the classical property of *hypercontractivity*, which is equivalent to the logarithmic Sobolev inequality (LSI) [Gro75].

of *hypocoercivity*, inspired by Hörmander's groundbreaking work on hypoellipticity [Hör67], was laid down by Villani in the monograph [Vil09a] as a principled framework for the study of degenerate diffusions. However, this is still a relatively nascent area of PDE and many important questions remain wide open; see the prior work in §6.1.3 for further background.

In contrast, we note that it is well-known how to obtain fast rates of convergence in the Wasserstein metric via standard coupling arguments. Consequently, the state-of-the-art $\widetilde{O}(d^{1/2})$ guarantees for the ULMC algorithm hold in the Wasserstein metric or the KL divergence [Che+18b; SL19; DR20; Ma+21; Zha+23], whereas for Rényi divergence bounds, it was previously unknown how to obtain rates which are better than even $\widetilde{O}(d^2)$.

■ 6.1.2.2 Settling the warm start conjecture: regularization via privacy

Our approach to hyperequilibration. To settle the warm start conjecture, we adopt a fundamentally different perspective. Namely, instead of trying to directly establish hyperequilibration via hypocoercivity techniques, we ask whether it can be deduced from simpler Wasserstein coupling arguments. At the heart of this approach is the fact that diffusions often enjoy strong *regularizing* properties, which allow for bounding stronger metrics (e.g., Rényi) in terms of weaker ones (e.g., Wasserstein). Such regularization results are typically established for continuous-time diffusions via abstract calculus methods, such as the theory of Markov semigroups [BGL14]. However, as discussed above, these techniques do not extend to ULD due to the fundamental issue of degeneracy.

Our key insight is to prove a regularization result for the *discrete-time* algorithm directly. This is enabled by the fact that although the noise added to each iteration of ULMC is *nearly* degenerate—and indeed degenerates as the step size $h \searrow 0$, as it must because ULD is degenerate—this ULMC noise remains non-degenerate for any positive step size $h > 0$. Hence, we can expect some mild amount of regularization for ULMC, a fact that we establish for the first time. On a technical level, we accomplish this via a more sophisticated version of techniques from the differential privacy literature—namely, the shifted Rényi analysis—which we describe next.

Rényi divergences with Orlicz–Wasserstein shifts. The regularization result we seek is of the following form: if we initialize two copies of our process of interest at the distributions μ_0, ν_0, and arrive at distributions μ_n, ν_n respectively at iteration n, we wish to control $\mathcal{R}_q(\mu_n \parallel \nu_n)$ in terms of an initial Wasserstein distance $W(\mu_0, \nu_0)$. In our application, the process of interest—namely ULMC—is an instance of what is sometimes called a "contractive noisy iteration" (CNI): an algorithm that interleaves Lipschitz mappings with (Gaussian) noise convolution

steps. This notion of a contractive noisy iteration is of broad interest as it captures algorithms in differential privacy (e.g., noisy optimization algorithms) and in sampling (e.g., discretizations of diffusions), and we therefore place our results in a framework which encompasses these various use cases.

A generalization of the regularization result we seek is to prove that for a CNI,

$$\mathcal{R}_q(\mu_n \,\|\, \nu_n) \lesssim \mathcal{R}_q^{(w)}(\mu_0 \,\|\, \nu_0) + [\text{error term depending on } w]\,, \tag{$\star$}$$

where $\mathcal{R}_q^{(w)}$ is the *shifted Rényi divergence*, defined as

$$\mathcal{R}_q^{(w)}(\mu \,\|\, \nu) \coloneqq \inf_{\mu' \text{ s.t. } W(\mu,\mu') \le w} \mathcal{R}_q(\mu' \,\|\, \nu)\,,$$

see §6.3 for details. Indeed, if we take $w = W(\mu_0, \nu_0)$ in ($\star$), then the term $\mathcal{R}_q^{(w)}(\mu_0 \,\|\, \nu_0)$ vanishes, and we will have controlled $\mathcal{R}_q(\mu_n \,\|\, \nu_n)$ in terms of $W(\mu_0, \nu_0)$ as desired. However, ($\star$) is more general, as it allows for carefully tracking the shift parameter w throughout. This proof technique, called *shifted divergence analysis*, was first introduced in the context of differential privacy by [Fel+18] for the purpose of establishing privacy amplification by iteration, and was recently honed into a form amenable to sampling analyses in [AT22a; AT22b].

A subtle yet essential technical issue that arises in establishing ($\star$) is: which Wasserstein metric W do we use? All previous versions of ($\star$) required the W_∞ metric, which is problematic for our setting as the W_∞ metric is infinite at initialization. Here, our main insight is to use a non-standard Wasserstein metric, called the *Orlicz–Wasserstein metric*, based on the sub-Gaussian Orlicz norm. As we discuss in Remark 6.3.9, this is exactly the right metric to use: in fact, ($\star$) cannot hold for any weaker metric (e.g., W_p for any finite p), and the initialization bound cannot be finite for any stronger metric. We then show that for Orlicz–Wasserstein shifts, ($\star$) indeed holds, with the caveat that the order of the shifted Rényi divergence on the right-hand side of ($\star$) is increased. This increase in the order also means that additional care is required when applying ($\star$), as the inequality cannot be iterated too many times, but we bypass this issue by showing that it suffices to only exploit the regularization from a *single* step.

Finally, we note that our analysis answers the open question raised in [AT22b] of how to use the shifted divergence technique in order to obtain sampling guarantees for discretized diffusions w.r.t. the true target distribution π, rather than w.r.t. the biased limit of the algorithm.

■ 6.1.2.3 From warm starts to faster high-accuracy samplers

In light of the discussion thus far, combining our warm start result with the results of §5 and [WSC22] immediately improves the dimension dependence of high-accuracy log-concave sampling to $\widetilde{O}(d^{1/2})$. However, two further issues remain.

First, thus far we have ignored the dependence on the condition number κ for simplicity of exposition, but the combined approach of ULMC and MALA incurs suboptimal dependence on κ, namely $\kappa^{3/2}$ rather than κ. Second, the result only holds for strongly log-concave targets. We address both of these issues simultaneously by adding a third algorithmic building block: the proximal sampler. Below, we briefly overview the proximal sampler and the final remaining technical challenges in its application.

Algorithmic framework. The proximal sampler [TP18; LST21c] is a Gibbs sampling method that can be viewed as the sampling analog of the proximal point method from optimization (see §4 for further background). Each iteration requires sampling from a regularized distribution called the restricted Gaussian oracle (RGO), parametrized by $y \in \mathbb{R}^d$:

$$\pi^{X|Y=y}(x) \propto \exp\Big(-V(x) - \frac{1}{2h}\|y-x\|^2\Big).$$

If V is β-smooth, and the step size h is chosen as $h \asymp \frac{1}{\beta}$, one can check that $\pi^{X|Y=y}$ is strongly log-concave and log-smooth with condition number $O(1)$. Hence:

complexity of the proximal sampler	=	# outer loops	×	complexity of sampling from $O(1)$-conditioned distributions to *high* accuracy

The requirement of sampling from the RGO to high accuracy arises to avoid accumulation of the errors from inexact implementation of the RGO.

So far, we have not made use of any assumptions on π beyond smoothness of V. Additional assumptions on π, such as log-concavity, can then used to control the number of outer loops. This program was carried out in §4, in which we studied the outer loop complexity of the proximal sampler under a variety of assumptions on the target π which, when combined with the implementation of the RGO via existing high-accuracy samplers, yielded state-of-the-art complexity bounds for sampling under those assumptions. Our faster high-accuracy log-concave sampler provides a better implementation of the RGO, and hence we improve upon these prior results by a factor of roughly $d^{1/2}$ in each setting. Moreover, in the strongly log-concave setting, the number of outer iterations of the proximal sampler is shown to be $\widetilde{O}(\kappa)$ [LST21c], so using ULMC + MALA to implement the RGO boosts the condition number dependence of the overall sampler to near-linear. This resolves the two issues described above, but in doing so we must also develop an inexact error analysis for the proximal sampler.

Inexact error analysis. In order to apply the proximal reduction framework, we must understand how the error from inexact implementation of the RGO propagates into the final sampling error. This was carried out in [LST21c] for the TV distance via a simple coupling argument, which amounts to a union bound over failure events at each iteration. Similarly, it is straightforward to carry out the inexact error analysis in the Wasserstein metric due to the availability of the triangle inequality. However, to establish our guarantees in §6.5, which hold also in the KL and χ^2 metrics, we must perform an error analysis in χ^2 (or equivalently, in Rényi). This is also complicated by the fact while the outer loop of the proximal sampler converges exponentially fast in the strongly log-concave setting, which facilitates summing up the geometrically decaying errors from each iteration, the convergence in the weakly log-concave setting does not have an exponential rate and moreover uses a modified Lyapunov functional, changing the nature of the error analysis. We remark that we did not encounter such issues in §4, since the rejection sampling implementation of the RGO is *exact*. Therefore, we believe that our inexact error analysis will also be useful for any future applications of the proximal sampler.

We also remark that our application of the proximal sampler, and the ensuing need for careful inexact error analysis, resembles the use of the (accelerated) proximal point method in optimization, e.g., [Fro+15; LMH15].

6.1.3 Related work

Low-accuracy sampling and Rényi guarantees. Rényi guarantees for sampling are relatively recent. Indeed, [VW19] proved fast Rényi mixing for LD and LMC to their respective stationary distributions, and this was translated into Rényi sampling guarantees for LMC in [GT20; EHZ22] and §3, for the proximal sampler in §4, and for ULMC in [GT20; Zha+23]. These lines of work have led to $\widetilde{O}(d)$ dimension dependence for LMC and the proximal sampler, but for ULMC the rates are much worse, namely $\widetilde{O}(d^{5/2})$ dependence [GT20] and only very recently $\widetilde{O}(d^2)$ dependence [Zha+23]. In Theorem 6.4.1, we obtain the first $\widetilde{O}(d^{1/2})$ rate in Rényi divergence.

In contrast, there are many more works which break the $\widetilde{O}(d)$ barrier in the Wasserstein metric: the randomized midpoint discretization of Langevin [HBE20], unadjusted Hamiltonian Monte Carlo (HMC) [CV19], ULMC [Che+18b; DR20; Mon21], and sophisticated discretizations of ULMC and HMC [SL19; FLO21; BM22]. Among these algorithms, at present we only understand how to perform Rényi discretization analysis for ULMC, but for ULMC it is the convergence of the corresponding *continuous-time* diffusion which remains elusive.

Underdamped Langevin, hypoellipticity, and hypocoercivity. The underdamped (or kinetic) Langevin diffusion has a rich history, dating back to Kolmogorov [Kol34]. The PDE governing the evolution of its marginal density is referred to as the kinetic Fokker–Planck equation. Unlike the Langevin diffusion, which is driven by a full-dimensional Brownian motion and for which regularity and convergence fall within the purview of classical elliptic and parabolic PDE theory, the underdamped Langevin diffusion is the canonical example of a degenerate diffusion for which these and related questions remain active areas of research within PDE. See §6.4.1 for background.

The question of regularity for these equations was largely solved by landmark work of Hörmander [Hör67] in arguably one of the most influential breakthroughs in PDE theory of the last century through the introduction of the theory of *hypoellipticity*. In turn, it inspired Villani to coin the study of the convergence of such equations *hypocoercivity* in his seminal monograph [Vil09a].

While convergence of this diffusion has been studied for nearly a century, early convergence results were qualitative in nature. It took intensive developments in the PDE community to get to a point where quantitative rates could be extracted, beginning in the 1970s [Tro77]. We do not attempt to comprehensively survey the extensive literature here. We refer to the monograph [Vil09a] for history; see also, e.g., the papers [DMS09; Bau17; RS18] for more modern references. We also mention the recent space-time Poincaré approach of [CLW20; Alb+21], which is also directly inspired by Hörmander's hypoelliptic theory. As we discuss in §6.1.2, however, all of these approaches fall short of establishing the key property of hyperequilibration.

MALA. MALA has been intensely studied over the past three decades since its introduction in [Bes+95], in large part due to its strong practical performance—in fact, it and its variants comprise the default implementations of sampling routines in many modern software packages [GLG15; Aba+16b]. Many classical works studied the geometric ergodicity and asymptotic properties of MALA. With regards to the dimension dependence, particularly influential was the optimal scaling result of [RR98], which showed that taking step size $h \propto d^{-1/3}$ leads to a non-trivial diffusion limit for MALA as $d \to \infty$, at least for product measures π satisfying strong regularity assumptions and when initialized at stationarity. Modern analysis techniques have enabled an understanding of the *non-asymptotic* complexity of, see [Dwi+19; Che+20a; LST20; LST21a; WSC22], §5, and Table 6.1 for a summary of the progress in this direction. Our work closes this line of work by showing that the warm start rate of [WSC22], which is tight due to their matching lower bound, is achievable. Moreover, our work provides theoretical justification for the improved empirical performance of MALA after using a low-accuracy

algorithm for warm starts, as observed in [WSC22].

Proximal sampler. The proximal sampler is an algorithmic framework introduced in [TP18; LST21c]. In [LST21c], it was used as a mechanism for boosting the condition number dependence of any high-accuracy log-concave sampler to near-linear, which was then used to design samplers for composite and finite-sum potentials. In §4, we showed that the proximal sampler reduces the problem of sampling from distributions satisfying weak log-concavity or functional inequalities to the problem of high-accuracy log-concave sampling. In this work, we exploit both these properties of the proximal sampler (see §6.1.2).

We also mention that in recent work, the proximal sampler has been connected to stochastic localization, leading to recent progress on the KLS conjecture [KP21; CE22; KL22], as well as to diffusion models (see §17). There are also applications to sampling from semi-smooth or non-smooth potentials [LC22; LC23], and to differential privacy [GLL22; Gop+23a; Gop+23b].

Zigzag sampler. The zigzag sampler is an alternative high-accuracy sampler that was recently proposed in [BFR19]. Instead of using a Metropolis–Hastings filter, the zigzag sampler is a piecewise deterministic Markov process which can be implemented without discretization bias. It was recently shown in [LW22] that similarly to MALA, the zigzag sampler has a dimension dependence of $\widetilde{O}(d^{1/2})$ from a warm start. Morever, in [LW22, Corollary 1.4], Lu and Wang show that by using LMC with a large step size to warm start the algorithm, one obtains a high-accuracy log-concave sampler with dimension dependence $\widetilde{O}(d^{4/5})$. Indeed, the same strategy can be used with the warm start results of §5 and [WSC22] to obtain complexities strictly better than $\widetilde{O}(d)$; however, it is clear that such an approach can never reach the desired complexity of $\widetilde{O}(d^{1/2})$—and in fact there is a fundamental barrier even at $\widetilde{O}(d^{3/4})$ because it is bottlenecked by the discretization bias of LMC. The goal of this chapter is achieving $\widetilde{O}(d^{1/2})$ complexity as this is this a natural barrier for high-accuracy samplers given a warm start, and LMC cannot work for this goal.[7] In analogy to our use of MALA, our new complexity result for ULMC (Theorem 6.4.1) can also be used to warm start the zigzag sampler, leading to the same final complexity bound of $\widetilde{O}(\kappa d^{1/2} \log^{O(1)}(1/\varepsilon))$. This answers the open questions in [LW22] regarding warm starting the zigzag sampler.

[7]With regards to dimension dependence, running LMC with step size h for $1/h$ steps yields a distribution μ with $\log \chi^2(\mu \,\|\, \pi) \le \widetilde{O}(dh)$ (§3). By optimizing the step size h and combining this with the best known complexity $\widetilde{O}(d^{1/2} \log^{3/2} \chi^2(\mu \,\|\, \pi))$ of the zigzag sampler, one obtains the final complexity $\widetilde{O}(d^{4/5})$ [LW22]. Even if the complexity of the zigzag sampler were improvable to $\widetilde{O}(d^{1/2} \log \chi^2(\mu \,\|\, \pi))$, the total complexity would still be at least $1/h + d^{1/2}\,(dh) \ge \widetilde{\Omega}(d^{3/4})$. Thus $d^{3/4}$ is a natural barrier for any warm start approach using LMC.

Differential privacy and sampling. Sampling algorithms have been widely used in differential privacy ever since the invention of the exponential mechanism [MT07]; for an exposition of the surrounding history and applications, see the textbook [DR13]. Sampling-inspired analyses have also been recently used to prove privacy properties of optimization algorithms [CYS21; RBP22; YS22]. Most related to this paper are connections in the other direction: the use of techniques from differential privacy in order to analyze sampling. There are two lines of work in this direction. One involves the technique of adaptive composition for Rényi divergences and its use for establishing Rényi bias bounds for LMC and ULMC [GT20; EHZ22; Zha+23]. The other involves the technique of privacy amplification by iteration (PABI), which was originally used to bound the privacy loss of differentially private optimization algorithms [Fel+18; Bal+19; ADC20; FKT20; SBD21; AT22a], and its recent use for analyzing the mixing time of LMC to its biased stationary distribution [AT22b]. In this chapter, we build upon this technique in several key ways: we show how to improve mixing results for the biased distribution to mixing results for the target distribution, we show how to use these ideas for ULMC rather than LMC, and most importantly we overcome the key issue of unboundedness by replacing W_∞ shifts by Orlicz–Wasserstein shifts, see §6.1.2.

Simultaneous work. We also mention the concurrent paper [FYC23], which also achieves $\widetilde{O}(d^{1/2})$ dimension dependence for high-accuracy log-concave sampling via an approximate rejection sampling implementation of the RGO.

6.1.4 Organization

We recall preliminaries in §6.2, especially regarding Rényi divergences. We isolate in §6.3 our key new technique involving Orlicz–Wasserstein shifted Rényi divergences. We use this technique to obtain faster algorithmic warm starts in §6.4, and then use these warm starts to develop faster high-accuracy samplers in §6.5. We conclude in §6.9 by discussing several future research directions that are motivated by our results.

6.2 Preliminaries

Throughout, $\pi \propto \exp(-V)$ denotes the target density and $V : \mathbb{R}^d \to \mathbb{R}$ denotes the potential. We assume V is twice continuously differentiable for simplicity. We reserve the symbol N for the number of iterations that the Markov Chain Monte Carlo algorithm is run, h for the discretization step size, and $T = Nh$ for the total elapsed time. We write $\mathcal{P}(\mathbb{R}^d)$ to denote the space of probability distributions over $\mathbb{R}^d$, and we write $\mathcal{P}_2(\mathbb{R}^d)$ to denote the subset of $\mathcal{P}(\mathbb{R}^d)$ with finite second

moment. All logarithms are natural.

For simplicity of exposition, we assume throughout that we have access to an algorithm for generating independent standard Gaussian random variables. We use the standard notation $\widetilde{O}(g) = g \log^{O(1)}(g)$ to suppress low-order terms. Note that since our final results depend polynomially on the dimension d and condition number κ, the $\widetilde{O}$ hides polylogarithmic factors in these terms—on the other hand, since ε occurs only polylogarithmically in our high-accuracy results, we do not hide the polylogarithmic factors in ε.

We say that f is α-strongly convex if $\nabla^2 f \succeq \alpha I_d$, and that f is β-smooth if $\|\nabla^2 f\|_{\mathrm{op}} \leq \beta$. If f is convex, then f is β-smooth if and only if $\nabla^2 f \preceq \beta I_d$. We always denote by κ the condition number $\kappa \coloneqq \beta/\alpha$. If $\pi \propto \exp(-V)$ where V is α-strongly convex (resp. β-smooth), we say that π is α-strongly log-concave (resp. β-log-smooth). All other notation is introduced in the main text.

We refer to §2.2.3 for the definition and basic properties of Rényi divergences.

6.3 Improved shifted divergence analysis

In this section we isolate from our analysis a key new technique of independent interest. As overviewed in §6.1.2, this technique is a strengthening of the "shifted divergence" analysis, a.k.a., "privacy amplification by iteration" (PABI), in which we crucially improve the ∞-Wasserstein shift to an Orlicz–Wasserstein shift. This enables obtaining Rényi divergence bounds on the mixing of any Markov chain which interleaves Lipschitz mapping steps (e.g., gradient descent steps) and noise convolution steps (e.g., adding a Gaussian). This notion captures a variety of algorithms from the differential privacy and sampling communities, often called "contractive noisy iterations".

The main result of this section is formally stated as follows. This result makes use of the Wasserstein metric W_{ψ_2} that evaluates a coupling's quality via the sub-Gaussian Orlicz norm; see §6.3.1 for background on this notion.

Theorem 6.3.1 (Shifted Rényi divergence analysis)**.** *Consider two Markov chains $\{\mu_n\}_{n\geq 0}$ and $\{\mu'_n\}_{n\geq 0}$ with possibly different initialization, but with the same update transitions*

$$\mu_{n+1} = (\mu_n P_n) * \mathsf{normal}(0, \sigma^2 I_d)$$
$$\mu'_{n+1} = (\mu'_n P_n) * \mathsf{normal}(0, \sigma^2 I_d)$$

where P_n is a Markov transition kernel that is c-Lipschitz in the W_{ψ_2} metric. Then for any Rényi order $q \geq 1$,

$$\mathcal{R}_q(\mu_N \,\|\, \mu'_N) \leq c^{2N} \, \frac{q \, W^2_{\psi_2}(\mu_0, \mu'_0)}{2\sigma^2} \,, \tag{6.1}$$

so long as $N \geq \log_{1/c}\left(\frac{\sqrt{q\,(q-1)}\,W_{\psi_2}(\mu_0,\mu_0')}{\sigma\sqrt{2}}\right)$.

We remark that unlike previous versions of the shifted divergence technique, Theorem 6.3.1 requires a restriction on the number of iterations N.[8] But this restriction is mild due to the logarithmic dependence. In fact, it is equivalent to requiring the upper bound in (6.1) to be at most $1/(q-1)$.

The rest of this section is devoted to proving Theorem 6.3.1. In §6.3.1 we define a new Lyapunov function, and in §6.3.2 we use it to prove Theorem 6.3.1.

■ 6.3.1 Shifted Rényi divergence using Orlicz–Wasserstein shifts

Key to our proof of Theorem 6.3.1 is a new Lyapunov function for tracking how indistinguishable the Markov chains become as they evolve. This new Lyapunov function is a shifted Rényi divergence, but unlike the standard shifted divergence technique, here we measure the shift using an "Orlicz–Wasserstein metric" rather than W_∞.

We begin by recalling the definition of a sub-Gaussian Orlicz norm. For shorthand, we drop the adjective "sub-Gaussian" as this is the only Orlicz norm considered in this paper. For further background on Orlicz norms, we refer the reader to, e.g., the textbooks [RR91; Ver18], and we mention that the standard significance of this particular (sub-Gaussian) Orlicz norm is that a random variable is sub-Gaussian if and only if this norm is finite [Ver18, Example 2.7.13].

Definition 6.3.2 (Orlicz norm)**.** *The Orlicz norm of a random variable* X *is*

$$\|X\|_{\psi_2} \coloneqq \inf\left\{\lambda > 0 \;:\; \mathbb{E}\,\psi_2\left(\frac{\|X\|}{\lambda}\right) \leq 1\right\},$$

where the function $\psi_2 : \mathbb{R} \to \mathbb{R}$ *is defined as* $\psi_2(x) \coloneqq \exp(x^2) - 1$.

Our proof of Theorem 6.3.1 uses the Orlicz norm for defining an optimal transport metric between probability distributions.

Definition 6.3.3 (Orlicz–Wasserstein metric)**.** *The Orlicz–Wasserstein metric between distributions* μ, ν *is*

$$W_{\psi_2}(\mu,\nu) \coloneqq \inf_{\substack{X\sim\mu\\ Y\sim\nu}} \|X - Y\|_{\psi_2}\,,$$

where the infimum is taken over all pairs of jointly defined random variables (X, Y) *with* $X \sim \mu$ *and* $Y \sim \nu$.

[8] This restriction comes from the fact that with this new Orlicz–Wasserstein shifted Rényi divergence, the new shift-reduction lemma (Lemma 6.3.7) does not apply to arbitrarily large shifts.

Note that since the Orlicz norm satisfies the triangle inequality, the standard gluing lemma from classical optimal transport theory shows that W_{ψ_2} is indeed a metric; see, e.g., [Vil09b, §6]. The Orlicz–Wasserstein metric has also been considered in prior works [Stu11; Kel17; GHN23], but to our knowledge our work constitutes the first use of this metric for sampling analysis.

Remark 6.3.4 (Comparison to W_p)**.** *Let W_p denote the standard p-Wasserstein metric on $\mathbb{R}^d$. Then*

$$\frac{1}{\sqrt{p}} W_p \lesssim W_{\psi_2} \leq \frac{1}{\sqrt{\log 2}} W_\infty .$$

The first inequality is because a finite Orlicz norm implies sub-Gaussianity with a related bound on the concentration parameter, which implies related moment bounds; and the second inequality is because an ess sup *bound implies compact support, which implies sub-Gaussianity by Hoeffding's lemma. Proofs of these bounds are provided in §6.6.1. We also note that the reverse inequalities do not hold, even if one weakens them by an arbitrarily large amount. For example, for the first inequality, take $\mu = \delta_0$ and ν the Laplace distribution with density $\nu(x) = \frac{1}{2}\exp(-|x|)$ on $\mathbb{R}$; then, $W_p(\mu, \nu)$ is finite for any $1 \leq p < \infty$, but $W_{\psi_2}(\mu, \nu) = \infty$. And for the second inequality, take $\mu = \delta_0$ and $\nu = \mathcal{N}(0, 1)$; then $W_{\psi_2}(\mu, \nu)$ is finite but $W_\infty(\mu, \nu) = \infty$.*

Definition 6.3.5 (Orlicz–Wasserstein shifted Rényi divergence)**.** *For any Rényi order $q \geq 1$ and shift $w \geq 0$, the W_{ψ_2}-shifted Rényi divergence between probability distributions μ and ν is defined as*

$$\mathcal{R}_q^{(w)}(\mu \,\|\, \nu) := \inf_{\mu' \text{ s.t. } W_{\psi_2}(\mu, \mu') \leq w} \mathcal{R}_q(\mu' \,\|\, \nu) .$$

■ 6.3.2 Proof of Theorem 6.3.1

Here we describe how the standard shifted divergence analysis is modified when using shifts in W_{ψ_2} rather than W_∞, and how this modified argument leads to a proof of Theorem 6.3.1.

The shifted divergence technique—in both its original form and the new form here—is built upon two key lemmas. These two lemmas track how the shifted Rényi divergence evolves when both distributions are either (1) pushed forward through a Lipschitz map; or (2) convolved with Gaussian noise. These two lemmas are called the "contraction-reduction lemma[9]" and the "shift-reduction lemma".

[9] Although we use this name to be consistent with the previous literature on the shifted divergence technique, we note that this map need not be a contraction, i.e., the Lipschitz constant can be greater than 1.

The contraction-reduction lemma is the simpler of these two lemmas, and extends unchanged—in terms of both statement and proof—when the standard W_∞ shift is replaced by our proposed W_{ψ_2} shift. For completeness, we provide a brief proof.

Lemma 6.3.6 (New contraction-reduction lemma, for Orlicz–Wasserstein shifted Rényi)**.** *For any Rényi order $q \geq 1$, any shift $w \geq 0$, any Markov transition kernel P that is W_{ψ_2}-Lipschitz with parameter c, and any distributions $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$,*

$$\mathcal{R}_q^{(w)}\left(\mu P \,\|\, \nu P\right) \leq \mathcal{R}_q^{(w/c)}\left(\mu \,\|\, \nu\right) .$$

Proof. Let μ' be the surrogate for μ in $\mathcal{R}_q^{(w/c)}\left(\mu \,\|\, \nu\right)$. Then by definition, we have $W_{\psi_2}(\mu, \mu') \leq w/c$ and $\mathcal{R}_q\left(\mu' \,\|\, \nu\right) = \mathcal{R}_q^{(w/c)}\left(\mu \,\|\, \nu\right)$. Thus

$$\mathcal{R}_q^{(w)}(\mu P \,\|\, \nu P) \leq \mathcal{R}_q(\mu' P \,\|\, \nu P) \leq \mathcal{R}_q(\mu' \,\|\, \nu) = \mathcal{R}_q^{(w/c)}(\mu \,\|\, \nu)\,,$$

where the first step is because $W_{\psi_2}(\mu P, \mu' P) \leq c\, W_{\psi_2}(\mu, \mu') \leq c\,(w/c) = w$ by Lipschitzness of P; the second step is by the data-processing inequality for Rényi divergences (Lemma 2.2.19); and the third step is by construction of μ'. □

The shift-reduction lemma, however, requires substantial modification.

Lemma 6.3.7 (New shift-reduction lemma, for Orlicz–Wasserstein shifted Rényi)**.** *For any Rényi order $q \geq 1$, any noise variance $\sigma^2 > 0$, any initial shift $w \geq 0$, any shift increase $\delta \leq \sigma/\sqrt{(2q-1)\,(q-1)}$, and any distributions $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$,*

$$\begin{aligned}\mathcal{R}_q^{(w)}\big(\mu * \mathsf{normal}(0, \sigma^2 I_d) \,\big\|\, \nu * \mathsf{normal}(0, \sigma^2 I_d)\big) \\ \leq \mathcal{R}_{2q-1}^{(w+\delta)}(\mu \,\|\, \nu) + \frac{(2q-1)\,\delta^2}{2\sigma^2} \log 2\,.\end{aligned}$$

Proof. Case 1: initial shift $w = 0$. For shorthand, let γ denote $\mathsf{normal}(0, \sigma^2 I_d)$. We bound the Rényi divergence between $\mathrm{law}(X + Z)$ and $\mathrm{law}(Y + Z)$, where $X \sim \mu$, $Y \sim \nu$, and $Z \sim \gamma$. Let μ' be the surrogate for $\mathcal{R}_{2q-1}^{(\delta)}(\mu \,\|\, \nu)$, so that $\mathcal{R}_{2q-1}^{(\delta)}(\mu \,\|\, \nu) = \mathcal{R}_{2q-1}(\mu' \,\|\, \nu)$ and $W_{\psi_2}(\mu, \mu') \leq \delta$. Let $X' \sim \mu'$ be optimally coupled with $X \sim \mu$ with respect to the Orlicz–Wasserstein metric $W_{\psi_2}(\mu, \mu')$ so that

$$\iint \psi_2\Big(\frac{\|x - x'\|}{\delta}\Big)\, p_{X,X'}(\mathrm{d}x, \mathrm{d}x') \leq 1\,, \tag{6.2}$$

where here and henceforth we write p_η as shorthand for the law of a random variable η.

Note that $X + Z$ and $Y + Z$ are the result of the same function applied to the tuples $(X', X - X' + Z)$ and (Y, Z) respectively. Thus, by the data-processing inequality for Rényi divergences (Lemma 2.2.19),

$$\begin{aligned}\mathcal{R}_q(\mu * \gamma \parallel \nu * \gamma) &= \mathcal{R}_q\big(\operatorname{law}(X + Z) \,\big\|\, \operatorname{law}(Y + Z)\big) \\ &\le \mathcal{R}_q\big(\operatorname{law}(X', X - X' + Z) \,\big\|\, \operatorname{law}(Y, Z)\big)\,.\end{aligned}$$

By expanding the definition of Rényi divergence and applying Hölder's inequality, we bound this by

$$\begin{aligned}\cdots &= \frac{1}{q-1} \log \iint \Big(\frac{p_{X',X-X'+Z}(x',z)}{p_{Y,Z}(x',z)}\Big)^{q-1} p_{X',X-X'+Z}(\mathrm{d}x', \mathrm{d}z) \\ &= \frac{1}{q-1} \log \iint \Big(\frac{p_{X'}(x')}{p_Y(x')} \frac{p_{X-X'+Z|X'=x'}(z)}{p_{Z|Y=x'}(z)}\Big)^{q-1} p_{X-X'+Z|X'=x'}(\mathrm{d}z)\, p_{X'}(\mathrm{d}x') \\ &= \frac{1}{q-1} \log \iint \Big(\frac{\mu'(x')}{\nu(x')} \frac{p_{X-X'+Z|X'=x'}(z)}{\gamma(z)}\Big)^{q-1} p_{X-X'+Z|X'=x'}(\mathrm{d}z)\, \mu'(\mathrm{d}x') \\ &\le \underbrace{\frac{1}{2\,(q-1)} \log \int \Big(\frac{\mu'(x')}{\nu(x')}\Big)^{2\,(q-1)} \mu'(\mathrm{d}x')}_{①} \\ &\qquad + \underbrace{\frac{1}{2\,(q-1)} \log \iint \Big(\frac{p_{X-X'+Z|X'=x'}(z)}{\gamma(z)}\Big)^{2\,(q-1)} p_{X-X'+Z|X'=x'}(\mathrm{d}z)\, \mu'(\mathrm{d}x')}_{②}\,.\end{aligned}$$

By definition of Rényi divergence and then the construction of μ', the first term ① simplifies to

$$① = \mathcal{R}_{2q-1}(\mu' \parallel \nu) = \mathcal{R}^{(\delta)}_{2q-1}(\mu \parallel \nu)\,.$$

Writing $\gamma_z \coloneqq \mathsf{normal}(z, \sigma^2 I_d)$, the second term ② can be bounded as

$$\begin{aligned}② &= \frac{1}{2\,(q-1)} \log \int \exp\big(2\,(q-1)\,\mathcal{R}_{2q-1}(p_{X-X'+Z|X'=x'} \parallel \gamma)\big)\, \mu'(\mathrm{d}x') \\ &\le \frac{1}{2\,(q-1)} \log \iint \exp\big(2\,(q-1)\,\mathcal{R}_{2q-1}(p_{x-x'+Z} \parallel \gamma)\big)\, p_{X,X'}(\mathrm{d}x, \mathrm{d}x') \\ &= \frac{1}{2\,(q-1)} \log \iint \exp\big(2\,(q-1)\,\mathcal{R}_{2q-1}(\gamma_{x-x'} \parallel \gamma)\big)\, p_{X,X'}(\mathrm{d}x, \mathrm{d}x') \\ &= \frac{1}{2\,(q-1)} \log \iint \exp\Big(\frac{(q-1)\,(2q-1)\,\|x-x'\|^2}{\sigma^2}\Big)\, p_{X,X'}(\mathrm{d}x, \mathrm{d}x')\end{aligned}$$

$$\leq \frac{\delta^2\,(2q-1)}{2\sigma^2} \log \iint \exp\Big(\frac{\|x-x'\|^2}{\delta^2}\Big)\, p_{X,X'}(\mathrm{d}x, \mathrm{d}x')$$
$$\leq \frac{\delta^2\,(2q-1)}{2\sigma^2} \log 2\,.$$

Above, the first step is by definition of the Rényi divergence; the second step is by noting that $p_{X-X'+Z|X'=x'} = \int p_{x-x'+Z}\, p_{X|X'}(\mathrm{d}x \mid x')$ and using convexity of f-divergences (Lemma 2.2.22); the fourth step is by the closed-form expression for the Rényi divergence between Gaussians (Lemma 2.2.21); the fifth step is by the assumption that $\rho \coloneqq \delta^2\,(q-1)\,(2q-1)/\sigma^2 \leq 1$, which enables us to use Jensen's inequality to bound $\mathbb{E}[R^\rho] \leq \mathbb{E}[R]^\rho$ where $R \coloneqq \exp(\|X-X'\|^2/\delta^2)$; and the final step is by the property (6.2) of the coupling (X, X'). Combining these bounds on (1) and (2) completes the proof for case 1.

Case 2: initial shift $w > 0$. Let μ' denote the surrogate for $\mathcal{R}^{(w+\delta)}_{2q-1}(\mu \parallel \nu)$, so that $\mathcal{R}^{(w+\delta)}_{2q-1}(\mu \parallel \nu) = \mathcal{R}_{2q-1}(\mu' \parallel \nu)$ and $W_{\psi_2}(\mu, \mu') \leq w + \delta$. Let $X' \sim \mu'$ be optimally coupled with $X \sim \mu$ with respect to the Orlicz–Wasserstein metric $W_{\psi_2}(\mu, \mu')$ so that $\|X - X'\|_{\psi_2} = W_{\psi_2}(\mu, \mu') \leq w + \delta$. Decompose

$$X' = \underbrace{\tau\, X + (1-\tau)\, X'}_{X'_1} + \underbrace{\tau\,(X' - X)}_{X'_2}\,,$$

where $\tau \coloneqq \delta/(w+\delta)$. Then

$$\begin{aligned}
\mathcal{R}^{(w)}_q(\mu * \gamma \parallel \nu * \gamma) &\leq \mathcal{R}_q(p_{X'_1} * \gamma \parallel \nu * \gamma) \\
&\leq \mathcal{R}^{(\delta)}_q(p_{X'_1} \parallel \nu) + \frac{\delta^2\,(2q-1)}{2\sigma^2} \log 2 \\
&\leq \mathcal{R}_{2q-1}(\mu' \parallel \nu) + \frac{\delta^2\,(2q-1)}{2\sigma^2} \log 2 \\
&= \mathcal{R}^{(w+\delta)}_{2q-1}(\mu \parallel \nu) + \frac{\delta^2\,(2q-1)}{2\sigma^2} \log 2\,.
\end{aligned}$$

Above, the first step is by using $p_{X'_1} * \gamma$ as a surrogate for $\mu * \gamma$, since

$$\begin{aligned}
W_{\psi_2}(\mu * \gamma,\, p_{X'_1} * \gamma) &\leq W_{\psi_2}(\mu, p_{X'_1}) \leq \|X - X'_1\|_{\psi_2} \\
&= (1-\tau)\, \|X - X'\|_{\psi_2} \leq (1-\tau)\,(w+\delta) = w\,.
\end{aligned}$$

The second step is by using the result from case 1; the third step is by using μ' as a surrogate for $p_{X'_1}$, since $W_{\psi_2}(\mu', p_{X'_1}) \leq \|X' - X'_1\|_{\psi_2} = \|X'_2\|_{\psi_2} = \tau\, \|X - X'\|_{\psi_2} \leq \tau\,(w+\delta) = \delta$; and the final step is by construction of μ'. □

Remark 6.3.8. *Lemma 6.3.7 can be generalized to*

$$\begin{aligned}&\mathcal{R}_q^{(w)}\big(\mu * \mathsf{normal}(0,\sigma^2 I_d) \,\big\|\, \nu * \mathsf{normal}(0,\sigma^2 I_d)\big)\\ &\qquad \le \mathcal{R}^{(w+\delta)}_{(q+\lambda-1)/\lambda}(\mu \,\|\, \nu) + \frac{(q-\lambda)\,\delta^2}{2\,(1-\lambda)\,\sigma^2}\log 2\end{aligned} \tag{6.3}$$

for any $\lambda \in [0,1]$ and any $\delta \le (1-\lambda)\,\sigma\sqrt{\frac{2}{(q-1)\,(q-\lambda)}}$. Choosing $\lambda = 1/2$ recovers Lemma 6.3.7. Different choices of λ enable trading off the increase in the Rényi order in the first term against the penalty in the second term. The proof of this generalized bound is identical except for one change: replace the Cauchy–Schwarz inequality in the proof with Hölder's inequality $\int fg \le \left(\int |f|^{1/\lambda}\right)^{\lambda}\left(\int |g|^{1/(1-\lambda)}\right)^{1-\lambda}$.

We conjecture that (6.3) is tight in that for any input parameters q, σ^2, w, there exist distributions μ, ν such that this bound holds with equality when it is optimized over the knobs δ, λ. Details on this conjecture are provided in §6.6.2.

Remark 6.3.9 (Failure of shift reduction for weaker notions of shift)**.** *The Orlicz–Wasserstein metric is the "right" metric to use for the shifted Rényi analysis in the sense that (1) Lemma 6.3.7 holds with this Orlicz–Wasserstein shifted Rényi divergence; and (2) the Orlicz–Wasserstein distance at initialization is bounded for sampling algorithms (shown in Lemma 6.4.7).*

In contrast, (2) fails for all previous versions of the shifted Rényi divergence analysis, since they use W_∞ shifts. And (1) fails for other natural candidates of the Wasserstein metric for which the initialization distance is bounded. This includes the W_p metric for any finite p, as well as the Orlicz–Wasserstein metric for any Orlicz norm that is weaker than sub-Gaussian. See §6.6.3 for details. Finally, we remark that this discussion is tailored to the fact that we are analyzing Markov chains with Gaussian noise; if for example, this were replaced by Laplacian noise, then the right notion of shift would be the Orlicz–Wasserstein metric with the sub-exponential Orlicz norm, and our techniques would extend straightforwardly.

We now use Lemmas 6.3.6 and 6.3.7 to prove Theorem 6.3.1.

Proof of Theorem 6.3.1. Let $\delta = c^N\, W_{\psi_2}(\mu_0, \mu_0')$. By using, in order: the definition of the Markov chain update, Lemma 6.3.7 (in the form of Remark 6.3.8 with $\lambda = 0$), and then Lemma 6.3.6, we obtain

$$\begin{aligned}&\mathcal{R}_q(\mu_N \,\|\, \mu_N')\\ &\quad = \mathcal{R}_q\big((\mu_{N-1}P_{N-1}) * \mathsf{normal}(0,\sigma^2 I_d) \,\big\|\, (\mu_{N-1}'P_{N-1}) * \mathsf{normal}(0,\sigma^2 I_d)\big)\\ &\quad \le \mathcal{R}_\infty^{(\delta)}(\mu_{N-1}P_{N-1} \,\|\, \mu_{N-1}'P_{N-1}) + \frac{q\delta^2}{2\sigma^2}\log 2\end{aligned}$$

$$\leq \mathcal{R}_\infty^{(\delta/c)}(\mu_{N-1} \,\|\, \mu'_{N-1}) + \frac{q\delta^2}{2\sigma^2} \log 2 \,.$$

Note that the use of Lemma 6.3.7 is valid since $\delta \leq \sigma \sqrt{\frac{2}{q\,(q-1)}}$ by the assumption on N.

It suffices to show that the Rényi term in the above display vanishes. To this end, let Q_n denote the transition kernel for the n-th step of the Markov chain, i.e., $\rho Q_n = (\rho P_n) * \mathsf{normal}(0, \sigma^2 I_d)$. Clearly Q_n is W_{ψ_2}-Lipschitz with parameter c since $W_{\psi_2}(\rho Q_n, \rho' Q_n) \leq W_{\psi_2}(\rho P_n, \rho' P_n) \leq c\, W_{\psi_2}(\rho, \rho')$ for any distributions ρ, ρ'. Thus we may apply Lemma 6.3.7 $N-1$ times to argue that

$$\begin{aligned}
\mathcal{R}_\infty^{(\delta/c)}(\mu_{N-1} \,\|\, \mu'_{N-1}) &= \mathcal{R}_\infty^{(\delta/c)}(\mu_0 Q^{N-1} \,\|\, \mu'_0 Q^{N-1}) \\
&\leq \mathcal{R}_\infty^{(\delta/c^N)}(\mu_0 \,\|\, \mu'_0) \\
&= \mathcal{R}_\infty^{(W_{\psi_2}(\mu_0, \mu'_0))}(\mu_0 \,\|\, \mu'_0) \\
&\leq \mathcal{R}_\infty(\mu'_0 \,\|\, \mu'_0) \\
&= 0 \,.
\end{aligned}$$

Here, the third step is by the choice of δ, and the fourth step is by definition of $\mathcal{R}_\infty$. The proof is complete by combining the above displays. □

■ 6.4 Low-accuracy sampling with $O(\sqrt{d})$ complexity

The main result of the section is the first Rényi convergence guarantee for log-concave sampling that requires a number of first-order queries that scales in the dimension d only as $\sqrt{d}$. This improves over the state-of-the-art which has d^2 scaling. This result is formally stated as follows.

Theorem 6.4.1 (Low accuracy sampling with $O(\sqrt{d})$ complexity)**.** *Suppose that $\pi \propto \exp(-V)$ where V is α-strongly-convex and β-smooth, and let $0 < \varepsilon \lesssim \frac{1}{\sqrt{q}}$. There is a randomized algorithm that, given knowledge of the minimizer of V and access to*

$$N = \widetilde{O}\Big(\frac{\kappa^{3/2} d^{1/2} q^{1/2}}{\varepsilon}\Big)$$

gradient queries for V, outputs a random point in $\mathbb{R}^d$ with μ satisfying

$$\mathcal{R}_q(\mu \,\|\, \pi) \leq \varepsilon^2 \,.$$

Remark 6.4.2 (Extension to arbitrary initialization)**.** *The algorithm in Theorem 6.4.1 initializes at the Dirac distribution δ_{x^*}. This is reasonable because the*

cost of using gradient descent to compute x^* approximately, using the same first-order oracle access, is dominated by the cost of subsequently running the sampling algorithm. If the algorithm is initialized at some other point x, the runtime only increases by a logarithmic factor of $\log W_{\psi_2}(\delta_x, \pi)$, which is lower order unless x is exponentially far from x^*, since $W_{\psi_2}(\delta_x, \pi) \lesssim \sqrt{d/\alpha} + \|x - x^*\|$ by Lemma 6.4.7 and the triangle inequality.

While the results of [Che+18b; DR20; Ma+21; Zha+23] have also shown iteration complexities that scale in the dimension d as $\widetilde{O}(d^{1/2})$, a key difference is that these results do not hold for Rényi divergences. In particular, past work has only proven weaker mixing results in the Wasserstein metric or the KL divergence. As discussed in the introduction, Wasserstein and KL guarantees are insufficient for the purpose of warm-starting high-accuracy sampling algorithms—for this, it is essential to have guarantees in the more stringent Rényi divergence. (Note that Wasserstein bounds are weaker than KL bounds by Talagrand's T_2 inequality, and moreover KL bounds are weaker than Rényi bounds by monotonicity of Rényi divergences.) See §6.1.2 for a detailed discussion of this and of the many longstanding technical difficulties involved with establishing Rényi guarantees.

The algorithm we use is underdamped Langevin Monte Carlo (ULMC) with certain parameters (stated explicitly in the proof in §6.4.3). Background on this algorithm is recalled in §6.4.1. At a high level, the proof of Theorem 6.4.1 uses the weak triangle inequality for Rényi divergences to decompose the sampling error of ULMC into the following two terms, both measured in Rényi divergence:

1. The "bias" error between the stationary distribution of ULMC and the target distribution π.

2. The "discrete mixing" error of ULMC to its biased stationary distribution.

The bias error (1) is readily handled by recent results such as [GT20; Zha+23]. Bounding the discrete mixing error (2) is the key technical challenge; see §6.1.2 for a detailed discussion of the technical obstacles related to this, and the connections to open problems about hypocoercivity in the PDE literature. The key contribution of this section is to bound this quantity in Theorem 6.4.4 below. To do this, we use the new shifted divergence technique developed in §6.3.

The section is organized as follows. In §6.4.1 we recall relevant background about ULMC, in §6.4.2 we bound the discrete mixing error of ULMC, and in §6.4.3 we use this to prove Theorem 6.4.1. Remark about notation in this section: the ULMC algorithm studied naturally operates on the augmented Hamiltonian state space $\mathbb{R}^{2d}$, so in this section we use boldface to denote probability distributions on $\mathbb{R}^{2d}$ from non-boldfaced distributions on $\mathbb{R}^d$. For example, we write $\boldsymbol{\pi}$ to denote the target distribution $\pi \otimes \mathsf{normal}(0, I_d)$ in this augmented space.

6.4.1 Background on underdamped Langevin Monte Carlo

The exposition in this subsection is based on the corresponding section in [Che23]; we refer the interested reader there for further details.

Underdamped Langevin diffusion. Studied since the work of Kolmogorov in the 1930s [Kol34], the underdamped Langevin diffusion—sometimes also called the kinetic Langevin diffusion—is the solution to the stochastic differential equation

$$\begin{aligned} \mathrm{d}X_t &= Y_t \,\mathrm{d}t\,, \\ \mathrm{d}Y_t &= -\nabla V(X_t)\,\mathrm{d}t - \gamma Y_t\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t\,, \end{aligned} \tag{6.4}$$

where $(B_t)_{t\geq 0}$ is a standard d-dimensional Brownian motion. Analogous to the classical theory of convex optimization, here the auxiliary state variable Y_t has the physical interpretation of momentum, and the linking parameter γ has the physical interpretation of friction. A related interpretation of the underdamped Langevin diffusion is as a variant of the idealized Hamiltonian Monte Carlo algorithm, in which the momentum is refreshed continuously rather than periodically. The stationary distribution for this SDE is the joint distribution

$$\boldsymbol{\pi}(x,y) \propto \exp\Big(-V(x) - \frac{1}{2}\,\|y\|^2\Big)\,. \tag{6.5}$$

In this section, we slightly abuse notation by using the boldface $\boldsymbol{\pi}$ to distinguish this joint distribution on $\mathbb{R}^{2d}$ from the target distribution $\pi \propto \exp(-V)$ on $\mathbb{R}^d$; of course the latter is the x-marginalization of the former.

A major obstacle for analyzing the convergence of the underdamped Langevin diffusion is that this process exhibits hypocoercive dynamics, i.e., the standard Markov semigroup approach based on isoperimetric inequalities does not work. It is a longstanding question in PDE theory to develop general tools for establishing fast convergence of hypocoercive dynamics; see §6.1.2. We bypass these issues by instead developing tools for analyzing a discrete-time version of this diffusion.

Underdamped Langevin Monte Carlo. There are several ways to discretize the underdamped Langevin diffusion. Perhaps the simplest way is the Euler–Maruyama discretization, as is standard for defining (unadjusted) Langevin Monte Carlo. However, for these underdamped Langevin dynamics, there is a better discretization which dates back at least to 1980 [EB80], namely:

$$\begin{aligned} \mathrm{d}X_t &= Y_t \,\mathrm{d}t \\ \mathrm{d}Y_t &= -\nabla V(X_{nh})\,\mathrm{d}t - \gamma Y_t\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t\,, \end{aligned}$$

for $t \in [nh, (n+1)h]$. This process is called underdamped Langevin Monte Carlo (ULMC) or kinetic Langevin Monte Carlo; we use the former term in this chapter.

The point of this discretization is that since the gradient is refreshed periodically rather than continuously, the SDE is linear within these periods, and thus can be integrated exactly in closed form (see, e.g., [Che+18b, Appendix A]). This is called an "exponential integrator" in the lingo of numerical analysis, and is formalized in the following lemma.

Lemma 6.4.3 (Explicit Gaussian law for ULMC iterates)**.** *Conditioned on the previous iterate (X_{nh}, Y_{nh}), the law of $(X_{(n+1)h}, Y_{(n+1)h})$ is the Gaussian distribution* $\mathsf{normal}(F(X_{nh}, Y_{nh}), \Sigma \otimes I_d)$ *where*

$$F(x,y) \coloneqq \big(x + \gamma^{-1}(1-a)\,y - \gamma^{-1}(h - \gamma^{-1}(1-a))\,\nabla V(x), \\ ay - \gamma^{-1}(1-a)\,\nabla V(x)\big)$$

and

$$\Sigma \coloneqq \begin{bmatrix} \frac{2}{\gamma}(h - \frac{2}{\gamma}(1-a) + \frac{1}{2\gamma}(1-a^2)) & \frac{1}{\gamma}(1-2a+a^2) \\ \frac{1}{\gamma}(1-2a+a^2) & 1-a^2 \end{bmatrix}.$$

Above, we use the notational shorthand $a \coloneqq \exp(-\gamma h)$.

In the rest of this section, we write $\boldsymbol{P} \coloneqq \boldsymbol{P}_{h,\gamma}$ to denote the Markov transition kernel on $\mathbb{R}^{2d}$ that corresponds to an iteration of ULMC. We suppress the dependence of $\boldsymbol{P}$ on the parameters h and γ for simplicity of notation.

■ 6.4.2 Discrete mixing of underdamped Langevin Monte Carlo

Theorem 6.4.4 (Discrete mixing of ULMC)**.** *Suppose that $\pi \propto \exp(-V)$ where V is α-strongly-convex and β-smooth. Let $\boldsymbol{P}$ denote the Markov transition kernel for ULMC when run with friction parameter $\gamma = \sqrt{2\beta}$ and step size $h \lesssim 1/(\kappa\sqrt{\beta})$. Then, for any target accuracy $0 < \varepsilon \le \sqrt{\frac{\log 2}{q-1}}$, any Rényi order $q \ge 1$, and any two initial distributions $\boldsymbol{\mu_0}, \boldsymbol{\mu'_0} \in \mathcal{P}(\mathbb{R}^{2d})$,*

$$\mathcal{R}_q(\boldsymbol{\mu_0}\boldsymbol{P}^N \,\|\, \boldsymbol{\mu'_0}\boldsymbol{P}^N) \le \varepsilon^2\,,$$

if the number of ULMC iterations is

$$N \gtrsim \frac{\sqrt{\beta}}{\alpha h} \log\Big(\frac{q\, W_{\psi_2}^2(\mathcal{M}_\# \boldsymbol{\mu_0}, \mathcal{M}_\# \boldsymbol{\mu'_0})}{\beta^{1/2}\,\varepsilon^2\,h^3}\Big)\,,$$

where $\mathcal{M}$ is the linear map defined in (6.6).

We remark that in a typical use case of this discrete mixing result, $\boldsymbol{\mu_0}$ is initialized at a product distribution of the form $\mu_0 \otimes \mathsf{normal}(0, I_d)$, and is compared to the target distribution $\boldsymbol{\mu_0'} = \boldsymbol{\pi} = \pi \otimes \mathsf{normal}(0, I_d)$. In this setting, the Orlicz–Wasserstein metric in the upper bound can be simplified to $W_{\psi_2}(\mathcal{M}_{\#}\boldsymbol{\mu_0}, \mathcal{M}_{\#}\boldsymbol{\mu_0'}) \le 2W_{\psi_2}(\mu_0, \pi)$.

To prove Theorem 6.4.4, we appeal to our shifted divergence technique developed in §6.3. This requires analyzing the ULMC iterates in a *twisted norm*, since an iteration of the ULMC algorithm (or more precisely, the mean-shifting function F defined in Lemma 6.4.3) is not contractive w.r.t. the standard Euclidean norm. This twisted norm is the Euclidean norm after the change of coordinates

$$(u, v) \coloneqq \mathcal{M}(x, y) \coloneqq \left(x, x + \frac{2}{\gamma} y\right). \tag{6.6}$$

In these new coordinates, the mean of the next iterate of ULMC started at (u, v) is $\bar{F}(u, v)$, where $\bar{F} = \mathcal{M} \circ F \circ \mathcal{M}^{-1}$. Since $\mathcal{M}^{-1}(u, v) = (u, \frac{\gamma}{2}(v - u))$, we can explicitly write

$$\begin{aligned} \bar{F}(u, v) = \Big(&u + \frac{1 - a}{2}(v - u) - \frac{h - \gamma^{-1}(1 - a)}{\gamma} \nabla V(u), \\ &u + \frac{1 + a}{2}(v - u) - \frac{h + \gamma^{-1}(1 - a)}{\gamma} \nabla V(u)\Big). \end{aligned} \tag{6.7}$$

By Lemma 6.4.3, conditioned on (U_{nh}, V_{nh}), the law of $(U_{(n+1)h}, V_{(n+1)h})$ is the Gaussian distribution $\mathsf{normal}(\bar{F}(U_{nh}, V_{nh}), \bar{\Sigma} \otimes I_d)$ where

$$\bar{\Sigma} = \mathcal{M} \Sigma \mathcal{M}^{\mathsf{T}}. \tag{6.8}$$

We make use of the following two helper lemmas about the dynamics of ULMC in this twisted norm. The first helper lemma shows that the ULMC algorithm sends two iterates to Gaussians with means that are closer in the twisted norm than the original iterates. Since the two Gaussians have the same covariance $\bar{\Sigma} \otimes I_d$, this implies that the ULMC Markov transition kernel P is contractive in W_{ψ_2}, which will allow us to use our new shifted divergence technique from §6.3. This lemma first appeared in the recent paper [Zha+23]; for completeness, a proof is provided in §6.7.1.

Lemma 6.4.5 (Contractivity of ULMC in the twisted norm)**.** *Suppose that $\pi \propto \exp(-V)$ where V is α-strongly-convex and β-smooth. For step size $h \lesssim 1$ and friction parameter $\gamma = \sqrt{2\beta}$, the function $\bar{F} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ defined in (6.7) is a contraction with*

$$\|\bar{F}\|_{\mathrm{Lip}} \le 1 - \frac{\alpha}{\sqrt{2\beta}} h + O(\beta h^2).$$

The second helper lemma estimates the noise of ULMC in this twisted norm. The proof is an explicit computation and is provided in §6.7.2.

Lemma 6.4.6 (Noise of ULMC in the twisted norm)**.** *Suppose that* $h \lesssim 1/\gamma$. *Then the matrix* $\bar{\Sigma}$ *defined in* (6.8) *satisfies*

$$\lambda_{\min}(\bar{\Sigma}) = \frac{\gamma h^3}{6}\left(1 - O(\gamma h)\right).$$

Armed with Lemmas 6.4.5 and 6.4.6, we are now ready to prove Theorem 6.4.4.

Proof of Theorem 6.4.4. We let $\{\boldsymbol{\mu_n}\}_{n\geq 0}$ and $\{\boldsymbol{\mu'_n}\}_{n\geq 0}$ denote the two processes $\boldsymbol{\mu_n} = \boldsymbol{\mu_0}\boldsymbol{P}^n$ and $\boldsymbol{\mu'_n} = \boldsymbol{\mu'_0}\boldsymbol{P}^n$ obtained by running ULMC from initialization distributions $\boldsymbol{\mu_0}$ and $\boldsymbol{\mu'_0}$, respectively. Define twisted processes $\{\boldsymbol{\nu_n}\}_{n\geq 0}$ and $\{\boldsymbol{\nu'_n}\}_{n\geq 0}$ by $\boldsymbol{\nu_n} = \mathcal{M}_{\#}\boldsymbol{\mu_n}$ and $\boldsymbol{\nu'_n} = \mathcal{M}_{\#}\boldsymbol{\mu'_n}$, where $\mathcal{M}$ is the change-of-coordinates matrix defined in (6.6). Since $\mathcal{M}$ is invertible, applying the data-processing inequality for Rényi divergences (Lemma 2.2.19) in both directions implies

$$\mathcal{R}_q(\boldsymbol{\mu_N} \,\|\, \boldsymbol{\mu'_N}) = \mathcal{R}_q(\boldsymbol{\nu_N} \,\|\, \boldsymbol{\nu'_N}).$$

We now show that the latter term is at most ε^2. For shorthand, define $\lambda \coloneqq \lambda_{\min}(\bar{\Sigma})$ and define $\boldsymbol{Q}$ to be the Markov operator given by $\boldsymbol{\nu Q} = \bar{F}_{\#}\boldsymbol{\nu} * \mathsf{normal}(0, \bar{\Sigma} \otimes I_d - \lambda I_{2d})$. Then by Lemma 6.4.3 and a change of measure, the law of $\boldsymbol{\nu_{n+1}}$ is

$$\bar{F}_{\#}\boldsymbol{\nu_n} * \mathsf{normal}(0, \bar{\Sigma} \otimes I_d) = \boldsymbol{\nu_n Q} * \mathsf{normal}(0, \lambda I_{2d}).$$

Similarly, the law of $\boldsymbol{\nu'_{n+1}}$ is

$$\bar{F}_{\#}\boldsymbol{\nu'_n} * \mathsf{normal}(0, \bar{\Sigma} \otimes I_d) = \boldsymbol{\nu'_n Q} * \mathsf{normal}(0, \lambda I_{2d}).$$

Thus, letting c denote the Lipschitz constant of the Markov operator $\boldsymbol{Q}$ w.r.t. the W_{ψ_2} metric, we may invoke[10] the new shifted divergence result (Theorem 6.3.1) to bound

$$\mathcal{R}_q(\boldsymbol{\nu_N} \,\|\, \boldsymbol{\nu'_N}) \leq c^{2N}\,\frac{q}{2\lambda}\,W^2_{\psi_2}(\boldsymbol{\nu}_0, \boldsymbol{\nu}'_0). \tag{6.9}$$

We now use the two helper lemmas to quantify the various terms in (6.9). First, by a simple coupling argument and then an application of Lemma 6.4.5, the Markov operator $\boldsymbol{Q}$ is W_{ψ_2}-contractive with parameter c, where

$$c \leq \|\bar{F}\|_{\mathrm{Lip}} \leq \exp\Big(-\Omega\Big(\frac{\alpha h}{\sqrt{\beta}}\Big)\Big).$$

[10]The application of this result requires the number of iterations N to be large enough that the right hand side of (6.9), later set to ε^2, is at most $(\log 2)/(q-1)$. But this holds by assumption.

Second, because ULMC is run with step size $h \lesssim 1/\gamma$, Lemma 6.4.6 implies

$$\lambda = \Omega\big(\sqrt{\beta}\, h^3\big)\,.$$

Therefore, by combining the above displays, we conclude that

$$\mathcal{R}_q(\boldsymbol{\mu}_{\boldsymbol{N}} \,\|\, \boldsymbol{\mu}'_{\boldsymbol{N}}) = \mathcal{R}_q(\boldsymbol{\nu}_{\boldsymbol{N}} \,\|\, \boldsymbol{\nu}'_{\boldsymbol{N}}) \lesssim \exp\bigg(-\Omega\Big(\frac{\alpha h}{\sqrt{\beta}}\, N\Big)\bigg)\, \frac{q}{\sqrt{\beta}\, h^3}\, W_{\psi_2}^2(\mathcal{M}_{\#}\boldsymbol{\mu}_{\mathbf{0}}, \mathcal{M}_{\#}\boldsymbol{\mu}'_{\mathbf{0}})\,.$$

Setting this bound to ε^2 and solving for N completes the proof. □

■ 6.4.3 Warm start with underdamped Langevin Monte Carlo

We begin by bounding the distance to the target at initialization. We emphasize that this initialization is *not* a warm start. Indeed, this initialization is even weaker than what is typically called a "feasible start" in the literature (namely $\mathsf{normal}(x^*, \beta^{-1} I_d)$), and moreover can be further relaxed to an arbitrary initialization x_0 so long as the distance between x_0 and the mode x^* of the target distribution is sub-exponentially large (since our final bound depends only logarithmically on this distance).

Lemma 6.4.7 (Orlicz–Wasserstein distance at initialization)**.** *Suppose that $\pi \propto \exp(-V)$ where V is α-strongly convex. Let x^* denote the minimizer of V. Then*

$$W_{\psi_2}(\delta_{x^*}, \pi) \le 6\sqrt{d/\alpha}\,.$$

Proof. Let $X \sim \pi$ and define $Y \coloneqq \|X - x^*\|$. By definition of the Orlicz–Wasserstein metric,

$$W_{\psi_2}(\delta_{x^*}, \pi) = \|X - x^*\|_{\psi_2} = \|Y\|_{\psi_2} = \inf\bigg\{\lambda > 0 \;:\; \mathbb{E}\exp\Big(\frac{Y^2}{\lambda^2}\Big) \le 2\bigg\}\,.$$

By the elementary inequality $(a+b)^2 \le 2\,(a^2 + b^2)$, we can bound

$$Y^2 = (Y - \mathbb{E}\,Y + \mathbb{E}\,Y)^2 \le 2\,\mathbb{E}[Y]^2 + 2\,(Y - \mathbb{E}\,Y)^2\,.$$

Thus

$$\mathbb{E}\exp\Big(\frac{Y^2}{\lambda^2}\Big) \le \exp\Big(\frac{2\,\mathbb{E}[Y]^2}{\lambda^2}\Big) \cdot \mathbb{E}\exp\Big(\frac{2\,(Y - \mathbb{E}\,Y)^2}{\lambda^2}\Big)\,.$$

For the first term, use the basic inequality $\mathbb{E}[Y]^2 \le \mathbb{E}[Y^2]$ and a standard second-moment-type bound for strongly log-concave distributions (Lemma 2.2.13) to obtain

$$\exp\Big(\frac{2\,\mathbb{E}[Y]^2}{\lambda^2}\Big) \le \exp\Big(\frac{2d}{\alpha\lambda^2}\Big)\,.$$

For the second term, we use sub-Gaussian concentration. Specifically, combining the log-Sobolev inequality on π with the fact that $Y = \|X - x^*\|$ is a 1-Lipschitz function of $X \sim \pi$, it follows that $Y - \mathbb{E}\, Y$ is sub-Gaussian with variance proxy $1/\alpha$ (Lemma 2.2.9). Therefore, by a standard moment generating function bound for the square of a sub-Gaussian random variable (see, e.g., [BLM13, §2.3]),

$$\mathbb{E} \exp\Big(\frac{2\,(Y - \mathbb{E}\, Y)^2}{\lambda^2}\Big) \le 2^{16/(\alpha\lambda^2)}\,,$$

for any $\lambda^2 \ge 16/\alpha$. By combining the above displays and setting $\lambda = \sqrt{32d/\alpha}$, we conclude that

$$\mathbb{E} \exp\Big(\frac{Y^2}{\lambda^2}\Big) \le \exp\Big(\frac{2d}{\alpha\lambda^2} + \frac{16 \log 2}{\alpha\lambda^2}\Big) \le \exp(\log 2) = 2\,.$$

Therefore this choice of λ is an upper bound on $W_{\psi_2}(\delta_{x^*}, \pi)$. □

The second lemma uses Girsanov's theorem to bound the bias of ULMC. Here, we build upon recent advances in the literature on Rényi discretization of stochastic processes. Beginning with the works [GT20; EHZ22] and culminating in the results of §3, it is now understood that the Girsanov discretization technique leads to bias bounds for LMC in Rényi divergence matching prior results which only held for weaker divergences, and yet remains flexible enough to cover varying assumptions. The recent paper [Zha+23] extends this technique for ULMC. Since the results of [Zha+23] hold under more general assumptions at the expense of a more involved analysis, and in the interest of keeping our derivations more self-contained, in §6.7.3 we simplify and streamline the Girsanov argument of [Zha+23] for our setting of interest. In order to clarify where the $d^{1/2}$ comes from, we write the final bound in terms of the total elapsed continuous time $T = Nh$ rather than the number of iterations N.

Lemma 6.4.8 (Bias of ULMC)**.** *Suppose that V is α-strongly-convex and β-smooth. Let $\boldsymbol{\pi}(x, y) \propto \exp(-V(x) - \frac{1}{2}\|y\|^2)$, and let $\boldsymbol{P}$ denote the Markov transition kernel corresponding to an iteration of ULMC with friction parameter $\gamma \asymp \sqrt{\beta}$ and step size $h \lesssim \frac{1}{\beta^{3/4} d^{1/2} q\,(T \log N)^{1/2}}$, where N is the total number of iterations and $T = Nh$ is the total elapsed time. Then,*

$$\mathcal{R}_q(\boldsymbol{\pi}\boldsymbol{P}^N \,\|\, \boldsymbol{\pi}) \lesssim \beta^{3/2} d h^2 q T\,.$$

Proof of Theorem 6.4.1. Consider the following algorithm: run ULMC for N iterations from initialization $\boldsymbol{\mu_0} = \delta_{x^*} \otimes \mathsf{normal}(0, I_d)$ to obtain an iterate $(X, Y) \in \mathbb{R}^{2d}$, and then output X. We show that for a certain setting of the ULMC parameters, this algorithm satisfies the guarantees of the theorem.

To this end, given a joint distribution $\boldsymbol{\nu} \in \mathcal{P}(\mathbb{R}^{2d}) \cong \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$, let $(\Pi^x)_{\#}\boldsymbol{\nu} \in \mathcal{P}(\mathbb{R}^d)$ denote the marginal on the first d coordinates, where $\Pi^x : \mathbb{R}^{2d} \to \mathbb{R}^d$ maps $(x, y) \mapsto x$. Then in particular, the law of the algorithm's output is $\mu \coloneqq (\Pi^x)_{\#}(\boldsymbol{\mu_0}\boldsymbol{P}^N)$, and the target distribution is $\pi = (\Pi^x)_{\#}\boldsymbol{\pi}$. Thus by the data-processing inequality for Rényi divergences (Lemma 2.2.19), the sampling error of the algorithm is bounded by

$$\mathcal{R}_q(\mu \,\|\, \pi) \le \mathcal{R}_q(\boldsymbol{\mu_0}\boldsymbol{P}^N \,\|\, \boldsymbol{\pi})\,.$$

By the weak triangle inequality for Rényi divergence (Lemma 2.2.23), we can further bound this by

$$\mathcal{R}_q(\boldsymbol{\mu_0}\boldsymbol{P}^N \,\|\, \boldsymbol{\pi}) \le \frac{q-1/2}{q-1}\,\mathcal{R}_{2q}(\boldsymbol{\mu_0}\boldsymbol{P}^N \,\|\, \boldsymbol{\pi}\boldsymbol{P}^n) + \mathcal{R}_{2q-1}(\boldsymbol{\pi}\boldsymbol{P}^N \,\|\, \boldsymbol{\pi})\,. \tag{6.10}$$

The coefficient $(q-1/2)/(q-1)$ can be crudely bounded by 2, say, since it suffices to bound the Rényi divergence error for $q \ge 3/2$ (indeed, monotonicity of Rényi divergences in the order q then implies the same bound for $q < 3/2$).

Now, by combining our discrete mixing result for ULMC (Theorem 6.4.4), our initialization bound (Lemma 6.4.7), and the ULMC bias bound (Lemma 6.4.8), we conclude that

$$\mathcal{R}_q(\boldsymbol{\mu_0}\boldsymbol{P}^N \,\|\, \boldsymbol{\pi}) \le \varepsilon^2\,, \tag{6.11}$$

if ULMC is run with friction parameter γ, step size h, and iteration complexity N that satisfy:

$$\gamma = \sqrt{2\beta} \quad \text{and} \quad h \lesssim \frac{\varepsilon}{\beta^{3/4} d^{1/2} q^{1/2} T^{1/2}} \quad \text{and} \quad N \gtrsim \frac{\sqrt{\beta}}{\alpha h} \log\Big(\frac{dq}{\alpha \beta^{1/2} \varepsilon^2 h^3}\Big)\,.$$

By recalling that $T \coloneqq Nh$, solving for these choices of parameters, and omitting logarithmic factors, we conclude that it suffices to run ULMC with the following choices of parameters:

$$\gamma = \sqrt{2\beta} \qquad \text{and} \qquad h = \widetilde{\Theta}\Big(\frac{\varepsilon \alpha^{1/2}}{\beta d^{1/2} q^{1/2}}\Big) \qquad \text{and} \qquad N = \widetilde{\Theta}\Big(\frac{\kappa^{3/2} d^{1/2} q^{1/2}}{\varepsilon}\Big)\,. \quad \square$$

■ 6.5 High-accuracy sampling with $O(\sqrt{d})$ complexity

Establishing fast mixing results for MALA is a longstanding problem. As detailed in §6.1, recent breakthroughs have made it clear that the key barrier for fast mixing of MALA is the question of warm starts. In this section, we use the faster

low-accuracy sampling result developed in §6.4 to efficiently warm start MALA. This leads to the fastest known high-accuracy sampling algorithms not only in strongly log-concave settings (details in §6.5.1), but also in weakly-log-concave and isoperimetric, non-log-concave settings (details in §6.5.2), for which we improve over state-of-the-art query complexity results by a factor of $\sqrt{d}$.

6.5.1 Strongly-log-concave setting

Here we improve the query complexity for high-accuracy sampling from a strongly log-concave target distribution to $\widetilde{O}(\kappa d^{1/2} \operatorname{polylog}(1/\varepsilon))$. This result is formally stated as follows.

Theorem 6.5.1 (Faster high-accuracy sampler for well-conditioned targets)**.** *Suppose that $\pi \propto \exp(-V)$ where V is α-strongly-convex and β-smooth. There is an algorithm with randomized runtime that, given knowledge of the minimizer of V and access to N first-order queries for V, outputs a random point in $\mathbb{R}^d$ with law μ satisfying $\mathsf{d}(\mu \,\|\, \pi) \le \varepsilon$ for any of the following metrics:*

$$\mathsf{d} \in \{\mathsf{TV}, \sqrt{\mathsf{KL}}, \sqrt{\chi^2}, \sqrt{\alpha}\, W_2\}\,.$$

Moreover, for any $\delta \in (0,1)$ with probability at least $1-\delta$, the number of queries made satisfies

$$N \le \widetilde{O}\Big(\kappa d^{1/2} \log^4 \max\Big\{\frac{1}{\varepsilon}, \log\frac{1}{\delta}\Big\}\Big)\,.$$

In analogy to the familiar concept from algorithm design for deterministic problems [Cor+09], the algorithm in Theorem 6.5.1 may be called a "Las Vegas" algorithm because it has a randomized runtime which is small with high probability. The fact that this runtime is randomized is not an issue in practice because the iteration complexity depends on a quantity that is efficiently estimable during the execution of the algorithm.

In the rest of this subsection, we overview the algorithm in Theorem 6.5.1 and its analysis; see §6.8 for full technical details. This algorithm combines three algorithms as building blocks: ULMC, MALA, and the proximal sampler algorithm. Let us explain this by building up to the full complexity in two steps—both because this will motivate why all three algorithmic components are needed, and also because this is how our analysis actually proceeds.

Weak version of Theorem 6.5.1 (full details in §6.8.1). First, consider the following simplified version of the algorithm in Theorem 6.5.1 which is only comprised of two algorithmic components: run ULMC, and then use this as a warm start for MALA.

In order to argue that this two-phase algorithm mixes rapidly, we crucially use our result from §6.4 to guarantee that ULMC mixes to constant Rényi divergence error in a number of iterations that scales in the dimension d as $\widetilde{O}(d^{1/2})$ rather than $\widetilde{O}(d)$. This allows us to provide an algorithm for warm-starting MALA which is not significantly slower than MALA is when initialized from a warm start. In other words, this lets us exploit, for the first time, the results of §5 and [WSC22] which show that the mixing time of MALA scales in the dimension as $\widetilde{O}(d^{1/2})$ rather than $\widetilde{O}(d)$ when it is initialized at a warm start rather than a feasible start. We recall that such an improvement *cannot* be obtained without the warm start, due to the lower bound of [LST21a]. This leads to a final runtime of roughly $\widetilde{O}(\kappa^{3/2}d^{1/2} + \kappa d^{1/2}\,\mathrm{polylog}(1/\varepsilon))$; a formal statement is given in Theorem 6.8.1.

However, while this simple combination of ULMC and MALA achieves the desired dependence on the dimension d, it leads to a suboptimal dependence on the condition number κ, namely $\widetilde{O}(\kappa^{3/2})$ rather than $\widetilde{O}(\kappa)$. This worsened dependence in κ arises from the state-of-the-art bounds on the discretization of ULMC [Zha+23]. For full details on this weak version of Theorem 6.5.1, see §6.8.1.

The full version of Theorem 6.5.1 (full details in §6.8.2). In order to improve the condition number dependence of the weak version of Theorem 6.5.1, we require an extra algorithmic component: the recently proposed proximal sampler algorithm. See §4 for background on this proximal sampler algorithm. Briefly, this algorithm reduces the problem of sampling a strongly log-concave distribution with condition number κ, to the the problem of sampling $\widetilde{O}(\kappa)$ related strongly log-concave distributions each with constant condition number. The upshot is that the latter can be accomplished in the desired $\widetilde{O}(\kappa d^{1/2}\,\mathrm{polylog}(1/\varepsilon))$ runtime by using the weak version of the algorithm since each sampling subproblem is well-conditioned.

Mixing in Rényi divergence. While the main conceptual innovation here is the high-level strategy of combining these three algorithmic building blocks, we remark that an additional technical obstacle for proving Theorem 6.5.1 is showing mixing in more stringent notions of distance than TV. See the discussion in §6.1.2. Indeed, while our new ULMC result proves fast mixing in Rényi divergence, existing results on MALA and its combination with the proximal sampler are limited to TV. We boost this mixing in TV to Rényi divergences (and thus all the other desired metrics by standard comparison inequalities) using two additional ideas.

The first improves mixing bounds for the proximal sampler from TV to Rényi divergence. To do this, we control the propagation of error when each step of the proximal sampler algorithm is performed approximately in Rényi divergence. As we show, this is readily accomplished by appealing to the "strong composition rule" of Rényi divergences from the differential privacy literature.

The second improves mixing bounds for MALA from TV to Rényi divergence. We accomplish this by further exploiting the fact that MALA is warm started in Rényi divergence. Note that this means we use the Rényi warm start in two ways: first to show that MALA mixes fast in TV, which is what we can conclude from the above argument and appealing to §5 and [WSC22]; and second, to boost the TV bound at the final iterate to a more stringent bound. We isolate this TV-to-Rényi boosting technique in the following simple lemma as it may be of independent interest: indeed, since it uses the TV mixing bound in an entirely black-box way, this lemma may be useful for establishing Rényi mixing of other warm-started algorithms. This lemma improves upon Lemma 5.6.16 because that result required a warm-start in $\mathcal{R}_\infty$ which is currently unavailable algorithmically, whereas this lemma here only requires the weaker condition of a warm start in a Rényi divergence of finite order (stated here with $q = 3$ for simplicity).

Lemma 6.5.2 (Boosting TV to Rényi mixing given Rényi warm start)**.** *Let P be a Markov transition kerrnel which has stationary distribution π. Consider running P from any initialization distribution for N steps to obtain a distribution $\mu_N \coloneqq \mu_0 P^N$. Then*

$$\chi^2(\mu_N \,\|\, \pi) \le \sqrt{\mathsf{TV}(\mu_N, \pi) \cdot \left(\exp(2\mathcal{R}_3(\mu_0 \,\|\, \pi)) + 1\right)}\,.$$

See §4 and §5 for background on the proximal sampler and MALA, respectively; and see §6.8.1 and §6.8.2 for proofs of the weak version and full version of Theorem 6.5.1, respectively.

6.5.2 Extensions to weakly-log-concave and non-log-concave settings

Our faster algorithm for sampling from well-conditioned targets (Theorem 6.5.1) yields faster samplers for a variety of other settings, due essentially to the reductions in §4. We present here several such extensions that concern target distributions which satisfy isoperimetric inequalities, which is quite flexible in the sense that this allows for non-log-concavity and also is preserved under, e.g., bounded perturbations and Lipschitz mappings. See §2.2 for background on these isoperimetric inequalities.

A comment on notation for these isoperimetric settings: we still use the condition number κ to denote the ratio $\kappa = \beta/\alpha$, but now α denotes the (inverse) parameter of an isoperimetric bound, rather than the parameter for strong convexity. The motivation behind this notation is that α-strong-convexity implies the log-Sobolev inequality with parameter $1/\alpha$, which in turn implies the Poincaré inequality with parameter $1/\alpha$ (see Lemma 2.2.8).

In this section, for simplicity of exposition, in addition to the first-order oracle for V we also assume access to a prox oracle which can compute the proximal operator for V with step size $h = \frac{1}{2\beta}$, namely for any $y \in \mathbb{R}^d$ the oracle returns return $\mathrm{prox}_{hV}(y) := \arg\min_{x\in\mathbb{R}^d}\{V(x) + \frac{1}{2h}\|y - x\|^2\}$. Note that for this choice of step size, computing the proximal operator is a strongly convex and smooth optimization problem with condition number $O(1)$, so this can be done with off-the-shelf optimization methods such as gradient descent. We emphasize, however, that this assumption is only made to ease the presentation of the results, and more detailed results without this assumption are provided in [AC23].

Theorem 6.5.3 (Faster high-accuracy sampling from LSI targets)**.** *Suppose that $\pi \propto \exp(-V)$ satisfies $1/\alpha$-LSI and that V is β-smooth. There is an algorithm that, given access to a first-order + prox oracle for V and initialized at μ_0, outputs a random point with law μ satisfying $\mathsf{d}(\mu \,\|\, \pi) \le \varepsilon$ for any of the following metrics:*

$$\mathsf{d} \in \{\mathsf{TV}, \sqrt{\mathsf{KL}}, \sqrt{\chi^2}, \sqrt{\alpha}\, W_2\}\,,$$

using at most

$$N = \widetilde{O}\bigg(\kappa d^{1/2}\, \log\Big(\frac{\mathcal{R}_2(\mu_0 \,\|\, \pi)}{\varepsilon^2}\Big)\, \log^3\Big(\frac{1}{\varepsilon}\Big)\bigg) \qquad \textit{queries}\,.$$

Theorem 6.5.4 (Faster high-accuracy sampling from PI targets)**.** *Suppose that $\pi \propto \exp(-V)$ satisfies $1/\alpha$-PI and that V is β-smooth. There is an algorithm that, given access to a first-order + prox oracle for V and initialized at μ_0, outputs a random point with law μ satisfying $\mathsf{d}(\mu \,\|\, \pi) \le \varepsilon$ for any of the following metrics:*

$$\mathsf{d} \in \{\mathsf{TV}, \sqrt{\mathsf{KL}}, \sqrt{\chi^2}, \sqrt{\alpha}\, W_2\}\,,$$

using at most

$$N \le \widetilde{O}\bigg(\kappa d^{1/2}\, \log\Big(\frac{\chi^2(\mu_0 \,\|\, \pi)}{\varepsilon^2}\Big)\bigg) \qquad \textit{queries}\,.$$

Remark 6.5.5 (Extensions to Latała–Oleszkiewicz targets)**.** *Just as in §4, we could also obtain a result for distributions satisfying a Latała–Oleszkiewicz inequality, which interpolates between PI and LSI. In this setting, we again improve over the previous state-of-the-art bounds by a factor of $d^{1/2}$. However, for the sake of brevity, we omit this extension as it is conceptually similar but requires more involved technical details.*

These results are the direct analogs of the complexity results in Corollary 4.3.7, but here with a dimension dependence that is improved by a factor of $d^{1/2}$.

We mention another consequence of our improved high-accuracy sampler for the strongly-log-concave setting. Namely, via the same proximal reduction framework, this gives the following alternative complexity bound for target distributions which are (non-strongly) log-concave, sometimes called weakly-log-concave. This bound is a direct analog of Corollary 4.3.6, but here with a dimension dependence that is also improved by a factor of $d^{1/2}$. Note that this theorem is a low-accuracy guarantee; one can also obtain high-accuracy samplers from our results in this log-concave setting by using the fact that log-concavity implies a Poincaré inequality, albeit with a function-dependent constant [KLS95], and then appealing to Theorem 6.5.4. The resulting low-accuracy and high-accuracy results are incomparable in the sense that each can dominate in different settings—but in any case, our theorems for both settings yield improvements by a factor of $d^{1/2}$.

Theorem 6.5.6 (Faster low-accuracy sampling from log-concave targets)**.** *Suppose that $\pi \propto \exp(-V)$, where V is convex and β-smooth. There is an algorithm that, given access to a first-order + prox oracle for V and initialized at μ_0, outputs a random point in $\mathbb{R}^d$ with law μ satisfying $\mathsf{KL}(\mu \,\|\, \pi) \le \varepsilon^2$, using at most*

$$N \le \widetilde{O}\left(\frac{\beta d^{1/2}\, W_2^2(\mu_0, \pi)}{\varepsilon^2}\right) \qquad \textit{queries}\,.$$

Proofs for the results in this section are provided in §6.8. At a high level, the proof of all these results use the same reduction to the problem of sampling from well-conditioned distributions. This reduction is based on the proximal sampler of §4 and lets us apply our improved sampler for the well-conditioned case (Theorem 6.5.1). In each case, however, we must track the propagation of error due to the inexact implementation of the backwards step of the proximal sampler, which was not previously done in any work except for in the TV distance.

In [AC23], we provide more explicit, albeit more complicated, statements of these results to address the following two points. (1) The above results depend on the initialization (through $W_2(\mu_0, \pi)$, $\mathcal{R}_2(\mu_0 \,\|\, \pi)$, or $\chi^2(\mu_0 \,\|\, \pi)$) and it may be unclear how large these quantities are in a given application. (2) We assumed that the algorithm has access to a stronger oracle than just a first-order oracle for V, namely, we also assumed access to a prox oracle for hV with $h = \frac{1}{2\beta}$. We address (1) by explicitly bounding these initialization quantities in terms of other, more easily computable problem parameters, and we address (2) by removing the assumption of a prox oracle.

6.6 Deferred details for §6.3

6.6.1 Proof for Remark 6.3.4

Here, we prove the inequality in Remark 6.3.4, repeated here for convenience:

$$\frac{1}{\sqrt{p}}\, W_p \lesssim W_{\psi_2} \le \frac{1}{\sqrt{\log 2}}\, W_\infty\,. \tag{6.12}$$

Bounding W_p by W_{ψ_2}. By [Ver18, Proposition 2.5.2], if (X, Y) is an optimal coupling of μ and ν for the W_{ψ_2} distance, then

$$W_p(\mu,\nu) \le \mathbb{E}[\|X-Y\|^p]^{1/p} \lesssim \sqrt{p}\,\|X-Y\|_{\psi_2} = W_{\psi_2}(\mu,\nu)\,.$$

Bounding W_{ψ_2} by W_∞. Observe that for any random variable Z, if we denote $\|Z\|_\infty \coloneqq \operatorname{ess\,sup}\|Z\|$, then we can bound

$$\mathbb{E}\left[\psi_2\Big(\frac{Z}{\|Z\|_\infty/\sqrt{\log 2}}\Big)\right] = \mathbb{E}\left[\exp\Big(\frac{\|Z\|^2}{\|Z\|_\infty^2}\cdot\log 2\Big) - 1\right] \le \exp(\log 2) - 1 = 1\,,$$

and therefore $\|Z\|_{\psi_2} \le \|Z\|_\infty/\sqrt{\log 2}$ by the definition of the Orlicz norm. Now, applying this bound to the random variable $Z = X - Y$, we conclude that

$$\begin{aligned} W_{\psi_2}(\mu,\nu) = \inf_{(X,Y)\in\mathcal{C}(\mu,\nu)} \|X-Y\|_{\psi_2} &\le \frac{1}{\sqrt{\log 2}} \inf_{(X,Y)\in\mathcal{C}(\mu,\nu)} \|X-Y\|_\infty \\ &= \frac{1}{\sqrt{\log 2}}\, W_\infty(\mu,\nu)\,. \end{aligned}$$

6.6.2 Remarks on tightness of Lemma 6.3.7

Here we remark that, conditional on the following plausible conjecture, the generalized version of Lemma 6.3.7 (as stated in Remark 6.3.8) is tight. This conjecture states that the shifted Rényi divergence between two isotropic Gaussians with same covariance is achieved by a deterministic shift. Understanding this simple case could be more broadly helpful for understanding tightness of other inequalities and analyses using the shifted Rényi divergence.

Conjecture 6.6.1. *For any Rényi order $q \ge 1$, noise variance $\sigma^2 > 0$, shift $w \ge 0$, and mean $x \in \mathbb{R}^d$,*

$$\begin{aligned} \mathcal{R}_q^{(w)}\big(\mathsf{normal}(x,\sigma^2 I_d) \,\big\|\, \mathsf{normal}(0,\sigma^2 I_d)\big) &= \mathcal{R}_q\big(\mathsf{normal}(cx,\sigma^2 I_d) \,\big\|\, \mathsf{normal}(0,\sigma^2 I_d)\big) \\ &= \frac{c^2 q\,\|x\|^2}{2\sigma^2}\,, \end{aligned}$$

where $c \coloneqq \max(0, 1 - w\sqrt{\log 2}/\|x\|)$.

Of course, the conjecture here is the first equality (the second equality is just the closed-form expression for the Rényi divergence between Gaussians in Lemma 2.2.21). The direction "$\leq$" is clear because $\mathsf{normal}(cx, \sigma^2 I_d)$ satisfies

$$W_{\psi_2}\big(\mathsf{normal}(x, \sigma^2 I_d),\, \mathsf{normal}(cx, \sigma^2 I_d)\big) \leq \|x - cx\|_{\psi_2} = \frac{1-c}{\sqrt{\log 2}} \|x\| \leq w\,,$$

and therefore is feasible for the optimization problem defining the shifted Rényi divergence. The direction "$\geq$" is the one requiring justification.

In the rest of this subsection, we show the claimed tightness assuming Conjecture 6.6.1. Fix any Rényi order $q \geq 1$, noise variance $\sigma^2 > 0$, and initial shift $w \geq 0$. Consider distributions $\mu = \delta_{ae_1}$ and $\nu = \delta_0$, where $a > w\sqrt{\log 2}$. We claim that the bound in Remark 6.3.8 holds with equality when its parameters δ, λ are optimized; that is,

$$\begin{aligned} &\mathcal{R}_q^{(w)}\big(\mu * \mathsf{normal}(0, \sigma^2 I_d) \;\big|\; \nu * \mathsf{normal}(0, \sigma^2 I_d)\big) \\ &\qquad = \inf_{\delta > 0,\, \lambda \in [0,1]} \left[\mathcal{R}^{(w+\delta)}_{(q+\lambda-1)/\lambda}(\mu \,\|\, \nu) + \frac{(q-\lambda)\,\delta^2}{2\,(1-\lambda)\,\sigma^2} \log 2 \right]. \end{aligned} \tag{6.13}$$

To this end, supposing Conjecture 6.6.1 holds, the left hand side of (6.13) is equal to

$$\begin{aligned} &\mathcal{R}_q^{(w)}\big(\mathsf{normal}(ae_1, \sigma^2 I_d) \;\big\|\; \mathsf{normal}(0, \sigma^2 I_d)\big) \\ &\qquad = \mathcal{R}_q\big(\mathsf{normal}((a - w\sqrt{\log 2})\, e_1, \sigma^2 I_d) \;\big\|\; \mathsf{normal}(0, \sigma^2 I_d)\big) \\ &\qquad = \frac{q\,(a - w\sqrt{\log 2})^2}{2\sigma^2}\,. \end{aligned}$$

On the other hand, note that $\mathcal{R}^{(w+\delta)}_{(q+\lambda-1)/\lambda}(\mu\|\nu)$ is equal to 0 if $a \leq (w+\delta)\sqrt{\log 2}$, and otherwise is equal to ∞. This means that the optimal value of δ is $a/\sqrt{\log 2} - w$. Thus the right hand side of (6.13) simplifies to

$$\inf_{\lambda \in [0,1]} \frac{(q-\lambda)\,(a - w\sqrt{\log 2})^2}{2\,(1-\lambda)\,\sigma^2} = \frac{q\,(a - w\sqrt{\log 2})^2}{2\sigma^2}\,,$$

where the final step is because the optimal value of λ is at $\lambda = 0$. We conclude that the left- and right-hand sides of (6.13) indeed match, as desired.

■ 6.6.3 Proof for Remark 6.3.9

Here we provide details for why Lemma 6.3.7 fails if the shifted Rényi divergence is defined using the W_p metric for any p finite, or alternatively with the Orlicz–Wasserstein metric with any Orlicz norm that is weaker than sub-Gaussian (i.e.,

with Orlicz function $\psi_b(x) \coloneqq \exp(x^b) - 1$ for $b < 2$). Specifically, for any of these Wasserstein metrics W, we claim that if μ is the distribution on $\mathbb{R}$ with density proportional to $\exp(-|\cdot|^a)$ for an appropriate constant $a < 2$, and $\nu = \delta_0$, then:

(i) $W(\mu, \nu) < \infty$.

(ii) $\mathcal{R}_q(\mu * \mathsf{normal}(0, \sigma^2) \,\|\, \nu * \mathsf{normal}(0, \sigma^2)) = \infty$.

This comprises a counterexample to Lemma 6.3.7 by taking $w = 0$ and $\delta = W(\mu, \nu)$ and noting that if $\delta > \sigma/\sqrt{(2q-1)(q-1)}$, then we can simply dilate the space (i.e., replace $\mu(x)$ by $\mu_b(x) \propto \mu(bx)$ for a sufficiently large $b > 0$) and repeat the same argument.

Proof of (i). In the case that $W = W_p$, then $W_p(\mu, \nu)$ is equal to the p-th norm of μ, which is finite for any $p < \infty$. In the case that W is an Orlicz–Wasserstein norm with Orlicz norm weaker than sub-Gaussian, then $W(\mu, \nu)$ is equal to the Orlicz norm of ν, which is finite if we choose $a = b$.

Proof of (ii). Note that $\mu * \mathsf{normal}(0, \sigma^2)$ is not sub-Gaussian, yet $\nu * \mathsf{normal}(0, \sigma^2) = \mathsf{normal}(0, \sigma^2)$ is sub-Gaussian. We may therefore appeal to the fact that the Rényi divergence is infinite whenever the first argument is not sub-Gaussian, but the second argument is. For a proof of this fact in the case that $q = 2$, see Lemma 3.6.9; this proof readily extends to any finite $q \in (1, \infty)$ by replacing the Cauchy–Schwarz inequality by Hölder's inequality.

6.7 Deferred details for §6.4

6.7.1 Proof of Lemma 6.4.5

Compute the partial derivatives

$$\begin{aligned}
\partial_u \bar{F}(u, v)_u &= \frac{1+a}{2} I_d - \frac{h - \gamma^{-1}(1-a)}{\gamma} \nabla^2 V(u)\,, \\
\partial_u \bar{F}(u, v)_v &= \frac{1-a}{2} I_d - \frac{h + \gamma^{-1}(1-a)}{\gamma} \nabla^2 V(u)\,, \\
\partial_v \bar{F}(u, v)_u &= \frac{1-a}{2} I_d\,, \\
\partial_v \bar{F}(u, v)_v &= \frac{1+a}{2} I_d\,.
\end{aligned}$$

Since $\frac{1}{\gamma}(h - \gamma^{-1}(1-a)) = O(h^2)$, we have

$$\|\nabla \bar{F}(u, v)\|_{\mathrm{op}} \le \frac{1}{2} \underbrace{\left\| \begin{bmatrix} (1+a) I_d & (1-a) I_d - b \nabla^2 V(u) \\ (1-a) I_d & (1+a) I_d \end{bmatrix} \right\|_{\mathrm{op}}}_{=:A} + O(\beta h^2)\,,$$

where we use the notational shorthand $b \coloneqq \frac{2}{\gamma}\left(h + \gamma^{-1}(1-a)\right)$.

To bound this operator norm, we compute:

$$AA^{\mathsf{T}} = \begin{bmatrix} (1+a)^2 I_d + \left((1-a) I_d - b\,\nabla^2 V(u)\right)^2 & 2(1-a^2) I_d - (1+a)\, b\,\nabla^2 V(u) \\ 2(1-a^2) I_d - (1+a)\, b\,\nabla^2 V(u) & \left((1-a)^2 + (1+a)^2\right) I_d \end{bmatrix}$$

Since $1 - a = O(\gamma h)$ and $b = O(h/\gamma)$, we can approximate AA^{T} by the following matrix B with error

$$\begin{aligned} &\Big\| AA^{\mathsf{T}} - 2 \underbrace{\begin{bmatrix} (1+a^2) I_d & (1-a^2) I_d - b\,\nabla^2 V(u) \\ (1-a^2) I_d - b\,\nabla^2 V(u) & (1+a^2) I_d \end{bmatrix}}_{=:B} \Big\|_{\mathrm{op}} \\ &\quad = O\Big(\frac{\beta^2 h^2}{\gamma^2} + \beta h^2\Big)\,. \end{aligned}$$

By a direct computation, the eigenvalues of B are $1 + a^2 \pm (1 - a^2 - b\lambda)$, where λ ranges over the eigenvalues of $\nabla^2 V(u)$. The strong-convexity and smoothness of V implies that $\lambda \in [\alpha, \beta]$. Thus

$$\|B\|_{\mathrm{op}} \le \max\{2a^2 + \beta b, 2 - \alpha b\}\,.$$

We note that

$$\begin{aligned} 2a^2 + \beta b &= 2\exp(-2\gamma h) + \frac{2\beta\left(h + \gamma^{-1}(1 - \exp(-\gamma h))\right)}{\gamma} \\ &= 2\left(1 - 2\gamma h + \frac{2\beta h}{\gamma} + O(\gamma^2 h^2 + \beta h^2)\right). \end{aligned}$$

In order for this to be strictly smaller than 2, we must take $\gamma > \sqrt{\beta}$. We choose $\gamma = \sqrt{2\beta}$, whereby

$$\begin{aligned} \|B\|_{\mathrm{op}} &\le 2\max\Big\{1 - h\sqrt{2\beta},\ 1 - \alpha h\sqrt{2/\beta}\Big\} + O(\beta h^2) \\ &= 2\left(1 - \alpha h\sqrt{2/\beta}\right) + O(\beta h^2)\,. \end{aligned}$$

We deduce that

$$\|AA^{\mathsf{T}}\|_{\mathrm{op}} \le 4\left(1 - \alpha h\sqrt{2/\beta}\right) + O(\beta h^2)$$

and therefore

$$\|\nabla \bar{F}(u,v)\|_{\mathrm{op}} \le \sqrt{1 - \alpha h\sqrt{\frac{2}{\beta}} + O(\beta h^2)} \le 1 - \frac{\alpha}{\sqrt{2\beta}}\, h + O(\beta h^2)\,.$$

6.7.2 Proof of Lemma 6.4.6

By definition of $\mathcal{M}$ and Σ,

$$\bar{\Sigma} = \mathcal{M}\Sigma\mathcal{M}^T = \frac{1}{\gamma^2}\begin{bmatrix} 2\gamma h + 4a - a^2 - 3 & 2\gamma h + a^2 - 1 \\ 2\gamma h + a^2 - 1 & 2\gamma h + 5 - a^2 - 4a \end{bmatrix}.$$

The smallest eigenvalue of this matrix is

$$\begin{aligned}\lambda_{\min}(\bar{\Sigma}) &= \frac{1}{2}\left(\mathrm{tr}(\bar{\Sigma}) - \sqrt{\mathrm{tr}(\bar{\Sigma})^2 - 4\det(\bar{\Sigma})}\right) \\ &= \frac{1}{\gamma^2}\left(2\gamma h + 1 - a^2 - \sqrt{17 - 32a + 14a^2 + a^4 - 4\gamma h + 4a^2\gamma h + 4\gamma^2 h^2}\right) \\ &= \frac{\gamma h^3}{6}\left(1 - \frac{\gamma h}{2} + \frac{\gamma^2 h^2}{240} - \cdots\right).\end{aligned}$$

Above, the first step is by the explicit formula for the eigenvalues of a 2×2 matrix; the second step is by plugging in the entries of $\bar{\Sigma}$ and simplifying; and the third step is by performing a Taylor expansion in the variable γh.

6.7.3 Proof of Lemma 6.4.8

We invoke the following result, which appears as Lemma 26 in [Zha+23].

Proposition 6.7.1 (Movement bound for underdamped Langevin)**.** *Let $(X_t, Y_t)_{t\geq 0}$ denote the continuous-time underdamped Langevin diffusion (6.4) with potential V that is β-smooth and minimized at x^*. Assume that $0 < h \lesssim \frac{1}{\sqrt{\beta}\vee\gamma}$ and $0 \leq \lambda \lesssim \frac{1}{\gamma d h^3}$. Then, conditioned on (X_0, Y_0),*

$$\log \mathbb{E}\exp\bigl(\lambda \sup_{t\in[0,h]} \|X_t - X_0\|^2\bigr) \lesssim \bigl(\beta^2 h^4 \|X_0 - x^\star\|^2 + h^2\|Y_0\|^2 + \gamma d h^3\bigr)\lambda\,.$$

Proof. This result can be easily adapted from the proof of [Zha+23, Lemma 26], noting that in our situation the bound simplifies as we are assuming ∇V is β-Lipschitz rather than merely Hölder continuous. □

We let $\widehat{\mathbf{\Pi}}_T$, $\mathbf{\Pi}_T$ denote the path measures (i.e., probability measures over $\mathcal{C}([0,T];\mathbb{R}^d)$) for the discretized and continuous underdamped Langevin processes respectively, both started at the stationary measure $\boldsymbol{\pi}$. Girsanov's theorem [Le 16, Theorem 5.22] yields

$$\frac{\mathrm{d}\widehat{\mathbf{\Pi}}_T}{\mathrm{d}\mathbf{\Pi}_T} = \exp\sum_{k=0}^{N-1}\Bigl(\frac{1}{\sqrt{2\gamma}}\int_{kh}^{(k+1)h}\langle \nabla V(X_t) - \nabla V(X_{kh}), \mathrm{d}B_t\rangle$$

$$
\begin{aligned}
&\qquad\qquad - \frac{1}{4\gamma} \int_{kh}^{(k+1)h} \|\nabla V(X_t) - \nabla V(X_{kh})\|^2 \, \mathrm{d}t \Big) \\
&=: \exp M_T \,,
\end{aligned}
$$

provided that Novikov's condition (see [Le 16, Theorem 5.23]) holds:

$$
\mathbb{E}_{\mathbf{\Pi}_T} \exp \sum_{k=0}^{N-1} \frac{1}{4\gamma} \int_{kh}^{(k+1)h} \|\nabla V(X_t) - \nabla V(X_{kh})\|^2 \, \mathrm{d}t < \infty \,. \tag{6.14}
$$

Assuming for the moment that (6.14) is indeed verified, Itô's formula yields

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{\Pi}_T} \big[\big(\frac{\mathrm{d}\widehat{\mathbf{\Pi}}_T}{\mathrm{d}\mathbf{\Pi}_T} \big)^q \big] - 1 \\
&\quad = \frac{q\,(q-1)}{4\gamma} \, \mathbb{E}_{\mathbf{\Pi}_T} \sum_{k=0}^{N-1} \int_{kh}^{(k+1)h} \exp(qM_t) \, \|\nabla V(X_t) - \nabla V(X_{kh})\|^2 \, \mathrm{d}t \\
&\quad \le \frac{q^2}{4\gamma} \sum_{k=0}^{N-1} \int_{kh}^{(k+1)h} \sqrt{\mathbb{E}_{\mathbf{\Pi}_T} \exp(2qM_t) \; \mathbb{E}_{\mathbf{\Pi}_T} [\|\nabla V(X_t) - \nabla V(X_{kh})\|^4]} \, \mathrm{d}t \,.
\end{aligned}
$$

Bounding the first term. For $t \in [kh, (k+1)h]$, let $\Delta_t := \nabla V(X_t) - \nabla V(X_{kh})$. By the Cauchy–Schwarz inequality,

$$
\begin{aligned}
\mathbb{E}_{\mathbf{\Pi}_T} \exp(2qM_t) \le & \underbrace{\sqrt{\mathbb{E}_{\mathbf{\Pi}_T} \exp \sum_{k=0}^{N-1} \int_{kh}^{(k+1)h\wedge t} \Big(\frac{2\sqrt{2}\,q}{\sqrt{\gamma}} \, \langle \Delta_s, \mathrm{d}B_s \rangle - \frac{4q^2}{\gamma} \, \|\Delta_s\|^2 \, \mathrm{d}s \Big)}}_{(\dagger)} \\
& \times \sqrt{\mathbb{E}_{\mathbf{\Pi}_T} \exp \sum_{k=0}^{N-1} \int_{kh}^{(k+1)h\wedge t} \big(\frac{4q^2}{\gamma} - \frac{q}{\gamma} \big) \, \|\Delta_s\|^2 \, \mathrm{d}s} \,.
\end{aligned}
$$

We claim that the term marked (†) equals 1; this would follow if the quantity inside the expectation is a martingale. In general, it is only a local martingale, but it is a bona fide martingale provided that Novikov's condition holds: it suffices if

$$
\mathbb{E}_{\mathbf{\Pi}_T} \exp \sum_{k=0}^{N-1} \frac{4q^2}{\gamma} \int_{kh}^{(k+1)h} \|\nabla V(X_t) - \nabla V(X_{kh})\|^2 \, \mathrm{d}t < \infty \,. \tag{6.15}
$$

Note that this condition is stronger than (6.14).

Towards this end, we bound, for a parameter $p \geq 1$ to be chosen later,[11]

$$\begin{aligned}
&\mathbb{E}_{\mathbf{\Pi}_T} \exp \sum_{k=0}^{N-1} \frac{4q^2}{\gamma} \int_{kh}^{(k+1)h} \|\nabla V(X_t) - \nabla V(X_{kh})\|^2 \,\mathrm{d}t \\
&\quad \leq \mathbb{E}_{\mathbf{\Pi}_T} \exp \sum_{k=0}^{N-1} \frac{4\beta^2 q^2}{\gamma} \int_{kh}^{(k+1)h} \|X_t - X_{kh}\|^2 \,\mathrm{d}t \\
&\quad \leq \mathbb{E}_{\mathbf{\Pi}_T} \exp \max_{k=0,1,\ldots,N-1} \sup_{t\in[kh,(k+1)h]} \frac{4\beta^2 q^2 T}{\gamma} \|X_t - X_{kh}\|^2 \\
&\quad \leq \left\{ \mathbb{E}_{\mathbf{\Pi}_T} \exp \max_{k=0,1,\ldots,N-1} \sup_{t\in[kh,(k+1)h]} \frac{4\beta^2 pq^2 T}{\gamma} \|X_t - X_{kh}\|^2 \right\}^{1/p} \\
&\quad \leq N^{1/p} \left\{ \max_{k=0,1,\ldots,N-1} \mathbb{E}_{\mathbf{\Pi}_T} \exp \sup_{t\in[kh,(k+1)h]} \frac{4\beta^2 pq^2 T}{\gamma} \|X_t - X_{kh}\|^2 \right\}^{1/p}.
\end{aligned}$$

By conditioning on (X_{kh}, Y_{kh}) and applying Proposition 6.7.1, this is bounded by

$$\cdots \leq N^{1/p} \left\{ \max_{k=0,1,\ldots,N-1} \mathbb{E}_{\mathbf{\Pi}_T} \exp\Big(O\Big(\frac{\beta^2 pq^2 T}{\gamma} \big(\beta^2 h^4 \|X_{kh} - x^*\|^2 + h^2 \|Y_{kh}\|^2 + \gamma d h^3\big)\Big)\Big) \right\}^{1/p}$$

provided that $h \lesssim \frac{1}{\beta^{2/3} d^{1/3} p^{1/3} q^{2/3} T^{1/3}}$. We choose $p \asymp \log N$ so that $N^{1/p} \asymp 1$. We now need tail bounds for $\|X_{kh} - x^*\|$ and $\|Y_{kh}\|$.

Using the argument in the proof of Lemma 6.4.7, for $c > 0$,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{\Pi}_T} \exp(c \|X_{kh} - x^*\|^2) \\
&\quad \leq \exp(2c\, \mathbb{E}_{\mathbf{\Pi}_T}[\|X_{kh} - x^*\|^2])\, \mathbb{E}_{\mathbf{\Pi}_T} \exp\big(2c\, (\|X_{kh} - x^*\| - \mathbb{E}_{\mathbf{\Pi}_T}\|X_{kh} - x^*\|)^2\big) \\
&\quad \leq \exp\Big(\frac{2cd}{\alpha}\Big) \Big(\mathbb{E}_{\mathbf{\Pi}_T} \exp\Big(\frac{2\,(\|X_{kh} - x^*\| - \mathbb{E}_{\mathbf{\Pi}_T}\|X_{kh} - x^*\|)^2}{36/\alpha}\Big)\Big)^{36c/\alpha} \\
&\quad \leq \exp\Big(O\Big(\frac{cd}{\alpha}\Big)\Big)
\end{aligned}$$

provided that $c \leq \alpha/36$. Therefore,

$$\mathbb{E}_{\mathbf{\Pi}_T} \exp\Big(O\Big(\frac{\beta^4 h^4 q^2 T \log N}{\gamma} \|X_{kh} - x^*\|^2\Big)\Big) \leq \exp\Big(O\Big(\frac{\beta^4 d h^4 q^2 T \log N}{\alpha\gamma}\Big)\Big)$$

[11] This argument avoids the use of the "conditioning lemma" from [GT20] (Lemma 23 in [Zha+23]).

provided that $h \lesssim \frac{\alpha^{1/4}\gamma^{1/4}}{\beta q^{1/2}\,(T\log N)^{1/4}}$.

The same argument applied to $\|Y_{kh}\|$ yields

$$\mathbb{E}_{\mathbf{\Pi}_T} \exp\Big(O\Big(\frac{\beta^2 h^2 q^2 T\log N}{\gamma}\,\|Y_{kh}\|^2\Big)\Big) \le \exp\Big(O\Big(\frac{\beta^2 d h^2 q^2 T\log N}{\gamma}\Big)\Big)$$

provided that $h \lesssim \frac{\gamma^{1/2}}{\beta q\,(T\log N)^{1/2}}$.

If we put these bounds together and take $\gamma \asymp \sqrt{\beta}$, we deduce that if $h \lesssim \frac{\gamma^{1/2}}{\beta d^{1/3} q\,(T\log N)^{1/2}}$ and $T \gtrsim \frac{\sqrt{\beta}}{\alpha}$, then it holds that

$$\begin{aligned}\mathbb{E}_{\mathbf{\Pi}_T} \exp \sum_{k=0}^{N-1} \frac{4q^2}{\gamma} \int_{kh}^{(k+1)h} \|\nabla V(X_t) - \nabla V(X_{kh})\|^2\,\mathrm{d}t \\ \le \exp\big(O(\beta^{3/2} d h^2 q^2 T\log N)\big)\,.\end{aligned}$$

This verifies (6.14) and (6.15), and moreover shows that for $h \lesssim \frac{1}{\beta^{3/4} d^{1/2} q\,(T\log N)^{1/2}}$,

$$\sup_{t\in[0,T]} \mathbb{E}_{\mathbf{\Pi}_T} \exp(2qM_t) \lesssim 1\,.$$

Bounding the second term. Next,

$$\sqrt{\mathbb{E}_{\mathbf{\Pi}_T}[\|\nabla V(X_t) - \nabla V(X_{kh})\|^4]} \le \beta^2 \sqrt{\mathbb{E}_{\mathbf{\Pi}_T}[\|X_t - X_{kh}\|^4]}\,.$$

Applying Proposition 6.7.1 and the above tail estimates,

$$\begin{aligned}\cdots &\lesssim \beta^2\ \mathbb{E}_{\mathbf{\Pi}_T}[\beta^2 h^4\,\|X_{kh} - x^*\|^2 + h^2\,\|Y_{kh}\|^2 + \gamma d h^3] \lesssim \frac{\beta^4 d h^4}{\alpha} + \beta^2 d h^2 + \beta^2\gamma d h^3 \\ &\lesssim \beta^2 d h^2\,.\end{aligned}$$

Concluding the proof of Lemma 6.4.8. In summary, we have shown

$$\mathbb{E}_{\mathbf{\Pi}_T}\big[\big(\frac{\mathrm{d}\widehat{\mathbf{\Pi}}_T}{\mathrm{d}\mathbf{\Pi}_T}\big)^q\big] - 1 \lesssim \frac{\beta^2 d h^2 q^2 T}{\gamma}\,.$$

Hence, by definition of the Rényi divergence, we conclude that

$$\mathcal{R}_q(\widehat{\mathbf{\Pi}}_T \,\|\, \mathbf{\Pi}_T) \lesssim \frac{\beta^2 d h^2 q T}{\gamma}\,.$$

6.8 Deferred details for §6.5

6.8.1 Weak version of Theorem 6.5.1

Here, we show the following weaker version of Theorem 6.5.1 as it provides the key building block to prove it. Like Theorem 6.5.1, this result here shows that from a feasible start, the query complexity of high-accuracy sampling from a strongly-log-concave distribution scales in the dimension d as $\widetilde{O}(d^{1/2})$ rather than $\widetilde{O}(d)$. The difference from Theorem 6.5.1 is that the weaker result here has a suboptimal dependence on the condition number κ, namely $\widetilde{O}(\kappa^{3/2})$ rather than $\widetilde{O}(\kappa)$. This dependence will later be boosted using the proximal sampler in §6.8.2, allowing us to prove the full Theorem 6.5.1.

Theorem 6.8.1 (Weak version of Theorem 6.5.1)**.** *Suppose that $\pi \propto \exp(-V)$ where V is α-strongly-convex and β-smooth. Let x^* denote the minimizer of V. For any sampling error $\varepsilon > 0$ and any initialization distribution $\delta_{x_0} \in \mathcal{P}(\mathbb{R}^d)$, there is an algorithm that uses*

$$N = \widetilde{O}\big(\kappa^{3/2} d^{1/2} \log(\|x_0 - x^*\|) \;+\; \kappa d^{1/2} \log^3(1/\varepsilon)\big)$$

first-order queries for V to output a random point in $\mathbb{R}^d$ with law μ satisfying $\mathrm{d}(\mu \,\|\, \pi) \le \varepsilon$ for any of the following metrics:

$$\mathrm{d} \in \{\mathsf{TV}, \sqrt{\mathsf{KL}}, \sqrt{\chi^2}, \sqrt{\alpha}\, W_2\}\,.$$

Remark 6.8.2 (Handling other metrics)**.** *To prove convergence in the various metrics, due to standard comparison inequalities it usually suffices to prove a convergence result in the strongest metric, namely, the chi-squared divergence. Indeed, convergence in the KL divergence follows from the monotonicity of Rényi divergences (Lemma 2.2.20) and convergence in the TV distance follows from Pinsker's inequality. If π satisfies an LSI, then convergence in W_2 follows from Talagrand's T_2 inequality (Lemma 2.2.11); otherwise, if π only satisfies a PI, then convergence in W_2 follows from the quadratic transport-variance inequality (Lemma 2.2.12).*

The proof has three steps:

1. Run ULMC from this arbitrary initialization δ_{x_0} to obtain a $\mathcal{R}_3$ warm start (which also implies a χ^2 warm start).

2. Use the χ^2 warm start to argue that MALA mixes rapidly in TV.

3. Use the $\mathcal{R}_3$ warm start to argue that the TV mixing guarantee implies mixing guarantees in χ^2 (and therefore also the other desired metrics by Remark 6.8.2).

Our main technical contribution here is the ability to implement step 1 in $\widetilde{O}(d^{1/2})$ queries—this is an immediate application of our result on ULMC (Theorem 6.4.1). Step 2 follows from the results in §5 and [WSC22], with only minor modification as described below, and step 3 follows from the helper Lemma 6.5.2.

The rest of this Appendix section is organized as follows. Step 2 is described in §6.8.1.1, step 3 is proved in §6.8.1.2, and then we combine these to prove Theorem 6.8.1 in §6.8.1.3.

■ 6.8.1.1 Rapid mixing of MALA from a Rényi warm start

We first formally state the result in step 2, namely, that MALA mixes rapidly in TV from a χ^2 warm start. This is [WSC22, Theorem 1], except with a less stringent assumption on the initialization μ_0. Specifically, [WSC22, Theorem 1] assumes that μ_0 is an M-warm start with respect to the target π (or equivalently, $\mathcal{R}_\infty(\mu_0 \,\|\, \pi) \le \log M$), whereas the following lemma only assumes that μ_0 has bounded χ^2 (or equivalently, bounded $\mathcal{R}_2$) distance to π. This requires only very minor modification to their analysis, but is essential for our purposes, since our result in §6.4 can only produce warm starts in Rényi divergences of finite order.

Theorem 6.8.3 (Runtime of MALA from warm start; implicit from [WSC22])**.** *Let $\pi \propto \exp(-V)$ where V is α-strongly-convex and β-smooth. For any error $\varepsilon \in (0,1)$, the $1/2$-lazy MALA algorithm with appropriate step size requires*

$$N = \widetilde{O}\Big(\kappa d^{1/2} \log^3\Big(\frac{\chi^2(\mu_0 \,\|\, \pi)}{\varepsilon^2}\Big)\Big)$$

first-order queries to V to output a random point in $\mathbb{R}^d$ whose law μ_N satisfies $\mathsf{TV}(\mu_N, \pi) \le \varepsilon$.

Proof. We appeal to Lemma 5.6.1 and the Cauchy–Schwarz inequality

$$\begin{aligned}|\mu_0(A) - \pi(A)| = \Big|\int \mathbb{1}_A \Big(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi} - 1\Big)\,\mathrm{d}\pi\Big| &\le \sqrt{\int \mathbb{1}_A \,\mathrm{d}\pi \cdot \int \Big(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi} - 1\Big)^2 \,\mathrm{d}\pi} \\ &= \sqrt{\pi(A)\,\chi^2(\mu_0 \,\|\, \pi)}\,,\end{aligned}$$

which implies $H_s \le \sqrt{s\,\chi^2(\mu_0 \,\|\, \pi)}$ and hence

$$\|\mu_N - \pi\|_{\mathsf{TV}} \le \sqrt{s\,\chi^2(\mu_0 \,\|\, \pi)} + \sqrt{\frac{\chi^2(\mu_0 \,\|\, \pi)}{s}}\exp\Big(-\frac{\mathsf{C}_s^2 N}{2}\Big)\,. \tag{6.16}$$

Thus the TV error is at most ε if we set $s = \frac{\varepsilon^2}{4\chi^2(\mu_0\|\pi)}$ and $N = \frac{2}{\mathsf{C}_s^2}\log(\frac{8\chi^2(\mu_0\|\pi)}{\varepsilon^2})$. Plugging in the bound [WSC22, equation (39)] on C_s completes the proof.[12] □

■ 6.8.1.2 Proof of Lemma 6.5.2

By the stationarity property of P and the data-processing inequality for Rényi divergences (Lemma 2.2.19), we have $\mathcal{R}_3(\mu_n \,\|\, \pi) = \mathcal{R}_3(\mu_0 P^n \,\|\, \pi P^n) \le \mathcal{R}_3(\mu_0 \,\|\, \pi)$. It now suffices to argue that the following inequality holds for any μ, π:

$$\chi^2(\mu \,\|\, \pi) \le \sqrt{\mathsf{TV}(\mu,\pi)\cdot\big(\exp(2\mathcal{R}_3(\mu \,\|\, \pi)) + 1\big)}\,. \tag{6.17}$$

To prove (6.17), we use the Cauchy–Schwarz inequality to bound

$$\begin{aligned}\chi^2(\mu \,\|\, \pi) &= \int\Big|\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\Big|^2 \mathrm{d}\pi = \int\Big|\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\Big|^{1/2}\Big|\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\Big|^{3/2}\mathrm{d}\pi \\ &\le \sqrt{\int\Big|\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\Big|\,\mathrm{d}\pi \;\cdot\; \int\Big|\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\Big|^3\mathrm{d}\pi}\,.\end{aligned}$$

The first integral is precisely $\mathsf{TV}(\mu,\pi)$. The second integral can be bounded by

$$\int\Big|\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\Big|^3\mathrm{d}\pi \le \int\Big|\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\Big|^3\mathrm{d}\pi + 1 = \exp(2\mathcal{R}_3(\mu \,\|\, \pi)) + 1\,,$$

where above, the first step is by the elementary inequality $|a-1|^3 \le a^3 + 1$, which holds for $a \ge 0$; and the second inequality is by the definition of Rényi divergence. This completes the proof of (6.17) and thus also the proof of the lemma.

■ 6.8.1.3 Proof of Theorem 6.8.1

By Theorem 6.4.1—or rather the extension in Remark 6.4.2—ULMC outputs a distribution ν satisfying $\mathcal{R}_3(\nu \,\|\, \pi) \le \log 2$, say, using

$$\widetilde{O}\big(\kappa^{3/2} d^{1/2} \log\|x_0 - x^*\|\big)$$

gradient queries, where δ_{x_0} is its initial distribution. By monotonicity of Rényi divergences (Lemma 2.2.20) and the identity between χ^2 and $\mathcal{R}_2$ (Remark 2.2.18), this guarantee implies $\chi^2(\nu \,\|\, \pi) = \exp(\mathcal{R}_2(\nu \,\|\, \pi)) - 1 \le \exp(\mathcal{R}_3(\nu \,\|\, \pi)) - 1 \le 1$, so ν is a warm start in χ^2 divergence. Thus we may invoke Theorem 6.8.3 to run MALA from initialization ν in order to produce a distribution μ satisfying $\mathsf{TV}(\mu,\pi) \le \varepsilon^4/5$, say, using

$$\widetilde{O}\big(\kappa d^{1/2} \log^3(1/\varepsilon)\big)$$

[12] In fact, one can simply set $M = 2\chi^2(\mu_0 \,\|\, \pi)/\varepsilon$ in their final bounds to obtain Theorem 6.8.3.

first-order queries. Now by Lemma 6.5.2, we can use the warm start property of ν to boost the TV guarantee on MALA's output μ to the following χ^2 guarantee:

$$\chi^2(\mu \,\|\, \pi) \le \sqrt{\mathsf{TV}(\mu,\pi) \cdot \big(\exp(2\mathcal{R}_3(\nu \,\|\, \pi)) + 1\big)} \le \varepsilon^2 \,.$$

This implies the desired χ^2 mixing bound. Mixing in the other metrics then follows from Remark 6.8.2.

■ 6.8.2 Proof of Theorems 6.5.1 and 6.5.3

Here we prove our main results about faster high-accuracy sampling algorithms in the setting that the target distribution π is strongly-log-concave (Theorem 6.5.1) or satisfies a log-Sobolev inequality (Theorem 6.5.3). Since our analysis only relies upon the LSI property, we are able to prove both theorems simultaneously. (Indeed, recall that strong-log-concavity implies a log-Sobolev inequality by the Bakry–Émery theorem, see the first part of Lemma 2.2.8). See §6.5.2 for a high-level overview of the algorithm and analysis.

We begin with a helper lemma, which is similar to the Orlicz–Wasserstein initialization bound for π in Lemma 6.4.7, but now generalized to the RGO $\pi^{X|Y=y} \propto \exp(-V - \frac{1}{2h}\|\cdot - y\|^2)$ that is used in the proximal sampler.

Lemma 6.8.4 (Orlicz–Wasserstein distance at initialization of RGO step)**.** *Suppose that $\pi \propto \exp(-V)$ where V is β-smooth. Let x^* denote the mode of π. Then for any $y \in \mathbb{R}^d$ and any proximal step size $h \le 1/(2\beta)$,*

$$W_{\psi_2}(\delta_y, \pi^{X|Y=y}) \le 9\sqrt{dh} + 3\beta h \,\|y - x^*\| \,.$$

Proof. Let x_y denote the mode of $\pi^{X|Y=y}$. By the triangle inequality,

$$W_{\psi_2}(\pi^{X|Y=y}, \delta_y) \le W_{\psi_2}(\pi^{X|Y=y}, \delta_{x_y}) + W_{\psi_2}(\delta_{x_y}, \delta_y) \,.$$

The former term is bounded above by $9\sqrt{dh}$ by an application of Lemma 6.4.7 and the observation that $\pi^{X|Y=y} \propto \exp(-V - \frac{1}{2h}\|\cdot - y\|^2)$ is strongly-log-concave with parameter $-\beta + \frac{1}{h} \ge \frac{1}{2h}$. Next, we bound the latter term

$$W_{\psi_2}(\delta_{x_y}, \delta_y) = \|x_y - y\|_{\psi_2} = \frac{\|x_y - y\|}{\sqrt{\log 2}} \,.$$

Since x_y is the mode of $\pi^{X|Y=y}$, it is the minimizer of the convex log-density, thus by first-order optimality conditions we have $0 = \nabla V(x_y) + \frac{1}{h}(x_y - y)$. By rearranging this identity, using the smoothness of V, and then using the triangle inequality,

$$\|x_y - y\| = h\,\|\nabla V(x_y) - \nabla V(x^*)\| \le \beta h\,\|x_y - x^*\| \le \beta h\,\big(\|x_y - y\| + \|y - x^*\|\big) \,.$$

Now by the assumption on the step size, $\beta h \le 1/2$. Plugging this in and rearranging yields

$$\|x_y - y\| \le 2\beta h \, \|y - x^*\| \, .$$

Combining the above displays completes the proof. □

Armed with this initialization lemma, we are now ready to prove the main results of this section.

Proof of Theorems 6.5.1 and 6.5.3. Recall from the discussion at the beginning of this subsection that it suffices to prove Theorem 6.5.3. Hence, in this proof we assume that π is $1/\alpha$-LSI but do not necessarily assume that it is α-strongly-log-concave. We prove the mixing time for the χ^2 divergence, which suffices by Remark 6.8.2.

First, suppose that the RGO in the proximal sampler algorithm is implemented exactly. Let X_n and Y_n denote the proximal sampler iterates at iteration n; and let μ_n^X and μ_n^Y denote their respective laws. Then after initializing at $\mu_0^X := \mathsf{normal}(x^*, (2\beta)^{-1} I_d)$, the laws of the iterates are given by $\mu_n^Y = \mu_n^X * \mathsf{normal}(0, hI_d)$, and $\mu_{n+1}^X = \int \pi^{X|Y}(\cdot \mid y) \, \mu_n^Y(\mathrm{d}y)$. By analyzing the simultaneous heat flow, it was shown in §4.3.3.1 that the forwards step of the proximal algorithm is a contraction in Rényi divergence, in the sense that

$$\mathcal{R}_q(\mu_n^Y \parallel \pi^Y) \le \frac{1}{(1+\alpha h)^{1/q}} \, \mathcal{R}_q(\mu_n^X \parallel \pi^X) \, . \tag{6.18}$$

Now suppose that we have oracle access to an approximate RGO in the sense that given any point $y \in \mathbb{R}^d$, we can sample from a distribution $\tilde{\pi}^{X|Y=y}$ with

$$\mathcal{R}_q(\tilde{\pi}^{X|Y=y} \parallel \pi^{X|Y=y}) \le \varepsilon_{\mathsf{RGO}}^2 \, , \tag{6.19}$$

using $N_{\mathsf{RGO}}(y)$ first-order queries to V. Let $\tilde{X}_n$, $\tilde{Y}_n$ denote the iterates with inexact implementation of the RGO, and let $\tilde{\mu}_n^X$, $\tilde{\mu}_n^Y$ denote their laws respectively.

We can bound the error of a backwards step using this approximate RGO as follows. Let $\tilde{Y}_n \sim \tilde{\mu}_n^Y$, so that $\tilde{X}_{n+1}$ is a sample from the approximate RGO $\tilde{\pi}^{X|Y=\tilde{Y}_n}$. Then

$$\begin{aligned}
\mathcal{R}_q(\tilde{\mu}_{n+1}^X \parallel \pi^X) &\le \mathcal{R}_q\big(\mathrm{law}(\tilde{X}_{n+1}, \tilde{Y}_n) \,\big\|\, \boldsymbol{\pi}\big) \\
&\le \mathcal{R}_q(\tilde{\mu}_n^Y \parallel \pi^Y) + \sup_{y_n \in \mathbb{R}^d} \mathcal{R}_q(\tilde{\pi}^{X|Y=y_n} \parallel \pi^{X|Y=y_n}) \\
&\le \mathcal{R}_q(\tilde{\mu}_n^Y \parallel \pi^Y) + \varepsilon_{\mathsf{RGO}}^2 \, .
\end{aligned} \tag{6.20}$$

Above, the first step is by the data-processing inequality for Rényi divergences (Lemma 2.2.19); the second step is by the "strong composition rule" for Rényi

differential privacy (this lemma has appeared in many equivalent forms, see, e.g., [Aba+16a; DR16; Mir17]; here we apply the version from [AT22a, Lemma 2.9]); and the final step is by the guarantee (6.19) of the approximate RGO.

By combining the error bounds (6.18) and (6.20) for the forward step and approximate backwards step of the proximal sampler, we conclude the bound

$$\mathcal{R}_q(\tilde{\mu}_{n+1}^X \,\|\, \pi^X) \le \frac{1}{(1+\alpha h)^{1/q}} \mathcal{R}_q(\tilde{\mu}_n^X \,\|\, \pi^X) + \varepsilon_{\mathsf{RGO}}^2 \,. \tag{6.21}$$

Iterating this bound N_{prox} times gives the following Rényi divergence bound on the mixing error of the proximal sampler when using this approximate RGO:

$$\mathcal{R}_q(\tilde{\mu}_{N_{\mathsf{prox}}}^X \,\|\, \pi^X) \le \frac{1}{(1+\alpha h)^{N_{\mathsf{prox}}/q}} \mathcal{R}_q(\mu_0^X \,\|\, \pi^X) + \varepsilon_{\mathsf{RGO}}^2 \sum_{n=0}^{N_{\mathsf{prox}}-1} \frac{1}{(1+\alpha h)^{n/q}} \,. \tag{6.22}$$

This error is at most ε^2 if we run the proximal sampler with step size $h \asymp 1/\beta$ for

$$N_{\mathsf{prox}} \asymp \kappa q \log \frac{\mathcal{R}_q(\mu_0^X \,\|\, \pi^X)}{\varepsilon^2}$$

iterations and perform each approximate RGO to accuracy

$$\varepsilon_{\mathsf{RGO}} \asymp \frac{\varepsilon}{\sqrt{\kappa q}} \,. \tag{6.23}$$

Henceforth, consider $q = 2$, so that $\mathcal{R}_2 \le \chi^2$ (see Remark 2.2.18). Observe that if the step size $h < 1/(2\beta)$, say, then the RGO is strongly-log-concave and has condition number of size at most

$$\frac{\beta + 1/h}{-\beta + 1/h} = \frac{1+\beta h}{1-\beta h} = \Theta(1) \,.$$

We next consider the complexity of implementing the RGO. Assume that we can compute the proximal operator for hV exactly, so we can compute the mode $x(\tilde{Y}_n)$ of $\pi^{X|Y=\tilde{Y}_n}$ and initialize at $\delta_{x(\tilde{Y}_n)}$. By Theorem 6.8.1, we can implement the approximate RGO $\tilde{\pi}^{X|Y=\tilde{Y}_n}$ in the n-th iteration by using $N_{\mathsf{RGO}}(\tilde{Y}_n)$ first-order queries, where

$$N_{\mathsf{RGO}}(\tilde{Y}_n) = \widetilde{O}\Big(d^{1/2} \log^3\Big(\frac{W_{\psi_2}(\delta_{x(\tilde{Y}_n)},\ \pi^{X|Y=\tilde{Y}_n})}{\varepsilon_{\mathsf{RGO}}}\Big)\Big) \,. \tag{6.24}$$

By Lemma 6.8.4, $W_{\psi_2}(\delta_{x(\tilde{Y}_n)},\ \pi^{X|Y=\tilde{Y}_n}) \lesssim \sqrt{d/\beta}$. We conclude that the total number of gradient queries required by this inexact proximal sampler algorithm is

$$N = \sum_{n=0}^{N_{\mathsf{prox}}-1} N_{\mathsf{RGO}}(\tilde{Y}_n) = \widetilde{O}\bigg(\kappa d^{1/2} \log\Big(\frac{\mathcal{R}_2(\mu_0^X \,\|\, \pi^X)}{\varepsilon^2}\Big) \log^3\Big(\frac{1}{\varepsilon}\Big)\bigg) \,. \tag{6.25}$$

To remove the assumption that we can exactly compute the proximal operator of hV, we refer to [AC23]. □

6.8.3 Proof of Theorem 6.5.4

We prove the χ^2 mixing bound; the other desired mixing bounds then follow immediately due to standard comparison inequalities (see Remark 6.8.2). We consider the same inexact RGO algorithm as in the LSI setting (see §6.8.2). Under the present Poincaré assumption, the forwards step of the proximal algorithm is known to be a contraction in χ^2—in direct analog to (6.18). Specifically, by analyzing the simultaneous heat flow, it was shown in §4.4.5 that

$$\chi^2(\mu_n^Y \,\|\, \pi^Y) \le \frac{1}{1+\alpha h}\, \chi^2(\mu_n^X \,\|\, \pi^X)\,. \tag{6.26}$$

The bound (6.20) on the error of a backwards step of the proximal sampler using an approximate RGO (6.19) remains unchanged (as it never uses the LSI assumption). This Rényi bound is equivalent to the χ^2 bound

$$\chi^2(\tilde{\mu}_{n+1}^X \,\|\, \pi^X) \le \exp(\varepsilon_{\mathsf{RGO}}^2)\, \chi^2(\tilde{\mu}_n^Y \,\|\, \pi^Y)\,, \tag{6.27}$$

by using the relationship $\mathcal{R}_2 = \log(1+\chi^2)$ between the chi-squared and Rényi divergences (see Remark 2.2.18). By combining the above two displays, we obtain the following convergence bound for one full step of the proximal sampler:

$$\chi^2(\tilde{\mu}_{n+1}^X \,\|\, \pi^X) \le \exp\bigl(-\Theta(\alpha h)\bigr)\, \chi^2(\tilde{\mu}_n^X \,\|\, \pi^X)\,, \tag{6.28}$$

if we solve each approximate RGO to accuracy

$$\varepsilon_{\mathsf{RGO}} \lesssim \sqrt{\alpha h} \asymp \frac{1}{\sqrt{\kappa}}\,. \tag{6.29}$$

By iterating this one-step bound (6.28), we conclude that the final mixing error $\chi^2(\tilde{\mu}_{N_{\mathsf{prox}}}^X \,\|\, \pi^X)$ of this inexact proximal sampler is at most ε^2 if it is run for N_{prox} iterations, where

$$N_{\mathsf{prox}} \asymp \kappa \log\Bigl(\frac{\chi^2(\mu_0^X \,\|\, \pi^X)}{\varepsilon^2}\Bigr)\,.$$

Now by the same argument as the LSI setting (see §6.8.2), by the choice of the proximal step size h, the RGO $\tilde{\pi}^{X|Y=\tilde{Y}_n}$ has condition number $\Theta(1)$, and thus can be implemented using $N_{\mathsf{RGO}}(\tilde{Y}_n)$ gradient queries by Theorem 6.8.1, where $N_{\mathsf{RGO}}(\tilde{Y}_n)$ is the quantity in (6.24). Therefore the total number of gradient queries required by this inexact proximal sampler algorithm is

$$N = \sum_{n=0}^{N_{\mathsf{prox}}-1} N_{\mathsf{RGO}}(\tilde{Y}_n) = \widetilde{O}\biggl(\kappa d^{1/2} \log\Bigl(\frac{\chi^2(\mu_0^X \,\|\, \pi^X)}{\varepsilon^2}\Bigr)\biggr)\,. \tag{6.30}$$

6.8.4 Proof of Theorem 6.5.6

The proof for the weakly log-concave case is similar to the proofs of Theorems 6.5.1, 6.5.3, and 6.5.4, in that we carefully keep track of the error from inexact implementation of the RGO, but the proof requires key modifications. It was shown in §4.4.3 that along the simultaneous heat flow,

$$\frac{1}{\mathsf{KL}(\mu_n^Y \,\|\, \pi^Y)} \geq \frac{1}{\mathsf{KL}(\mu_n^X \,\|\, \pi^X)} + \frac{h}{2W_2^2(\mu_n^X, \pi^X)}\,. \tag{6.31}$$

Let us assume that the RGO is implemented inexactly, so that for each $y \in \mathbb{R}^d$ we sample from $\tilde{\pi}^{X|Y=y}$ satisfying

$$\mathsf{KL}(\tilde{\pi}^{X|Y=y} \,\|\, \pi^{X|Y=y}) \leq \varepsilon_{\mathsf{RGO}}^2\,, \qquad W_2(\tilde{\pi}^{X|Y=y}, \pi^{X|Y=y}) \leq \sqrt{2\beta}\,\varepsilon_{\mathsf{RGO}}\,. \tag{6.32}$$

In fact, the second guarantee follows from the first together with Talagrand's T_2 inequality (see Lemma 2.2.11) if we choose step size $h = \frac{1}{2\beta}$, because the RGO is then β-strongly log-concave.

By applying (6.20) for the proximal sampler with inexact RGO implementation, convexity of the map $x \mapsto 1/x$, and (6.31), we deduce that

$$\begin{aligned}\frac{1}{\mathsf{KL}(\tilde{\mu}_{n+1}^X \,\|\, \pi^X)} &\geq \frac{1}{\mathsf{KL}(\tilde{\mu}_n^Y \,\|\, \pi^Y) + \varepsilon_{\mathsf{RGO}}^2} \geq \frac{1}{\mathsf{KL}(\tilde{\mu}_n^Y \,\|\, \pi^Y)} - \frac{\varepsilon_{\mathsf{RGO}}^2}{\mathsf{KL}(\tilde{\mu}_n^Y \,\|\, \pi^Y)^2} \\ &\geq \frac{1}{\mathsf{KL}(\tilde{\mu}_n^X \,\|\, \pi^X)} + \frac{h}{2W_2^2(\tilde{\mu}_n^X, \pi^X)} - \frac{\varepsilon_{\mathsf{RGO}}^2}{\mathsf{KL}(\tilde{\mu}_n^Y \,\|\, \pi^Y)^2}\,. \end{aligned} \tag{6.33}$$

We now split into two cases. In the first case, suppose that $\mathsf{KL}(\tilde{\mu}_n^Y \,\|\, \pi^Y) \leq \sqrt{\varepsilon_{\mathsf{RGO}}}$ for some $n = 0, 1, \ldots, N_{\mathsf{prox}} - 1$. By repeatedly applying (6.20) and the data-processing inequality (Lemma 2.2.19), we obtain

$$\begin{aligned}\mathsf{KL}(\tilde{\mu}_{N_{\mathsf{prox}}}^X \,\|\, \pi^X) &\leq \mathsf{KL}(\tilde{\mu}_{N_{\mathsf{prox}}-1}^Y \,\|\, \pi^Y) + \varepsilon_{\mathsf{RGO}}^2 \\ &\leq \mathsf{KL}(\tilde{\mu}_{N_{\mathsf{prox}}-1}^X \,\|\, \pi^X) + \varepsilon_{\mathsf{RGO}}^2 \leq \cdots \\ &\leq \mathsf{KL}(\tilde{\mu}_n^Y \,\|\, \pi^Y) + N_{\mathsf{prox}}\,\varepsilon_{\mathsf{RGO}}^2 \leq \sqrt{\varepsilon_{\mathsf{RGO}}} + N_{\mathsf{prox}}\,\varepsilon_{\mathsf{RGO}}^2\,. \end{aligned}$$

For the other case, assume $\mathsf{KL}(\tilde{\mu}_n^Y \,\|\, \pi^Y) \geq \sqrt{\varepsilon_{\mathsf{RGO}}}$ for all $n = 0, 1, \ldots, N_{\mathsf{prox}} - 1$. Then, from (6.33), we have

$$\frac{1}{\mathsf{KL}(\tilde{\mu}_{n+1}^X \,\|\, \pi^X)} \geq \frac{1}{\mathsf{KL}(\tilde{\mu}_n^X \,\|\, \pi^X)} + \frac{h}{2W_2^2(\tilde{\mu}_n^X, \pi^X)} - \varepsilon_{\mathsf{RGO}}\,.$$

Iterating this,

$$\frac{1}{\mathsf{KL}(\tilde{\mu}_{N_{\mathsf{prox}}}^X \,\|\, \pi^X)} \geq \frac{1}{\mathsf{KL}(\mu_0^X \,\|\, \pi^X)} + \frac{h}{2}\sum_{n=0}^{N_{\mathsf{prox}}-1} \frac{1}{W_2^2(\tilde{\mu}_n^X, \pi^X)} - N_{\mathsf{prox}}\,\varepsilon_{\mathsf{RGO}}\,.$$

Moreover, from the second condition in (6.32), a standard coupling argument (see, e.g., §4.4.2), and Wasserstein contractivity of the exact proximal sampler steps under log-concavity (Theorem 4.3.1), we obtain

$$\begin{aligned} W_2(\tilde{\mu}^X_{n+1}, \pi^X) &\le W_2\Big(\tilde{\mu}^X_{n+1}, \int \pi^{X|Y=y}\, \tilde{\mu}^Y_n(\mathrm{d}y)\Big) + W_2\Big(\int \pi^{X|Y=y}\, \tilde{\mu}^Y_n(\mathrm{d}y),\, \pi^X\Big) \\ &\le \sqrt{2\beta}\, \varepsilon_{\mathsf{RGO}} + W_2(\tilde{\mu}^Y_n, \pi^Y) \\ &\le \sqrt{2\beta}\, \varepsilon_{\mathsf{RGO}} + W_2(\tilde{\mu}^X_n, \pi^X) \le \cdots \\ &\le \sqrt{2\beta}\, N_{\mathsf{prox}}\, \varepsilon_{\mathsf{RGO}} + W_2(\mu^X_0, \pi^X)\,. \end{aligned}$$

Therefore, we obtain

$$\begin{aligned} &\frac{1}{\mathsf{KL}(\tilde{\mu}^X_{N_{\mathsf{prox}}} \,\|\, \pi^X)} \\ &\quad \ge \frac{1}{\mathsf{KL}(\mu^X_0 \,\|\, \pi^X)} + \frac{N_{\mathsf{prox}}\, h}{2\left(W_2(\mu^X_0, \pi^X) + \sqrt{2\beta}\, N_{\mathsf{prox}}\, \varepsilon_{\mathsf{RGO}}\right)^2} - N_{\mathsf{prox}}\, \varepsilon_{\mathsf{RGO}}\,. \end{aligned}$$

Let us assume that $\varepsilon_{\mathsf{RGO}} \lesssim W_2(\mu^X_0, \pi^X)/(\sqrt{\beta}\, N_{\mathsf{prox}})$ and $\varepsilon_{\mathsf{RGO}} \lesssim h/W_2(\mu^X_0, \pi^X)$. This reads

$$\frac{1}{\mathsf{KL}(\tilde{\mu}^X_{N_{\mathsf{prox}}} \,\|\, \pi^X)} \ge \frac{1}{\mathsf{KL}(\mu^X_0 \,\|\, \pi^X)} + \Omega\Big(\frac{N_{\mathsf{prox}}\, h}{2W_2^2(\mu^X_0, \pi^X)}\Big)\,.$$

Upon rearranging this and taking $h = \frac{1}{2\beta}$, it implies

$$\mathsf{KL}(\tilde{\mu}^X_{N_{\mathsf{prox}}} \,\|\, \pi^X) \lesssim \frac{\beta W_2^2(\mu^X_0, \pi^X)}{N_{\mathsf{prox}}}\,.$$

Therefore, we obtain $\mathsf{KL}(\tilde{\mu}^X_{N_{\mathsf{prox}}} \,\|\, \pi^X) \le \varepsilon^2$ if we take

$$N_{\mathsf{prox}} \asymp \frac{\beta W_2^2(\mu^X_0, \pi^X)}{\varepsilon^2}\,.$$

To summarize, after considering both cases, we obtain $\mathsf{KL}(\tilde{\mu}^X_{N_{\mathsf{prox}}} \,\|\, \pi^X) \le \varepsilon^2$ if

$$N_{\mathsf{prox}} \asymp \frac{d + \beta \mathfrak{m}^2}{\varepsilon^2}$$

and

$$\varepsilon_{\mathsf{RGO}} \lesssim \min\Big\{\varepsilon^4,\, \frac{\varepsilon^4}{\beta W_2^2(\mu^X_0, \pi^X)},\, \frac{\varepsilon^2}{\beta^{3/2} W_2(\mu^X_0, \pi^X)},\, \frac{1}{\beta W_2(\mu^X_0, \pi^X)}\Big\}\,.$$

By invoking Theorem 6.8.1, the total number of first-order queries is

$$N = \widetilde{O}\Big(\frac{\beta d^{1/2}\, W_2^2(\mu^X_0, \pi^X)}{\varepsilon^2}\Big)\,.$$

6.9 Conclusion

Here we mention several interesting questions for future research that are inspired by our results.

- **Are warm starts essential for future progress in high-accuracy sampling?** Our work is the first to show the achievability of the faster rates proven for high-accuracy samplers under a warm start assumption. We do this by exhibiting an efficient algorithm for producing the warm start. We believe that this general strategy may be important for future progress in high-accuracy sampling. Indeed, the natural next candidate for improving upon MALA is Metropolized Hamiltonian Monte Carlo [Nea11], or related variants. For such Metropolized algorithms, we suspect that much of the intuition from §6.1.2 remains true; namely, that the algorithm can benefit from a more aggressive step size near stationarity. Hence, to extract the full potential of these algorithms, it seems likely that we must again pursue the dual plan of improving the rates under a warm start, and efficiently computing that warm start. Insofar as warm starts continue to play an important role in sampling analysis, the Rényi analysis techniques that we developed in §6.3 and §6.4 could prove useful for future progress in this direction.

- **Can we leverage shifted divergence techniques for further advances in differential privacy and beyond?** Core to our results is an improved version of the shifted Rényi divergence technique that uses Orlicz–Wasserstein shifts rather than W_∞ shifts. Since their introduction, shifted divergences have been instrumental for advances in differentially private optimization (see the prior work discussion in §6.1.3), and also very recently in the context of sampling ([AT22b] and Theorem 6.4.1). We believe that we are only scratching the surface of potential applications, extensions, and refinements of this technique, and we are optimistic that a deeper understanding of our Rényi analysis toolbox will have implications far beyond.

Chapter 7

Lower bound in one dimension

In the previous chapters, we have focused on obtaining *upper bounds* on the complexity of sampling via algorithmic guarantees. However, a full understanding of the query complexity also requires matching *lower bounds*, which have been largely and conspicuously absent from the literature.

In this chapter, we establish the first tight lower bound of $\Omega(\log\log\kappa)$ on the query complexity of sampling from the class of strongly log-concave and log-smooth distributions with condition number κ in one dimension. Whereas existing guarantees for MCMC-based algorithms scale polynomially in κ, we introduce a novel rejection sampling algorithm that closes this doubly exponential gap.

This chapter is based on [Che+22d], joint with Patrik R. Gerber, Chen Lu, Thibaut Le Gouic, and Philippe Rigollet.

7.1 Introduction

The task of sampling from a target probability distribution known up to a normalizing constant is of fundamental importance in fields such as Bayesian statistics, randomized algorithms, and online learning. Recently, there has been a resurgence of interest in sampling and its interplay with the more well-developed field of optimization. On the one hand, the extensive optimization toolkit has inspired the development of novel sampling algorithms [Ber18; Wib18; Zha+20; DL21; LST21c]; on the other hand, the theory of optimization has motivated researchers to provide quantitative and non-asymptotic convergence guarantees for sampling methods, which depend on parameters that describe the problem complexity (e.g., the condition number and the dimension); see the previous chapters of this thesis.

Conspicuously absent from this interplay, however, are *lower bounds* for the complexity of sampling, in analogy to the oracle lower bounds initiated in the seminal work by Nemirovski and Yudin for optimization [NY83]. Besides charting the fundamental limits of optimization, such lower bounds have been instrumental in the development of faster algorithms, most notably Nesterov's acceleration,

which was "found mainly because the investigating of complexity enforced to believe that such a method should exist" [Nem94, §10].

A canonical structured class of distributions is that of strongly log-concave and log-smooth distributions on $\mathbb{R}^d$, i.e., the class of distributions with a density $\pi \propto \exp(-V)$, where the potential $V : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable, α-strongly convex, and β-smooth. The relevant parameters of this class are the dimension d, as well as the *condition number* $\kappa := \beta/\alpha$, and we seek to understand the number of queries to V (and its derivatives) necessary to generate a sample close in total variation distance to π. We call a solution to this problem a *general sampling lower bound*.

Related works. Despite several attempts at establishing query complexity lower bounds for sampling, we are not aware of a general sampling lower bound. Whereas sampling upper bounds are derived using techniques that are close to those employed in optimization [Dal17b; DMM19], it is unclear how to use lower bound techniques for optimization [Nes18] to derive general sampling lower bounds. Note that sampling upper bounds typically assume that the minimizer of V is known a priori; thus, a direct reduction of the sampling task to apply existing optimization lower bounds would likely capture the complexity of finding the mode of V rather than the intrinsic difficulty of the sampling task itself. In lieu of a direct reduction, it is possible to envision, at least in principle, an approach which adapts the optimization lower bound constructions to the sampling setting, but we are not aware of any successful results in this direction.

Another family of approaches is based on information-theoretic ideas which have been highly successful for developing a minimax theory of statistics [Le 86; LY00; Tsy09]; however, prior works applying these ideas have largely focused on various adjacent questions which do not imply a lower bound for the sampling task itself. A notable example is the estimation of the normalizing constant of a strongly log-concave distribution, for which a lower bound was established in [GLL20]. However, this lower bound does not yield a general sampling lower bound; in fact, the two problems differ in difficulty. Indeed, the randomized midpoint discretization of the underdamped Langevin dynamics [SL19] obtains samples in $\widetilde{O}(d^{1/3})$ queries, whereas the lower bound for estimating the normalizing constant in [GLL20] grows as $\widetilde{\Omega}(d)$. Another example is the paper [CBL22], which studies sampling with access to stochastic gradient queries. However, the resulting lower bound arises primarily out of the need to overcome the noise in the gradient queries, and it again does not yield sampling lower bounds for our setting of precise gradient queries.

To circumvent the difficulties in establishing general sampling lower bounds, various works have focused on establishing lower bounds for specific and popular

algorithms such as underdamped Langevin Monte Carlo (ULMC) [CLW21] and the Metropolis-adjusted Langevin algorithm (MALA) [see §5 and LST21a; WSC22]. In particular, the latter results establish that the minimax query complexity for MALA over the class of strongly log-concave and log-smooth distributions is $\widetilde{\Theta}(\kappa d)$ from a "cold start" and $\widetilde{\Theta}(\kappa\sqrt{d})$ from a "warm start".

A lower complexity bound in one dimension. Recall that for convex optimization, there are two relevant regimes [see, e.g., Bub15]: (1) the low-dimensional regime, in which algorithms such as the cutting plane method achieve the rate $O(d\log(1/\varepsilon))$ (where ε is the accuracy parameter), and (2) the high-dimensional regime, in which algorithms such as gradient descent achieve dimension-free rates at the cost of inverse polynomial dependency on the accuracy. In this paper, we study the low-dimensional regime for sampling; in particular, we consider $d = 1$.

We prove that for the class of α-strongly log-concave and β-log-smooth distributions in one dimension (with mode at 0), any algorithm, which can produce a sample that is at total variation distance at most $\frac{1}{64}$ from the target distribution π (uniformly over π belonging to the class), must make at least $\Omega(\log\log\kappa)$ queries to V or any of its derivatives. To our knowledge, this is the first lower complexity bound for this problem class.

Achievability of the lower bound. The lower bound of $\Omega(\log\log\kappa)$ is surprisingly small, and existing guarantees for standard algorithms such as the Langevin algorithm (or its variants), the Metropolis-adjusted Langevin algorithm, or Hamiltonian Monte Carlo, all have a dependence that scales polynomially with the condition number κ [see for instance the comparison in SL19].

To provide an algorithm which matches the lower bound, we return to the fundamental idea of rejection sampling, developed by John von Neumann and Stan Ulam [Neu51; Eck87]. We develop an algorithm which uses $O(\log\log\kappa)$ queries in order to build a proposal distribution. Once the proposal distribution is constructed, new samples which are ε-close to π in total variation distance can be generated using $O(\log(1/\varepsilon))$ additional queries per sample.

7.2 Lower bound

We begin by formally defining the class of strongly log-concave and log-smooth distributions in one dimension, which is the focus of this paper.

Definition 7.2.1. *The class of univariate* α-strongly log-concave and β-log-smooth *distributions, for constants* $0 < \alpha \le \beta$, *is the class of continuous distributions* π *supported on* $\mathbb{R}$, *whose density is of the form* $\pi(x) = \exp(-V(x))$, *for a potential function* $V : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ *which is twice continuously differentiable*

and satisfies

$$\alpha \le V''(x) \le \beta\,, \qquad \forall x \in \mathbb{R}\,. \tag{7.1}$$

In addition, we always assume[1] *that the mode of the distribution is at* 0*, or equivalently* $V'(0) = 0$*.*

We study the *query complexity* of sampling from this class. Formally, suppose that the target distribution is $\pi = \exp(-V)$. The sampling algorithm is allowed to make queries to the following oracle: given a point $x \in \mathbb{R}$, the oracle returns some or all of (1) the evaluation of the potential $V(x) + C$ up to a constant C, which is unknown to the algorithm but does not change from query to query; (2) the evaluation of the gradient $V'(x)$; or (3) the evaluation of the Hessian $V''(x)$. Depending on what information the oracle returns, it may be described as providing 0^{th}-, 1^{st}-, or 2^{nd}-order information. For instance, the Langevin algorithm uses 1^{st}-order information, whereas the Metropolis-adjusted Langevin algorithm uses both 0^{th}-order and 1^{st}-order information. Our lower bound will in fact apply to the strongest of these oracles, namely the one that returns all three pieces of information (or more generally, any *local* oracle [NY83]).

We now state our lower bound.

Theorem 7.2.2. *Consider the class* $\mathcal{P}$ *of univariate* α*-strongly log-concave and* β*-log-smooth distributions as defined in Definition 7.2.1, and let* $\kappa := \beta/\alpha$ *denote the condition number. Suppose that an algorithm satisfies the following guarantee: for any* $p \in \mathcal{P}$*, the algorithm makes* n *queries to the oracle providing* 0^{th}*-,* 1^{st}*-, and* 2^{nd}*-order information for* π*, and outputs a random variable whose law is at most* $\frac{1}{64}$ *away from* π *in total variation distance. Then,* $n \gtrsim \log\log\kappa$*.*

We now give some intuition for the lower bound construction, and defer the proof to §7.4. The strategy is to construct a family of distributions $\{\pi_1, \dots, \pi_m\}$ which forms a packing of the class $\mathcal{P}$ in total variation distance. Because the family is well-separated, if an algorithm can accurately sample from each π_i, it can also *identify* π_i. We construct the family $\{\pi_i\}_{i=1}^m$ in such a way that identifying π_i from queries to local oracles requires at least $\Omega(\log m)$ queries, e.g., via bisection.

With the strategy in place, we now describe motivation for the construction of the family $\{\pi_i\}_{i=1}^m$. Suppose that we have a distribution $\pi \propto \exp(-V)$ which is rescaled to satisfy $1 \le V'' \le \kappa$. The bound $V'' \ge 1$ implies that a substantial fraction of the mass of π is supported on the interval $[-1, 1]$. On the other hand,

[1] This localization assumption is common in the sampling literature; without some knowledge of the mode (e.g., that the mode is contained in an interval) it is impossible to even find the mode in the query model.

the bound $V'' \leq \kappa$ allows for the density π to suddenly drop from ≈ 1 to nearly 0 over an interval of much smaller length, $\asymp 1/\sqrt{\kappa}$. Hence, as a first approximation, we can imagine dividing the interval $[-1, 1]$ into $\asymp \sqrt{\kappa}$ bins, and thinking of each π_i as piecewise constant on each bin. While keeping the log-concavity constraint in mind, for the purpose of this heuristic discussion we will consider the family $\{\pi_i\}_{i=1}^m$ of $m \asymp \sqrt{\kappa}$ distributions, where π_i is the uniform distribution on $[-i/\sqrt{\kappa}, i/\sqrt{\kappa}]$; see Figure 7.1.

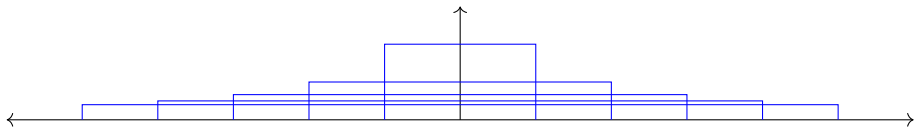


Figure 7.1: A family of uniform distributions.

However, this family is not well-separated in total variation distance. Indeed, it can be checked that for $i < j$, in order for the total variation distance between π_i and π_j to be appreciable, we require $j \geq 2i$. This motivates us to consider the subfamily $\{\pi_{2^i}, 1 \leq i \leq \log_2 \sqrt{\kappa}\}$, of which there are $O(\log \kappa)$ elements. For this subfamily, we can hope to reduce the task of sampling to that of identifying π_{2^i} via queries, and binary search for this problem requires only $O(\log \log \kappa)$ queries. This is the basis for our somewhat unusual lower bound.

The uniform distributions involved in this informal discussion do not belong to the class $\mathcal{P}$, as they are neither strongly log-concave nor log-smooth. The main technical challenge in our lower bound is to produce distributions which lie in $\mathcal{P}$ but still behaves similarly to uniform distributions, in the sense of requiring $\Omega(\log \log \kappa)$ oracle queries to identify a distribution via queries. We defer these details to §7.4.

7.3 Upper bound

In this section, we show that the $\Omega(\log \log \kappa)$ lower bound in the previous section is achievable. Note that the existing guarantees for standard sampling algorithms (c.f. the comparison in [SL19]) usually scale polynomially in the condition number κ, so they are not optimal for our setting.

Moreover, the heuristic discussion of the lower bound construction motivates choosing the query points according to a binary search strategy. In order to implement this idea, we turn towards the classical idea of rejection sampling: first, we make queries in order to construct a *proposal* distribution q. To generate new samples from π, we repeatedly draw samples from q, and each sample is

accepted with a carefully chosen acceptance probability (which can be computed via additional queries to the oracle for the density up to normalization).

Algorithm 7.1 ENVELOPE

Use binary search to find the first index $i_+ \in \{0, 1, \ldots, \lceil \frac{1}{2} \log_2 \kappa \rceil\}$ with $V(2^{i_+}/\sqrt{\kappa}) \geq \frac{1}{2}$.
Use binary search to find the first index $i_- \in \{0, 1, \ldots, \lceil \frac{1}{2} \log_2 \kappa \rceil\}$ with $V(-2^{i_-}/\sqrt{\kappa}) \geq \frac{1}{2}$.
Set $x_- \coloneqq -2^{i_-}/\sqrt{\kappa}$ and $x_+ \coloneqq 2^{i_+}/\sqrt{\kappa}$.
return

$$\tilde{q}(x) \coloneqq \begin{cases} \exp\Big[-\dfrac{x - x_-}{2x_-} - \dfrac{(x - x_-)^2}{2}\Big], & x \leq x_-\,, \\ 1\,, & x_- \leq x \leq x_+\,, \\ \exp\Big[-\dfrac{x - x_+}{2x_+} - \dfrac{(x - x_+)^2}{2}\Big], & x \geq x_+\,. \end{cases}$$

We give the high-level pseudocode for building an upper envelope in Algorithm 7.1, and for generating new samples in Algorithm 7.2. Note that while our lower bound applies to algorithms using 0^{th}-, 1^{st}-, and 2^{nd}-order information, our upper bound algorithm in fact only requires 0^{th}-order information. We next proceed to discuss details of the algorithms.

Algorithm 7.2 SAMPLE

Normalize $\tilde{q}$ to form q.
while sample is not accepted **do**
  Sample $X \sim q$.
  Accept X w.p. $\tilde{\pi}(X)/\tilde{q}(X)$.
return X

Before implementing Algorithm 7.1, we first perform several preprocessing steps. Recall that the mode of the distribution π is assumed to be at 0, and that $\pi \propto \exp(-V)$. We also assume that $1 \leq V'' \leq \kappa$. To reduce to this case, say we start with $\alpha \leq V'' \leq \beta$, and the bounds α, β are known. Then, observe that the rescaled potential $\bar{V}(x) \coloneqq V(x/\sqrt{\alpha})$ satisfies $1 \leq \bar{V}'' \leq \kappa = \beta/\alpha$. Given access to an oracle for V (up to additive constant), we can simulate an oracle to $\bar{V}$ (up to additive constant) and apply our algorithm to generate a sample $\bar{X}$ from the density $\bar{\pi} \propto \exp(-\bar{V})$; it can be checked that $\bar{X}/\sqrt{\alpha}$ is a sample from π. Finally,

we assume that the oracle, when given a query point x, returns $V(x)$, where V is normalized to satisfy $V(0) = 0$; this is achieved by replacing the output $V(x)$ of the oracle by $V(x) - V(0)$.

Implementing the first step of Algorithm 7.1 requires performing binary search over an array of size $O(\log \kappa)$, which requires only $O(\log \log \kappa)$ queries; similar comments apply to the second step. We prove in §7.5 that the indices i_- and i_+ always exist under our assumptions. We also prove in §7.5 that the output $\tilde{q}$ of Algorithm 7.1 is an *upper envelope* for the oracle, i.e., $\tilde{q} \geq \exp(-V)$. The upper envelope $\tilde{q}$ constructed in Algorithm 7.1 is the input to Algorithm 7.2; see Figure 7.2 for a visualization.

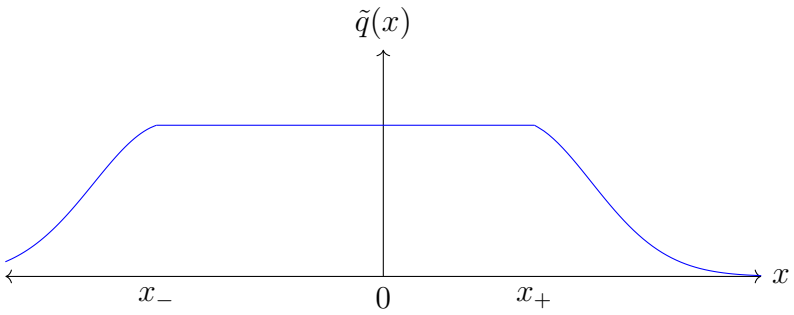


Figure 7.2: The upper envelope $\tilde{q}$ constructed in Algorithm 7.1.

In Algorithm 7.2, we normalize $\tilde{q}$ to a probability distribution q, which requires computing a one-dimensional integral for the normalizing constant: $\int_{\mathbb{R}} \tilde{q}$. Once normalized, we must also be able to draw samples from the distribution q. These steps can be implemented with low computational burden, but we do not dwell on this point here because we are primarily interested in the *query complexity* in this work. Note that the steps of normalizing q and drawing new samples from q do not require additional queries to the oracle.

The framework of rejection sampling provides a flexible guarantee: if we desire an *exact* sample from π, then we can continue drawing samples from q until one is accepted, yielding an exact sample with a guarantee on the expected total number of queries. On the other hand, if we are content with producing a sample whose law is at a fixed distance ε away from π in total variation distance, then we can force the algorithm to stop after a prespecified number of iterations, declaring failure if no sample from q is accepted, and achieve the total variation guarantee. We describe both of these guarantees in the following theorem, which summarizes the query complexity of our algorithm.

Theorem 7.3.1. *Suppose that the target distribution π belongs to the class of univariate strongly log-concave and log-smooth distributions (Definition 7.2.1).*

Algorithm 7.1 uses $O(\log\log\kappa)$ queries to build the upper envelope $\tilde{q}$. Once $\tilde{q}$ is constructed, we can use it for either of the following tasks.

1. *(exact sampling) Algorithm 7.2 returns an exact sample from π after an additional $O(1)$ expected queries to the oracle.*
2. *(approximate sampling) Fix an accuracy parameter $0 < \varepsilon < 1$. If we limit Algorithm 7.2 to use at most $O(\log(1/\varepsilon))$ queries, then the output of Algorithm 7.2 (or 'FAILURE', if Algorithm 7.2 fails to accept a sample within the allowed number of queries) has a distribution which is at total variation distance at most ε away from π.*

We give the proof in §7.5.

Although our algorithm is tailored to distributions in one dimension, we remark that the task of sampling from a one-dimensional log-concave distribution is a subroutine for the hit-and-run algorithm, which is explored in [Che+22d].

7.4 Proof of the lower bound

7.4.1 The construction

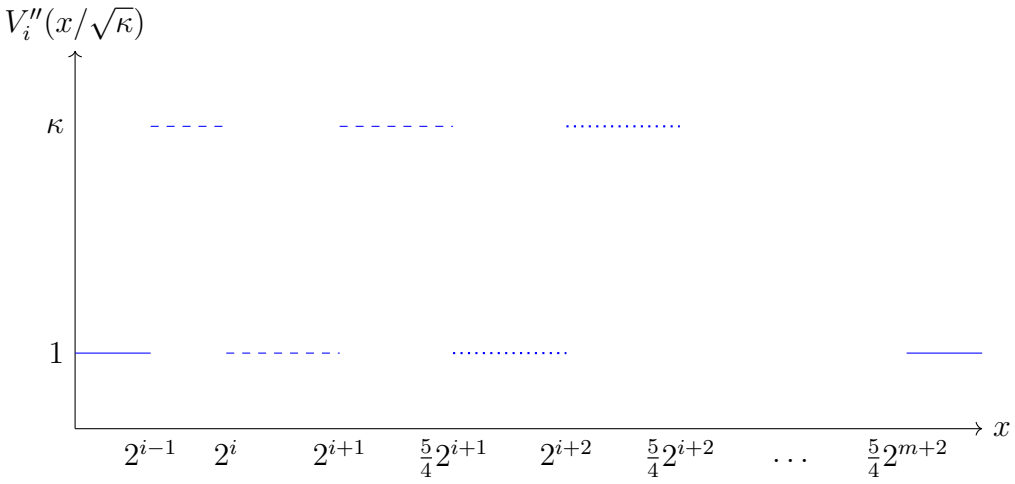


Figure 7.3: The dashed lines correspond to ϕ and the dotted lines correspond to ψ.

Let m be the largest integer such that

$$\exp\Big(-\frac{2^{2m-2}}{2\kappa}\Big) \geq \frac{1}{2}\,. \tag{7.2}$$

Define two auxiliary functions

$$\phi(x) := \begin{cases} \kappa\,, & 1/2 \le x < 1\,, \\ 1\,, & 1 \le x < 2\,, \\ \kappa\,, & 2 \le x < 5/2\,, \\ 0 & \text{otherwise}\,, \end{cases} \qquad \psi(x) := \begin{cases} 1\,, & 5/2 \le x < 4\,, \\ \kappa\,, & 4 \le x < 5\,, \\ 0\,, & \text{otherwise}\,. \end{cases}$$

We define a family $(V_i)_{i=1}^m$ of 1-strongly convex and κ-smooth potentials as follows. We require that $V_i(0) = V_i'(0) = 0$ and that V_i be an even function, so it suffices to specify V_i'' on $\mathbb{R}_+$. For $x \ge 0$, the second derivative is given by

$$V_i''(x) := \mathbb{1}\{x \le \kappa^{-\frac{1}{2}} 2^{i-1}\} + \phi\Big(\frac{x}{\kappa^{-\frac{1}{2}} 2^i}\Big) + \sum_{j=i}^{m-1} \psi\Big(\frac{x}{\kappa^{-\frac{1}{2}} 2^j}\Big) + \mathbb{1}\{x \ge 5\kappa^{-\frac{1}{2}} 2^{m-1}\}\,.$$

Observe that all of the terms in the above summation have disjoint supports, see Figure 7.3.

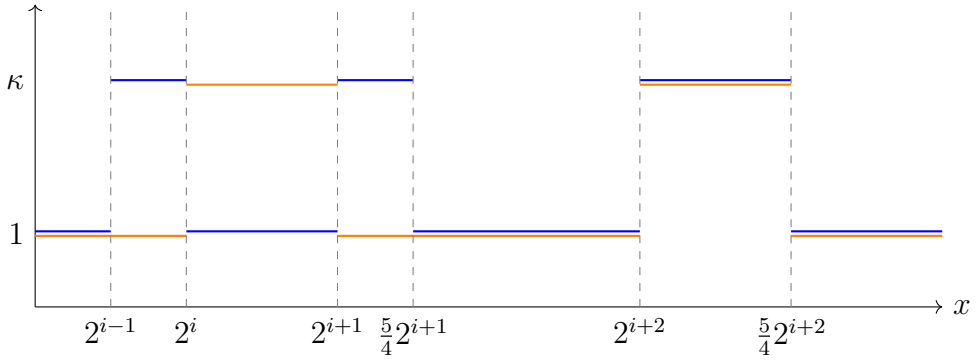


Figure 7.4: We plot V_i'' (in blue) and V_{i+1}'' (in orange). In this figure, we do not distort the horizontal axis lengths to make it easier to visually compare the relative lengths of intervals on which the second derivatives are constant.

The following lemma provides intuition for the construction.

Lemma 7.4.1. *We have the equalities*

$$\begin{aligned} V_i &= V_{i+1}\,, \\ V_i' &= V_{i+1}'\,, \\ V_i'' &= V_{i+1}''\,, \end{aligned}$$

outside of the set $\{x \in \mathbb{R} : \kappa^{-\frac{1}{2}} 2^{i-1} \le |x| \le \frac{5}{4}\kappa^{-\frac{1}{2}} 2^{i+1}\}$.

Proof. Refer to Figure 7.4 for a visual aid for the proof.

Clearly the potentials and derivatives match when $|x| \le \kappa^{-\frac{1}{2}} 2^{i-1}$. Since the second derivatives match when $|x| \ge \frac{5}{4}\kappa^{-\frac{1}{2}} 2^{i+1}$, it suffices to show that $V_i'(\frac{5}{4}\kappa^{-\frac{1}{2}} 2^{i+1}) = V_{i+1}'(\frac{5}{4}\kappa^{-\frac{1}{2}} 2^{i+1})$ and $V_i(\frac{5}{4}\kappa^{-\frac{1}{2}} 2^{i+1}) = V_{i+1}(\frac{5}{4}\kappa^{-\frac{1}{2}} 2^{i+1})$.

To that end, note that for $x \ge 0$,

$$\begin{aligned} V_{i+1}''(x) - V_i''(x) &= \mathbb{1}\{\kappa^{-\frac{1}{2}} 2^{i-1} < x \le \kappa^{-\frac{1}{2}} 2^i\} \\ &\qquad - \phi\big(\frac{x}{\kappa^{-\frac{1}{2}} 2^i}\big) + \phi\big(\frac{x}{\kappa^{-\frac{1}{2}} 2^{i+1}}\big) - \psi\big(\frac{x}{\kappa^{-\frac{1}{2}} 2^i}\big) \\ &= \begin{cases} -(\kappa - 1)\,, & \kappa^{-\frac{1}{2}} 2^{i-1} \le x \le \kappa^{-\frac{1}{2}} 2^i\,, \\ +(\kappa - 1)\,, & \kappa^{-\frac{1}{2}} 2^i \le x \le \kappa^{-\frac{1}{2}} 2^{i+1}\,, \\ -(\kappa - 1)\,, & \kappa^{-\frac{1}{2}} 2^{i+1} \le x \le \frac{5}{4}\kappa^{-\frac{1}{2}} 2^{i+1}\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \end{aligned}$$

A little algebra shows that the above expression integrates to zero, hence we deduce the equality $V_i'(\frac{5}{4}\kappa^{-\frac{1}{2}} 2^{i+1}) = V_{i+1}'(\frac{5}{4}\kappa^{-\frac{1}{2}} 2^{i+1})$. Also, by integrating this expression twice, we see that

$$\begin{aligned} &V_{i+1}\big(\frac{5}{4}\,\kappa^{-\frac{1}{2}} 2^{i+1}\big) - V_i\big(\frac{5}{4}\,\kappa^{-\frac{1}{2}} 2^{i+1}\big) \\ &\quad = \underbrace{-\frac{\kappa - 1}{2}\,(\kappa^{-\frac{1}{2}} 2^{i-1})^2}_{\text{integral on } [\kappa^{-\frac{1}{2}} 2^{i-1}, \kappa^{-\frac{1}{2}} 2^i]} \underbrace{- (\kappa - 1)\,\kappa^{-\frac{1}{2}} 2^{i-1}\,\kappa^{-\frac{1}{2}} 2^i + \frac{\kappa - 1}{2}\,(\kappa^{-\frac{1}{2}} 2^i)^2}_{\text{integral on } [\kappa^{-\frac{1}{2}} 2^i, \kappa^{-\frac{1}{2}} 2^{i+1}]} \\ &\qquad \underbrace{+ (\kappa - 1)\,\kappa^{-\frac{1}{2}} 2^{i-1}\,\frac{1}{4}\,\kappa^{-\frac{1}{2}} 2^{i+1} - \frac{\kappa - 1}{2}\,\big(\frac{1}{4}\,\kappa^{-\frac{1}{2}} 2^{i+1}\big)^2}_{\text{integral on } [\kappa^{-\frac{1}{2}} 2^{i+1}, \frac{5}{4}\kappa^{-\frac{1}{2}} 2^{i+1}]} \\ &\quad = \frac{\kappa - 1}{\kappa}\,\{-2^{2i-3} - 2^{2i-1} + 2^{2i-1} + 2^{2i-2} - 2^{2i-3}\} \\ &\quad = 0\,, \end{aligned}$$

as desired. □

We also need a lemma showing that each probability distribution $\pi_i \propto \exp(-V_i)$ places a substantial amount of mass on the interval $(\kappa^{-\frac{1}{2}} 2^{i-2}, \kappa^{-\frac{1}{2}} 2^{i-1}]$.

Lemma 7.4.2. *For each $i \in [m]$,*

$$\pi_i\big((\kappa^{-\frac{1}{2}} 2^{i-2}, \kappa^{-\frac{1}{2}} 2^{i-1}]\big) \ge \frac{1}{32}\,.$$

Proof. According to the definition of π_i, we have

$$\pi_i\big((\kappa^{-\frac{1}{2}}2^{i-2}, \kappa^{-\frac{1}{2}}2^{i-1}]\big) = \frac{\int_{\kappa^{-\frac{1}{2}}2^{i-2}}^{\kappa^{-\frac{1}{2}}2^{i-1}} \exp(-x^2/2)\,\mathrm{d}x}{Z_{\pi_i}}\,, \qquad Z_{\pi_i} \coloneqq \int_{\mathbb{R}} \exp(-V_i)\,.$$

Recalling that m is chosen so that $\exp(-x^2/2) \geq 1/2$ whenever $|x| \leq \kappa^{-\frac{1}{2}}2^{m-1}$ (see (7.2)), we can conclude that

$$\int_{\kappa^{-\frac{1}{2}}2^{i-2}}^{\kappa^{-\frac{1}{2}}2^{i-1}} \exp\big(-\frac{x^2}{2}\big)\,\mathrm{d}x \geq \frac{1}{2}\,\kappa^{-\frac{1}{2}}2^{i-2}\,.$$

For the normalizing constant, observe that

$$\int_0^\infty \exp(-V_i) = \int_0^{\kappa^{-\frac{1}{2}}2^i} \exp(-V_i) + \int_{\kappa^{-\frac{1}{2}}2^i}^\infty \exp(-V_i) \leq \kappa^{-\frac{1}{2}}2^i + \int_{\kappa^{-\frac{1}{2}}2^i}^\infty \exp(-V_i)\,.$$

Since $V_i'' = \kappa$ on $[\kappa^{-\frac{1}{2}}2^{i-1}, \kappa^{-\frac{1}{2}}2^i]$, it follows that $V_i'(\kappa^{-\frac{1}{2}}2^i) \geq \kappa^{\frac{1}{2}}2^{i-1}$, and so

$$V_i(x) \geq \kappa^{\frac{1}{2}}2^{i-1}\,(x - \kappa^{-\frac{1}{2}}2^i) + \frac{\big(x - \kappa^{-\frac{1}{2}}2^i\big)^2}{2}\,, \qquad x \geq \kappa^{-\frac{1}{2}}2^i\,.$$

Therefore,

$$\begin{aligned}\int_{\kappa^{-\frac{1}{2}}2^i}^\infty \exp(-V_i) &\leq \int_{\kappa^{-\frac{1}{2}}2^i}^\infty \exp\Big(-\kappa^{\frac{1}{2}}2^{i-1}\,(x - \kappa^{-\frac{1}{2}}2^i) - \frac{\big(x - \kappa^{-\frac{1}{2}}2^i\big)^2}{2}\Big)\,\mathrm{d}x \\ &\leq \frac{1}{\kappa^{\frac{1}{2}}2^{i-1}} \leq \frac{1}{\sqrt{\kappa}}\,,\end{aligned}$$

where we applied a standard tail estimate for Gaussian densities (Lemma 7.4.3). Putting it together,

$$\pi_i\big((\kappa^{-\frac{1}{2}}2^{i-2}, \kappa^{-\frac{1}{2}}2^{i-1}]\big) \geq \frac{2^{i-3}}{2\,(2^i+1)} \geq \frac{1}{32}\,,$$

which proves the result. □

Above, we used the following elementary lemma about Gaussian integrals.

Lemma 7.4.3. *Let $a, x_0 > 0$. Then,*

$$\int_{x_0}^\infty \exp\big(-a\,(x - x_0) - \frac{1}{2}\,(x - x_0)^2\big)\,\mathrm{d}x \leq \frac{1}{a}\,.$$

Proof. Completing the square,

$$\int_{x_0}^{\infty} \exp\bigl(-a\,(x-x_0) - \frac{1}{2}\,(x-x_0)^2\bigr)\,\mathrm{d}x = \int_0^{\infty} \exp\bigl(-ax - \frac{1}{2}\,x^2\bigr)\,\mathrm{d}x$$
$$= \sqrt{2\pi}\exp\bigl(\frac{a^2}{2}\bigr)\,\mathbb{P}(Z > a)\,,$$

where $Z \sim \mathsf{normal}(0,1)$. The result follows from the Mills ratio inequality [Gor41]. □

■ 7.4.2 Lower bound via Fano's inequality

In this section, we use the densities $\{\pi_i\}_{i=1}^m$ constructed in the previous section together with Fano's inequality from information theory in order to prove the lower bound.

Proof of Theorem 7.2.2. Let $Z \sim \mathsf{uniform}([m])$ be an index chosen uniformly at random. Suppose that an algorithm makes n queries to the oracle for π_Z, and given $Z = i$, outputs a sample Y whose law μ_i is at total variation distance at most $\frac{1}{64}$ from π_i. In light of Lemma 7.4.2, a good candidate estimator for Z from the observation of Y is given by

$$\widehat{Z} := \{k \in \mathbb{N} : Y \in (\kappa^{-\frac{1}{2}}2^{k-2}, \kappa^{-\frac{1}{2}}2^{k-1}]\}\,.$$

On the one hand, the probability that the estimator is correct is bounded by

$$\begin{aligned}
\mathbb{P}\{\widehat{Z} = Z\} &= \frac{1}{m}\sum_{i=1}^m \mathbb{P}\{\widehat{Z} = i \mid Z = i\} \\
&= \frac{1}{m}\sum_{i=1}^m \mathbb{P}\{Y \in (\kappa^{-\frac{1}{2}}2^{i-2}, \kappa^{-\frac{1}{2}}2^{i-1}] \mid Z = i\} \\
&= \frac{1}{m}\sum_{i=1}^m \mu_i\bigl((\kappa^{-\frac{1}{2}}2^{i-2}, \kappa^{-\frac{1}{2}}2^{i-1}]\bigr) \\
&\geq \frac{1}{m}\sum_{i=1}^m \pi_i\bigl((\kappa^{-\frac{1}{2}}2^{i-2}, \kappa^{-\frac{1}{2}}2^{i-1}]\bigr) - \frac{1}{64} \geq \frac{1}{64}\,, \qquad (7.3)
\end{aligned}$$

where the last inequality uses Lemma 7.4.2.

On the other hand, we can lower bound $\mathbb{P}\{\widehat{Z} \neq Z\}$ using Fano's inequality. Let $x_1, \ldots, x_n$ denote the query points of the algorithm, and let W_i be a shorthand for the triple (V_i, V_i', V_i''). We will first prove the lower bound for deterministic

algorithms, i.e., assuming that each query point x_j is a deterministic function of the previous query points and query values. Since

$$Z \to \{x_j, W_Z(x_j),\ j \in [n]\} \to \widehat{Z}$$

forms a Markov chain, Fano's inequality [CT06] yields

$$\mathbb{P}\{\widehat{Z} \neq Z\} \geq 1 - \frac{I(\{x_j, W_Z(x_j)\}_{j\in[n]}; Z) + \log 2}{\log m}\,,$$

where I denotes the mutual information. By the chain rule for mutual information [CT06],

$$I\big(\{x_j, W_Z(x_j)\}_{j\in[n]}; Z\big) = \sum_{j=1}^{n} I\big(x_j, W_Z(x_j); Z \;\big|\; x_1, W_Z(x_1), \ldots, x_{j-1}, W_Z(x_{j-1})\big)\,.$$

Observe that, conditioned on $\{x_i, W_Z(x_i)\}_{i=1}^{j-1}$, the query point x_j is deterministic. Also, from the construction of the family of potentials, we know that $W_Z(x_j) = W_1(x_j)$ if $x_j \leq \kappa^{-\frac{1}{2}} 2^{Z-1}$, and $W_Z(x_j) = W_m(x_j)$ if $x_j \geq \frac{5}{4}\kappa^{-\frac{1}{2}} 2^{Z+1}$. It yields that:

- for $Z \leq \log_2(\frac{4}{5}\sqrt{\kappa}x_j) - 1$, $W_Z(x_j)$ takes a unique value given by $W_m(x_j)$,
- for $Z \geq \log_2(\sqrt{\kappa}x_j) + 1$, $W_Z(x_j)$ takes a unique value given by $W_1(x_j)$,

and otherwise, Z lives in an interval of size at most $\log_2(\sqrt{\kappa}x_j)+1-(\log_2(\frac{4}{5}\sqrt{\kappa}x_j)-1) \leq 2+\log_2(5/4)$ which covers at most three integers, say z_0-1, z_0, z_0+1. Hence, the conditional distribution of $W_Z(x_j)$ can be supported on at most 5 points given respectively by

$$W_1(x_j), W_m(x_j), W_{z_0-1}(x_j), W_{z_0}(x_j),\ \text{and}\ W_{z_0+1}(x_j)\,.$$

Since the mutual information is upper bounded by the conditional entropy of $W_Z(x_j)$, we can conclude

$$I\big(\{x_j, W_Z(x_j)\}_{j\in[n]}; Z\big) \leq n \log 5\,.$$

Substituting this into Fano's inequality yields

$$\mathbb{P}\{\widehat{Z} \neq Z\} \geq 1 - \frac{n\log 5 + \log 2}{\log m}\,. \tag{7.4}$$

In general, if the algorithm is randomized, then we can apply the inequality (7.4) conditioned on the random seed ξ of the algorithm, since ξ is independent of Z. It yields

$$\mathbb{P}\{\widehat{Z} \neq Z \mid \xi\} \geq 1 - \frac{n\log 5 + \log 2}{\log m}\,,$$

and upon taking expectations we see that (7.4) holds for randomized algorithms as well.

Combined with (7.3), we obtain $n \gtrsim \log m \gtrsim \log\log\kappa$ as desired. □

■ 7.5 Proof of the upper bound

Let π be the target distribution and let $\tilde{\pi} = \pi Z_\pi$ denote the unnormalized distribution which we access via oracle queries. We recall our preprocessing steps: we assume that the query values take the form $\tilde{\pi}(x) = \exp(-V(x))$, with $V(0) = V'(0) = 0$ and V satisfying (7.1). This is without loss of generality because we can query $\tilde{\pi}(0)$ and replace subsequent queries $\tilde{\pi}(x)$ with $\tilde{\pi}(x)/\tilde{\pi}(0)$, thereby normalizing V to satisfy $V(0) = 0$. By rescaling the distribution, we can assume that $1 \le V'' \le \kappa$. Also, we can assume that the target distribution is only supported on the positive reals $\mathbb{R}_+$, because we can then construct an upper envelope on all of $\mathbb{R}$ by repeating our algorithm on the negative reals, which only doubles the number of queries and does not change the complexity.

Proof of Theorem 7.3.1. Our goal is to use the oracle queries to construct an upper envelope $\tilde{q}$ that satisfies $\tilde{q} \ge \tilde{\pi}$, and $Z_q \lesssim Z_\pi$, where

$$Z_\pi \coloneqq \int_{\mathbb{R}} \tilde{\pi}\,, \qquad Z_q \coloneqq \int_{\mathbb{R}} \tilde{q}$$

are the normalizing constants. The guarantees of Theorem 7.3.1 will then follow from standard results on rejection sampling, see Theorem 4.4.6.

Let i_0 denote the smallest integer such that $V(2^{i_0}/\sqrt{\kappa}) \ge 1/2$. Note that $x^2/2 \le V(x) \le \kappa x^2/2$ implies that $0 \le i_0 \le (\log_2 \kappa)/2$. Using binary search over an array of size $O(\log\kappa)$, we can find i_0 using only $O(\log\log\kappa)$ queries to $\tilde{\pi}$.

Let $x_0 \coloneqq 2^{i_0}/\sqrt{\kappa}$. We first claim that

$$\int_0^{x_0} \tilde{\pi} \gtrsim x_0\,. \tag{7.5}$$

When $i_0 = 0$, this holds because

$$\int_0^{x_0} \tilde{\pi} = \int_0^{1/\sqrt{\kappa}} \exp(-V) \ge \int_0^{1/\sqrt{\kappa}} \exp\Big(-\frac{\kappa x^2}{2}\Big)\,\mathrm{d}x \ge \frac{1}{3\sqrt{\kappa}} = \frac{x_0}{3}\,.$$

When $i_0 > 0$, this holds because, by definition of i_0, we have $V(x_0/2) \le 1/2$ and

$$\int_0^{x_0} \tilde{\pi} \ge \int_0^{x_0/2} \exp(-V) \gtrsim x_0\,.$$

Next, define the upper envelope as follows:

$$\tilde{q}(x) = \begin{cases} 1\,, & x \le x_0\,, \\ \exp\{-(x-x_0)/(2x_0) - (x-x_0)^2/2\}\,, & x > x_0\,. \end{cases}$$

To see that $\tilde{q} \ge \tilde{\pi}$ and hence that $\tilde{q}$ is a valid upper envelope, observe first that since $\tilde{\pi}(0) = 1$, and $\tilde{\pi}$ is decreasing, we get that $\tilde{\pi}(x) \le 1 = \tilde{q}(x)$ for all $x \in [0, x_0]$.

Next, if $x > x_0$, using the fact that V is convex and $V(x_0) \ge 1/2$ by the definition of x_0,

$$V'(x_0) \ge \frac{V(x_0) - V(0)}{x_0} \ge \frac{1}{2x_0}\,.$$

Hence, for any $x > x_0$ we have

$$\begin{aligned} V(x) &\ge V(x_0) + V'(x_0)\,(x - x_0) + \frac{1}{2}\,(x - x_0)^2 \\ &\ge \frac{1}{2x_0}\,(x - x_0) + \frac{1}{2}\,(x - x_0)^2\,. \end{aligned}$$

It implies that $\tilde{\pi}(x) \le \tilde{q}(x)$ also for the tail $x > x_0$.

To complete the proof, we show that $Z_q \lesssim Z_\pi$. In light of (7.5) it is sufficient to show that $Z_q \lesssim x_0$. To see this, observe that by Lemma 7.4.3, we have

$$Z_q = \int_0^{x_0} \tilde{q} + \int_{x_0}^{\infty} \tilde{q} \le x_0 + \int_{x_0}^{\infty} \exp\Big(-\frac{1}{2x_0}\,(x - x_0) - \frac{1}{2}\,(x - x_0)^2\Big)\,\mathrm{d}x \le 3x_0\,.$$

This completes the proof. □

■ 7.6 Conclusion

In this paper, we established the oracle complexity of sampling from the class of univariate strongly log-concave and log-smooth distributions, in analogy with the now pervasive oracle lower bounds for optimization initiated by Nemirovski and Yudin [NY83]. A clear future direction suggested by this work is to extend this result to higher dimensions, and to ultimately develop a theory of lower complexity bounds and optimal algorithms for sampling. Towards that goal, in the forthcoming work [Che+23b], we pin down the complexity of sampling in two further relevant regimes: in any constant dimension, the complexity is $\Theta(\log \kappa)$, and for the task of sampling from Gaussians, the complexity is $\widetilde{\Theta}(\min\{d, \sqrt{\kappa}\log d\})$. In particular, since Gaussians are a subfamily of the class of strongly log-concave and log-smooth distributions, a lower bound for sampling from the former readily

furnishes a lower bound for sampling from the latter. In another direction, we investigate lower bounds for *non-log-concave* sampling in §14.

Recently, an intense amount of research has been devoted to the use of Markov chain Monte Carlo-based methods for sampling, and it may come as a surprise that the complexity lower bound we have proven in this paper is attained by an entirely different type of algorithm, namely rejection sampling. Our result highlights that standard algorithms may not be optimal, and that the search for optimal algorithms goes hand-in-hand with lower bound constructions.

In particular, our work motivates revisiting the idea of rejection sampling through the modern lens of minimax optimality. See [Che+22e] for an investigation of the complexity of rejection sampling in a discrete setting.

Part II

Constrained sampling and Brascamp–Lieb

Chapter 8

Continuous-time analysis of the mirror Langevin diffusion

In the first part of the thesis, we focused on *unconstrained* sampling, i.e., sampling from densities on $\mathbb{R}^d$ with full support. In this chapter, we begin our investigations of the mirror Langevin diffusion, the sampling analogue of the mirror descent algorithm from optimization. Our main result highlights the role of relative convexity of the potential w.r.t. the mirror map for establishing convergence of the diffusion in continuous time, and this same assumption will be used in §9 to obtain non-asymptotic guarantees for *constrained* sampling. As a special case of this framework, we propose a class of diffusions called Newton Langevin diffusions and prove that they converge to stationarity exponentially fast with a rate that is independent of the condition number of the problem.

This chapter is based on [Che+20e], joint with Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet, and Austin J. Stromme.

8.1 Introduction

Sampling from a target distribution is a central task in statistics and machine learning with applications ranging from Bayesian inference [RC04; DM19] to deep generative models [Goo+14]. Owing to a firm mathematical grounding in the theory of Markov processes [MT09], as well as its great versatility, Markov chain Monte Carlo (MCMC) has emerged as a fundamental sampling paradigm. While traditional analyses are anchored in the asymptotic framework of ergodic theory, this work focuses on finite-time results that better witness the practical performance of MCMC for high-dimensional problems arising in machine learning.

This perspective parallels an earlier phenomenon in the much better understood field of optimization where convexity has played a preponderant role for both theoretical and methodological advances [Bub15; Nes18]. In fact, sampling and optimization share deep conceptual connections that have contributed to a renewed

understanding of the theoretical properties of sampling algorithms [Dal17a; Wib18] building on the seminal work of Jordan, Kinderlehrer and Otto [JKO98].

We consider the following canonical sampling problem. Let π be a log-concave probability measure over $\mathbb{R}^d$ so that π has density equal to $\exp(-V)$, where the potential $V : \mathbb{R}^d \to \mathbb{R}$ is convex. Throughout this paper, we also assume that V is twice continuously differentiable for convenience, though many of our results hold under weaker conditions.

Many MCMC algorithms for this task are based on the *Langevin diffusion* (LD), that is the solution $(X_t)_{t\geq 0}$ to the stochastic differential equation (SDE)

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \tag{LD}$$

with $(B_t)_{t\geq 0}$ a standard Brownian motion in $\mathbb{R}^d$. Indeed, π is the unique invariant distribution of (LD) and suitable discretizations result in algorithms that can be implemented when V is known only up to an additive constant, which is crucial for applications in Bayesian statistics and machine learning.

A first connection between sampling from log-concave measures and optimizing convex functions is easily seen from (LD): omitting the Brownian motion term yields the gradient flow $\dot{x}_t = -\nabla V(x_t)$, which results in the celebrated gradient descent algorithm when discretized in time [Dal17a; Dal17b]. There is, however, a much deeper connection involving the distribution of X_t rather than X_t itself, and this latter connection has been substantially more fruitful: the marginal distribution of a Langevin diffusion process $(X_t)_{t\geq 0}$ evolves according to a *gradient flow*, over the Wasserstein space of probability measures, that minimizes the Kullback–Leibler (KL) divergence $\mathsf{KL}(\cdot \,\|\, \pi)$ [JKO98; AGS08; Vil09b]. This point of view has led not only to a better theoretical understanding of the Langevin diffusion [Ber18; CB18; Wib18; DMM19; VW19] but it has also inspired new sampling algorithms based on classical optimization algorithms, such as proximal/splitting methods [Ber18; Wib18; Wib19; SR20], mirror descent [Hsi+18; Zha+20], Nesterov's accelerated gradient descent [Che+18b; DR20; Ma+21], and Newton methods [Mar+12; Sim+16; WL20].

Our contributions. This paper further exploits the optimization perspective on sampling by establishing a theoretical framework for a large class of stochastic processes called *mirror Langevin diffusions* (MLD) introduced in [Zha+20]. These processes correspond to alternative optimization schemes that minimize the KL divergence over the Wasserstein space by changing its geometry. They show better dependence in key parameters such as the condition number and the dimension.

Our theoretical analysis is streamlined by a technical device which is unexpected at first glance, yet proves to be elegant and effective: we track the progress of these schemes not by measuring the objective function itself, the KL divergence,

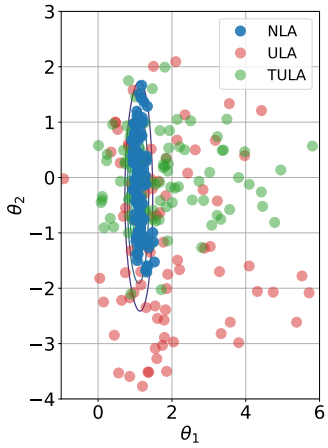


Figure 8.1: Samples from the posterior distribution of a 2D Bayesian logistic regression model using the Newton Langevin algorithm (NLA), the unadjusted Langevin algorithm (ULA), and the tamed unadjusted Langevin algorithm (TULA) [Bro+19]. For details, see §8.7.2.

but rather by measuring the chi-squared divergence to the target distribution π as a surrogate. This perspective highlights the central role of mirror Poincaré inequalities (MP) as sufficient conditions for exponentially fast convergence of the mirror Langevin diffusion to stationarity in chi-squared divergence, which readily yields convergence in other well-known information divergences, such as the Kullback–Leibler divergence, the Hellinger distance, and the total variation distance [Tsy09, §2.4].

We also specialize our results to the case when the mirror map equals the potential V. This can be understood as the sampling analogue of Newton's method, and we therefore call it the *Newton Langevin diffusion* (NLD). In this case, the mirror Poincaré inequality translates into the Brascamp–Lieb inequality which automatically holds when V is twice-differentiable and *strictly* convex. In turn, it readily implies exponential convergence of the Newton Langevin diffusion (Corollary 8.4.1) and can be used for approximate sampling even when the second derivative of V vanishes (Corollary 8.4.2). Strikingly, the rate of convergence *has no dependence on π or on the dimension d* and, in particular, is robust to cases where $\nabla^2 V$ is arbitrarily close to zero. This *scale-invariant* convergence parallels that of Newton's method in convex optimization.

This invariance property is useful for approximately sampling from the uniform distribution over a convex body $\mathcal{C}$, which has been well-studied in the computer science literature [FKP94; KLS95; LV07]. By taking the target distribution

$\pi \propto \exp(-\beta V)$, where V is any strictly convex *barrier function*, and β, the inverse temperature parameter, is taken to be small (depending on the target accuracy), we can use the Newton Langevin diffusion, much in the spirit of interior point methods (as promoted by [LLV20]), to output a sample which is approximately uniformly distributed on $\mathcal{C}$; see Corollary 8.4.3.

Throughout this chapter, we work exclusively in the setting of continuous-time diffusions such as (LD), and we leave the question of obtaining discretization error bounds to §9.

Related work. The discretized Langevin algorithm, and the Metropolis–Hastings adjusted version, have been well-studied when used to sample from strongly log-concave distributions, or distributions satisfying a log-Sobolev inequality [Dal17b; DM17; CB18; DK19; DM19; Dwi+19; VW19; Che+20c; Mou+22], see also §3. Moreover, various ways of adapting Langevin diffusion to sample from bounded domains have been proposed [BEL18; Hsi+18; Zha+20]; in particular, [Zha+20] studied the discretized mirror Langevin diffusion. Finally, we note that while our analysis and methods are inspired by the optimization perspective on sampling, it connects to a more traditional analysis based on coupling stochastic processes. Quantitative analysis of the continuous Langevin diffusion process associated to SDE (LD) has been performed with Poincaré and log-Sobolev inequalities [BGG12; BGL14; VW19], and with couplings of stochastic processes [CL89; Ebe16].

Notation. The Euclidean norm over $\mathbb{R}^d$ is denoted by $\|\cdot\|$. Throughout, we simply write $\int g$ to denote the integral with respect to the Lebesgue measure: $\int g(x)\,\mathrm{d}x$. When the integral is with respect to a different measure μ, we explicitly write $\int g\,\mathrm{d}\mu$. The expectation and variance of $g(X)$ when $X \sim \mu$ are respectively denoted $\mathbb{E}_\mu g = \int g\,\mathrm{d}\mu$ and $\mathrm{var}_\mu g \coloneqq \int (g - \mathbb{E}_\mu g)^2\,\mathrm{d}\mu$. When clear from context, we sometimes abuse notation by identifying a measure μ with its Lebesgue density.

8.2 Mirror Langevin diffusions

Before introducing mirror Langevin diffusions, our main objects of interest, we provide some intuition by drawing a parallel with convex optimization.

8.2.1 Gradient flows, mirror flows, and Newton's method

We briefly recall some background on gradient flows and mirror flows; we refer readers to the monograph [Bub15] for the convergence analysis of the corresponding discrete-time algorithms.

Suppose we want to minimize a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$. The *gradient flow* of f is the curve $(x_t)_{t\geq 0}$ on $\mathbb{R}^d$ solving $\dot{x}_t = -\nabla f(x_t)$. A suitable

time discretization of this curve yields the well-known *gradient descent* (GD).

Although the gradient flow typically works well for optimization over Euclidean spaces, it may suffer from poor dimension scaling in more general cases such as Banach space optimization; a notable example is the case when f is defined over the probability simplex equipped with the ℓ_1 norm. This observation led Nemirovski and Yudin [NY83] to introduce the *mirror flow*, which is defined as follows. Let $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a *mirror map*, that is a strictly convex twice continuously differentiable function of *Legendre type*[1]. The mirror flow $(x_t)_{t\geq 0}$ satisfies $\partial_t \nabla\phi(x_t) = -\nabla f(x_t)$, or equivalently, $\dot{x}_t = -[\nabla^2\phi(x_t)]^{-1}\nabla f(x_t)$. The corresponding discrete-time algorithms, called *mirror descent* (MD) algorithms, have been successfully employed in varied tasks of machine learning [Bub15] and online optimization [BC12] where the entropic mirror map plays an important role. In this work, we are primarily concerned with the following choices for the mirror map:

1. When $\phi = \|\cdot\|^2/2$, then the mirror flow reduces to the gradient flow.
2. Taking $\phi = f$ and the discretization $x_{k+1} = x_k - h_k\,[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$ yields another popular optimization algorithm known as (damped) *Newton's method*. Newton's method has the important property of being invariant under affine transformations of the problem, and its local convergence is known to be much faster than that of GD; see [Bub15, §5.3].

■ 8.2.2 Mirror Langevin diffusions

We now introduce the *mirror Langevin diffusion* (MLD) of [Zha+20]. Just as LD corresponds to the gradient flow, the MLD is the sampling analogue of the mirror flow. To describe it, let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a mirror map as in the previous section. Then, the mirror Langevin diffusion satisfies the SDE

$$X_t = \nabla\phi^\star(Y_t)\,, \qquad \mathrm{d}Y_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,[\nabla^2\phi(X_t)]^{1/2}\,\mathrm{d}B_t\,, \tag{MLD}$$

where $\phi^\star$ denotes the convex conjugate of ϕ [BL06, §3.3]. In particular, if we choose the mirror map ϕ to equal the potential V, then we arrive at a sampling analogue of *Newton's method*, which we call the *Newton Langevin diffusion* (NLD),

$$X_t = \nabla V^\star(Y_t)\,, \qquad \mathrm{d}Y_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,[\nabla^2 V(X_t)]^{1/2}\,\mathrm{d}B_t\,. \tag{NLD}$$

From our intuition gained from optimization, we expect that NLD has special properties, such as affine invariance and faster convergence. We validate this

[1] This ensures that $\nabla\phi$ is invertible, c.f. [Roc97, §26].

intuition in Corollary 8.4.1 below by showing that, provided π is strictly log-concave, the NLD converges to stationarity exponentially fast, with no dependence on π. This should be contrasted with the vanilla Langevin diffusion (LD), for which the convergence rate depends on the Poincaré constant of π, as we discuss in the next section.

We now compare MLD and NLD with similar sampling algorithms proposed in the literature inspired by mirror descent and Newton's method.

Mirrored Langevin dynamics. A variant of MLD, called "mirrored Langevin dynamics", was introduced in [Hsi+18]. The mirrored Langevin dynamics is motivated by constrained sampling and corresponds to the vanilla Langevin algorithm applied to the new target measure $(\nabla\phi)_{\#}\pi$. In contrast, MLD can be understood as a Riemannian diffusion w.r.t. the Riemannian metric induced by the mirror map ϕ. Thus, the motivations and properties of the two algorithms are different, and we refer to [Zha+20] for further comparison of the two algorithms.

Quasi-Newton diffusion. The paper [Sim+16] proposes a quasi-Newton sampling algorithm, based on L-BFGS, which is partly motivated by the desire to avoid computation of the third derivative $\nabla^3 V$ while implementing the Newton Langevin diffusion. We remark, however, that the form of NLD employed above, which treats V as a mirror map, does not in fact require the computation of $\nabla^3 V$, and thus can be implemented practically; see §8.7.1. Moreover, since we analyze the full NLD, rather than a quasi-Newton implementation, we are able to give a clean convergence result.

Information Newton's flow. Inspired by the perspective of [JKO98], which views the Langevin diffusion as a gradient flow in the Wasserstein space of probability measures, the paper [WL20] proposes an approach termed "information Newton's flow" that applies Newton's method directly on the space of probability measures equipped with either the Fisher–Rao or the Wasserstein metric. However, unlike LD and NLD that both operate at the level of particles, information Newton's flow faces significant challenges at the level of both implementation and analysis.

8.3 Convergence analysis

8.3.1 Convergence of gradient flows and mirror flows

We provide a brief reminder about the convergence analysis of gradient flows and mirror flows defined in §8.2.1 to provide intuition for the next section. Throughout, let f be a differentiable function with minimizer x^*.

Consider first the gradient flow for f: $\dot{x}_t = -\nabla f(x_t)$. From straightforward computation, $\partial_t[f(x_t) - f(x^*)] = -\|\nabla f(x_t)\|^2$. From this identity, it is natural

to assume a *Polyak–Łojasiewicz* (PL) *inequality*, which is well-known in the optimization literature [KNS16] and can be employed even when f is not convex. Indeed, if there exists a constant $C_{\mathsf{PL}} > 0$ with

$$f(x) - f(x^*) \le \frac{C_{\mathsf{PL}}}{2} \|\nabla f(x)\|^2 \qquad \forall\, x \in \mathbb{R}^d\,, \tag{PL}$$

then $\partial_t[f(x_t) - f(x^*)] \le -\frac{2}{C_{\mathsf{PL}}}[f(x_t) - f(x^*)]$. Together with Grönwall's inequality, it readily yields exponentially fast convergence in objective value: $f(x_t) \le f(x_0)\exp(-2t/C_{\mathsf{PL}})$.

A similar analysis may be carried out for the mirror flow. Fix a mirror map ϕ and consider the mirror flow: $\dot{x}_t = -[\nabla^2\phi(x_t)]^{-1}\nabla f(x_t)$. Similarly, it holds that $\partial_t[f(x_t) - f(x^*)] = -\langle \nabla f(x_t), [\nabla^2\phi(x_t)]^{-1}\nabla f(x_t)\rangle$. Therefore, the analogue of (PL) which guarantees exponential decay in the objective value is the following inequality, which we call a *mirror PL inequality*:

$$f(x) - f(x^*) \le \frac{C_{\mathsf{MPL}}}{2} \langle \nabla f(x), [\nabla^2\phi(x)]^{-1}\nabla f(x)\rangle \qquad \forall\, x \in \mathbb{R}^d\,. \tag{MPL}$$

Next, we describe analogues of (PL) and (MPL) that guarantee convergence of LD and MLD.

■ 8.3.2 Convergence of mirror Langevin diffusions

The above analysis employs the objective function f to measure the progress of both the gradient and mirror flows. While this is the most natural choice, our approach below crucially relies on measuring progress via a *different functional* F. What should we use as F? To answer this question, we first consider the simpler case of the vanilla Langevin diffusion (LD), which is a special case of MLD when the mirror map is $\phi = \|\cdot\|^2/2$. We keep this discussion informal and postpone rigorous arguments to §8.5.

Since the work of [JKO98], it has been known that the marginal distribution μ_t at time $t \ge 0$ of LD evolves according to the *gradient flow* of the KL divergence $\mathsf{KL}(\cdot\,\|\,\pi)$ with respect to the 2-Wasserstein distance W_2; we refer readers to [San17] for an overview of this work, and to [AGS08; Vil09b] for comprehensive treatments. Therefore, the most natural choice for F is, as in §8.3.1, the objective function $\mathsf{KL}(\cdot\,\|\,\pi)$ itself. Following this approach, one can compute [Vil03, §9.1.5]

$$\partial_t\,\mathsf{KL}(\mu_t\,\|\,\pi) = -\int \Big\|\nabla \ln \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\Big\|^2 \mathrm{d}\mu_t = -4\int \Big\|\nabla\sqrt{\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}}\Big\|^2 \mathrm{d}\pi\,.$$

In this setup, the role of the PL inequality (PL) is played by a *log-Sobolev inequality* of the form

$$\mathrm{ent}_\pi(g^2) := \int g^2 \ln(g^2)\,\mathrm{d}\pi - \Big(\int g^2\,\mathrm{d}\pi\Big)\ln\Big(\int g^2\,\mathrm{d}\pi\Big) \le 2C_{\mathsf{LSI}}\int \|\nabla g\|^2\,\mathrm{d}\pi\,. \tag{LSI}$$

When $g = \sqrt{\mathrm{d}\mu_t/\mathrm{d}\pi}$, (LSI) reads $\mathsf{KL}(\mu_t \,\|\, \pi) \le 2C_{\mathsf{LSI}} \int \big\|\nabla \sqrt{\mathrm{d}\mu_t/\mathrm{d}\pi}\big\|^2 \,\mathrm{d}\pi$, which implies exponentially fast convergence: $\mathsf{KL}(\mu_t \,\|\, \pi) \le \mathsf{KL}(\mu_0 \,\|\, \pi) \exp(-2t/C_{\mathsf{LSI}})$ by Grönwall's inequality.

A disadvantage of this approach, however, is that the log-Sobolev inequality (LSI) does not hold for any log-concave measure π, or it may hold with a poor constant C_{LSI}. For example, it is known that the log-Sobolev constant of an isotropic log-concave distribution must in general depend on the diameter of its support [LV18b]. In contrast, we work below with a *Poincaré inequality*, which is conjecturally satisfied by such distributions with a *universal constant* [KLS95].

Motivated by [BCG08; CG09], we instead consider the *chi-squared divergence*

$$F(\mu) = \chi^2(\mu \,\|\, \pi) \coloneqq \mathrm{var}_\pi \frac{\mathrm{d}\mu}{\mathrm{d}\pi} = \int \Big(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\Big)^2 \mathrm{d}\pi - 1\,, \qquad \text{if } \mu \ll \pi\,,$$

and $F(\mu) = \infty$ otherwise. It is well-known that the law $(\mu_t)_{t\ge 0}$ of LD satisfies the Fokker–Planck equation in the weak sense [KS91, §5.7]:

$$\partial_t \mu_t = \mathrm{div}\big(\mu_t \,\nabla \ln \frac{\mu_t}{\pi}\big)\,.$$

Using this, we can compute the derivative of the chi-squared divergence:

$$\begin{aligned}\frac{1}{2}\,\partial_t F(\mu_t) &= \int \frac{\mu_t}{\pi}\,\partial_t \mu_t = \int \frac{\mu_t}{\pi}\,\mathrm{div}\big(\mu_t \nabla \ln \frac{\mu_t}{\pi}\big) = -\int \big\langle \nabla \ln \frac{\mu_t}{\pi}, \nabla \frac{\mu_t}{\pi}\big\rangle\,\mu_t \\ &= -\int \big\|\nabla \frac{\mu_t}{\pi}\big\|^2\,\pi\,,\end{aligned}$$

and exponential convergence of the chi-squared divergence follows if π satisfies a Poincaré inequality:

$$\mathrm{var}_\pi\, g \le C_{\mathsf{P}}\, \mathbb{E}_\pi[\|\nabla g\|^2] \qquad \text{for all locally Lipschitz } g \in L^2(\pi)\,. \tag{P}$$

Thus, when using the chi-squared divergence to track progress, the role of the PL inequality is played by a Poincaré inequality. As we discuss in §8.4.1 and §8.4.3 below, the Poincaré inequality is significantly weaker than (LSI).

A similar analysis may be carried out for MLD using an appropriate variation of Poincaré inequalities.

Definition 8.3.1 (Mirror Poincaré inequality)**.** *Given a mirror map ϕ, we say that the distribution π satisfies a* mirror Poincaré inequality *with constant C_{MP} if*

$$\mathrm{var}_\pi\, g \le C_{\mathsf{MP}}\, \mathbb{E}_\pi \langle \nabla g, (\nabla^2 \phi)^{-1}\, \nabla g\rangle \qquad \textit{for all locally Lipschitz } g \in L^2(\pi)\,. \tag{MP}$$

When $\phi = \|\cdot\|^2/2$, (MP) is simply called a Poincaré inequality *and the smallest C_{MP} for which the inequality holds is the* Poincaré constant *of π, denoted C_{P}.*

Using a similar argument as the one above, we show exponential convergence of MLD in chi-squared divergence $\chi^2(\cdot \,\|\, \pi)$ under (MP). Together with standard comparison inequalities between information divergences [Tsy09, §2.4], it implies exponential convergence in a variety of commonly used divergences, including the total variation (TV) distance $\|\cdot - \pi\|_{\mathsf{TV}}$, the Hellinger distance $H(\cdot, \pi)$, and the KL divergence $\mathsf{KL}(\cdot \,\|\, \pi)$.

Theorem 8.3.2. *For each $t \geq 0$, let μ_t be the marginal distribution of MLD with target distribution π at time t. Then if π satisfies the mirror Poincaré inequality (MP) with constant C_{MP}, it holds*

$$2\,\|\mu_t - \pi\|_{\mathsf{TV}}^2,\ H^2(\mu_t, \pi),\ \mathsf{KL}(\mu_t \| \pi),\ \chi^2(\mu_t \| \pi) \leq \exp\bigl(-\frac{2t}{C_{\mathsf{MP}}}\bigr)\,\chi^2(\mu_0 \| \pi)\,, \quad \forall\, t \geq 0\,.$$

We give two proofs of this result in §8.5.

Recall that LD can be understood as a gradient flow for the KL divergence on the 2-Wasserstein space. In light of this interpretation, the above bound for the KL divergence yields a convergence rate *in objective value*, and it is natural to wonder whether a similar rate holds for the iterates themselves in terms of 2-Wasserstein distance. From the works [Din15; Led18; Liu20], it is known that a Poincaré inequality (P) implies the transport inequality

$$W_2^2(\mu, \pi) \leq 2C_{\mathsf{P}}\,\chi^2(\mu \,\|\, \pi)\,, \qquad \forall \mu \ll \pi\,. \tag{8.1}$$

A proof of a weaker version of this inequality is also given in [Che+20e].

The inequality (8.1) implies that if π has a finite Poincaré constant C_{P} then Theorem 8.3.2 also yields exponential convergence in Wasserstein distance. In the rest of the paper, we write this as

$$\frac{1}{2C_{\mathsf{P}}}\,W_2^2(\mu_t, \pi) \leq \exp\bigl(-\frac{2t}{C_{\mathsf{MP}}}\bigr)\,\chi^2(\mu_0 \,\|\, \pi)\,,$$

for *any* target measure π that satisfies a mirror Poincaré inequality, with the convention that $C_{\mathsf{P}} = \infty$ when π fails to satisfy a Poincaré inequality. In this case, the above inequality is simply vacuous.

8.4 Applications

We specialize Theorem 8.3.2 to the following important applications.

8.4.1 Newton Langevin diffusion

For NLD, we assume that V is strictly convex and twice continuously differentiable; take $\phi = V$. In this case, the mirror Poincaré inequality (MP) reduces to the

Brascamp–Lieb inequality, which is known to hold with constant $C_{\mathsf{MP}} = 1$ for any strictly log-concave distribution π [BL76; BL00; Gen08]. It yields the following remarkable result where the exponential contraction rate has no dependence on π nor on the dimension d.

Corollary 8.4.1. *Suppose that V is strictly convex and twice continuously differentiable. Then, the law $(\mu_t)_{t\geq 0}$ of* NLD *satisfies*

$$2\,\|\mu_t - \pi\|_{\mathsf{TV}}^2,\ H^2(\mu_t,\pi),\ \mathsf{KL}(\mu_t \parallel \pi),\ \chi^2(\mu_t \parallel \mu),\ \frac{1}{2C_{\mathsf{P}}}\,W_2^2(\mu_t,\pi) \leq \exp(-2t)\,\chi^2(\mu_0 \parallel \pi)\,.$$

If π is log-concave, then it satisfies a Poincaré inequality [AB15] so that the result in Wasserstein distance holds. In fact, contingent on the *Kannan–Lovász–Simonovitz* (KLS) conjecture [KLS95], the Poincaré constant of any log-concave distribution π is upper bounded by a dimension-free constant times the largest eigenvalue of the covariance matrix of π.

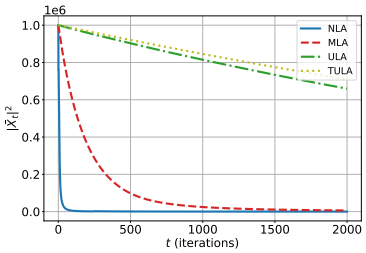


Figure 8.2: Approximately sampling from $\pi \propto \exp(-\|\cdot\|)$ by sampling from $\pi_\beta \propto \exp(-\|\cdot\| - \beta\,\|\cdot - \mathbf{1}\|^2)$ ($\beta = .0005$). Algorithms are initialized at a random X_0 with $\|X_0\| = 1000$. The plot shows the squared distance of the running means to 0.

At this point, one may wonder, under the same assumptions as the Brascamp–Lieb inequality, whether a mirror version of the log-Sobolev inequality (LSI) holds. This question was answered negatively in [BL00], reinforcing our use of the chi-squared divergence as a surrogate for the KL divergence.

If the potential V is convex, but degenerate (i.e., not strictly convex) we cannot use NLD directly with π as the target distribution. Instead, we perturb π slightly to a new measure π_β, which is strongly log-concave, and for which we can use NLD. Crucially, due to the scale invariance of NLD, the time it takes for NLD to mix does not depend on β, the parameter which governs the approximation error.

Corollary 8.4.2. *Fix a target accuracy $\varepsilon > 0$. Suppose $\pi = \exp(-V)$ is log-concave and set $\pi_\beta \propto \exp(-V - \beta\,\|\cdot\|^2)$, where $\beta \leq \varepsilon^2/(2\int \|\cdot\|^2\,\mathrm{d}\pi)$. Then, the law $(\mu_t)_{t\geq 0}$ of* NLD *with target distribution π_β satisfies $\|\mu_t - \pi\|_{\mathsf{TV}} \leq \varepsilon$ by time $t = \frac{1}{2}\ln[2\chi^2(\mu_0 \parallel \pi_\beta)] + \ln(1/\varepsilon)$.*

Proof. From our assumption, it holds

$$\mathsf{KL}(\pi \parallel \pi_\beta) = \int \ln \frac{\mathrm{d}\pi}{\mathrm{d}\pi_\beta}\,\mathrm{d}\pi = \beta \int \|\cdot\|^2\,\mathrm{d}\pi + \ln \int \exp(-\beta\,\|\cdot\|^2)\,\mathrm{d}\pi$$

$$\leq \beta \int \|\cdot\|^2 \,\mathrm{d}\pi \leq \frac{\varepsilon^2}{2}\,.$$

Moreover, Theorem 8.3.2 with the above choice of t yields $\mathsf{KL}(\mu_t \,\|\, \pi_\beta) \leq \varepsilon^2/2$. To conclude, we use Pinsker's inequality and the triangle inequality for $\|\cdot\|_{\mathsf{TV}}$. □

Convergence guarantees for other cases where ϕ is only a *proxy* for V are presented in §8.6.1.

■ 8.4.2 Sampling from the uniform distribution on a convex body

Next, we consider an application of NLD to the problem of sampling from the uniform distribution π on a convex body $\mathcal{C}$. A natural method of outputting an approximate sample from π is to take a strictly convex function $\widetilde{V} : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ such that $\operatorname{dom} \widetilde{V} = \mathcal{C}$ and $\widetilde{V}(x) \to \infty$ as $x \to \partial\mathcal{C}$, and to run NLD with target distribution $\pi_\beta \propto \exp(-\beta \widetilde{V})$, where the inverse temperature β is taken to be small (so that $\pi_\beta \approx \pi$). The function $\widetilde{V}$ is known as a *barrier function*.

Although we can take any choice of barrier function $\widetilde{V}$, we obtain a clean theoretical result if we assume that $\widetilde{V}$ is ν^{-1}-exp-concave, that is, the mapping $\exp(-\nu^{-1}\widetilde{V})$ is concave. Interestingly, this assumption further deepens the rich analogy between sampling and optimization, since such barriers are widely studied in the optimization literature. There, the property of exp-concavity is typically paired with the property of *self-concordance*, and barrier functions satisfying these two properties are a cornerstone of the theory of *interior point algorithms* (see [Bub15, §5.3] and [Nes18, §4]).

We now formulate our sampling result. In our continuous framework, it does not require self-concordance of the barrier function.

Corollary 8.4.3. *Fix a target accuracy $\varepsilon > 0$. Let π be the uniform distribution over a convex body $\mathcal{C}$ and let $\widetilde{V}$ be a ν^{-1}-exp-concave barrier for $\mathcal{C}$. Then, the law $(\mu_t)_{t\geq 0}$ of NLD with target density $\pi_\beta \propto \exp(-\beta\widetilde{V})$ for $\beta \leq \varepsilon^2/(2\nu)$ satisfies $\|\mu_t - \pi\|_{\mathsf{TV}} \leq \varepsilon$ by time $t = \frac{1}{2}\ln[2\chi^2(\mu_0 \,\|\, \pi_\beta)] + \ln(1/\varepsilon)$.*

Proof. Lemma 8.6.3 in §8.6.2 ensures that $\mathsf{KL}(\pi_\beta \,\|\, \pi) \leq \varepsilon^2/2$. We conclude as in the proof of Corollary 8.4.2, by using Theorem 8.3.2, Pinsker's inequality, and the triangle inequality for $\|\cdot\|_{\mathsf{TV}}$. □

We demonstrate the efficacy of NLD in a simple simulation: sampling uniformly from the ill-conditioned rectangle $[-a, a] \times [-1, 1]$ with $a = 0.01$ (Figure 8.3). We compare NLA with the projected Langevin algorithm (PLA) [BEL18], both with 200 iterations and $h = 10^{-4}$. For NLA, we take $\widetilde{V}(x) = -\ln(1 - x_1^2) - \ln(a^2 - x_2^2)$ and $\beta = 10^{-4}$.

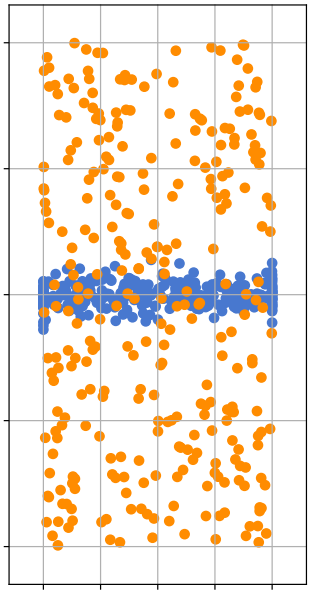


Figure 8.3: Uniform sampling from the set $[-0.01, 0.01] \times [-1, 1]$: PLA (blue) vs. NLA (orange). See §8.7.3.

■ 8.4.3 Langevin diffusion under a Poincaré inequality

We conclude this section by giving some implications of Theorem 8.3.2 to the Langevin diffusion (LD) when $\phi = \|\cdot\|^2/2$. In this case, the mirror Poincaré inequality (MP) reduces to the classical Poincaré inequality (P) as in §8.3.2.

Corollary 8.4.4. *Suppose that π satisfies a Poincaré inequality (P) with constant $C_{\mathsf{P}} > 0$. Then, the law $(\mu_t)_{t\geq 0}$ of the Langevin diffusion (LD) satisfies*

$$2\,\|\mu_t - \pi\|_{\mathsf{TV}}^2,\ H^2(\mu_t, \pi),\ \mathsf{KL}(\mu_t \,\|\, \pi),\ \chi^2(\mu_t \,\|\, \mu),\ \frac{1}{2C_{\mathsf{P}}}\, W_2^2(\mu_t, \pi)$$
$$\leq \exp\Big(-\frac{2t}{C_{\mathsf{P}}}\Big)\, \chi^2(\mu_0 \,\|\, \pi)\,.$$

The convergence in TV distance recovers results of [Dal17b; DM17]. Bounds for the stronger error metric $\chi^2(\cdot\|\pi)$ have appeared explicitly in [CLL19; VW19] and is implicit in the work of [BCG08; CG09] on which the TV bound of [DM17] is based. Discretization guarantees under a Poincaré inequality and (weak) smoothness were obtained in §3.

Moreover, it is classical that if π satisfies a log-Sobolev inequality (LSI) with constant C_{LSI} then it has Poincaré constant $C_{\mathsf{P}} \leq C_{\mathsf{LSI}}$ (Lemma 2.2.8). Thus, the

choice of the chi-squared divergence as a surrogate for the KL divergence when tracking progress indeed requires weaker assumptions on π.

8.5 Proof of the main convergence result

The law $(\mu_t)_{t\geq 0}$ of MLD satisfies the Fokker–Planck equation

$$\partial_t \mu_t = \operatorname{div}\big(\mu_t \, (\nabla^2 \phi)^{-1} \, \nabla \ln \frac{\mu_t}{\pi}\big) \,. \tag{8.2}$$

A unique solution to this equation, with enough regularity to justify our computations below, exists under fairly benign conditions on ϕ and V, see [LL08, Proposition 6].

As discussed in §8.3.2, it suffices to prove the convergence result in chi-squared divergence. The convergence results for total variation distance, Hellinger distance, and KL divergence follow from the inequalities [Tsy09, §2.4]

$$2 \, \|\mu - \pi\|_{\mathsf{TV}}^2, \;\; H^2(\mu, \pi), \;\; \mathsf{KL}(\mu \parallel \pi) \leq \chi^2(\mu \parallel \pi) \,, \qquad \forall \mu \ll \pi \,,$$

while the convergence in Wasserstein distance follows from (8.1).

Proof of Theorem 8.3.2. Using the Fokker–Planck equation (8.2), we compute

$$\begin{aligned} \partial_t \chi^2(\mu_t \parallel \pi) &= \partial_t \int \frac{\mu_t^2}{\pi} = 2 \int \frac{\mu_t}{\pi} \, \partial_t \mu_t = 2 \int \frac{\mu_t}{\pi} \, \operatorname{div}\big(\mu_t \, (\nabla^2 \phi)^{-1} \, \nabla \ln \frac{\mu_t}{\pi}\big) \\ &= -2 \int \big\langle \nabla \frac{\mu_t}{\pi}, (\nabla^2 \phi)^{-1} \, \nabla \ln \frac{\mu_t}{\pi} \big\rangle \, \mu_t = -2 \int \big\langle \nabla \frac{\mu_t}{\pi}, (\nabla^2 \phi)^{-1} \, \nabla \frac{\mu_t}{\pi} \big\rangle \, \pi \,. \end{aligned}$$

The mirror Poincaré inequality (MP) implies that this quantity is upper bounded by $-2 C_{\mathsf{MP}}^{-1} \, \chi^2(\mu_t \parallel \pi)$, which completes the proof via Grönwall's inequality. □

We may reinterpret this proof within Markov semigroup theory.

Proof of Theorem 8.3.2 from a Markov semigroup perspective. In this proof, we denote the semigroup of MLD by $(P_t)_{t\geq 0}$; we refer readers to [BGL14; Han16] for background on Markov semigroup theory. The Dirichlet form $\mathcal{E}$ is given by

$$\mathcal{E}(f, g) = \int \langle \nabla f, (\nabla^2 \phi)^{-1} \, \nabla g \rangle \, \mathrm{d}\pi \,.$$

Since it is a self-adjoint semigroup, we get for all $f \in L^2(\pi)$,

$$\int P_t\big(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}\big) \, f \, \mathrm{d}\pi = \int \big(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}\big) \, P_t f \, \mathrm{d}\pi = \int P_t f \, \mathrm{d}\mu_0 = \int f \, \mathrm{d}\mu_t = \int \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} \, f \, \mathrm{d}\pi \,,$$

so that

$$P_t\big(\frac{\mu_0}{\pi}\big) = \frac{\mu_t}{\pi}\,.$$

Therefore,

$$\chi^2(\mu_t \,\|\, \pi) \coloneqq \mathrm{var}_\pi\big(\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\big) = \mathrm{var}_\pi\, P_t\big(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}\big)\,.$$

Then, using a classical result of Markov semigroup theory (see for instance [CG09, Theorem 2.1] or [BGL14, Theorem 4.2.5]),

$$\chi^2(\mu_t \,\|\, \pi) = \mathrm{var}_\pi\, P_t\big(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}\big) \le \exp\big(-\frac{2t}{C}\big)\,\mathrm{var}_\pi\big(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}\big) = \exp\big(-\frac{2t}{C}\big)\,\chi^2(\mu_0 \,\|\, \pi)$$

if and only if the semigroup $(P_t)_{t\ge 0}$ satisfies

$$\mathrm{var}_\pi(f) \le C\,\mathcal{E}(g,g)\,, \qquad \text{for all} \quad g \in D(\mathcal{E})\,, \tag{8.3}$$

where $\mathcal{E}$ is the Dirichlet form of $(P_t)_{t\ge 0}$ with domain $D(\mathcal{E})$. To conclude the proof, it suffices to note that (8.3) is precisely our assumption (MP) with $C = C_{\mathsf{MP}}$. □

8.6 Auxiliary results

8.6.1 Additional choices for the mirror map

We extend our results to other choices of the mirror map ϕ that serve as proxies for V and that also lead to exponential convergence of MLD.

The first result below is useful in situations when there exists a strictly convex mirror map ϕ such $\nabla\phi$ is easier to invert than ∇V. It ensures exponential ergodicity of (MLD) when $\nabla^2 V$ dominates $\nabla^2\phi$ in the sense of the Loewner order.

Corollary 8.6.1. *Suppose that π is strictly log-concave and that $\nabla^2\phi \preceq C\,\nabla^2 V$, where $\preceq$ denotes the Loewner order. Then, the law $(\mu_t)_{t\ge 0}$ of MLD satisfies*

$$\begin{aligned} &2\,\|\mu_t - \pi\|_{\mathsf{TV}}^2,\ H^2(\mu_t,\pi),\ \mathsf{KL}(\mu_t \,\|\, \pi),\ \chi^2(\mu_t \,\|\, \mu),\ \frac{1}{2C_{\mathsf{P}}} W_2^2(\mu_t,\pi) \\ &\qquad \le \exp\big(-\frac{2t}{C}\big)\,\chi^2(\mu_0 \,\|\, \pi)\,. \end{aligned}$$

Proof. The assumption implies

$$C\,\mathbb{E}_\pi\langle \nabla f, (\nabla^2\phi)^{-1}\,\nabla f\rangle \ge \mathbb{E}_\pi\langle \nabla f, (\nabla^2 V)^{-1}\,\nabla f\rangle \ge \mathrm{var}_\pi\, f\,,$$

where again we apply the Brascamp–Lieb inequality. This verifies (MP) with constant $C_{\mathsf{MP}} = C$. □

Our second result does not require π to be log-concave but only that it is close to a strictly log-concave distribution $\widetilde{\pi}$ in the following sense: the density of π with respect to $\widetilde{\pi}$ is uniformly bounded away from 0 and ∞.

Corollary 8.6.2. *Suppose that $\widetilde{\pi} = \exp(-\widetilde{V})$ is strictly log-concave and suppose that π has density ρ w.r.t. $\widetilde{\pi}$. Let $M := (\sup \rho)/(\inf \rho)$. Then, the law $(\mu_t)_{t\geq 0}$ of* MLD *with mirror map $\phi = \widetilde{V}$ and target density π satisfies*

$$2\,\|\mu_t - \pi\|_{\mathsf{TV}}^2,\ H^2(\mu_t,\pi),\ \mathsf{KL}(\mu_t \parallel \pi),\ \chi^2(\mu_t \parallel \mu),\ \frac{1}{2C_{\mathsf{P}}M} W_2^2(\mu_t,\pi) \leq \exp\bigl(-\frac{2t}{M}\bigr)\,\chi^2(\mu_0 \parallel \pi)\,,$$

where C_{P} is the Poincaré constant of $\widetilde{\pi}$.

Proof. It is standard that the Poincaré inequality (P), and the mirror Poincaré inequality (MP), are stable under bounded perturbations of the measure. It implies that π satisfies a Poincaré inequality with constant $C_{\mathsf{P}}M$, and a mirror Poincaré inequality with constant M. We prove the latter statement for completeness; for the former statement, see [Han16, Problem 3.20].

Observe that

$$\begin{aligned}\int \langle \nabla f, (\nabla^2 \widetilde{V})^{-1}\,\nabla f\rangle \,\mathrm{d}\pi &= \int \langle \nabla f, (\nabla^2 \widetilde{V})^{-1}\,\nabla f\rangle \,\frac{\mathrm{d}\pi}{\mathrm{d}\widetilde{\pi}}\,\mathrm{d}\widetilde{\pi} \\ &\geq (\inf \rho) \int \langle \nabla f, (\nabla^2 \widetilde{V})^{-1}\,\nabla f\rangle \,\mathrm{d}\widetilde{\pi}\end{aligned}$$

and

$$\begin{aligned}\mathrm{var}_{\widetilde{\pi}}\, f &= \inf_{m\in\mathbb{R}} \int |f-m|^2\,\mathrm{d}\widetilde{\pi} = \inf_{m\in\mathbb{R}} \int |f-m|^2\,\frac{\mathrm{d}\widetilde{\pi}}{\mathrm{d}\pi}\,\mathrm{d}\pi \\ &\geq \frac{1}{\sup \rho}\,\inf_{m\in\mathbb{R}} \int |f-m|^2\,\mathrm{d}\pi = \frac{1}{\sup\rho}\,\mathrm{var}_{\pi}\, f\,.\end{aligned}$$

Combining these inequalities with the Brascamp–Lieb inequality for $\widetilde{\pi}$,

$$\int \langle \nabla f, (\nabla^2 \widetilde{V})^{-1}\,\nabla f\rangle \,\mathrm{d}\widetilde{\pi} \geq \mathrm{var}_{\widetilde{\pi}}\, f\,,$$

yields (MP) with constant $C_{\mathsf{MP}} = M$. □

■ 8.6.2 Stability in KL with respect to exp-concave perturbations

The following lemma quantifies the approximation error of replacing π by π_β in §8.4.2 and, more generally provides a simple bound to control the KL divergence between a log-concave distribution and its perturbation by a ν-exp-concave barrier function. Its proof uses crucially displacement convexity of the KL divergence to a log-concave measure [Vil03, §5], and it can be viewed as the sampling analogue of [Nes18, (4.2.17)].

Recall that b is *ν-exp-concave* if the mapping $\exp(-\nu^{-1}b)$ is concave.

Lemma 8.6.3. *Let π be a log-concave distribution on a convex set $\mathcal{K} \subset \mathbb{R}^d$. Fix $\nu > 0$, and let $\widetilde{\pi}$ have density $\exp(-b)$ with respect to π, where $b : \mathcal{K} \to \mathbb{R}$ is ν-exp-concave. Then it holds that*

$$\mathsf{KL}(\widetilde{\pi} \,\|\, \pi) \le \nu \,.$$

Proof. On $\operatorname{int}\mathcal{K}$, we have

$$-\nabla \ln \frac{\mathrm{d}\widetilde{\pi}}{\mathrm{d}\pi} = \nabla b \,. \tag{8.4}$$

The measure π is log-concave, so by displacement convexity of entropy [AGS08, Theorem 9.4.11] and the "above-tangent" formulation of convexity [Vil03, Proposition 5.29], we have

$$0 = \mathsf{KL}(\pi \,\|\, \pi) \ge \mathsf{KL}(\widetilde{\pi} \,\|\, \pi) + \mathbb{E}\big\langle \nabla \ln \frac{\mathrm{d}\widetilde{\pi}}{\mathrm{d}\pi}(\widetilde{X}), X - \widetilde{X} \big\rangle \,,$$

where $(X, \widetilde{X})$ are optimally coupled for π and $\widetilde{\pi}$. If we rearrange this inequality and use the identities in (8.4), we get

$$\mathsf{KL}(\widetilde{\pi} \,\|\, \pi) \le -\,\mathbb{E}\big\langle \nabla \ln \frac{\mathrm{d}\widetilde{\pi}}{\mathrm{d}\pi}(\widetilde{X}), X - \widetilde{X} \big\rangle = \mathbb{E}\langle \nabla b(\widetilde{X}), X - \widetilde{X} \rangle \,. \tag{8.5}$$

We now use the fact that b is ν-exp-concave. To that end, define the convex function

$$\varphi(t) = -\exp\big(-\frac{1}{\nu}\, b(\widetilde{X} + t\,(X - \widetilde{X}))\big) \,, \qquad t \in [0,1] \,.$$

By convexity, we have

$$\varphi'(0) \cdot (1 - 0) \le \varphi(1) - \varphi(0) \le -\varphi(0) = \exp\big(-\frac{1}{\nu}\, b(\widetilde{X})\big) \,.$$

Since

$$\varphi'(0) = \frac{1}{\nu} \exp\big(-\frac{1}{\nu}\, b(\widetilde{X})\big)\, \langle \nabla b(\widetilde{X}), X - \widetilde{X} \rangle\,,$$

the above inequality reads $\langle \nabla b(\widetilde{X}), X - \widetilde{X} \rangle \leq \nu$, which completes the proof together with (8.5). □

Remark 8.6.4. *It is known that given any convex body $\mathcal{C} \subset \mathbb{R}^d$, there exists a standard self-concordant ν^{-1}-exp-concave barrier with $\nu \leq d$ [NN94; BE19; LY21]; see also §10.*

■ 8.7 Numerical experiments

In this section, we gather additional details and figures to support our numerical experiments. First, in §8.7.1, we display the samples for a Gaussian experiment. Then, §8.7.2 gives details of the Bayesian logistic regression experiment displayed in Figure 8.1 and shows the effect of varying step size. §8.7.3 gives details of sampling from an ill-conditioned convex set. Finally, §8.7.4 shows an experiment where we use the NLA and a mirror Langevin algorithm MLA to approximately sample from a degenerate log-concave distribution.

■ 8.7.1 Sampling from a Gaussian distribution

In this section, we examine the numerical performance of the *Newton Langevin algorithm* (NLA), which is given by the following Euler discretization of NLD:

$$\nabla V(X_{k+1}) = (1-h)\, \nabla V(X_k) + \sqrt{2h}\, [\nabla^2 V(X_k)]^{1/2}\, \xi_k\,, \qquad \text{(NLA)}$$

where $(\xi_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. $\mathsf{normal}(0, I_d)$ variables. In cases where ∇V does not have a closed-form inverse, such as the logistic regression case of §8.7.2, we invert it numerically by solving the convex optimization problem $\nabla V^\star(y) = \arg\max_{x\in\mathbb{R}^d} \{\langle x, y\rangle - V(x)\}$.

We focus here on sampling from an ill-conditioned generalized Gaussian distribution on $\mathbb{R}^{100}$ with $V(x) = \langle x, \Sigma^{-1} x\rangle^\gamma/2$ for $\gamma = 3/4$ to demonstrate the scale invariance of NLD established in Corollary 8.4.1.

Figure 8.4 compares the performance of NLA to that of the unadjusted Langevin algorithm (ULA) and of the tamed unadjusted Langevin algorithm (TULA) from [Bro+19]. We run the algorithms 50 times and compute running estimates for the mean and scatter matrix of the family following [ZWG13]. Convergence is measured in terms of squared distance between means and relative squared distance between scatter matrices, $\|\hat{\Sigma} - \Sigma\|^2/\|\Sigma\|^2$. NLA generates samples that

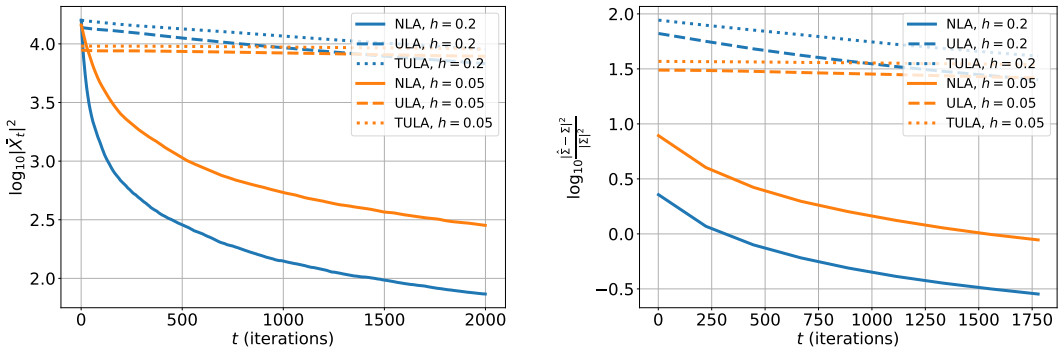


Figure 8.4: $V(x) = \langle x, \Sigma^{-1}x\rangle^{\frac{3}{4}}/2$, $\Sigma = \mathrm{diag}(1, 2, 3, \ldots, 100)$. Left: absolute squared error of the mean 0. Right: relative squared error for the scatter matrix Σ.

rapidly approximate the true distribution and also displays stability to the choice of the step size.

Next, we repeat the example in Figure 8.4 for the simpler case of the Gaussian distribution ($\gamma = 1$) on $\mathbb{R}^{100}$ with the same scatter matrix $\Sigma = \mathrm{diag}(1, 2, 3, \ldots, 100)$ in Figure 8.5. We again see the superiority of NLA over the ULA and TULA. The additional parameter of TULA (denoted γ in [Bro+19]) is chosen equal to 0.1.

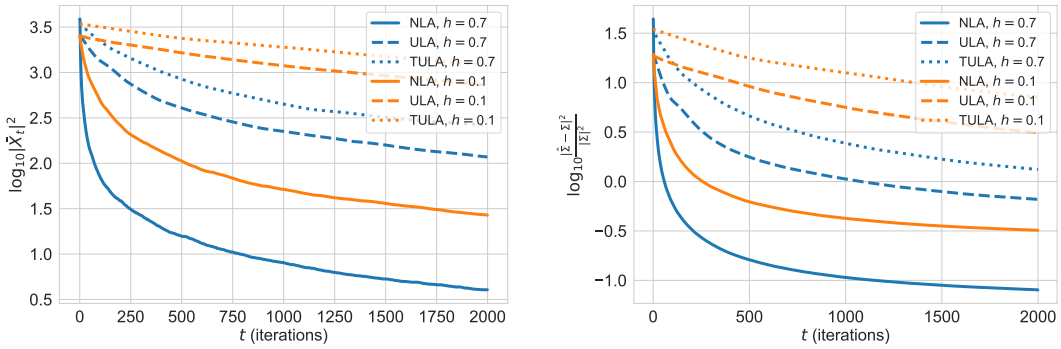


Figure 8.5: We display convergence of the various algorithms for an ill-conditioned Gaussian distribution, with $d = 100$ and $\Sigma = \mathrm{diag}(1, 2, 3, \ldots, 100)$. Left: error is the squared distance from 0. Right: error is the relative distance between scatter matrices. As in the experiment displayed in Figure 8.4, NLA rapidly converges both in terms of location and scale for large step sizes.

We also display some samples from the Gaussian experiment of Figure 8.5 in Figure 8.6. NLA maintains good performance for a wide range of step sizes, while ULA and TULA require a small step size to accurately sample from the target

distribution. In fact, even with a small step size, ULA and TULA often jump to small probability regions, while NLA avoids these regions even for large step sizes.

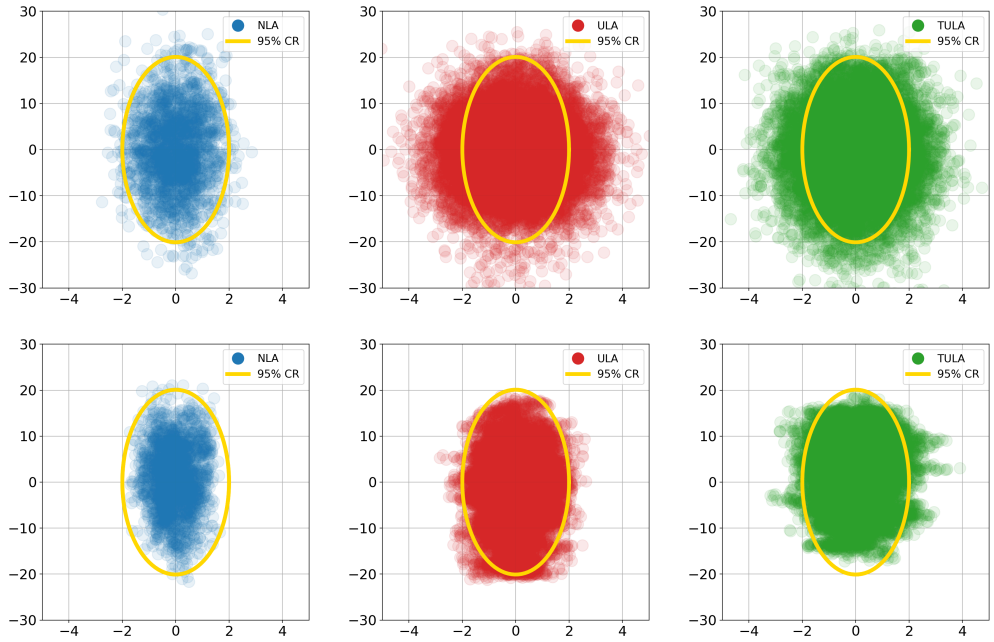


Figure 8.6: Samples from NLA, ULA, and TULA for the ill-conditioned Gaussian example of Figure 8.5, with $\Sigma = \mathrm{diag}(1, 2, 3, \ldots, 100)$. We display the projection onto the first (least spread) and last (most spread) population principal components, along with the projection of a 95% confidence region. Top: the step size for all algorithms is $h = 0.7$. Bottom: the step size for all algorithms is $h = 0.05$.

■ 8.7.2 Bayesian logistic regression

We give details for the two-dimensional Bayesian logistic regression example in Figure 8.1. In the Bayesian logistic regression model, covariates are drawn as $X_i \sim \mathsf{normal}(0, \mathrm{diag}(10, 0.1))$, the response variables are $Y_i \sim \mathsf{Bernoulli}(\mathrm{logit}(\langle \theta, X_i \rangle))$, and the parameters θ have a $\mathsf{normal}(0, 10I_2)$ prior. We consider using NLA to sample from the posterior distribution of θ given the observations (X_i, Y_i), $i = 1, \ldots, n$, which is

$$\pi(\theta) \propto \exp\Big[-\frac{1}{20}\,\|\theta\|^2 + \sum_{i=1}^{n}\big(Y_i\,\langle \theta, X_i\rangle - \ln(1 + \exp\langle \theta, X_i\rangle)\big)\Big]\,,$$

which is strongly log-concave. While the gradient of the potential is invertible, it has no closed-form, and so in our experiments we invert it numerically by

solving $\nabla V^\star(y) = \arg\max_{x \in \mathbb{R}^d} \{\langle x, y\rangle - V(x)\}$ with Newton's method. We find that, with a warm start from the current iterate X_t, it suffices to run Newton's method for a small number of iterations to approximately invert the gradient.

For visualization, we generate 100 samples $X_i \sim \mathsf{normal}(0, \mathrm{diag}(10, 0.1))$ and $Y_i \sim \mathsf{Bernoulli}(\mathrm{logit}(\langle \theta^\star, X_i\rangle))$, where we set $\theta^\star = (1, 1)$.

We display the result for various sampling algorithms in Figure 8.1. All algorithms are implemented with $h = 0.1$ and a burn-in time of 10^4 steps. This example shows the advantage of taking a large step size with NLA in this ill-conditioned model, while ULA and TULA create samples that are overdispersed. In Figure 8.7, we also show the effect of decreasing step size in this example. In this case, we see that ULA and TULA still step into low probability regions or fail to explore the underlying density well. On the other hand, NLA remains constrained in the high probability region.

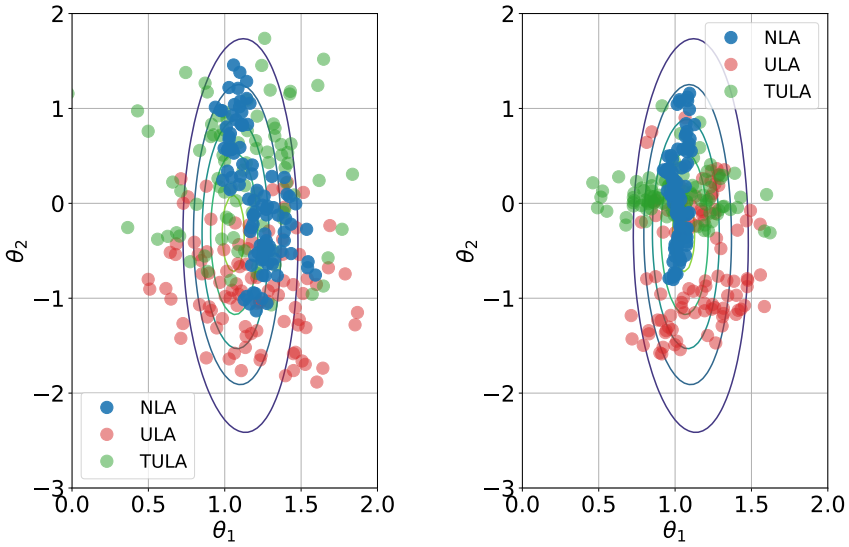


Figure 8.7: Samples from the posterior distribution of a Bayesian logistic regression model using one run of NLA, ULA, and TULA after a burn-in of 10^4. Left: large step size (all algorithms use $h = 0.05$); NLA remains within the high-density contours, while the ULA and TULA take steps into low-density areas. Right: small step size (all algorithms use $h = 0.01$); NLA explores the underlying distribution faster than its competitors.

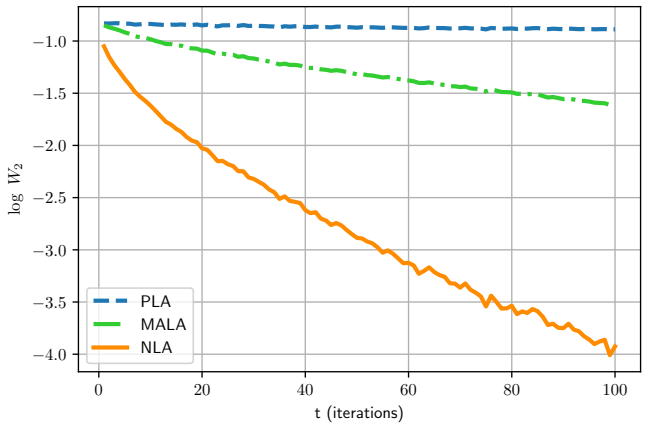


Figure 8.8: W_2 distance (on logarithmic scale) between the uniform distribution on the rectangle $[-0.01, 0.01] \times [-1, 1]$, and samples produced by NLA, PLA, and MALA.

8.7.3 Uniform sampling on a convex body

This section contains details for the simulations in Figure 8.3. We sample from the uniform distribution on the rectangle $[-0.01, 0.01] \times [-1, 1]$ using NLA, PLA, and the Metropolis-adjusted Langevin algorithm (MALA) [Bes+95]. PLA and MALA target the uniform distribution directly. NLA samples from an approximate distribution, given in §8.4.2. The step sizes are chosen as $h = 10^{-5}$ for NLA and PLA and $h = 0.01$ for MALA. The step sizes for PLA and MALA are tuned to allow the algorithm to reach approximate stationarity in the fewest number of iterations. MALA can use a larger step size because it is unbiased (its stationary distribution coincides with the target distribution, due to the Metropolis–Hastings adjustment). On the other hand, samples from PLA tend to cluster around the boundary for larger step sizes, so we use a smaller step size for both PLA (and NLA for fair comparison).

To evaluate the performance of the algorithms, we estimate the 2-Wasserstein distance between the samples drawn by the algorithms and samples drawn from the uniform distribution on the rectangle; see Figure 8.8. We use the Sinkhorn distance ($\varepsilon = 0.01$) as an approximation for the 2-Wasserstein distance [Cut13; ANR17]. Specifically, we sample 1000 points in parallel, using the three algorithms of interest. At each iteration, we also draw 1000 points from the uniform distribution on the rectangle, and we compute the Sinkhorn distance between these points and the samples produced by the algorithms. The convergence estimates are averaged over 30 runs.

8.7.4 Approximate sampling from degenerate log-concave distributions

In this section, we explore further the problem of approximately sampling according to the measure $\pi(x) \propto \exp(-\|x\|)$ in $\mathbb{R}^2$ considered in Figure 8.2. To that end, we use the penalization strategy outlined in §8.4.1 and sample instead from the strongly log-concave measure $\pi_\beta(x) \propto \exp(-\|x\| - \beta \|x - \mathbf{1}\|^2)$ as in Corollary 8.4.2, where $\beta = 0.0005$, using discretizations of either NLD or MLD with a customized mirror map. Here, $\mathbf{1}$ is the vector of all ones, which simulates the effect of not knowing the true mean.

We initialize all algorithms with a random point X_0 with $\|X_0\| = 1000$. The initialization is chosen so that the gradients of the potential at initialization are extremely small. In these circumstances, we expect ULA to mix slowly.

Through this experiment, we demonstrate two empirical observations:

1. Initially, the iterates of NLA converge extremely rapidly to the vicinity of the origin. This suggests that NLA can be useful for initializing other sampling algorithms in highly ill-conditioned settings.

2. However, once the iterates of NLA are near the origin, NLA becomes unstable. Specifically, since the Hessian of the potential degenerates rapidly near 0, the iterates of NLA occasionally make large jumps away from 0. This is due to the fact that the Hessian of $V(x) = \|x\| + \beta \|x - \mathbf{1}\|^2$ is given by

$$\nabla^2 V(x) = \frac{1}{\|x\|} \left[I_2 - \left(\frac{x}{\|x\|} \right) \left(\frac{x}{\|x\|} \right)^\mathsf{T} \right] + 2\beta \, I_2 \tag{8.6}$$

   which blows up to infinity around $x = 0$. We remark that Newton's method in optimization can also exhibit unstable behavior [CGT00; NP06], so this phenomenon is not unexpected.

   To rectify this behavior, we also consider the Euler discretization of MLD, which we call MLA (see below). We demonstrate that with an appropriate choice of mirror map, the iterates of MLA are stable, yet still enjoy faster convergence than ULA.

Now we proceed to the details of the experiment. We compare four different methods for sampling from this distribution: NLA, ULA, TULA, and the mirror Langevin algorithm (MLA)

$$\nabla \phi(X_{k+1}) = \nabla \phi(X_k) - h \, \nabla V(X_k) + \sqrt{2h} \, [\nabla^2 \phi(X_k)]^{1/2} \, \xi_k \,, \tag{MLA}$$

with mirror map $\phi(x) = \|x\|^{3/2}$ and potential $V(x) = \|x\| + \beta \|x - \mathbf{1}\|^2$. Notice that this mirror map corresponds to that used in the generalized Gaussian case of §8.7.1.

In Figure 8.9, we display the results of the first 1000 iterations of the four algorithms. In this stage of the experiment, we observe rapid convergence of NLA towards the origin (around which the mass is concentrated), and MLA also exhibits faster convergence than ULA and TULA. However, already in Figure 8.9 (Right) we observe the instability of NLA witnessed through large jumps of the iterates.

Next, in Figure 8.10, we treat the samples from the first 1000 iterations as burn-in, and we look at the performance of the next 1000 samples. Here we see that the flexible framework of the more general MLD allows us to design algorithms which can outperform NLA with superior stability in specific scenarios.

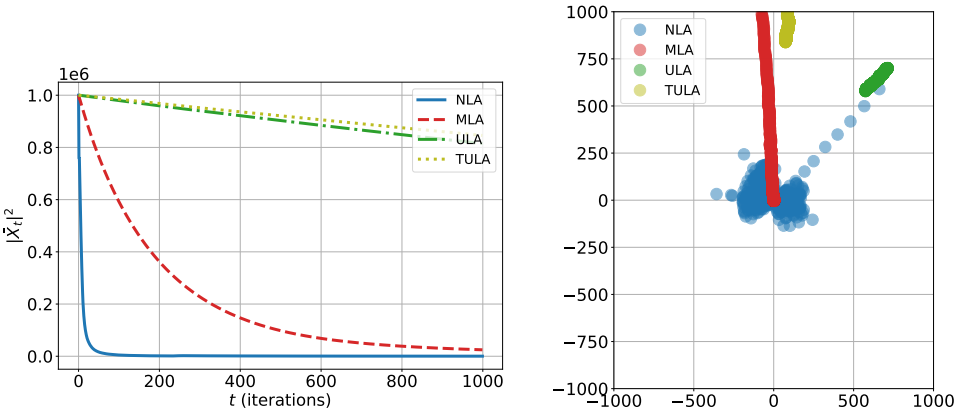


Figure 8.9: First stage of the experiment. Left: We plot the norm of the running mean versus the iteration number for the target measure $\pi_\beta(x) \propto \exp(-\|x\| - 0.0005\,\|x - \mathbf{1}\|^2)$. Right: We display the corresponding samples.

Recall that the Hessian of the potential V is given in (8.6) while the potential of the mirror map ϕ is given by

$$\nabla^2 \phi(x) = \frac{3}{2\,\|x\|^{1/2}} \left[I_2 - \frac{3}{4} \left(\frac{x}{\|x\|} \right) \left(\frac{x}{\|x\|} \right)^{\mathsf{T}} \right].$$

From these expressions, it can be checked that Corollary 8.6.1 holds with $C \leq 3/(4\sqrt{2\beta})$. On the other hand, the measure π_β satisfies a Poincaré inequality (P) with constant $C_{\mathsf{P}} \leq 1/(2\beta)$. Heuristically, we therefore expect the mixing time of ULA to scale as $O(\beta^{-1})$, and the mixing time of MLA to scale as $O(\beta^{-1/2})$, which provides an explanation for the rates of convergence observed in Figure 8.9. In comparison, the mixing time of NLA is scale-invariant, i.e., $O(1)$, as we demonstrated in Corollary 8.4.1, as witnessed by the initial rapid convergence in Figure 8.9.

As mentioned in our open questions, this points to the intriguing possibility of developing more stable variants of NLA, which would mirror the development of such strategies for Newton's method [CGT00; NP06].

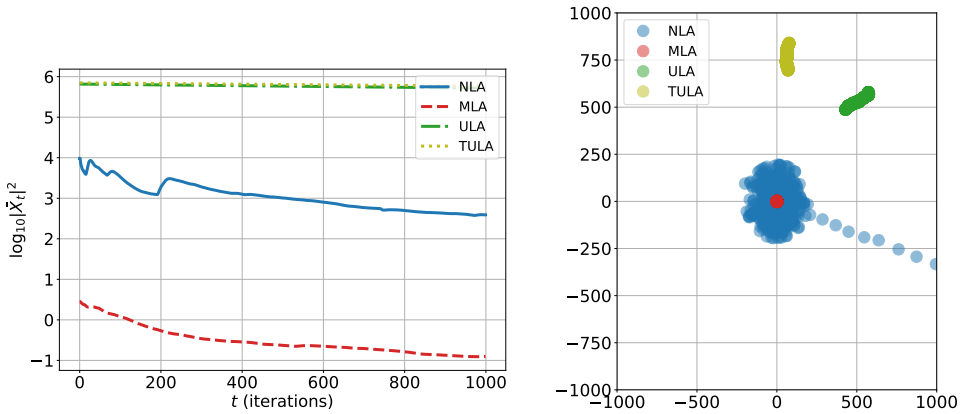


Figure 8.10: Second stage of the experiment. In this stage, we treat the 1000 samples from the first stage of the experiment as burn-in and look at the performance of the next 1000 samples. Left: We plot the logarithm of the norm of the running mean versus iteration. Right: We again display the corresponding samples.

■ 8.8 Conclusion

We conclude this chapter by discussing several intriguing directions for future research. In this chapter, we focused on giving clean convergence results for the continuous-time diffusions MLD and NLD, and the problem of obtaining discretization error bounds is deferred to §9. In discrete time, Newton's method can be unstable, and one uses methods such as damped Newton, Levenburg–Marquardt, or cubic-regularized Newton [CGT00; NP06]; it is an interesting question to develop sampling analogues of these optimization methods. In a different direction, we ask the following question: are there appropriate variants of other popular sampling methods, such as accelerated Langevin [Ma+21] or Hamiltonian Monte Carlo [Nea11], which also enjoy the scale invariance of NLD?

Chapter 9

Discretization analysis of mirror Langevin Monte Carlo

In §8, we focused on the mirror Langevin diffusion in continuous time. In this chapter, we propose a new discretization of the mirror Langevin diffusion and give a crisp proof of its convergence. Our analysis uses relative convexity/smoothness and self-concordance, ideas which originated in convex optimization, together with a new result in optimal transport that generalizes the displacement convexity of the entropy. Unlike prior works, our result both (1) requires much weaker assumptions on the mirror map and the target distribution, and (2) has vanishing bias as the step size tends to zero. In particular, for the task of sampling from a log-concave distribution supported on a compact set, our theoretical results are significantly better than the existing guarantees.

This chapter is based on [AC21], joint with Kwangjun Ahn.

9.1 Introduction

We consider the following canonical sampling problem. Let $V : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a convex function and let π be the density on $\mathbb{R}^d$ which is proportional to $\exp(-V)$. The task is to output a sample which is (approximately) distributed according to π, given query access to the gradients of V.

As in previous chapters, we are motivated by the deep and fruitful connection between sampling and the field of *optimization*, introduced in the seminal work [JKO98]. To describe this connection, we recall the *Langevin diffusion*, which is the solution to the following stochastic differential equation (SDE):

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,. \tag{LD}$$

Under standard assumptions on the potential V, the SDE is well-posed and it converges in distribution, as $t \to \infty$, to its unique stationary distribution π. Thus, once suitably discretized, it yields a popular algorithm for the sampling prob-

lem. The Langevin diffusion is classically studied using techniques from Markov semigroup theory [see, e.g., BGL14; Pav14], but there is a more insightful perspective which views the diffusion (LD) through the lens of optimization [JKO98]. Specifically, if μ_t denotes the law of the process (LD) at time t, then the curve $(\mu_t)_{t\geq 0}$ is the *gradient flow* of the KL divergence $\mathsf{KL}(\cdot \,\|\, \pi)$ in the Wasserstein space of probability measures. This perspective has not only inspired new analyses of Langevin [CB18; Wib18; DMM19; VW19], but has also emboldened the possibility of bringing to bear the extensive toolkit of optimization onto the problem of sampling (see, e.g., §4).

However, the vanilla Langevin diffusion notably fails when the support of the target distribution π is not all of $\mathbb{R}^d$. This task of *constrained sampling*, named in analogy to constrained optimization, arises in applications such as Bayesian matrix factorization [PBJ15], latent Dirichlet allocation [BNJ03], ordinal data models [JA99], and regularized regression [Cel+12]. Despite such a broad range of applications, the constrained sampling problem has proven to be challenging. In particular, most prior works have focused on domain-specific algorithms [GSL92; PP14; LS16], and the first general-purpose algorithms for this task are recent [Bro+17; BEL18].

In this work, we tackle the constrained sampling problem via *mirror Langevin Monte Carlo* (MLMC). MLMC is a discretization of the mirror Langevin diffusion [Hsi+18; Zha+20], which is the sampling analogue of *mirror descent*. Namely, if $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a mirror map, then the mirror Langevin diffusion is the solution to the SDE

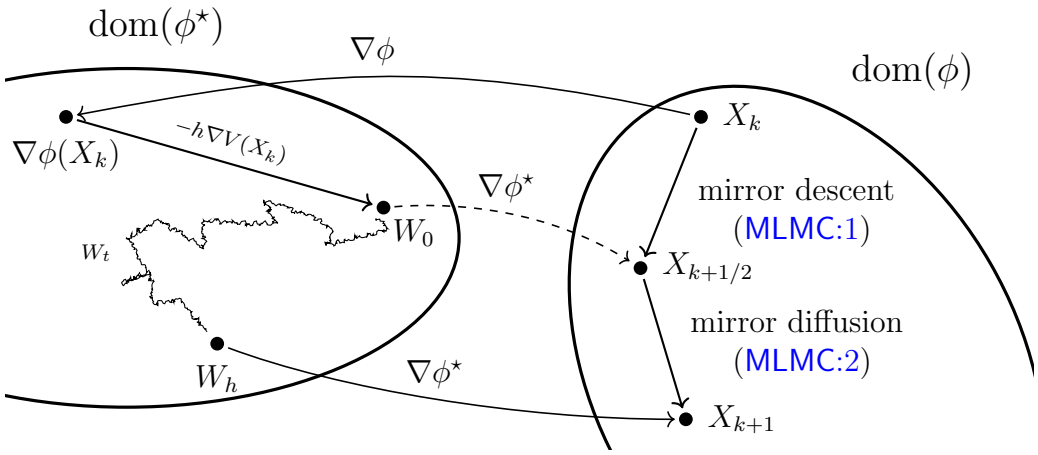
$$X_t = \nabla\phi^\star(Y_t)\,, \qquad \mathrm{d}Y_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,[\nabla^2\phi(X_t)]^{1/2}\,\mathrm{d}B_t\,. \tag{MLD}$$

Technical motivation. Zhang et al. [Zha+20] analyzed an Euler–Maruyama discretization of MLD; see (9.2) for details. The most curious aspect of their result is that their convergence guarantee has a bias term that does not vanish even when step size tends to zero and the number of iterations tends to infinity. Moreover, they conjecture that this bias term is unavoidable. This is in contrast to known results for standard Langevin, which raises the main question of this paper:

Can a different discretization of MLD *lead to a vanishing bias?*

Our contributions. We propose a new discretization of the mirror Langevin diffusion, given in (MLMC) and illustrated in Figure 9.1. Our proposed discretization has the same cost as the standard Euler–Maruyama discretization of MLD in terms of the number of queries to the gradient oracle for V. We remark that our scheme for the case $\phi = \frac{\|\cdot\|^2}{2}$ recovers the Langevin Monte Carlo algorithm. The

Figure 9.1: Illustration of mirror Langevin Monte Carlo (MLMC). This illustration is adapted from [Bub15, Figure 4.1].



most important aspect of our result is that the bias of our algorithm vanishes as the step size tends to zero unlike the result by Zhang et al. [Zha+20].

By adapting the analysis of Durmus et al. [DMM19], we provide a clean convergence analysis of our algorithm which theoretically validates our discretization scheme. Notably, our analysis only requires standard assumptions/definitions which are well-studied in optimization. In particular, we establish a stronger link between sampling and optimization without relying on technical assumptions of [Zha+20] (e.g., commutation conditions for Hessians; see (A5) therein).

Moreover, our analysis combines ideas from optimization with the calculus of optimal transport. In particular, we establish a new generalization of a celebrated fact, namely that the entropy functional is *displacement convex* along Wasserstein geodesics, to the setting of Bregman divergences (Theorem 9.4.1). This inequality has interesting consequences in its own right; as we discuss in Corollary 9.4.2, our result already implies the transport inequality of Cordero-Erausquin [Cor17].

We provide convergence guarantees for the following classes of potentials: (1) convex and relatively smooth (Theorem 9.3.5); (2) strongly relatively convex and relatively smooth (Theorem 9.3.6); and (3) convex and Lipschitz (Theorem 9.3.7). Our results largely match state-of-the-art results for the discretization of the Langevin algorithm for unconstrained sampling. Our work paves the way for the practical deployment of mirror Langevin methods for sampling applications, paralleling the successes of mirror descent in optimization [NY83; Bub15].

In §9.5, we demonstrate the strength of our convergence guarantees compared

with the previous works [Bro+17; BEL18] in various applications such as Bayesian logistic regression.

Other related works. Recently, a few works have proposed modifications of the Langevin algorithm for the task of constrained sampling. Bubeck et al. [BEL18] studied the *projected Langevin algorithm*, which simply projects each step of the Langevin algorithm onto $\mathrm{dom}(V)$. A different approach was taken in Brosse et al. [Bro+17], which applies the Langevin algorithm to a smooth approximation of V given by the Moreau–Yosida envelope. The latter approach was later interpreted and further analyzed by Salim and Richtarik [SR20] using the primal-dual optimality framework from convex optimization.

A different line of work, more closely related to ours, uses a mirror map to change the *geometry* of the sampling problem, see [Hsi+18; Zha+20] and §8. In particular, the mirror Langevin diffusion (MLD) was first introduced in an earlier draft of [Hsi+18], as well as in [Zha+20]. The diffusion was further studied in §8, which provided a simple convergence analysis in continuous time using the sampling analog of *Polyak–Łojasiewicz inequalities* [KNS16]. We also remark that the idea of changing the geometry via a mirror map also played an crucial role for the problem of sampling from the uniform distribution over a polytope [KN12; Che+18a; LV18a; LLV20; GN22; LV22].

Lastly, our work follows the trend of applying ideas from optimization to the task of sampling. Specifically, our analysis adopts the framework of relative convexity and smoothness, which was advocated as a more flexible framework for optimization in [BBT17; LFN18].

9.2 Mirror Langevin Monte Carlo

9.2.1 Background

In this section, we list basic definitions and assumptions that we employ.

Convex functions of Legendre type. Throughout, we assume familiarity with the basic notions of convex analysis [see, e.g., Roc97; BL06].

Definition 9.2.1 (Convex functions of Legendre type [Roc97, §26])**.** *A proper convex lower semicontinuous function $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is* of Legendre type *if*

(i) $\mathcal{Q} := \mathrm{int}(\mathrm{dom}(\phi)) \neq \varnothing$,

(ii) ϕ *is strictly convex and differentiable on* $\mathcal{Q}$*, and*

(iii) $\lim_{k\to\infty} \|\nabla\phi(x_k)\| = \infty$ *whenever* $\{x_k\}_{k\in\mathbb{N}}$ *is a sequence in* $\mathcal{Q}$ *converging to* $\partial\mathcal{Q}$.

The key properties of convex functions of Legendre type are listed below:

- The subdifferential $\partial\phi$ is single-valued and hence $\partial\phi = \{\nabla\phi\}$ [Roc97, Theorem 26.1].
- ϕ is a convex function of Legendre type if and only if its Fenchel conjugate $\phi^\star$ is a convex function of Legendre type [Roc97, Theorem 26.5].
- The gradient $\nabla\phi$ forms a bijection between $\mathrm{int}(\mathrm{dom}(\phi))$ and $\mathrm{int}(\mathrm{dom}(\phi^\star))$ with $\nabla\phi^\star = (\nabla\phi)^{-1}$ [Roc97, Theorem 26.5].

We refer readers to [Roc97, §26] for more details. We henceforth assume that our mirror map ϕ is a convex function of Legendre type.

The natural notion of "distance" associated with the mirror map ϕ is given by the Bregman divergence [see, e.g., Bub15, §4]:

Definition 9.2.2 (Bregman divergence [Brè67])**.** *For a convex function ϕ of Legendre type, the Bregman divergence $D_\phi(\cdot \,\|\, \cdot)$ associated to ϕ is defined as*

$$D_\phi(x \,\|\, y) \coloneqq \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle\,, \qquad \forall x, y \in \mathcal{Q} \coloneqq \mathrm{int}(\mathrm{dom}(\phi))\,.$$

The Bregman divergence behaves like a squared distance; indeed, as $x \to y$ a Taylor expansion shows that $D_\phi(x, y) \sim \frac{1}{2}\langle x - y, \nabla^2\phi(y)\,(x - y)\rangle$. We refer to [Bub15, §4] for other basic properties of the Bregman divergence. Hereinafter, when we use a Bregman divergence, we implicitly assume that its associated function ϕ is a mirror map.

An important case to keep in mind is the mirror map $\phi = \frac{\|\cdot\|^2}{2}$, where $\|\cdot\|$ denotes the Euclidean norm, in which case the Bregman divergence simply becomes $D_\phi(x, y) = \frac{1}{2}\|x - y\|^2$.

Self-concordance. We recall the definition of self-concordance, which has been extensively used in applications such as interior-point methods [NN94]. Given a $\mathcal{C}^2$ strictly convex function ϕ, the local norm at $x \in \mathrm{int}(\mathrm{dom}(\phi))$ with respect to ϕ is defined as

$$\|u\|_{\nabla^2\phi(x)} = \sqrt{\langle \nabla^2\phi(x)u, u\rangle} \qquad \text{for all } u \in \mathbb{R}^d\,.$$

The dual local norm at $x \in \mathrm{int}(\mathrm{dom}(\phi))$ with respect to ϕ is

$$\|u\|_{[\nabla^2\phi(x)]^{-1}} = \sqrt{\langle [\nabla^2\phi(x)]^{-1}u, u\rangle} \qquad \text{for all } u \in \mathbb{R}^d\,.$$

Definition 9.2.3 (Self-concordant function [Nes18, §5.1.3])**.** *We say that a $\mathcal{C}^3$ convex function ϕ is* self-concordant *with a constant $M_\phi \geq 0$ if for any $x \in \mathrm{int}(\mathrm{dom}(\phi))$,*

$$|\nabla^3\phi(x)[u, u, u]| \leq 2M_\phi \|u\|^3_{\nabla^2\phi(x)} \qquad \textit{for all } u \in \mathbb{R}^d\,.$$

Relative convexity/smoothness. We recall the following definitions:

Definition 9.2.4 (Relative convexity [BBT17; LFN18]). *V is α-convex relative to ϕ if*

$$V(y) \geq V(x) + \langle \nabla V(x), y - x \rangle + \alpha D_\phi(y \,\|\, x) \qquad \forall x, y \in \mathcal{Q}\,.$$

Definition 9.2.5 (Relative smoothness [BBT17; LFN18]). *V is β-smooth relative to ϕ if*

$$V(y) \leq V(x) + \langle \nabla V(x), y - x \rangle + \beta D_\phi(y \,\|\, x) \qquad \forall x, y \in \mathcal{Q}\,.$$

For the reader's convenience, we list basic facts regarding relative convexity and smoothness.

Proposition 9.2.6 ([LFN18, Proposition 1.1]). *The following conditions are equivalent:*

- *f is β-smooth relative to h.*
- *$\beta h - f$ is convex on $\mathcal{Q}$.*
- *Under twice differentiability, $\nabla^2 f(x) \preceq \beta\, \nabla^2 h(x)$ for any $x \in \operatorname{int}(\mathcal{Q})$.*
- *$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \beta \langle \nabla h(x) - \nabla h(y), x - y \rangle$ for all $x, y \in \operatorname{int}(\mathcal{Q})$.*

Furthermore, the following conditions are equivalent:

- *f is α-convex relative to h.*
- *$f - \alpha h$ is convex on $\mathcal{Q}$.*
- *Under twice differentiability, $\nabla^2 f(x) \succeq \alpha\, \nabla^2 h(x)$ for any $x \in \operatorname{int}(\mathcal{Q})$.*
- *$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \alpha \langle \nabla h(x) - \nabla h(y), x - y \rangle$ for all $x, y \in \operatorname{int}(\mathcal{Q})$.*

In the rest of this work, we assume that $V \in \mathcal{C}^2(\mathcal{X})$ where $\mathcal{X} := \operatorname{int}(\operatorname{dom}(V))$, that $\mathcal{X} \subseteq \overline{\mathcal{Q}}$, and $\mathcal{X} \cap \mathcal{Q} \neq \varnothing$. Also, we assume that $\exp(-V)$ is integrable so that π is well-defined; this holds if and only if $V(x) \geq a\,\|x\| - b$ for some $a, b > 0$ [Bra+14, Lemma 2.2.1].

Optimal transport. Given a lower semicontinuous cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty]$, we can define the *optimal transport* cost between two probability measures μ and ν on $\mathbb{R}^d$ to be

$$\inf\{\mathbb{E}\, c(X, Y) \mid X \sim \mu,\ Y \sim \nu\}\,. \tag{9.1}$$

Here, the infimum is taken over pairs of random variables (X, Y) defined on the same probability space, with marginal laws μ and ν respectively. It is known that the infimum in (9.1) is always attained; we refer to the standard introductory texts [Vil03; Vil09b; San15] for this and other basic facts in optimal transport.

In this work, we are most concerned with the case when the cost function c is the Bregman divergence associated with a mirror map:

Definition 9.2.7 (Bregman transport cost)**.** *The Bregman transport cost is defined as*

$$\mathcal{D}_\phi(\mu \,\|\, \nu) \coloneqq \inf\{\mathbb{E}\, D_\phi(X \,\|\, Y) \mid X \sim \mu,\ Y \sim \nu\}\,.$$

The Bregman transport cost was also studied in [Cor17].

In particular, when $\phi = \frac{\|\cdot\|^2}{2}$, we obtain an important special case:

Definition 9.2.8 (2-Wasserstein distance)**.** *The 2-Wasserstein distance W_2 is defined as*

$$W_2^2(\mu, \nu) \coloneqq \inf\{\mathbb{E}[\|X - Y\|^2] \mid X \sim \mu,\ Y \sim \nu\}\,.$$

The W_2 optimal transport cost indeed defines a metric over the space of probability measures on $\mathbb{R}^d$ with finite second moment [Vil03, Theorem 7.3]; we refer to this metric space as the *Wasserstein space*. The W_2 metric is particularly important because it arises from a formal Riemannian structure on the Wasserstein space. This perspective was introduced in [Ott01] and applied to the Langevin diffusion in [JKO98; OV00]; in particular, these latter two works justify the perspective of the Langevin diffusion as a gradient flow of the Kullback-Leibler divergence in Wasserstein space. A rigorous exposition to Wasserstein calculus can be found in [AGS08; Vil09b]. See also §2.1 for further background.

Here, we give a brief and informal introduction to the calculation rules of optimal transport. For any regular curve of measures $(\mu_t)_{t\geq 0}$, there is a corresponding family of *tangent vectors* $(v_t)_{t\geq 0}$ [see AGS08, Theorem 8.3.1]; here, $v_t : \mathbb{R}^d \to \mathbb{R}^d$ is a vector field on $\mathbb{R}^d$. Also, if $\mathcal{F}$ is any well-behaved functional defined over Wasserstein space, then at each regular measure μ one can define the *Wasserstein gradient* of $\mathcal{F}$ at μ, which we denote $\nabla_{W_2}\mathcal{F}(\mu)$; it is also a mapping $\mathbb{R}^d \to \mathbb{R}^d$. Then, we have the calculation rule

$$\partial_t \mathcal{F}(\mu_t) = \mathbb{E}\langle \nabla_{W_2}\mathcal{F}(\mu_t)(X_t), v_t(X_t)\rangle$$

for any regular curve of measures $(\mu_t)_{t\geq 0}$ with corresponding tangent vectors $(v_t)_{t\geq 0}$, where $X_t \sim \mu_t$. We will use this calculation rule in §9.6.

9.2.2 Discretization of the mirror Langevin diffusion

In order to turn a continuous-time diffusion such as (MLD) into an implementable algorithm, it is necessary to first discretize the stochastic process. The discretization considered in [Zha+20] and in §8 is a simple Euler–Maruyama discretization: fixing $h > 0$, we define a sequence of iterates $(X_k)_{k\in\mathbb{N}}$ via

$$\nabla\phi(X_{k+1}) = \nabla\phi(X_k) - h\,\nabla V(X_k) + \sqrt{2h}\,[\nabla^2\phi(X_k)]^{1/2}\,\xi_k\,, \tag{9.2}$$

where $(\xi_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. standard Gaussians in $\mathbb{R}^d$.

However, many other discretizations are possible. Indeed, in many machine learning applications, the most costly step is the evaluation of ∇V, which may require a sum over a large training set, whereas the mirror map ϕ may be chosen to have a simple form. For the purpose of obtaining a more efficient sampling algorithm, it may therefore be a favorable trade-off to use a high-precision implementation of the diffusion step at the cost of additional computation time (which nonetheless does not require additional query access to the gradients of V). Motivated by these considerations, we propose a new discretization (see Figure 9.1 for an illustration):

Mirror Langevin Monte Carlo (MLMC):

$$X_{k+1/2} \coloneqq \operatorname*{arg\,min}_{x\in\mathcal{Q}} \left[\langle h\,\nabla V(X_k), x\rangle + D_\phi(x \,\|\, X_k)\right], \tag{MLMC:1}$$

$$X_{k+1} \coloneqq \nabla\phi^\star(W_h)\,, \qquad \text{where } \begin{cases} \mathrm{d}W_t = \sqrt{2}\,[\nabla^2\phi^\star(W_t)]^{-1/2}\,\mathrm{d}B_t\,, \\ W_0 = \nabla\phi(X_{k+1/2})\,. \end{cases} \tag{MLMC:2}$$

In MLMC:2, the stochastic processes $(W_t)_{t\geq 0}$ are assumed to be driven by independent Brownian motions at each iteration. When $\phi = \frac{\|\cdot\|^2}{2}$, MLMC recovers the unadjusted Langevin algorithm.

Practicality. Although MLMC:2 is defined using the exact solution of an SDE (and we analyze the exact step MLMC:2 for simplicity), it should be understood as capturing the idea of discretizing the diffusion step more finely (e.g., through multiple inner iterations of an Euler–Maruyama discretization) than the gradient step. This is indeed amenable to practical implementation since, as previously discussed, the gradient step is typically much more costly than the diffusion step. Moreover, this is justified by our theoretical results in the next section, which together with the conjecture of [Zha+20] suggest that fine discretization of MLMC:2 is potentially crucial for attaining vanishing bias. Nevertheless, it is

indeed the case that a single iteration of MLMC is more costly than a single step of the Euler–Maruyama discretization of MLD, and this represents a limitation of our work.

Remark 9.2.9. *Our proposed discretization can be understood as a more faithful discretization of the mirror Langevin diffusion* (MLD), à la *[GWS21]. It can also be understood as the forward-flow discretization of* MLD *in the interpretation of [Wib18].*

Remark 9.2.10. *In order for* MLMC*:2 to be well-defined, we require assumptions on ϕ such that the diffusion $(W_t)_{t\geq 0}$ is non-explosive, i.e., it does not exit* $\mathrm{int}(\mathrm{dom}(\phi^\star))$ *in finite time. This holds under mild assumptions on ϕ; see [GK96]. For situations of interest, the assumptions of [GK96] can be checked directly.*

■ 9.3 Convergence analysis for mirror Langevin Monte Carlo

First, we state the main assumptions which are used for our main results.

Assumption 9.3.1 (Self-concordance of ϕ)**.** *We assume that the mirror map ϕ is M_ϕ-self-concordant (Definition 9.2.3).*

Assumption 9.3.2 (Relative Lipschitzness)**.** *We assume that V is L-relatively Lipschitz with respect to ϕ, in the sense that $\|\nabla V(x)\|_{[\nabla^2\phi(x)]^{-1}} \leq L$ for all $x \in \mathcal{X}$ (see §9.2.1 for the definition of the local norm used here).*

Assumption 9.3.3 (Relative convexity and smoothness)**.** *We assume that V is α-convex relative to ϕ and β-smooth relative to ϕ, where $0 \leq \alpha \leq \beta \leq \infty$ (Definitions 9.2.4 and 9.2.5).*

Remark 9.3.4. *We work under weaker assumptions than those of [Zha+20]. In particular, our analysis does not assume the moment condition on the Hessian ((A2) therein) and the bound on the commutator between $\nabla^2\phi$ and $\nabla^2 V$ ((A5) therein). Moreover, our analysis uses weaker (and more standard) definitions of self-concordance.*

Throughout this section, we assume the conditions listed above, and we present convergence results for MLMC under various sets of assumptions. Our first two results pertain to the smooth case, i.e., $\beta < \infty$. Define the parameter

$$\boxed{\beta' := \beta + 2M_\phi L\,.}$$

One might wonder how large β' is for typical applications. First, we only need ϕ to be self-concordant, *not* a self-concordant *barrier*. Hence the appearance of

M_ϕ is typically not problematic; for instance, a log barrier with m constraints is $O(1)$-self-concordant. The smoothness parameter could be large and dimension-dependent in general. However, such situations are actually where our approach could be potentially advantageous. In §9.5.2, we demonstrate an example where the smoothness parameter becomes much smaller by choosing ϕ carefully.

Theorem 9.3.5 (Weakly convex case)**.** *Suppose that Assumptions 9.3.1, 9.3.2, and 9.3.3 hold with $\alpha = 0$ and $\beta' > 0$. For a target accuracy $\varepsilon > 0$, let $X_k \sim \mu_k$ denote the iterates of* MLMC *with step size $h = \min\{\frac{\varepsilon^2}{2\beta' d}, \frac{1}{\beta'}\}$. Then, the following convergence rate holds for the mixture distribution $\overline{\mu}_N := \frac{1}{N}\sum_{k=1}^N \mu_k$:*

$$\mathsf{KL}(\overline{\mu}_N \,\|\, \pi) \le \varepsilon^2\,, \qquad \text{provided that } N \ge \frac{4\beta' d\, \mathcal{D}_\phi(\pi \,\|\, \mu_0)}{\varepsilon^4} \max\Big\{1, \frac{\varepsilon^2}{2d}\Big\}\,. \tag{9.4}$$

Proof. See §9.6.2. □

Theorem 9.3.6 (Strongly relatively convex case)**.** *Suppose that the Assumptions 9.3.1, 9.3.2, 9.3.3 hold with $\alpha, \beta' > 0$.*

1. *(Convergence in Bregman transport cost) For a target accuracy $\varepsilon > 0$, let $X_k \sim \mu_k$ denote the iterates of* MLMC *with step size $h = \min\{\frac{\alpha\varepsilon^2}{2\beta' d}, \frac{1}{\beta'}\}$. Then,*

$$\mathcal{D}_\phi(\pi \,\|\, \mu_N) \le \varepsilon^2\,, \quad \text{provided that } N \ge \frac{2\beta' d}{\alpha^2\varepsilon^2} \ln\Big(\frac{2\mathcal{D}_\phi(\pi \,\|\, \mu_0)}{\varepsilon^2}\Big) \max\Big\{1, \frac{\alpha\varepsilon}{2d}\Big\}\,.$$

2. *(Convergence in KL divergence) For a target accuracy $\varepsilon > 0$, suppose that $X_0 \sim \mu_0$ satisfies $\mathcal{D}_\phi(\pi \,\|\, \mu_0) \le \varepsilon^2/\alpha$. Let $X_k \sim \mu_k$ denote the iterates of* MLMC *with step size $h = \min\{\frac{\alpha\varepsilon^2}{2\beta' d}, \frac{1}{\beta'}\}$. Then, the following convergence rate holds for the mixture distribution $\overline{\mu}_N := \frac{1}{N}\sum_{k=1}^N \mu_k$,*

$$\mathsf{KL}(\overline{\mu}_N \,\|\, \pi) \le \varepsilon^2\,, \qquad \text{provided that } N \ge \frac{4\beta' d}{\alpha\varepsilon^2} \max\Big\{1, \frac{\varepsilon^2}{2d}\Big\}\,.$$

Proof. See §9.6.3. □

Note that the initialization assumption $\mathcal{D}_\phi(\pi, \mu_0) \le \varepsilon^2/\alpha$ in the second assertion of Theorem 9.3.6 can be obtained from the first guarantee of Theorem 9.3.6. Chaining together the two parts of the theorem, we therefore obtain the following guarantee: suppose that we initialize MLMC at a distribution μ_0. Then, with step size $h = \min\{\frac{\alpha\varepsilon^2}{2\beta' d}, \frac{1}{\beta'}\}$, we obtain

$$\mathcal{D}_{\mathsf{KL}}\Big(\frac{1}{N_1}\sum_{k=N_0+1}^{N_0+N_1} \mu_k \,\Big\|\, \pi\Big) \le \varepsilon^2\,, \qquad \text{provided that } \begin{cases} N_0 \ge \widetilde{\Omega}(\frac{\beta' d}{\alpha\varepsilon^2})\,, \\ N_1 \ge \Omega(\frac{\beta' d}{\alpha\varepsilon^2})\,. \end{cases}$$

Observe also that for the case $\phi = \frac{\|\cdot\|^2}{2}$, Theorems 9.3.5 and 9.3.6 recover the corresponding convergence guarantees for the unadjusted Langevin algorithm [Corollary 7, and Corollaries 10 and 11 respectively in DMM19].[1]

Next, we present our guarantee for the non-smooth case $\beta = \infty$. For this result, we assume that ϕ is strongly convex w.r.t. a norm $\|\cdot\|$ on $\mathbb{R}^d$, and that V is $\tilde{L}$-Lipschitz in this norm. Since the norm of the gradient should be measured in the dual norm $\|\cdot\|_\star$, this means precisely that

$$\|\nabla V(x)\|_\star \le \tilde{L}\,, \qquad \text{for all } x \in \mathcal{X}\,. \tag{9.5}$$

We also note that the next result does not require self-concordance of ϕ.

Theorem 9.3.7 (Non-smooth case)**.** *Assume ϕ is 1-strongly convex w.r.t. a norm $\|\cdot\|$ on $\mathbb{R}^d$, that V is $\tilde{L}$-Lipschitz in this norm (in the sense of (9.5)), and that $\alpha = 0$ (i.e., V is convex). For a target accuracy $\varepsilon > 0$, let $X_k \sim \mu_k$ denote the iterates of* MLMC *with step size $h = \varepsilon^2/\tilde{L}^2$. Then, the following convergence rate holds for the mixture distribution $\overline{\mu}_N := \frac{1}{N}\sum_{k=1}^N \mu_k$:*

$$\mathsf{KL}(\overline{\mu}_N \,\|\, \pi) \le \varepsilon^2\,, \qquad \textit{provided that } N \ge \frac{2\tilde{L}^2 \mathcal{D}_\phi(\pi \,\|\, \mu_{1/2})}{\varepsilon^4}\,.$$

Proof. See §9.6.4. □

The assumption (9.5) is stronger than relative Lipschitzness: if V satisfies (9.5) and ϕ is 1-strongly convex w.r.t. $\|\cdot\|$, then V is L-relatively Lipschitz with respect to ϕ with $L \le \tilde{L}$. When $\|\cdot\|$ is the Euclidean norm and $\phi = \frac{\|\cdot\|^2}{2}$, then we recover a special case of [DMM19, Corollary 14].

We now make a number of remarks about our result.

Remark 9.3.8 (Implementing $\overline{\mu}_N$)**.** *One can output a sample from $\overline{\mu}_N$ by simply outputting one of the iterates $\{X_k\}_{k=1}^N$ chosen uniformly at random.*

Remark 9.3.9 (Convergence in other metrics)**.** *Using standard inequalities, our results for convergence in KL divergence imply convergence in a number of other information divergences such as the total variation distance, see [Tsy09, §2.4].*

When V is α-strongly convex (w.r.t. $\frac{\|\cdot\|^2}{2}$), then the T_2 transport inequality [Vil09b, Theorem 22.14] $\alpha W_2^2(\mu,\pi) \le \mathsf{KL}(\mu\,\|\,\pi)$ implies convergence in the W_2 distance as well. In general, we do not have convergence in W_2, but we can always obtain convergence with respect to a different optimal transport cost, namely, the Bregman transport cost $\mathcal{D}_V$ associated with V. This is a consequence of Corollary 9.4.2 [also see Cor17, Proposition 1], which asserts that $\mathcal{D}_V(\mu,\pi) \le \mathsf{KL}(\mu\,\|\,\pi)$.

[1] In this case, $M_\phi = 0$, so the Lipschitz constant L does not enter the final result. In particular, it is not contradictory to assume strong convexity ($\alpha > 0$).

Remark 9.3.10 (Dimension dependence)**.** *Ignoring for now the dependence on β' (which may also have a dimension dependence depending on the application), the Bregman divergence $\mathcal{D}_\phi(\pi \,\|\, \mu_0)$ term is typically of size $O(d)$ (see §9.5.1 for a particular instance of this). Thus, our overall dimension dependence is $O(d^2)$ for the weakly convex case and $O(d)$ for the strongly convex case. Overall, this is a significantly better dependence on the dimension as compared to the previous works [Bro+17; BEL18]; we perform a comparison in §9.5.1 for a specific setting.*

We also remark that mirror descent has classically been used for dimension reduction by changing the geometry of the algorithm from ℓ_2 to ℓ_1. We investigate the possibility of doing the same for sampling in §9.5.2.

Remark 9.3.11 (Comparison with [Hsi+18])**.** *[Hsi+18] show that for strictly log-concave targets, there exists a good mirror map ϕ for which the pushforward of the target distribution via $\nabla\phi$ enjoys the same guarantees as ordinary Langevin. However, this result is only existential and gives no guidance on how to construct the mirror map. In contrast, our theorems hold for any choice of mirror map which satisfies our assumptions, and provide guidance on how to choose the mirror map. Also, our relative smoothness condition allows for potentials which blow up at the boundary of their domain (i.e., the target distribution vanishes near the boundary of its support), whereas this is forbidden by the assumptions of [Hsi+18]. Lastly, our algorithm does not require computing the third derivative of the mirror map, whereas this is required for [Hsi+18].*

Remark 9.3.12 (Comparison with [Zha+20])**.** *[Zha+20] performs an analysis of the Euler–Maruyama discretization of* MLD*, which we temporarily refer to as* MLMC′*. Our result guarantees that for any desired accuracy ε, it is possible to choose the step size sufficiently small so that* MLMC *achieves the target accuracy; in contrast, the result of [Zha+20] only guarantees that* MLMC′ *contracts to within a ball around π of radius $O(\sqrt{d})$ (measured w.r.t. a modified Wasserstein distance).*[2] *Moreover, [Zha+20] conjecture that their bias term is unavoidable.*

In light of our result, we believe that it is an interesting open question to resolve their conjecture. Their conjecture, if true, suggests that replacing MLMC*:2 in our algorithm by a single step of the Euler–Maruyama discretization has disastrous effects on the convergence of the algorithm, and therefore provides further support for considering* MLMC *instead of* MLMC′*.*

Recently, after the first draft of our paper was published online, Li et al. [Li+22] gave an analysis of MLMC′ *under a subset of the assumptions of [Zha+20] which indeed exhibits vanishing bias, provided that the relative strong convexity parameter of the potential is sufficiently large compared to the modified self-concordance*

[2] Notably, the radius of this ball is comparable to the distance at initialization.

parameter of the mirror map. It remains an open question to remove this latter restriction from their work, and moreover to obtain similar results under the more usual definitions of relative convexity/smoothness and self-concordance that we adopt in this work.

In order to generalize the discretization analysis from the vanilla Langevin algorithm to the mirror Langevin algorithm, in the next section we prove a new *displacement convexity* result for the entropy with respect to the Bregman transport cost which may be of independent interest.

9.4 Convexity of the entropy with respect to the Bregman divergence

It is well-known that the entropy functional $\mathcal{H}$ is displacement convex along W_2 geodesics [AGS08, Theorem 9.4.11]. In fact, this displacement convexity is crucial in showing that $\mathsf{KL}(\cdot \,\|\, \pi) = \mathcal{E} + \mathcal{H}$ is displacement convex (when π is log-concave), which in turn is used to analyze the convergence of LD to the target measure. Therefore, in order to understand the convergence of MLD, it is crucial to see if such a result is true when W_2 is replaced by $\mathcal{D}_\phi$. We prove that indeed the displacement convexity-like property holds for $\mathcal{H}$ under $\mathcal{D}_\phi$-optimal couplings.

Theorem 9.4.1 ("Convexity" of the entropy with respect to the Bregman divergence)**.** *Let μ, ν be probability measures on $\mathbb{R}^d$ and let $X \sim \mu$, $Y \sim \nu$ be coupled according to the Bregman transport cost $\mathcal{D}_\phi(\mu \,\|\, \nu)$. Then, it holds that*

$$\mathcal{H}(\nu) \geq \mathcal{H}(\mu) + \mathbb{E}\big\langle [\nabla_{W_2}\mathcal{H}(\mu)](X), Y - X \big\rangle \,.$$

As a corollary, we can use the calculus of optimal transport in order to recover the transport inequality of [Cor17]. In the following, we do not carry out the approximation arguments necessary to make the proof fully correct because a rigorous proof of the statement is already given in [Cor17].[3] Rather, our main purpose in giving this argument is simply to point out the convexity principle which underlies the transport inequality.

Corollary 9.4.2 ([Cor17, Proposition 1])**.** *For any probability measure μ on $\mathbb{R}^d$,*

$$\mathsf{KL}(\mu \,\|\, \pi) \geq \mathcal{D}_V(\mu \,\|\, \pi) \,.$$

Proof sketch. Let (X, Y) be optimally coupled according to the Bregman transport cost $\mathcal{D}_V(\cdot \,\|\, \cdot)$ between μ and π. We decompose $\mathsf{KL}(\mu \,\|\, \pi) = \mathbb{E}\, V(X) + \mathcal{H}(\mu)$. On one hand, the first term is

$$\mathbb{E}\, V(X) = \mathbb{E}\, V(Y) + \mathbb{E}\langle \nabla V(Y), X - Y \rangle + \mathbb{E}\, D_V(X \,\|\, Y) \,.$$

[3] In fact, the proof of [Cor17] does not even require convexity of V.

On the other hand, the convexity result (Theorem 9.4.1) shows that

$$\mathcal{H}(\mu) \geq \mathcal{H}(\pi) + \mathbb{E}\langle \nabla_{W_2}\mathcal{H}(\pi)(Y), X - Y\rangle\,.$$

Putting these together, we obtain

$$\begin{aligned}\mathsf{KL}(\mu \,\|\, \pi) &\geq \mathbb{E}\, V(Y) + \mathbb{E}\langle \nabla V(Y), X - Y\rangle + \mathbb{E}\, D_V(X \,\|\, Y) \\ &\quad + \mathcal{H}(\pi) + \mathbb{E}\langle \nabla_{W_2}\mathcal{H}(\pi)(Y), X - Y\rangle \\ &= \mathsf{KL}(\pi \,\|\, \pi) + \mathbb{E}\big\langle [\nabla V + \nabla_{W_2}\mathcal{H}(\pi)](Y), X - Y\big\rangle + \mathbb{E}\, D_V(X \,\|\, Y) \\ &= \mathbb{E}\, D_V(X \,\|\, Y)\,,\end{aligned}$$

since $\nabla V + \nabla_{W_2}\mathcal{H}(\pi)$, the W_2 gradient of $\mathsf{KL}(\cdot \,\|\, \pi)$ at π, is zero. □

9.5 Applications

In this section, we provide examples which illustrate our results; see [AC21] for numerical experiments.

9.5.1 Bayesian logistic regression

In this section, we apply our main result to Bayesian logistic regression.

We recall the setting of Bayesian logistic regression: we observe pairs (X_i, Y_i), $i = 1, \ldots, n$, where $X_i \in \mathbb{R}^d$ and $Y_i \in \{0, 1\}$. The data follow the model

$$Y_i \sim \mathsf{Bernoulli}\Big(\frac{\exp\langle \theta, X_i\rangle}{1 + \exp\langle \theta, X_i\rangle}\Big)\,, \qquad \text{independently for } i = 1, \ldots, n\,. \tag{9.6}$$

Here, the parameter θ itself is assumed to be a random variable taking values in $\mathbb{R}^d$. If we assume that θ has a prior density λ with respect to Lebesgue measure, then the posterior distribution is

$$\pi(\theta) \propto \lambda(\theta) \exp\Big[\sum_{i=1}^{n}\big(Y_i \langle \theta, X_i\rangle - \ln(1 + \exp\langle \theta, X_i\rangle)\big)\Big]\,.$$

Since it may be computationally infeasible to explicitly compute the normalizing constant for the posterior distribution, we turn towards sampling algorithms.

When we take a prior λ which has full support on $\mathbb{R}^d$, e.g., a Gaussian prior, then we may apply off-the-shelf methods such as the Langevin diffusion (LD). However, if we choose a prior which has compact support, then the unadjusted Langevin algorithm is no longer an acceptable option because it outputs samples outside the support of the posterior. In this case, we must turn to other methods,

such as the projected Langevin algorithm [BEL18]. Here, we explore the use of the mirror Langevin algorithm (MLMC) for constrained sampling.

For the rest of this section, we will focus on a particular problem for concreteness and interpretability: we consider the uniform prior λ on the ℓ_∞ ball $[-1,1]^d$. By duality, this is an attractive model when the data $(X_i)_{i=1}^n$ have small ℓ_1-norm, i.e., are approximately sparse. A natural choice of mirror map for this problem is the logarithmic barrier

$$\phi(\theta) = \sum_{i=1}^{d} \Big(\ln \frac{1}{1-\theta[i]} + \ln \frac{1}{1+\theta[i]} \Big) ,$$

where we use $\theta[\cdot]$ to denote the coordinates of $\theta \in \mathbb{R}^d$. Then, ϕ is 1-self-concordant ([Nes18, §5.1.3]). We remark that the separability of the mirror map in this example implies that the diffusion step MLMC:2 can be simulated in $O(d)$ steps, rather than $O(d^2)$.

For this setting, we compare the guarantees of MLMC with the Moreau–Yosida unadjusted Langevin algorithm (MYULA) [Bro+17] and the projected Langevin algorithm (PLA) [BEL18]; see Table 9.1.

Algorithm	Guarantee
MLMC	$O(d/\varepsilon^4)$
MYULA	$O(d^9/\varepsilon^6)$
PLA	$O(d^{15}/\varepsilon^{12})$

Table 9.1: Comparison of MLMC with other constrained sampling algorithms for the number of iterations required to output a sample whose total variation distance to π is at most ε. For simplicity, we focus on the dependence with respect to dimension and the accuracy ε.

Details. We may compute

$$\begin{aligned} V(\theta) &= \sum_{i=1}^{n} \big(-Y_i \langle \theta, X_i \rangle + \ln(1 + \exp \langle \theta, X_i \rangle) \big) , \\ \nabla V(\theta) &= -\sum_{i=1}^{n} \Big(Y_i - \frac{\exp \langle \theta, X_i \rangle}{1 + \exp \langle \theta, X_i \rangle} \Big) X_i , \\ \nabla^2 V(\theta) &= \sum_{i=1}^{n} \frac{\exp \langle \theta, X_i \rangle}{(1 + \exp \langle \theta, X_i \rangle)^2} X_i X_i^\mathsf{T} , \end{aligned}$$

and

$$\phi(\theta) = \sum_{i=1}^{d} \Big(\ln \frac{1}{1-\theta[i]} + \ln \frac{1}{1+\theta[i]} \Big) ,$$
$$\nabla \phi(\theta) = \sum_{i=1}^{d} \Big(\frac{1}{1-\theta[i]} - \frac{1}{1+\theta[i]} \Big) e_i ,$$
$$\nabla^2 \phi(\theta) = \operatorname{diag}\Big[\frac{1}{(1-\theta)^2} + \frac{1}{(1+\theta)^2} \Big] .$$

From these expressions, we see that

$$0 \preceq \nabla^2 V \preceq \sum_{i=1}^{n} X_i X_i^{\mathsf{T}} , \qquad 2 I_d \preceq \nabla^2 \phi .$$

Let $L \coloneqq \sup_{[-1,1]^d} \|\nabla V\|$ denote the (ordinary) Lipschitz constant of V, and let β denote the (ordinary) smoothness parameter of V (from above we see that β can be taken to be the largest eigenvalue of $\sum_{i=1}^{n} X_i X_i^{\mathsf{T}}$). Note that the 2-strong convexity of ϕ implies that V is $L/\sqrt{2}$-relatively Lipschitz and $\beta/2$-relatively smooth with respect to ϕ, so Theorem 9.3.5 holds with

$$\beta' = \frac{\beta}{2} + \sqrt{2} L .$$

In order to fully understand the quantitative convergence rate provided by Theorem 9.3.5, we must also bound the Bregman divergence $\mathcal{D}_\phi(\pi, \mu_0)$. We have:

Lemma 9.5.1. *Let $\mu_0 = \delta_0$ be the point mass at 0. Then, for the logarithmic barrier mirror map ϕ defined above, we have $\mathcal{D}_\phi(\pi \parallel \mu_0) \le 4.1\,(1 + \beta + L)\, d$.*

Proof. See §9.5.4. □

From Theorem 9.3.5, we can deduce that using N iterations of MLMC, we can obtain a distribution μ_N^{MLMC} such that for $\varepsilon \le \sqrt{2d}$,

$$2 \, \|\mu_N^{\mathsf{MLMC}} - \pi\|_{\mathsf{TV}}^2 \le \mathsf{KL}(\mu_N^{\mathsf{MLMC}} \parallel \pi) \le \varepsilon^2 , \qquad \text{provided } N \ge \frac{23\,(1+\beta+L)^2 d^2}{\varepsilon^4} ,$$

where β is the largest eigenvalue of $\sum_{i=1}^{n} X_i X_i^{\mathsf{T}}$ and $L \coloneqq \sup_{[-1,1]^d} \|\nabla V\|$ is the usual Lipschitz constant of V. In fact, if we use the non-smooth guarantee in Theorem 9.3.7, then we can improve this to $O(L^2 d/\varepsilon^4)$ iterations. For comparison

purposes, the Moreau–Yosida unadjusted Langevin algorithm (MYULA) [Bro+17, Theorem 2] with $R = \sqrt{d}$ provides the guarantee[4]

$$\|\mu_N^{\mathsf{MYULA}} - \pi\|_{\mathrm{TV}} \le \varepsilon\,, \qquad \text{provided } N \ge \widetilde{\Omega}\Big(\frac{\beta^4 d^9}{\varepsilon^6}\Big)\,.$$

On the other hand, the corresponding guarantee for the projected Langevin algorithm (PLA) [BEL18, Theorem 1] with $R = \sqrt{d}$ implies

$$\|\mu_N^{\mathsf{PLA}} - \pi\|_{\mathrm{TV}} \le \varepsilon\,, \qquad \text{provided } N \ge \widetilde{\Omega}\Big(\frac{(\sqrt{d} + \beta + L)^{12} d^9}{\varepsilon^{12}}\Big)\,.$$

9.5.2 Better dimension dependency via mirror Langevin

As described in [Bub15, §4.3], a classical application of mirror descent is to obtain better dependence on the dimension by changing the geometry of the optimization algorithm from ℓ_2 to ℓ_1. We investigate the possibility of analogous improvements in the setting of constrained sampling.

We consider a simple toy problem in which the constraint set is the interior of the filled-in simplex $\mathcal{Q} \coloneqq \{x \in \mathbb{R}^d \mid x > 0,\ \sum_{i=1}^d x[i] < 1\}$, and we take the potential to be a quadratic

$$V(x) \coloneqq \frac{1}{2}\,\langle x, Ax\rangle\,,$$

where $A \in \mathbb{R}^{d\times d}$ is a symmetric positive semidefinite matrix with all entries bounded in magnitude by 1. We choose as our mirror map the barrier:

$$\phi(x) \coloneqq \sum_{i=1}^d \ln\frac{1}{x[i]} + \ln\frac{1}{1 - \sum_{i=1}^d x[i]}\,.$$

This map is self-concordant with parameter 1.

We can compute

$$\nabla\phi(x) = \sum_{i=1}^d \Big(-\frac{1}{x[i]} + \frac{x[i]}{1 - \sum_{j=1}^d x[j]}\Big)\,e_i\,,$$
$$\nabla^2\phi(x) = \operatorname{diag}\frac{1}{x^2} + \frac{I_d}{1 - \sum_{i=1}^d x[i]} + \frac{xx^{\mathsf{T}}}{\big(1 - \sum_{i=1}^d x[i]\big)^2}\,.$$

[4]To be precise, their bound on the number of iterates required reads $N \ge \widetilde{\Omega}(\Delta_2^4 d^7/\varepsilon^6)$, where Δ_2 is a parameter measuring how close the domain $\operatorname{dom}(V)$ is to an isotropic convex body. For concreteness, we bound this parameter by βR following [Bro+17, pg. 7].

Since $x[i] < 1$ for all $i \in [d]$, it follows that $\langle v, \operatorname{diag}(1/x^2)\, v\rangle \geq \langle v, \operatorname{diag}(1/x)\, v\rangle \geq \|v\|_1^2$, where the second inequality follows from the strong convexity of the entropy with respect to the ℓ_1-norm. Hence, ϕ is 1-strongly convex with respect to the ℓ_1-norm. From our assumption on A,

$$\|\nabla V(x)\|_{[\nabla^2\phi(x)]^{-1}} \leq \|\nabla V(x)\|_\infty \leq 1\,,$$
$$\langle v, \nabla^2 V(x)\, v\rangle \leq \Big|\sum_{i,j=1}^d A_{i,j} v_i v_j\Big| \leq \sum_{i,j=1}^d |v_i|\,|v_j| \leq \|v\|_1^2\,,$$

which implies that V is 1-relatively Lipschitz and 1-relatively smooth with respect to ϕ, and the assumptions of Theorem 9.3.5 hold with $\beta' = 3$. In contrast, if we had instead considered the ℓ_2-norm, then the Lipschitz constant of V could be as large as $\sqrt{d}$, and the smoothness parameter of V could be as large as d. Together with a warm start, this suggests that MLMC could attain a better dimension dependence for this example.

Remark 9.5.2. *Alternatively, we can apply Theorem 9.3.7 with the entropic mirror map*

$$\phi(x) = \sum_{i=1}^d x[i] \ln x[i] + \Big(1 - \sum_{i=1}^d x[i]\Big) \ln\Big(1 - \sum_{i=1}^d x[i]\Big)$$

to the above setting; note that Theorem 9.3.7 only requires standard assumptions for mirror descent guarantees (e.g., [Bub15, Theorem 4.2]), and does not require the mirror map to be self-concordant. In particular, V is 1-Lipschitz w.r.t. $\|\cdot\|_1$ and ϕ is strongly convex w.r.t. $\|\cdot\|_1$, so Theorem 9.3.7 implies that $\mathsf{KL}(\overline{\mu}_N \,\|\, \pi) \leq \varepsilon^2$ after $N = O\big(\frac{\mathcal{D}_\phi(\pi\|\mu_{1/2})}{\varepsilon^4}\big)$ iterations. For comparison, note that the approach of Hsieh et al. [Hsi+18] does not apply to this example, because the pushforward of the distribution via the entropic mirror map is not log-concave.

■ 9.5.3 Sampling from non-smooth distributions

Thus far, we have focused on distributions whose potential V is bounded within its domain $\operatorname{dom}(V)$. However, in many applications, one is required to sample from a distribution whose potential V blows up near the boundary of its domain. Such distributions violate the standard assumptions of Lipschitzness and smoothness and hence are beyond the scope of the existing guarantees. In this subsection, we demonstrate that one can still sample from such distributions via MLMC together with the relative Lipschitzness and relative smoothness.

Consider the Dirichlet distribution π which is defined on the interior of the filled-in simplex $\mathcal{Q} \coloneqq \{x \in \mathbb{R}^d \mid x > 0,\ \sum_{i=1}^d x[i] < 1\}$ by the potential

$$V(x) = a_0 \ln \frac{1}{1 - \sum_{i=1}^d x[i]} + \sum_{i=1}^d a_i \ln \frac{1}{x[i]}\,,$$

for some constants $a_0, a_1, \dots, a_d > 0$, and we take $V = \phi$. Then, it is well-known that V is $(\max_{i=0,1,\dots,d} {a_i}^{-1/2})$-self-concordant [Nes18, Theorem 5.1.1]. Also, from $(\sum_{i=0}^d a_i)$-exp-concavity of V [Nes18, Theorem 5.3.2], $\|\nabla V(x)\|_{[\nabla^2 V(x)]^{-1}} \leq \left(\sum_{i=0}^d a_i\right)^{1/2}$. Therefore, it follows that V is $\left(\sum_{i=0}^d a_i\right)^{1/2}$-Lipschitz, 1-convex, and 1-smooth relative to V, so that the assumptions of Theorem 9.3.6 hold with

$$\beta' = 1 + 2\left(\max_{i=0,1,\dots,d} {a_i}^{-1/2}\right)\left(\sum_{i=0}^d a_i\right)^{1/2} \leq 3\sqrt{d}\,\sqrt{\frac{a_{\max}}{a_{\min}}}\,,$$

where $a_{\max} \coloneqq \max_{i=0,1,\dots,d} a_i$ and $a_{\max} \coloneqq \min_{i=0,1,\dots,d} a_i$. Therefore, one can obtain a mixture distribution $\overline{\mu}_N$ after N iterations of MLMC such that

$$\mathsf{KL}(\overline{\mu}_N \,\|\, \pi) \leq \varepsilon^2\,, \quad \text{provided that } N \geq \widetilde{\Omega}\left(\sqrt{\frac{a_{\max}}{a_{\min}}}\,\frac{d^{3/2}}{\varepsilon^2}\right).$$

■ 9.5.4 Auxiliary results

Lemma 9.5.3. *Let π be a probability distribution supported on $[-1,1]^d$ which has density proportional to $\exp(-V)$. Assume that $V : [-1,1]^d \to \mathbb{R}^d$ is L-Lipschitz and β-smooth. Then, we have the following bound on the marginal density π_1 of π on the first coordinate:*

$$\sup_{[-1,1]} \pi_1 \leq 3\,(1 + \sqrt{\beta} + L)\,.$$

Proof. Let $Z \coloneqq \int_{[-1,1]^d} \exp(-V)$ denote the normalizing constant, let $\theta_1^\star \in [-1,1]$ be the maximizer of π_1, and let $\theta \in [-1,1]^d$. We can write $\theta = (\theta_1, \theta_{-1})$, where $\theta_{-1} \in \mathbb{R}^{d-1}$.[5] Then,

$$\begin{aligned} V(\theta) &\leq V(\theta_1^\star, \theta_{-1}) + \partial_1 V(\theta_1^\star, \theta_{-1})\,(\theta_1 - \theta_1^\star) + \frac{\beta}{2}\,(\theta_1 - \theta_1^\star)^2 \\ &\leq V(\theta_1^\star, \theta_{-1}) + L\,|\theta_1 - \theta_1^\star| + \frac{\beta}{2}\,(\theta_1 - \theta_1^\star)^2 \end{aligned}$$

[5] In §9.5.1 we used the notation $\theta[i]$ for the ith coordinate of θ, but for the sake of simplicity we switch to the notation θ_i for this proof.

$$\leq \frac{1}{2} + V(\theta_1^\star, \theta_{-1}) + \frac{\beta + L^2}{2} (\theta_1 - \theta_1^\star)^2 .$$

This yields the lower bound

$$\begin{aligned}
\pi_1(\theta_1) &= \frac{1}{Z} \int_{[-1,1]^{d-1}} \exp\bigl(-V(\theta_1, \theta_{-1})\bigr) \, \mathrm{d}\theta_{-1} \\
&\geq \frac{\exp[-(\beta + L^2)(\theta_1 - \theta_1^\star)^2/2 - 1/2]}{Z} \int_{[-1,1]^{d-1}} \exp\bigl(-V(\theta_1^\star, \theta_{-1})\bigr) \, \mathrm{d}\theta_{-1} \\
&= \exp\biggl[-\frac{1}{2}(\beta + L^2)(\theta_1 - \theta^\star)^2 - \frac{1}{2}\biggr] \sup_{[-1,1]} \pi_1 .
\end{aligned}$$

Next,

$$\begin{aligned}
1 &= \int_{[-1,1]} \pi_1(\theta_1) \, \mathrm{d}\theta_1 \geq \frac{\sup_{[-1,1]} \pi_1}{\sqrt{\mathrm{e}}} \int_{[-1,1]} \exp\bigl[-\frac{1}{2}(\beta + L^2)(\theta_1 - \theta^\star)^2\bigr] \, \mathrm{d}\theta_1 \\
&\geq \frac{\sup_{[-1,1]} \pi_1}{\sqrt{\mathrm{e}}} \int_0^1 \exp\bigl[-\frac{1}{2}(\beta + L^2) x^2\bigr] \, \mathrm{d}x .
\end{aligned}$$

Let $c \coloneqq \int_0^1 \exp(-x^2) \, \mathrm{d}x$. By splitting into the two cases $\beta + L^2 \leq 1$ and $\beta + L^2 \geq 1$, we can deduce the inequality

$$1 \geq \frac{c \sup_{[-1,1]} \pi_1}{\sqrt{\mathcal{E}}} \Bigl(\frac{1}{\sqrt{\beta + L^2}} \wedge 1\Bigr) .$$

It yields

$$\sup_{[-1,1]} \pi_1 \leq \frac{\sqrt{\mathrm{e}}}{c} \Bigl(\sqrt{\beta + L^2} \vee 1\Bigr) \leq \frac{\sqrt{\mathrm{e}}}{c} \bigl((\sqrt{\beta} + L) \vee 1\bigr) \leq \frac{\sqrt{\mathrm{e}}}{c} (1 + \sqrt{\beta} + L) ,$$

which is the result. □

Proof of Lemma 9.5.1. Let π_i denote the i-th marginal of π. Then, since $\phi(0) = 0$ and $\nabla\phi(0) = 0$, we must estimate

$$\begin{aligned}
\mathcal{D}_\phi(\pi \parallel \mu_0) &= \int_{[-1,1]^d} \sum_{i=1}^d \ln \frac{1}{1 - \theta[i]^2} \, \pi(\theta) \, \mathrm{d}\theta = \sum_{i=1}^d \int_{[-1,1]} \ln \frac{1}{1 - \theta[i]^2} \, \pi_i(\theta[i]) \, \mathrm{d}\theta[i] \\
&\leq C (1 + \sqrt{\beta} + L) \, d \int_{[-1,1]} \ln \frac{1}{1 - x^2} \, \mathrm{d}x \\
&\leq \frac{3}{2} C (1 + \beta + L) \, d \int_{[-1,1]} \ln \frac{1}{1 - x^2} \, \mathrm{d}x ,
\end{aligned}$$

where C is the constant from the proof of Lemma 9.5.3. It yields the result. □

9.6 Proof of the convergence rates

9.6.1 Per-iteration progress bound

For the convergence rates of MLMC, we first prove the following per-iterate progress bound, from which Theorems 9.3.5 and 9.3.6 will be easily deduced.

Lemma 9.6.1 (Per-iteration progress bound)**.** *Assume $\beta > 0$. For $0 \le h \le \frac{1}{\beta}$, let $X_k \sim \mu_k$ be the iterates of MLMC with step size h. Then, under Assumptions 9.3.1-9.3.3, the following holds:*

$$h\,\mathsf{KL}(\mu_{k+1} \,\|\, \pi) \le (1 - \alpha h)\,\mathcal{D}_\phi(\pi \,\|\, \mu_k) - \mathcal{D}_\phi(\pi \,\|\, \mu_{k+1}) + (\beta + 2M_\phi L)\,dh^2\,. \quad (9.7)$$

Proof. We decompose the KL divergence into two parts:

$$\mathsf{KL}(\mu \,\|\, \pi) = \underbrace{\int_{\mathcal{Q}} V(x)\,\mathrm{d}\mu(x)}_{=:\mathcal{E}(\mu)} + \underbrace{\int_{\mathcal{Q}} \mu(x) \ln \mu(x)\,\mathrm{d}x}_{=:\mathcal{H}(\mu)}\,.$$

Here and throughout the paper, we abuse notation by identifying a measure μ with its density.

The first term above has the interpretation of *energy*, while the second term has the interpretation of (negative) *entropy*. The basic scheme of the proof follows the method in [DMM19], which views the two steps of the update rule MLMC as alternately dissipating the energy and the entropy. More specifically, we will show that MLMC:1 dissipates $\mathcal{E}$ and MLMC:2 dissipates $\mathcal{H}$, while the two steps do not badly interfere with each other.

Our analysis proceeds by controlling each term in the following decomposition:

$$\begin{aligned}\mathsf{KL}(\mu_{k+1} \,\|\, \pi) &= \mathcal{E}(\mu_{k+1}) + \mathcal{H}(\mu_{k+1}) - \mathcal{E}(\pi) - \mathcal{H}(\pi) \\ &= \underbrace{\mathcal{E}(\mu_{k+1/2}) - \mathcal{E}(\pi)}_{①} + \underbrace{\mathcal{E}(\mu_{k+1}) - \mathcal{E}(\mu_{k+1/2})}_{②} + \underbrace{\mathcal{H}(\mu_{k+1}) - \mathcal{H}(\pi)}_{③}\,.\end{aligned}$$

Before we go into the analysis of each term, we outline our proof strategy. Term ① corresponds to a deterministic step of the mirror descent algorithm, and we adapt the analysis of mirror descent based on the Bregman proximal inequality [CT93, Lemma 3.2].

For terms ② and ③, it will be important to understand the stochastic process $(Z_t)_{t\in[0,h]}$ in MLMC, where $Z_t \coloneqq \nabla\phi^\star(W_t)$, along with the corresponding marginal laws $(\nu_t)_{t\in[0,h]}$. There are two important and distinct perspectives we can adopt. On one hand, the stochastic process $(Z_t)_{t\in[0,h]}$ is a diffusion, and can be studied

via stochastic calculus. On the other hand, the laws $(\nu_t)_{t\in[0,h]}$ follow a Wasserstein "mirror flow" of the entropy functional $\mathcal{H}$, in the sense that it evolves continuously in Wasserstein space with tangent vector $-[\nabla^2\phi]^{-1}\nabla_{W_2}\mathcal{H}(\nu_t)$ (see §9.2.1 for a brief introduction to Wasserstein calculus, and §8 for a discussion of MLD from this perspective). In turn, these two perspectives offer different calculation rules: stochastic calculus provides Itô's formula (see [Le 16, Theorem 5.10] or [Str18, §3.3]), while Wasserstein calculus provides the rule

$$\partial_t\mathcal{F}(\nu_t) = -\,\mathbb{E}\langle\nabla_{W_2}\mathcal{F}(\nu_t)(Z_t), [\nabla^2\phi(Z_t)]^{-1}\nabla_{W_2}\mathcal{H}(\nu_t)(Z_t)\rangle\,,$$

for any sufficiently well-behaved functional $\mathcal{F}$ on Wasserstein space. Both of these perspectives are insightful, and we will employ both.

For term ②, we show that MLMC:2 does not greatly increase the energy, and we accomplish this via calculations using Itô's formula together with the relative smoothness and self-concordance assumptions. Finally, we control term ③ by developing a new displacement convexity result (Theorem 9.4.1) for the entropy functional $\mathcal{H}$, which is crucial for applying Wasserstein calculus.

①: Let Y be a random variable (defined on the same probability space) which is distributed according to π. Then,

$$\begin{aligned}
\mathcal{E}(\mu_{k+1/2}) - \mathcal{E}(\pi) &= \mathbb{E}[V(X_{k+1/2})] - \mathbb{E}[V(Y)] \\
&= \mathbb{E}[V(X_{k+1/2})] - \mathbb{E}[V(X_k)] + \mathbb{E}[V(X_k)] - \mathbb{E}[V(Y)] \\
&\le \mathbb{E}[\langle\nabla V(X_k), X_{k+1/2} - X_k\rangle + \beta D_\phi(X_{k+1/2}\,\|\,X_k)] \\
&\quad + \mathbb{E}[\langle\nabla V(X_k), X_k - Y\rangle - \alpha D_\phi(Y\,\|\,X_k)] \\
&= \mathbb{E}[\langle\nabla V(X_k), X_{k+1/2} - Y\rangle \\
&\qquad + \beta D_\phi(X_{k+1/2}\,\|\,X_k) - \alpha D_\phi(Y\,\|\,X_k)]\,, \qquad (9.8)
\end{aligned}$$

where the inequality follows due to the α-relative strong convexity and β-relative smoothness of V. Now to control (9.8), we invoke a standard tool from optimization:

Lemma 9.6.2 (Bregman proximal inequality [CT93, Lemma 3.2])**.** *For a convex function f and a convex function ϕ of Legendre type, suppose that*

$$x_+ := \operatorname*{arg\,min}_{z\in\mathcal{Q}} [f(z) + D_\phi(z\,\|\,x)]\,.$$

Then,

$$f(x_+) - f(y) \le D_\phi(y\,\|\,x) - D_\phi(y\,\|\,x_+) - D_\phi(x_+\,\|\,x) \qquad \forall y\in\mathcal{Q}\,.$$

Applying the Bregman proximal inequality (Lemma 9.6.2) with $f(x) = h\langle \nabla V(X_k), x\rangle$,

$$\begin{aligned}(9.8) &\le \mathbb{E}\Big[\big(\frac{1}{h}-\alpha\big)\, D_\phi(Y \,\|\, X_k) - \frac{1}{h}\, D_\phi(Y \,\|\, X_{k+1/2}) \\ &\qquad + \big(\beta - \frac{1}{h}\big)\, D_\phi(X_{k+1/2} \,\|\, X_k)\Big] \\ &\le \mathbb{E}\Big[\big(\frac{1}{h}-\alpha\big)\, D_\phi(Y \,\|\, X_k) - \frac{1}{h}\, D_\phi(Y \,\|\, X_{k+1/2})\Big]\,,\end{aligned}$$

provided that $\frac{1}{h} \ge \beta \Leftrightarrow h \le \beta^{-1}$. Choosing Y so that the coupling (Y, X_k) minimizes $\mathbb{E}[D_\phi(Y \,\|\, X_k)]$, we obtain

$$\begin{aligned}h\,\{\mathcal{E}(\mu_{k+1/2}) - \mathcal{E}(\pi)\} &\le (1-\alpha h)\, \mathcal{D}_\phi(\pi \,\|\, \mu_k) - \mathbb{E}[D_\phi(Y \,\|\, X_{k+1/2})] \\ &\le (1-\alpha h)\, \mathcal{D}_\phi(\pi \,\|\, \mu_k) - \mathcal{D}_\phi(\pi \,\|\, \mu_{k+1/2})\,.\end{aligned}$$

(2): First, note from MLMC that

$$\mathcal{E}(\mu_{k+1}) - \mathcal{E}(\mu_{k+1/2}) = \mathbb{E}\big[V\big(\nabla\phi^\star(W_h)\big) - V\big(\nabla\phi^\star(W_0)\big)\big]\,.$$

To compute the above term, we define $f(x) \coloneqq V(\nabla\phi^\star(x))$ and apply Itô's formula to the random variable $f(W_h) - f(W_0)$. To that end, we first compute the Hessian of f:

$$\begin{aligned}\nabla f &= \nabla V(\nabla\phi^\star)^{\mathsf{T}} \nabla^2\phi^\star = \nabla V(\nabla\phi^\star)^{\mathsf{T}} [\nabla^2\phi(\nabla\phi^\star)]^{-1}\,, \\ \nabla^2 f &= \nabla^2 V(\nabla\phi^\star)\, [\nabla^2\phi(\nabla\phi^\star)]^{-1}\, [\nabla^2\phi^\star] \\ &\qquad + \nabla V(\nabla\phi^\star)^{\mathsf{T}}\, [\nabla^2\phi(\nabla\phi^\star)]^{-1}\, [\nabla^3\phi(\nabla\phi^\star)]\, [\nabla^2\phi(\nabla\phi^\star)]^{-2}\,.\end{aligned}$$

Itô's formula now decomposes $f(W_h) - f(W_0)$ into the sum of an integral and a stochastic integral. Intuitively, the stochastic integral has mean zero (since it is a local martingale), and this can be rigorously argued using the standard technique of localization; we give the argument at the end of this step. Thus, we concentrate on the expectation of the first term. Writing $Z_t \coloneqq \nabla\phi^\star(W_t)$, the above Hessian calculation gives

$$\begin{aligned}&\mathbb{E}[f(W_h) - f(W_0)] \qquad (9.9)\\ &\quad = \mathbb{E}\int_0^h \big\langle \nabla^2 V(Z_t)\, [\nabla^2\phi(Z_t)]^{-2}, \nabla^2\phi(Z_t)\big\rangle\, \mathrm{d}t \\ &\qquad + \mathbb{E}\int_0^h \big\langle \nabla V(Z_t)^{\mathsf{T}}\, [\nabla^2\phi(Z_t)]^{-1}\, [\nabla^3\phi(Z_t)]\, [\nabla^2\phi(Z_t)]^{-2}, \nabla^2\phi(Z_t)\big\rangle\, \mathrm{d}t\end{aligned}$$

$$= \mathbb{E} \int_0^h \left\langle \nabla^2 V(Z_t), [\nabla^2 \phi(Z_t)]^{-1} \right\rangle \mathrm{d}t \tag{9.10}$$

$$+ \mathbb{E} \int_0^h \mathrm{tr}\big(\nabla V(Z_t)^\mathsf{T} [\nabla^2 \phi(Z_t)]^{-1} [\nabla^3 \phi(Z_t)] [\nabla^2 \phi(Z_t)]^{-1}\big) \, \mathrm{d}t \,. \tag{9.11}$$

We can control (9.10) easily based on the relative smoothness of V: indeed, since $\nabla^2 V \preceq \beta \nabla^2 \phi$ (see Proposition 9.2.6),

$$(9.10) \le \beta d h \,.$$

To control (9.11), we use the self-concordance of ϕ. We recall here the following result:

Proposition 9.6.3 ([Nes18, Corollary 5.1.1])**.** *A function ϕ is self-concordant with a constant $M_\phi \ge 0$ if and only if for any $x \in \mathrm{dom}(\phi)$ and any direction $u \in \mathbb{R}^n$ we have*

$$\nabla^3 \phi(x)\, u \preceq 2 M_\phi \, \|u\|_{\nabla^2 \phi(x)} \, \nabla^2 \phi(x) \,.$$

Using Proposition 9.6.3, it follows that

$$\begin{aligned} &\frac{1}{2M_\phi} \times (9.11) \\ &\quad \le \int_0^h \mathbb{E}\big[\big\| [\nabla^2 \phi(Z_t)]^{-1} \nabla V(Z_t) \big\|_{\nabla^2 \phi(Z_t)} \, \mathrm{tr}\big([\nabla^2 \phi(Z_t)] \, [\nabla^2 \phi(Z_t)]^{-1} \big) \big] \, \mathrm{d}t \\ &\quad \le d \int_0^h \mathbb{E}\big[\|\nabla V(Z_t)\|_{[\nabla^2 \phi(Z_t)]^{-1}} \big] \, \mathrm{d}t \le 2 M_\phi L d h \,. \end{aligned}$$

Thus, our calculation shows that

$$\mathcal{E}(\mu_{k+1}) - \mathcal{E}(\mu_{k+1/2}) \le (\beta + 2 M_\phi L) \, dh \,. \tag{9.12}$$

We now sketch the localization argument. Let $(\tau_\ell)_{\ell \in \mathbb{N}}$ be a localizing sequence for $(W_t)_{t \in [0,h]}$. The argument above may be applied rigorously for the stopped process $(W_{t \wedge \tau_\ell})_{t \in [0,h]}$ to obtain $\mathbb{E}\, V(Z_{h \wedge \tau_\ell}) - \mathbb{E}\, V(Z_0) \le (\beta + 2 M_\phi L)\, dh$. Since V is bounded below, we use Fatou's lemma to pass $\ell \to \infty$ and deduce (9.12).

(3): Let ν_t denote the law of $Z_t \coloneqq \nabla \phi^\star(W_t)$. For this step, we calculate the derivative of $t \mapsto \mathcal{D}_\phi(\pi \parallel \nu_t)$. Noting that $\nabla_2 D_\phi(y \parallel x) = -\nabla^2 \phi(x)\,(y - x)$

and that ν_t follows the Wasserstein tangent vector $-[\nabla^2\phi]^{-1}\nabla_{W_2}\mathcal{H}(\nu_t)$, we expect that

$$\begin{aligned}\partial_t\mathcal{D}_\phi(\pi\,\|\,\nu_t) &= \mathbb{E}\big\langle[\nabla^2\phi(Z_t)]^{-1}\nabla_{W_2}\mathcal{H}(\nu_t)(Z_t), \nabla^2\phi(Z_t)\,(Y-Z_t)\big\rangle \\ &= \mathbb{E}\langle\nabla_{W_2}\mathcal{H}(\nu_t)(Z_t), Y-Z_t\rangle\,,\end{aligned}$$

where (Y, Z_t) are optimally coupled for π and ν_t for the Bregman transport cost. In general, the differentiability properties of optimal transport costs can be quite subtle, but thankfully it is much easier to establish the superdifferentiability

$$\partial_t^+\mathcal{D}_\phi(\pi\,\|\,\nu_t) \le \mathbb{E}\langle\nabla_{W_2}\mathcal{H}(\nu_t)(Z_t), Y-Z_t\rangle$$

at almost all t, which is all that will be needed for the subsequent argument. The superdifferentiability result is proven along the lines of [OV00, Lemma 2]; see also [AGS08, Theorem 10.2.2] or the proof of [Vil09b, Theorem 23.9].

Next, we apply a result which can be interpreted as convexity of the entropy functional with respect to the Bregman divergence, given as Theorem 9.4.1. It implies that for $t \in [0, h]$,

$$\begin{aligned}\partial_t^+\mathcal{D}_\phi(\pi\,\|\,\nu_t) &\le \mathbb{E}\langle\nabla_{W_2}\mathcal{H}(\nu_t)(Z_t), Y-Z_t\rangle \le \mathcal{H}(\pi)-\mathcal{H}(\nu_t) \\ &\le \mathcal{H}(\pi)-\mathcal{H}(\nu_h)\,,\end{aligned}$$

where the last inequality follows since

$$\partial_t\mathcal{H}(\nu_t) = -\,\mathbb{E}\big[\big\langle\nabla_{W_2}\mathcal{H}(\nu_t)(Z_t), [\nabla^2\phi(Z_t)]^{-1}\nabla_{W_2}\mathcal{H}(\nu_t)(Z_t)\big\rangle\big] \le 0\,,$$

which implies $\mathcal{H}(\nu_h) \le \mathcal{H}(\nu_t)$ for any $t \in [0, h]$. Integrating from 0 to h,

$$\mathcal{D}_\phi(\pi\,\|\,\nu_h) - \mathcal{D}_\phi(\pi\,\|\,\nu_0) \le h\,\{\mathcal{H}(\pi)-\mathcal{H}(\nu_h)\}\,,$$

which is the same as

$$h\,\{\mathcal{H}(\mu_{k+1})-\mathcal{H}(\pi)\} \le \mathcal{D}_\phi(\pi\,\|\,\mu_{k+1/2}) - \mathcal{D}_\phi(\pi\,\|\,\mu_{k+1})\,.$$

Combining the bounds from ①, ②, and ③, the proof is complete. □

■ 9.6.2 Proof of Theorem 9.3.5

From the per-iteration progress bound (Lemma 9.6.1), we have for any $k \in \mathbb{N}$

$$h\,\mathsf{KL}(\mu_{k+1}\,\|\,\pi) \le \mathcal{D}_\phi(\pi\,\|\,\mu_k) - \mathcal{D}_\phi(\pi\,\|\,\mu_{k+1}) + \beta' d h^2\,. \tag{9.13}$$

Summing (9.13) over $k = 0, 1, \ldots, N-1$,

$$h \sum_{k=1}^{N} \mathsf{KL}(\mu_k \,\|\, \pi) \le \mathcal{D}_\phi(\pi \,\|\, \mu_0) - \mathcal{D}_\phi(\pi \,\|\, \mu_N) + \beta' d h^2 N \,.$$

Using the convexity of the KL divergence [which follows from the Gibbs variational principle; see RS15, §5.1],

$$\mathsf{KL}(\overline{\mu}_N \,\|\, \pi) \le \frac{\mathcal{D}_\phi(\pi \,\|\, \mu_0)}{Nh} + \beta' d h \le \frac{\varepsilon^2}{2} + \frac{\varepsilon^2}{2} \,,$$

where the last inequality follows from the choice $N \ge \frac{2\mathcal{D}_\phi(\pi\|\mu_0)}{h\varepsilon^2}$ and $h \le \frac{\varepsilon^2}{2\beta d}$.

■ 9.6.3 Proof of Theorem 9.3.6

Let us first prove the convergence in Bregman transport cost. For any $k \in \mathbb{N}$, the per-iteration progress bound (Lemma 9.6.1) together with $\mathsf{KL}(\mu_{k+1} \,\|\, \pi) \ge 0$ imply

$$\mathcal{D}_\phi(\pi \,\|\, \mu_{k+1}) \le (1 - \alpha h)\, \mathcal{D}_\phi(\pi \,\|\, \mu_k) + \beta' d h^2 \,. \tag{9.14}$$

Recursively applying (9.14) for $k = 0, 1, \ldots, N-1$, we obtain

$$\begin{aligned}
\mathcal{D}_\phi(\pi \,\|\, \mu_N) &\le (1 - \alpha h)^N\, \mathcal{D}_\phi(\pi \,\|\, \mu_0) + \beta' d h^2 \sum_{k=0}^{N-1} (1 - \alpha h)^k \\
&\le (1 - \alpha h)^N\, \mathcal{D}_\phi(\pi \,\|\, \mu_0) + \beta' d h^2 \sum_{k=0}^{\infty} (1 - \alpha h)^k \\
&\le \exp(-\alpha h N)\, \mathcal{D}_\phi(\pi \,\|\, \mu_0) + \frac{\beta' d h}{\alpha} \le \frac{\varepsilon^2}{2} + \frac{\varepsilon^2}{2} \,,
\end{aligned}$$

where the last inequality follows since $N \ge \frac{1}{\alpha h} \ln \frac{2\mathcal{D}_\phi(\pi\|\mu_0)}{\varepsilon^2}$ and $h \le \frac{\alpha\varepsilon^2}{2\beta' d}$. Having proved the convergence in terms of the Bregman transport cost, the convergence in terms of the KL divergence follows by applying Theorem 9.3.5.

■ 9.6.4 Analysis for the non-smooth case (Theorem 9.3.7)

The analysis for the non-smooth case proceeds in a similar manner to the smooth case. We first prove the following per-iterate progress bound.

Lemma 9.6.4 (Per-iteration progress bound; non-smooth case)**.** *Let $X_k \sim \mu_k$ be the iterates of* MLMC *with step size $h > 0$. Assume that ϕ is 1-strongly convex w.r.t $\|\cdot\|$, and that V is convex and $\tilde{L}$-Lipschitz w.r.t $\|\cdot\|$. Then,*

$$h\, \mathsf{KL}(\mu_{k+1} \,\|\, \pi) \le \mathcal{D}_\phi(\pi \,\|\, \mu_{k+1/2}) - \mathcal{D}_\phi(\pi \,\|\, \mu_{k+3/2}) + \frac{h^2 \tilde{L}^2}{2} \,. \tag{9.15}$$

Proof. Our analysis proceeds by controlling each term in the decomposition

$$\mathsf{KL}(\mu_{k+1} \,\|\, \pi) = \underbrace{\mathcal{E}(\mu_{k+1}) - \mathcal{E}(\pi)}_{\text{(A)}} + \underbrace{\mathcal{H}(\mu_{k+1}) - \mathcal{H}(\pi)}_{\text{(B)}} .$$

For term (B), we invoke the following upper bound (from the analysis of term (3) in the proof of Lemma 9.6.1):

$$h\left\{\mathcal{H}(\mu_{k+1}) - \mathcal{H}(\pi)\right\} \le \mathcal{D}_\phi(\pi \,\|\, \mu_{k+1/2}) - \mathcal{D}_\phi(\pi \,\|\, \mu_{k+1}) . \tag{9.16}$$

Let us turn to (A), and let Y be a random variable (defined on the same probability space) which is distributed according to π. Since we have

$$X_{k+3/2} = \arg\min_{x\in\mathcal{Q}} \left[\langle h\,\nabla V(X_{k+1}), x\rangle + D_\phi(x \,\|\, X_{k+1})\right] ,$$

applying the Bregman proximal inequality (Lemma 9.6.2) with the choice $f(x) = h\,\langle \nabla V(X_{k+1}), x\rangle$ gives

$$\begin{aligned} &h\,\langle \nabla V(X_{k+1}), X_{k+3/2} - Y\rangle \\ &\quad \le D_\phi(Y \,\|\, X_{k+1}) - D_\phi(Y \,\|\, X_{k+3/2}) - D_\phi(X_{k+3/2} \,\|\, X_{k+1}) , \end{aligned}$$

which after rearranging becomes

$$\begin{aligned} &D_\phi(Y \,\|\, X_{k+3/2}) - D_\phi(Y \,\|\, X_{k+1}) \\ &\quad \le h\,\langle \nabla V(X_{k+1}), Y - X_{k+3/2}\rangle - D_\phi(X_{k+3/2} \,\|\, X_{k+1}) . \end{aligned} \tag{9.17}$$

On the other hand, the right hand side of (9.17) can be controlled using the convexity, the Lipschitzness of V, and strong convexity of ϕ:

$$\begin{aligned} &\text{RHS of (9.17)} \\ &\quad = h\,\langle \nabla V(X_{k+1}), Y - X_{k+1}\rangle + h\,\langle \nabla V(X_{k+1}), X_{k+1} - X_{k+3/2}\rangle \\ &\qquad - D_\phi(X_{k+3/2} \,\|\, X_{k+1}) \\ &\quad \le h\,[V(Y) - V(X_{k+1})] + h\,\|\nabla V(X_{k+1})\|_\star\,\|X_{k+1} - X_{k+3/2}\| \\ &\qquad - \frac{1}{2}\,\|X_{k+1} - X_{k+3/2}\|^2 \\ &\quad \le h\,[V(Y) - V(X_{k+1})] + \frac{h^2}{2}\,\|\nabla V(X_{k+1})\|_\star^2 \\ &\quad \le h\,[V(Y) - V(X_{k+1})] + \frac{h^2\tilde{L}^2}{2} . \end{aligned}$$

For the LHS of (9.17), choose Y so that the coupling (Y, X_{k+1}) minimizes the cost $\mathbb{E}[D_\phi(Y \,\|\, X_{k+1})]$ to obtain

$$\begin{aligned}\mathbb{E}[\text{LHS of (9.17)}] &= \mathbb{E}[D_\phi(Y \,\|\, X_{k+3/2})] - \mathcal{D}_\phi(\pi \,\|\, \mu_{k+1}) \\ &\geq \mathcal{D}_\phi(\pi \,\|\, \mu_{k+3/2}) - \mathcal{D}_\phi(\pi \,\|\, \mu_{k+1})\,.\end{aligned}$$

Combining these upper and lower bounds, (9.17) becomes:

$$h\,[\mathcal{E}(\mu_{k+1}) - \mathcal{E}(\pi)] \leq \mathcal{D}_\phi(\pi \,\|\, \mu_{k+1}) - \mathcal{D}_\phi(\pi \,\|\, \mu_{k+3/2}) + \frac{h^2\tilde{L}^2}{2}\,.$$

Together with (9.16), the proof is complete. □

Now using Lemma 9.6.4, we prove Theorem 9.3.7.

Proof of Theorem 9.3.7. From Lemma 9.6.4, we have for any $k \in \mathbb{N}$

$$h\,\mathsf{KL}(\mu_{k+1} \,\|\, \pi) \leq \mathcal{D}_\phi(\pi \,\|\, \mu_{k+1/2}) - \mathcal{D}_\phi(\pi \,\|\, \mu_{k+3/2}) + \frac{h^2\tilde{L}^2}{2}\,. \tag{9.18}$$

Summing (9.18) over $k = 0, 1, \ldots, N-1$,

$$h\sum_{k=1}^{N} \mathsf{KL}(\mu_k \,\|\, \pi) \leq \mathcal{D}_\phi(\pi \,\|\, \mu_{1/2}) - \mathcal{D}_\phi(\pi \,\|\, \mu_{N+1/2}) + \frac{h^2\tilde{L}^2}{2}\,N\,.$$

Again using the convexity of the KL divergence, we obtain

$$\mathsf{KL}(\overline{\mu}_N \,\|\, \pi) \leq \frac{\mathcal{D}_\phi(\pi \,\|\, \mu_{1/2})}{Nh} + \frac{h\tilde{L}^2}{2} \leq \frac{\varepsilon^2}{2} + \frac{\varepsilon^2}{2}\,,$$

where the last inequality follows from the choice $N \geq \frac{2\mathcal{D}_\phi(\pi\|\mu_{1/2})}{h\varepsilon^2}$ and $h \leq \frac{\varepsilon^2}{\tilde{L}^2}$. □

■ 9.7 Proofs for the convexity of entropy

To prove Theorem 9.4.1, we will use the known result about the convexity of $\mathcal{H}$ along generalized geodesics [AGS08, Theorem 9.4.11]. To that end, the first step is to obtain a characterization of the optimal Bregman transport coupling which is analogous to Brenier's theorem. The following theorem is of independent interest:

Theorem 9.7.1 (Brenier's theorem for the Bregman transport cost)**.** *Let μ, ν be probability measures on $\mathbb{R}^d$. The optimal Bregman transport coupling (X, Y) for μ and ν is of the form*

$$\nabla\phi(X) - \nabla\phi(Y) = \nabla h(X)\,,$$

where $h : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\}$ is such that $\phi - h$ is convex.

Proof. From the general theory of optimal transport duality, it holds that

$$\nabla_1 D_\phi(X, Y) = \nabla h(X)\,,$$

where h is a D_ϕ-concave function [see Vil09b, Theorem 10.28].[6] The left-hand side of this equation evaluates to $\nabla\phi(X) - \nabla\phi(Y)$, so we simply have to check that D_ϕ-concavity of h implies that $\phi - h$ is convex (which is in fact equivalent to saying that h is 1-relatively smooth with respect to ϕ, see Proposition 9.2.6).

Recall that the D_ϕ-concavity of h means there exists a function $\tilde{h} : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\}$ such that

$$h(x) = \inf_{y \in \mathbb{R}^d} \{D_\phi(x, y) - \tilde{h}(y)\}\,,$$

see [Vil09b, Definition 5.2].[7] If we expand out the definition of the Bregman divergence, we can rewrite this as

$$\phi(x) - h(x) = \sup_{y \in \mathbb{R}^d} \{\langle \nabla\phi(y), x - y\rangle + \tilde{h}(y) + \phi(y)\}\,. \tag{9.19}$$

As a supremum of affine functions, we see that $\phi - h$ is convex, which completes the proof. □

We now prove Theorem 9.4.1 using Theorem 9.7.1:

Proof of Theorem 9.4.1. By Theorem 9.7.1, the optimal Bregman transport coupling is of the form $\nabla\phi(Y) = \nabla(\phi - h)(X) = \nabla\zeta(X)$, where we have defined the convex function $\zeta := \phi - h$. Hence, letting $\overline{\nu}$ denote the law of $\overline{Y} := \nabla\phi(Y)$, it follows that $(X, \nabla\zeta(X))$ is a W_2 optimal coupling between μ and $\overline{\nu}$. Furthermore, since ϕ is a convex function of Legendre type, $(\nabla\zeta(X), \nabla\phi^\star \circ \nabla\zeta(X))$ is also a W_2 optimal coupling between $\overline{\nu}$ and ν. Noting that

$$\nabla\phi^\star \circ \nabla\zeta(X) = \nabla\phi^\star \circ \nabla\phi(Y) = Y\,,$$

it follows that (X, Y) is a generalized geodesic according to W_2. Therefore, the convexity of $\mathcal{H}$ along generalized geodesics [AGS08, Theorem 9.4.11] concludes the proof (see [SR20, Lemma 4]). □

[6] In fact, there are three assumptions for [Vil09b, Theorem 10.28]. Here, we explicitly check them one by one for clarity. (i) *Super-differentiability:* D_ϕ is clearly differentiable on $\mathcal{Q}$ as ϕ is of class C^3. (ii) *Injectivity of gradient:* $\nabla_1 D_\phi(x, \cdot) = \nabla\phi(x) - \nabla\phi(\cdot)$ is injective as ϕ is of Legendre type. (iii) *μ-almost-sure differentiability of D_ϕ-concave functions:* In (9.19), we actually show that for any D_ϕ-concave function h, $\phi - h$ is convex and thus differentiable Lebesgue a.e. [Roc97, Theorem 25.5]. Since μ is absolutely continuous w.r.t. Lebesgue measure and ϕ is differentiable, h must be differentiable μ-almost surely.

[7] In Villani's book, he works with the definition of *c-convexity* rather than *c-concavity*, but this is merely a matter of convention; c.f. [Vil03, §2.4] for the conventions regarding c-concavity.

Remark 9.7.2. *For reader's convenience, we provide a direct calculation that (formally) shows the convexity result. Since we have shown* $Y = \nabla\phi^\star \circ \nabla\zeta(X)$*, the change of variable formula gives*

$$\begin{aligned}\mathcal{H}(\nu) = \int \nu(y) \ln \nu(y) \,\mathrm{d}y &= \int \mu(x) \ln \nu\big(\nabla\phi^\star \circ \nabla\zeta(x)\big) \,\mathrm{d}x \\ &= \int \mu(x) \ln \frac{\mu(x)}{\det \nabla(\nabla\phi^\star \circ \nabla\zeta)(x)} \,\mathrm{d}x.\end{aligned}$$

Here the change of variables is valid since

$$\det\big([\nabla(\nabla\phi^\star \circ \nabla\zeta)](x)\big) = \det\big([\nabla^2\phi^\star(\nabla\zeta(x))]\,[\nabla^2\zeta(x)]\big) > 0\,.$$

Thus, using the convexity of $-\ln\det$ *and integrating by parts, we obtain*

$$\begin{aligned}\mathcal{H}(\mu) - \mathcal{H}(\nu) &= -\int \mu(x) \ln\det \nabla(\nabla\phi^\star \circ \nabla\zeta)(x) \,\mathrm{d}x \\ &\geq -\int \mu(x) \,\mathrm{tr}[\nabla(\nabla\phi^\star \circ \nabla\zeta)(x) - I_d] \,\mathrm{d}x \\ &= \int \langle \nabla\mu(x), (\nabla\phi^\star \circ \nabla\zeta)(x) - x\rangle \,\mathrm{d}x \\ &= \int \langle \nabla \ln \mu(x), (\nabla\phi^\star \circ \nabla\zeta)(x) - x\rangle \,\mathrm{d}\mu(x)\,.\end{aligned}$$

Recalling that $\nabla_{W_2}\mathcal{H}(\mu) = \nabla \ln \mu$ *and* $Y = \nabla\phi^\star(\nabla\zeta(X))$*, the result follows.*

■ 9.8 Conclusion

We conclude by discussing some questions for future research.

1. As we discuss in Remark 9.3.12, it is an open question to determine if the analyses of [Zha+20; Li+22] can be improved to obtain vanishing bias for the Euler–Maruyama discretization of MLD under weaker assumptions.

   We remark that [Jia21] also obtained similar conclusions as our work, namely that the Euler–Maruyama discretization potentially incurs non-vanishing bias, whereas MLMC does not. See also the recent work of [GV22].

2. In our work, we analyze the sampling analogue of mirror descent under the assumption that the mirror map is self-concordant. This notably bears resemblance to the development of interior-point methodology in optimization [NN94], and it is an interesting problem to develop further sampling analogues of interior-point algorithms.

3. In §9.5.2, we investigated the possibility that MLMC can alleviate the dependence on dimension for some sampling problems. However, Metropolis-adjusted variants of the Langevin algorithm enjoy significantly better dependence on the dimension as compared to their unadjusted counterparts; see [RR98; PST12] and §5. Thus, the Metropolis-adjusted version of MLMC may be a more appropriate setting in which to investigate this dimension reduction question, which we leave to future work.

Chapter 10

Interlude: two applications of Brascamp–Lieb inequalities

We saw in §8 that the Brascamp–Lieb inequality is key for establishing convergence of the mirror Langevin diffusion. In turn, the geometry underlying the mirror Langevin diffusion can be used to recover the Brascamp–Lieb inequality (Corollary 9.4.2). In this chapter, we explore two further applications of the Brascamp–Lieb inequality to optimization and optimal transport.

For any convex body $K \subseteq \mathbb{R}^n$, S. Bubeck and R. Eldan introduced the entropic barrier on K in [BE19] and showed that it is a $(1 + o(1))\, n$-self-concordant barrier. In §10.1, we prove that the optimal bound of n on the self-concordance parameter holds as a consequence of the dimensional Brascamp–Lieb inequality. This is based on [Che21b].

The optimal transport map between the standard Gaussian measure and an α-strongly log-concave probability measure is $\alpha^{-1/2}$-Lipschitz, as first observed in a celebrated theorem of Caffarelli. In §10.2, we apply dual covariance inequalities (the Brascamp–Lieb and Cramér–Rao inequalities) to prove a sharp bound on the Lipschitz constant of the map that arises from *entropically regularized* optimal transport. In the limit as the regularization tends to zero, we obtain an elegant and short proof of Caffarelli's original result. We also extend Caffarelli's theorem to the setting in which the Hessians of the log-densities of the measures are bounded by arbitrary positive definite commuting matrices. This is based on [CP22], joint with Aram-Alexandre Pooladian.

10.1 Optimal self-concordance of the entropic barrier

10.1.1 Introduction

Let $K \subseteq \mathbb{R}^n$ be a convex body. In [BE19], S. Bubeck and R. Eldan introduced the *entropic barrier* $f^\star : \operatorname{int} K \to \mathbb{R}$, defined as follows. First, let $f : \mathbb{R}^n \to \mathbb{R}$

denote the logarithmic Laplace transform of the uniform measure on K,

$$f(\theta) := \ln \int_K \exp \langle \theta, x \rangle \, \mathrm{d}x \,. \tag{10.1}$$

Then, define $f^\star$ to be the Fenchel conjugate of f,

$$f^\star(x) := \sup_{\theta \in \mathbb{R}^n} \{ \langle \theta, x \rangle - f(\theta) \} \,.$$

They proved the following result.

Theorem 10.1.1 ([BE19, Theorem 1])**.** *The function $f^\star$ is strictly convex on* $\operatorname{int} K$*. Also, the following statements hold.*

1. *$f^\star$ is self-concordant, i.e.,*

$$\nabla^3 f^\star(x)[h,h,h] \le 2 \, |\langle h, \nabla^2 f^\star(x) \, h \rangle|^{3/2} \,, \qquad \text{for all } x \in \operatorname{int} K \,,\ h \in \mathbb{R}^n \,.$$

2. *$f^\star$ is a ν-self-concordant barrier, i.e.,*

$$\nabla^2 f^\star(x) \succeq \frac{1}{\nu} \nabla f^\star(x) \, \nabla f^\star(x)^\mathsf{T} \,, \qquad \text{for all } x \in \operatorname{int} K \,,$$

with $\nu = (1 + o(1))\, n$.

Self-concordant barriers are most well-known for their prominent role in the theory of interior-point methods for optimization [NN94], but they also find applications to numerous other problems such as online linear optimization with bandit feedback [AHR08] (indeed, the latter was a motivating example for the introduction of the entropic barrier in [BE19]).

A central theoretical question in the study of self-concordant barriers is: for any convex domain $K \subseteq \mathbb{R}^n$, does there exist a ν-self-concordant barrier for K, and if so, what the optimal value of the parameter ν? In their seminal work [NN94], Y. Nesterov and A. Nemirovskii constructed for each K a *universal barrier* with $\nu = O(n)$. On the other hand, explicit examples (e.g., the simplex and the cube) show that the best possible self-concordance parameter is $\nu = n$ [NN94, Proposition 2.3.6]. The situation was better understood for convex cones, on which the *canonical barrier* was shown to be n-self-concordant independently by R. Hildebrand and D. Fox [Hil14; Fox15]. Then, in [BE19], S. Bubeck and R. Eldan introduced the entropic barrier and showed that it is $(1 + o(1))\, n$-self-concordant on general convex bodies, and n-self-concordant on convex cones; further, they showed that the universal barrier is also n-self-concordant on convex cones. Subsequently, Y. T. Lee and M.-C. Yue settled the question of obtaining optimal self-concordant barriers for general convex bodies by proving that the universal barrier is always n-self-concordant [LY21].

The purpose of this section is to describe the following observation.

Theorem 10.1.2. *The entropic barrier on any convex body $K \subseteq \mathbb{R}^n$ is an n-self-concordant barrier.*

Besides improving the result of [BE19], the theorem shows that the entropic barrier provides a second example of an optimal self-concordant barrier for general convex bodies; to the best of the author's knowledge, no other optimal self-concordant barriers are known.

We will provide two distinct proofs of Theorem 10.1.2. First, we will observe that Theorem 10.1.2 is an immediate consequence of the following theorem, which was obtained independently in [Ngu14; Wan14]; see also [FMW16].

Theorem 10.1.3. *Let $\mu \propto \exp(-V)$ be a log-concave density on $\mathbb{R}^n$. Then,*

$$\operatorname{var}_\mu V \le n\,.$$

In turn, as discussed in [Ngu14; BGG18], Theorem 10.1.3 is related to certain dimensional improvements of the *Brascamp–Lieb inequality*. We state a version of this inequality which is convenient for the present discussion.

Theorem 10.1.4 ([BGG18, Proposition 4.1])**.** *Let $\mu \propto \exp(-V)$ be a log-concave density on $\mathbb{R}^n$, where V is of class $\mathcal{C}^2$ and $\nabla^2 V \succ 0$. Then, for all $\mathcal{C}^1$ compactly supported $g : \mathbb{R}^n \to \mathbb{R}$, it holds that*

$$\operatorname{var}_\mu g \le \mathbb{E}_\mu \langle \nabla g, (\nabla^2 V)^{-1} \nabla g \rangle - \frac{\operatorname{cov}_\mu(g, V)^2}{n - \operatorname{var}_\mu V}\,.$$

It is straightforward to see that Theorem 10.1.4 implies Theorem 10.1.3. Indeed, via a routine approximation argument, we may assume that μ satisfies the hypothesis of Theorem 10.1.4. Taking $g = V$ (which is justified via another approximation argument) and rearranging the inequality of Theorem 10.1.4 yields

$$\operatorname{var}_\mu V \le \frac{n\,\mathbb{E}_\mu \langle \nabla V, (\nabla^2 V)^{-1} \nabla V \rangle}{n + \mathbb{E}_\mu \langle \nabla V, (\nabla^2 V)^{-1} \nabla V \rangle} \le n\,.$$

Next, in our second approach to Theorem 10.1.2, we observe that a key step in the proof of Theorem 10.1.3 given by [Wan14] is a tensorization principle. It is then natural to wonder whether such a principle can be applied directly to deduce Theorem 10.1.2. Indeed, we have the following elementary lemma.

Lemma 10.1.5. *Suppose that for each $n \in \mathbb{N}^+$ and each convex body $K \subseteq \mathbb{R}^n$, we have a function $\phi_{n,K} : \operatorname{int} K \to \mathbb{R}$ such that $\phi_{n,K}$ is a $\nu(n)$-self-concordant barrier for K. Also, suppose that the following consistency condition holds:*

$$\phi_{m+n,K\times K'}(x, x') = \phi_{m,K}(x) + \phi_{n,K'}(x')\,, \tag{10.2}$$

for all $m, n \in \mathbb{N}^+$, all convex bodies $K \subseteq \mathbb{R}^m$, $K' \subseteq \mathbb{R}^n$, and all $x \in K$, $x' \in K'$. Then, $\phi_{n,K}$ is a $\inf_{k\in\mathbb{N}^+} \nu(kn)/k$-self-concordant barrier for K.

We will check that the entropic barrier satisfies the consistency condition described in the previous lemma in §10.1.4. Combined with the second statement in Theorem 10.1.1, it yields another proof of Theorem 10.1.2.

The remainder of this section is organized as follows. In §10.1.2, we will explain the connection between Theorem 10.1.2 and Theorem 10.1.3, thereby deducing the former from the latter. Then, so as to make this section more self-contained, in §10.1.3 we will provide two proofs of the dimensional Brascamp–Lieb inequality (Theorem 10.1.4). The first proof follows [BGG18] and proceeds via a dimensional improvement of Hörmander's L^2 method. The second "proof", which is only sketched, shows how the dimensional Brascamp–Lieb inequality may be obtained from a convexity principle: the entropy functional is convex along generalized Wasserstein geodesics which arise from Bregman divergence couplings (Theorem 9.4.1). The second argument appears to be new. Finally, in §10.1.4, we present the tensorization argument as encapsulated in Lemma 10.1.5.

■ 10.1.2 From the entropic barrier to the dimensional Brascamp–Lieb inequality

In this section, we follow [BE19]. The entropic barrier has a fruitful interpretation in terms of an exponential family of probability distributions defined over the convex body $K \subseteq \mathbb{R}^n$. For each $\theta \in \mathbb{R}^n$, we define the density p_θ on K via

$$p_\theta(x) := \frac{\exp\langle \theta, x\rangle}{\int_K \exp\langle \theta, x'\rangle \, \mathrm{d}x'} \, \mathbb{1}\{x \in K\} \,. \tag{10.3}$$

Since f (defined in (10.1)) is essentially the logarithmic moment-generating function of p_θ, then the derivatives of f yield cumulants of p_θ. In particular,

$$\nabla f(\theta) = \mathbb{E}_{p_\theta} X \,, \qquad \nabla^2 f(\theta) = \mathrm{cov}_{p_\theta} X \,.$$

By convex duality, the mappings $\nabla f : \mathbb{R}^n \to \operatorname{int} K$ and $\nabla f^\star : \operatorname{int} K \to \mathbb{R}^n$ are inverses of each other. From the classical duality between the logarithmic moment-generating function and entropy, we can also deduce that

$$f^\star(x) = \mathcal{H}(p_{\nabla f^\star(x)}) \,,$$

where $\mathcal{H}$ denotes the entropy functional[1]

$$\mathcal{H}(p) := \int p \ln p \,. \tag{10.4}$$

[1]Note the sign convention, which is opposite the usual one in information theory. We use this convention as it is convenient for $\mathcal{H}$ to be convex.

The self-concordance parameter of $f^\star$ is the least $\nu \geq 0$ such that

$$\langle \nabla f^\star(x), [\nabla^2 f^\star(x)]^{-1} \nabla f^\star(x) \rangle \leq \nu \,, \qquad \text{for all } x \in \operatorname{int} K \,.$$

Taking $x = \nabla f(\theta)$, equivalently we require

$$\langle \theta, \nabla^2 f(\theta)\, \theta \rangle \leq \nu \,, \qquad \text{for all } \theta \in \mathbb{R}^n \,,$$

which has the probabilistic interpretation

$$\operatorname{var}_{p_\theta} \langle \theta, X \rangle \leq \nu \,, \qquad \text{for all } \theta \in \mathbb{R}^n \,. \tag{10.5}$$

From the definition (10.3), we see that the density $p_\theta \propto \exp(-V)$ is log-concave, where $V(x) = \langle \theta, x \rangle$ for $x \in \operatorname{int} K$. By applying Theorem 10.1.3 to p_θ, we immediately deduce that (10.5) holds with $\nu = n$.

■ 10.1.3 Proof of the dimensional Brascamp–Lieb inequality

Next, we wish to give some proofs of the dimensional Brascamp–Lieb inequality (Theorem 10.1.4). Classically, the Brascamp–Lieb inequality reads as follows.

Theorem 10.1.6 ([BL76])**.** *Let $\mu \propto \exp(-V)$ be a density on $\mathbb{R}^n$, where V is a convex function of class $\mathcal{C}^2$. Then, for every locally Lipschitz $g : \mathbb{R}^n \to \mathbb{R}$,*

$$\operatorname{var}_\mu g \leq \mathbb{E}_\mu \langle \nabla g, (\nabla^2 V)^{-1} \nabla g \rangle \,. \tag{10.6}$$

The Brascamp–Lieb inequality is a Poincaré inequality for the measure μ corresponding to the Newton–Langevin diffusion (§8). When V is strongly convex, $\nabla^2 V \succeq \alpha I_n$, it recovers the usual Poincaré inequality

$$\operatorname{var}_\mu g \leq \frac{1}{\alpha}\, \mathbb{E}_\mu[\|\nabla g\|^2] \,.$$

See [BL00; BGL14; Cor17] for various proofs of Theorem 10.1.6.

Since the inequality (10.6) makes no explicit reference to the dimension, it actually holds in infinite-dimensional space. In contrast, Theorem 10.1.4 asserts that (10.6) can be improved by subtracting an additional non-negative term from the right-hand side in any finite dimension. This is referred to as a *dimensional improvement* of the Brascamp–Lieb inequality.

■ 10.1.3.1 Proof by Hörmander's L^2 method

We now present the proof of Theorem 10.1.4 given in [BGG18]. The starting point for Hörmander's L^2 method is to first dualize the Poincaré inequality.

Proposition 10.1.7 ([BC13, Lemma 1])**.** *Let $\mu \propto \exp(-V)$ be a probability density on $\mathbb{R}^n$, where V is of class $\mathcal{C}^1$. Define the corresponding generator $\mathscr{L}$ on smooth functions $g : \mathbb{R}^n \to \mathbb{R}$ via*

$$\mathscr{L} g := \Delta g - \langle \nabla V, \nabla g \rangle \,.$$

Suppose $A : \mathbb{R}^n \to \mathrm{PD}(n)$ is a matrix-valued function mapping into the space of symmetric positive definite matrices such that for all smooth $u : \mathbb{R}^n \to \mathbb{R}$,

$$\mathbb{E}_\mu[(\mathscr{L} u)^2] \geq \mathbb{E}_\mu \langle \nabla u, A \nabla u \rangle \,. \tag{10.7}$$

Then, for all $g \in L^2(\mu)$, it holds that

$$\mathrm{var}_\mu \, g \leq \mathbb{E}_\mu \langle \nabla g, A^{-1} \nabla g \rangle \,.$$

Proof. We may assume $\mathbb{E}_\mu \, g = 0$. This condition is certainly necessary for the Poisson equation $-\mathscr{L} u = g$ to be solvable; in order to streamline the proof, we will assume that a solution u exists. (This assumption can be avoided by invoking [CFM04] and using a density argument; see [BC13] for details.)

Using the integration by parts formula for the generator,

$$-\mathbb{E}_\mu[g \, \mathscr{L} u] = \mathbb{E}_\mu \langle \nabla g, \nabla u \rangle \,,$$

we obtain

$$\begin{aligned} \mathrm{var}_\mu \, g = \mathbb{E}_\mu[g^2] &= -2 \, \mathbb{E}_\mu[g \, \mathscr{L} u] - \mathbb{E}_\mu[(\mathscr{L} u)^2] \\ &\leq 2 \, \mathbb{E}_\mu \langle \nabla g, \nabla u \rangle - \mathbb{E}_\mu \langle \nabla u, A \nabla u \rangle \,. \end{aligned}$$

Next, since $2 \langle x, y \rangle \leq \langle x, A \, x \rangle + \langle y, A^{-1} \, y \rangle$ for all $x, y \in \mathbb{R}^n$, it implies

$$\mathrm{var}_\mu \, g \leq \mathbb{E}_\mu \langle \nabla g, A^{-1} \nabla g \rangle \,. \qquad \square$$

The key idea now is that the condition (10.7) can be verified with the help of the *curvature* of the potential V. Indeed, assume now that V is of class $\mathcal{C}^2$ and that $\nabla^2 V \succ 0$. By direct calculation, one verifies the commutation relation

$$\nabla \mathscr{L} u = (\mathscr{L} - \nabla^2 V) \, \nabla u \,. \tag{10.8}$$

Hence,

$$\begin{aligned} \mathbb{E}_\mu[(\mathscr{L} u)^2] &= -\mathbb{E}_\mu \langle \nabla u, \nabla \mathscr{L} u \rangle = -\mathbb{E}_\mu \langle \nabla u, (\mathscr{L} - \nabla^2 V) \, \nabla u \rangle \\ &= \mathbb{E}_\mu \langle \nabla u, \nabla^2 V \, \nabla u \rangle + \mathbb{E}_\mu[\|\nabla^2 u\|_{\mathrm{HS}}^2] \,, \end{aligned} \tag{10.9}$$

where the last equality follows from the integration by parts formula for the generator applied to each coordinate separately: $-\mathbb{E}_\mu\langle\nabla u, \mathscr{L}\nabla u\rangle = \mathbb{E}_\mu[\|\nabla^2 u\|_{\mathrm{HS}}^2]$. Since the second term is non-negative, Proposition 10.1.7 now implies the Brascamp–Lieb inequality (Theorem 10.1.6).

In order to obtain the dimensional improvement of the Brascamp–Lieb inequality (Theorem 10.1.4), we will imitate the proof of Proposition 10.1.7, only now we will use the additional term $\mathbb{E}_\mu[\|\nabla^2 u\|_{\mathrm{HS}}^2]$ in the above identity.

Proof of Theorem 10.1.4. As before, let $\mathbb{E}_\mu g = 0$. However, we introduce an additional trick and consider u not necessarily satisfying $-\mathscr{L}u = g$; this will help to optimize the bound at the end of the argument. Following the computations in Proposition 10.1.7 and using the key identity (10.9), we obtain

$$\begin{aligned}\operatorname{var}_\mu g = \mathbb{E}_\mu[g^2] &= \mathbb{E}_\mu[(g+\mathscr{L}u)^2] - 2\,\mathbb{E}_\mu[g\,\mathscr{L}u] - \mathbb{E}_\mu[(\mathscr{L}u)^2] \\ &= \mathbb{E}_\mu[(g+\mathscr{L}u)^2] + 2\,\mathbb{E}_\mu\langle\nabla g, \nabla u\rangle - \mathbb{E}_\mu\langle\nabla u, \nabla^2 V\,\nabla u\rangle - \mathbb{E}_\mu[\|\nabla u\|_{\mathrm{HS}}^2] \\ &\le \mathbb{E}_\mu[(g+\mathscr{L}u)^2] + \mathbb{E}_\mu\langle\nabla g, (\nabla^2 V)^{-1}\,\nabla g\rangle - \mathbb{E}_\mu[\|\nabla u\|_{\mathrm{HS}}^2]\,.\end{aligned}$$

For the second term, we use the inequality

$$\mathbb{E}_\mu[\|\nabla u\|_{\mathrm{HS}}^2] \ge \frac{1}{n}\,(\mathbb{E}_\mu\,\Delta u)^2\,.$$

From integration by parts,

$$\mathbb{E}_\mu\,\Delta u = \mathbb{E}_\mu\langle\nabla V, \nabla u\rangle = -\,\mathbb{E}_\mu[V\,\mathscr{L}u] = \operatorname{cov}_\mu(g, V) - \mathbb{E}_\mu[V\,(\mathscr{L}u + g)]\,.$$

We now choose $-\mathscr{L}u = g + a\,(V - \mathbb{E}_\mu V)$ for some $a \ge 0$ to be chosen later. For brevity of notation, write $\mathbf{C} \coloneqq \operatorname{cov}_\mu(g, V)$ and $\mathbf{V} \coloneqq \operatorname{var}_\mu V$. Then,

$$\begin{aligned}\operatorname{var}_\mu g - \mathbb{E}_\mu\langle\nabla g, (\nabla^2 V)^{-1}\,\nabla g\rangle &\le a^2\mathbf{V} - \frac{1}{n}\,(\mathbf{C} + a\mathbf{V})^2 \\ &= -\frac{\mathbf{V}\,(n-\mathbf{V})}{n}\left(a - \frac{\mathbf{C}}{n-\mathbf{V}}\right)^2 - \frac{\mathbf{C}^2\mathbf{V}}{n\,(n-\mathbf{V})} - \frac{\mathbf{C}^2}{n}\,.\end{aligned}$$

Observe that this inequality entails $\mathbf{V} \le n$, or else we could send $a \to \infty$ and arrive at a contradiction. Optimizing over a, we obtain

$$\operatorname{var}_\mu g \le \mathbb{E}_\mu\langle\nabla g, (\nabla^2 V)^{-1}\,\nabla g\rangle - \frac{\mathbf{C}^2}{n-\mathbf{V}}\,. \qquad \square$$

■ 10.1.3.2 Proof by convexity of the entropy along Bregman divergence couplings

It is well-known that Poincaré inequalities are obtained from linearizing transportation inequalities. In [Cor17], D. Cordero-Erausquin obtained the Brascamp–Lieb inequality (Theorem 10.1.6) by linearizing the following inequality:

$$\mathcal{D}_V(\rho \,\|\, \mu) \le \mathsf{KL}(\rho \,\|\, \mu)\,, \qquad \text{for all } \rho \in \mathcal{P}(\mathbb{R}^n)\,. \tag{10.10}$$

Here, $\mu \propto \exp(-V)$ on $\mathbb{R}^n$; $\mathcal{P}(\mathbb{R}^n)$ denotes the space of probability measures on $\mathbb{R}^n$; $\mathsf{KL}(\cdot\,\|\,\cdot)$ is the Kullback–Leibler (KL) divergence; and $\mathcal{D}_V(\cdot\,\|\,\cdot)$ is the Bregman divergence coupling cost, defined as

$$\mathcal{D}_V(\rho \,\|\, \mu) = \inf_{\gamma \in \mathsf{couplings}(\rho,\mu)} \int D_V(x \,\|\, y)\, \mathrm{d}\gamma(x,y)\,,$$

with

$$D_V(x \,\|\, y) \coloneqq V(x) - V(y) - \langle \nabla V(y), x - y \rangle\,.$$

On the other hand, we obtained the transport inequality (10.10) as Corollary 9.4.2 as a consequence of a convexity principle in optimal transport. It is therefore natural to ask whether the dimensional Brascamp–Lieb inequality (Theorem 10.1.4) can be obtained directly from (a strengthening of) this principle. This is indeed the case, and we now describe this argument.

Making the argument fully rigorous, however, would entail substantial technical complications which would detract from the focus of this section. In any case, a complete proof of the dimensional Brascamp–Lieb inequality is already present in [BGG18]. Hence, we will work on a purely formal level and assume that everything is smooth, bounded, etc. Also, the computations are rather similar to the proof of Theorem 10.1.4 given in the previous section. Nevertheless, the argument seems interesting enough to warrant presenting it here.

The main difference with the preceding proof is that the Bochner formula (implicit in the commutation relation (10.8)) is replaced by the convexity principle.

Proof sketch of Theorem 10.1.4. Throughout the proof, let $\varepsilon > 0$ be small. Let h be bounded and satisfy $\mathbb{E}_\mu h = 0$, so that $\mu_\varepsilon \coloneqq (1+\varepsilon h)\,\mu$ defines a valid probability density on $\mathbb{R}^n$. Our aim is to first strengthen the transportation inequality (10.10), at least infinitesimally, and then to linearize it.

Let (X_ε, X) be an optimal coupling for the Bregman divergence coupling cost $\mathcal{D}_V(\mu_\varepsilon \,\|\, \mu)$. In §9, we proved the following facts:

1. There is a function $u_\varepsilon : \mathbb{R}^n \to \mathbb{R}$ such that $\nabla V(X) = \nabla V(X_\varepsilon) - \nabla u_\varepsilon(X_\varepsilon)$, and $V - u_\varepsilon$ is convex.

2. The entropy functional (defined in (10.4)) is convex in the sense that

$$\mathcal{H}(\mu_\varepsilon) \geq \mathcal{H}(\mu) + \mathbb{E}\langle [\nabla_{W_2}\mathcal{H}(\mu)](X), X_\varepsilon - X\rangle\,. \tag{10.11}$$

   Here, $\nabla_{W_2}\mathcal{H}(\mu) = \nabla \ln \mu$ is the Wasserstein gradient of the entropy functional, c.f. [AGS08; Vil09b; San15].

Write $T_\varepsilon(x) \coloneqq (\nabla V - \nabla u_\varepsilon)^{-1}(\nabla V(x))$. Since $(T_\varepsilon)_{\#}\mu = \mu_\varepsilon$, the change of variables formula implies

$$\frac{\mu(x)}{\mu_\varepsilon(T_\varepsilon(x))} = \frac{\mu(x)}{\mu(T_\varepsilon(x))\,(1+\varepsilon h(T_\varepsilon(x)))} = \det \nabla T_\varepsilon(x)\,. \tag{10.12}$$

To linearize this equation, write $u_\varepsilon = \varepsilon u + o(\varepsilon)$ and $T_\varepsilon(x) = x + \varepsilon T(x) + o(\varepsilon)$. Then, the definition of T_ε yields

$$\begin{aligned}\nabla V(x) &= (\nabla V - \nabla u_\varepsilon)\big(x + \varepsilon T(x) + o(\varepsilon)\big)\\ &= \nabla V(x) + \varepsilon\,\nabla^2 V(x)\,T(x) - \varepsilon\,\nabla u(x) + o(\varepsilon)\end{aligned}$$

which implies

$$T_\varepsilon(x) = x + \varepsilon\,[\nabla^2 V(x)]^{-1}\,\nabla u(x) + o(\varepsilon)\,.$$

Taking logarithms and expanding to first order in ε,

$$\begin{aligned}&\ln \mu(x) - \ln \mu(T_\varepsilon(x)) - \ln(1+\varepsilon h(T_\varepsilon(x)))\\ &\quad= -\varepsilon\,\langle \nabla \ln \mu(x), [\nabla^2 V(x)]^{-1}\,\nabla u(x)\rangle - \varepsilon h(x) + o(\varepsilon)\\ &\quad= \varepsilon\,\langle \nabla V(x), [\nabla^2 V(x)]^{-1}\,\nabla u(x)\rangle - \varepsilon h(x) + o(\varepsilon)\end{aligned}$$

and

$$\begin{aligned}\ln\det \nabla T_\varepsilon(x) &= \ln\det \nabla\big(\mathrm{id} + \varepsilon\,[\nabla^2 V]^{-1}\,\nabla u + o(\varepsilon)\big)(x)\\ &= \ln\det\big(I_n + \varepsilon\,\nabla([\nabla^2 V]^{-1}\,\nabla u)(x) + o(\varepsilon)\big)\\ &= \varepsilon\,\mathrm{div}([\nabla^2 V]^{-1}\,\nabla u)(x) + o(\varepsilon)\,.\end{aligned}$$

To interpret this, we introduce a new generator, denoted $\hat{\mathscr{L}}$ to avoid confusion with the previous section, defined by

$$\hat{\mathscr{L}} u \coloneqq \mathrm{div}([\nabla^2 V]^{-1}\,\nabla u) - \langle \nabla V, [\nabla^2 V]^{-1}\,\nabla u\rangle\,.$$

This new generator satisfies the integration by parts formula

$$\mathbb{E}_\mu[u\,\hat{\mathscr{L}} v] = \mathbb{E}_\mu\langle \nabla u, [\nabla^2 V]^{-1}\,\nabla v\rangle\,.$$

In this notation, the preceding computations yield

$$\hat{\mathscr{L}} u = -h + o(1)\,.$$

Next, to strengthen (10.11), we repeat the proof. From (10.12),

$$\begin{aligned}\mathcal{H}(\mu_\varepsilon) &= \int \mu_\varepsilon \ln \mu_\varepsilon = \int \mu \ln(\mu_\varepsilon \circ T_\varepsilon) = \int \mu \ln \frac{\mu}{\det \nabla T_\varepsilon} \\ &= \mathcal{H}(\mu) - \int \mu \ln \det \nabla T_\varepsilon\,.\end{aligned}$$

From the second-order expansion of $-\ln\det$ around I_n,

$$\begin{aligned}&-\int \mu \ln\det \nabla T_\varepsilon \\ &\quad\geq -\int \mu \ln\det I_n - \int \mu\,\langle I_n, \nabla T_\varepsilon - I_n\rangle + \frac{1}{2}\int \mu\,\|\nabla T_\varepsilon - I_n\|_{\mathrm{HS}}^2 + o(\varepsilon^2) \\ &\quad\geq -\int \mu \,\mathrm{tr}(\nabla T_\varepsilon - I_n) + \frac{1}{2n}\left(\int \mu\,\mathrm{tr}(\nabla T_\varepsilon - I_d)\right)^2 + o(\varepsilon^2) \\ &\quad= -\int \mu \,\mathrm{div}(T_\varepsilon - \mathrm{id}) + \frac{1}{2n}\left(\int \mu\,\mathrm{div}(T_\varepsilon - \mathrm{id})\right)^2 + o(\varepsilon^2) \\ &\quad= \int \mu\,\langle \nabla \ln \mu, T_\varepsilon - \mathrm{id}\rangle + \frac{1}{2n}\left(\int \mu\,\langle \nabla \ln \mu, T_\varepsilon - \mathrm{id}\rangle\right)^2 + o(\varepsilon^2)\,.\end{aligned}$$

Recalling that $\nabla_{W_2}\mathcal{H}(\mu) = \nabla \ln \mu$, we have established

$$\begin{aligned}\mathcal{H}(\mu_\varepsilon) - \mathcal{H}(\mu) - \mathbb{E}\langle [\nabla_{W_2}\mathcal{H}(\mu)](X), X_\varepsilon - X\rangle &\\ \geq \frac{1}{2n}\left(\int \mu\,\langle \nabla V, T_\varepsilon - \mathrm{id}\rangle\right)^2 + o(\varepsilon^2) &\\ = \frac{\varepsilon^2}{2n}\left(\int \mu\,\langle \nabla V, [\nabla^2 V]^{-1}\nabla u\rangle\right)^2 + o(\varepsilon^2) &\\ = \frac{\varepsilon^2}{2n}\left\{\mathbb{E}_\mu[V\,\hat{\mathscr{L}} u]\right\}^2 + o(\varepsilon^2)\,.&\end{aligned}$$

The next step is to write down the strengthened transportation inequality. Indeed, if we add a suitable additive constant to V so that $\mu = \exp(-V)$, then

$$\begin{aligned}\mathsf{KL}(\mu_\varepsilon \,\|\, \mu) &= \mathbb{E}_{\mu_\varepsilon} V + \mathcal{H}(\mu_\varepsilon) \\ &\geq \underbrace{\mathbb{E}\,V(X) + \mathcal{H}(\mu)}_{=\mathsf{KL}(\mu\|\mu)=0} + \underbrace{\mathbb{E}\langle [\nabla V + \nabla_{W_2}\mathcal{H}(\mu)](X), X_\varepsilon - X\rangle}_{=[\nabla_{W_2}\mathsf{KL}(\cdot\|\mu)](\mu)=0} \\ &\quad + \underbrace{\mathbb{E}[V(X_\varepsilon) - V(X) - \langle \nabla V(X), X_\varepsilon - X\rangle]}_{=\mathcal{D}_V(\mu_\varepsilon\|\mu)}\end{aligned}$$

$$+ \frac{\varepsilon^2}{2n} \{\mathbb{E}_\mu[hV]\}^2 + o(\varepsilon^2)$$
$$\geq \mathcal{D}_V(\mu_\varepsilon \,\|\, \mu) + \frac{\varepsilon^2}{2n} \{\mathbb{E}_\mu[hV]\}^2 + o(\varepsilon^2)\,.$$

Finally, it remains to linearize the transportation inequality. On one hand, it is classical that

$$\mathsf{KL}(\mu_\varepsilon \,\|\, \mu) = \frac{\varepsilon^2}{2}\, \mathbb{E}_\mu[h^2] + o(\varepsilon^2)\,.$$

On the other hand, we can guess that

$$\begin{aligned}
\mathcal{D}_V(\mu_\varepsilon \,\|\, \mu) &= \frac{1}{2}\, \mathbb{E}\langle X_\varepsilon - X, \nabla^2 V(X)\,(X_\varepsilon - X)\rangle + o(\varepsilon^2) \\
&= \frac{\varepsilon^2}{2}\, \mathbb{E}_\mu\langle \nabla u, (\nabla^2 V)^{-1}\, \nabla u\rangle + o(\varepsilon^2) \\
&\geq \frac{\varepsilon^2}{2}\, \frac{\{\mathbb{E}_\mu\langle \nabla g, (\nabla^2 V)^{-1}\, \nabla u\rangle\}^2}{\mathbb{E}_\mu\langle \nabla g, (\nabla^2 V)^{-1}\, \nabla g\rangle} + o(\varepsilon^2) \\
&= \frac{\varepsilon^2}{2}\, \frac{\{\mathbb{E}_\mu[g\, \hat{\mathscr{L}} u]\}^2}{\mathbb{E}_\mu\langle \nabla g, (\nabla^2 V)^{-1}\, \nabla g\rangle} + o(\varepsilon^2) \\
&= \frac{\varepsilon^2}{2}\, \frac{\{\mathbb{E}_\mu[gh]\}^2}{\mathbb{E}_\mu\langle \nabla g, (\nabla^2 V)^{-1}\, \nabla g\rangle} + o(\varepsilon^2)\,.
\end{aligned}$$

A rigorous proof of this inequality is given as [Cor17, Lemma 3.1].

Thus, we obtain

$$\frac{1}{2}\, \frac{\{\mathbb{E}_\mu[gh]\}^2}{\mathbb{E}_\mu\langle \nabla g, (\nabla^2 V)^{-1}\, \nabla g\rangle} + \frac{1}{2n} \{\mathbb{E}_\mu[hV]\}^2 \leq \frac{1}{2}\, \mathbb{E}_\mu[h^2] + o(1)\,.$$

Now we let $\varepsilon \searrow 0$ and choose $h = g + a\,(V - \mathbb{E}_\mu V)$ for some $a \in \mathbb{R}$. Writing $\mathbf{C} \coloneqq \mathrm{cov}_\mu(g, V)$ and $\mathbf{V} \coloneqq \mathrm{var}_\mu V$, it yields

$$\frac{(\mathrm{var}_\mu\, g + a\mathbf{C})^2}{\mathbb{E}_\mu\langle \nabla g, (\nabla^2 V)^{-1}\, \nabla g\rangle} + \frac{1}{n}\, (\mathbf{C} + a\mathbf{V})^2 \leq \mathrm{var}_\mu\, g + 2a\mathbf{C} + a^2\mathbf{V}\,.$$

Actually, choosing a to optimize this inequality and simplifying the resulting expression may be cumbersome, so with our foresight from the earlier proof of Theorem 10.1.4, we now take $a = \mathbf{C}/(n - \mathbf{V})$. After some algebra,

$$\frac{(\mathrm{var}_\mu\, g + \mathbf{C}^2/(n - \mathbf{V}))^2}{\mathbb{E}_\mu\langle \nabla g, (\nabla^2 V)^{-1}\, \nabla g\rangle} \leq \mathrm{var}_\mu\, g + \frac{\mathbf{C}^2}{n - \mathbf{V}}\,,$$

which of course yields

$$\mathrm{var}_\mu\, g \leq \mathbb{E}_\mu\langle \nabla g, (\nabla^2 V)^{-1}\, \nabla g\rangle - \frac{\mathbf{C}^2}{n - \mathbf{V}}\,. \qquad \square$$

■ 10.1.4 A tensorization trick

We begin by verifying that the entropic barrier has the consistency property (10.2). Let f_K denote the function (10.1), where we now explicitly denote the dependence on the convex body K. Also, let $f_K^\star$ denote the corresponding entropic barrier. Then, we see that

$$\begin{aligned} f_{K\times K'}(\theta,\theta') &= \ln \int_{K\times K'} \exp(\langle\theta,x\rangle + \langle\theta',x'\rangle)\,\mathrm{d}x\,\mathrm{d}x' \\ &= \ln \int_K \exp\langle\theta,x\rangle\,\mathrm{d}x + \ln\int_{K'} \exp\langle\theta',x'\rangle\,\mathrm{d}x' = f_K(\theta) + f_{K'}(\theta')\,. \end{aligned}$$

Hence,

$$f^\star_{K\times K'}(x,x') = \sup_{\theta,\theta'\in\mathbb{R}^n}\{\langle\theta,x\rangle + \langle\theta',x'\rangle - f_K(\theta) - f_{K'}(\theta')\} = f_K^\star(x) + f_{K'}^\star(x')\,.$$

Finally, we check that the tensorization property automatically improves the bound on the self-concordance parameter of $f_K^\star$ obtained in [BE19].

Proof of Lemma 10.1.5. Let $\boldsymbol{x} \coloneqq (x_1,\dots,x_k) \in (\mathbb{R}^n)^k$. By assumption, the self-concordant barrier ϕ_{kn,K^k} on K^k satisfies $\phi_{kn,K^k}(\boldsymbol{x}) = \sum_{j=1}^k \phi_{n,K}(x_j)$. Also, we are given that

$$\nabla^2\phi_{kn,K^k}(\boldsymbol{x}) \succeq \frac{1}{\nu(kn)}\,\nabla\phi_{kn,K^k}(\boldsymbol{x})\,\nabla\phi_{kn,K^k}(\boldsymbol{x})^{\mathsf{T}}\,. \tag{10.13}$$

Via elementary calculations,

$$\nabla\phi_{kn,K^k}(\boldsymbol{x}) = \big(\nabla\phi_{n,K}(x_1),\dots,\nabla\phi_{n,K}(x_k)\big)$$

and

$$\nabla^2\phi_{kn,K^k}(\boldsymbol{x}) = \begin{bmatrix} \nabla^2\phi_{n,K}(x_1) & & \\ & \ddots & \\ & & \nabla^2\phi_{n,K}(x_k) \end{bmatrix}.$$

Let $v\in\mathbb{R}^n$ and let $\boldsymbol{v} \coloneqq (v,\dots,v) \in (\mathbb{R}^n)^k$. Also, take $x_1 = \cdots = x_k = x$. By (10.13), we know that

$$\begin{aligned} k\,\langle v, \nabla^2\phi_{n,K}(x)\,v\rangle = \langle \boldsymbol{v}, \nabla^2\phi_{kn,K^k}(\boldsymbol{x})\,\boldsymbol{v}\rangle &\ge \frac{1}{\nu(kn)}\,\langle \boldsymbol{v}, \nabla\phi_{kn,K^k}(\boldsymbol{x})\rangle^2 \\ &= \frac{k^2}{\nu(kn)}\,\langle v, \nabla\phi_{n,K}(x)\rangle^2 \end{aligned}$$

which proves

$$\nabla^2 \phi_{n,K}(x) \succeq \frac{k}{\nu(kn)} \nabla \phi_{n,K}(x) \nabla \phi_{n,K}(x)^\mathsf{T}$$

and gives the claim. □

Proof of Theorem 10.1.2. According to Theorem 10.1.1, we know that the entropic barrier in n dimensions is $(1+\varepsilon_n)$ n-self-concordant, with $\varepsilon_n \to 0$ as $n \to \infty$. By Lemma 10.1.5, it is actually $(1+\varepsilon_{kn})$ n-self-concordant, for any $k \in \mathbb{N}^+$. Let $k \to \infty$ to deduce that it is in fact n-self-concordant. □

10.2 An entropic generalization of Caffarelli's contraction theorem

10.2.1 Introduction

In [Caf00], Caffarelli proved the following seminal result.

Theorem 10.2.1 (Caffarelli's contraction theorem)**.** *Let $P = \exp(-V)$ and $Q = \exp(-W)$ have smooth densities on $\mathbb{R}^d$, with $\nabla^2 V \preceq \beta_V I$ and $\nabla^2 W \succeq \alpha_W I \succ 0$. Then, the optimal transport map $\nabla \phi_0$ from P to Q is $\sqrt{\beta_V/\alpha_W}$-Lipschitz.*

Here, $\phi_0 : \mathbb{R}^d \to \mathbb{R}$ is a convex function, known as a *Brenier potential*. The optimal transport map $\nabla \phi_0 : \mathbb{R}^d \to \mathbb{R}^d$ pushes forward P to Q, in the sense that if X is a random variable with law P, then $\nabla \phi_0(X)$ is a random variable with law Q. See §10.2.2.2 and the textbook [Vil03] for background on optimal transport.

Caffarelli's contraction theorem can be used to transfer functional inequalities, such as a Poincaré inequality, from the standard Gaussian measure on $\mathbb{R}^d$ to other probability measures [BGL14]. Towards this end, recent works have also constructed and studied alternative Lipschitz transport maps (e.g. [KM12; MS21; MS22; Nee22]), but still the properties of the original optimal transport map remain of fundamental interest, with many questions unresolved [Val07; CFJ17].

Indeed, besides the application to functional inequalities, the structural properties of optimal transport maps play a fundamental role in theoretical and methodological advances in optimal transport, such as the control of the curvature of the Wasserstein space through the notion of extendible geodesics [ALP20; Le +22], the stability of Wasserstein barycenters (see §15), and the statistical estimation of optimal transport maps [HR21].

In applied domains, however, the inauspicious computational and statistical burden of solving the original optimal transport problem has instead led practitioners to consider *entropically regularized* optimal transport, as pioneered by

Cuturi in [Cut13]. In addition to its practical merits, entropic optimal transport enjoys a rich mathematical theory, rooted in its connection to the classical Schrödinger bridge problem [Léo14], which has led to powerful applications to high-dimensional probability [Led18; FGP20; Gen+20]. As such, it is natural to study the properties of the entropic analogue of the optimal transport map.

In this section, we prove a generalization of Caffarelli's contraction theorem to the setting of entropic optimal transport. Namely, we study the Hessian of the *entropic Brenier potential* (see §10.2.3), which admits a representation as a covariance matrix (Lemma 10.2.6). By applying two well-known inequalities for covariance matrices (the Brascamp–Lieb inequality and the Cramér–Rao inequality), we quickly deduce a sharp upper bound on the operator norm of the Hessian which holds for any value $\varepsilon > 0$ of the regularization parameter.

As a by-product of our analysis, by sending $\varepsilon \searrow 0$ and appealing to recent convergence results for the entropic Brenier potentials [BGN22], we obtain the shortest proof of Caffarelli's contraction theorem to date. Notably, our argument allows us to sidestep the regularity of the optimal transport map, which is a key obstacle in Caffarelli's original proof and many others in the literature (see, e.g., [Kol11]).

Recently, in [FGP20] (see also [Pro21]), Fathi, Gozlan, and Prod'homme gave a proof of Caffarelli's theorem using a surprising equivalence between Theorem 10.2.1 and a statement about Wasserstein projections, which was discovered through the theory of weak optimal transport [GJ20]. In order to verify the latter, their proof also used ideas from entropic optimal transport.[2] In comparison, we note that our argument is much more direct.

To further demonstrate the applicability of our technique, in §10.2.5 we prove a generalization of Caffarelli's result which reveals a remarkable extremal property of optimal transport maps between Gaussians. Namely, if $\nabla^2 V \preceq A^{-1}$ and $\nabla^2 W \succeq B^{-1}$, where A and B are arbitrary commuting positive definite matrices, then the Hessian of the Brenier potential from P to Q is pointwise upper bounded (in the PSD ordering) by $A^{-1/2}B^{1/2}$, the Hessian of the Brenier potential from $\mathsf{normal}(0, A)$ to $\mathsf{normal}(0, B)$. To the best of our knowledge, this result is new.

10.2.2 Background

10.2.2.1 Assumptions

We study probability measures P, Q on $\mathbb{R}^d$ satisfying the following mild regularity assumptions.

[2] In particular, with some effort, a bound on the Hessian of the entropic Brenier potential can also be read off from their proof.

Assumption 10.2.2 (Regularity conditions)**.** *We henceforth refer to the source measure as P and the target measure as Q. We say that (P, Q) satisfies our regularity conditions if:*

1. *P has full support on $\mathbb{R}^d$ and Q is supported on a convex subset of $\mathbb{R}^d$. Let Ω_Q denote the interior of the support of Q, so that Ω_Q is a convex open set.*
2. *P and Q admit positive Lebesgue densities on $\mathbb{R}^d$ and Ω_Q, which we can therefore be written $\exp(-V)$ and $\exp(-W)$ respectively for functions $V, W : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. We abuse notation and identify the measures with their densities, thus writing $P = \exp(-V)$ and $Q = \exp(-W)$.*
3. *We assume that V and W are twice continuously differentiable on $\mathbb{R}^d$ and Ω_Q respectively.*

Some of these assumptions can be eventually relaxed, but they suffice for the purposes of this work. Throughout the rest of the chapter and for the sake of simplicity, these regularity assumptions are assumed to hold for the probability measures under consideration.

■ 10.2.2.2 Optimal transport without regularization

Let P and Q be probability measures with finite second moment. The *optimal transport problem* is the following optimization problem:

$$\underset{\pi \in \Pi(P,Q)}{\text{minimize}} \quad \int \tfrac{1}{2} \|x - y\|^2 \, \mathrm{d}\pi(x, y) \tag{10.14}$$

where $\Pi(P, Q)$ is the set of joint probability measures with marginals P and Q. The following fundamental result characterizes the optimal solution to (10.14).

Theorem 10.2.3 (Brenier's theorem)**.** *Suppose that P admits a density with respect to Lebesgue measure. Then, there exists a proper, convex, lower semicontinuous function $\phi_0 : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ such that the optimal transport plan in (10.14) can be written $\pi_0 = (\mathrm{id}, \nabla\phi_0)_{\#}P$. The function ϕ_0 is called the* Brenier potential*, and the mapping $\nabla\phi_0$ is called the* optimal transport map *from P to Q. Moreover, the optimal transport map $\nabla\phi_0$ is unique up to P-a.e. equality.*

The Brenier potential ϕ_0 is obtained as the solution to the dual problem

$$\underset{\phi \in \Gamma_0}{\text{maximize}} \quad \int \left(\frac{\|\cdot\|^2}{2} - \phi\right) \mathrm{d}P + \int \left(\frac{\|\cdot\|^2}{2} - \phi^*\right) \mathrm{d}Q \,, \tag{10.15}$$

where ϕ^* is the convex conjugate to ϕ, and Γ_0 is the set of proper, convex, lower semicontinuous functions on $\mathbb{R}^d$.

We refer to [Vil03] for further background.

■ 10.2.3 Optimal transport with entropic regularization

We recall that *entropic optimal transport* is the problem that arises when we add the Kullback–Leibler (KL) divergence, $\mathsf{KL}(\cdot \,\|\, \cdot)$, as a regularizer to (10.14):

$$\underset{\pi \in \Pi(P,Q)}{\text{minimize}} \quad \int \tfrac{1}{2} \|x-y\|^2 \,\mathrm{d}\pi(x,y) + \varepsilon\, \mathsf{KL}(\pi \,\|\, P \otimes Q)\,. \tag{10.16}$$

The following theorem characterizes the solution to (10.16) [Csi75; PC19; BGN22].

Theorem 10.2.4 (Entropic optimal transport)**.** *Let P and Q be probability measures on $\mathbb{R}^d$ and fix $\varepsilon > 0$. Then there exists a unique solution $\pi_\varepsilon \in \Pi(P,Q)$ to (10.16). Moreover, π_ε has the form*

$$\pi_\varepsilon(\mathrm{d}x, \mathrm{d}y) = \exp\Big(\frac{f_\varepsilon(x) + g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon}\Big)\, P(\mathrm{d}x)\, Q(\mathrm{d}y)\,, \tag{10.17}$$

where $(f_\varepsilon, g_\varepsilon)$ are maximizers for the dual problem

$$\underset{(f,g)\in L^1(P)\times L^1(Q)}{\text{maximize}} \quad \int f \,\mathrm{d}P + \int g \,\mathrm{d}Q - \varepsilon \iint \exp\Big(\frac{f(x)+g(y)-\frac{1}{2}\|x-y\|^2}{\varepsilon}\Big)\,\mathrm{d}P(x)\,\mathrm{d}Q(y) + \varepsilon\,.$$

The constraint that π_ε has marginals P and Q implies the following dual optimality conditions for $(f_\varepsilon, g_\varepsilon)$ (see [MN19; BGN22] for more details):

$$f_\varepsilon(x) = -\varepsilon \log \int \exp\Big(\frac{g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon}\Big)\,\mathrm{d}Q(y) \qquad (x \in \mathbb{R}^d)\,, \tag{10.18}$$

$$g_\varepsilon(y) = -\varepsilon \log \int \exp\Big(\frac{f_\varepsilon(x) - \frac{1}{2}\|x-y\|^2}{\varepsilon}\Big)\,\mathrm{d}P(x) \qquad (y \in \mathbb{R}^d)\,. \tag{10.19}$$

In particular, f_ε and g_ε are smooth. In this work, it is more convenient to work with the *entropic Brenier potentials*, defined as

$$(\phi_\varepsilon, \psi_\varepsilon) \coloneqq (\tfrac{1}{2}\|\cdot\|^2 - f_\varepsilon,\ \tfrac{1}{2}\|\cdot\|^2 - g_\varepsilon)\,. \tag{10.20}$$

Since $(f_\varepsilon, g_\varepsilon)$ are only unique up to adding a constant to f_ε and subtracting the same constant from g_ε, we fix the normalization convention $\int f_\varepsilon \,\mathrm{d}P = \int g_\varepsilon \,\mathrm{d}Q$. Under this condition, it was shown by Nutz and Wiesel in [NW22] that we have convergence to the Brenier potential $\varphi_\varepsilon \to \varphi_0$ as $\varepsilon \searrow 0$; we recall an abbreviated version of the statement for the convenience of the reader:

Theorem 10.2.5. *For any choice of regularization parameter $\varepsilon > 0$, let $(\phi_\varepsilon, \psi_\varepsilon)$ be the unique entropic Brenier potentials with the normalization condition*

$$\int (\tfrac{1}{2}\|\cdot\|^2 - \phi_\varepsilon)\,\mathrm{d}P = \int (\tfrac{1}{2}\|\cdot\|^2 - \psi_\varepsilon)\,\mathrm{d}Q\,.$$

If (ϕ_0, ϕ_0^*) are unique, it holds that $\lim_{\varepsilon \searrow 0} \phi_\varepsilon = \phi_0$ in $L^1(P)$ and $\lim_{\varepsilon \searrow 0} \psi_\varepsilon = \phi_0^*$ in $L^1(Q)$.

Adopting this new notation, with $P = \exp(-V)$ and $Q = \exp(-W)$, we can rewrite the entropic optimal plan as

$$\pi_\varepsilon(\mathrm{d}x, \mathrm{d}y) = \exp\Big(-\frac{\varphi_\varepsilon(x) + \psi_\varepsilon(y) - \langle x, y\rangle}{\varepsilon} - V(x) - W(y)\Big)\,\mathrm{d}x\,\mathrm{d}y\,.$$

The entropic Brenier potentials were first introduced to develop a computationally tractable estimator of the optimal transport map $\nabla\phi_0$ [Seg+18; PCN22; PN22]. Indeed, this is motivated by the following observation, which acts as an entropic version of Brenier's theorem. Write $\pi_\varepsilon^{Y|X=x}$ for the conditional distribution of Y given $X = x$ for $(X, Y) \sim \pi_\varepsilon$, and similarly define $\pi_\varepsilon^{X|Y=y}$. Then, by [PN22, Proposition 1], $\nabla\phi_\varepsilon$ is the barycentric projection

$$\nabla\phi_\varepsilon(x) = \int y\,\mathrm{d}\pi_\varepsilon^{Y|X=x}(y)\,. \tag{10.21}$$

For clarity, we abuse notation and abbreviate $\pi_\varepsilon^{Y|X=x}$ by π_ε^x and $\pi_\varepsilon^{X|Y=y}$ by π_ε^y when there is no danger of confusion.

The following lemma is a straightforward computation using (10.17), (10.18), and (10.19).

Lemma 10.2.6. *It holds that*

$$\nabla^2\phi_\varepsilon(x) = \varepsilon^{-1}\operatorname{cov}_{Y\sim\pi_\varepsilon^x}(Y)\,, \qquad \text{and} \qquad \nabla^2\psi_\varepsilon(y) = \varepsilon^{-1}\operatorname{cov}_{X\sim\pi_\varepsilon^y}(X)\,.$$

In particular, φ_ε and ψ_ε are convex. Moreover, under our regularity conditions,

$$\begin{aligned}\nabla_y^2 \log(1/\pi_\varepsilon^x)(y) &= \varepsilon^{-1}\nabla^2\psi_\varepsilon(y) + \nabla^2 W(y)\,,\\ \nabla_x^2 \log(1/\pi_\varepsilon^y)(x) &= \varepsilon^{-1}\nabla^2\phi_\varepsilon(x) + \nabla^2 V(x)\,.\end{aligned}$$

■ 10.2.3.1 Covariance inequalities

In our proofs, we make use of the following key inequalities.

Lemma 10.2.7. *Let $P = \exp(-V)$ be a probability measure on $\mathbb{R}^d$ and assume that V is twice continuously differentiable on the interior of its domain. Then, the following hold.*

1. *(Brascamp–Lieb inequality) If in addition we assume that P is strictly log-concave, then it holds that*

$$\operatorname{cov}_{X\sim P}(X) \preceq \mathbb{E}_{X\sim P}[(\nabla^2 V(X))^{-1}] .$$

2. *(Cramér–Rao inequality)*

$$\operatorname{cov}_{X\sim P}(X) \succeq (\mathbb{E}_{X\sim P}[\nabla^2 V(X)])^{-1} .$$

The Brascamp–Lieb inequality is classical, and we refer readers to [BL00; BGL14; Cor17] for several proofs. To make our exposition more self-contained, we provide a proof of the Cramér–Rao inequality.

Proof of Lemma 10.2.7, Cramér–Rao inequality. For any smooth and compactly supported test function $h : \mathbb{R}^d \to \mathbb{R}$, integration by parts yields

$$\mathbb{E}_P \nabla h = \int \nabla h \, \mathrm{d}P = -\int (h \, \nabla \ln P) \, \mathrm{d}P = \int (h - \mathbb{E}_P h) \, \nabla V \, \mathrm{d}P$$

where we used the fact that $\mathbb{E}_P \nabla \ln P = 0$. Therefore,

$$\langle \mathbb{E}_P \nabla h, (\mathbb{E}_P \nabla^2 V)^{-1} \mathbb{E}_P \nabla h \rangle = \int (h - \mathbb{E}_P h) \, \langle \nabla V, (\mathbb{E}_P \nabla^2 V)^{-1} \mathbb{E}_P \nabla h \rangle \, \mathrm{d}P . \tag{10.22}$$

Applying the Cauchy–Schwarz inequality,

$$(10.22) \le \sqrt{(\operatorname{var}_P h) \int \langle \mathbb{E}_P \nabla h, (\mathbb{E}_P \nabla^2 V)^{-1} (\nabla V)^{\otimes 2} (\mathbb{E}_P \nabla^2 V)^{-1} \mathbb{E}_P \nabla h \rangle \, \mathrm{d}P} .$$

Integration by parts shows that $\int \nabla V^{\otimes 2} \, \mathrm{d}P = \int \nabla^2 V \, \mathrm{d}P$, and upon rearranging we deduce that

$$\operatorname{var}_P h \ge \langle \mathbb{E}_P \nabla h, (\mathbb{E}_P \nabla^2 V)^{-1} \mathbb{E}_P \nabla h \rangle . \tag{10.23}$$

By approximation, this continues to hold for any locally Lipschitz $h : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}_P \|\nabla h\| < \infty$.

Specializing the inequality (10.23) to $h := \langle e, \cdot \rangle$ for a unit vector $e \in \mathbb{R}^d$ then recovers the Cramér–Rao inequality of Lemma 10.2.7. □

10.2.4 Main theorem

We now state and prove our main theorem.

Theorem 10.2.8. *Let $P = \exp(-V)$ and $Q = \exp(-W)$.*

1. *Suppose that (P, Q) satisfy our regularity assumptions, as well as*

$$\nabla^2 V \preceq \beta_V I\,, \qquad \text{and} \qquad \nabla^2 W \succeq \alpha_W I \succ 0\,.$$

Then, for every $\varepsilon > 0$ and all $x \in \mathbb{R}^d$, the Hessian of the entropic Brenier potential satisfies

$$\nabla^2 \varphi_\varepsilon(x) \preceq \frac{1}{2}\left(\sqrt{4\beta_V/\alpha_W + \varepsilon^2 \beta_V^2} - \varepsilon \beta_V\right) I\,.$$

2. *Suppose that (Q, P) satisfy our regularity assumptions, as well as*

$$\nabla^2 V \succeq \alpha_V I \succ 0\,, \qquad \text{and} \qquad \nabla^2 W \preceq \beta_W I\,.$$

Then, for every $\varepsilon > 0$ and all $x \in \Omega_P := \operatorname{int}(\operatorname{supp}(P))$, the Hessian of the entropic Brenier potential satisfies

$$\nabla^2 \varphi_\varepsilon(x) \succeq \frac{1}{2}\left(\sqrt{4\alpha_V/\beta_W + \varepsilon^2 \alpha_V^2} - \varepsilon \alpha_V\right) I\,.$$

As $\varepsilon \searrow 0$, we formally expect the following bounds on the Brenier potential:

$$\sqrt{\alpha_V/\beta_W}\, I \preceq \nabla^2 \varphi_0(x) \preceq \sqrt{\beta_V/\alpha_W}\, I\,.$$

In particular, this recovers Caffarelli's contraction theorem (Theorem 10.2.1). We make this intuition rigorous below by appealing to convergence results for the entropic potentials as the regularization parameter ε tends to zero.

Proof of Theorem 10.2.8. **Upper bound.** Fix $x \in \mathbb{R}^d$. Recall from Lemma 10.2.6:

$$\nabla^2 \varphi_\varepsilon(x) = \varepsilon^{-1} \operatorname{cov}_{Y \sim \pi_\varepsilon^x}(Y)\,.$$

By an application of the Brascamp–Lieb inequality, this results in the upper bound

$$\begin{aligned} \nabla^2 \varphi_\varepsilon(x) &= \varepsilon^{-1} \operatorname{cov}_{Y \sim \pi_\varepsilon^x}(Y) \\ &\preceq \varepsilon^{-1}\, \mathbb{E}_{Y \sim \pi_\varepsilon^x}\left[\left(\varepsilon^{-1} \nabla^2 \psi_\varepsilon(Y) + \nabla^2 W(Y)\right)^{-1}\right] \\ &\preceq \mathbb{E}_{Y \sim \pi_\varepsilon^x}\left[\left(\nabla^2 \psi_\varepsilon(Y) + \varepsilon \alpha_W I\right)^{-1}\right], \end{aligned} \tag{10.24}$$

where in the last inequality we also used the lower bound on the spectrum of $\nabla^2 W$. Next, using Lemma 10.2.6 and the Cramér–Rao inequality (Lemma 10.2.7), we obtain the lower bound

$$\begin{aligned}\nabla^2 \psi_\varepsilon(Y) &= \varepsilon^{-1} \operatorname{cov}_{X \sim \pi_\varepsilon^Y}(X) \\ &\succeq \varepsilon^{-1} \left(\mathbb{E}_{X \sim \pi_\varepsilon^Y}\left[\varepsilon^{-1} \nabla^2 \varphi_\varepsilon(X) + \nabla^2 V(X)\right]\right)^{-1} \\ &\succeq \left(\mathbb{E}_{X \sim \pi_\varepsilon^Y}\left[\nabla^2 \varphi_\varepsilon(X) + \varepsilon \beta_V I\right]\right)^{-1},\end{aligned}$$

where we used the upper bound on the spectrum of $\nabla^2 V$. Combining these inequalities,

$$\nabla^2 \varphi_\varepsilon(x) \preceq \mathbb{E}_{Y \sim \pi_\varepsilon^x}\left[\left(\left(\mathbb{E}_{X \sim \pi_\varepsilon^Y}\left[\nabla^2 \varphi_\varepsilon(X) + \varepsilon \beta_V I\right]\right)^{-1} + \varepsilon \alpha_W I\right)^{-1}\right].$$

Now, define the quantity

$$L_\varepsilon \coloneqq \sup_{x \in \mathbb{R}^d} \lambda_{\max}\big(\nabla^2 \varphi_\varepsilon(x)\big).$$

From (10.24) and the fact that ψ_ε is convex (Lemma 10.2.6), it follows that L_ε is finite: $L_\varepsilon \le (\varepsilon \alpha_W)^{-1}$. Then, we have shown

$$\lambda_{\max}\big(\nabla^2 \varphi_\varepsilon(x)\big) \le \left((L_\varepsilon + \varepsilon \beta_V)^{-1} + \varepsilon \alpha_W\right)^{-1}.$$

Taking the supremum over $x \in \mathbb{R}^d$,

$$L_\varepsilon \le \left((L_\varepsilon + \varepsilon \beta_V)^{-1} + \varepsilon \alpha_W\right)^{-1}.$$

Solving the inequality yields

$$L_\varepsilon \le \frac{1}{2}\left(\sqrt{4\beta_V/\alpha_W + \varepsilon^2 \beta_V^2} - \varepsilon \beta_V\right). \tag{10.25}$$

Lower bound. The lower bound argument is symmetric, but we give the details for completeness. Using Lemma 10.2.6 and the Cramér–Rao inequality (Lemma 10.2.7),

$$\begin{aligned}\nabla^2 \varphi_\varepsilon(x) &= \varepsilon^{-1} \operatorname{cov}_{Y \sim \pi_\varepsilon^x}(Y) \\ &\succeq \varepsilon^{-1} \left(\mathbb{E}_{Y \sim \pi_\varepsilon^x}\left[\varepsilon^{-1} \nabla^2 \psi_\varepsilon(Y) + \nabla^2 W(Y)\right]\right)^{-1} \\ &\succeq \left(\mathbb{E}_{Y \sim \pi_\varepsilon^x}\left[\nabla^2 \psi_\varepsilon(Y) + \varepsilon \beta_W I\right]\right)^{-1}.\end{aligned}$$

Applying Lemma 10.2.6 and the Brascamp–Lieb inequality (Lemma 10.2.7),

$$\nabla^2 \psi_\varepsilon(Y) = \varepsilon^{-1} \operatorname{cov}_{X \sim \pi_\varepsilon^Y}(X)$$

$$\preceq \varepsilon^{-1}\,\mathbb{E}_{X\sim\pi_\varepsilon^Y}\left[\left(\varepsilon^{-1}\,\nabla^2\varphi_\varepsilon(X)+\nabla^2 V(X)\right)^{-1}\right]$$
$$\preceq \mathbb{E}_{X\sim\pi_\varepsilon^Y}\left[\left(\nabla^2\varphi_\varepsilon(X)+\varepsilon\alpha_V I\right)^{-1}\right].$$

Combining the two inequalities and setting

$$\ell_\varepsilon := \inf_{x\in\Omega_P}\lambda_{\min}\left(\nabla^2\varphi_\varepsilon(x)\right),$$

we deduce that

$$\ell_\varepsilon \geq \left(\left(\ell_\varepsilon+\varepsilon\alpha_V\right)^{-1}+\varepsilon\beta_W\right)^{-1}.$$

On the other hand, from Lemma 10.2.6, we know that $\ell_\varepsilon \geq 0$. Solving the inequality then yields

$$\ell_\varepsilon \geq \frac{1}{2}\left(\sqrt{4\alpha_V/\beta_W+\varepsilon^2\alpha_V^2}-\varepsilon\alpha_V\right). \qquad \square$$

Next, we rigorously deduce Caffarelli's contraction theorem.

Proof of Caffarelli's contraction (Theorem 10.2.1). For every $\varepsilon > 0$, by Theorem 10.2.8, we have proven that $\nabla^2\varphi_\varepsilon \preceq L_\varepsilon I$, with L_ε as in (10.25). Equivalently, this can be reformulated as saying that $\frac{L_\varepsilon\|\cdot\|^2}{2}-\varphi_\varepsilon$ is convex. Fix some $\delta > 0$; in particular, for ε sufficiently small, $\frac{(\sqrt{\beta_V/\alpha_W}+\delta)\|\cdot\|^2}{2}-\varphi_\varepsilon$ is convex.

Upon passing to a sequence $\varepsilon_k \searrow 0$, existing results on the convergence of entropic optimal transport potentials show that $\varphi_{\varepsilon_k}\to\varphi_0$ in $L^1(P)$ (see Theorem 10.2.5). Passing to a further subsequence, we obtain $\varphi_{\varepsilon_k}\to\varphi_0$ (P-almost surely). It follows that $\frac{(\sqrt{\beta_V/\alpha_W}+\delta)\|\cdot\|^2}{2}-\varphi_0$ is convex for every $\delta > 0$ (see the remark after [Roc97, Theorem 25.7]), and thus for $\delta = 0$. $\square$

Remark 10.2.9. *Our main theorem provides both upper and lower bounds for $\nabla^2\varphi_\varepsilon$. In the case when $\varepsilon = 0$, the lower bound follows from the upper bound. Indeed, if φ_0 is the Brenier potential for the optimal transport from P to Q, then the convex conjugate φ_0^* is the Brenier potential for the optimal transport from Q to P. By applying Caffarelli's contraction theorem to φ_0^* and appealing to convex duality, it yields a lower bound on $\nabla^2\varphi_0$. However, we are not aware of a method of deducing the lower bound from the upper bound for positive values of ε.*

Remark 10.2.10. *In §10.2.6, by inspecting the Gaussian case, we show that Theorem 10.2.8 is sharp for every $\varepsilon > 0$.*

Remark 10.2.11. *In the proof of Theorem 10.2.8, we do not use the full force of the Brascamp–Lieb inequality. Rather, we use the covariance inequality in Lemma 10.2.7 which is a corollary of the usual Brascamp–Lieb inequality obtained by applying it to linear test functions.*

An inspection of the proof of the upper bound in Theorem 10.2.8 reveals the following more general pair of inequalities.

Proposition 10.2.12. *Let (P, Q) be probability measures satisfying our regularity conditions. Then, for all $x \in \mathbb{R}^d$, $y \in \Omega_Q$,*

$$\nabla^2 \varphi_\varepsilon(x) \preceq \mathbb{E}_{Y \sim \pi_\varepsilon^x} \left[\left(\nabla^2 \psi_\varepsilon(Y) + \varepsilon \, \nabla^2 W(Y) \right)^{-1} \right],$$
$$\nabla^2 \psi_\varepsilon(y) \succeq \left(\mathbb{E}_{X \sim \pi_\varepsilon^y} \left[\nabla^2 \varphi_\varepsilon(X) + \varepsilon \, \nabla^2 V(X) \right] \right)^{-1}.$$

In the next section, we use these inequalities to prove a generalization of Caffarelli's theorem.

■ 10.2.5 A generalization to commuting positive definite matrices

In the next result, we replace the main assumptions of Caffarelli's theorem, namely $\nabla^2 V \preceq \beta_V I$ and $\nabla^2 W \succeq \alpha_W I$, by the conditions

$$\nabla^2 V \preceq A^{-1} \qquad \text{and} \qquad \nabla^2 W \succeq B^{-1}, \tag{10.26}$$

where A and B are commuting positive definite matrices. Recall that the Hessian of the Brenier potential between the Gaussian distributions $\mathsf{normal}(0, A)$ and $\mathsf{normal}(0, B)$ is the matrix $A^{-1/2} B^{1/2}$ [Gel90]. In light of this observation, the following theorem is sharp for every pair of commuting positive definite (A, B), and shows that the Brenier potential between Gaussians achieves the largest possible Hessian among all source and target measures obeying the constraint (10.26).

Theorem 10.2.13. *Let (P, Q) satisfy our regularity conditions as well as the condition (10.26). Then, the Hessian of the Brenier potential satisfies the uniform bound: for all $x \in \mathbb{R}^d$, it holds that*

$$\nabla^2 \varphi_0(x) \preceq A^{-1/2} B^{1/2}.$$

As in Theorem 10.2.8, the proof technique also yields a lower bound on $\nabla^2 \varphi_0$ under appropriate assumptions. We omit this result because it is straightforward.

Proof. Let C_ε be the smallest constant $C \geq 0$ such that $\nabla^2 \varphi_\varepsilon(x) \preceq A^{-1/2} B^{1/2} + CI$ for all $x \in \mathbb{R}^d$. In light of Theorem 10.2.8, C_ε is well-defined and finite:

$$C_\varepsilon = \sup_{x \in \mathbb{R}^d} \sup_{e \in \mathbb{R}^d, \, \|e\| = 1} \left\langle e, \left[\nabla^2 \varphi_\varepsilon(x) - A^{-1/2} B^{1/2} \right] e \right\rangle.$$

Let (x, e) achieve the above supremum. Using our assumptions and Proposition 10.2.12, we obtain

$$C_\varepsilon = \left\langle e, \left[\nabla^2 \varphi_\varepsilon(x) - A^{-1/2} B^{1/2} \right] e \right\rangle$$

$$\leq \left\langle e, \left[\left(\mathbb{E}_{Y \sim \pi_\varepsilon^x} \nabla^2 \psi_\varepsilon(Y) + \varepsilon B^{-1} \right)^{-1} - A^{-1/2} B^{1/2} \right] e \right\rangle$$
$$\leq \left\langle e, \left[\left((A^{-1/2} B^{1/2} + C_\varepsilon I + \varepsilon A^{-1})^{-1} + \varepsilon B^{-1} \right)^{-1} - A^{-1/2} B^{1/2} \right] e \right\rangle .$$

From our assumptions and Theorem 10.2.8, we know that the spectrum of $M_\varepsilon := A^{-1/2} B^{1/2} + C_\varepsilon I$ is bounded away from zero and infinity as $\varepsilon \searrow 0$, which justifies the Taylor expansion

$$\begin{aligned} \left((M_\varepsilon + \varepsilon A^{-1})^{-1} + \varepsilon B^{-1} \right)^{-1} &= \left(M_\varepsilon^{-1} - \varepsilon M_\varepsilon^{-1} A^{-1} M_\varepsilon^{-1} + \varepsilon B^{-1} + O(\varepsilon^2)\, I \right)^{-1} \\ &= M_\varepsilon + \varepsilon A^{-1} - \varepsilon M_\varepsilon B^{-1} M_\varepsilon + O(\varepsilon^2)\, I . \end{aligned}$$

Hence,

$$\begin{aligned} C_\varepsilon &\leq \left\langle e, \left[M_\varepsilon + \varepsilon A^{-1} - \varepsilon M_\varepsilon B^{-1} M_\varepsilon + O(\varepsilon^2)\, I - A^{-1/2} B^{1/2} \right] e \right\rangle \\ &\leq C_\varepsilon + \varepsilon \left\langle e, \left[A^{-1} - M_\varepsilon B^{-1} M_\varepsilon \right] e \right\rangle + O(\varepsilon^2) \\ &= C_\varepsilon - \varepsilon \left\langle e, \left[2 C_\varepsilon A^{-1/2} B^{-1/2} + C_\varepsilon^2 B^{-1} \right] e \right\rangle + O(\varepsilon^2) . \end{aligned}$$

This shows that $\lim_{\varepsilon \searrow 0} C_\varepsilon = 0$ (otherwise $(C_\varepsilon)_{\varepsilon > 0}$ would have a strictly positive cluster point which would contradict the above inequality for small enough $\varepsilon > 0$).

By combining this fact with convergence of the entropic Brenier potentials as in the proof of Theorem 10.2.1, we deduce the result. □

Next, we recover and extend a result of Valdimarsson [Val07], which was used to derive new forms of the Brascamp–Lieb inequality.[3]

Theorem 10.2.14. *Suppose that*

- $\bar{A}$, $\bar{B}$, *and* G *are positive definite matrices;*
- $\bar{A} \preceq G$ *and* $\bar{B}$ *commutes with* G*;*
- $P = \exp(-\tilde{V}) * \mu$, *where* $\nabla^2 \tilde{V} \preceq \bar{B}^{-1} G$, $*$ *denotes convolution, and* μ *is an arbitrary probability measure on* $\mathbb{R}^d$*;*
- $Q = \exp(-W)$ *with* $\nabla^2 W \succeq \bar{B}^{-1/2} \bar{A}^{-1} \bar{B}^{-1/2}$.

Then, the Brenier potential satisfies $\nabla^2 \varphi_0 \preceq G$.

Remark 10.2.15. *Valdimarsson's result required that* $P = \mathsf{normal}(0, \bar{B} G^{-1}) * \mu$.

To prove this result, we check that convolution with any probability measure only makes the density more log-smooth.

[3] This is a different Brascamp–Lieb inequality than the one in Lemma 10.2.7.

Lemma 10.2.16. *Let $\widetilde{P} \propto \exp(-\widetilde{V})$ be a probability measure, where $\widetilde{V} : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable. Let $P := \widetilde{P} * \mu = \exp(-V)$ where μ is any probability measure on $\mathbb{R}^d$. Suppose that for some positive definite matrix A^{-1}, we have $\nabla^2 \widetilde{V} \preceq A^{-1}$. Then, $\nabla^2 V \preceq A^{-1}$ as well.*

Proof. An elementary computation shows that if we define the probability measure

$$\nu_y(\mathrm{d}x) := \frac{\exp(-\widetilde{V}(y-x))\,\mu(\mathrm{d}x)}{\int \exp(-\widetilde{V}(y-x'))\,\mu(\mathrm{d}x')}$$

then

$$\nabla^2 V(y) = \mathbb{E}_{X\sim\nu_y}\left[\nabla^2 \widetilde{V}(y-X)\right] - \mathrm{cov}_{X\sim\nu_y}\left(\nabla \widetilde{V}(y-X)\right),$$

from which the result follows. □

Proof of Theorem 10.2.14. Under Lemma 10.2.16 and the third assumption, it holds that $P \propto \exp(-V)$ with $\nabla^2 V \preceq \bar{B}^{-1} G$. The other assumptions imply that $Q \propto \exp(-W)$ with

$$\nabla^2 W \succeq \bar{B}^{-1/2} \bar{A}^{-1} \bar{B}^{-1/2} \succeq \bar{B}^{-1/2} G^{-1} \bar{B}^{-1/2} = \bar{B}^{-1} G^{-1}.$$

By Theorem 10.2.13, it holds that $\nabla^2 \phi_0 \preceq G$. □

Remark 10.2.17. *It is natural to ask whether Theorem 10.2.13 can be obtained by first applying Caffarelli's contraction theorem to show that the optimal transport map $\widetilde{T}_0$ between the measures $(A^{-1/2})_{\#}P$ and $(B^{-1/2})_{\#}Q$ is 1-Lipschitz, and then considering the mapping $T_0(x) := B^{1/2}\widetilde{T}_0(A^{-1/2}x)$. Although T_0 is indeed a valid transport mapping from P to Q, under our assumptions ∇T_0 is not guaranteed to be symmetric, so it does not make sense to ask that $\nabla T_0 \preceq A^{-1/2}B^{1/2}$.*

In Valdimarsson's application to Brascamp–Lieb inequalities, it is crucial that the transport map T_0 is chosen so that ∇T_0 is symmetric and positive definite. Symmetry of ∇T_0 implies that T_0 is the gradient $\nabla \phi_0$ of a function $\phi_0 : \mathbb{R}^d \to \mathbb{R}$, and positive definiteness implies that ϕ_0 is convex. By Brenier's theorem, the unique gradient of a convex function that pushes forward P to Q is the optimal transport map. Thus, it is crucial that we consider the optimal *transport map here; alternative maps such as the ones in [KM12; MS21] cannot be applied.*

■ 10.2.6 Gaussian case

Suppose $P = \mathsf{normal}(0, A)$ and $Q = \mathsf{normal}(0, B)$ are Gaussians. Then, it is known that the Hessian of the Brenier potential is given by [Gel90]

$$\nabla^2 \varphi_0(x) = A^{-1/2}\left(A^{1/2} B A^{1/2}\right)^{1/2} A^{-1/2}.$$

If we have

$$A^{-1} \preceq \beta I \qquad \text{and} \qquad B^{-1} \succeq \alpha I \succ 0\,,$$

then Caffarelli's contraction theorem (Theorem 10.2.1) implies

$$\|\nabla^2 \phi_0\|_{\mathrm{op}} \leq \sqrt{\beta/\alpha}\,.$$

For $\varepsilon > 0$, the upper bound from Theorem 10.2.8 implies

$$\|\nabla^2 \phi_\varepsilon\|_{\mathrm{op}} \leq \frac{1}{2}\left(\sqrt{4\beta/\alpha + \varepsilon^2\beta^2} - \varepsilon\beta\right). \tag{10.27}$$

On the other hand, from [Jan+20; MGM22], it is known that

$$\nabla^2 \phi_\varepsilon(x) = A^{-1/2}\left(A^{1/2} B A^{1/2} + \frac{\varepsilon^2}{4}\, I\right)^{1/2} A^{-1/2} - \frac{\varepsilon}{2}\, A^{-1}\,.$$

In particular, if we take $A = \beta^{-1} I$ and $B = \alpha^{-1} I$, then (10.27) is an equality. Hence, Theorem 10.2.8 is sharp for every $\varepsilon > 0$.

Part III

Optimization and sampling without convexity

Chapter 11

Dimension-free log-Sobolev inequalities for mixtures

We now turn towards optimization and sampling without convexity assumptions. Recall that in §3, §4, and §6, we already initiated a study of non-log-concave sampling by assuming that the target distribution satisfies a functional inequality, such as a log-Sobolev inequality. However, even for the canonical case of a Gaussian mixture, sharp bounds on the log-Sobolev constant were unknown.

In this chapter, we prove that if $(P_x)_{x \in \mathscr{X}}$ is a family of probability measures which satisfy the log-Sobolev inequality and whose pairwise chi-squared divergences are uniformly bounded, and μ is any mixing distribution on $\mathscr{X}$, then the mixture $\int P_x \, \mathrm{d}\mu(x)$ satisfies a log-Sobolev inequality. In various settings of interest, the resulting log-Sobolev constant is dimension-free. In particular, our result implies a conjecture of Zimmermann and Bardet et al. that Gaussian convolutions of measures with bounded support enjoy dimension-free log-Sobolev inequalities.

This chapter is based on [CCN21], joint with Hong-Bin Chen and Jonathan Niles-Weed.

11.1 Introduction

Functional inequalities, such as the Poincaré inequality and the log-Sobolev inequality, have played a key role in the study of subjects such as concentration of measure and quantitative convergence analysis of Markov processes [BGL14; Han16] (in particular for spin systems [Mar99; Wei04]), as well as the geometry of metric measure spaces [Led00]. It is therefore of considerable interest to identify situations in which such inequalities hold, and furthermore to identify simple criteria which imply their validity.

We begin with a few motivating examples. Suppose that μ is a probability measure on $\mathbb{R}^d$ whose support is contained in the Euclidean ball of radius R, and let $\gamma_{0,t}$ denote the centered Gaussian distribution with variance tI_d. What

functional inequalities can we expect the convolution measure $\mu * \gamma_{0,t}$ to satisfy? This question, motivated by random matrix theory, was initiated in [Zim13; Zim16], and further investigated in [WW16; Bar+18]. These works prove that $\mu * \gamma_{0,t}$ satisfies both a Poincaré inequality and a log-Sobolev inequality; moreover, the Poincaré inequality holds with a constant depending only on R and t, and not on the dimension d. Furthermore, [Bar+18] conjectures that the same holds true for the log-Sobolev constant, and they verify the conjecture in special cases.

Another line of work [CM10; Sch19] studies the following question: let P_0 and P_1 be two probability measures on $\mathbb{R}^d$, and consider the mixture distribution $(1-p)\,P_0 + p\,P_1$ with mixing weight $p \in (0,1)$. If both P_0 and P_1 satisfy log-Sobolev inequalities, when does the mixture satisfy a log-Sobolev inequality too?

Although the two preceding examples may at first glance appear to be different in nature, we can in fact place them in the same framework, as follows. Let $(P_x)_{x \in \mathscr{X}}$ be a family of probability measures satisfying the log-Sobolev inequality, and let μ be a mixture distribution on $\mathscr{X}$; here, $\mathscr{X}$ may be finite or infinite. When does the mixture $\int P_x \,\mathrm{d}\mu(x)$ satisfy a log-Sobolev inequality?

- For the Gaussian convolution example, we take P_x to be the Gaussian distribution with mean x and variance tI_d.
- For the mixture example, we take μ to be the Bernoulli distribution with parameter p.

In this chapter, we identify general conditions which ensure that a mixture distribution satisfies a log-Sobolev inequality. Our main contribution can be summarized as follows.

Theorem 11.1.1 (Informal)**.** *Let $(P_x)_{x \in \mathscr{X}}$ be a family of probability measures satisfying the log-Sobolev inequality with a uniform constant C_1. Assume that the pairwise chi-squared divergences $\chi^2(P_x \,\|\, P_{x'})$ are uniformly bounded by C_2. Then, the mixture $\int P_x \,\mathrm{d}\mu(x)$ satisfies a log-Sobolev inequality with a constant depending only on C_1 and C_2.*

In fact, in our main result, we will relax the assumption that the chi-squared divergences are uniformly bounded into a moment condition; see Theorem 11.3.1. In turn, this will allow us to prove log-Sobolev inequalities for Gaussian convolutions of measures with sub-Gaussian tails, provided that the variance of the Gaussians is sufficiently large.

Crucially, the log-Sobolev constant has no dependence on the mixing distribution μ. As we show in §11.4, our general theorem yields dimension-free log-Sobolev inequalities in various settings; in particular, our result implies the conjecture of [Zim13; Zim16; Bar+18].

The rest of the chapter is organized as follows. In §11.2, we describe the setting of our general investigation and recall the definitions of a Poincaré inequality and a log-Sobolev inequality. We then state and prove our main theorem in §11.3.

In §11.4, we illustrate our general result in a number of applications. §11.4.1 is devoted to the proof of the aforementioned conjecture, and §11.4.2 and §11.4.3 generalize the result to Gaussian convolutions of measures with sub-Gaussian tails and other diffusion semigroups. In §11.4.4, we compare our results to prior work on functional inequalities for mixtures of two distributions. Then, in §11.4.5, we discuss analogues of our result on the Boolean hypercube.

■ 11.2 Background and notation

To state our results in a form that applies to both discrete and continuous mixture distributions, we adopt the general framework of [BGL14] and let Γ be a suitable notion of a gradient operator. More precisely, let $\mathscr{Y}$ be a Polish space equipped with the Borel σ-algebra $\mathcal{B}_{\mathscr{Y}}$, and let $\mathcal{A}$ be a subspace of bounded measurable functions on E containing all constant functions. Let $\Gamma : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$ be a symmetric bilinear operator satisfying $\Gamma(f, f) \geq 0$ everywhere on $\mathscr{Y}$ for every $f \in \mathcal{A}$. In addition, we require Γ to satisfy

$$\Gamma(1, 1) = 0\,, \tag{11.1}$$

where 1 is understood as a constant function. From bilinearity and positivity of Γ follows the Cauchy–Schwarz inequality

$$\Gamma(f, g)^2 \leq \Gamma(f, f)\,\Gamma(g, g) \qquad \text{for all } f, g \in \mathcal{A}\,,$$

which in turn shows that (11.1) is equivalent to the condition $\Gamma(1, f) = 0$ for all $f \in \mathcal{A}$. For brevity, we write $\Gamma(f) = \Gamma(f, f)$.

Important examples include the squared gradient $\Gamma(f) = \|\nabla f\|^2$ on $\mathbb{R}^d$, and $\Gamma(f) = \sum_{i=1}^d (D_i f)^2$ on a product space $\mathscr{X} = \mathcal{X}^d$, where $D_i f$ is given by

$$\begin{aligned} D_i f(x) \coloneqq \sup_{x_i' \in \mathcal{X}} & f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_d) \\ & - \inf_{x_i' \in \mathcal{X}} f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_d)\,. \end{aligned}$$

For any probability measure ρ on $(\mathscr{Y}, \mathcal{B}_{\mathscr{Y}})$, we write

$$\mathbb{E}_\rho[f] \coloneqq \int_{\mathscr{Y}} f \,\mathrm{d}\rho$$

for a ρ-integrable function f. In addition, we define

$$\begin{aligned}\operatorname{var}_\rho(f) &\coloneqq \mathbb{E}_\rho[(f - \mathbb{E}_\rho f)^2]\,,\\ \operatorname{ent}_\rho(g) &\coloneqq \mathbb{E}_\rho(g \log g) - \mathbb{E}_\rho g \log \mathbb{E}_\rho g\,,\end{aligned}$$

for suitable measurable functions f and g, with g non-negative. When there is no confusion, we often omit the brackets and parentheses in these expressions. If X is a random variable with law μ, we also write $\mathbb{E} f(X) = \mathbb{E}_\mu f$ and similarly for var and ent.

We say ρ satisfies a Poincaré inequality (PI) if there exists $C \geq 0$ such that

$$\operatorname{var}_\rho(f) \leq C\,\mathbb{E}_\rho\,\Gamma(f)\,, \qquad \forall f \in \mathcal{A}\,. \tag{PI}$$

The optimal constant in this inequality is denoted $C_{\mathrm{P}}(\rho)$. In addition, ρ is said to satisfy a logarithmic Sobolev inequality (LSI) if there exists $C \geq 0$ such that

$$\operatorname{ent}_\rho(f^2) \leq 2C\,\mathbb{E}_\rho\,\Gamma(f)\,, \qquad \forall f \in \mathcal{A}\,. \tag{LSI}$$

Similarly, we let $C_{\mathrm{LS}}(\rho)$ denote the optimal constant in this inequality.

For probability measures ρ_1 and ρ_2 on $(\mathscr{Y}, \mathcal{B}_{\mathscr{Y}})$, the Kullback–Leibler (KL) divergence and the chi-squared divergence are defined as

$$\begin{aligned}\mathsf{KL}(\rho_1 \,\|\, \rho_2) &\coloneqq \operatorname{ent}_{\rho_2}\Big(\frac{\mathrm{d}\rho_1}{\mathrm{d}\rho_2}\Big) = \int_{\mathscr{Y}} \frac{\mathrm{d}\rho_1}{\mathrm{d}\rho_2} \ln \frac{\mathrm{d}\rho_1}{\mathrm{d}\rho_2}\,\mathrm{d}\rho_2 = \int_{\mathscr{Y}} \Big(\ln \frac{\mathrm{d}\rho_1}{\mathrm{d}\rho_2}\Big)\,\mathrm{d}\rho_1\,,\\ \chi^2(\rho_1 \,\|\, \rho_2) &\coloneqq \operatorname{var}_{\rho_2}\Big(\frac{\mathrm{d}\rho_1}{\mathrm{d}\rho_2}\Big) = \int_{\mathscr{Y}} \big(\frac{\mathrm{d}\rho_1}{\mathrm{d}\rho_2} - 1\big)^2\,\mathrm{d}\rho_2 = \int_{\mathscr{Y}} \frac{\mathrm{d}\rho_1}{\mathrm{d}\rho_2}\,\mathrm{d}\rho_1 - 1\,.\end{aligned}$$

The expressions above are understood to be $+\infty$ if ρ_1 is not absolutely continuous w.r.t. ρ_2.

■ 11.3 Main theorem

In addition to $(\mathscr{Y}, \mathcal{B}_{\mathscr{Y}})$, let $\mathscr{X}$ be a polish space with Borel σ-algebra $\mathcal{B}_{\mathscr{X}}$. We consider a Markov kernel $P : \mathscr{X} \times \mathcal{B}_{\mathscr{Y}} \to [0,1]$ satisfying: (1) for each $x \in \mathscr{X}$, $P(x,\cdot)$ is a probability measure on $(\mathscr{Y}, \mathcal{B}_{\mathscr{Y}})$, and (2) for each $B \in \mathcal{B}_{\mathscr{Y}}$, $P(\cdot, B)$ is a $\mathcal{B}_{\mathscr{X}}$-measurable function on $\mathscr{X}$. We also write $P_x \coloneqq P(x,\cdot)$ for convenience. This kernel naturally induces a transition map which maps bounded measurable functions on $\mathscr{X}$ to bounded measurable functions on $\mathscr{Y}$:

$$Pf(x) \coloneqq \int_{\mathscr{X}} f\,\mathrm{d}P_x\,, \qquad \forall x \in \mathscr{X}\,.$$

For a probability measure μ on $(\mathscr{X}, \mathcal{B}_{\mathscr{X}})$, we denote by μP the probability measure on $(\mathscr{Y}, \mathcal{B}_{\mathscr{Y}})$ defined by the duality

$$\int_{\mathscr{Y}} f \, \mathrm{d}\mu P = \int_{\mathscr{X}} P f \, \mathrm{d}\mu \, .$$

Lastly, we introduce the following quantities.

$$K_{\mathrm{P}}(P;\mu) \coloneqq \operatorname{ess\,sup}_{\mu\text{-a.s. } x \in \mathscr{X}} C_{\mathrm{P}}(P_x) \, , \tag{11.2}$$

$$K_{\mathrm{LS}}(P;\mu) \coloneqq \operatorname{ess\,sup}_{\mu\text{-a.s. } x \in \mathscr{X}} C_{\mathrm{LS}}(P_x) \, , \tag{11.3}$$

$$K_{p,\chi^2}(P;\mu) \coloneqq \mathbb{E}\big[\big(1 + \chi^2(P_X \parallel P_{X'})\big)^p\big]^{\frac{1}{p}} \, , \tag{11.4}$$

for $p \geq 1$, where X and X' are i.i.d. with law μ. Since (LSI) implies (PI) with the same constant, we have $K_{\mathrm{P}}(P;\mu) \leq K_{\mathrm{LS}}(P;\mu)$. Throughout, for $p \geq 1$, we set $p^* = \frac{p}{p-1}$ to be the dual exponent.

Theorem 11.3.1.

1. *If $K_{\mathrm{P}}(P;\mu)$ and $K_{p,\chi^2}(P;\mu)$ are finite for some $p > 1$, then μP satisfies (PI) with constant*
$$C_{\mathrm{P}}(\mu P) \leq K_{\mathrm{P}} \left\{ p^* + K_{p,\chi^2}^{p^*} \right\} ,$$
where $K_{\mathrm{P}} = K_{\mathrm{P}}(P;\mu)$ and $K_{p,\chi^2} = K_{p,\chi^2}(P;\mu)$.

2. *If $K_{\mathrm{LS}}(P;\mu)$ and $K_{p,\chi^2}(P;\mu)$ are finite for some $p > 1$, then μP satisfies (LSI) with constant*
$$C_{\mathrm{LS}}(\mu P) \leq 3 K_{\mathrm{LS}} \left(p^* + K_{p,\chi^2}^{p^*} \right) \left(1 + \log K_{p,\chi^2}^{p^*} \right) ,$$
where $K_{\mathrm{LS}} = K_{\mathrm{LS}}(P;\mu)$ and $K_{p,\chi^2} = K_{p,\chi^2}(P;\mu)$.

Remark 11.3.2. *Our theorem is stated with a simpler constant for readability. A slightly sharper constant can be read off from the proof. Our results clearly extend to the case $p = \infty$ ($p^* = 1$) with*

$$K_{\infty,\chi^2}(P;\mu) \coloneqq 1 + \operatorname{ess\,sup}_{\mu\text{-a.s. } x,x' \in \mathscr{X}} \chi^2(P_x \parallel P_{x'}) \, .$$

For both steps, our starting point is to apply classical decompositions for the variance and the entropy, which have been used to prove functional inequalities for spin systems (see, e.g., the appendix of [Wei04]). If X is a random variable drawn according to μ, then

$$\operatorname{var}_{\mu P} f = \mathbb{E} \operatorname{var}_{P_X} f + \operatorname{var} \mathbb{E}_{P_X} f \, , \tag{11.5}$$

$$\operatorname{ent}_{\mu P} f^2 = \mathbb{E} \operatorname{ent}_{P_X} f^2 + \operatorname{ent} \mathbb{E}_{P_X} f^2 . \tag{11.6}$$

In both of these decompositions, the first term is easy to handle because we can apply the PI, resp. LSI, for the family $(P_x)_{x \in \mathscr{X}}$ inside the expectation. The crux of the proof is therefore the second terms.

Proof of Theorem 11.3.1 (1). In the case $p = \infty$ (i.e., the pairwise chi-squared divergences are uniformly bounded), the Poincaré inequality can be proven via a straightforward generalization of [Bar+18]. However, the case $1 < p < \infty$ requires non-trivial modifications, and we present a complete proof.

Let X be a random variable with law μ. As described above, we use the decomposition (11.5), and we focus on the problematic second term

$$\operatorname{var} \mathbb{E}_{P_X} f = \mathbb{E}[|\mathbb{E}_{P_X} f - \mathbb{E}_{\mu P} f|^2] .$$

The finiteness of $K_{p,\chi^2}(P;\mu)$ implies that P_x and $P_{x'}$ are mutually absolutely continuous for μ-a.e. x, x'. In particular, it implies that the Radon–Nikodym derivatives $\frac{\mathrm{d}\mu P}{\mathrm{d}P_X}$ and $\frac{\mathrm{d}P_X}{\mathrm{d}\mu P}$ are well-defined almost surely. We can therefore write

$$\begin{aligned} \mathbb{E}_{P_X} f - \mathbb{E}_{\mu P} f &= \int f \left(1 - \frac{\mathrm{d}\mu P}{\mathrm{d}P_X}\right) \mathrm{d}P_X \\ &= -\int f \left(1 - \frac{\mathrm{d}P_X}{\mathrm{d}\mu P}\right) \mathrm{d}\mu P . \end{aligned}$$

For brevity, we write $\chi^2_{\rho,\rho'} \coloneqq \chi^2(\rho \,\|\, \rho')$. Applying the Cauchy–Schwarz inequality to the above display, we have

$$\begin{aligned} \operatorname{var} \mathbb{E}_{P_X} f &\le \mathbb{E} \min\{(\operatorname{var}_{\mu P} f)\, \chi^2_{P_X,\mu P},\ (\operatorname{var}_{P_X} f)\, \chi^2_{\mu P, P_X}\} \\ &\le \mathbb{E}\big[(\operatorname{var}_{\mu P} f)^{1/p} \, (\chi^2_{P_X,\mu P})^{1/p} \, (\operatorname{var}_{P_X} f)^{1/p^*} \, (\chi^2_{\mu P, P_X})^{1/p^*}\big] . \end{aligned}$$

Then, Young's inequality implies that for all $\lambda > 0$,

$$\operatorname{var} \mathbb{E}_{P_X} f \le \frac{\lambda^p}{p} (\operatorname{var}_{\mu P} f)\, \mathbb{E}\big[(\chi^2_{P_X,\mu P})\, (\chi^2_{\mu P, P_X})^{p-1}\big] + \frac{\lambda^{-p^*}}{p^*} \mathbb{E} \operatorname{var}_{P_X} f .$$

Setting

$$\lambda = \mathbb{E}\big[(\chi^2_{P_X,\mu P})\, (\chi^2_{\mu P, P_X})^{p-1}\big]^{-\frac{1}{p}}$$

and substituting the above into (11.5) yields

$$\operatorname{var}_{\mu P} f \le \big\{p^* + \mathbb{E}\big[(\chi^2_{P_X,\mu P})\, (\chi^2_{\mu P, P_X})^{p-1}\big]^{\frac{1}{p-1}}\big\}\, \mathbb{E} \operatorname{var}_{P_X} f .$$

The joint convexity of the chi-squared divergence follows from, e.g., its dual characterization as a supremum of linear functionals (see, e.g., [AGS08, Lemma 9.4.4]), which implies

$$\max\{\mathbb{E}\,\chi^2_{P_X,\mu P}, \mathbb{E}\,\chi^2_{\mu P,P_X}\} \le \mathbb{E}\,\chi^2_{P_X,P_{X'}}\,,$$

where X' is an i.i.d. copy of X. Hölder's inequality then implies

$$\mathbb{E}\big[(\chi^2_{P_X,\mu P})\,(\chi^2_{\mu P,P_X})^{p-1}\big] \le \mathbb{E}\big[(\chi^2_{P_X,P_{X'}})^p\big]\,.$$

The desired result follows from the definitions of $K_{\mathrm{P}}(P;\mu)$ in (11.3) and $K_{p,\chi^2}(P;\mu)$ in (11.4). □

To prove the second assertion in Theorem 11.3.1, we derive a so-called defective LSI for μP, which can be tightened to yield a full LSI. In order to control the second term in (11.6), we need a lemma.

Lemma 11.3.3. *Let π and ρ be two probability measures. Then, the following holds for every non-negative function f satisfying $\mathbb{E}_\pi(f) < \infty$:*

$$\mathbb{E}_\pi\, f \log \frac{\mathbb{E}_\pi\, f}{\mathbb{E}_\rho\, f} \le \mathrm{ent}_\pi(f) + \mathbb{E}_\pi(f) \log\big(1+\chi^2(\pi\,\|\,\rho)\big)\,,$$

where by convention both sides vanish if $\mathbb{E}_\pi\, f = 0$.

Proof. Recall the Donsker–Varadhan theorem[1]: for any probability measures μ and ν, it holds

$$\mathsf{KL}(\mu\,\|\,\nu) = \sup_g \{\mathbb{E}_\mu\, g - \log \mathbb{E}_\nu \exp(g)\}\,, \tag{11.7}$$

where the supremum is taken over all g for which the expectations on the right side make sense.

We may assume that π is absolutely continuous with respect to ρ and that $\mathbb{E}_\pi(f\log f) < \infty$; otherwise, the expression on the right side is infinite. Since the expression is vacuous if $\mathbb{E}_\pi\, f = 0$, we may assume that $0 < \mathbb{E}_\pi\, f < \infty$, and, since each term in the lemma statement is homogeneous in f, we may assume without loss of generality that $\mathbb{E}_\pi\, f = 1$.

Define a new probability measure π_f by $\frac{\mathrm{d}\pi_f}{\mathrm{d}\pi} = f$. Then,

$$\begin{aligned}\mathbb{E}_\pi\Big[f \log \frac{f}{\mathbb{E}_\rho\, f}\Big] = \mathbb{E}_{\pi_f} \log \frac{f}{\mathbb{E}_\rho\, f} &\le \mathsf{KL}(\pi_f\,\|\,\rho) + \log \mathbb{E}_\rho \exp \log \frac{f}{\mathbb{E}_\rho\, f} \\ &= \mathsf{KL}(\pi_f\,\|\,\rho)\,,\end{aligned}$$

[1] See [RS15, Theorem 5.4] or [DZ10, Lemma 6.2.13].

where we have used (11.7). Since

$$\mathsf{KL}(\pi_f \,\|\, \rho) = \mathbb{E}_\pi\Big[f \log\big(f \,\frac{\mathrm{d}\pi}{\mathrm{d}\rho}\big)\Big]\,,$$

subtracting $\mathbb{E}_\pi(f \log f)$ from both sides of the inequality above and recalling that we have assumed that $\mathbb{E}_\pi f = 1$ yields

$$\mathbb{E}_\pi f \log \frac{\mathbb{E}_\pi f}{\mathbb{E}_\rho f} \le \mathbb{E}_\pi\Big[f \log \frac{\mathrm{d}\pi}{\mathrm{d}\rho}\Big]\,.$$

Continuing, we have again by (11.7) that

$$\begin{aligned}\mathbb{E}_\pi\Big[f \log \frac{\mathrm{d}\pi}{\mathrm{d}\rho}\Big] = \mathbb{E}_{\pi_f} \log \frac{\mathrm{d}\pi}{\mathrm{d}\rho} &\le \mathsf{KL}(\pi_f \,\|\, \pi) + \log \mathbb{E}_\pi \exp \log \frac{\mathrm{d}\pi}{\mathrm{d}\rho} \\ &= \mathbb{E}_\pi(f \log f) + \log\big(1 + \chi^2(\pi \,\|\, \rho)\big)\,,\end{aligned}$$

as claimed. □

Proof of Theorem 11.3.1 (2). Assume that $p < \infty$. If not, we may apply the argument below with p finite and send $p \to \infty$ to obtain the desired bound.

Let X, X' be i.i.d. copies with law μ. The second term $\operatorname{ent} \mathbb{E}_{P_X}(f^2)$ in (11.6) can be written as

$$\operatorname{ent} \mathbb{E}_{P_X}(f^2) = \mathbb{E}\Big[\mathbb{E}_{P_X}(f^2) \log \frac{\mathbb{E}_{P_X}(f^2)}{\mathbb{E}_{\mu P}(f^2)}\Big]\,.$$

Setting $\pi = P_X$ and $\rho = \mu P$ in Lemma 11.3.3, we obtain

$$\operatorname{ent} \mathbb{E}_{P_X}(f^2) \le \mathbb{E} \operatorname{ent}_{P_X}(f^2) + \mathbb{E}\big[\mathbb{E}_{P_X}(f^2) \log\big(1 + \chi^2(P_X \,\|\, P_{X'})\big)\big] \tag{11.8}$$

where we also used the convexity of the chi-squared divergence in the second inequality. The definition of $K_{p,\chi^2}(P;\mu)$ in (11.4) ensures that

$$\mathbb{E} \exp\big\{p \log\big(1 + \chi^2(P_X \,\|\, P_{X'})\big) - p \log K_{p,\chi^2}(P;\mu)\big\} = 1\,.$$

Using the variational principle for the entropy [Han16, Lemma 3.15]:

$$\operatorname{ent} Y = \sup\{\mathbb{E}(YZ) \mid Z \text{ is a random variable with } \mathbb{E} \exp Z = 1\}\,,$$

we obtain

$$\mathbb{E}\big[\mathbb{E}_{P_X}(f^2) \log\big(1 + \chi^2(P_X \,\|\, P_{X'})\big)\big]$$

$$\leq \frac{1}{p} \operatorname{ent} \mathbb{E}_{P_X}(f^2) + \log K_{p,\chi^2}(P;\mu)\, \mathbb{E}_{\mu P}(f^2)\,.$$

Substituting this into (11.8) yields

$$\operatorname{ent} \mathbb{E}_{P_X}(f^2) \leq p^* \left\{ \mathbb{E} \operatorname{ent}_{P_X}(f^2) + \log K_{p,\chi^2}(P;\mu)\, \mathbb{E}_{\mu P}(f^2) \right\}.$$

We insert this into (11.6) to obtain:

$$\begin{aligned} \operatorname{ent}_{\mu P}(f^2) &\leq (p^*+1)\, \mathbb{E} \operatorname{ent}_{P_X}(f^2) + p^* \log K_{p,\chi^2}(P;\mu)\, \mathbb{E}_{\mu P}(f^2) \\ &\leq 4p^* K_{\mathrm{LS}}(P;\mu)\, \mathbb{E}_{\mu P}\, \Gamma(f) + p^* \log K_{p,\chi^2}(P;\mu)\, \mathbb{E}_{\mu P}(f^2)\,. \end{aligned}$$

This inequality is known as a *defective LSI* (see [BGL14, §5]). It is standard that a defective LSI together with a Poincaré inequality implies a full LSI; this is known as *tightening* the LSI, and we refer to §11.5 for details. Together with the PI in the first assertion of Theorem 11.3.1, this completes the proof. □

11.4 Applications

11.4.1 Gaussian convolutions

Set $\mathscr{Y} = \mathbb{R}^d$, $\mathcal{A} = \mathcal{C}_{\mathrm{b}}^\infty(\mathbb{R}^d)$ (infinitely differentiable functions with bounded derivatives), and $\Gamma(f) = \|\nabla f\|^2$. Let μ be a probability measure supported on $B(0,R) := \{x \in \mathbb{R}^d : \|x\| \leq R\}$, and for $x \in \mathbb{R}^d$ and $t > 0$, let

$$\gamma_{x,t}(y) = \frac{1}{(2\pi t)^{\frac{d}{2}}} \exp\Big(-\frac{\|y-x\|^2}{2t}\Big)$$

be the Gaussian with mean x and variance tI_d. If we take $P_x = \gamma_{x,t}$, then the measure μP is the convolution $\mu * \gamma_{0,t}$.

Functional inequalities for the measure $\mu * \gamma_{0,t}$ were studied in [Zim13; Zim16], and further investigated in [WW16; Bar+18]. In particular, [Bar+18] proves that $C_{\mathrm{P}}(\mu * \gamma_{0,t})$ is bounded above by a function of R and t, and is therefore dimension-free.

For the log-Sobolev constant, these works also show that $C_{\mathrm{LS}}(\mu * \gamma_{0,t})$ is finite, but the precise dependence of this constant (in particular on the dimension) was previously unknown. Bardet et al. [Bar+18] verify in several cases that $C_{\mathrm{LS}}(\mu * \gamma_{0,t})$ is dimension-free, and they conjecture that this is true in general. We now show that their conjecture is an immediate consequence of Theorem 11.3.1.

It is well-known that $\gamma_{x,t}$ satisfies (LSI) with $C_{\mathrm{LS}}(\gamma_{x,t}) = t$. Also, for $x, x' \in \mathbb{R}^d$ and $t \geq 0$, a straightforward computation shows that

$$\chi^2(\gamma_{x,t} \,\|\, \gamma_{x',t}) = \exp\Big(\frac{\|x-x'\|^2}{t}\Big) - 1\,.$$

Hence, $K_{\infty,\,\chi^2}(P;\mu) \le \exp(4R^2/t)$ and we deduce the following result.

Corollary 11.4.1. *Let μ be a probability measure on $\mathbb{R}^d$ supported on $B(0,R)$ for some $R \ge 0$. Then, for each $t \ge 0$, $\mu * \gamma_{0,t}$ satisfies (LSI) with*

$$C_{\mathrm{LS}}(\mu * \gamma_{0,t}) \le 6\,(4R^2 + t)\exp\Big(\frac{4R^2}{t}\Big)\,.$$

Bardet et al. also prove that $\mu * \gamma_{0,t}$ satisfies a T_2 transport-entropy inequality with a dimension-dependent constant; see [Vil03, §9.3] for the relevant background. Since a log-Sobolev inequality implies a T_2 inequality with the same constant [OV00], we immediately obtain the following improvement.

Corollary 11.4.2. *Let μ be a probability measure on $\mathbb{R}^d$ supported on $B(0,R)$ for some $R \ge 0$. Then, for each $t \ge 0$, $\mu * \gamma_{0,t}$ satisfies a T_2 transport-entropy inequality with constant*

$$C_{\mathsf{T}_2}(\mu * \gamma_{0,t}) \le 6\,(4R^2 + t)\exp\Big(\frac{4R^2}{t}\Big)\,.$$

Remark 11.4.3. *These results show that evolving a compactly supported measure for a short time under the heat flow yields dimension-free functional inequalities, which can be interpreted as a strong regularizing effect of the heat flow. This is in line with other results on the smoothing behavior of the heat flow, e.g., [EL18].*

Remark 11.4.4 (Sharpness of the result)**.** *As $t \to \infty$, Corollary 11.4.1 implies*

$$\limsup_{t\to\infty} \frac{C_{\mathrm{LS}}(\mu * \gamma_{0,t})}{t} \le 6\,.$$

It is easy to improve this to 1, which is sharp. Indeed, from the subadditivity of the log-Sobolev constant under convolution, for $t \ge 4R^2$,

$$C_{\mathrm{LS}}(\mu * \gamma_{0,t}) \le C_{\mathrm{LS}}(\mu * \gamma_{0,4R^2}) + C_{\mathrm{LS}}(\gamma_{0,t-4R^2}) \le t + 130\,R^2\,.$$

On the other hand, as $t \searrow 0$, the exponential dependence on R^2/t cannot be avoided, as a simple example shows. Indeed, consider the measure $\mu = \frac{1}{2}\,\delta_{-R} + \frac{1}{2}\,\delta_R$ in one dimension and $0 < t \ll R$. Define the function $f : \mathbb{R} \to [-1,1]$ via

$$f(x) := \begin{cases} -1 \text{ for } x < -R/2, \\ +1 \text{ for } x > +R/2, \\ \text{linear interpolation in between.} \end{cases}$$

Let g denote a standard Gaussian variable. Then, $\mathbb{E}_{\mu*\gamma_{0,t}} f = 0$, so

$$\begin{aligned}\operatorname{var}_{\mu*\gamma_{0,t}} f &= \mathbb{E}_{\mu*\gamma_{0,t}}(f^2) \\ &\geq \frac{1}{2}\,\mathbb{P}\{-R+\sqrt{t}\,g \leq -\frac{R}{2}\} + \frac{1}{2}\,\mathbb{P}\{R+\sqrt{t}\,g \geq \frac{R}{2}\} \\ &= \mathbb{P}\{g \leq \frac{R}{2\sqrt{t}}\} \geq \frac{1}{2}\,.\end{aligned}$$

On the other hand, $|f'| = 2/R$ on $[-R/2, R/2]$, so

$$\mathbb{E}_{\mu*\gamma_{0,t}}(|f'|^2) \leq \frac{4}{R^2}\,\mathbb{P}\{g \geq \frac{R}{2\sqrt{t}}\} \leq \frac{2}{R^2}\exp\bigl(-\frac{R^2}{8t}\bigr)\,,$$

by standard Gaussian tail bounds. This yields the following lower bound on the Poincaré constant of $\mu * \gamma_{0,t}$:

$$C_{\mathrm{P}}(\mu * \gamma_{0,t}) \geq \frac{1}{4}\,R^2 \exp\frac{R^2}{8t}\,.$$

Hence, the exponential dependence on R^2/t is already present in the Poincaré constant. However, it is worth noting that the $\exp(4R^2/t)$ *dependence in the log-Sobolev constant enters* only *via the Poincaré constant through the method of tighening a defective log-Sobolev inequality. In particular, if μ is known* a priori *to satisfy a Poincaré inequality with constant $C_{\mathrm{P}}(\mu)$, then $\mu * \gamma_{0,t}$ satisfies a Poincaré inequality with constant $C_{\mathrm{P}}(\mu * \gamma_{0,t}) \leq C_{\mathrm{P}}(\mu) + t$, and the log-Sobolev inequality no longer suffers an explicit exponential dependence on R^2/t.*

■ 11.4.2 Extension to sub-Gaussian tails

Consider the setting in the previous section. However, we now relax the assumption that μ has bounded support, and instead assume that μ has sub-Gaussian tails. More specifically, assume that there exist constants $\sigma^2, C_{\mathrm{SG}}$ such that

$$\iint \exp\bigl(\frac{\|x-x'\|^2}{\sigma^2}\bigr)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x') \leq C_{\mathrm{SG}}\,. \tag{11.9}$$

Since a log-Sobolev inequality implies sub-Gaussian tails [BGL14, §5.4], the existence of such constants σ^2, C_{SG} are certainly necessary in order for $\mu * \gamma_{0,t}$ to satisfy (LSI). We will show that if t is greater than σ^2, then we indeed obtain a log-Sobolev constant for $\mu * \gamma_{0,t}$, and we will explicitly estimate the constant.

The main point is to estimate, for X, X' i.i.d. from μ,

$$\mathbb{E}\bigl[\{1+\chi^2(\gamma_{X,t} \parallel \gamma_{X',t})\}^p\bigr] = \iint \exp\bigl(\frac{p\,\|x-x'\|^2}{t}\bigr)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x') \leq C_{\mathrm{SG}}\,,$$

provided that $t/p \geq \sigma^2$; then, $K_{p,\,\chi^2}(P;\mu)^{p^*} \leq C_{\mathrm{SG}}^{p^*/p}$. We therefore take $p = t/\sigma^2$ and we obtain as an immediate consequence of Theorem 11.3.1 the following.

Theorem 11.4.5. *Suppose μ is a probability measure on $\mathbb{R}^d$ satisfying (11.9) and that $t > \sigma^2$. Then, $\mu * \gamma_{0,t}$ satisfies both (PI) and (LSI), with*

$$C_{\mathrm{P}}(\mu * \gamma_{0,t}) \leq t \left\{ \frac{t}{t-\sigma^2} + C_{\mathrm{SG}}^{\sigma^2/(t-\sigma^2)} \right\},$$

and

$$C_{\mathrm{LS}}(\mu * \gamma_{0,t}) \leq 3t \left\{ \frac{t}{t-\sigma^2} + C_{\mathrm{SG}}^{\sigma^2/(t-\sigma^2)} \right\} \left\{ 1 + \frac{\sigma^2}{t-\sigma^2} \log C_{\mathrm{SG}} \right\}.$$

Remark 11.4.6. *The first part of Theorem 11.4.5 was observed without proof in [Cou20].*

Remark 11.4.7. *The result of Theorem 11.4.5 recovers the result of Corollary 11.4.1, albeit with worse constants. Indeed, if μ has support contained in the ball $B(0,R)$ and $t > 0$, then we can take $\sigma^2 = t/2$ and*

$$C_{\mathrm{SG}} = \iint \exp\Big(\frac{2\,\|x-x'\|^2}{t}\Big)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x') \leq \exp\frac{8R^2}{t}\,.$$

Then, Theorem 11.4.5 yields a log-Sobolev inequality for $\mu * \gamma_{0,t}$ with a similar dependence as Corollary 11.4.1.

Remark 11.4.8. *The sub-Gaussian tail condition (11.9) is equivalent to μ satisfying a T_1 transportation-cost inequality [BV05]. Hence, our result shows that sufficient Gaussian smoothing upgrades a T_1 inequality to a log-Sobolev inequality.*

Note that the condition $t > \sigma^2$ is similar to the one in [WW16, Theorem 1.2].

Remark 11.4.9. *As in Remark 11.4.4, the Poincaré and log-Sobolev constants here can easily be improved when $t \to \infty$ to improve the constant factor in front of t to 1.*

■ 11.4.3 General diffusions

We now consider a different extension of the setting in §11.4.1. Let $(P^t)_{t\geq 0}$ be a Markov semigroup on $(\mathscr{Y}, \mathcal{B}_{\mathscr{Y}})$ with invariant measure π and infinitesimal generator $\mathscr{L}$. Let $\mathcal{A}$ be an algebra of bounded measurable functions such that $\mathcal{A}$ is dense in $L^2(\mathscr{Y},\pi)$; $\mathcal{A}$ is contained in the domain of $\mathscr{L}$; and the *carré du champ* operator $\Gamma : \mathcal{A}\times\mathcal{A} \to \mathcal{A}$ given by

$$\Gamma(f,g) = \frac{1}{2}\big(\mathscr{L}(fg) - f\mathscr{L}g - g\mathscr{L}f\big)$$

is well defined for $f, g \in \mathcal{A}$. We assume these objects satisfy the conditions specified in [BGL14, §1.14] so that results therein are applicable. For $\kappa \in \mathbb{R}$ and $t \geq 0$, we set

$$C_{\mathrm{loc}}(\kappa, t) := \begin{cases} (1 - \exp(-2\kappa t))/\kappa\,, & \kappa \neq 0\,, \\ 2t\,, & \kappa = 0\,. \end{cases}$$

We recall the following result ([BGL14, Theorem 5.5.2]).

Lemma 11.4.10. *For every $\kappa \in \mathbb{R}$, the following statements are equivalent.*

1. *The curvature-dimension condition* $\mathrm{CD}(\kappa, \infty)$ *holds.*
2. *For all $x \in \mathscr{X}$ and $t \geq 0$,*

$$C_{\mathrm{LS}}(P_x^t) \leq C_{\mathrm{loc}}(\kappa, t)\,.$$

The following result is then a special case of Theorem 11.3.1.

Corollary 11.4.11. *Suppose that the curvature-dimension condition* $\mathrm{CD}(\kappa, \infty)$ *holds for some $\kappa \in \mathbb{R}$. Let μ be a probability measure on $(\mathscr{Y}, \mathcal{B}_{\mathscr{Y}})$. Then, for every $t \geq 0$, it holds:*

$$C_{\mathrm{LS}}(\mu P^t) \leq 6 C_{\mathrm{loc}}(\kappa, t)\, K_{\infty,\chi^2}(P^t; \mu)\, \{1 + \log K_{\infty,\chi^2}(P^t; \mu)\}\,.$$

Remark 11.4.12. *If $\mathscr{Y}$ is a complete connected Riemannian manifold and the diffusion has generator $\mathscr{L} = \Delta - \langle \nabla V, \nabla \cdot \rangle$ which satisfies the curvature-dimension condition, then under mild conditions the constant $K_{\infty,\chi^2}(P^t; \mu)$ is finite for any measure μ with bounded support, as a consequence of heat kernel estimates in [GW01].*

■ 11.4.4 Mixtures of two distributions

In this section, we consider the case when $\mathscr{X} = \{0, 1\}$ is the two-point space. Then, the mixing distribution μ is a Bernoulli distribution with a mixing weight $p \in [0, 1]$, and the measure μP is the convex combination

$$\mu P = (1 - p) P_0 + p P_1\,. \tag{11.10}$$

Functional inequalities for such mixtures were studied in [CM10; Sch19]. One of the interesting findings of these papers is that as the mixing weight p tends to $\{0, 1\}$, the Poincaré constant can remain bounded whereas the log-Sobolev constant diverges logarithmically. Specifically, [Sch19] shows that if P_0 and P_1 satisfy (LSI), $p \in (0, 1)$, and either $\chi^2(P_0 \,\|\, P_1)$ or $\chi^2(P_1 \,\|\, P_0)$ is finite, then μP

satisfies (LSI). Note that this last assumption is weaker than ours, which requires both $\chi^2(P_0 \,\|\, P_1)$ and $\chi^2(P_1 \,\|\, P_0)$ to be finite. However, even under our stronger assumption, the bound of [Sch19] on the log-Sobolev constant diverges in general as $p \to \{0, 1\}$.

We now present our results for this setting for comparison.

Corollary 11.4.13. *For all $p \in [0, 1]$, the mixture (11.10) satisfies (LSI) with*

$$C_{\mathrm{LS}}(\mu P) \le 6 \max\{C_{\mathrm{LS}}(P_0), C_{\mathrm{LS}}(P_1)\}\, K_{\chi^2}\, \{1 + \log(1 + K_{\chi^2})\}\,,$$

where $K_{\chi^2} \coloneqq \max\{\chi^2(P_0 \,\|\, P_1), \chi^2(P_1 \,\|\, P_0)\}$.

In particular, our assumption $K_{\chi^2} < \infty$ guarantees that the mixture satisfies (LSI) with a constant independent of p, and hence does not exhibit a logarithmic divergence as $p \to \{0, 1\}$. We refer to the aforementioned papers for further discussion and examples of mixtures.

■ 11.4.5 Analogues on the hypercube

We now present another interesting illustration of our results. Here, we take $\mathscr{X} = \{0, 1\}^n$ to be the Boolean hypercube, and we take $\mathscr{Y} \coloneqq \mathcal{Y}^n$ to be a product space. We also require the Γ operator on $\mathscr{Y}$ to be consistent with the product structure; for simplicity of presentation, we omit this discussion and instead think of Γ as being either the squared gradient operator $\Gamma(f) = \|\nabla f\|^2$ on Euclidean space, or the discrete gradient $\Gamma(f) = (Df)^2$ as described in §11.2. Let π_0, π_1 be two probability measures on $\mathcal{Y}$ with

$$\begin{aligned} K_{\mathrm{LS}}(\pi) &\coloneqq \max\{C_{\mathrm{LS}}(\pi_0), C_{\mathrm{LS}}(\pi_1)\} < \infty\,, \\ K_{\chi^2}(\pi) &\coloneqq \max\{\chi^2(\pi_0 \,\|\, \pi_1), \chi^2(\pi_1 \,\|\, \pi_0)\} < \infty\,. \end{aligned}$$

Given $x \in \{0, 1\}^n$, define the measure

$$P_x = \bigotimes_{i=1}^{n} \pi_{x_i}\,. \tag{11.11}$$

From the tensorization of the chi-squared divergence,

$$\chi^2(P_x \,\|\, P_{x'}) = \prod_{i=1}^{n} \{1 + \chi^2(\pi_{x_i} \,\|\, \pi_{x_i'})\} - 1 \le \{1 + K_{\chi^2}(\pi)\}^{d(x,x')} - 1\,,$$

where $d(\cdot, \cdot)$ denotes the Hamming metric on $\{0, 1\}^n$. Moreover, each P_x satisfies (LSI) with a constant at most $K_{\mathrm{LS}}(\pi)$, due to the classical tensorization of log-Sobolev inequalities. As a consequence, we deduce the following result from Theorem 11.3.1.

Corollary 11.4.14. *Suppose μ is a probability measure on $\{0,1\}^n$ which is supported on a set of diameter at most k in the Hamming metric. Then, the mixture distribution $\mu P := \sum_{x\in\{0,1\}^n} \mu(x) P_x$, with P_x as in (11.11), satisfies (LSI) with*

$$C_{\mathrm{LS}}(\mu P) \le 6k\, K_{\mathrm{LS}}(\pi) \left\{1 + K_{\chi^2}(\pi)\right\}^k \left\{1 + \log\bigl(1 + K_{\chi^2}(\pi)\bigr)\right\}.$$

Importantly, the log-Sobolev inequality is dimension-free in the sense that it depends only on properties of π_0 and π_1 as well as the diameter k of the support of μ. An example of such a measure μ is any measure which is supported on $k/2$-sparse strings.

We now specialize this result to obtain an analogue of the result for Gaussian convolutions in §11.4.1 to the setting of the Boolean hypercube. Let $0 < p < 1/2$, and we take π_0 and π_1 to be the Bernoulli distributions with parameters p and $1-p$ respectively. Also, we take the Γ operator to be the square of discrete gradient. The optimal log-Sobolev inequality for these distributions is given in [Han16, Problem 8.3], and a quick computation yields

$$K_{\mathrm{LS}}(\pi) = \frac{p\,(1-p)}{2\,(1-2p)} \log \frac{1-p}{p}\,, \qquad K_{\chi^2}(\pi) = \frac{(1-p)^2}{p} + \frac{p^2}{1-p} - 1\,.$$

Note that the mixture μP can be interpreted as the result of evolving the initial measure μ for a short time under the natural semigroup on the hypercube. We obtain the following result.

Corollary 11.4.15. *Suppose μ is a probability measure on $\{0,1\}^n$ which is supported on a set of diameter at most k in the Hamming metric. Then, the mixture distribution μP with $0 < p < 1/2$ satisfies (LSI) with*

$$C_{\mathrm{LS}}(\mu P) \le \frac{6k}{p^{k-1}\,(1-2p)} \log^2 \frac{1}{p}\,.$$

■ 11.5 Tightening of the LSI

The following proposition is a standard result, see [BGL14, Proposition 5.1.3]. It is straightforward to see that bilinearity of Γ and our assumption (11.1) are sufficient for the proof to go through.

Proposition 11.5.1.

1. *If ρ satisfies (LSI), then ρ satisfies (PI) with $C_{\mathrm{P}}(\rho) \le C_{\mathrm{LS}}(\rho)$.*

2. If ρ satisfies the following defective LSI

$$\mathrm{ent}_\rho(f^2) \le 2C\,\mathbb{E}_\rho\,\Gamma(f) + D\,\mathbb{E}_\rho(f^2) \qquad \forall f \in \mathcal{A}\,,$$

together with (PI)*, then ρ satisfies* (LSI) *with*

$$C_{\mathrm{LS}}(\rho) \le C + C_{\mathrm{P}}(\rho)\,\big(\frac{D}{2} + 1\big)\,.$$

Chapter 12

Lower bounds for finding stationary points in optimization

In general, finding the global minimizer of a non-convex function is intractable, and the best we can hope for algorithmically is to locate a stationary point.

In this chapter, we characterize the query complexity of finding stationary points of one-dimensional non-convex but smooth functions. We consider four settings, based on whether the algorithms under consideration are deterministic or randomized, and whether the oracle outputs 1^{st}-order or both 0^{th}- and 1^{st}-order information. Our results show that algorithms for this task provably benefit by incorporating either randomness or 0^{th}-order information. Our results also show that, for every dimension $d \geq 1$, gradient descent is optimal among deterministic algorithms using 1^{st}-order queries only.

This chapter is based on [CBS23], joint with Sébastien Bubeck and Adil Salim.

12.1 Introduction

We consider optimizing a non-convex but smooth function $f : \mathbb{R}^d \to \mathbb{R}$, a task which underlies the spectacular successes of modern machine learning. Despite the fundamental nature of this question, there are still important aspects which remain poorly understood.

To set the stage for our investigation, let $f : \mathbb{R}^d \to \mathbb{R}$ be a β-smooth function with bounded objective gap: $f(0) - \inf f \leq \Delta$. Since global minimization of f is, in general, computationally intractable [c.f. NY83], we focus on the task of outputting an *ε-stationary point*, that is, a point $x^\star \in \mathbb{R}^d$ such that $\|\nabla f(x^\star)\| < \varepsilon$. By a standard rescaling argument (see Lemma 12.2.1), it suffices to consider the case $\beta = \Delta = 1$. Then, it is well-known [see, e.g., Nes18], that the standard gradient descent (GD) algorithm solves this task in $O(1/\varepsilon^2)$ queries to an oracle for the gradient ∇f. Conversely, [Car+20] proved that if the dimension d is sufficiently large, then any randomized algorithm for this task must use at least

$\Omega(1/\varepsilon^2)$ queries to a local oracle for f, thereby establishing the optimality of GD in high dimension.

However, the *low-dimensional complexity* of computing stationary points remains open. Indeed, the main limitation of [Car+20] is that their lower bound constructions require the ambient dimension to be large: more precisely, they require $d \geq \Omega(1/\varepsilon^2)$ for deterministic algorithms, and $d \geq \widetilde{\Omega}(1/\varepsilon^4)$ for randomized algorithms. The large dimensionality arises because they adapt to the non-convex smooth setting a "chain-like" lower bound construction for optimization of a convex non-smooth function [Nes18]. The chain-like construction forces certain natural classes of iterative algorithms to explore only one new dimension per iteration, and hence the dimension of the "hard" function in the construction is at least as large as the iteration complexity.

In fact, the non-convex and smooth setting shares interesting parallels with the convex and non-smooth setting, despite their apparent differences (in the former setting, we seek an ε-stationary point, whereas in the latter setting, we seek an ε-minimizer). Namely, in both settings the optimal oracle complexity is $\Theta(1/\varepsilon^2)$ in high dimension, and the optimal algorithm is (sub)gradient descent (as opposed to the convex smooth setting, for which accelerated gradient methods outperform GD). However, for the convex non-smooth setting, we know that the large dimensionality $d \geq \Omega(1/\varepsilon^2)$ of the lower bound construction is almost necessary, because of the existence of cutting-plane methods [see, e.g., Bub15; Nes18] which achieve a better complexity of $O(d \log(1/\varepsilon))$ in dimension $d \leq \widetilde{O}(1/\varepsilon^2)$. This raises the question of whether or not there exist analogues of cutting-plane methods for *non-convex* optimization.

A negative answer to this question would substantially improve our understanding of non-convex optimization, as it would point towards fundamental algorithmic obstructions. As such, the low-dimensional complexity of finding stationary points for non-convex optimization was investigated in a series of works [Vav93; Hin18; BM20]. These results show the existence of algorithms which improve upon GD in dimension $d \leq O(\log(1/\varepsilon))$. This suggests that GD is actually optimal for all $d \geq \Omega(\log(1/\varepsilon))$. To date, there has been little progress on this tantalizing conjecture because the existing low-dimensional lower bounds are delicate, relying on the theory of unpredictable random walks [Vav93; BPP98; BM20].

Our contributions. In this chapter, we study the task of finding an ε-stationary point of a smooth and univariate function $f : \mathbb{R} \to \mathbb{R}$. Our results, which are summarized as Table 12.1, provide a complete characterization of the oracle complexity of this task in four settings, based on whether or not the algorithm is allowed to use external randomness and whether or not the oracle outputs zeroth-order information. In particular, our lower bounds, which hold in dimension one,

Alg. Class	Oracle	Complexity	Lower Bound	Upper Bound
Deterministic	1^{st}	$\Theta(1/\varepsilon^2)$	Theorem 12.2.4	GD (well-known)
Randomized	1^{st}	$\Theta(1/\varepsilon)$	Theorem 12.2.2	Theorem 12.2.3
Deterministic	$0^{\text{th}} + 1^{\text{st}}$	$\Theta(\log(1/\varepsilon))$	Theorem 12.2.5	Theorem 12.2.6
Randomized	$0^{\text{th}} + 1^{\text{st}}$	$\Theta(\log(1/\varepsilon))$	Theorem 12.2.5	Theorem 12.2.6

Table 12.1: Summary of the results of this chapter.

also hold in every dimension $d \geq 1$. In spite of the simplicity of the setting, we can draw a number of interesting conclusions from the results.

- **Optimality of GD for any dimension $d \geq 1$.** Our results imply that, among algorithms which are deterministic and only use first-order queries, GD is optimal in every dimension $d \geq 1$. This was previously known only for $d \geq \Omega(1/\varepsilon^2)$ [Car+20].
- **Separations between algorithm classes and oracles**. Our results exhibit a natural setting in which both randomization and zeroth-order queries provably improve the query complexity of optimization. It shows, in particular, that at least one of these additional ingredients is *necessary* to improve upon the basic GD algorithm.
- **Finding stationary points for unconstrained optimization**. The methods of [Vav93; BM20] for improving upon the complexity of GD in low dimension are applicable to the constrained case in which the domain of f is the cube $[0,1]^d$, and it is not obvious that they can be applied to unbounded domains. We address this question by characterizing the oracle complexity for the unconstrained case.

Related works. Usually, optimization lower bounds are established for specific classes of algorithms, such as algorithms for which each iterate lies in the span of the previous iterates and gradients [Nes18]. As noted in [WS17], lower bounds against arbitrary randomized algorithms for convex optimization are trickier and are often loose with regards to the dimension in which the construction is embedded. The complexity of finding stationary points is further studied in [Car+21].

Conventions and notation. A function $f : \mathbb{R}^d \to \mathbb{R}$ is *β-smooth* if it is continuously differentiable and its gradient ∇f is β-Lipschitz. If $d = 1$, we shall write f' instead of ∇f. We use the standard asymptotic notation $\Omega(\cdot)$, $O(\cdot)$, and $\Theta(\cdot)$.

12.2 Results

In this section, we give detailed statements of our results as well as proof sketches. The full proofs are deferred to §12.3. We also record the following lemma, which allows us to reduce to the case of $\beta = \Delta = 1$.

Lemma 12.2.1. *Let $\mathscr{C}_*(\varepsilon; \beta, \Delta, d, \mathscr{O}) \geq 0$ denote the complexity of finding an ε-stationary point over the class of β-smooth functions $f : \mathbb{R}^d \to \mathbb{R}$ with $f(0) - \inf f \leq \Delta$ using an oracle $\mathscr{O}$, where given $x \in \mathbb{R}^d$ the oracle $\mathscr{O}$ returns either $\nabla f(x)$ (first-order information) or $(f(x), \nabla f(x))$ (zeroth- and first-order information). Here, $* \in \{det, rand\}$ is a subscript denoting whether or not the algorithm is allowed to use external randomness; when $* = rand$, the randomized complexity refers to the minimum number of queries required to find an ε-stationary point with probability at least $1/2$. Then, for any $\beta, \Delta, \varepsilon > 0$,*

$$\mathscr{C}_*(\varepsilon; \beta, \Delta, d, \mathscr{O}) = \mathscr{C}_*\big(\frac{\varepsilon}{\sqrt{\beta\Delta}}; 1, 1, d, \mathscr{O}\big)\,.$$

Proof. Given a β-smooth function $f : \mathbb{R}^d \to \mathbb{R}$ with $f(0) - \inf f \leq \Delta$, define $g : \mathbb{R}^d \to \mathbb{R}$ via $g(x) \coloneqq \Delta^{-1} f(\sqrt{\Delta/\beta}\, x)$. Then, g is 1-smooth with $g(0) - \inf g \leq 1$, and it is clear that the oracle for g can be simulated using the oracle for f. Moreover, an $\varepsilon/\sqrt{\beta\Delta}$-stationary point for g translates into an ε-stationary point for f. Obviously, the reduction is reversible. □

Often, we will assume without loss of generality that $f(0) = 1$ and $\beta = \Delta = 1$, so that $f \geq 0$. Also, we may assume that $f'(0) \leq -\varepsilon$, since if $f'(0) \in (-\varepsilon, \varepsilon)$ then 0 is an ε-stationary point of f, and if $f'(0) \geq \varepsilon$ we can replace f by $x \mapsto f(-x)$. We abbreviate $\mathscr{C}_*(\varepsilon; \mathscr{O}) \coloneqq \mathscr{C}_*(\varepsilon; 1, 1, 1, \mathscr{O})$, and from now on we consider $d = 1$.

Let $\mathscr{O}^{1^{\text{st}}}$ denote the oracle which returns first-order information (given $x \in \mathbb{R}$, it outputs $f'(x)$), and let $\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}$ denote the oracle which returns zeroth- and first-order information (given $x \in \mathbb{R}$, it outputs $(f(x), f'(x))$). We remark that in the one-dimensional setting, we could instead assume access to an oracle $\mathscr{O}^{0^{\text{th}}}$ which only outputs zeroth-order information, rather than $\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}$; this is because we can simulate $\mathscr{O}^{1^{\text{st}}}$ to arbitrary accuracy given $\mathscr{O}^{0^{\text{th}}}$ with only a constant factor overhead in the number of oracle queries by using finite differences. For simplicity, we work with $\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}$ and we will not consider $\mathscr{O}^{0^{\text{th}}}$ further.

12.2.1 Lower bound for randomized algorithms

We begin with a lower bound construction for randomized algorithms which only use first-order queries. For simplicity, assume that $1/\varepsilon$ is an integer. We construct a family of functions $(f_j)_{j \in [1/\varepsilon]}$, with the following properties. On the negative

half-line $\mathbb{R}_-$, each f_j decreases with slope $-\varepsilon$, with $f_j(0) = 1$. We also set the slope of f_j on the positive half-line $\mathbb{R}_+$ to be $-\varepsilon$, but this entails that $f_j(x) < 0$ for $x > 1/\varepsilon$, violating the constraint $f_j(0) - \inf f \le 1$. Instead, on the interval $[j-1, j]$, we modify f_j to increase as much as possible while remaining $O(1)$-smooth, so that $f_j(1/\varepsilon) = f_j(0) = 1$; we can then periodically extend f_j on the rest of $\mathbb{R}_+$.

Due to the periodicity of the construction, we can restrict our attention to the interval $[0, 1/\varepsilon]$. Without prior knowledge of the index j, any algorithm only has a "probability" (made precise in §12.3.2) of at most ε of finding the interval $[j-1, j]$, which contains all of the ε-stationary points in $[0, 1/\varepsilon]$. Hence, we expect that any randomized algorithm must require at least $\Omega(1/\varepsilon)$ queries to find an ε-stationary point of f_j.

To make this formal, let $\Phi : [0,1] \to \mathbb{R}$ be a smooth function such that $\Phi(0) = 0$, $\Phi(1) = 1$, and $\Phi'(0) = \Phi'(1) = -\varepsilon$. For example, we can take

$$\Phi(x) = \begin{cases} 2\,(1+\varepsilon)\,x^2 - \varepsilon\,x\,, & x \in [0, \frac{1}{2}]\,, \\ 2\,\Phi(\frac{1}{2}) - \Phi(1-x)\,, & x \in [\frac{1}{2}, 1]\,. \end{cases}$$

We can check that Φ satisfies the desired properties and that Φ is β-smooth with $\beta = 4\,(1+\varepsilon) \le 5$ for $\varepsilon \le \frac{1}{4}$. Then, let

$$f_j(x) \coloneqq \begin{cases} 1 - \varepsilon\,x\,, & x \in (-\infty, j-1]\,, \\ 1 - \varepsilon\,(j-1) + (1-\varepsilon)\,\Phi(x-(j-1))\,, & x \in [j-1, j]\,, \\ f_j(j) - \varepsilon\,(x-j)\,, & x \in [j, 1/\varepsilon]\,, \\ f_j(x - 1/\varepsilon)\,, & x \in [1/\varepsilon, \infty)\,. \end{cases}$$

It follows that f_j is also 5-smooth, with $f_j(0) - \inf f_j \le 1$; see Figure 12.1.

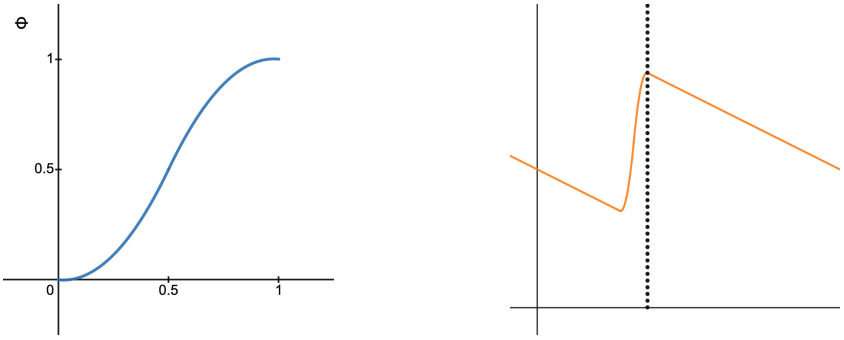


Figure 12.1: (Left) A plot of Φ. (Right) A plot of f_j, where the dotted line indicates the value of j.

We prove the following theorem in §12.3.2.

Theorem 12.2.2. *For all $\varepsilon \in (0, \frac{1}{8})$, it holds that*

$$\mathscr{C}_{\mathrm{rand}}\big(\varepsilon; \mathscr{O}^{1^{\mathrm{st}}}\big) \geq \Omega\Big(\frac{1}{\varepsilon}\Big).$$

■ 12.2.2 An optimal randomized algorithm

The lower bound construction of the previous section suggests a simple strategy for computing an ε-stationary point of f: namely, just repeatedly pick points uniformly at random in the interval $[0, 1/\varepsilon]$. We now show that such a strategy (together with some additional processing steps) succeeds at obtaining an ε-stationary point in $O(1/\varepsilon)$ queries.

Algorithm 12.1 RandomSearch

Require: oracle $\mathscr{O}^{1^{\mathrm{st}}}$ for f
Ensure: ε-stationary point x
  while true **do**
    draw $x \sim \mathsf{uniform}([0, 2/\varepsilon])$
    if $|f'(x)| < \varepsilon$ **then**
      output x
    else if $f'(x) > 0$ **then**
      call BinarySearch$(\mathscr{O}^{1^{\mathrm{st}}}, 0, x)$

Algorithm 12.2 BinarySearch

Require: oracle $\mathscr{O}^{1^{\mathrm{st}}}$ for f; initial points $x_0 < x_1$ with $f'(x_0) \leq -\varepsilon$ and $f'(x_1) > 0$
Ensure: ε-stationary point x
  set $m \leftarrow \frac{x_0 + x_1}{2}$
  if $|f'(m)| < \varepsilon$ **then**
    output m
  else if $f'(m) \leq -\varepsilon$ **then**
    call BinarySearch$(\mathscr{O}^{1^{\mathrm{st}}}, m, x_1)$
  else if $f'(m) > 0$ **then**
    call BinarySearch$(\mathscr{O}^{1^{\mathrm{st}}}, x_0, m)$

The pseudocode for the algorithms is given as Algorithms 12.1 and 12.2. In short, RandomSearch (Algorithm 12.1) uses $O(1/\varepsilon)$ queries to find a "good point", i.e., either an ε-stationary point or a point x with $f'(x) > 0$. In the latter case, BinarySearch (Algorithm 12.2) then locates an ε-stationary point using an additional $O(\log(1/\varepsilon))$ queries.

We prove the following theorem in §12.3.3.

Theorem 12.2.3. *Assume that $f : \mathbb{R} \to \mathbb{R}$ is 1-smooth, $f \geq 0$, $f(0) = 1$, and $f'(0) \leq -\varepsilon$. Then,* RandomSearch *(Algorithm 12.1) terminates with an ε-stationary point for f using at most $O(1/\varepsilon)$ queries to the oracle with probability at least $1/2$.*

As usual, the success probability can be boosted by rerunning the algorithm. In Figure 12.2, we demonstrate the performance of RandomSearch in a numerical experiment as a sanity check.

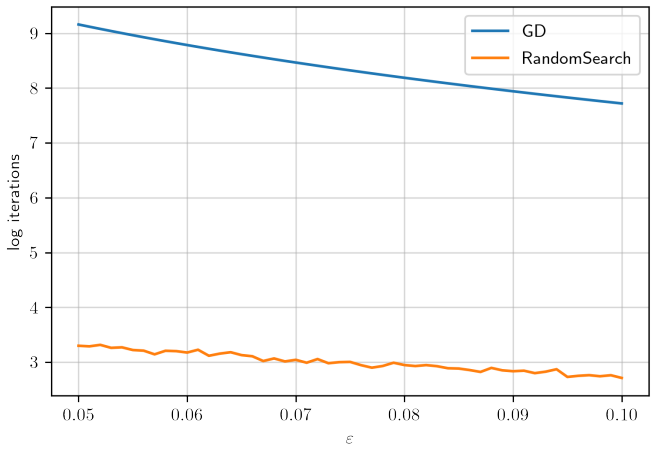


Figure 12.2: Iteration complexity of gradient descent (GD) vs. one run of RandomSearch (Algorithm 12.1) for various choices of ε on an instance of the construction in §12.2.1. The flatter slope of the orange line reflects the improved $O(1/\varepsilon)$ complexity of RandomSearch over the $O(1/\varepsilon^2)$ complexity of GD.

■ 12.2.3 Lower bound for deterministic algorithms

Against the class of deterministic algorithms, the construction of Theorem 12.2.2 can be strengthened to yield a $\Omega(1/\varepsilon^2)$ lower bound. The idea is based on the concept of a *resisting oracle* $\mathcal{O}^{\text{resist}}$ from [Nes18] which, regardless of the query point x, outputs "$f'(x) = -\varepsilon$". The goal then is to show that for any deterministic sequence of queries $x_1, \ldots, x_N$, if $N \leq O(1/\varepsilon^2)$, there exists a 1-smooth function $f : \mathbb{R} \to \mathbb{R}$ with $f(0) - \inf f \leq \Delta$ which is consistent with the output of the oracle, i.e., satisfies $f'(x_i) = -\varepsilon$ for all $i \in [N]$. Note that this strategy necessarily only

provides a lower bound against deterministic algorithms.[1]

For simplicity of notation, since the order of the queries does not matter here, we assume that the queries are sorted: $x_1 < \cdots < x_N$. The function f that we construct has slope $-\varepsilon$ at the query points, but rapidly rises in between the query points to ensure that the condition $f(0) - \inf f \leq 1$ holds. Moreover, we will ensure that $f'(x) = -\varepsilon$ for $x \leq 0$ and that f' is periodic on $\mathbb{R}_+$ with period $1/\varepsilon$; hence, we may assume that all of the queries lie in the informative interval $(0, 1/\varepsilon)$. The key here is that for deterministic algorithms, the intervals on which the function f rises can be adapted to the query points, rather than being selected in advance.

The intuition is as follows. If the algorithm has made fewer than $O(1/\varepsilon^2)$ queries, then there must be $\Omega(1/\varepsilon^2)$ disjoint intervals in $[0, 1/\varepsilon]$ of length at least $\Omega(\varepsilon)$ in which there are no query points. On each such interval, we can grow our function value by $\Omega(\varepsilon^2)$ while staying smooth and with slope $-\varepsilon$ at the start and end of the interval. Hence, we can guarantee that the constructed function f remains above $f(0) - 1$, while answering $f'(x) = -\varepsilon$ at every query point x.

To make this precise, let $\ell_i \coloneqq x_{i+1} - x_i$ and define the function

$$\Phi_i(x) \coloneqq -\varepsilon\,(x - x_i) + \begin{cases} \frac{1}{2}\,(x - x_i)^2\,, & x \in [x_i, x_i + \frac{\ell_i}{2}]\,, \\ \frac{\ell_i^2}{8} + \frac{\ell_i}{2}\,(x - x_i - \frac{\ell_i}{2}) - \frac{1}{2}\,(x - x_i - \frac{\ell_i}{2})^2\,, & x \in [x_i + \frac{\ell_i}{2}, x_{i+1}]\,. \end{cases}$$

The construction of Φ_i satisfies the following properties:

1. Φ_i is continuously differentiable and 1-smooth on $[x_i, x_{i+1}]$.
2. $\Phi_i(x_i) = 0$ and $\Phi_i(x_{i+1}) = \ell_i\,(\frac{\ell_i}{4} - \varepsilon)$.
3. $\Phi_i'(x_i) = \Phi_i'(x_{i+1}) = -\varepsilon$.

Write $x_0 \coloneqq 0$ and $x_{N+1} \coloneqq 1/\varepsilon$. Recall that $x_i \in (0, 1/\varepsilon)$, for all $i \in [N]$. We now

[1] In more detail, the argument is as follows. Let $x_1, \ldots, x_N$ be the sequence of query points generated by the algorithm when run with $\mathscr{O}^{\text{resist}}$, and suppose we can find a function f which is consistent with the responses of $\mathscr{O}^{\text{resist}}$. Then, for a deterministic algorithm, we can be sure that *had the algorithm been run with the oracle* $\mathscr{O}^{1^{\text{st}}}$ *for* f, it would have generated the same sequence of query points $x_1, \ldots, x_N$, and hence would have never found an ε-stationary point of f among the N query points. This argument fails if the algorithm incorporates external randomness.

define

$$f(x) \coloneqq \begin{cases} 1 - \varepsilon\, x\,, & x \in (-\infty, 0]\,, \\ f(x_i) - \varepsilon\,(x - x_i)\,, & x \in [x_i, x_{i+1}] \text{ and } \ell_i < 8\varepsilon \;\; (0 \le i \le N)\,, \\ f(x_i) + \Phi_i(x)\,, & x \in [x_i, x_{i+1}] \text{ and } \ell_i \ge 8\varepsilon \;\; (0 \le i \le N)\,, \\ f(x - 1/\varepsilon) + a\,, & x \in [1/\varepsilon, \infty)\,, \end{cases}$$

where $a \coloneqq f(1/\varepsilon) - f(0)$. See Figure 12.3 for an illustration of f. We shall prove that when $N \le O(1/\varepsilon^2)$, then the function f is 1-smooth and satisfies $f(0) - \inf f \le 1$, thus completing the resisting oracle construction. It yields the following theorem, which we prove in §12.3.4.

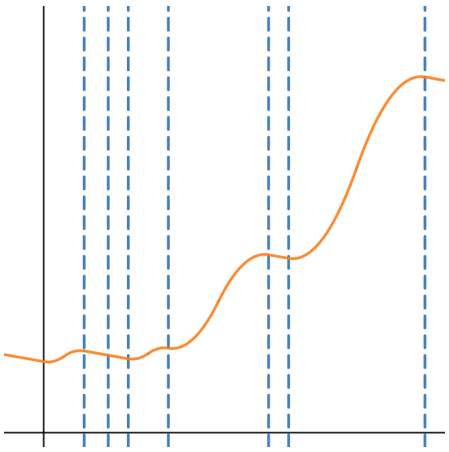


Figure 12.3: We plot an example of the function f. The dashed lines indicate the query points made by the algorithms.

Theorem 12.2.4. *For all $\varepsilon \in (0, 1)$, it holds that*

$$\mathscr{C}_{\text{det}}\big(\varepsilon; \mathscr{O}^{1^{\text{st}}}\big) \ge \Omega\Big(\frac{1}{\varepsilon^2}\Big)\,.$$

The lower bound is matched by gradient descent. For the sake of completeness, we provide a proof of the matching $O(1/\varepsilon^2)$ upper bound via gradient descent as Theorem 12.3.2 in §12.3.1.

12.2.4 Lower bound for randomized algorithms with zeroth-order information

We now turn towards algorithms which use the $0^{\text{th}} + 1^{\text{st}}$-order oracle $\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}$. For the lower bound, we again use the family of functions $(f_j)_{j \in [1/\varepsilon]}$ introduced in

§12.2.1. The main difference is that given a query point $x \in [0, 1/\varepsilon]$, the value of $f_j(x)$ reveals whether or not the interval $[j-1, j]$ lies to the left of x and hence allows for binary search to determine j. Consequently, the lower bound is only of order $\Omega(\log(1/\varepsilon))$.

We prove the following theorem in §12.3.5.

Theorem 12.2.5. *For all $\varepsilon \in (0, \frac{1}{8})$, it holds that*

$$\mathscr{C}_{\text{det}}\left(\varepsilon; \mathscr{O}^{0^{\text{th}}+1^{\text{st}}}\right) \geq \mathscr{C}_{\text{rand}}\left(\varepsilon; \mathscr{O}^{0^{\text{th}}+1^{\text{st}}}\right) \geq \Omega\left(\log \frac{1}{\varepsilon}\right).$$

■ 12.2.5 An optimal deterministic algorithm with zeroth-order information

Finally, we provide a deterministic algorithm whose complexity matches the lower bound in Theorem 12.2.5. At a high level, the idea is to use the zeroth-order information to perform binary search, but the actual algorithm is slightly more involved and requires the consideration of various cases.

We summarize the idea behind the algorithm. First, as described earlier, we may freely assume $f \geq 0$, $f(0) = 1$, and $f'(0) \leq -\varepsilon$. Also, we recall that if the algorithm ever sees a point x with either $|f'(x)| < \varepsilon$ or $f'(x) > 0$, then we are done (in the latter case, we can call Algorithm 12.2: BinarySearch).

1. DecreaseGap (Algorithm 12.4) checks the value of $f(2/\varepsilon)$. If $f(2/\varepsilon) \leq \frac{3}{4} f(0)$, then we have made progress on the objective gap and we may treat $2/\varepsilon$ as the new origin. This can happen at most $O(\log(1/\varepsilon))$ times. Otherwise, we have $f(2/\varepsilon) \geq \frac{3}{4} f(0)$, and we move on to the next phase of the algorithm.

2. Set $x_- \coloneqq 0$ and $x_+ \coloneqq 2/\varepsilon$. There are two cases: either $\frac{3}{4} f(x_-) \leq f(x_+) \leq f(x_-)$, in which case $f(x_-) - f(x_+) \leq \frac{\varepsilon}{4}(x_+ - x_-)$, or $f(x_+) \geq f(x_-)$.

3. The first case is handled by BinarySearchII (Algorithm 12.5). A simple calculation reveals that the condition $0 \leq f(x_-) - f(x_+) \leq \frac{3}{4}(x_+ - x_-)$ together with $f'(x_-) \leq -\varepsilon$ implies the existence of an ε-stationary point in $[x_-, x_+]$. We now check the midpoint m of x_- and x_+. If $f(m) \notin [f(x_+), f(x_-)]$, then we arrive at the second case. Otherwise, we replace either x_- or x_+ with m; one of these two choices will cut the value of $f(x_-) - f(x_+)$ by at least half, thereby ensuring that the condition $0 \leq f(x_-) - f(x_+) \leq \frac{3}{4}(x_+ - x_-)$ continues to hold. This can happen at most $O(\log(1/\varepsilon))$ times.

4. Finally, the second case is handled by BinarySearchIII (Algorithm 12.6). In this case, $f(x_+) \geq f(x_-)$ together with $f'(x_-) \leq -\varepsilon$ ensures that there is a stationary point in $[x_-, x_+]$. We then check the value of $f(m)$ where m is

the midpoint of x_- and x_+. It is straightforward to check that we can replace either x_- or x_+ with m and preserve the condition $f(x_+) \geq f(x_-)$. This can happen at most $O(\log(1/\varepsilon))$ times.

Algorithm 12.3 ZEROTHORDER

Require: oracle $\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}$ for f
Ensure: ε-stationary point x
  set $x_- \leftarrow$ DECREASEGAP($\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}, 0$)
  set $x_+ \leftarrow x_- + 2/\varepsilon$
  if $|f'(x_-)| < \varepsilon$ **then**
    output x_-
  else if $f(x_+) \leq f(x_-)$ **then**
    call BINARYSEARCHII($\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}, x_-, x_+$)
  else if $f(x_+) > f(x_-)$ **then**
    call BINARYSEARCHIII($\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}, x_-, x_+$)

Algorithm 12.4 DECREASEGAP

Require: oracle $\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}$ for f; point x_0
Ensure: either an ε-stationary point x or a point x such that $f(x) \leq f(x_0)$, $f'(x) \leq -\varepsilon$, and $f(x + 2/\varepsilon) \geq \frac{3}{4} f(x)$
  if $|f'(x_0 + 2/\varepsilon)| < \varepsilon$ **then**
    output $x_0 + 2/\varepsilon$
  else if $f'(x_0 + 2/\varepsilon) > 0$ **then**
    call BINARYSEARCH($\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}, x_0, x_0 + 2/\varepsilon$)
  else if $f(x_0 + 2/\varepsilon) \geq \frac{3}{4} f(x_0)$ **then**
    output x_0
  else
    call DECREASEGAP($\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}, x_0 + 2/\varepsilon$)

We prove the following theorem in §12.3.6.

Theorem 12.2.6. *Assume that $f : \mathbb{R} \to \mathbb{R}$ is 1-smooth, $f \geq 0$, $f(0) = 1$, and $f'(0) \leq -\varepsilon$. Then,* ZEROTHORDER *(Algorithm 12.3) terminates with an ε-stationary point for f using at most $O(\log(1/\varepsilon))$ queries to the oracle.*

Algorithm 12.5 BinarySearchII

Require: oracle $\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}$ for f; points $x_- < x_+$ with $f'(x_-) \le -\varepsilon$ and $0 \le f(x_-) - f(x_+) \le \frac{\varepsilon}{4}(x_+ - x_-)$
Ensure: an ε-stationary point x
  set $m \leftarrow \frac{x_- + x_+}{2}$
  if $|f'(m)| < \varepsilon$ **then**
    output m
  else if $f'(m) > 0$ **then**
    call BinarySearch($\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}, x_-, m$)
  else if $f(m) \ge f(x_-)$ **then**
    call BinarySearchIII($\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}, x_-, m$)
  else if $f(m) \le f(x_+)$ **then**
    call BinarySearchIII($\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}, m, x_+$)
  else if $f(x_-) - f(m) \le \frac{1}{2}(f(x_-) - f(x_+))$ **then**
    call BinarySearchII($\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}, x_-, m$)
  else if $f(m) - f(x_+) \le \frac{1}{2}(f(x_-) - f(x_+))$ **then**
    call BinarySearchII($\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}, m, x_+$)

Algorithm 12.6 BinarySearchIII

Require: oracle $\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}$ for f; points $x_- < x_+$ with $f'(x_-) \le -\varepsilon$, $f(x_+) \ge f(x_-)$
Ensure: an ε-stationary point x
  set $m \leftarrow \frac{x_- + x_+}{2}$
  if $|f'(m)| < \varepsilon$ **then**
    output m
  else if $f'(m) > 0$ **then**
    call BinarySearch($\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}, x_-, m$)
  else if $f(m) \ge f(x_-)$ **then**
    call BinarySearchIII($\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}, x_-, m$)
  else
    call BinarySearchIII($\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}, m, x_+$)

12.3 Proofs

12.3.1 Preliminaries

The standard approach for proving lower bounds against randomized algorithms is to reduce the task under consideration to a statistical estimation problem, for which we can bring to bear tools from information theory. Namely, we use *Fano's inequality*; we refer readers to [CT06, §2] for background.

Theorem 12.3.1 (Fano's inequality)**.** *Let m be a positive integer and let $J \sim \mathsf{uniform}([m])$. Then, for any estimator $\widehat{J}$ of J which is measurable w.r.t. some data Y, it holds that*

$$\mathbb{P}\{\widehat{J} \neq J\} \geq 1 - \frac{I(J;Y) + \ln 2}{\ln m}\,,$$

where I denotes the mutual information.

For the sake of completeness, we also include a proof of the $O(1/\varepsilon^2)$ complexity bound for gradient descent.

Theorem 12.3.2. *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is 1-smooth with $f(0) - \inf f \leq 1$. Set $x_0 \coloneqq 0$ and for $k \in \mathbb{N}$, consider the iterates of GD with step size 1:*

$$x_{k+1} \coloneqq x_k - \nabla f(x_k)\,.$$

Then,

$$\min_{k=0,1,\ldots,N-1} \|\nabla f(x_k)\| \leq \sqrt{\frac{2}{N}}\,.$$

Proof. Due to the 1-smoothness of f,

$$f(x_{k+1}) - f(x_k) \leq \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{1}{2}\|x_{k+1} - x_k\|^2 = -\frac{1}{2}\|\nabla f(x_k)\|^2\,. \tag{12.1}$$

Rearranging this and summing,

$$\begin{aligned}\min_{k=0,1,\ldots,N-1} \|\nabla f(x_k)\|^2 &\leq \frac{1}{N}\sum_{k=0}^{N-1}\|\nabla f(x_k)\|^2 \leq \frac{2}{N}\sum_{k=0}^{N-1}\{f(x_k) - f(x_{k+1})\} \\ &\leq \frac{2}{N}\{f(0) - f(x_N)\} \leq \frac{2}{N}\{f(0) - \inf f\} \leq \frac{2}{N}\,. \qquad \square\end{aligned}$$

12.3.2 Proof of Theorem 12.2.2

Proof of Theorem 12.2.2. By making the value of ε larger (up to a factor of 2), we may assume that $1/\varepsilon$ is an integer.

We reduce the optimization task to a statistical estimation problem. Let $J \sim \mathsf{uniform}([1/\varepsilon])$. Since the only regions in which $|f_j'| < \varepsilon$ are contained in intervals of the form $k/\varepsilon + [j-1, j]$ for some $k \in \mathbb{N}$, then finding an ε-stationary point of f_J implies that the algorithm can guess the value of J (exactly).

On the other hand, we lower bound the number of queries required to guess the value of J. Let $x_1, \ldots, x_N$ denote the query points of the algorithm, which may also depend on an external source of randomness U. Write $\mathscr{O}_{f_j}(x) = f_j'(x)$ for the output of the oracle for f_j on the query x (we omit the superscript 1^{st} for brevity). Let $\widehat{J}$ be any estimator of J based on $\{x_i, \mathscr{O}_{f_J}(x_i) : i \in [N]\}$. Then, by Fano's inequality (Theorem 12.3.1),

$$\mathbb{P}\{\widehat{J} \neq J\} \geq 1 - \frac{I(\{x_i, \mathscr{O}_{f_J}(x_i) : i \in [N]\}; J) + \ln 2}{\ln(1/\varepsilon)}\,.$$

First, suppose that the algorithm is deterministic. This means that each x_i is a deterministic function of $\{x_{i'}, \mathscr{O}_{f_J}(x_{i'}) : i' \in [i-1]\}$. The chain rule for the mutual information implies that

$$\begin{aligned} &I\big(\{x_i, \mathscr{O}_{f_J}(x_i) : i \in [N]\}; J\big) \\ &\qquad \leq \sum_{i=1}^{N} I\big(\mathscr{O}_{f_J}(x_i); J \;\big|\; \{x_{i'}, \mathscr{O}_{f_J}(x_{i'}) : i' \in [i-1]\}\big)\,. \end{aligned}$$

On the other hand, there are two possibilities for the i-th term in the summation. Either one of the previous queries already landed in an interval corresponding to J, in which case J is already known and the mutual information is zero, or none of the previous queries have hit an interval corresponding to J. In the latter case, conditionally on the information up to iteration i, J is uniformly distributed on $1/\varepsilon - i$ remaining intervals, and so

$$\begin{aligned} &I\big(\mathscr{O}_{f_J}(x_i); J \;\big|\; \{x_{i'}, \mathscr{O}_{f_J}(x_{i'}) : i' \in [i-1]\}\big) \\ &\qquad \leq H\big(\mathscr{O}_{f_J}(x_i) \;\big|\; \{x_{i'}, \mathscr{O}_{f_J}(x_{i'}) : i' \in [i-1]\}\big) = h\big(\frac{1}{1/\varepsilon - i}\big)\,, \end{aligned}$$

with h denoting the entropy function $p \mapsto p \ln \frac{1}{p} + (1-p) \ln \frac{1}{1-p}$. The last inequality follows because conditionally, $\mathscr{O}_{f_J}(x_i)$ can only be one of two possible values with probabilities $\frac{1}{1/\varepsilon - i}$ and $1 - \frac{1}{1/\varepsilon - i}$ respectively. If $N \leq 1/(2\varepsilon)$, then

$$I\big(\{x_i, \mathscr{O}_{f_J}(x_i) : i \in [N]\}; J\big) \leq 2 \sum_{i=1}^{N} \frac{1}{1/\varepsilon - i} \ln\big(\frac{1}{\varepsilon} - i\big) \leq 4N\varepsilon \ln \frac{1}{\varepsilon}\,.$$

Hence,

$$\mathbb{P}\{\widehat{J} \neq J\} \geq 1 - \frac{4N\varepsilon \ln(1/\varepsilon) + \ln 2}{\ln(1/\varepsilon)} > \frac{1}{2} \tag{12.2}$$

provided that $\varepsilon \leq \frac{1}{8}$ and $N \leq O(1/\varepsilon)$ for a sufficiently small implied constant. Although we have proven the bound (12.2) for deterministic algorithms, the bound (12.2) continues to hold for randomized algorithms simply by conditioning on the random seed U which is independent of J.

We have proven that any randomized algorithm which is guaranteed to find an ε-stationary point of f_J must use at least $N \geq \Omega(1/\varepsilon)$ queries, or

$$\mathscr{C}(\varepsilon; 5, 1, 1, \mathscr{O}^{\text{1st}}) \geq \Omega\big(\frac{1}{\varepsilon}\big)\,.$$

We conclude by applying the rescaling lemma (Lemma 12.2.1). □

■ 12.3.3 Proof of Theorem 12.2.3

First, we analyze the subroutine BinarySearch.

Lemma 12.3.3. *Suppose that f is 1-smooth. Then,* BinarySearch *(Algorithm 12.2) terminates with an ε-stationary point for f using at most $O(\log \frac{x_1 - x_0}{\varepsilon})$ queries to the oracle.*

Proof. Since f is 1-smooth, $f(x_0) \leq -\varepsilon$ and $f(x_1) > 0$ cannot hold if $x_1 - x_0 \leq \varepsilon$. Moreover, each time that BinarySearch fails to find an ε-stationary point for f, the length of the interval $[x_0, x_1]$ is cut in half. The result follows. □

We also need one lemma about continuous functions on $\mathbb{R}$.

Lemma 12.3.4. *Let $g : \mathbb{R} \to \mathbb{R}$ be continuous, let I be a compact and non-empty interval, and let $\varepsilon > 0$. Then, there is a finite collection of disjoint closed intervals which cover $I \cap \{g \geq \varepsilon\}$ and which are contained in $I \cap \{g \geq 0\}$.*

Proof. For each $x \in S \coloneqq I \cap \{g \geq \varepsilon\}$, by continuity of g there exists a closed interval $I_x \subseteq I$ such that x belongs to the interior of I_x and such that $g \geq 0$ on I_x. The collection $(I_x)_{x \in S}$ covers the compact set S, so we can extract a finite subcover. The connected components of the union of the finite subcover consist of disjoint closed intervals. □

We are now ready to prove Theorem 12.2.3.

Proof of Theorem 12.2.3. Let $x \sim \mathsf{uniform}([0, 2/\varepsilon])$. If $|f'(x)| < \varepsilon$, then we are done, and if $f'(x) > 0$, then Lemma 12.3.3 shows that BinarySearch terminates with an ε-stationary point of f using $O(\log(1/\varepsilon))$ queries. What remains to show is that x satisfies either $|f'(x)| < \varepsilon$ or $f'(x) > 0$ with probability at least $\Omega(\varepsilon)$, which implies that Algorithm 12.1 succeeds using $O(1/\varepsilon)$ queries with probability at least $1/2$.

Let $\mathfrak{m}$ denote the Lebesgue measure restricted to $[0, 2/\varepsilon]$. Then,

$$\begin{aligned} 1 \geq f(0) - f(2/\varepsilon) &= -\int_{[0,2/\varepsilon]} f' \\ &\geq \varepsilon\, \mathfrak{m}\{f' \leq -\varepsilon\} - \varepsilon\, \mathfrak{m}\{|f'| < \varepsilon\} - \int_{[0,2/\varepsilon] \cap \{f' \geq \varepsilon\}} f' \,. \end{aligned}$$

From Lemma 12.3.4, we can cover the set $[0, 2/\varepsilon] \cap \{f' \geq \varepsilon\}$ with a union of disjoint closed intervals $\bigcup_{k=1}^{K} I_k \subseteq [0, 2/\varepsilon] \cap \{f' \geq 0\}$. On I_k, the smoothness of f ensures that

$$-\int_{I_k} f' \geq -\mathfrak{m}(I_k)\, \underbrace{f'(\inf I_k)}_{\leq \varepsilon} - \int_{I_k} (x - \inf I_k)\, \mathrm{d}x \geq -\varepsilon\, \mathfrak{m}(I_k) - \frac{1}{2}\, \mathfrak{m}(I_k)^2 \,.$$

Write $\ell_k \coloneqq \mathfrak{m}(I_k) = \sup I_k - \inf I_k$. Note that $\sum_{k=1}^{K} \ell_k \leq \mathfrak{m}\{f' \geq 0\}$. Thus,

$$\begin{aligned} -\int_{[0,2/\varepsilon] \cap \{f' \geq \varepsilon\}} f' &\geq -\varepsilon \sum_{k=1}^{K} \ell_k - \frac{1}{2} \sum_{k=1}^{K} \ell_k^2 \geq -\varepsilon \sum_{k=1}^{K} \ell_k - \frac{1}{2} \Big(\sum_{k=1}^{K} \ell_k \Big)^2 \\ &\geq -\varepsilon\, \mathfrak{m}\{f' \geq 0\} - \frac{1}{2}\, \mathfrak{m}\{f' \geq 0\}^2 \,. \end{aligned}$$

Now suppose that $\mathfrak{m}\{|f'| < \varepsilon \text{ or } f' \geq \varepsilon\} \leq c_0$, where $c_0 > 0$ is a constant to be chosen later. In this case, the inequalities above imply

$$1 + 2c_0\varepsilon + \frac{1}{2}\, c_0^2 \geq \varepsilon\, \mathfrak{m}\{f' \leq -\varepsilon\} \geq \varepsilon \left(\frac{2}{\varepsilon} - \mathfrak{m}\{|f'| < \varepsilon \text{ or } f' \geq \varepsilon\} \right)$$

which, when rearranged, yields

$$1 + 3c_0\varepsilon + \frac{1}{2}\, c_0^2 \geq 2 \,.$$

If c_0 is a sufficiently small absolute constant, we arrive at a contradiction.

We conclude that $\mathfrak{m}\{|f'| < \varepsilon \text{ or } f' \geq \varepsilon\} \geq c_0$, which means that the random point x will be good in the sense that either $|f'(x)| < \varepsilon$ or $f'(x) \geq \varepsilon$. The probability that Algorithm 12.1 fails to obtain a good random point in N tries

is at most $(1 - c_0\varepsilon/2)^N$, which can be made at most $1/2$ by taking $N = \Theta(1/\varepsilon)$. We conclude that with probability at least $1/2$, using

$$O\big(\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon}\big) = O\big(\frac{1}{\varepsilon}\big) \quad \text{queries}\,,$$

Algorithm 12.1 finds an ε-stationary point. □

■ 12.3.4 Proof of Theorem 12.2.4

Proof of Theorem 12.2.4. The goal is to show that when $N \le O(1/\varepsilon^2)$, the resisting oracle construction succeeds, and hence no deterministic algorithm can find an ε-stationary point of an arbitrary 1-smooth function with objective gap at most 1 using N queries.

For the resisting oracle construction, the crux of the matter is to show that $a = f(1/\varepsilon) - f(0) \ge 0$. Indeed, if this holds, then since f is clearly bounded below by 0 on $[0, 1/\varepsilon]$ it will follow that $f \ge 0$ on all of $\mathbb{R}$, and hence $f(0) - \inf f \le 1$.

Let I be the set of indices $i \in [N]$ for which $\ell_i \ge 8\varepsilon$. Since f has slope $-\varepsilon$ on all of the linear pieces, then over all of the linear pieces the value of f drops by at most 1 on the interval $[0, 1/\varepsilon]$. The goal is to show that

$$\sum_{i\in I}\{f(x_{i+1}) - f(x_i)\} \overset{!}{\ge} 1\,.$$

To prove this, write

$$\frac{1}{\varepsilon} = \sum_{i=1}^{N}\ell_i = \sum_{i\in I}\ell_i + \sum_{i\in I^{\mathsf{c}}}\ell_i \le \sum_{i\in I}\ell_i + 8\varepsilon\,|I^{\mathsf{c}}|\,.$$

There are two cases to consider. If $|I^{\mathsf{c}}| \ge \frac{1}{16\varepsilon^2}$ queries, then we are done, as the algorithm has made $\Omega(1/\varepsilon^2)$ queries. Otherwise, $|I^{\mathsf{c}}| \le \frac{1}{16\varepsilon^2}$, in which case

$$\frac{1}{2\varepsilon} \le \sum_{i\in I}\ell_i\,.$$

In this second case, we now have

$$\begin{aligned}\sum_{i\in I}\{f(x_{i+1}) - f(x_i)\} = \sum_{i\in I}\Phi_i(x_{i+1}) = \sum_{i\in I}\ell_i\big(\frac{\ell_i}{4} - \varepsilon\big) &\ge \frac{1}{8}\sum_{i\in I}\ell_i^2 \\ &\ge \frac{1}{8\,|I|}\Big(\sum_{i\in I}\ell_i\Big)^2 \ge \frac{1}{32\varepsilon^2\,|I|}\,.\end{aligned}$$

This is greater than 1 provided $|I| \le \frac{1}{32\varepsilon^2}$.

In summary, the resisting oracle construction is valid provided $|I| \le \frac{1}{32\varepsilon^2}$ and $|I^{\mathsf{c}}| \le \frac{1}{16\varepsilon^2}$. Since $|I| + |I^{\mathsf{c}}| = N$, any deterministic algorithm which finds an ε-stationary point must use at least $N \ge \min\{\frac{1}{32\varepsilon^2}, \frac{1}{16\varepsilon^2}\} = \frac{1}{32\varepsilon^2}$ queries, or

$$\mathscr{C}_{\text{det}}(\varepsilon; \mathscr{O}^{1^{\text{st}}}) \ge \frac{1}{32\varepsilon^2}\,. \qquad \square$$

■ 12.3.5 Proof of Theorem 12.2.5

Proof of Theorem 12.2.5. The proof is very similar to the proof of Theorem 12.2.2. We follow the proof up to the point where

$$\begin{aligned} &I\big(\mathscr{O}_{f_J}(x_i); J \;\big|\; \{x_{i'}, \mathscr{O}_{f_J}(x_{i'}) : i' \in [i-1]\}\big) \\ &\qquad \le H\big(\mathscr{O}_{f_J}(x_i) \;\big|\; \{x_{i'}, \mathscr{O}_{f_J}(x_{i'}) : i' \in [i-1]\}\big)\,, \end{aligned}$$

where now $\mathscr{O}_{f_J}(x) = \{f_J(x), f_J'(x)\}$ returns zeroth- and first-order information. The key point now is that since x_i is deterministic (conditioned on previous queries), $\mathscr{O}_{f_J}(x_i)$ can only take a constant number of possible values, and so the above entropy term is $O(1)$ (as opposed to Theorem 12.2.2, in which the entropy term was of order $O(\varepsilon \log(1/\varepsilon))$). Plugging this into Fano's inequality (Theorem 12.3.1), we obtain

$$\mathbb{P}\{\widehat{J} \ne J\} \ge 1 - \frac{O(N) + \ln 2}{\ln(1/\varepsilon)} > \frac{1}{2}\,,$$

provided that $\varepsilon \le \frac{1}{8}$ and $N \le O(\log(1/\varepsilon))$. Hence, $\Omega(\log(1/\varepsilon))$ queries to $\mathscr{O}^{0^{\text{th}}+1^{\text{st}}}$ are necessary to find an ε-stationary point, even for a randomized algorithm. $\square$

■ 12.3.6 Proof of Theorem 12.2.6

We prove the correctness of the algorithms in reverse order, beginning with BinarySearchIII.

Lemma 12.3.5. *Let $f : \mathbb{R} \to \mathbb{R}$ be 1-smooth. Then,* BinarySearchIII *(Algorithm 12.6) terminates with an ε-stationary point of f using $O(\log \frac{x_+ - x_-}{\varepsilon})$ queries to the oracle.*

Proof. Due to the 1-smoothness of f, if $x_+ - x_- < \varepsilon$, then $f' < 0$ on the interval $[x_-, x_+]$, which contradicts the hypothesis $f(x_+) \ge f(x_-)$. Hence, BinarySearchIII recursively calls itself $O(\log \frac{x_+ - x_-}{\varepsilon})$ times. If it calls BinarySearch, then by Lemma 12.3.3 this uses an additional $O(\log \frac{x_+ - x_-}{\varepsilon})$ queries. $\square$

Lemma 12.3.6. *Let $f : \mathbb{R} \to \mathbb{R}$ be 1-smooth. Then,* BinarySearchII *(Algorithm 12.5) terminates with an ε-stationary point of f using $O(\log \frac{x_+ - x_-}{\varepsilon})$ queries to the oracle.*

Proof. First, we check that when BinarySearchII calls itself, the preconditions of BinarySearchII continue to be met. Suppose for instance that $0 \le f(x_-) - f(m) \le \frac{1}{2}(f(x_-) - f(x_+))$. Since $0 \le f(x_-) - f(x_+) \le \frac{\varepsilon}{4}(x_+ - x_-)$ by hypothesis,

$$0 \le f(x_-) - f(m) \le \frac{\varepsilon}{8}(x_+ - x_-) = \frac{\varepsilon}{4}(x_- - m),$$

which is what we wanted to show. The other case is similar.

Next, we argue that BinarySearchII terminates. The hypotheses of BinarySearchII imply that there is an $\varepsilon/2$-stationary point in the interval $[x_-, x_+]$. Indeed, if this were not the case, then $f' \le -\varepsilon/2$ on the entire interval, so $f(x_+) = f(x_-) + \int_{[x_-, x_+]} f' \le f(x_-) - \frac{\varepsilon}{2}(x_+ - x_-)$, but this contradicts the assumption $f(x_-) - f(x_+) \le \frac{\varepsilon}{4}(x_+ - x_-)$. Therefore, if $x_+ - x_- < \frac{\varepsilon}{2}$, it would follow that $f'(x_-) > -\varepsilon$, which contradicts the hypothesis $f'(x_-) \le -\varepsilon$. Since the value of $x_+ - x_-$ is cut in half each time that BinarySearchII calls itself, we conclude that this can happen at most $O(\log \frac{x_+ - x_-}{\varepsilon})$ times. If BinarySearchII calls either BinarySearch or BinarySearchIII, then by Lemma 12.3.3 and Lemma 12.3.5, this uses at most an additional $O(\log \frac{x_+ - x_-}{\varepsilon})$ queries to the oracle. □

Lemma 12.3.7. *Let $f : \mathbb{R} \to \mathbb{R}$ be 1-smooth. Then,* DecreaseGap *(Algorithm 12.4) terminates, either with an ε-stationary point of f, or with a point x such that $f(x) \le f(x_0)$, $f(x + \frac{2}{\varepsilon}) \ge \frac{3}{4} f(x)$, using $O(\log \frac{1}{\varepsilon})$ queries to the oracle.*

Proof. Each time DecreaseGap calls itself, the value of $f(x_0)$ decreases by a factor of $\frac{3}{4}$. If $f'(x_0) \le -\varepsilon$, then from (12.1) we deduce that $f(x_0) \ge \frac{1}{2}|f'(x_0)|^2 \ge \varepsilon^2/2$. Hence, DecreaseGap can call itself at most $O(\log \frac{1}{\varepsilon^2}) = O(\log \frac{1}{\varepsilon})$ times. If it calls BinarySearch, then by Lemma 12.3.3 this uses an additional $O(\log \frac{1}{\varepsilon})$ queries to the oracle. □

Finally, we are ready to verify the correctness of ZerothOrder (Algorithm 12.3).

Proof of Theorem 12.2.6. From Lemma 12.3.7, if $|f'(x_-)| > \varepsilon$ then we must have $f'(x_-) \le -\varepsilon$ and $f(x_+) \ge \frac{3}{4} f(x_-)$. There are two cases. If $f(x_+) \le f(x_-)$, then we know that

$$0 \le f(x_-) - f(x_+) \le \frac{1}{4} f(x_-) \le \frac{1}{4} = \frac{\varepsilon}{8}(x_+ - x_-)$$

so the preconditions of BinarySearchII are met; by Lemma 12.3.6, the algorithm ZerothOrder terminates with an ε-stationary point of f using $O(\log \frac{1}{\varepsilon})$ additional queries. In the other case $f(x_+) \geq f(x_-)$, by Lemma 12.3.5, ZerothOrder again terminates with an ε-stationary point of f using $O(\log \frac{1}{\varepsilon})$ additional queries. This concludes the proof. □

12.4 Conclusion

We have characterized the oracle complexity of finding an ε-stationary point of a smooth univariate function $f : \mathbb{R} \to \mathbb{R}$ in four natural settings of interest. Besides providing insight into the limitations of gradient descent, our results exhibit surprising separations between the power of deterministic and randomized algorithms, and between algorithms that use zeroth-order information and algorithms (like gradient descent) which only use first-order information.

We conclude with a number of open directions for future research.

- The main question motivating this work remains open, namely, for randomized algorithms using zeroth- and first-order information, **is it possible to prove a $\Omega(1/\varepsilon^2)$ complexity lower bound with a construction in dimension $d = O(\log(1/\varepsilon))$?** An affirmative answer to this question would likely build upon the lower bound techniques used in [Vav93; BM20].

  An even more ambitious goal is to fully characterize the query complexity of finding stationary points using zeroth- and first-order information in every fixed dimension d.

- Towards the above question, we also ask: **is there an analogue of gradient flow trapping [BM20] for unconstrained optimization?**

- We have established that among deterministic algorithms which only use first-order queries, gradient descent is optimal already in dimension one. Although randomized algorithms outperform GD in our setting of investigation, it is unclear to what extent randomness helps in higher dimension. Hence, we make the following bold conjecture: **can one prove a $\Omega(1/\varepsilon^2)$ complexity lower bound for randomized algorithms which only make first-order queries in dimension two?**

Chapter 13

Sampling upper bounds in the Fisher information metric

Just as we studied the complexity of finding stationary points for non-convex optimization in §12, in this chapter we explore the concept of first-order stationarity in the context of non-log-concave sampling.

For the task of sampling from a density $\pi \propto \exp(-V)$ on $\mathbb{R}^d$, where V is possibly non-convex but β-smooth, we prove that averaged Langevin Monte Carlo outputs a sample with ε^2-relative Fisher information after $O(\beta^2 d^2/\varepsilon^4)$ iterations. This is the sampling analogue of complexity bounds for finding an ε-approximate first-order stationary points in non-convex optimization and therefore constitutes a first step towards the general theory of non-log-concave sampling.

This chapter is based on [Bal+22], joint with Krishnakumar Balasubramanian, Murat A. Erdogdu, Adil Salim, and Matthew Zhang.

13.1 Introduction

Consider the canonical task of sampling from a density $\pi \propto \exp(-V)$ on $\mathbb{R}^d$, given query access to the gradients of V. In the case where V is strongly convex and smooth, this task is well-studied, with a number of works giving precise and non-asymptotic complexity bounds which scale polynomially in the problem parameters. In contrast, there are comparatively few works which study the case when V is non-convex. In this chapter, we take a first step towards developing a general theory of non-log-concave sampling by formulating the sampling analogue of *stationary point analysis*, which has been highly successful in the non-convex optimization [Nes18].

Recall that the *Langevin diffusion*

$$\mathrm{d}Z_t = -\nabla V(Z_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \tag{13.1}$$

has π as its unique stationary distribution and converges to it as $t \to \infty$ under

mild conditions. Here, $(B_t)_{t\geq 0}$ is a standard d-dimensional Brownian motion. Discretizing this stochastic process with step size $h > 0$ yields the standard *Langevin Monte Carlo* (LMC) algorithm

$$X_{(k+1)h} \coloneqq X_{kh} - h\,\nabla V(X_{kh}) + \sqrt{2}\,(B_{(k+1)h} - B_{kh})\,. \tag{LMC}$$

Several extensions of LMC have been considered in the literature. For instance, a stochastic gradient can be used as an estimate of the "full" gradient $\nabla V(x_{kh})$ at each iteration.

Although LMC and its extensions are ostensibly sampling algorithms, they find applications in optimization. Indeed, LMC and its extensions can be viewed as a variant of (stochastic) gradient descent in which Gaussian noise is explicitly injected in the (stochastic) gradient in each iteration. As explored, for example, in [RRT17; Jin+21], the presence of noise allows the iteration to escape local minima and allows for establishing global non-asymptotic convergence guarantees on well-behaved yet non-convex objectives.

Perhaps surprisingly, the connection between optimization and sampling also goes in the other direction: the theory of optimization can be used to understand the performance of sampling algorithms. On a superficial level, this is anticipated because the Langevin diffusion (13.1) is simply a standard gradient flow to which a Brownian noise has been added. However, there is a much deeper connection, due to [JKO98], which interprets the Langevin diffusion as an *exact* gradient flow in the space of probability measures equipped with the geometry of optimal transport, where the objective functional is the Kullback–Leibler (KL) divergence $\mathsf{KL}(\cdot \,\|\, \pi)$. This perspective has spurred researchers to provide novel optimization-inspired analyses of sampling [Ber18; Wib18; DMM19].

For example, the Wasserstein gradient of $\mathsf{KL}(\cdot \,\|\, \pi)$ at μ is $\nabla \ln(\mu/\pi)$, and the calculation rules for gradient flows imply that if π_t denotes the law of the Langevin diffusion (13.1) at time t, then $\partial_t\, \mathsf{KL}(\pi_t \,\|\, \pi) = -\,\mathbb{E}_{\pi_t}[\|\nabla \ln(\pi_t/\pi)\|^2]$ [see AGS08; Vil09b; San15]. As this quantity is important in what follows, we explicitly write $\mathsf{FI}(\mu \,\|\, \pi) \coloneqq \mathbb{E}_{\mu}[\|\nabla \ln(\mu/\pi)\|^2]$ for the (relative) *Fisher information* of μ w.r.t. π. If V is convex (resp. strongly convex), then the objective functional $\mathsf{KL}(\cdot \,\|\, \pi)$ is convex (resp. strongly convex) in the Wasserstein geometry, which in turn implies that $\mathsf{KL}(\pi_t \,\|\, \pi)$ decays to zero at the rate $O(1/t)$ (resp. exponentially fast).

In the case when V is non-convex, however, less is known. Of course, just like non-convex optimization, it is in general impossible to obtain polynomial sampling guarantees for non-log-concave distributions. Recently, [VW19; Ma+21] study tractable cases of non-log-concave sampling in which the target π satisfies a *functional inequality*, such as the *log-Sobolev inequality* (LSI); see also §3, §4, and §6. Indeed, if LSI holds, then $\mathsf{FI}(\mu \,\|\, \pi) \gtrsim \mathsf{KL}(\mu \,\|\, \pi)$ for all μ. In light of the

Wasserstein calculus described above, this is the analogue of the *gradient domination* condition (or *Polyak–Łojasiewicz inequality*) in non-convex optimization: $\|\nabla V(x)\|^2 \gtrsim V(x) - \min V$ [Łoj63; Pol63; KNS16]. Furthermore, [DM17; Li+19; Che+20b; MMS20; EH21; HBE22] study tractable classes of non-log-concave sampling based on certain tail-growth conditions. However, the assumptions made in these works are far from capturing the breadth of non-log-concave sampling.

Instead, in general non-convex optimization, the standard approach is to prove convergence to a *stationary point* of the objective function, or from a more quantitative perspective, to determine the complexity of obtaining a point x satisfying $\|\nabla V(x)\| \leq \varepsilon$. This complexity is typically $O(1/\varepsilon^2)$ [Nes18]. Following this paradigm, we propose to use the Fisher information as the sampling analogue of the squared norm of the gradient. Our main result (Theorem 13.4.2) establishes that under the sole assumption that ∇V is β-gradient Lipschitz, an averaged version of the LMC algorithm (LMC) outputs a sample whose law μ satisfies $\mathsf{FI}(\mu \parallel \pi) \leq \varepsilon^2$ after $O(\beta^2 d^2/\varepsilon^4)$ iterations. Intuitively, the Fisher information captures the rapid local mixing of the Langevin diffusion near modes of the distribution π, while ignoring the metastability effects which occur between the modes [Bov+02; Bov+04; BGK05]. We give an illustrative example in §13.2 which expands upon this intuition.

Organization and contributions. The rest of the chapter is organized as follows. In §13.2, we provide intuitions on Fisher information guarantees in sampling. In §13.3, we formally define the Fisher information, and in §13.4, we state our main result in Theorem 13.4.2. In §13.5, we consider applications of our main result:

- We show the weak convergence of averaged LMC with decaying step size under general assumptions (§13.5.1).
- We provide new sampling guarantees in the total variation distance under a Poincaré inequality (§13.5.2). These guarantees can be compared with the ones obtained in §3 (in fact, here we obtain substantially better dimension dependence).
- In an effort to stick to the main ideas, we have omitted other applications and extensions of our results which can be found in the full paper [Bal+22]. We give an overview of these additional results in §13.5.3.

Finally, we conclude with open directions in §13.7.

■ 13.2 Interpretation of approximate first-order stationarity in sampling

Intriguingly, unlike the situation in non-convex optimization, in sampling there are no "spurious stationary points": if μ and π have positive and smooth densities and $\mathsf{FI}(\mu \parallel \pi) = 0$, then $\mu = \pi$. However, for $\varepsilon > 0$, it may be unclear what the guarantee $\mathsf{FI}(\mu \parallel \pi) \le \varepsilon^2$ entails. In this section, we give an example illustrating what conclusions may be drawn from a bound on the Fisher information, which helps to better interpret our result in the next sections.

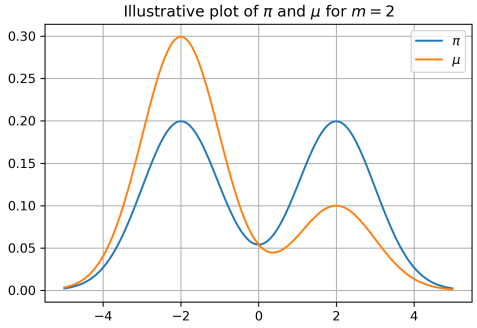


Figure 13.1

Consider a mixture of two Gaussians in one dimension as the target:

$$\pi = \frac{1}{2} \underbrace{\mathsf{normal}(-m, 1)}_{\pi_-} + \frac{1}{2} \underbrace{\mathsf{normal}(+m, 1)}_{\pi_+},$$

where $m \gg 0$. Also, consider a mixture of two Gaussians with different weights:

$$\mu := \frac{3}{4}\pi_- + \frac{1}{4}\pi_+ .$$

An illustrative plot of π and μ is provided in Figure 13.1 for the sake of easier visualization. In §13.6.1, we will prove the following.

Proposition 13.2.1. *Let π and μ be as defined above. For all $m \ge 0$,*

$$\|\mu - \pi\|_{\mathrm{TV}} \ge \frac{1}{4}\left[1 - \exp\left(-\frac{m^2}{2}\right)\right].$$

On the other hand,

$$\mathsf{FI}(\mu \parallel \pi) \le 4m^2 \exp\left(-\frac{m^2}{2}\right) \to 0 \qquad \text{as } m \to \infty .$$

In the next section, we will show that averaged LMC can drive the Fisher information to zero at a polynomial rate. For large m, the measure μ has small Fisher information with respect to π, so μ serves as a model for the kind of distribution that averaged LMC can reach. We can draw a few conclusions:

1. Although the Fisher information $\mathsf{FI}(\mu \parallel \pi)$ is very small, the total variation distance remains bounded away from zero. This shows that a Fisher information guarantee does *not* ensure fast convergence of averaged LMC in other metrics without further assumptions (anyway, polynomial guarantees for non-log-concave sampling in other metrics are impossible in general).

2. Here, μ *locally* captures the correct shape of π at the two modes. On the other hand, μ has different mixing weights than π, which means that μ is *globally* different from π. Since $\mathsf{FI}(\mu \parallel \pi)$ is small for this example, it shows that the Fisher information is not sensitive to the latter effect. Hence, our Fisher information guarantee for averaged LMC captures the fact that the algorithm rapidly gets the local structure of π correct.

3. After a few steps of LMC started at the distribution $\frac{3}{4}\,\delta_{-m} + \frac{1}{4}\,\delta_{+m}$, the algorithm arrives at a measure which closely resembles μ, rather than the true stationary measure π. Indeed, the iterates of LMC do not need to jump from one mode to another to approximate μ. This jumping takes an exponentially long time and is the main barrier to the mixing of LMC, but it is necessary for LMC to learn the global mixing weights—this is known as the *metastability* phenomenon [Bov+02; Bov+04; BGK05]. Our analysis provides a convenient way to quantify this effect.

Remark 13.2.2. *In the context of Bayesian inference, the choice of relative Fisher information metric between the prior and the exact posterior distribution has been proposed by [Wal16; HW17; Sha+19], as a measure of robustness of the overall inferential procedure. In this regard, our results provide a computational angle to this paradigm: in practice we rarely have access to the exact posterior distribution. Our results algorithmically quantify the distance (in relative Fisher information) between the posterior distribution obtained after a certain number of iterations of LMC and the exact posterior.*

■ 13.3 Preliminaries

Throughout the paper, we assume that the potential $V : \mathbb{R}^d \to \mathbb{R}$ is a smooth (i.e., twice continuously differentiable) function such that $\int \exp(-V) < \infty$. The target distribution $\pi \propto \exp(-V)$ is therefore well-defined.

For a probability measure μ with a smooth density, we can define the *Fisher information* of μ relative to π via $\mathsf{FI}(\mu \parallel \pi) := \mathbb{E}_\mu[\|\nabla \ln(\mu/\pi)\|^2]$. To extend this definition to other probability measures, we recall from Markov semigroup theory [see BGL14] that we associate with the Langevin diffusion (13.1) a *Dirichlet energy* $f \mapsto \mathscr{E}(f)$ which maps a subspace $\operatorname{dom}\mathscr{E} \subseteq L^2(\pi)$ to $\mathbb{R}_+$. If f is smooth and compactly supported, then $f \in \operatorname{dom}\mathscr{E}$ and the Dirichlet energy has the explicit expression $\mathscr{E}(f) = \mathbb{E}_\pi[\|\nabla f\|^2]$. The Fisher information is defined from the Dirichlet energy as follows. For an arbitrary probability measure μ, set

$$\mathsf{FI}(\mu \parallel \pi) := \begin{cases} 4\,\mathscr{E}(\sqrt{f})\,, & \text{if } f := \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \text{ exists and } \sqrt{f} \in \operatorname{dom}\mathscr{E}\,, \\ +\infty\,, & \text{otherwise}\,. \end{cases}$$

In particular, if $f = \frac{\mathrm{d}\mu}{\mathrm{d}\pi}$ is positive and smooth, one can check that

$$\mathsf{FI}(\mu \parallel \pi) = \int \|\nabla \ln(f)\|^2 \,\mathrm{d}\mu\,, \qquad \text{or} \qquad \mathsf{FI}(\mu \parallel \pi) = \int \frac{\|\nabla f\|^2}{f}\,\mathrm{d}\pi\,.$$

Using the convexity of $(a, b) \mapsto \|a\|^2/b$ on $\mathbb{R}^d \times \mathbb{R}_+$, the latter formula implies that the Fisher information $\mu \mapsto \mathsf{FI}(\mu \,\|\, \pi)$ is convex in the classical sense on the space of probability measures. Besides, the Fisher information is also lower semicontinuous in its first argument with respect to the weak topology of measures [see, e.g., Wu00, Appendix B].

■ 13.4 Main result

Recall that the LMC algorithm is given by

$$X_{(k+1)h} := X_{kh} - h\,\nabla V(X_{kh}) + \sqrt{2}\,(B_{(k+1)h} - B_{kh})\,.$$

Our main result is stated for the following continuous interpolation of LMC:

$$X_t := X_{kh} - (t - kh)\,\nabla V(X_{kh}) + \sqrt{2}\,(B_t - B_{kh}) \qquad \text{for } t \in [kh, (k+1)h]\,. \tag{13.2}$$

We write μ_t for the law of X_t.

Assumption 13.4.1. *The gradient of V is β-Lipschitz continuous: for some $\beta > 0$, $\|\nabla V(x_1) - \nabla V(x_2)\| \le \beta\,\|x_1 - x_2\|$ for all $x_1, x_2 \in \mathbb{R}^d$.*

Theorem 13.4.2. *Let $(\mu_t)_{t\ge 0}$ denote the law of the interpolation (13.2) of LMC, and let the potential V satisfy Assumption 13.4.1. Then, for any step size $h \in (0, \frac{1}{6\beta})$, it holds that*

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \parallel \pi)\,\mathrm{d}t \le \frac{2\,\mathsf{KL}(\mu_0 \parallel \pi)}{Nh} + 8\beta^2 dh\,.$$

In particular, if $\mathsf{KL}(\mu_0 \,\|\, \pi) \le K_0$ *and we choose* $h = \sqrt{K_0}/(2\beta\sqrt{dN})$*, then for* $N \ge 9K_0/d$*,*

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \le \frac{8\beta\sqrt{dK_0}}{\sqrt{N}}\,.$$

By the convexity of the Fisher information, it follows that the averaged distribution $\bar{\mu}_{Nh} \coloneqq (Nh)^{-1}\int \mu_t\,\mathrm{d}t$ satisfies $\mathsf{FI}(\bar{\mu}_{Nh} \,\|\, \pi) \le 8\beta\sqrt{dK_0/N}$ as well. Also, it is possible to output a sample from $\bar{\mu}_{Nh}$, as follows:

1. Pick a time $t \in [0, Nh]$ uniformly at random.

2. Let k be the largest integer such that $kh \le t$, and let X_{kh} be the iterate of LMC at time kh. Then, perform a partial LMC update for time $t - kh$, i.e.,

$$X_t \coloneqq X_{kh} - (t - kh)\,\nabla V(X_{kh}) + \sqrt{2}\,(B_t - B_{kh})\,.$$

   Then, X_t is a sample from $\bar{\mu}_{Nh}$. Note that it is possible to sample the Brownian increments exactly as long as one can sample standard Gaussians.

Remark 13.4.3. *Since we can usually take* K_0 *to be of order* d*, see, e.g., [VW19, Lemma 1] or §3.6.6, in order for averaged LMC to reach* ε^2 *accuracy in terms of the Fisher information w.r.t. the target, the iteration complexity is* $O(\beta^2 d^2/\varepsilon^4)$*.*

13.5 Applications

13.5.1 Asymptotic convergence of averaged LMC with vanishing step size

Our main result immediately implies asymptotic convergence of averaged LMC with decreasing step size under very general conditions. Let $(h_k)_{k=1}^{\infty}$ be a sequence of positive step sizes such that

$$\sum_{k=1}^{\infty} h_k = \infty \qquad \text{and} \qquad \sum_{k=1}^{\infty} h_k^2 < \infty\,. \tag{13.3}$$

Write $\tau_n \coloneqq \sum_{k=1}^{n} h_k$, and denote by $\bar{\mu}_{\tau_n} \coloneqq \tau_n^{-1}\int_0^{\tau_n} \mu_t\,\mathrm{d}t$, where μ_t is the law of X_t defined by

$$X_t = X_{\tau_{n-1}} - (t - \tau_{n-1})\,\nabla V(X_{\tau_{n-1}}) + \sqrt{2}\,(B_t - B_{\tau_{n-1}})\,, \qquad t \in [\tau_{n-1}, \tau_n]\,.$$

Then, we have the following convergence result.

Theorem 13.5.1. *Let $(\mu_t)_{t\geq 0}$ denote the law of the interpolation (13.2) of LMC, and let the potential V satisfy Assumption 13.4.1. Suppose that LMC is initialized at μ_0 with $\mathsf{KL}(\mu_0 \parallel \pi) < \infty$ and that the step size sequence $(h_k)_{k=1}^{\infty}$ satisfies $h_k \in (0, \frac{1}{6\beta})$ for every k, as well as the conditions (13.3). Then, $\bar{\mu}_{\tau_n} \to \pi$ weakly.*

While it might be possible to prove the weak convergence of LMC using other techniques, for example, the ordinary differential equation method from the stochastic approximation literature [KY03] or general results on the analysis of Markov chains [BGL14; Dou+18], we emphasize that Theorem 13.5.1 follows immediately from our main result in Theorem 13.4.2 and the connectedness property that $\mathsf{FI}(\mu \parallel \pi) = 0$ implies $\mu = \pi$. To the best of our knowledge, explicit results available in the literature on the weak convergence of LMC [e.g., LP02; PP12] require Lyapunov-type conditions. In comparison, Theorem 13.5.1 holds just under the Lipschitz gradient assumption on the potential V. See also the recent results of [KHK23].

■ 13.5.2 New sampling guarantees under a Poincaré inequality

In this section, we show that if we additionally assume that π satisfies a Poincaré inequality, then we obtain sampling guarantees in total variation distance as a corollary of our main theorem. Surprisingly, the rates we obtain in this manner are competitive with (and arguably better than) the state-of-the-art results for LMC, for these classes of target distributions. To present our result, we recall the following transport inequality.

Lemma 13.5.2 ([Gui+09, Theorem 3.1])**.** *Suppose that π satisfies a Poincaré inequality: for all smooth compactly supported functions $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\operatorname{var}_\pi f \leq C_{\mathsf{PI}}\, \mathbb{E}_\pi[\|\nabla f\|^2]\,. \tag{PI}$$

Then, for all probability measures μ,

$$\|\mu - \pi\|_{\mathrm{TV}}^2 \leq 4 C_{\mathsf{PI}}\, \mathsf{FI}(\mu \parallel \pi)\,.$$

When combined with Theorem 13.4.2, we immediately obtain:

Corollary 13.5.3. *Let $(\mu_t)_{t\geq 0}$ denote the law of the interpolation (13.2) of LMC, and let the potential V satisfy Assumption 13.4.1. If $\mathsf{KL}(\mu_0 \parallel \pi) \leq K_0$ and we choose $h = \sqrt{K_0}/(2\beta\sqrt{dN})$, then for $N \geq 9K_0/d$ and $\bar{\mu}_{Nh} \coloneqq (Nh)^{-1} \int_0^{Nh} \mu_t \,\mathrm{d}t$,*

$$\|\bar{\mu}_{Nh} - \pi\|_{\mathrm{TV}}^2 \leq \frac{32 C_{\mathsf{PI}} \beta \sqrt{d K_0}}{\sqrt{N}}\,.$$

Remark 13.5.4. *If $K_0 = O(d)$, Corollary 13.5.3 implies an iteration complexity of $O(C_{\mathsf{PI}}^2 \beta^2 d^2/\varepsilon^4)$ to output a sample whose total variation distance to π is at most ε. In contrast, Theorem 3.3.4 yields an iteration complexity of $\widetilde{O}(C_{\mathsf{PI}}^2 \beta^2 d^3/\varepsilon^2)$ for LMC (without averaging). Corollary 13.5.3 has worse dependence on the inverse accuracy, but better dependence on the dimension.*

13.5.3 Further applications and extensions

In this section, we describe further results contained in [Bal+22].

- If the potential V satisfies an additional Hessian smoothness condition, i.e., $\nabla^2 V$ is Lipschitz in the operator norm, then under an additional mild dissipativity condition we improve the complexity to reach ε^2 Fisher information to $O(d^2/\varepsilon^3)$ [Bal+22, Theorem 12].

- We also extend Theorem 13.4.2 to cover the use of stochastic gradients with bounded bias and variance [Bal+22, Theorem 15].

- A particular application of the stochastic gradient result is to cover non-smooth potentials by applying the Gaussian smoothing technique [following Cha+20]. By choosing the smoothing level to balance the bias and variance [see NS17] and incorporating mini-batching, we extend Theorem 13.4.2 and Corollary 13.5.3 to the non-smooth case [Bal+22, Corollaries 18 and 19].

- Finally, when the potential is a finite sum $V = \sum_{i=1}^n V_i$, then we consider using a variance-reduced stochastic gradient given by PAGE [Li+21] in order to provide a guarantee in terms of the number of individual gradient evaluations [Bal+22, Theorem 21].

13.6 Proofs

13.6.1 Proof for the illustrative example

Proof of Proposition 13.2.1. The total variation distance is

$$\|\mu - \pi\|_{\mathrm{TV}} = \frac{1}{2}\int |\mu - \pi| = \frac{1}{8}\int |\pi_+ - \pi_-| = \frac{1}{4}\,\|\pi_+ - \pi_-\|_{\mathrm{TV}}\,.$$

Since $\pi_- = \mathsf{normal}(-m, 1)$ and $\pi_+ = \mathsf{normal}(m, 1)$, standard Gaussian tail estimates yield

$$\pi_-(\mathbb{R}_+) \le \frac{1}{2}\exp\bigl(-\frac{m^2}{2}\bigr)\,, \qquad \pi_+(\mathbb{R}_+) \ge 1 - \frac{1}{2}\exp\bigl(-\frac{m^2}{2}\bigr)\,,$$

and the lower bound on $\|\mu - \pi\|_{\mathrm{TV}}$ follows.

Next, we have

$$\begin{aligned}\nabla \ln \frac{\mu}{\pi} &= \frac{1}{\mu}\left(\frac{3}{4}\nabla\pi_- + \frac{1}{4}\nabla\pi_+\right) - \frac{1}{\pi}\left(\frac{1}{2}\nabla\pi_- + \frac{1}{2}\nabla\pi_+\right) \\ &= \frac{1}{\mu\pi}\left[\pi\left(\frac{3}{4}\nabla\pi_- + \frac{1}{4}\nabla\pi_+\right) + \mu\left(\frac{1}{2}\nabla\pi_- + \frac{1}{2}\nabla\pi_+\right)\right].\end{aligned}$$

Writing $s_\mp := \nabla \ln \pi_\mp$, some algebra reveals that

$$\nabla \ln \frac{\mu}{\pi} = \frac{1}{4\mu\pi}\left(\pi_+\nabla\pi_- - \pi_-\nabla\pi_+\right) = \frac{\pi_-\pi_+}{4\mu\pi}\left(s_- - s_+\right) = -\frac{\pi_-\pi_+}{2\mu\pi}\, m\,.$$

Therefore,

$$\begin{aligned}\mathsf{FI}(\mu \,\|\, \pi) &= \frac{m^2}{4}\int \frac{\pi_-^2\pi_+^2}{\mu^2\pi^2}\,\mathrm{d}\mu = \frac{m^2}{4}\int \frac{\pi_-^2\pi_+^2}{\mu\pi^2} = \frac{m^2}{4}\int \frac{\pi_-^2\pi_+^2}{\left(\frac{3}{4}\,\pi_- + \frac{1}{4}\,\pi_+\right)\left(\frac{1}{2}\,\pi_- + \frac{1}{2}\,\pi_+\right)^2} \\ &\le 4m^2\int \frac{\pi_-^2\pi_+^2}{\left(\pi_- + \pi_+\right)^3} \le 4m^2\left[\int_{\mathbb{R}_-}\frac{\pi_+^2}{\pi_-} + \int_{\mathbb{R}_+}\frac{\pi_-^2}{\pi_+}\right].\end{aligned}$$

Writing $Z := (2\pi)^{d/2}$ for the normalizing constant,

$$\begin{aligned}\int_{\mathbb{R}_+}\frac{\pi_-^2}{\pi_+} &= \frac{1}{Z}\int_0^\infty \exp\bigl(-|x+m|^2 + \frac{1}{2}\,|x-m|^2\bigr)\,\mathrm{d}x \\ &= \frac{\exp(4m^2)}{Z}\int_0^\infty \exp\bigl(-\frac{1}{2}\,|x+3m|^2\bigr)\,\mathrm{d}x = \exp(4m^2)\,\mathbb{P}\{\xi \ge 3m\}\end{aligned}$$

where ξ is a standard Gaussian random variable. Using a Gaussian tail bound,

$$\mathbb{P}\{\xi \ge 3m\} \le \frac{1}{2}\exp\bigl(-\frac{9m^2}{2}\bigr)\,.$$

A symmetric argument holds for the other integral, and hence

$$\mathsf{FI}(\mu \,\|\, \pi) \le 4m^2\exp\bigl(-\frac{m^2}{2}\bigr)$$

which completes the proof. □

■ 13.6.2 Proof of the main theorem

Our proof follows the interpolation argument of [VW19] which proceeds by obtaining a differential inequality for the KL divergence along an interpolation of the algorithm. Although these lemmas also appeared in §3, we reproduce them here for convenience.

Lemma 13.6.1. *Along the interpolation (13.2), writing μ_t for the law of X_t, it holds that*

$$\partial_t \mathsf{KL}(\mu_t \,\|\, \pi) \le -\frac{3}{4}\mathsf{FI}(\mu_t \,\|\, \pi) + \mathbb{E}[\|\nabla V(X_t) - \nabla V(X_{kh})\|^2]\,.$$

Proof. See Proposition 3.6.1. □

Lemma 13.6.2. *Assume that ∇V is β-Lipschitz. For any probability measure μ, it holds that*

$$\mathbb{E}_\mu[\|\nabla V\|^2] \le \mathsf{FI}(\mu \,\|\, \pi) + 2\beta d\,.$$

Proof. See Lemma 3.6.3. □

We now prove our main result.

Proof of Theorem 13.4.2. Let $(X_t)_{t\ge 0}$ denote the interpolation of LMC (defined in (13.2)). For $t \in [kh, (k+1)h]$, Lemma 13.6.1 yields

$$\partial_t \mathsf{KL}(\mu_t \,\|\, \pi) \le -\frac{3}{4}\mathsf{FI}(\mu_t \,\|\, \pi) + \mathbb{E}[\|\nabla V(X_t) - \nabla V(X_{kh})\|^2]$$

and the error term is

$$\begin{aligned}\mathbb{E}[\|\nabla V(X_t) - \nabla V(X_{kh})\|^2] &\le \beta^2\,\mathbb{E}[\|X_t - X_{kh}\|^2] \\ &\le 2\beta^2\,(t-kh)^2\,\mathbb{E}[\|\nabla V(X_{kh})\|^2] + 4\beta^2\,\mathbb{E}[\|B_t - B_{kh}\|^2]\,.\end{aligned}$$

Next, since ∇V is Lipschitz,

$$\begin{aligned}\|\nabla V(X_{kh})\| &\le \|\nabla V(X_t)\| + \beta\,\|X_t - X_{kh}\| \\ &\le \|\nabla V(X_t)\| + \beta h\,\|\nabla V(X_{kh})\| + \sqrt{2}\beta\,\|B_t - B_{kh}\|\,,\end{aligned}$$

and for $h \le 1/(3\beta)$ we can rearrange this to yield

$$\|\nabla V(X_{kh})\| \le \frac{3}{2}\,\|\nabla V(X_t)\| + \frac{3\beta}{\sqrt{2}}\,\|B_t - B_{kh}\|\,.$$

Plugging this in,

$$\|\nabla V(X_t) - \nabla V(X_{kh})\|^2 \le 9\beta^2\,(t-kh)^2\,\|\nabla V(X_t)\|^2 + 6\beta^2\,\|B_t - B_{kh}\|^2\,. \quad (13.4)$$

For the expectation of the first term, we can use Lemma 13.6.2 to bound

$$\mathbb{E}_{\mu_t}[\|\nabla V\|^2] \le \mathsf{FI}(\mu_t \,\|\, \pi) + 2\beta d\,.$$

Hence, for $h \le 1/(6\beta)$,

$$\begin{aligned}\partial_t \mathsf{KL}(\mu_t \,\|\, \pi) &\le -\big(\frac{3}{4} - 9\beta^2 h^2\big)\, \mathsf{FI}(\mu_t \,\|\, \pi) + 18\beta^3 d\,(t-kh)^2 + 6\beta^2 d\,(t-kh) \\ &\le -\frac{1}{2}\,\mathsf{FI}(\mu_t \,\|\, \pi) + 18\beta^3 d\,(t-kh)^2 + 6\beta^2 d\,(t-kh)\,. \end{aligned} \tag{13.5}$$

Integrating, we obtain

$$\begin{aligned}\mathsf{KL}(\mu_{(k+1)h} \,\|\, \pi) - \mathsf{KL}(\mu_{kh} \,\|\, \pi) &\le -\frac{1}{2}\int_{kh}^{(k+1)h} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t + 6\beta^3 dh^3 + 3\beta^2 dh^2 \\ &\le -\frac{1}{2}\int_{kh}^{(k+1)h} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t + 4\beta^2 dh^2\,. \end{aligned} \tag{13.6}$$

Now by summing, we have

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \le \frac{2\,\mathsf{KL}(\mu_0 \,\|\, \pi)}{Nh} + 8\beta^2 dh\,.$$

This concludes the proof. □

■ 13.6.3 Asymptotic convergence of averaged LMC

Proof of Theorem 13.5.1. The one-step recursion (13.6) in the proof of Theorem 13.4.2 yields

$$\mathsf{KL}(\mu_{\tau_n} \,\|\, \pi) - \mathsf{KL}(\mu_{\tau_{n-1}} \,\|\, \pi) \le -\frac{1}{2}\int_{\tau_{n-1}}^{\tau_n} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t + 4\beta^2 d h_n^2\,.$$

Iterating the above bound, we obtain

$$\mathsf{KL}(\mu_{\tau_n} \,\|\, \pi) \le \mathsf{KL}(\mu_0 \,\|\, \pi) - \frac{1}{2}\int_0^{\tau_n} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t + 4\beta^2 d \sum_{k=1}^{n} h_k^2\,.$$

Rearranging the terms, dividing by τ_n, and using the convexity of the Fisher information,

$$\mathsf{FI}(\bar{\mu}_{\tau_n} \,\|\, \pi) \le \frac{1}{\tau_n}\int_0^{\tau_n} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \le \frac{2\,\mathsf{KL}(\mu_0 \,\|\, \pi)}{\tau_n} + \frac{8\beta^2 d}{\tau_n}\,S\,, \tag{13.7}$$

where $S \coloneqq \sum_{k=1}^{\infty} h_k^2 < \infty$. On the other hand, if $t \in [\tau_n, \tau_{n+1}]$, integrating (13.5) between τ_n and t shows that

$$\mathsf{KL}(\mu_t \,\|\, \pi) \le \mathsf{KL}(\mu_{\tau_n} \,\|\, \pi) + 4\beta^2 d\,(t-\tau_n)^2 \le \mathsf{KL}(\mu_0 \,\|\, \pi) + 8\beta^2 dS < \infty\,,$$

so that $\{\mathsf{KL}(\mu_t \,\|\, \pi) \mid t \geq 0\}$ is bounded. By convexity of the KL divergence, it also implies that $\{\mathsf{KL}(\bar{\mu}_{\tau_n} \,\|\, \pi) \mid n \in \mathbb{N}\}$ is uniformly bounded. Recalling that the sublevel sets of $\mathsf{KL}(\cdot \,\|\, \pi)$ are weakly compact we obtain that $(\bar{\mu}_{\tau_n})_{n\in\mathbb{N}}$ is tight. To show that $\bar{\mu}_{\tau_n} \to \pi$ weakly, it suffices to show that every cluster point of $(\bar{\mu}_{\tau_n})_{n\in\mathbb{N}}$ is equal to π.

Consider a subsequence of $(\bar{\mu}_{\tau_n})_{n\in\mathbb{N}}$ converging to some cluster point $\bar{\mu}$. Taking $n \to \infty$ in (13.7) and noting that $\tau_n \to \infty$ by our assumptions, $\mathsf{FI}(\bar{\mu}_{\tau_n} \,\|\, \pi) \to 0$, therefore this is still true along the subsequence. Using the weak lower semicontinuity of the Fisher information along the subsequence, $\mathsf{FI}(\bar{\mu} \,\|\, \pi) = 0$. This means that for $f \coloneqq \frac{\mathrm{d}\bar{\mu}}{\mathrm{d}\pi}$, we have $\sqrt{f} \in \operatorname{dom} \mathscr{E}$ and $\mathscr{E}(\sqrt{f}) = 0$. Since ∇V is Lipschitz, then π has a continuous and strictly positive density on $\mathbb{R}^d$, so $\mathscr{E}(\sqrt{f}) = 0$ implies that f is a constant π-a.e., and hence $\bar{\mu} = \pi$. □

13.7 Conclusion

In this chapter, we have initiated the study of non-log-concave sampling by proving that, under the sole assumption that the potential has a Lipschitz gradient, averaged LMC drives the Fisher information w.r.t. the target to zero after polynomially many iterations. We have argued that this is the natural sampling analogue of finding approximate first-order stationary points in non-convex optimization.

Although our focus was to work under the minimal assumption of smoothness, surprisingly our analysis yielded new results for sampling from targets satisfying a Poincaré inequality, and moreover our results attain state-of-the-art dimension dependence for these settings for LMC.

We believe there are many intriguing directions for future work, and we list a few to conclude.

1. (lower bounds) We ask whether one can prove matching lower bounds on the complexity of outputting a sample whose Fisher information w.r.t. the target is ε^2. Since the setting of this work is fully non-log-concave, it may be easier to produce lower bound constructions than the strongly log-concave case, in which the theory of lower bounds is nascent (see §7). In §14, we investigate this lower bound question in detail and in doing so we establish further connections between non-convex optimization and non-log-concave sampling, although pinning down the complexity of obtaining Fisher information guarantees is still an open question in many regimes.

2. (improved results and further extensions) Although we have provided results under Hessian smoothness and via variance reduction, our investigation is still preliminary and we believe that these results can be strengthened. Ad-

ditionally, there are other important extensions to consider; for instance, is there an analogue of second-order stationarity in sampling?

3. (Poincaré case) The iteration complexity we obtained for smooth potentials which satisfy a Poincaré inequality (focusing only on dimension and accuracy) is $O(d^2/\varepsilon^4)$, whereas in §3 we obtained a complexity of $\widetilde{O}(d^3/\varepsilon^2)$. Is it possible to achieve $\widetilde{O}(d^2/\varepsilon^2)$ with a variant of LMC? If so, is averaging necessary?

Chapter 14

Sampling lower bounds in the Fisher information metric

We prove two lower bounds for the complexity of non-log-concave sampling within the framework of §13, which introduced the use of Fisher information (FI) bounds as a notion of approximate first-order stationarity in sampling. Our first lower bound shows that averaged Langevin Monte Carlo (LMC) is optimal for the regime of large FI by reducing the problem of finding stationary points in non-convex optimization to sampling. Our second lower bound shows that in the regime of small FI, obtaining a FI of at most ε^2 from the target distribution requires $\mathrm{poly}(1/\varepsilon)$ queries, which is surprising as it rules out the existence of high-accuracy algorithms (e.g., algorithms using Metropolis–Hastings filters) in this context.

This chapter is based on [Che+23c], joint with Patrik R. Gerber, Holden Lee, and Chen Lu.

14.1 Introduction

What is the query complexity of sampling from a β-log-smooth but possibly non-log-concave target distribution π on $\mathbb{R}^d$? Until recently, this question was only investigated from an upper bound perspective, and only for restricted classes of distributions, such as distributions satisfying functional inequalities [see §3, §4, §6, and VW19; Wib19; Ma+21], distributions with tail decay conditions [DM17; Xu+18; Li+19; Che+20b; MMS20; EH21; ZXG21; HBE22], or mixtures of log-concave distributions [LRG18].

In §13, we developed a general framework to investigate non-log-concave sampling. Motivated by stationary point analysis in non-convex optimization [see, e.g., Nes18] and the interpretation of sampling as optimization over the space of probability measures [JKO98; Wib18], we proposed to call any measure μ satisfying $\sqrt{\mathsf{FI}(\mu \parallel \pi)} \le \varepsilon$ an ε-stationary point for sampling, where $\mathsf{FI}(\mu\|\pi) \coloneqq \mathbb{E}_\mu[\|\nabla \ln \frac{\mu}{\pi}\|^2]$ denotes the *relative Fisher information* of μ from π. In §13.2, we explained

the interpretation of this condition via the classical phenomenon of *metastability* [Bov+02; Bov+04; BGK05]; in particular, for a multimodal distribution, small Fisher information means that the distribution locally approximates the shape at each mode, but not necessarily the relative weights between the modes. We further showed (Theorem 13.4.2) that averaged Langevin Monte Carlo (LMC) can find an ε-stationary point in $O(\beta^2 d K_0/\varepsilon^4)$ iterations, where $K_0 := \mathsf{KL}(\mu_0 \,\|\, \pi)$ is the initial Kullback–Leibler (KL) divergence to the target π.

In the field of optimization, however, there are also corresponding *lower bounds* on the complexity of finding stationary points [see §12 and Vav93; Nes12; BM20; Car+20; Car+21]. Such lower bounds are important for identifying optimal algorithms and understanding the fundamental difficulty of the task at hand. For example, the work of [Car+20] shows that the standard gradient descent algorithm is optimal for finding stationary points of smooth functions, at least in high dimension.

In this chapter, we establish the first lower bounds for Fisher information guarantees for sampling. As we discuss below, our results also reveal a surprising equivalence between the task of obtaining a sample which has moderate Fisher information relative to a target distribution and the task of finding an approximate stationary point of a smooth function, thereby strengthening the connection between the fields of non-convex optimization and non-log-concave sampling.

Our contributions. We now informally describe our main results. Details on notation, our oracle model, and the definition of query complexity for sampling (Definition 14.2.2) are given in §14.2. Precise statements of our results are given in §14.3 and §14.4. For a density $\pi \propto \exp(-U)$ the function $U : \mathbb{R}^d \to \mathbb{R}$ is called the *potential*. Throughout, our notion of complexity is the number of queries made to an oracle that returns the value of U (up to an additive constant) and its gradients. For a 1-smooth function $V : \mathbb{R}^d \to \mathbb{R}$ and $\beta > 0$ let us define the density $\pi_\beta \propto \exp(-\beta V)$, assuming it is well-defined (i.e., $\int \exp(-\beta V) < \infty$).

Our first result connects the task of obtaining Fisher information guarantees with finding stationary points in non-convex optimization, for a particular regime of large smoothness β.

Theorem 14.1.1 (Equivalence, informal)**.** *The following problems are equivalent.*

1. *Output an ε-stationary point of V.*
2. *Output a sample from a measure μ such that $\mathsf{FI}(\mu \,\|\, \pi_\beta) \lesssim \beta d$, where $\beta \asymp d/\varepsilon^2$.*

By combining this equivalence with the lower bound of [Car+20] for finding ε-stationary points, we obtain:

Theorem 14.1.2 (First lower bound, informal)**.** *The number of queries required to obtain a sample from a measure μ satisfying $\sqrt{\mathsf{FI}(\mu \,\|\, \pi_\beta)} \lesssim \sqrt{\beta d}$, starting from an initial distribution μ_0 with KL divergence $K_0 \coloneqq \mathsf{KL}(\mu_0 \,\|\, \pi_\beta)$, is at least $\Omega(K_0/d)$. The lower bound is attained by averaged LMC (Langevin Monte Carlo) as given in Theorem 13.4.2.*

To our knowledge an optimality result for LMC was not previously known in any setting.

The first lower bound addresses the regime of large Fisher information, i.e., $\mathsf{FI}(\mu \,\|\, \pi_\beta) \lesssim \beta d$. In order to target the regime of small Fisher information, we give a construction based on hiding a bump of large mass and prove the following:

Theorem 14.1.3 (Second lower bound, informal)**.** *The number of queries required to obtain a sample from a measure μ satisfying $\sqrt{\mathsf{FI}(\mu \,\|\, \pi_\beta)} \le \varepsilon$, starting from an initial distribution μ_0 with KL divergence $K_0 \coloneqq \mathsf{KL}(\mu_0 \,\|\, \pi_\beta) \le 1$, is at least $(\sqrt{\beta}/\varepsilon)^{2d/(d+2)-o(1)}$ as $\varepsilon \searrow 0$.*

We give a more precise form of our lower bound in §14.4. In infinite dimension (actually, $d \ge \widetilde{\Omega}(\sqrt{\log(\beta/\varepsilon^2)})$ suffices, see §14.4), the lower bound reads $\widetilde{\Omega}(\beta/\varepsilon^2)$, which can be compared to the averaged LMC upper bound of $O(\beta^2 d/\varepsilon^4)$. It is an open question to close this gap.

In terms of technical novelty, we note that the difficulty of showing the first lower bound lies mainly in establishing the equivalence between optimization and sampling, after which lower bounds from optimization apply; on the other hand, the second lower bound requires significant technical work to establish.

We next discuss implications of our results.

- **Towards a theory of lower bounds for sampling.** The problem of obtaining sampling lower bounds is a notorious open problem raised in many prior works [see, e.g., Che+18b; GLL20; LST21a; CBL22]. So far, unconditional lower bounds have only been obtained in restricted settings such as in dimension 1; see §7 and the discussion therein, as well as the reduction to optimization in [GLL22]. Our lower bounds are the first of their kind for Fisher information guarantees, and are some of the *only* lower bounds for sampling in general. Hence, our results take a significant step towards a better understanding of the complexity of sampling. In particular, our first lower bound identifies a regime in which (averaged) LMC is *optimal*, which was not previously known in any setting.

- **Stronger connections between non-convex optimization and non-log-concave sampling.** The equivalence in Theorem 14.1.1 provides compelling

evidence that Fisher information guarantees are the correct analogue of stationary point guarantees in non-convex optimization, thereby supporting the framework of §13.

- **Obtaining an approximate stationary point in sampling is strictly harder for non-log-concave targets.** Ignoring the dependence on other parameters besides the accuracy, our second lower bound yields a $\mathrm{poly}(1/\varepsilon)$ lower bound for the Fisher information task for non-log-concave targets. In contrast, it is morally possible to solve this task in $\mathrm{polylog}(1/\varepsilon)$ queries for *log-concave* targets; see §14.5 for justification. This exhibits a stark separation between log-concave and non-log-concave sampling. Note that the analogous separation does not exist in the context of optimization, because there is a $\mathrm{poly}(1/\varepsilon)$ lower bound for finding an ε-stationary point of a convex and smooth function [Car+21].

- **A separation between optimization and sampling.** Finally, our second lower bound yields a $\mathrm{poly}(1/\varepsilon)$ lower bound, even in dimension one. In contrast, for the analogous question in optimization of finding an ε-stationary point of a univariate function, we exhibited in §12 an algorithm with $O(\log(1/\varepsilon))$ complexity. To our knowledge, this is one of the first instances in which sampling is provably harder than optimization.

■ 14.2 Notation and setting

Notation. Given a probability measure π on $\mathbb{R}^d$ which admits a density w.r.t. the Lebesgue measure, we abuse notation by identifying π with its density.

The class of distributions that we wish to sample from are the β-log-smooth distributions on $\mathbb{R}^d$, defined as follows:

Definition 14.2.1 (Log-smooth distributions)**.** *The class of β-log-smooth distributions consists of distributions π_β supported on $\mathbb{R}^d$ whose densities are of the form $\pi \propto \exp(-U_\beta)$, for potential functions $U_\beta : \mathbb{R}^d \to \mathbb{R}$ that are twice continuously differentiable, and satisfy*

$$\|\nabla U_\beta(x) - \nabla U_\beta(y)\| \le \beta \|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

Oracle model. We work under the following oracle model. The algorithm is given access to a target distribution π in our class via two oracles: initialization and local information. The initialization oracle outputs samples from some distribution μ_0 for which $\mathsf{KL}(\mu_0 \,\|\, \pi) \le K_0$. The local oracle for π, given a query point $x \in \mathbb{R}^d$, returns the value of the potential (up to an additive constant) and its gradient at

the query point x, i.e., the tuple $(U_\beta(x), \nabla U_\beta(x))$. Algorithms can access samples from μ_0 for free, and we care about the number of local information queries needed. The query complexity is defined as follows.

Definition 14.2.2 (Query complexity)**.** *Let $\mathscr{C}(d, K_0, \varepsilon; \beta)$ be the largest number $n \in \mathbb{N}$ such that any algorithm which works in the oracle model described above and outputs a sample from a measure μ_β satisfying $\sqrt{\mathsf{FI}(\mu_\beta \,\|\, \pi_\beta)} \le \varepsilon$, for any β-log-smooth target π_β and any valid initialization oracle for π_β, requires at least n queries to the local oracle for π_β.*

The upper bound of Theorem 13.4.2 shows that using averaged LMC,

$$\mathscr{C}(d, K_0, \varepsilon; \beta) \lesssim 1 \vee \frac{\beta^2 d K_0}{\varepsilon^4} \,. \tag{14.1}$$

We also note the following rescaling lemma.

Lemma 14.2.3 (Rescaling)**.** *It holds that*

$$\mathscr{C}(d, K_0, \varepsilon; \beta) = \mathscr{C}\big(d, K_0, \frac{\varepsilon}{\sqrt{\beta}}; 1\big) \,.$$

Proof. Suppose that $U_\beta : \mathbb{R}^d \to \mathbb{R}$ is β-smooth and that $\pi_\beta \propto \exp(-U_\beta)$ is a density. Define the rescaled potential $U : \mathbb{R}^d \to \mathbb{R}$ via $U(x) := U_\beta(x/\sqrt{\beta})$, and let $\pi \propto \exp(-U)$. (Note that the relationship between π and π_β is different from that in §14.1.) Note that U is 1-smooth; moreover, if $Z \sim \pi_\beta$ then $\sqrt{\beta}\, Z \sim \pi$. Suppose $\mathsf{KL}(\mu_\beta \,\|\, \pi_\beta) = K_0$ and that $X_\beta \sim \mu_\beta$ is a sample from μ_β, and let $\mu := \mathrm{law}(\sqrt{\beta}\, X_\beta)$. Since the KL divergence is invariant under bijective transformations, we have $\mathsf{KL}(\mu \,\|\, \pi) = K_0$, which shows that we can simulate an initialization oracle for π given an initialization oracle for π_β. We can also simulate the local oracle for π given a local oracle for π_β, as $\nabla U(x) = \frac{1}{\sqrt{\beta}} \nabla U_\beta(x/\sqrt{\beta})$. Finally, let $\hat{\mu}$ satisfy $\sqrt{\mathsf{FI}(\hat{\mu} \,\|\, \pi)} \le \varepsilon/\sqrt{\beta}$ and write $\hat{\mu}_\beta := \mathrm{law}(\hat{X}/\sqrt{\beta})$ where $\hat{X} \sim \hat{\mu}$. A straightforward calculation shows that $\sqrt{\mathsf{FI}(\hat{\mu}_\beta \,\|\, \pi_\beta)} \le \varepsilon$. This proves the upper bound $\mathscr{C}(d, K_0, \varepsilon; \beta) \le \mathscr{C}(d, K_0, \varepsilon/\sqrt{\beta}; 1)$, and the reverse bound follows because this reduction is reversible. □

From here on, we abbreviate $\mathscr{C}(d, K_0, \varepsilon) := \mathscr{C}(d, K_0, \varepsilon; 1)$.

■ 14.3 Reduction to optimization and the first lower bound

In this section, we show a perhaps surprising equivalence between obtaining Fisher information guarantees in sampling and finding stationary points of smooth functions in optimization. The formal statement of the equivalence is as follows.

Theorem 14.3.1 (Equivalence)**.** *Let $V : \mathbb{R}^d \to \mathbb{R}$ be a 1-smooth function such that for any $\beta > 0$, the function $\exp(-\beta V)$ is integrable. Let π_β be the probability measure with density $\pi_\beta \propto \exp(-\beta V)$, where $\beta = d/\varepsilon^2$.*

1. *Suppose that $x \in \mathbb{R}^d$ is a point with $\|\nabla V(x)\| \le \varepsilon$. Then, for $\mu_\beta \coloneqq \mathsf{normal}(x, \beta^{-1} I_d)$, it holds that $\mathsf{FI}(\mu_\beta \parallel \pi_\beta) \le 10\beta d$.*
2. *Conversely, suppose that μ is such that $\mathsf{FI}(\mu \parallel \pi_\beta) \le \beta d$. Let $X \sim \mu$ be a sample. Then, $\|\nabla V(X)\| \le 3\varepsilon$ with probability at least $1/2$.*

Proof. See §14.6.1. □

Note that an oracle for βV can be simulated from an oracle for V, so that the above theorem provides an exact equivalence between a sampling problem and an optimization problem within the oracle model, up to universal constants.

As a first application of this equivalence, we observe that averaged LMC yields an nearly optimal algorithm for finding stationary points of smooth functions. We recall the LMC algorithm for sampling from a density $\pi \propto \exp(-U)$. We fix a step size $h > 0$, initialize at $X_0 \sim \mu_0$, and for $t \in [kh, (k+1)h]$, we set

$$X_t = X_{kh} - (t - kh)\,\nabla U(X_{kh}) + \sqrt{2}\,(B_t - B_{kh})\,, \tag{14.2}$$

where $(B_t)_{t \ge 0}$ is a standard Brownian motion in $\mathbb{R}^d$. Let $\mu_t \coloneqq \mathrm{law}(X_t)$ denote the law of the algorithm at time t. Then, the *averaged* LMC algorithm at iteration N outputs a sample from the law of $\bar{\mu}_{Nh} \coloneqq (Nh)^{-1} \int_0^{Nh} \mu_t \,\mathrm{d}t$. This is obtained algorithmically as follows: first, we sample a time $t \in [0, Nh]$ uniformly at random (independently of all other random variables). Let k denote the largest integer such that $kh \le t$. We then compute $X_0, X_h, X_{2h}, \ldots, X_{kh}$ using the LMC recursion, and then output X_t which is obtained via the partial LMC update (14.2).

Corollary 14.3.2 (Averaged LMC is nearly optimal for finding stationary points)**.** *Let $V : \mathbb{R}^d \to \mathbb{R}$ be 1-smooth and satisfy $V(0) - \inf V \le \Delta$. Let $\varepsilon > 0$ be such that $\Delta/\varepsilon^2 \ge 1$. Assume that for $\beta = d/\varepsilon^2$, the probability measure with density $\pi_\beta \propto \exp(-\beta V)$ is well-defined and that $\int \|\cdot\|^2 \,\mathrm{d}\pi_\beta \le \mathrm{poly}(\Delta, d, 1/\varepsilon)$. Consider running averaged LMC with step size $h = \widetilde{\Theta}(1/\beta)$, initial distribution $\mu_0 = \mathsf{normal}(0, \beta^{-1} I_d)$, and target π_β, with*

$$N \ge \widetilde{\Omega}\Big(\frac{\Delta}{\varepsilon^2}\Big) \qquad \textit{iterations}\,.$$

Then, we obtain a sample X such that with probability at least $1/2$, it holds that $\|\nabla V(X)\| \lesssim \varepsilon$.

Proof. We combine Theorem 14.3.1 with the analysis of averaged LMC in Theorem 13.4.2; see §14.6.2. □

This matches the usual $O(\Delta/\varepsilon^2)$ complexity for the standard gradient descent algorithm to find an ε-stationary point [see, e.g., Bub15; Nes18]. On its own, this observation is not terribly surprising because as $\beta \to \infty$, the LMC iteration (14.2) recovers the gradient descent algorithm. However, it is remarkable that the analysis in §13 of averaged LMC in Fisher information nearly recovers the gradient descent guarantee.

This observation also suggests that the lower bound of [Car+20], which establishes optimality of gradient descent for finding stationary points in high dimension, also implies optimality of averaged LMC in a certain regime. We obtain the following theorem.

Theorem 14.3.3 (First lower bound)**.** *Suppose that the dimension d satisfies $\widetilde{O}(K_0) \geq d \geq \widetilde{\Omega}(K_0^{2/3})$. Then, it holds that*

$$\mathscr{C}\big(d, K_0, \varepsilon = \sqrt{\beta d}; \beta\big) \gtrsim \frac{K_0}{d}\,.$$

Proof. In the lower bound of [Car+20], the authors construct a family of functions $\mathscr{F}$ such that each $f \in \mathscr{F}$ is β-smooth and satisfies $f(0) - \inf f \leq \Delta$. Moreover, any randomized algorithm which, for any $f \in \mathscr{F}$, makes queries to a local oracle for f and outputs an δ-stationary point of f with probability at least $1/2$, requires at least $\Omega(\beta\Delta/\delta^2)$ queries. The dimension of the functions in the construction is $d = \widetilde{\Theta}(\beta^2\Delta^2/\delta^4)$. Setting $\beta V = f$ and using the equivalence from Theorem 14.3.1 completes the proof. Details are given in §14.6.3. □

The lower bound of Theorem 14.3.3 is matched by averaged LMC, see (14.1). In the theorem, the restriction $d \geq \widetilde{\Omega}(K_0^{2/3})$ arises because the lower bound construction of [Car+20] for finding a ε-stationary point of a smooth function requires a large dimension $d \geq \widetilde{\Omega}(1/\varepsilon^4)$. If, as conjectured in [BM20] and in §12, the lower bound construction can be embedded in dimension $d \gtrsim \log(1/\varepsilon)$, then the restriction in Theorem 14.3.3 would instead become $d \gtrsim \log K_0$.

■ 14.4 Bump construction and the second lower bound

The main drawback of the first lower bound (Theorem 14.3.3) is that it only provides a lower bound on the Fisher information for a specific value of the target accuracy, $\varepsilon = \sqrt{\beta d}$. To complement this result, we provide the following lower bound for the query complexity of sampling to high accuracy in Fisher information; recall that it suffices to consider $\beta = 1$ by the rescaling lemma (Lemma 14.2.3).

Theorem 14.4.1 (Second lower bound)**.** *For the class of* 1*-log-smooth distributions on* $\mathbb{R}^d$*, there exist universal constants* $c, c' > 0$*, such that for all* $\varepsilon < \exp(-c'd)$*, we have*

$$\mathscr{C}(d, K_0 = 1, \varepsilon) \gtrsim \left(\frac{cd}{\log(1/\varepsilon)}\right)^{d/2} \frac{1}{\varepsilon^{2d/(d+2)}}\,. \tag{14.3}$$

Proof. Here we sketch the main ideas of the proof. We construct a family of distributions in our class which put a constant fraction of their mass on disjoint bumps. Specifically, let B_r denote the ball of radius r in $\mathbb{R}^d$, and let $\mathscr{P}_{2r,R}$ be a maximal $2r$-packing of B_{R-r}. For any $\omega \in \mathscr{P}_{2r,R}$, let $\tilde{\pi}_\omega$ denote the unnormalized density

$$\tilde{\pi}_\omega(x) \coloneqq \exp\left(r^2\phi\Big(\frac{\|x-\omega\|}{r}\Big) - \frac{1}{2}\left(\|x\| - R\right)_+^2\right) =: \exp\big(-V_\omega(x)\big)\,, \tag{14.4}$$

where $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a decreasing, twice continuously differentiable function supported on $[0,1]$ with bounded second derivative, chosen such that $\tilde{\pi}_\omega$ is 1-log-smooth. We see that the mass of the distribution π_ω will be concentrated on B_R. Moreover, by a careful choice of r we can ensure that exactly half of the mass of π_ω is in the set $\omega + B_r$.

The key idea is the following reduction: being able to sample from π_ω within small Fisher information means that we can estimate $\omega \in \mathscr{P}_{2r,R}$. To make this reduction work, note that if we make a query within $\omega + B_r$, then we can immediately identify ω. Because π_ω puts half of its mass on $\omega + B_r$ by construction, if we can sample from a distribution within total variation distance less than $1/2$ from π_ω then we will sample a point in $\omega + B_r$ with constant probability. The last ingredient is to note that sampling close to π_ω in Fisher information implies that we are close in total variation distance due to the following functional inequality (see [Gui+09]): for any probability measure μ,

$$\mathsf{TV}(\mu, \pi_\omega)^2 \le \frac{1}{4}\, C_{\mathsf{PI}}(\pi_\omega)\, \mathsf{FI}(\mu \parallel \pi_\omega)\,,$$

where $C_{\mathsf{PI}}(\pi_\omega)$ is the Poincaré constant of π_ω.

As a result, a query complexity lower bound on sampling in Fisher information directly follows from a lower bound on the query complexity of estimating ω, which by standard information-theoretic arguments takes $\Omega(|\mathcal{P}_{2r,R}|)$ queries.

Although the scheme of the argument is straightforward, the proof requires careful balancing of the parameters r, R, d and ε and some delicate calculations to satisfy all of the desired properties. The details are given in §14.7. □

The lower bound in Theorem 14.4.1 deteriorates in high dimension; note that due to the restriction $\varepsilon \le \exp(-c'd)$, the first factor in (14.3) is exponentially small in d. However, we can remedy this by noting that a d-dimensional construction can be embedded into $\mathbb{R}^{d'}$ for any $d' \ge d$, and hence

$$\mathscr{C}(d, K_0 = 1, \varepsilon) \gtrsim \max_{d_\star \le d} \left[\left(\frac{c d_\star}{\log(1/\varepsilon)} \right)^{d_\star/2} \varepsilon^{4/(d_\star+2)} \right] \frac{1}{\varepsilon^2} .$$

By optimizing over $d_\star$, we show (§14.7.8) that if $\varepsilon \le 1/C$, then

$$\mathscr{C}(d, 1, \varepsilon) \gtrsim \begin{cases} \dfrac{1}{\varepsilon^{2d/(d+2)} \exp(C\sqrt{\log(1/\varepsilon)\log\log(1/\varepsilon)})} , & \text{if } d \lesssim \sqrt{\dfrac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}} , \\ \dfrac{1}{\varepsilon^2 \exp(C\sqrt{\log(1/\varepsilon)\log\log(1/\varepsilon)})} , & \text{if } d \gtrsim \sqrt{\dfrac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}} , \end{cases}$$

$$= \frac{1}{\varepsilon^{\min\{2d/(d+2),2\} - o(1)}} , \qquad \text{for all } d \ge 1 ,$$

as $\varepsilon \to 0$, where $C > 0$ is universal. Noting $2d/(d+2) < 2$, this yields the simplified bound in Theorem 14.1.3.

For $d = 1$, Theorem 14.4.1 reads $\mathscr{C}(1, 1, \varepsilon) \gtrsim 1/(\varepsilon^{2/3}\sqrt{\log(1/\varepsilon)})$. However, for the one-dimensional case we can in fact obtain better bounds on the Poincaré constants of the measures in our lower bound construction, leading to an improvement of the exponent from 2/3 to 1. This result is stated below.

Theorem 14.4.2 (Second lower bound, univariate case)**.** *For the class of* 1*-log-smooth distributions on* $\mathbb{R}$*, there exists a universal constant* $c > 0$*, such that for all* $\varepsilon < c$*, we have*

$$\mathscr{C}(d = 1, K_0 = 1, \varepsilon) \gtrsim \frac{1}{\varepsilon\sqrt{\log(1/\varepsilon)}} .$$

Proof. See §14.7.9. □

The univariate setting also provides a convenient setting in order to compare our lower bounds with algorithms such as rejection sampling, so we include a detailed discussion in §14.8. We highlight a few interesting conclusions of the discussion here.

- Although rejection sampling can indeed obtain Fisher information guarantees with complexity $O(\log(1/\varepsilon))$ (Proposition 14.8.1), this does not contradict our lower bounds because rejection sampling cannot be directly implemented within our oracle model. Instead of an initialization μ_0 satisfying $\mathsf{KL}(\mu_0 \,\|\, \pi) \le K_0$, rejection sampling requires the stronger assumption

$\max\{\sup \ln(\mu_0/\pi), \sup \ln(\pi/\mu_0)\} \le M_0$. Under this stronger initialization oracle, the complexity guarantee for rejection sampling is $O(\exp(3M_0)\log(1/\varepsilon))$.

- In the model with the stronger initialization oracle (i.e., bounded M_0), any algorithm which has polylog$(1/\varepsilon)$ dependence on the accuracy ε necessarily incurs exponential dependence on M_0 (Corollary 14.8.3). This demonstrates a fundamental trade-off between high accuracy (e.g., rejection sampling) and polynomial dependence on M_0 (e.g., averaged LMC).
- The initialization oracle with bounded M_0 is strictly stronger than the one with bounded K_0. In other words, sampling is strictly easier in the presence of an initialization with bounded density ratio to the target (i.e., a *warm start*) than an initialization with bounded KL divergence. This is consistent with intuition from prior work on the complexity of the Metropolis-adjusted Langevin algorithm [see §5, §6, and LST21a; WSC22].
- The effective radius R of our lower bound construction scales with $1/\varepsilon$. This is in fact necessary: if R is fixed then there is an algorithm with $O(\log(1/\varepsilon))$ complexity (Proposition 14.8.4).

14.5 Separation between log-concave and non-log-concave sampling

We show that $O(\log \frac{1}{\varepsilon})$ Fisher information query complexity is attainable for log-concave densities, by giving a generic post-processing method to turn χ^2-error guarantees into Fisher information guarantees.

14.5.1 Post-processing lemma

Let Q_t denote heat flow for time t (i.e., convolution with a Gaussian of variance t). We aim to bound $\mathsf{FI}(\mu Q_t \,\|\, \pi)$, where π is the distribution that we wish to sample from, and μ is the output of a sampling algorithm with chi-squared error guarantees.

Lemma 14.5.1 (Fisher information guarantee from a chi-squared guarantee)**.** *Suppose that μ and π are two probability measures on $\mathbb{R}^d$, that π is β-log-smooth, and that $\chi^2(\mu\,\|\,\pi) \le \varepsilon_\chi^2 \le 1$. Then, if $t \lesssim 1/\beta$ for a small enough implied constant, it holds that*

$$\mathsf{FI}(\mu Q_t \,\|\, \pi) \lesssim \frac{\varepsilon_\chi\,(d + \log(1/\varepsilon_\chi))}{t} + \beta^2 dt\,.$$

To prove Lemma 14.5.1, we start with

$$\begin{aligned}\mathsf{FI}(\mu Q_t \parallel \pi) &\coloneqq \int \|\nabla \ln(\mu Q_t)(x) - \nabla \ln \pi(x)\|^2 \, \mu Q_t(\mathrm{d}x) \\ &\le 2\,\mathsf{FI}(\mu Q_t \parallel \pi Q_t) + 2\int_{\mathbb{R}^d} \|\nabla \log(\pi Q_t)(x) - \nabla \log \pi(x)\|^2 \, \mu Q_t(\mathrm{d}x)\,.\end{aligned} \tag{14.5}$$

For the first term in (14.5), we use the following lemma on error in the score function (gradient of the log-density).

Lemma 14.5.2 (Score error under heat flow, [LLT23, Lemma 6.2])**.** *Let μ and π be probability measures on $\mathbb{R}^d$, and let Q_t denote the heat semigroup at time t. In addition, we assume that $\chi^2(\mu \parallel \pi) \le \varepsilon_\chi^2 \le 1$. Then,*

$$\mathsf{FI}(\mu Q_t \parallel \pi Q_t) = \int_{\mathbb{R}^d} \|\nabla \ln(\mu Q_t)(x) - \nabla \ln(\pi Q_t)(x)\|^2 \, \mu Q_t(\mathrm{d}x) \lesssim \frac{\varepsilon_\chi \left(d + \ln \frac{1}{\varepsilon_\chi}\right)}{t}\,.$$

For the second term in (14.5), we use the following score perturbation lemma.

Lemma 14.5.3 ([LLT22, Lemma C.11])**.** *Suppose that $\pi \propto \exp(-V)$ is a probability density on $\mathbb{R}^d$, where V is β-smooth. Then for $\beta \le \frac{1}{2t}$,*

$$\left\|\nabla \ln \frac{\pi(x)}{(\pi Q_t)(x)}\right\| \le 6\beta d^{1/2} t^{1/2} + 2\beta t \, \|\nabla V(x)\|\,.$$

We are now ready to prove Lemma 14.5.1.

Proof of Lemma 14.5.1. For the second term in (14.5), Lemma 14.5.3 yields

$$\mathbb{E}_{\mu Q_t}[\|\nabla \ln(\pi Q_t) - \nabla \ln \pi\|^2] \lesssim \beta^2 dt + \beta^2 t^2 \, \mathbb{E}_{\mu Q_t}[\|\nabla V\|^2]\,.$$

On the other hand, Lemma 14.6.1 below yields

$$\mathbb{E}_{\mu Q_t}[\|\nabla V\|^2] \lesssim \mathsf{FI}(\mu Q_t \parallel \pi) + \beta d\,.$$

Hence, from (14.5) and Lemma 14.5.2,

$$\begin{aligned}\mathsf{FI}(\mu Q_t \parallel \pi) &\lesssim \mathsf{FI}(\mu Q_t \parallel \pi Q_t) + \mathbb{E}_{\mu Q_t}[\|\nabla \ln(\pi Q_t) - \nabla \ln \pi\|^2] \\ &\lesssim \frac{\varepsilon_\chi \left(d + \log(1/\varepsilon_\chi)\right)}{t} + \beta^2 dt + \beta^2 t^2 \, \mathsf{FI}(\mu Q_t \parallel \pi)\,.\end{aligned}$$

If $t \lesssim 1/\beta$ for a small enough implied constant, it implies

$$\mathsf{FI}(\mu Q_t \parallel \pi) \lesssim \frac{\varepsilon_\chi \left(d + \log(1/\varepsilon_\chi)\right)}{t} + \beta^2 dt$$

as desired. □

14.5.2 High-accuracy Fisher information guarantees for log-concave targets

We now apply the post-processing lemma (Lemma 14.5.1). We recall the following high-accuracy guarantee for sampling from log-concave targets in chi-squared divergence, based on the proximal sampler.

Theorem 14.5.4 (Corollary 4.3.7)**.** *Suppose that the target distribution $\pi \propto \exp(-V)$ is β-log-smooth and satisfies a Poincaré inequality with constant C_{PI}. Then, the proximal sampler, with rejection sampling implementation of the restricted Gaussian oracle and initialized at μ_0, outputs a sample from a measure μ with $\chi^2(\mu \,\|\, \pi) \le \varepsilon_\chi^2$ using N queries to π in expectation, where N satisfies*

$$N \le \widetilde{O}\Big(C_{\mathsf{PI}}\beta d \left(\log(1 + \chi^2(\mu_0 \,\|\, \pi)) \vee \log \frac{1}{\varepsilon_\chi}\right)\Big) .$$

We now briefly justify why this morally leads to an $O(\log(1/\varepsilon))$ complexity guarantee in Fisher information, omitting details for brevity. Assume that $\beta = 1$ and that π is log-concave. If we set $t \asymp \varepsilon^2/d$ in Lemma 14.5.1, then we can ensure that $\mathsf{FI}(\mu Q_t \,\|\, \pi) \le \varepsilon^2$, where μ is the output of the proximal sampler, provided that $\varepsilon_\chi \le \widetilde{O}(\varepsilon^4/d^2)$. Applying Theorem 14.5.4, this can be achieved using

$$N = \widetilde{O}\Big(C_{\mathsf{PI}} d \left(\log(1 + \chi^2(\mu_0 \,\|\, \pi)) \vee \log \frac{\sqrt{d}}{\varepsilon}\right)\Big)$$

queries in expectation. Let us give crude bounds for these terms. First, let $\mathfrak{m}_2^2 \coloneqq \mathbb{E}_\pi[\|\cdot\|^2]$ denote the second moment of π. Then, we know that the Poincaré constant of π is bounded because π is log-concave, and in fact $C_{\mathsf{PI}} \lesssim \mathfrak{m}_2^2$ [see, e.g., Bob99]. Also, if $\nabla V(0) = 0$, then we can initialize with $\log(1 + \chi^2(\mu_0 \,\|\, \pi)) \le \widetilde{O}(d)$ (see §3.6.6). Putting this together, we see that $N = \mathrm{poly}(d, \mathfrak{m}_2, \log(\sqrt{d}/\varepsilon))$ queries suffice in expectation in order to obtain the guarantee $\sqrt{\mathsf{FI}(\mu Q_t \,\|\, \pi)} \le \varepsilon$. This is in contrast with our lower bound in Theorem 14.4.1, which shows that $\mathrm{poly}(1/\varepsilon)$ queries are necessary to obtain Fisher information guarantees for *non-log-concave* targets, thereby establishing a separation between log-concave and non-log-concave sampling in this context.

The astute reader will observe that there are some holes in this argument when comparing the lower and upper bounds. Namely, the upper bound uses further properties about the target distribution (e.g., $\nabla V(0) = 0$) and does not strictly hold in the oracle model that we describe in §14.2; the upper bound is in terms of the expected number of queries made, because the number of queries made by the algorithm is random; and the upper bound depends on other parameters such as $\mathfrak{m}_2$ which do not appear in the lower bound. In particular, the third point requires

some consideration because in our lower bound construction for Theorem 14.4.1, the effective radius R of the distributions depends on $1/\varepsilon$. We claim, however, that if we set $d, R = \mathrm{polylog}(1/\varepsilon)$, then the upper bound for log-concave targets is $\mathrm{polylog}(1/\varepsilon)$ (with the caveats just discussed) and the lower bound for non-log-concave targets is $\mathrm{poly}(1/\varepsilon)$. As this is not the focus of our work, we do not attempt to make this reasoning more rigorous; rather, we leave it as the sketch of an argument showing that non-log-concave sampling is fundamentally harder than log-concave sampling. We also note that our argument in fact shows that $\mathrm{polylog}(1/\varepsilon)$ query complexity is possible for distributions satisfying a Poincaré inequality, which form a strict superclass of log-concave distributions.

14.6 Proofs for the first lower bound

14.6.1 Proof of the equivalence

In order to prove the equivalence in Theorem 14.3.1, we recall the following useful lemma (see Lemma 3.6.3).

Lemma 14.6.1. *Let $\pi \propto \exp(-V)$ be a β-log-smooth density on $\mathbb{R}^d$. Then, for any probability measure μ,*

$$\mathbb{E}_\mu[\|\nabla V\|^2] \le \mathsf{FI}(\mu \parallel \pi) + 2\beta d\,.$$

With the lemma in hand, we are ready to prove Theorem 14.3.1.

Proof of Theorem 14.3.1. 1. We can explicitly compute

$$\begin{aligned}
\mathsf{FI}(\mu_\beta \parallel \pi_\beta) &= \int \|\nabla \ln \mu_\beta - \nabla \ln \pi_\beta\|^2 \,\mathrm{d}\mu_\beta \\
&= \int \|\beta\,(z-x) - \beta\,\nabla V(z)\|^2 \,\mathrm{d}\mu_\beta(z) \\
&\le 2\beta^2 \int \|z-x\|^2 \,\mathrm{d}\mu_\beta(z) + 2\beta^2 \int \|\nabla V(z)\|^2 \,\mathrm{d}\mu_\beta(z) \\
&\le 2\beta^2 \int \|z-x\|^2 \,\mathrm{d}\mu_\beta(z) + 4\beta^2 \int \{\|z-x\|^2 + \|\nabla V(x)\|^2\} \,\mathrm{d}\mu_\beta(z) \\
&\le 6\beta^2 \int \|z-x\|^2 \,\mathrm{d}\mu_\beta(z) + 4\beta^2 \underbrace{\|\nabla V(x)\|^2}_{\le \varepsilon^2}\,,
\end{aligned}$$

where we used smoothness of V. Also, $\int \|z-x\|^2 \,\mathrm{d}\mu_\beta(z) = d/\beta$. Hence,

$$\mathsf{FI}(\mu_\beta \parallel \pi_\beta) \le 6\beta d + 4\beta^2\varepsilon^2 = 10\beta d\,,$$

provided $\beta = d/\varepsilon^2$.

2. Conversely, since $\nabla \ln(1/\pi_\beta) = \beta \nabla V$ is β-Lipschitz, from Lemma 14.6.1

$$\mathbb{E}_\mu[\|\nabla V\|^2] = \frac{1}{\beta^2}\,\mathbb{E}_\mu[\|\nabla(\beta V)\|^2] \le \frac{1}{\beta^2}\,\{\mathsf{FI}(\mu \parallel \pi_\beta) + 2\beta d\} \le \frac{3d}{\beta}\,.$$

If we take $\beta = d/\varepsilon^2$, then $\mathbb{E}_\mu[\|\nabla V\|^2] \le 3\varepsilon^2$. By Chebyshev's inequality, $X \sim \mu$ satisfies $\|\nabla V(X)\| \le \sqrt{6}\,\varepsilon$ with probability at least $1/2$.

□

■ 14.6.2 Proof of the averaged LMC guarantee

In order to apply Theorem 13.4.2, we need a bound on the KL divergence at initialization. Such bounds are standard; however, since §3.6.6 assumes that we start at a stationary point of V (contrary to the present setting), we present an adapted version.

Lemma 14.6.2 (KL divergence at initialization)**.** *Suppose that $U : \mathbb{R}^d \to \mathbb{R}$ is a function such that $U(0) - \inf U \le \Delta$, ∇U is β-Lipschitz, and $\mathfrak{m} := \int \|\cdot\| \,\mathrm{d}\pi < \infty$ where $\pi \propto \exp(-U)$. Then, for $\mu_0 = \mathsf{normal}(0, \beta^{-1} I_d)$, we have the bound*

$$\mathsf{KL}(\mu_0 \parallel \pi) \lesssim \Delta + d\,\big(1 \vee \ln(\beta \mathfrak{m}^2)\big)\,.$$

Proof. Write

$$\frac{\mu_0}{\pi} = \exp\Big(U - \frac{\beta}{2}\,\|\cdot\|^2\Big)\,\frac{\int \exp(-U)}{\int \exp(-U - \delta\,\|\cdot\|^2)}\,\frac{\int \exp(-U - \delta\,\|\cdot\|^2)}{(2\pi/\beta)^{d/2}}\,,$$

where $\delta > 0$ is chosen later.

For the first term, by smoothness and Young's inequality,

$$U(x) - \frac{\beta}{2}\,\|x\|^2 \le U(0) + \langle \nabla U(0), x\rangle \le U(0) + \frac{\|\nabla U(0)\|^2}{2\beta} + \frac{\beta\,\|x\|^2}{2}\,.$$

Plugging in $x = -\frac{1}{\beta}\,\nabla U(0)$,

$$U\Big(-\frac{1}{\beta}\,\nabla U(0)\Big) - U(0) \le -\frac{1}{2\beta}\,\|\nabla U(0)\|^2$$

or

$$\|\nabla U(0)\|^2 \le 2\beta\,\Big(U(0) - U\big(-\frac{1}{\beta}\,\nabla U(0)\big)\Big) \le 2\beta\,\big(U(0) - \inf U\big) \le 2\beta\Delta\,.$$

Hence, for any x,

$$U(x) - \frac{\beta}{2}\|x\|^2 \le U(0) + \Delta + \frac{\beta\|x\|^2}{2}\,.$$

For the second term, Markov's inequality yields

$$\begin{aligned}\frac{\int \exp(-U - \delta\|\cdot\|^2)}{\int \exp(-U)} = \int \exp(-\delta\|\cdot\|^2)\,\mathrm{d}\pi &\ge \exp(-4\delta\mathfrak{m}^2)\,\pi\{\|\cdot\| \le 2\mathfrak{m}\} \\ &\ge \frac{1}{2}\exp(-4\delta\mathfrak{m}^2)\,.\end{aligned}$$

For the third term,

$$\frac{\int \exp(-U - \delta\|\cdot\|^2)}{(2\pi/\beta)^{d/2}} \le \frac{\exp(-\inf U)\int \exp(-\delta\|\cdot\|^2)}{(2\pi/\beta)^{d/2}} = \exp(-\inf U)\left(\frac{\beta}{2\delta}\right)^{d/2}.$$

Combining these bounds,

$$\begin{aligned}\mathsf{KL}(\mu_0 \,\|\, \pi) &= \mathbb{E}_{\mu_0} \ln \frac{\mu_0}{\pi} \\ &\le U(0) - \inf U + \Delta + \frac{\beta}{2}\,\mathbb{E}_{\mu_0}[\|\cdot\|^2] + \ln 2 + 4\delta\mathfrak{m}^2 + \frac{d}{2}\ln\frac{\beta}{2\delta}\,.\end{aligned}$$

Now we set $\delta = \frac{1}{4\mathfrak{m}^2}$ to obtain

$$\mathsf{KL}(\mu_0 \,\|\, \pi) \lesssim \Delta + d\left(1 \vee \ln(\beta\mathfrak{m}^2)\right)$$

as claimed. □

Proof of Corollary 14.3.2. Let V be 1-smooth and apply the above lemma to $U = \beta V$, which is β-smooth and satisfies $U(0) - \inf U \le \beta\Delta$, so that

$$K_0 \coloneqq \mathsf{KL}(\mu_0 \,\|\, \pi_\beta) \lesssim \beta\Delta + d\left(1 \vee \ln(\beta\,\mathbb{E}_{\pi_\beta}[\|\cdot\|^2])\right) = \widetilde{O}(\beta\Delta + d)\,. \tag{14.6}$$

The main result of §13 says that after N steps of averaged LMC, with an appropriate choice of step size h, we output a sample from μ satisfying

$$\mathsf{FI}(\mu \,\|\, \pi_\beta) \lesssim \frac{\beta\sqrt{K_0 d}}{\sqrt{N}}\,.$$

To apply this result, we find N such that this inequality implies $\mathsf{FI}(\mu \,\|\, \pi_\beta) \le \beta d$, where we recall that $\beta = d/\varepsilon^2$; this requires $N \gtrsim K_0/d$. From (14.6), it suffices to have $N \ge \widetilde{\Omega}(\Delta/\varepsilon^2)$, provided $\Delta/\varepsilon^2 \ge 1$. The result for finding stationary points via averaged LMC now follows from the equivalence in Theorem 14.3.1. □

■ 14.6.3 Proof of the first lower bound

Proof of Theorem 14.3.3. Let $\mathscr{F}$ be the family constructed in the lower bound of [Car+20], and let $f \in \mathscr{F}$. Recall that $\mathscr{F}$ satisfies the following properties: each $f \in \mathscr{F}$ is β-smooth with $f(0) - \inf f \le \Delta$; any randomized algorithm which, for any $f \in \mathscr{F}$, makes queries to a local oracle for f and outputs an δ-stationary point of f with probability at least $1/2$, requires at least $\Omega(\beta\Delta/\delta^2)$ queries.

We set $\delta \coloneqq 4\sqrt{\beta d}$. From the Fisher information lemma (Lemma 14.6.1), if we can obtain a sample from a measure μ such that for $\pi_f \propto \exp(-f)$, it holds that $\mathsf{FI}(\mu \,\|\, \pi_f) \le \beta d$, then a sample from μ is a δ-stationary point of f with probability at least $1/2$.

We set the initialization oracle to simply output samples from the distribution $\mu_0 \coloneqq \mathsf{normal}(0, \beta^{-1} I_d)$. We need to compute the value of $K_0 \coloneqq \sup_{f \in \mathscr{F}} \mathsf{KL}(\mu_0 \| \pi_f)$, and for this we use Lemma 14.6.2. First, we must bound the second moment $\mathbb{E}_{\pi_f}[\|\cdot\|^2]$. Since we only care about polynomial dependencies for this calculation, let poly denote any positive quantity for which both the quantity and its inverse are bounded above by polynomials in β, Δ, d, and $1/\delta$. Inspecting the proof of [Car+20] and using the notation therein, each $f \in \mathscr{F}$ is of the form

$$f(x) = \mathsf{poly} \cdot \tilde{f}_{T,U}\big(\rho(x/\mathsf{poly})\big) + \frac{1}{2\tau^2}\|x\|^2, \qquad \text{where } \tau = \mathsf{poly}.$$

Also, $\tilde{f}_{T,U}$ is bounded; thus, $\pi_f \propto \exp(-f)$ is well-defined. To bound the second moment of π_f, we can use the Donsker–Varadhan variational principle to write, for any $\lambda > 0$,

$$\mathbb{E}_{\pi_f}[\|\cdot\|^2] \le \frac{1}{\lambda}\left\{\mathsf{KL}(\pi_f \parallel \nu) + \ln \mathbb{E}_\nu \exp(\lambda \|\cdot\|^2)\right\},$$

where $\nu \coloneqq \mathsf{normal}(0, \tau I_d)$. By choosing $\lambda = 1/\mathsf{poly}$, we can ensure that

$$\ln \mathbb{E}_\nu \exp(\lambda \|\cdot\|^2) \le 1.$$

Next, since ν satisfies a log-Sobolev inequality with constant poly, we obtain

$$\mathbb{E}_{\pi_f}[\|\cdot\|^2] \le \mathsf{poly} \cdot \big(1 + \mathsf{FI}(\pi_f \parallel \nu)\big).$$

The Fisher information is computed to be

$$\mathsf{FI}(\pi_f \parallel \nu) = \mathsf{poly} \cdot \mathbb{E}_{\pi_f}\Big[\big\|\nabla\big(\tilde{f}_{T,U}\big(\rho(\cdot/\mathsf{poly})\big)\big)\big\|^2\Big].$$

Here, $\tilde{f}_{T,U} : \mathbb{R}^d \to \mathbb{R}$ and $\rho : \mathbb{R}^d \to \mathbb{R}^d$ are poly-Lipschitz, and hence

$$\big\|\nabla\big(\tilde{f}_{T,U}\big(\rho(\cdot/\mathsf{poly})\big)\big)\big\| \le \mathsf{poly}.$$

Putting everything together, we deduce that $\mathbb{E}_{\pi_f}[\|\cdot\|^2] \le \mathsf{poly}$.

From Lemma 14.6.2, we can take $K_0 \lesssim \Delta + \widetilde{O}(d)$. If $K_0 \ge \widetilde{\Omega}(d)$, then this shows that $\Delta \gtrsim K_0$. From the lower bound of [Car+20], we obtain

$$\mathscr{C}(d, K_0, \sqrt{\beta d}; \beta) \gtrsim \frac{\beta\Delta}{\delta^2} \gtrsim \frac{\beta K_0}{\beta d} = \frac{K_0}{d}\,.$$

Finally, in order for the construction of [Car+20] to be valid, the functions must be defined in dimension $d \ge \widetilde{\Omega}((K_0/d)^2)$, which holds provided $d \ge \widetilde{\Omega}(K_0^{2/3})$. □

14.7 Proofs for the second lower bound

14.7.1 Proof of Theorem 14.4.1

Throughout the proof, we will often work with unnormalized densities. For a distribution π, which we identify with its density, we denote by $\tilde{\pi}$ an unnormalized density, where $\pi = \frac{\tilde{\pi}}{Z}$ and the normalizing constant is given by $Z \coloneqq \int_{\mathbb{R}^d} \tilde{\pi}(x)\,\mathrm{d}x$.

We reduce the task of estimating the distribution from queries to the task of sampling. Namely, we construct a set of distributions π that are 1-log-smooth, such that if we can sample well from π in Fisher information, then we can estimate its identity. Let B_r denote the ball of radius r in $\mathbb{R}^d$; let $V_d \coloneqq \pi^{d/2}/\Gamma(\frac{d}{2}+1)$ denote the volume of B_1, and let $A_{d-1} = dV_d$ denote the surface area of ∂B_1. Let $\mathscr{P}_{2r,R}$ be a maximal $2r$-packing of B_{R-r}, for some $R \ge r$ to be specified. By standard volume arguments [see, e.g., Ver18, §4.2], we know that $|\mathscr{P}_{2r,R}| \ge \left(\frac{R-r}{2r}\right)^d$. For any $\omega \in \mathscr{P}_{2r,R}$, let $\tilde{\pi}_\omega$ denote the unnormalized density

$$\tilde{\pi}_\omega(x) \coloneqq \exp\left(r^2\phi\Big(\frac{\|x-\omega\|}{r}\Big) - \frac{1}{2}\left(\|x\| - R\right)_+^2\right) =: \exp\big(-V_\omega(x)\big)\,, \tag{14.7}$$

where $(x)_+ \coloneqq \max(0, x)$, and $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a bump function with the following properties[1]:

(ϕ**.1**) ϕ is continuous, decreasing, supported on $[0,1]$, and twice continuously differentiable on the open interval $(0,\infty)$.

(ϕ**.2**) $\phi(x) = \phi(0) - \frac{1}{2}x^2$ for all $x \in [0,\alpha]$ for some $\alpha > 0$.

(ϕ**.3**) $\sup_{x>0} |\phi''(x)| \le 1$.

[1] One such function is $\phi(x) = \begin{cases} \frac{11}{64} - \frac{1}{2}x^2\,, & \text{for } x \in [0, 1/4]\,, \\ \frac{1}{27}\left(4 + 8x - 48x^2 + 56x^3 - 20x^4\right), & \text{for } x \in [1/4, 1]\,, \\ 0\,, & \text{otherwise}\,. \end{cases}$

The above implies that on $\mathbb{R}^d$, $x \mapsto \phi(\|x\|)$ is 1-smooth (see Lemma 14.7.10), and hence $\tilde{\pi}_\omega$ is 1-log-smooth. For a measurable set A, we will write $\tilde{\pi}_\omega(A) \coloneqq \int_A \tilde{\pi}_\omega(x)\,\mathrm{d}x$ and we let $Z_\omega \coloneqq \tilde{\pi}_\omega(\mathbb{R}^d)$ denote the normalizing constant for $\tilde{\pi}_\omega$.

We also define the null probability measure π_{init} to have unnormalized density

$$\tilde{\pi}_{\mathsf{init}}(x) \coloneqq \exp\Big(-\frac{1}{2}\left(\|x\| - R\right)_+^2\Big)\,,$$

with normalizing constant $Z_{\mathsf{init}} \coloneqq \tilde{\pi}_{\mathsf{init}}(\mathbb{R}^d)$.

The distribution π_ω is the combination of a flat, uniform distribution on B_R, fast decaying tails outside of B_R, and a bump of radius r around the point $\omega \in \mathscr{P}_{2r,R}$. The following lemma summarizes the properties that we need for the lower bound construction. Together, Properties **(P.1)** and **(P.2)** imply that if an algorithm outputs a sample X from a distribution which is close in Fisher information to π_ω, then X is likely to lie in the set $\omega + B_r$. Hence, an algorithm for sampling from π_ω can be used to *estimate* ω. Property **(P.4)** is then used to prove a lower bound on the number of queries to solve the estimation task. Finally, Property **(P.3)** is needed in order to ensure that there is a valid initialization oracle with $K_0 = 1$.

Lemma 14.7.1 (Lower bound construction)**.** *There exist universal constants $c_\varepsilon, c > 0$ such that for every $d \geq 1$ and $\varepsilon \leq \exp(-c_\varepsilon d)$ we can choose r, R such that the following properties hold.*

(P.1) *(most of the mass lies in the bump) For any $\omega \in \mathscr{P}_{2r,R}$,*

$$\pi_\omega(\omega + B_r) = \frac{1}{2}\,.$$

(P.2) *(FI guarantees imply TV guarantees) For any $\omega \in \mathcal{P}_{2r,R}$ and any probability measure μ,*

$$\sqrt{\mathsf{FI}(\mu \parallel \pi_\omega)} \leq \varepsilon \implies \mathsf{TV}(\mu, \pi_\omega) \leq \frac{1}{3}\,.$$

(P.3) *(initial KL divergence) There exists a probability measure π_{init} that satisfies*

$$\max_{\omega \in \mathscr{P}_{2r,R}} \mathsf{KL}(\pi_{\mathsf{init}} \parallel \pi_\omega) \leq \log 2\,.$$

(P.4) *(lower bound on the packing number) It holds that*

$$|\mathscr{P}_{2r,R}| \geq \Big(\frac{cd}{\log(1/\varepsilon)}\Big)^{d/2} \frac{1}{\varepsilon^{2d/(d+2)}}\,.$$

Proof. First, Property (**P.1**) holds by the definition of r and R, see (14.15) and Lemma 14.7.5. We prove Property (**P.2**) in §14.7.4, Property (**P.3**) in §14.7.5, and Property (**P.4**) in §14.7.6. □

Remark 14.7.2. *In order for the bound in Property (**P.4**) to be non-trivial, i.e.,* $|\mathscr{P}_{2r,R}| \gtrsim 1$, *we require* $\varepsilon^{-2d/(d+2)} \gtrsim (\sqrt{\log(1/\varepsilon)/(cd)})^d$, *i.e.,*

$$\frac{2d}{d+2}\log\frac{1}{\varepsilon} \overset{!}{\geq} \frac{d}{2}\log\log\frac{1}{\varepsilon} - \frac{d}{2}\log d + \Omega(d)\,.$$

Let γ *be such that* $\log(1/\varepsilon) = \gamma d$. *Substituting this in, we require*

$$\frac{2\gamma d^2}{d+2} \overset{!}{\geq} \frac{d}{2}\log\gamma + \Omega(d)\,.$$

This holds as long as γ *is larger than a universal constant, which is equivalent to* $\varepsilon \leq \exp(-c_\varepsilon d)$ *for a sufficiently large absolute constant* $c_\varepsilon > 0$.

Using the lemma, we can now apply a standard information theoretic argument. We recall the statement of Fano's inequality, see [CT06, §2] for background.

Theorem 14.7.3 (Fano's inequality)**.** *Let* $\omega \sim \mathsf{uniform}(\mathcal{X})$, *where* $\mathcal{X}$ *is a finite set. Then, for any estimator* $\widehat{\omega}$ *which is measurable w.r.t. the data* ξ, *it holds that*

$$\mathbb{P}\{\widehat{\omega} \neq \omega\} \geq 1 - \frac{I(\xi;\omega) + \ln 2}{\ln|\mathcal{X}|}\,,$$

where I *denotes the mutual information.*

With this theorem in hand, we are ready to prove Theorem 14.4.1.

Proof of Theorem 14.4.1. Let $\omega \sim \mathsf{uniform}(\mathscr{P}_{2r,R})$ and consider the task of estimating ω with randomized algorithms that have query access to π_ω. We first show that a sampling algorithm can solve this estimation task. Suppose that there is an algorithm that works under the oracle model specified in §14.2, with initialization oracle outputting samples from $\mu_0 = \pi_{\mathsf{init}}$ given in Property (**P.3**), which guarantees that $\mathsf{KL}(\mu_0 \,\|\, \pi_\omega) \leq \log 2$. For any $\omega \in \mathscr{P}_{2r,R}$ and target π_ω, the algorithm makes at most N queries to the local oracle, and outputs a sample from the measure μ_N with $\sqrt{\mathsf{FI}(\mu_N \,\|\, \pi_\omega)} \leq \varepsilon$. We can then estimate ω as follows: let $X \sim \mu_N$, and set

$$\widehat{\omega} := \operatorname*{arg\,min}_{\omega \in \mathscr{P}_{2r,R}} \|X - \omega\|\,.$$

Because the initialization oracle μ_0 is independent of the choice of ω, the estimator $\widehat{\omega}$ is the output of a randomized algorithm that only uses the query information to π_ω to estimate ω.

The probability that $\widehat{\omega}$ succeeds can be calculated as follows. By Property (**P.2**), we have $\mathsf{TV}(\mu_N, \pi_\omega) \le 1/3$. Let $X \sim \mu_N$; then,

$$\mathbb{P}\{X \in \omega + B_r\} = \mu_N(\omega + B_r) \ge \pi_\omega(\omega + B_r) - \mathsf{TV}(\mu_N, \pi_\omega) \ge \frac{1}{2} - \frac{1}{3} = \frac{1}{6}\,,$$

where we used Property (**P.1**). Hence we see that

$$\mathbb{P}\{\widehat{\omega} = \omega\} \ge \frac{1}{6}\,. \tag{14.8}$$

Now we prove a lower bound for the estimation task for any algorithm that succeeds with probability at least $\frac{1}{6}$. Let $x_1, \ldots, x_N$ denote the query points made by the algorithm. We first prove a lower bound for deterministic algorithms, where each query point x_i is a deterministic function of the previous queries and oracle outputs $(x_{i'}, V_\omega(x_{i'}), \nabla V_\omega(x_{i'}) : i' = 1, \ldots, i-1)$. Since the initialization oracle is independent of ω, the data available to the algorithm is

$$\xi_N \coloneqq \big(x_i, V_\omega(x_i), \nabla V_\omega(x_i) : i = 1, \ldots, N\big)\,.$$

We assume that the algorithm has made at most $N \le M/2$ queries where $M \coloneqq |\mathscr{P}_{2r,R}|$ (otherwise, $N \ge M/2$ and this is our desired lower bound).

Applying Fano's inequality (Theorem 14.7.3):, we therefore have

$$\mathbb{P}\{\widehat{\omega} \ne \omega\} \ge 1 - \frac{I(\xi_N; \omega) + \ln 2}{\ln M}. \tag{14.9}$$

Applying the chain rule for the mutual information,

$$I(\xi_N; \omega) \le \sum_{i=1}^{N} I\big(x_i, V_\omega(x_i), \nabla V_\omega(x_i);\ \omega \bigm| \xi_{i-1}\big)\,.$$

Given ξ_{i-1}, the query point x_i is deterministic. We can bound the mutual information via the conditional entropy,

$$I\big(x_i, V_\omega(x_i), \nabla V_\omega(x_i);\ \omega \bigm| \xi_{i-1}\big) \le H\big(V_\omega(x_i), \nabla V_\omega(x_i) \bigm| \xi_{i-1}\big)\,.$$

If one of the query points $x_1, \ldots, x_{i-1}$ landed in the ball $\omega + B_r$, then ω is fully known and the conditional entropy is zero. Otherwise, given the history ξ_{i-1}, the random variable ω is uniformly distributed on the set

$$\mathscr{P}_{2r,R}(i) \coloneqq \{\omega' \in \mathscr{P}_{2r,R} \mid x_{i'} \notin \omega' + B_r \text{ for } i = 1, \ldots, i-1\}\,.$$

If x_i does not belong to $\omega' + B_r$ for some $\omega' \in \mathscr{P}_{2r,R}(i)$, then the query point is useless and the conditional entropy is again zero. Otherwise, conditionally on ξ_{i-1}, the tuple $(V_\omega(x_i), \nabla V_\omega(x_i))$ can take on two possible values with probability $1/|\mathscr{P}_{2r,R}(i)|$ and $1 - 1/|\mathscr{P}_{2r,R}(i)|$ respectively, depending on whether or not $x_i \in \omega + B_r$. The conditional entropy is thus bounded by

$$H\big(V_\omega(x_i), \nabla V_\omega(x_i) \;\big|\; \xi_{i-1}\big) \le h\Big(\frac{1}{|\mathscr{P}_{2r,R}(i)|}\Big) \le h\big(\frac{2}{M}\big)\,,$$

where $h(p) \coloneqq p \ln \frac{1}{p} + (1-p) \ln \frac{1}{1-p}$ is the binary entropy function. Assuming that $M \ge 4$ (which can be ensured thanks to Remark 14.7.2),

$$h\big(\frac{2}{M}\big) \le \frac{4}{M} \ln \frac{M}{2}\,.$$

Substituting this into (14.9),

$$\mathbb{P}\{\widehat{\omega} \ne \omega\} \ge 1 - \frac{\frac{4N}{M} \ln(M/2) + \ln 2}{\ln M}\,. \tag{14.10}$$

If $M \ge 4$, and $N \le \frac{1}{12} M$, we would obtain $\mathbb{P}\{\widehat{\omega} \ne \omega\} > 5/6$, contradicting (14.8). Hence, we deduce that $N \gtrsim M$.

In general, if the algorithm is randomized, it can depend on a random seed ζ that is independent of ω. Then we can apply (14.10) conditional on ζ, and obtain

$$\mathbb{P}\{\widehat{\omega} \ne \omega \mid \zeta\} \ge 1 - \frac{\frac{4N}{M} \ln(M/2) + \ln 2}{\ln M}\,.$$

Taking expectation over ζ, we see that the lower bound (14.10), and hence $N \gtrsim M$, holds for randomized algorithms as well.

The proof of Theorem 14.4.1 is concluded by noting that the estimation lower bound gives a lower bound on sampling, and that Property (**P.4**) provides us with a lower bound on M. □

In the remaining sections, we focus on establishing Lemma 14.7.1.

■ 14.7.2 Estimates for integrals

In this section we provide useful estimates for integrals that appear in the normalizing constants for our lower bound construction. Notice that since $\tilde{\pi}_\omega = 1$ on $B_R \setminus (\omega + B_r)$,

$$Z_\omega = \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + \tilde{\pi}_\omega(B_R)$$

$$= \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + (R^d - r^d)\, V_d + \int_{B_r} \exp\Big(r^2 \phi\big(\frac{\|x\|}{r}\big)\Big)\, \mathrm{d}x$$
$$= \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + (R^d - r^d)\, V_d + r^d\, I_r\,,$$

where we define $I_r \coloneqq \int_{B_1} \exp(r^2 \phi(\|x\|))\, \mathrm{d}x$. We record some useful properties of the quantities defined thus far that will be used throughout the proof of Lemma 14.7.1.

Lemma 14.7.4 (Main estimates)**.** *For any number $c > 0$ there exists $c_r(c) > 0$ depending only on c such that for all $r \geq c_r(c)\sqrt{d}$, the following hold:*

1. *(asymptotics of I_r)*

$$\frac{1}{2} \leq \frac{r^d\, I_r}{(2\pi)^{d/2} \exp(r^2 \phi(0))} \leq 2\,. \tag{14.11}$$

2. *(mass outside B_R) There is a universal constant $c_0 > 2$, independent of c, such that*

$$\sqrt{\frac{\pi}{2}}\, V_d d R^{d-1} \leq \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) \leq V_d c_0^d\, (d R^{d-1} + d^{(d+1)/2})\,. \tag{14.12}$$

3. *(mass on the bump)*

$$\ln \frac{I_r}{V_d} \geq cd\,. \tag{14.13}$$

Proof. Because we have chosen ϕ to be a quadratic function on the range $[0, \alpha]$ (see (ϕ**.2**)), we can decompose I_r as follows:

$$I_r \coloneqq \int_{B_1} \exp\big(r^2 \phi(\|x\|)\big)\, \mathrm{d}x$$
$$= \underbrace{\int_{B_1 \setminus B_\alpha} \exp\big(r^2 \phi(\|x\|)\big)\, \mathrm{d}x}_{\mathsf{A}} + \underbrace{\exp\big(r^2 \phi(0)\big) \int_{B_\alpha} \exp\Big(-\frac{r^2 \|x\|^2}{2}\Big)\, \mathrm{d}x}_{\mathsf{B}}\,.$$

As ϕ is decreasing by (ϕ**.1**), clearly $\mathsf{A} \leq V_d \exp(r^2 \phi(\alpha))$, and the second term is given by

$$\mathsf{B} = \frac{(2\pi)^{d/2}}{r^d} \exp\big(r^2 \phi(0)\big)\, \mathbb{P}(\|X\| \leq \alpha r)\,,$$

where X is a standard Gaussian in $\mathbb{R}^d$. By standard concentration inequalities (e.g., Markov's inequality suffices), there exists a universal constant c_1 such that the above probability is at least $1/2$ provided $r \geq c_1\sqrt{d}$. Recall that $\log\Gamma(\frac{d}{2}+1) = \frac{d}{2}\log d + O(d)$. Thus, for $r \geq c_1\sqrt{d}$ we have

$$\begin{aligned}\log\frac{\mathsf{A}}{\mathsf{B}} &\leq \log\frac{2V_d\exp(r^2\phi(\alpha))}{(2\pi)^{d/2}\,r^{-d}\exp(r^2\phi(0))} = \log\frac{2\exp(r^2\phi(\alpha))}{2^{d/2}\,r^{-d}\exp(r^2\phi(0))\,\Gamma(\frac{d}{2}+1)}\\ &= O(d) - r^2\left(\phi(0)-\phi(\alpha)\right) + d\log r - \frac{d}{2}\log d\\ &= O(d) - d\left(\left(\frac{r}{\sqrt{d}}\right)^2\left(\phi(0)-\phi(\alpha)\right) - \log\left(\frac{r}{\sqrt{d}}\right)\right).\end{aligned}$$

From the above it is clear that there is a universal constant c_2 such that $r \geq c_2\sqrt{d}$ implies that $\mathsf{A} \leq \mathsf{B}$. Thus, for $r \geq (c_1 \vee c_2)\sqrt{d}$ the following holds:

$$\mathsf{B} \leq I_r \leq 2\mathsf{B}\,, \tag{14.14}$$

proving (14.11). We now turn to the proof of (14.12). By integrating in polar coordinates, and taking X to be a standard Gaussian on $\mathbb{R}$,

$$\begin{aligned}\tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) &= A_{d-1}\int_R^\infty s^{d-1}\exp\bigl(-\frac{1}{2}(s-R)^2\bigr)\,\mathrm{d}s\\ &= A_{d-1}\int_0^\infty (s+R)^{d-1}\exp\bigl(-\frac{s^2}{2}\bigr)\,\mathrm{d}s\\ &\leq \sqrt{2\pi}\,A_{d-1}\,\mathbb{E}[|X+R|^{d-1}]\\ &\leq \sqrt{2\pi}\,A_{d-1}2^d\,(R^{d-1}+\mathbb{E}[|X|^{d-1}])\\ &\leq A_{d-1}c_0^d\,(R^{d-1}+(d-1)^{(d-1)/2})\\ &\leq V_d c_0^d\,(dR^{d-1}+d^{(d+1)/2})\end{aligned}$$

for some universal constant $c_0 > 2$. For the other direction we can simply write

$$\begin{aligned}\tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) &= A_{d-1}\int_0^\infty (s+R)^{d-1}\exp\bigl(-\frac{s^2}{2}\bigr)\,\mathrm{d}s\\ &\geq \sqrt{\frac{\pi}{2}}\,A_{d-1}R^{d-1}\,.\end{aligned}$$

Finally, we prove (14.13). We again use the fact $\log\Gamma(\frac{d}{2}+1) = \frac{d}{2}\log d + O(d)$. Therefore, for $r \geq (c_1 \vee c_2)\sqrt{d}$ and using (14.11) we obtain

$$\log\frac{I_r}{V_d} \geq \log\frac{(2\pi)^{d/2}\,r^{-d}\exp(r^2\phi(0))/2}{\pi^{d/2}/\Gamma(\frac{d}{2}+1)}$$

$$= d\left(\left(\frac{r}{\sqrt{d}}\right)^2 \phi(0) - \log\left(\frac{r}{\sqrt{d}}\right)\right) + O(d)\,.$$

Clearly, there exists a constant c_3 (depending only on c) such that $r \geq c_3\sqrt{d}$ implies that the RHS is at least linear in d with a positive constant. Taking $c_r = c_1 \vee c_2 \vee c_3$ concludes the proof. □

■ 14.7.3 Proof of Property (P.1)

We choose r, R such that (P.1) holds, i.e., $\pi_\omega(\omega + B_r) = 1/2$. This holds if

$$f(r) \coloneqq (I_r + V_d)\, r^d \stackrel{!}{=} \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + V_d R^d =: g(R)\,. \tag{14.15}$$

Lemma 14.7.5 (Choice of r, R)**.** *For any $d \geq 1$, $R \geq 0$, there exists a corresponding value of r such that (14.15) holds. Moreover, there is a universal constant $c_R \geq 1$ such that for any $R \geq c_R\sqrt{d}$, the corresponding r solving (14.15) satisfies*

$$r \geq c_r\big(\log(6c_0)\big)\sqrt{d}\,, \tag{14.16}$$
$$R/r \geq 2\,, \tag{14.17}$$

where $c_r(\cdot)$ is the function defined in Lemma 14.7.4.

The argument $\log(6c_0)$ to $c_r(\cdot)$ in Lemma 14.7.5 is chosen for later convenience.

Proof. Notice that f and g are continuous and increasing in r, R respectively. Moreover, we check that $f(0) = 0$, $g(0) = (2\pi)^{d/2}$, and $f(\infty) = g(\infty) = \infty$. This tells us that for any value of $d \geq 1$ and $R \geq 0$, there exists a value of $r \geq 0$ for which $f(r) = g(R)$.

For the rest of the proof, we abbreviate $c_r \coloneqq c_r(\log(6c_0))$.

First, we prove (14.16). Note that since (14.16) is a hypothesis of Lemma 14.7.4, we cannot invoke Lemma 14.7.4 during the proof of (14.16) in order to avoid a circular argument.

By the definitions of r and R,

$$(I_r + V_d)\, r^d \geq V_d\, R^d\,.$$

Taking logarithms and using the definition of I_r, this rewrites as

$$\begin{aligned} d\log\frac{R}{r} &\leq \log\big(1 + \frac{I_r}{V_d}\big) = \log\Big(1 + \frac{\int_{B_1} \exp(r^2\phi(\|x\|))\,\mathrm{d}x}{V_d}\Big) \\ &\leq \log\big(1 + \exp\big(r^2\phi(0)\big)\big)\,. \end{aligned}$$

Suppose, for the sake of contradiction, that $r < c_r\sqrt{d}$. Then, we have

$$d \log \frac{R}{r} \leq c_r^2 d\phi(0) + \log 2 \,.$$

Rearranging,

$$R \leq \exp\Big(c_r^2\phi(0) + \frac{\log 2}{d}\Big)\, r \leq \exp\Big(c_r^2\phi(0) + \frac{\log 2}{d}\Big)\, c_r\sqrt{d} \,.$$

Hence, if $R \geq c_R\sqrt{d}$ for a large enough universal constant c_R, then we arrive at the desired contradiction. For later convenience we choose c_R to always be at least 1. This proves (14.16).

Next, we prove (14.17). We use the fact that $R \geq c_R\sqrt{d}$; so that in particular $c_R \geq 1$ and thus $\sqrt{d} \leq R$. Then, using (14.12) from Lemma 14.7.4,

$$\begin{aligned}(I_r + V_d)\, r^d &\leq V_d\, (c_0^d d R^{d-1} + c_0^d d^{(d+1)/2} + R^d) \leq V_d\, (c_0^d \sqrt{d} R^d + c_0^d \sqrt{d} R^d + R^d) \\ &\leq V_d \cdot 3 c_0^d \sqrt{d}\, R^d \,.\end{aligned}$$

Taking logarithms, rearranging, and using (14.13) from Lemma 14.7.4,

$$d \log \frac{R}{r} \geq \log\big(1 + \frac{I_r}{V_d}\big) - d \log c_0 - \log(3\sqrt{d}) \geq \big(c - \log c_0 - \underbrace{\frac{\log(3\sqrt{d})}{d}}_{\leq \log 3}\big)\, d \,.$$

Taking $c = \log c_0 + \log 3 + \log 2 = \log(6c_0)$, this implies $R/r \geq 2$ as desired. □

■ 14.7.4 Proof of Property (P.2)

The proof of Property (**P.2**) requires an upper bound on the Poincaré constant of π_ω. We recall that the Poincaré constant of a probability measure π is the smallest constant $C_{\mathsf{PI}}(\pi) > 0$ such that for all smooth and bounded test functions $f : \mathbb{R}^d \to \mathbb{R}$, it holds that

$$\mathrm{var}_\pi(f) \leq C_{\mathsf{PI}}(\pi)\, \mathbb{E}_\pi[\|\nabla f\|^2] \,.$$

We begin with a Poincaré inequality for π_{init}.

Lemma 14.7.6 (Poincaré inequality for π_{init})**.** *If $R \geq \sqrt{d}$, then the probability measure π_{init} has Poincaré constant at most $c_{\mathsf{PI}} R^2/d$ for a universal constant c_{PI}.*

Proof. From [Bob03] and the fact that π_{init} is a radially symmetric log-concave measure, the Poincaré constant of π_{init} is bounded by

$$C_{\mathsf{PI}}(\pi_{\mathsf{init}}) \leq \frac{13\, \mathbb{E}_{\pi_{\mathsf{init}}}[\|\cdot\|^2]}{d} \,.$$

The second moment is

$$\begin{aligned}\mathbb{E}_{\pi_{\mathsf{init}}}[\|\cdot\|^2] &= \frac{\int_{B_R}\|\cdot\|^2}{Z_{\mathsf{init}}} + \frac{\int_{\mathbb{R}^d\setminus B_R}\|\cdot\|^2 \exp(-\frac{1}{2}\left(\|\cdot\|-R\right)^2)}{Z_{\mathsf{init}}} \\ &\le \frac{\int_{B_R}\|\cdot\|^2}{V_d R^d} + \frac{A_{d-1}\int_0^\infty (r+R)^{d+1}\exp(-r^2/2)\,\mathrm{d}r}{A_{d-1}\int_0^\infty (r+R)^{d-1}\exp(-r^2/2)\,\mathrm{d}r} \\ &\le R^2 + \int (r+R)^2\,\nu(\mathrm{d}r)\,,\end{aligned}$$

where ν is the probability measure on $\mathbb{R}_+$ with density

$$\nu(r) \propto (r+R)^{d-1}\exp\bigl(-\frac{r^2}{2}\bigr)\,. \tag{14.18}$$

Note that ν is 1-strongly-log-concave. Hence, by Lemma 2.2.13,

$$\begin{aligned}\int (r+R)^2\,\nu(\mathrm{d}r) &\lesssim R^2 + \int r^2\,\nu(\mathrm{d}r) \lesssim R^2 + r_\star^2 + \int (r-r_\star)^2\,\nu(\mathrm{d}r) \\ &\le R^2 + r_\star^2 + 1\,,\end{aligned}$$

where $r_\star$ is the mode of ν. To find the mode, (14.18) and elementary calculus show that $r_\star$ satisfies $r_\star\,(r_\star+R) = d-1$, which implies $r_\star \le (d-1)/R$. If $R \ge \sqrt{d}$, then $r_\star \lesssim R$. Combining the bounds, we obtain $C_{\mathsf{PI}}(\pi_{\mathsf{init}}) \lesssim R^2/d$. □

Next, we recall the statement of the Holley–Stroock perturbation principle.

Theorem 14.7.7 (Holley–Stroock perturbation principle, [HS87])**.** *Let π be a probability measure which satisfies a Poincaré inequality. Suppose that μ is another probability measure such that*

$$0 < c \le \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \le C < \infty\,.$$

Then, μ also satisfies a Poincaré inequality, with

$$C_{\mathsf{PI}}(\mu) \le \frac{C}{c}\,C_{\mathsf{PI}}(\pi)\,.$$

Proof. See [BGL14, Lemma 5.1.7]. □

Corollary 14.7.8 (Poincaré inequality for π_ω)**.** *Assume that $R \ge \sqrt{d}$. Then, for each $\omega \in \mathscr{P}_{2r,R}$,*

$$C_{\mathsf{PI}}(\pi_\omega) \le \frac{2c_{\mathsf{PI}}R^2}{d}\exp\bigl(r^2\phi(0)\bigr)\,.$$

Proof. By (ϕ.1), we know that $\tilde{\pi}_\omega \geq \tilde{\pi}_{\mathsf{init}}$ and hence $Z_\omega \geq Z_{\mathsf{init}}$. It follows that

$$\frac{Z_{\mathsf{init}}}{Z_\omega} \leq \frac{\pi_\omega}{\pi_{\mathsf{init}}} = \frac{\tilde{\pi}_\omega}{\tilde{\pi}_{\mathsf{init}}} \frac{Z_{\mathsf{init}}}{Z_\omega} \leq \frac{\tilde{\pi}_\omega}{\tilde{\pi}_{\mathsf{init}}} \leq \exp\bigl(r^2 \phi(0)\bigr) .$$

Also, by (14.15),

$$\begin{aligned} Z_\omega = \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + V_d\, R^d + (I_r - V_d)\, r^d &\leq \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + V_d\, R^d + (I_r + V_d)\, r^d \\ &= 2\left(\tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + V_d\, R^d\right) = 2\, Z_{\mathsf{init}} . \end{aligned}$$

Hence, $Z_{\mathsf{init}}/Z_\omega \geq 1/2$. The result now follows from Lemma 14.7.6 and the Holley–Stroock perturbation principle (Theorem 14.7.7). □

To translate Fisher information guarantees into total variation guarantees, we use the following consequence of the Poincaré inequality.

Proposition 14.7.9 (Fisher information controls total variation)**.** *Suppose that a probability measure π satisfies a Poincaré inequality. Then, for any probability measure μ,*

$$\mathsf{TV}(\mu, \pi)^2 \leq \frac{C_{\mathsf{PI}}(\pi)}{4}\, \mathsf{FI}(\mu \parallel \pi) .$$

Proof. See [Gui+09]. □

We are finally ready to prove Property (**P.2**). More specifically, we will show that there is a universal constant $c_\varepsilon > 0$ such that if $\varepsilon \leq \exp(-c_\varepsilon d)$, then we can choose r and R (depending on ε) such that: (i) r and R are related according to (14.15), which is necessary for Property (**P.1**); (ii) $R \geq c_R\sqrt{d}$, which is necessary for Lemma 14.7.5; and (iii) Property (**P.2**) holds.

*Proof of Property (**P.2**).* For $\omega \in \mathscr{P}_{2r,R}$, suppose that μ satisfies $\sqrt{\mathsf{FI}(\mu \parallel \pi_\omega)} \leq \varepsilon$. Then, by Corollary 14.7.8 and Proposition 14.7.9, we have

$$\mathsf{TV}^2(\mu, \pi_\omega) \leq \frac{C_{\mathsf{PI}}(\pi_\omega)}{4}\, \mathsf{FI}(\mu \parallel \pi) \leq \frac{c_{\mathsf{PI}} R^2 \exp(r^2 \phi(0))}{2d}\, \varepsilon^2 . \tag{14.19}$$

Hence, if we choose

$$R^2 = \frac{2d}{9 c_{\mathsf{PI}} \varepsilon^2 \exp(r^2 \phi(0))} \tag{14.20}$$

then $\sqrt{\mathsf{FI}(\mu \parallel \pi_\omega)} \leq \varepsilon$ implies $\mathsf{TV}(\mu, \pi_\omega) \leq 1/3$, i.e., Property (**P.2**) holds.

To justify (14.20), note that thus far we have shown that for any choice of R, there exists a choice of r which depends on R, which we temporarily denote

by $r(R)$, such that (14.15) holds. Also, $r(\cdot)$ is an increasing function. In order for (14.20) to hold, it is equivalent to require

$$R^2 \exp\big(r(R)^2\,\phi(0)\big) = \frac{2d}{9c_{\mathsf{PI}}\varepsilon^2} \tag{14.21}$$

where the left-hand side is an increasing function of R. We also want R to satisfy $R \geq c_R\sqrt{d}$, where c_R is the universal constant in Lemma 14.7.5. From Lemma 14.7.5, for the choice of $R = c_R\sqrt{d}$,

$$r(c_R\sqrt{d}) \leq \frac{c_R\sqrt{d}}{2}\,.$$

Therefore, for this choice of R, the left-hand side of (14.21) is bounded by

$$c_R^2 d \exp\big(\frac{c_R^2 d}{4}\,\phi(0)\big)\,.$$

If it holds that

$$\varepsilon^2 \leq \frac{2}{9c_{\mathsf{PI}}c_R^2\exp(c_R^2 d\phi(0)/4)} \tag{14.22}$$

then the R satisfying (14.20) necessarily satisfies $R \geq c_R\sqrt{d}$. In turn, (14.22) holds if $\varepsilon \leq \exp(-c_\varepsilon d)$ for a universal constant $c_\varepsilon > 0$. □

■ 14.7.5 Proof of Property (P.3)

Proof of Property (P.3). In the proof of Corollary 14.7.8, we showed that $Z_\omega \leq 2\,Z_{\mathsf{init}}$. The KL divergence is bounded by

$$\mathsf{KL}(\pi_{\mathsf{init}} \,\|\, \pi_\omega) = \mathbb{E}_{\pi_{\mathsf{init}}} \ln\big(\underbrace{\frac{\tilde\pi_{\mathsf{init}}}{\tilde\pi_\omega}}_{\leq 1}\,\underbrace{\frac{Z_\omega}{Z_{\mathsf{init}}}}_{\leq 2}\big) \leq \log 2\,,$$

which is what we wanted to show. □

■ 14.7.6 Proof of Property (P.4)

Proof of Property (P.4). We choose r and R to satisfy (14.15) and (14.20). If $\varepsilon \leq \exp(-c_\varepsilon d)$, then we showed in the proof of Property (P.2) that $R \geq c_R\sqrt{d}$ and hence Lemmas 14.7.4 and 14.7.5 apply.

As in the proof of (14.17) in Lemma 14.7.5, $R \geq \sqrt{d}$ implies

$$(I_r + V_d)\,r^d \leq V_d \cdot 3c_0^d\sqrt{d}\,R^d\,.$$

Taking logarithms in (14.11) from Lemma 14.7.4 and using the above inequality, we obtain

$$r^2\phi(0) \le \log \frac{2r^d I_r}{(2\pi)^{d/2}} \le O(d) + \log V_d + d\log R\,.$$

From (14.20), we have

$$\log R = \frac{1}{2}\log d + \log\frac{1}{\varepsilon} - \frac{1}{2}r^2\phi(0) + O(1)\,.$$

Substituting this in and using $\log V_d = -\frac{d}{2}\log d + O(d)$,

$$r^2\phi(0) \le d\log\frac{1}{\varepsilon} - \frac{d}{2}r^2\phi(0) + O(d)$$

which is rearranged to yield

$$r^2\phi(0) \le \frac{2d}{d+2}\log\frac{1}{\varepsilon} + O(1)\,.$$

Then, the packing number is lower bounded by

$$\begin{aligned}|\mathscr{P}_{2r,R}| \ge \Big(\frac{R-r}{2r}\Big)^d &\ge \Big(\frac{R}{4r}\Big)^d \\ &\ge \Big(c\,\frac{\sqrt{d}\exp(-\frac{d}{d+2}\log(1/\varepsilon))}{\varepsilon\sqrt{\log(1/\varepsilon)}}\Big)^d \\ &\ge \Big(c\sqrt{\frac{d}{\log(1/\varepsilon)}}\Big)^d \frac{1}{\varepsilon^{2d/(d+2)}}\,,\end{aligned}$$

for some universal constant c. □

■ 14.7.7 Auxiliary lemmas

Lemma 14.7.10. *Suppose that $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies (ϕ**.1**), (ϕ**.2**), and (ϕ**.3**). Then, the map $x \mapsto \phi(\|x\|)$ is 1-smooth on $\mathbb{R}^d$.*

Proof. First, we claim that $|\phi'(x)|/x \le 1$ for all $x > 0$. This follows from (ϕ**.3**) because (ϕ**.2**) implies that the right derivative $\phi'(0+)$ exists and equals 0.

Next, we have for $x \ne 0$

$$\partial_{x_i}\partial_{x_j}\phi(\|x\|) = \partial_{x_j}\,\phi'(\|x\|)\,\frac{x_i}{\|x\|}$$

$$= \phi''(\|x\|)\,\frac{x_i x_j}{\|x\|^2} - \phi'(\|x\|)\,\frac{x_i x_j}{\|x\|^3} + \delta_{i,j}\,\phi'(\|x\|)\,\frac{1}{\|x\|}\,.$$

Thus, in matrix form we have

$$\nabla_x^2 \phi(\|x\|) = \frac{\phi'(\|x\|)}{\|x\|}\,I_d + \Big(\frac{\phi''(\|x\|)}{\|x\|^2} - \frac{\phi'(\|x\|)}{\|x\|^3}\Big)\,xx^{\mathsf{T}}.$$

In particular, the eigenvalues are always $\frac{\phi'(\|x\|)}{\|x\|}$ with multiplicity $d-1$ and $\phi''(\|x\|)$ with multiplicity 1. The fact that $\phi(\|\cdot\|)$ is 1-smooth follows. □

■ 14.7.8 Optimization of the bound

We wish to find d which maximizes

$$\Big(\frac{cd}{\log(1/\varepsilon)}\Big)^{d/2}\,\varepsilon^{4/(d+2)}\,,$$

or after taking logarithms, we wish to maximize

$$f(d) \coloneqq \frac{d}{2}\log d - \frac{4}{d+2}\log\frac{1}{\varepsilon} - \frac{d}{2}\log\log\frac{1}{\varepsilon} - \frac{d}{2}\log\frac{1}{c}\,.$$

Rather than maximizing this expression exactly, we shall ignore the last two terms and pick d to be the smallest integer such that the sum of the first two terms is non-negative, i.e.,

$$\frac{d\,(d+2)\log d}{8} \geq \log\frac{1}{\varepsilon}\,.$$

It suffices to find d such that $g(d) \coloneqq d^2 \log d \geq 8\log(1/\varepsilon)$. In order to invert g, let y be sufficiently large and consider finding x such that $g(x) = y$. We make the choice $x = \alpha\sqrt{y/(\log y)}$ and plug this into the expression for g in order to obtain

$$\begin{aligned}\log g\Big(\alpha\sqrt{\frac{y}{\log y}}\Big) &= 2\log\alpha + \log y - \log\log y + \log\log\Big(\alpha\sqrt{\frac{y}{\log y}}\Big)\\ &= 2\log\alpha + \log y + \underbrace{\log\frac{\frac{1}{2}\log y - \frac{1}{2}\log\log y + \log\alpha}{\log y}}_{\to\log(1/2)\ \text{as}\ y\to\infty}\,.\end{aligned}$$

From this expression, we see that provided y is sufficiently large, this expression is less than $\log y$ for $\alpha = 0$ and greater than $\log y$ for $\alpha = 3$. We conclude that $g^{-1}(y) \asymp \sqrt{y/(\log y)}$, and therefore that our choice of d satisfies

$$d \asymp \sqrt{\frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}}\,.$$

In particular, since $d = o(\log(1/\varepsilon))$, then the condition $\varepsilon \le \exp(-c_\varepsilon d)$ holds for all sufficiently small ε, and Theorem 14.4.1 holds. Then,

$$f(d) \ge -\frac{d}{2}\log\log\frac{1}{\varepsilon} - \frac{d}{2}\log\frac{1}{c} \asymp -\sqrt{\left(\log\frac{1}{\varepsilon}\right)\left(\log\log\frac{1}{\varepsilon}\right)}\,.$$

This verifies the expression in §14.4.

To justify the simplified expression of the bound that we gave in the informal statement of Theorem 14.1.3, note that in dimension

$$d \lesssim \sqrt{\frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}} \tag{14.23}$$

we have

$$\log\Big(\big(\frac{cd}{\log(1/\varepsilon)}\big)^{d/2}\Big) = \frac{d}{2}\underbrace{\left(\log d - \log\log\frac{1}{\varepsilon} - \log\frac{1}{c}\right)}_{\text{negative as } \varepsilon\searrow 0} \gtrsim -\sqrt{\left(\log\frac{1}{\varepsilon}\right)\left(\log\log\frac{1}{\varepsilon}\right)}\,.$$

In other words, we can simplify our bound as follows. For all $d \ge 1$ and all ε smaller than a universal constant, if the condition (14.23) holds, then we have the lower bound

$$\mathscr{C}(d,1,\varepsilon) \gtrsim \frac{1}{\varepsilon^{2d/(d+2)}\exp(C\sqrt{\log(1/\varepsilon)\log\log(1/\varepsilon)})}\,.$$

Otherwise, if the condition (14.23) fails, then we instead have the bound

$$\begin{aligned}\mathscr{C}(d,1,\varepsilon) &\gtrsim \frac{1}{\varepsilon^{2}\exp(C\sqrt{\log(1/\varepsilon)\log\log(1/\varepsilon)})}\\ &\ge \frac{1}{\varepsilon^{2d/(d+2)}\exp(C\sqrt{\log(1/\varepsilon)\log\log(1/\varepsilon)})}\,.\end{aligned}$$

In either case, we have $\mathscr{C}(d,1,\varepsilon) \ge (1/\varepsilon)^{2d/(d+2)-o(1)}$. Together with Theorem 14.4.2 on the univarate case and Lemma 14.2.3 on rescaling, it yields Theorem 14.1.3.

■ 14.7.9 Proof of Theorem 14.4.2

In the univarate case, we can sharpen Theorem 14.4.1 by obtaining a better bound on the Poincaré constant of π_ω. We use the following result.

Theorem 14.7.11 (Muckenhoupt's criterion)**.** *Let π be a probability density on $\mathbb{R}$ and let m be a median of π. Then,*

$$C_{\mathsf{PI}}(\pi) \asymp \max\Big\{\sup_{x<m} \pi\big((-\infty, x]\big) \int_x^m \frac{1}{\pi},\ \sup_{x>m} \pi\big([x, +\infty)\big) \int_m^x \frac{1}{\pi}\Big\}\,.$$

Proof. See [BGL14, Theorem 4.5.1]. □

Lemma 14.7.12 (Improved Poincaré inequality for π_ω)**.** *Suppose that $d = 1$ and $R \geq 1$. Then, for all $\omega \in \mathscr{P}_{2r,R}$,*

$$C_{\mathsf{PI}}(\pi_\omega) \lesssim R^2\,.$$

Proof. We use Muckenhoupt's criterion (Theorem 14.7.11). First, we note that by Property **(P.1)**, it holds that $\pi_\omega(\omega + B_r) = \frac{1}{2}$ which implies that $\omega - r \leq m \leq \omega + r$. We proceed to check that

$$\sup_{x>m} \pi_\omega\big([x, +\infty)\big) \int_m^x \frac{1}{\pi_\omega} \lesssim R^2\,.$$

The other condition is verified in the same way due to symmetry.

We split into three cases. First, suppose that $m < x < \omega + r$. Then, as in the proof of Corollary 14.7.8, we have $Z_\omega \leq 2\, Z_{\mathsf{init}} = 2\, \tilde{\pi}_{\mathsf{init}}(\mathbb{R} \setminus B_R) + 4R \leq 2\sqrt{2\pi} + 4R \lesssim R$. Then,

$$\pi_\omega\big([x, +\infty)\big) \int_m^x \frac{1}{\pi_\omega} \leq Z_\omega\,(x - m) \lesssim Rr \leq R^2\,.$$

Next, suppose that $\omega + r < x < R$. Then,

$$\pi_\omega\big([x, +\infty)\big) \int_m^x \frac{1}{\pi_\omega} = \tilde{\pi}_\omega\big([x, +\infty)\big) \int_m^x \frac{1}{\tilde{\pi}_\omega} \leq \Big(R - x + \sqrt{\frac{\pi}{2}}\Big)\,(x - m) \lesssim R^2\,.$$

Finally, suppose that $x > R$. Then, using standard Gaussian tail bounds,

$$\begin{aligned}
\pi_\omega\big([x, +\infty)\big) \int_m^x \frac{1}{\pi_\omega} &= \tilde{\pi}_\omega\big([x, +\infty)\big) \int_m^x \frac{1}{\tilde{\pi}_\omega} \\
&\leq \Big[\sqrt{2\pi}\,\big(\frac{1}{2} \wedge \frac{1}{x - R}\big) \exp\Big(-\frac{(x - R)^2}{2}\Big)\Big] \\
&\qquad \times \Big[R - m + (x - R) \exp\Big(\frac{(x - R)^2}{2}\Big)\Big]\,.
\end{aligned}$$

If $x - R \leq 1$, then this yields

$$\pi_\omega\big([x, +\infty)\big) \int_m^x \frac{1}{\pi_\omega} \lesssim R\,.$$

Otherwise, if $x - R \geq 1$, then we obtain

$$\pi_\omega\big([x, +\infty)\big) \int_m^x \frac{1}{\pi_\omega} \lesssim \frac{1}{x-R} \exp\Big(-\frac{(x-R)^2}{2}\Big) \Big[R + (x-R)\exp\Big(\frac{(x-R)^2}{2}\Big)\Big] \lesssim R\,.$$

This completes the proof. □

We now use the improved Poincaré inequality in order to establish Theorem 14.4.2.

Proof of Theorem 14.4.2. We follow the proof of Theorem 14.4.1. The proofs of Properties **(P.1)** and **(P.3)** remain unchanged.

In the proof of Property **(P.2)**, the equation (14.19) is replaced by

$$\mathsf{TV}^2(\mu, \pi_\omega) \leq c_{\mathsf{PI}} R^2 \varepsilon^2$$

for a different universal constant $c_{\mathsf{PI}} > 0$, using Lemma 14.7.12. Hence, we choose $R^2 = 1/(9c_{\mathsf{PI}}\varepsilon^2)$ in order to verify Property **(P.2)**. Since we require $R \geq c_R$ for a universal constant $c_R \geq 1$, this requires $\varepsilon \leq \exp(-c_\varepsilon)$ for a universal constant $c_\varepsilon > 0$.

Next, we turn towards the sharpened version of Property **(P.4)**. From (14.15), r is chosen so that

$$(I_r + 2)\, r = \tilde{\pi}_\omega(\mathbb{R} \setminus B_R) + 2R\,.$$

Using (14.11) from Lemma 14.7.4, we have

$$r I_r \asymp \exp\big(r^2 \phi(0)\big) \gtrsim r\,.$$

This implies that

$$\exp\big(r^2\phi(0)\big) \gtrsim (I_r + 2)\, r \gtrsim R\,,$$

or $r \gtrsim \sqrt{\log R} \asymp \sqrt{\log(1/\varepsilon)}$. Hence,

$$|\mathscr{P}_{2r,R}| \geq \frac{R}{4r} \gtrsim \frac{1}{\varepsilon\sqrt{\log(1/\varepsilon)}}\,.$$

By substituting this new bound on the packing number into the information theoretic argument of Theorem 14.4.1 (see (14.10), where $M = |\mathscr{P}_{2r,R}|$), we obtain Theorem 14.4.2. □

14.8 Further discussion of the univariate case

In this section, we provide further discussion of algorithms for the univariate case.

Rejection sampling. First of all, we note that the $\mathrm{poly}(1/\varepsilon)$ lower bounds of Theorems 14.4.1 and 14.4.2 may come as a surprise due to the existence of the rejection sampling algorithm. We briefly recall rejection sampling here. Let $\tilde{\pi}$ be an unnormalized density, let $Z_\pi \coloneqq \int \tilde{\pi}$ denote the normalizing constant, and let $\pi \coloneqq \tilde{\pi}/Z$ denote the target distribution. Rejection sampling requires knowledge of an upper envelope $\tilde{\mu}$ for $\tilde{\pi}$, i.e., a function $\tilde{\mu}$ satisfying $\tilde{\mu} \geq \tilde{\pi}$ pointwise. The algorithm proceeds by repeatedly drawing samples from the density $\mu \coloneqq \tilde{\mu}/Z_\mu$, where $Z_\mu \coloneqq \int \tilde{\mu}$; each sample X is accepted with probability $\tilde{\pi}(X)/\tilde{\mu}(X)$.

It is standard to show (see Theorem 4.4.6) that the accepted samples are drawn exactly from the target π, and that the number of queries made to $\tilde{\pi}$ until the first accepted sample is geometrically distributed with mean Z_μ/Z_π. To translate this into a total variation guarantee, we run the algorithm for N iterations and output "FAIL" if we have not accepted a sample by iteration N. The probability of failure is at most $(1 - Z_\pi/Z_\mu)^N$, so the number of iterations required for the output of the algorithm to be ε-close to the target π in total variation distance is $N \geq \log(1/\varepsilon)/\log(1 - Z_\pi/Z_\mu)$.

Although this is a total variation guarantee, rather than a Fisher information guarantee, it suggests (similarly to §14.5) that $\log(1/\varepsilon)$ rates are attainable using rejection sampling. The reason why this does not contradict our lower bounds in Theorems 14.4.1 and 14.4.2 is that the initialization oracle we consider, which provides a measure μ_0 such that $\mathsf{KL}(\mu_0 \,\|\, \pi) \leq K_0$, is not sufficient to construct an upper envelope of the unnormalized density $\tilde{\pi}$.

Indeed, consider instead a stronger initialization oracle which outputs a measure μ_0 such that

$$\max\Big\{\sup \ln \frac{\mu_0}{\pi},\ \sup \ln \frac{\pi}{\mu_0}\Big\} \leq M_0 < \infty\,.$$

Denote the complexity of obtaining $\sqrt{\mathsf{FI}(\mu \,\|\, \pi)} \leq \varepsilon$ over the class of 1-log-smooth distributions on $\mathbb{R}^d$ with this stronger initialization oracle by $\mathscr{C}_\infty(d, M_0, \varepsilon)$. Then, the rejection sampling algorithm can be implemented within this new oracle model. It yields the following.

Proposition 14.8.1 (Fisher information guarantees via rejection sampling)**.** *It holds that*

$$\mathscr{C}_\infty(d, M_0, \varepsilon) \leq \widetilde{O}\Big(\exp(3M_0) \log \frac{\sqrt{d}}{\varepsilon}\Big)\,.$$

Proof. For the algorithm, we use rejection sampling, which requires producing an upper envelope. Recall that in our oracle model, we can query the value of an unnormalized version $\tilde{\pi}$ of π. By replacing $\tilde{\pi}$ with $\tilde{\pi}/\tilde{\pi}(0)$, we can assume that $\tilde{\pi}(0) = 1$. Then,

$$\tilde{\pi} = \frac{\tilde{\pi}}{\tilde{\pi}(0)} = \frac{\pi}{\pi(0)} \le \frac{\exp(M_0)\,\mu_0}{\exp(-M_0)\,\mu_0(0)} = \underbrace{\frac{\exp(2M_0)}{\mu_0(0)}}_{\coloneqq Z_{\mu_0}} \mu_0\,.$$

This shows that $\tilde{\mu}_0 \coloneqq Z_{\mu_0}\,\mu_0$ is an upper envelope for $\tilde{\pi}$. Also, using $\pi(0) = 1/Z_\pi$,

$$\frac{Z_{\mu_0}}{Z_\pi} = \exp(2M_0)\,\frac{\pi(0)}{\mu_0(0)} \le \exp(3M_0)\,.$$

Hence, we can run rejection sampling, where we output a sample from μ_0 if the algorithm exceeds N iterations. Therefore, the law of the output of rejection sampling is $\mu = (1-p)\,\pi + p\,\mu_0$, where $p = (1 - Z_\pi/Z_{\mu_0})^N \le \exp(-NZ_\pi/Z_{\mu_0})$ is the probability of failure. We calculate

$$1 + \chi^2(\mu \,\|\, \pi) = \mathbb{E}_\mu\Big(\frac{\mu}{\pi}\Big) = 1 - p + p\,\mathbb{E}_\mu\Big(\frac{\mu_0}{\pi}\Big) \le 1 + p\exp(M_0)\,.$$

Applying Lemma 14.5.1 with $\varepsilon_\chi^2 = p\exp(M_0)$ (assuming that $p \le \exp(-M_0)$) and $t \lesssim 1$, we obtain

$$\mathsf{FI}(\mu Q_t \,\|\, \pi) \lesssim \frac{p\exp(M_0)\,(d + \log(1/p) - M_0)}{t} + dt\,.$$

We set $t \lesssim \varepsilon^2/d$ so that

$$\mathsf{FI}(\mu Q_t \,\|\, \pi) \lesssim \frac{d^2\exp(M_0)\,p\log(1/p)}{\varepsilon^2} + \varepsilon^2\,.$$

In order to make the first term at most $\varepsilon^2/2$, we take $p = \widetilde{\Theta}(\varepsilon^4/(d^2\exp(M_0)))$. In turn, this is satisfied provided

$$N \ge \frac{Z_{\mu_0}}{Z_\pi}\log\frac{1}{p} \asymp \exp(3M_0)\log\frac{d^2\exp(M_0)}{\varepsilon^4}\,,$$

which proves the desired result. □

Hence, under the stronger oracle model, $\log(1/\varepsilon)$ rates are indeed possible (albeit with exponential dependence on M_0). To see why this does not contradict the lower bound construction of Theorem 14.4.2, observe that if we take the initialization oracle to be π_{init}, then our construction satisfies $M_0 = r^2\phi(0)$. By inspecting the proof of Theorem 14.4.2, one sees that $r \asymp \sqrt{\log(1/\varepsilon)}$. Hence, our construction does not provide a lower bound for $\mathscr{C}_\infty(1, M_0, \varepsilon)$ for constant M_0. Instead, we obtain the following lower bound.

Corollary 14.8.2 (Lower bound for the stronger initialization oracle)**.** *There exists a universal constant $c > 0$ such that for all $\varepsilon \le 1/c$, it holds that*

$$\mathscr{C}_\infty\big(1, c\log(1/\varepsilon), \varepsilon\big) \gtrsim \frac{1}{\varepsilon\sqrt{\log(1/\varepsilon)}}\,.$$

Note also the following corollary.

Corollary 14.8.3 (High-accuracy Fisher information requires exponential dependence on M_0)**.** *Suppose that there exists an algorithm which works within the stronger oracle model and which, for any 1-log-smooth distribution π on $\mathbb{R}$, outputs a measure μ with $\sqrt{\mathsf{FI}(\mu \parallel \pi)} \le \varepsilon$ using N queries, with query complexity*

$$N \le f(M_0)\,\mathrm{polylog}\big(\frac{1}{\varepsilon}\big)$$

for some increasing function $f : [1, \infty) \to \mathbb{R}_+$. Then, there is a universal constant $c' > 0$ such that

$$f(M_0) \ge \widetilde{\Omega}\big(\exp(c' M_0)\big)\,.$$

Proof. Using Corollary 14.8.2 with $M_0 = c\log(1/\varepsilon)$, we have

$$f\big(c\log\frac{1}{\varepsilon}\big)\,\mathrm{polylog}\big(\frac{1}{\varepsilon}\big) \ge N \gtrsim \frac{1}{\varepsilon\sqrt{\log(1/\varepsilon)}}\,,$$

or

$$f\big(c\log\frac{1}{\varepsilon}\big) \ge \frac{1}{\varepsilon\,\mathrm{polylog}(1/\varepsilon)}\,.$$

Writing this in terms of $M_0 = c\log(1/\varepsilon)$, or $\varepsilon = \exp(-M_0/c)$,

$$f(M_0) \ge \frac{\exp(M_0/c)}{(M_0/c)^{O(1)}} = \widetilde{\Omega}\Big(\exp\big(\frac{M_0}{c}\big)\Big)$$

which establishes the result. □

Hence, we see that there is a fundamental trade-off in the stronger oracle model: any algorithm must either incur polynomial dependence on $1/\varepsilon$ (e.g., averaged LMC), or exponential dependence on M_0 (e.g., rejection sampling, see Proposition 14.8.1).

The stronger oracle model is strictly stronger. We also observe the following consequence of these observations. On one hand, our lower bound in Thoerem 14.4.2 shows that

$$\mathscr{C}(1, K_0 = 1, \varepsilon) \geq \Omega\Big(\frac{1}{\varepsilon\sqrt{\log(1/\varepsilon)}}\Big)\,.$$

On the other hand, for constant M_0, rejection sampling (Proposition 14.8.1) yields

$$\mathscr{C}_\infty(1, M_0, \varepsilon) \leq \widetilde{O}\Big(\exp(3M_0)\log\frac{1}{\varepsilon}\Big)\,.$$

Hence, the stronger oracle model is indeed stronger: *obtaining Fisher information guarantees is strictly easier with access to an oracle with bounded M_0, rather than an oracle with bounded K_0.*

On the effect of the radius of the effective support. In our lower bound construction, the distributions are "effectively" supported on a ball of radius R, where R scales with $1/\varepsilon$. Here, we show that this is in fact necessary, by showing that for any *fixed* d and R, it is possible to sample from such a distribution in Fisher information using $O(\log(1/\varepsilon))$ queries. The algorithm involves uses a simple grid search.

Proposition 14.8.4 (Sampling from bounded effective support)**.** *Suppose that the target distribution $\pi \propto \exp(-V)$ on $\mathbb{R}^d$ has the following properties:*

1. *$V(0) = 0$.*
2. *$V(x) = \frac{1}{2}\left(\|x\| - R\right)_+^2$, for $\|x\| \geq R$.*
3. *V is 1-smooth.*

Then, there is an algorithm which outputs μ with $\sqrt{\mathsf{FI}(\mu \parallel \pi)} \leq \varepsilon$ using N queries to $(V, \nabla V)$, where the number of queries satisfies

$$N \lesssim (cR)^d + \log\frac{\sqrt{d}}{\varepsilon}\,,$$

where $c > 0$ is a universal constant.

Proof. We use function approximation to build an upper envelope for $\tilde{\pi} \coloneqq \exp(-V)$, and then apply rejection sampling. Namely, let $\mathscr{N}$ be a 1-net of B_R, and for each $x \in B_R$ let $x_{\mathscr{N}}$ denote a closest point of $\mathscr{N}$ to x. Define the approximation

$$\widehat{V}(x) \coloneqq \begin{cases} \frac{1}{2}\left(\|x\| - R\right)_+^2\,, & \|x\| \geq R\,, \\ V(x_{\mathscr{N}}) + \langle \nabla V(x_{\mathscr{N}}), x - x_{\mathscr{N}}\rangle - \frac{1}{2}\|x - x_{\mathscr{N}}\|^2\,, & \|x\| < R\,. \end{cases}$$

By 1-smoothness of V, we have $V \geq \widehat{V}$, so that if we let $\tilde{\mu}_0 \coloneqq \exp(-\widehat{V})$, then $\tilde{\mu}_0 \geq \tilde{\pi}$. Also, for $\|x\| < R$, we have the bound

$$\begin{aligned}\tilde{\mu}_0(x) &= \exp\Big(-V(x_{\mathscr{N}}) - \langle \nabla V(x_{\mathscr{N}}), x - x_{\mathscr{N}} \rangle + \frac{1}{2}\|x - x_{\mathscr{N}}\|^2\Big) \\ &\leq \exp\big(-V(x) + \|x - x_{\mathscr{N}}\|^2\big) = \tilde{\pi}(x)\exp(\|x - x_{\mathscr{N}}\|^2) \leq \exp(1)\,\tilde{\pi}(x)\,,\end{aligned}$$

so that $Z_{\mu_0}/Z_\pi \lesssim 1$. We now perform rejection sampling using N' iterations with upper envelope $\tilde{\mu}_0$, outputting a sample from μ_0 if N' iterations are exceeded. Tracing through the proof of Proposition 14.8.1, one can show that for the output μ of rejection sampling, it holds that $\mathsf{FI}(\mu Q_t \parallel \pi) \leq \varepsilon^2$ for an appropriate choice of t. Moreover, the number of iterations of rejection sampling required to achieve this satisfies $N' \lesssim \log(\sqrt{d}/\varepsilon)$. Finally, since $|\mathscr{N}| \leq (cR)^d$ for a universal constant $c > 0$, it requires $O((cR)^d)$ queries in order to build the upper envelope $\tilde{\mu}_0$, which proves the result. □

To summarize the situation, if the effective radius R is known and fixed, then it is possible to obtain $O(\log(1/\varepsilon))$ complexity. However, if there is no a priori upper bound on the radius R, then the lower bounds of Theorem 14.4.2 and Corollary 14.8.2 apply.

■ 14.9 Conclusion

In this work, we have provided the first lower bounds for the query complexity of obtaining Fisher information guarantees for sampling. Due to the scarcity of general sampling lower bounds, our bounds are in fact some of the *only* known lower bounds for sampling. Our results have a number of interesting implications, which we discussed thoroughly in previous sections, and they advance our understanding of the fundamental task of non-log-concave sampling.

To conclude, we highlight a few problems left open in our work. Most notably, our lower bound in Theorem 14.4.1 does not match the upper bound of averaged LMC, and it is an important question to close this gap. We also note that our lower bounds in Theorems 14.4.1 and 14.4.2 do not capture the dependence of K_0, and this is also left for future work.

Part IV

Other applications of Wasserstein gradient flows

Chapter 15

Bures–Wasserstein barycenters

We study first-order methods to compute the barycenter of Gaussian distributions with respect to the optimal transport metric. We derive global rates of convergence for both gradient descent and stochastic gradient descent despite the fact that the barycenter functional is not geodesically convex. Our analysis overcomes this technical hurdle by developing a Polyak–Łojasiewicz (PL) inequality, which is built using tools from optimal transport and metric geometry.

This chapter is based on [Che+20f; Alt+21], joint with Jason M. Altschuler, Patrik R. Gerber, Tyler Maunu, Philippe Rigollet, and Austin J. Stromme.

15.1 Introduction

Averaging multiple data sources is among the most classical and fundamental subroutines in data science. However, a modern challenge is that data is often more complicated than points in $\mathbb{R}^d$. In this chapter, we study the task of averaging probability distributions on $\mathbb{R}^d$, a setting that commonly arises in computer vision and graphics [Rab+12; Sol+15], machine learning and statistics [CD14; Ho+17; SLD18; Dog+19], probability theory [KS94; RU02], and signal processing [Elv+20]; see also the surveys [PC19; PZ19] and the references within.

Namely, consider the following statistical problem. We observe n independent copies $\mu_1, \dots, \mu_n$ of a probability measure μ over $\mathbb{R}^d$. Assume furthermore that $\mu \sim P$, where P is an unknown distribution over probability measures. We wish to output a single probability measure on $\mathbb{R}^d$, $\bar{\mu}_n$, which represents the *average* measure under P in a suitable sense. For example, the measures $\mu_1, \dots, \mu_n$ may arise as representations of images, in which case the average of the measures with respect to the natural linear structure on the space of signed measures is unsuitable for many applications [CD14]. Instead, we study the *Wasserstein barycenter* [AC11], also known as a *Fréchet mean*, which has been proposed in the literature as a more desirable notion of average because it incorporates the geometry of the underlying space.

To formally set up the situation, let $\mathcal{P}_2(\mathbb{R}^d)$ be the set of all (Borel) probability measures on $\mathbb{R}^d$ with finite second moment, and let $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ be the subset of those measures in $\mathcal{P}_2(\mathbb{R}^d)$ that are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and thus admit a density. When endowed with the *2-Wasserstein metric*, W_2, this set forms a geodesic metric space $(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d), W_2)$. Throughout this chapter, we assume that P is a distribution over measures that is supported on a subset of $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ that consists only of certain multivariate Gaussian measures. We denote by P_n the empirical distribution of the sample $\mu_1, \ldots, \mu_n$.

A *barycenter* of P, denoted $b^\star$, is defined to be a minimizer of the functional

$$F(b) \coloneqq \frac{1}{2} P W_2^2(b, \cdot) = \frac{1}{2} \int W_2^2(b, \cdot) \, \mathrm{d}P \, .$$

A natural estimator of $b^\star$ is the *empirical barycenter* $\hat{b}_n$, defined as a minimizer of

$$F_n(b) \coloneqq \frac{1}{2} P_n W_2^2(b, \cdot) = \frac{1}{2n} \sum_{i=1}^{n} W_2^2(b, \mu_i) \, .$$

The many applications of Wasserstein barycenters (see, e.g., [CE10; Rab+12; CD14; GPC15; RP15; Sol+15; BPC16; SLD18; LLR20]) have inspired significant research into their mathematical and statistical properties since their introduction roughly a decade ago [AC11; Rab+12]. For instance, on the mathematical side it is known that under mild conditions, the barycenter exists, is unique, and admits a dual formulation related to a multi-marginal optimal transport problem [CE10; AC11; COO15].

Statistical consistency of the empirical barycenter in a general context was first established in [LL17] and further work has focused on providing effective rates of convergence for the quantity $W_2^2(\hat{b}_n, b^\star)$. A first step towards this goal was made in [ALP20] by deriving non-parametric rates of the form $W_2^2(\hat{b}_n, b^\star) \lesssim n^{-1/d}$ when $d \geq 3$. Moreover, in the same paper [ALP20], the authors establish parametric rates of the form $W_2^2(\hat{b}_n, b^\star) \lesssim n^{-1}$ when P is supported on a space of finite doubling dimension. An important example with this property arises when P is supported on mean-zero non-degenerate Gaussian measures. In this case, the Gaussians can be identified with their covariance matrices, and the Wasserstein metric induces a distance metric on the space of positive definite matrices. This distance metric, known as the *Bures metric* (or the *Bures–Wasserstein metric* when measured between probability measures), is equivalent to a Riemannian metric on the manifold of positive definite matrices, and the resulting Riemannian structure is known as the *Bures manifold* [Mod17; BJL19]. The name of the Bures manifold originates from quantum physics and quantum information theory, where it is used to model the space of density matrices [Bur69]. In fact, in

the Bures case, more precise statistical results, including central limit theorems, are known [AC17; KSS21]. It is worth noting that parametric rates are also achievable in the infinite-dimensional case under additional conditions. First, it is not surprising that such rates are achievable over $(\mathcal{P}_2(\mathbb{R}), W_2)$ since this space can be isometrically embedded in a Hilbert space [PZ16; Big+18]. Moreover, it was shown that, under additional regularity conditions, such rates are achievable for much more general infinite-dimensional spaces [Le +22], including $(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d), W_2)$ for any $d \geq 2$.

While these results for the empirical barycenter are satisfying from a statistical perspective, computing this object is challenging because of two fundamental obstacles. The first is that in general, barycenters can be complicated distributions which are much harder to represent (even approximately) than the input distributions. The second is that generically, these problems are computationally hard in high dimensions. For instance, Wasserstein barycenters and geometric medians of discrete distributions are NP-hard to compute (even approximately) in high dimension [AB22].

Algorithms for averaging on the Bures–Wasserstein manifold. Nevertheless, these computational obstacles can be potentially averted in parametric settings. This work as well as most of the literature [Álv+16; ZP19; Bac+22] on parametric settings focuses on the commonly arising setting where P is supported on Gaussian distributions.[1] As noted in [Álv+16], the Gaussian case also encompasses general location-scatter families.

There are two natural families of approaches for designing averaging algorithms in this setting. Both rely on iterative, first-order algorithms [CD14; Álv+16; CCS18; ZP19; Bac+22], exploiting the fact that modulo a simple re-centering of all distributions, the relevant space of probability distributions is isometric to the *Bures–Wasserstein manifold*, i.e., the cone of positive semidefinite matrices equipped with the Bures–Wasserstein metric (background is given in §2.3).

The first approach is simply to recognize the (regularized) Wasserstein barycenter problem as a convex optimization problem over the space of positive semidefinite matrices and apply off-the-shelf methods such as Euclidean gradient descent or semidefinite programming solvers. However, these methods have received little prior attention for good reason: they suffer from severe scalability and parameter-tuning issues (see §15.5.2 for numerics). Briefly, the underlying issue is that these algorithms operate in the standard Euclidean geometry rather than the natural geometry of optimal transport.

A much more effective approach in practice is to exploit the geometry of the

[1] In the setting of Gaussian distributions, the Wasserstein barycenter was first studied in the 1990s [OR93; KS94].

Bures–Wasserstein manifold via geodesic optimization. This approach is supported by the influential work of Otto [Ott01] who established that the geometry of Wasserstein space bears resemblance to a Riemannian manifold. In particular, one can define the gradient of the functional F, so it does indeed make sense to consider a *gradient descent*-based approach towards estimating $b^\star$. In the population setting (where the distribution P is known), such an algorithm was proposed in Álvarez-Esteban et al. [Álv+16], where it was introduced as a fixed-point algorithm. Álvarez-Esteban et al. prove that the fixed-point algorithm converges to the true barycenter as the number of iterations goes to infinity. The consistency results were further generalized in [ZP19; Bac+22] and extended to the non-population and stochastic gradient case. However, the literature previously did not provide any rates of convergence for these first-order methods. In fact, Álvarez-Esteban et al. empirically observed a linear rate of convergence for the gradient descent algorithm in the Gaussian setting and left open the theoretical study of this phenomenon for future study. One contribution of this chapter is to establish a dimension-free rate of convergence (Theorem 15.4.1), and we also provide multiple extensions including the first rate of convergence for stochastic gradient descent in this context.

Challenges for geodesic optimization over the Bures–Wasserstein manifold. Although geodesic optimization is natural for this problem, it comes with several important obstacles: the non-negative curvature of the Bures–Wasserstein manifold necessitates new tools for analysis, and moreover the barycenter problem is *non-convex* in the Bures–Wasserstein geometry. (These two issues are in fact intimately related, see §15.6.) This prevents applying standard results in the geodesic optimization literature (see, e.g., [ZS16; Bou23]) since in general it is only possible to prove local convergence guarantees for non-convex problems.

For the Wasserstein barycenter problem, it is possible to interpret Riemannian gradient descent (with step size one) as a fixed-point iteration, and through this lens establish asymptotic convergence [Álv+16; ZP19; Bac+22]. Obtaining non-asymptotic rates of convergence is more challenging because it requires developing quantitative proxies for the standard convexity inequalities needed to analyze gradient descent.

■ 15.1.1 Techniques

Here we briefly sketch the specific technical challenges we face and how we address them to analyze Riemannian gradient descent for the Bures–Wasserstein barycenter problem.

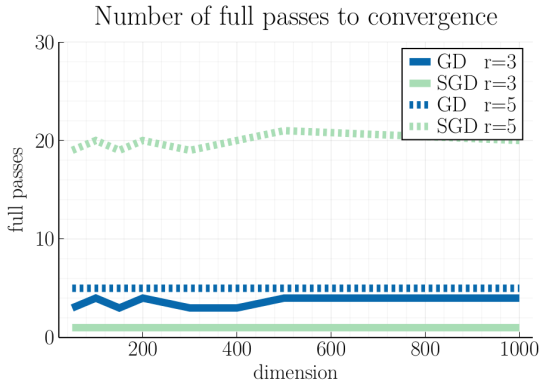


Figure 15.1: Passes until convergence error 10^{-r} to the barycenter, for $r \in \{3, 5\}$. This is *dimension independent* for Riemannian GD and SGD—consistent with our main results. Details in §15.4.

Overcoming non-convexity. As we discuss in §15.6, there is a close connection between the second-order behavior of these objective functionals and the non-negative curvature of the Bures–Wasserstein manifold. In particular, while non-negative curvature is used to prove smoothness properties for the three functionals, it also leads to them all being geodesically non-convex. To circumvent this issue, we establish gradient domination conditions, also known as Polyak–Łojasiewicz inequalities [KNS16], which intuitively are quantitative proxies for strong convexity in the non-convex setting. Proving such inequalities requires synthesizing general optimization principles with specialized arguments based on the theory of optimal transport. We ultimately show that these inequalities hold with constants depending on the conditioning of the iterates, i.e., the ratio between the maximum and minimum eigenvalues of the corresponding covariance matrices.

Overcoming ill-conditioned iterates. So long as smoothness and gradient domination inequalities hold at the current iterate, standard optimization results guarantee that the next iterate of gradient descent makes progress. However, the amount of progress degrades if the iterates are poorly conditioned. Thus the second major obstacle is to control the regularity of the iterates. Here, the primary technical tool is shared across the analyses. Informally, it states that if the objective is a sum of functions, each of whose gradients point towards well-conditioned matrices, then the gradient descent iterates remain well-conditioned. Formally, this is captured by the following geometric result, which may be of independent interest. Below, $\mathbf{S}^d_{++}$ denotes the set of $d \times d$ positive definite matrices. See §2.1 and §2.3 for a review of the relevant geometric concepts, and see §15.8 for the proof, discussion of tightness, and complementary results.

Theorem 15.1.1. *Let $0 < \alpha \le \beta < \infty$. Let Q be any distribution over $\mathbf{S}^d_{++}$ with*

$$\left(\int \sqrt{\lambda_{\min}(\Sigma)}\,\mathrm{d}Q(\Sigma)\right)^2 \ge \alpha \qquad \text{and} \qquad \int \lambda_{\max}(\Sigma)\,\mathrm{d}Q(\Sigma) \le \beta\,.$$

Then, for any matrix Σ_0 with eigenvalues bounded below by $\frac{\alpha}{4}$ and any $0 \le \eta \le \frac{\alpha}{2\beta}$, the generalized barycenter of $(1-\eta)\,\delta_{\Sigma_0} + \eta\, Q$ at Σ_0 also has eigenvalues lower bounded by $\frac{\alpha}{4}$.

Using this theorem together with careful analysis of the objective functions, we establish global convergence guarantees for first-order geodesic optimization.

In an earlier version of [Alt+21], we incorrectly claimed that $-\sqrt{\lambda_{\min}}$ and $\sqrt{\lambda_{\max}}$ are convex along generalized geodesics, which is stronger than Theorem 15.1.1. Here, we fix this issue; see Remark 15.8.3 for a detailed discussion.

15.1.2 Other related work

Averaging on curved spaces. Barycenters on curved spaces have become popular due to the applications in brain-computer interfaces [CBB17; YBL17], computer vision, machine learning, and radar signal processing [ABY13]. While their mathematical properties such as existence and uniqueness are fairly well-understood [Afs11], their computation is an active area of research [VZ00; Stu03; Vaz09; Yan10; BI13; Bač14a; OP15]. For the Wasserstein barycenter problem, there have been a multitude of approaches proposed for both the discrete setting (see, e.g., [CD14; Ben+15; COO15; Kro+19; Lin+20; AB21; Gum+21; Haa+21; Bor22; Dvi22; Lin+22]) and the continuous setting (see, e.g., [Li+20; CAD21; FTC21; Kor+21]).

15.2 Preliminaries

We write $\mathbf{S}^d$ for the space of symmetric $d \times d$ matrices, $\mathbf{S}^d_{++}$ for the open subset of $\mathbf{S}^d$ consisting of positive definite matrices, and $\mathbf{S}^d_+$ for the set of positive semidefinite matrices. We denote by $\lambda_1(\Sigma), \dots, \lambda_d(\Sigma) \ge 0$ the eigenvalues of a matrix $\Sigma \in \mathbf{S}^d_+$. The Gaussian measure on $\mathbb{R}^d$ with mean $m \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbf{S}^d_+$ is denoted $\gamma_{m,\Sigma}$. We reserve the notation log for the inverse of the Riemannian exponential map and use instead $\ln(\cdot)$ to denote the natural logarithm. The (convex analysis) indicator function $\iota_{\mathcal{C}}$ of a set $\mathcal{C}$ is defined by $\iota_{\mathcal{C}}(x) = 0$ if $x \in \mathcal{C}$ and $\iota_{\mathcal{C}}(x) = +\infty$ otherwise. We denote by id the identity map of $\mathbb{R}^d$.

Given probability measures μ and ν on $\mathbb{R}^d$ with finite second moment, the 2-Wasserstein distance between μ and ν is defined as

$$W_2^2(\mu,\nu) \coloneqq \inf_{\pi \in \mathcal{C}(\mu,\nu)} \int \|x-y\|^2\,\mathrm{d}\pi(x,y)\,, \tag{15.1}$$

where $\mathcal{C}(\mu, \nu)$ denotes the set of couplings of μ and ν, i.e., the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are respectively μ and ν. If μ and ν admit densities with respect to the Lebesgue measure on $\mathbb{R}^d$, then the infimum is attained, and the optimal coupling is supported on the graph of a map, i.e., there exists a map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that for π-a.e. $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, it holds that $y = T(x)$. The map T is called the *optimal transport map* from μ to ν.

We refer readers to [Vil03; San15] for an introduction to optimal transport, and to [Car92] and §2 for background on Riemannian geometry. The Riemannian structure of optimal transport was introduced in the seminal work [Ott01]; detailed treatments can be found in [AGS08; Vil09b], see also §2 for a quick overview.

In this chapter, we mainly work with centered Gaussians, which can be identified with their covariance matrices. (Extensions to the non-centered case are also discussed in the next sections.) We abuse notation via this identification: given $\Sigma, \Sigma' \in \mathbf{S}^d_{++}$, we write $W_2(\Sigma, \Sigma')$ for the 2-Wasserstein distance between centered Gaussians with covariance matrices Σ, Σ' respectively. Throughout, all Gaussians of interest are non-degenerate; that is, their covariances are non-singular.

The Wasserstein distance has a closed-form expression for Gaussians:

$$W_2^2(\Sigma, \Sigma') = \operatorname{tr}\left[\Sigma + \Sigma' - 2\left(\Sigma^{1/2}\Sigma'\Sigma^{1/2}\right)^{1/2}\right]. \tag{15.2}$$

Also, the optimal transport map from Σ to Σ' is the symmetric matrix

$$T_{\Sigma \to \Sigma'} = \Sigma^{-1/2}\left(\Sigma^{1/2}\Sigma'\Sigma^{1/2}\right)^{1/2}\Sigma^{-1/2} = \mathrm{GM}(\Sigma^{-1}, \Sigma'). \tag{15.3}$$

Above, $\mathrm{GM}(A, B) := A^{1/2}\left(A^{-1/2}BA^{-1/2}\right)^{1/2}A^{1/2}$ denotes the matrix geometric mean between two positive semidefinite matrices [Bha07, §4]. The Wasserstein distance on $\mathbf{S}^d_{++}$ in fact arises from a Riemannian metric, which was first introduced by Bures in [Bur69]. Hence, the Riemannian manifold $\mathbf{S}^d_{++}$ endowed with this Wasserstein distance is referred to as the *Bures–Wasserstein space*. The geometry of this space is studied in detail in [Mod17; BJL19]. For completeness, we provide additional background on the Bures–Wasserstein manifold in §2.3.

■ 15.3 General results for Wasserstein barycenters

In this section, we develop a general machinery to study first-order methods for optimizing the barycenter functional on Wasserstein space. Establishing fast convergence of first-order methods is intimately related to convexity. Since our setting is the curved Wasserstein space, we consider *geodesic convexity* rather than the usual convexity employed in flat, Euclidean spaces. Geodesic convexity has been used to study statistical efficiency in manifold-constrained estimation [AMR05; Wie12] and, more recently, optimization [Bon13; Bač14b; ZS16].

Barring a direct approach to establishing quantitative convergence guarantees, the barycenter functional is actually not geodesically convex on the Wasserstein space. In fact, the barycenter functional may even be *concave* along geodesics; see Figure 15.2. As such, it does not lend itself to the general techniques of geodesically convex optimization. This non-convexity is a manifestation of the non-negative curvature of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ [AGS08, §7.3].

Fortunately, the optimization literature describes conditions for global convergence of first order algorithms even for non-convex objectives. In this work, we employ a Polyak–Łojasiewicz (PL) inequality of the form (15.5), which is known to yield linear convergence for a variety of gradient methods on flat spaces even in absence of convexity [KNS16]. Theorems 15.3.1 and 15.3.2 below are proved using modifications of the usual proofs in the optimization literature. Their proofs make critical use of the non-negative curvature of the Wasserstein space and are deferred to §15.7.

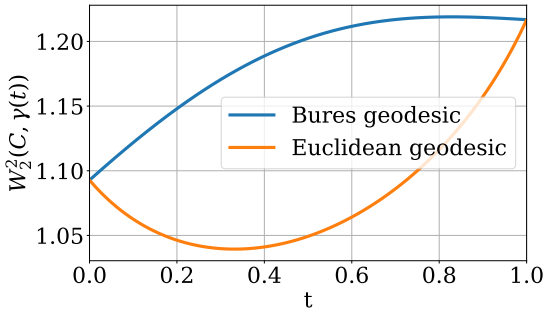


Figure 15.2: Example of the non-geodesic convexity of W_2^2. Displayed is the squared Bures distance along a Wasserstein geodesic and a Euclidean geodesic. Details are given in §15.5.4.

In this section, we study the barycenter functional

$$G(b) := \frac{1}{2} QW_2^2(b, \cdot) = \frac{1}{2} \int W_2^2(b, \cdot) \, \mathrm{d}Q \,, \tag{15.4}$$

for some generic distribution Q with barycenter $\bar{b}$. This notation allows us to treat simultaneously the cases where $Q = P$ and $Q = P_n$, which are the situations of interest for statisticians. The case when Q is an arbitrary discrete distribution supported on Gaussian measures has also been studied in the geodesic optimization literature: [AC11; Álv+16; BJL19; ZP19; WS22].

■ 15.3.1 Gradient descent algorithms over the Wasserstein space

We refer to §2.1 for background on optimal transport.

■ 15.3.1.1 Gradient descent

Let Q be a probability distribution over $(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d), W_2)$. In the sequel, we focus on the cases where $Q = P$, $Q = P_n$, or Q is a weighted atomic distribution, but our results apply generically to any Q that satisfy the conditions stated in the theorems below.

Using the techniques of [AGS08], the *gradient* of a barycenter functional G defined in (15.4) may be easily computed [ZP19]. It is given by the following map from $\mathbb{R}^d$ to $\mathbb{R}^d$:

$$\nabla G(b) \coloneqq -Q \log_b(\cdot) = -\int (T_{b\to\mu} - \mathrm{id})\, \mathrm{d}Q(\mu)\,.$$

Denote by $\bar{b}$ any minimizer of G.

The primary assumption we work with is common in the optimization literature. We say that G satisfies a *Polyak–Łojasiewicz (PL)* inequality at b if

$$\|\nabla G(b)\|_b^2 \geq 2C_{\mathsf{PL}}\,[G(b) - G(\bar{b})] \quad \text{for some } C_{\mathsf{PL}} > 0\,. \tag{15.5}$$

It follows from (15.12) below that $C_{\mathsf{PL}} \leq 1$ for any such Q.

The *gradient descent (GD)* iterates on G are defined as

$$b_0 \in \operatorname{supp} Q\,, \qquad b_{t+1} \coloneqq \exp_{b_t}\big(-\nabla G(b_t)\big) = [\mathrm{id} - \nabla G(b_t)]_{\#} b_t \quad \text{for } t \geq 1\,. \tag{15.6}$$

Note that this method employs a unit step size. This is in agreement with the observation made in [ZP19] that it leads to the maximum decrement in G.

We show that a PL inequality yields a linear rate of convergence.

Theorem 15.3.1 (Rate of convergence for gradient descent)**.** *If G satisfies the PL inequality (15.5) at all the iterates $(b_t)_{t<T}$, then*

$$G(b_T) - G(\bar{b}) \leq (1 - C_{\mathsf{PL}})^T\,[G(b_0) - G(\bar{b})]\,.$$

■ 15.3.1.2 Stochastic gradient descent

PL inequalities are also useful in the stochastic setting where we observe n independent copies $\mu_1, \ldots, \mu_n$ of $\mu \sim Q$. In this case, we consider the natural *stochastic gradient descent (SGD)* iterates defined by

$$\begin{aligned} b_0 &\coloneqq \mu_0\,, \\ b_{t+1} &\coloneqq \exp_{b_t}\big(-\eta_t \log_{b_t}(\mu_{t+1})\big) = [\mathrm{id} + \eta_t\,(T_{b_t\to\mu_{t+1}} - \mathrm{id})]_{\#} b_t \quad \text{for } t = 0, \ldots, n-1\,, \end{aligned} \tag{15.7}$$

where $\eta_t \in (0, 1)$ denotes the step size. At each iteration, SGD moves the iterate along the geodesic between b_t and μ_{t+1} by a distance η_t. Under the assumption of a PL inequality, we show that SGD achieves a parametric rate of convergence.

In the following result, we recall that the *variance* of Q is defined as

$$\operatorname{var}(Q) := \int W_2^2(\bar{b}, \cdot)\, \mathrm{d}Q = 2G(\bar{b})\,.$$

Theorem 15.3.2 (Rates of convergence for SGD)**.** *Assume that there exists a constant $C_{\mathsf{PL}} > 0$ such that the following holds: G satisfies the PL inequality (15.5) at all the iterates $(b_t)_{0 \le t \le n}$ of SGD run with step size*

$$\eta_t = C_{\mathsf{PL}} \left(1 - \sqrt{1 - \frac{2\,(t+k)+1}{C_{\mathsf{PL}}^2\,(t+k+1)^2}}\right) \le \frac{2}{C_{\mathsf{PL}}\,(t+k+1)}\,, \tag{15.8}$$

where we take $k = 2/C_{\mathsf{PL}}^2 - 1 \ge 0$. Then,

$$\mathbb{E}\, G(b_n) - G(\bar{b}) \le \frac{3 \operatorname{var}(Q)}{C_{\mathsf{PL}}^2 n}\,.$$

The parameter k in (15.8) ensures that the step size η_t is well-defined and less than 1.

15.3.2 Properties of the barycenter functional

Unlike results in generic optimization, this chapter focuses on a specific function to optimize: the barycenter functional. In fact, this is a vast family of functionals, each indexed by the distribution Q in (15.4). However, some structure is shared across this family. In the rest of this section, we extract properties that are relevant to our optimization questions: a variance inequality, smoothness, as well as an interated PL inequality. These properties are valid for general distributions Q over $\mathcal{P}_2(\mathbb{R}^d)$ and are specialized to the Bures manifold in the next section.

15.3.2.1 Variance inequality

Variance inequalities indicate quadratic growth of the barycenter functional around its minimum. More specifically, we say that Q satisfies a *variance inequality* with constant $C_{\mathsf{var}} > 0$ if

$$G(b) - G(\bar{b}) \ge \frac{C_{\mathsf{var}}}{2}\, W_2^2(b, \bar{b})\,, \qquad \forall b \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)\,. \tag{15.9}$$

In particular, (15.9) implies uniqueness of $\bar{b}$. The importance of variance inequalities for obtaining statistical rates of convergence for the empirical barycenter was

emphasized in [ALP20]. In [Le +22], it is shown that an assumption on the regularity of the transport maps from the barycenter $\bar{b}$ implies a variance inequality. Specifically, suppose that all of the Kantorovich potentials $\varphi_{\bar{b}\to\mu}$ for $\mu \in \operatorname{supp} Q$ are (α, β)-regular in the sense of (2.2). Then, a variance inequality holds with $C_{\mathsf{var}} = 1 - (\beta - \alpha)$.

It turns out that a variance inequality holds without needing to assume smoothness of $\varphi_{\bar{b}\to\mu}$: assuming that the potential $\phi_{\bar{b}\to\mu}$ is $(\alpha(\mu), \infty)$-regular for each $\mu \in \operatorname{supp} Q$ yields a variance inequality with $C_{\mathsf{var}} = \int \alpha(\mu)\, \mathrm{d}Q(\mu)$. The improvement here is critical for achieving global results on the Bures manifold. To formally state this result, we need the notion of an *optimal dual solution* for the barycenter problem. A discussion of this concept, along with a proof of the following theorem, is given in §15.7.2. We verify that the hypotheses of the theorem hold in the case when Q is supported on non-degenerate Gaussian measures in §15.9.1.

Theorem 15.3.3 (Variance inequality)**.** *Fix $Q \in \mathcal{P}_2(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d))$ be a distribution with barycenter $\bar{b} \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. Assume that there exists an optimal dual solution φ for the barycenter problem w.r.t. $\bar{b}$ such that, for Q-a.e. $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, the mapping φ_μ is $\alpha(\mu)$-strongly convex for some measurable function $\alpha : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}_+$. Then, Q satisfies a variance inequality (15.9) with constant*

$$C_{\mathsf{var}} = \int \alpha(\mu)\, \mathrm{d}Q(\mu)\,.$$

■ 15.3.2.2 Smoothness

Recall that a convex differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|y - x\|^2\,, \qquad \forall x, y \in \mathbb{R}^d\,. \tag{15.10}$$

A consequence of β-smoothness is the following inequality, which measures how much progress gradient descent makes in a single step [Bub15].

$$f\big(x - \beta^{-1}\nabla f(x)\big) - f(x) \le -\frac{1}{2\beta} \|\nabla f(x)\|^2\,. \tag{15.11}$$

In fact, only the latter inequality (15.11) is needed for the analysis of gradient descent methods. It was noted, first in [Álv+16, Proposition 3.3] and then in [ZP19, Lemma 2], that an analogue of (15.11) holds in Wasserstein space for the barycenter functional. Below, we provide a different, more geometric proof of this fact that emphasizes the collective role of smoothness and curvature. On the way, we also establish a smoothness inequality (15.12) that is used in the proof of Theorem 15.3.1 and also ensures that $C_{\mathsf{PL}} \le 1$ for any distribution Q supported on $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$.

Theorem 15.3.4. *For any $b_0, b_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ the barycenter functional satisfies the smoothness inequality*

$$G(b_1) \leq G(b_0) + \langle \nabla G(b_0), \log_{b_0} b_1 \rangle_{b_0} + \frac{1}{2} W_2^2(b_0, b_1)\,. \tag{15.12}$$

Moreover, for any $b \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ and $b^+ \coloneqq [\mathrm{id} - \nabla G(b)]_{\#} b$, it holds:

$$G(b^+) - G(b) \leq -\frac{1}{2} \|\nabla G(b)\|_b^2\,. \tag{15.13}$$

Proof. Let $(b_s)_{s\in[0,1]}$ be the constant-speed geodesic between arbitrary $b_0, b_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. From the non-negative curvature inequality (2.10), for any $s \in (0,1]$,

$$\begin{aligned}\int \frac{W_2^2(b_s,\mu) - W_2^2(b_0,\mu)}{s}\,\mathrm{d}Q(\mu) \\ \geq \int [W_2^2(b_1,\mu) - W_2^2(b_0,\mu)]\,\mathrm{d}Q(\mu) - (1-s)\,W_2^2(b_0,b_1)\,.\end{aligned}$$

By dominated convergence, the left-hand side converges to

$$\begin{aligned}\int \partial_s|_{s=0^+} W_2^2(b_s,\mu)\,\mathrm{d}Q(\mu) &= -2\int \langle T_{b_0\to\mu} - \mathrm{id}, T_{b_0\to b_1} - \mathrm{id}\rangle_{L_2(b_0)}\,\mathrm{d}Q(\mu) \\ &= 2\,\langle \nabla G(b_0), \log_{b_0}(b_1)\rangle_{b_0}\,,\end{aligned}$$

where in the first identity, we used the characterization of [AGS08, Proposition 7.3.6]. Rearranging terms yields (15.12).

Noticing that $W_2^2(b, b^+) = \|-\nabla G(b)\|_b^2$, Theorem 15.3.4 is now an immediate consequence of (15.12) applied to $b_0 = b$ and $b_1 = b^+$. □

■ 15.3.2.3 An integrated PL inequality

The main technical hurdle of this work is to provide sufficient conditions under which the PL inequality holds. The following lemma, proved in §15.7.3, is our main device to establish PL inequalities.

Lemma 15.3.5. *Let Q satisfy a variance inequality with constant C_{var} and let $b \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ be such that the barycenter $\bar{b}$ of Q is absolutely continuous w.r.t. b. Assume further the following measurability conditions: there exists a measurable mapping $\phi : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, $(\mu, x) \mapsto \phi_{b\to\mu}(x)$, such that, for Q-almost every $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, $\phi_{b\to\mu} : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a Kantorovich potential for the optimal transport from b to μ. Then,*

$$G(b) - G(\bar{b}) \leq \frac{2}{C_{\mathsf{var}}} \left(\int_0^1 \|\nabla G(b)\|_{L^2(b_s)}\,\mathrm{d}s \right)^2,$$

where $(b_s)_{s\in[0,1]}$ *is the constant-speed* W_2*-geodesic beginning at* $b_0 \coloneqq b$ *and ending at* $b_1 \coloneqq \bar{b}$.

This lemma can yield a PL inequality in quite general situations, but the crucial issue is whether these conditions hold uniformly for each iterate in the optimization trajectory. In the next section, we show how to turn an integrated PL inequality into a bona fide PL inequality when Q is supported on certain Gaussian measures.

15.4 Main results for Bures–Wasserstein barycenters

Identifying a centered non-degenerate Gaussian measure with its covariance matrix, the Wasserstein geometry induces a Riemannian structure on the space of positive definite matrices, known as the *Bures* geometry. Accordingly, we now refer to the barycenter of P as the *Bures–Wasserstein barycenter*:

$$\Sigma^\star \in \operatorname*{arg\,min}_{\Sigma\in\mathbf{S}^d_{++}} \int W_2^2(\Sigma,\cdot)\,\mathrm{d}P\,.$$

We refer to the introduction for a discussion of the past work on the Bures–Wasserstein barycenter. We also remark that the case when P is supported on possibly non-centered Gaussians is easily reduced to the centered case, as we discuss below.

15.4.1 Bures–Wasserstein gradient descent algorithms

We now specialize both GD and SGD when the distribution of interest is supported on mean-zero Gaussian measures. In this case, the updates of both algorithms take a remarkably simple form. To see this, for $m \in \mathbb{R}^d$, $\Sigma \in \mathbb{S}^D_+$, let $\gamma_{m,\Sigma}$ denote the Gaussian measure on $\mathbb{R}^d$ with mean m and covariance matrix Σ. The set of non-degenerate Gaussians constitutes a well-behaved subset of Wasserstein space, called the *Bures–Wasserstein* manifold [Bur69; BJL19]. In particular, the optimal coupling between γ_{m_0,Σ_0} and γ_{m_1,Σ_1} has the explicit form

$$x \mapsto T_{\gamma_{\mu_0,\Sigma_0}\to\gamma_{\mu_1,\Sigma_1}}(x) \coloneqq m_1 + \Sigma_0^{-1/2}(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\Sigma_0^{-1/2}\,(x-m_0)\,. \tag{15.14}$$

Observe that $T_{\gamma_{\mu_0,\Sigma_0}\to\gamma_{\mu_1,\Sigma_1}}$ is affine, and thus $\int T_{\gamma_{\mu_0,\Sigma_0}\to\gamma}\,\mathrm{d}P(\gamma)$ is affine.

This means that all of the GD (or SGD) iterates are Gaussian measures, so it suffices to keep track of the mean and covariance matrix of the current iterate. For both GD and SGD, the update equation for the descent step decomposes into two decoupled equations: an update equation for the mean, and an update

equation for the covariance matrix. Moreover, the update equation for the mean is trivial, corresponding to a simple GD or SGD procedure on the objective function $m \mapsto \int \|m - m(\mu)\|^2 \, \mathrm{d}P(\mu)$. Therefore, for simplicity and without loss of generality, we consider only mean-zero Gaussians throughout this work and we simply have to write down the update equations for the covariance matrix Σ_t of the iterate. They are summarized in Algorithms 15.1 and 15.2 below.

GD is useful for computing high-precision solutions due to its linear convergence (Theorem 15.4.1), and SGD is useful for large-scale or online settings because of its cheaper updates. Here, Σ_0 is the initialization, which can be taken to be any matrix in the support of P. For SGD, we also require a sequence $(\eta_t)_{t=1}^T$ of step sizes and a sequence $(K_t)_{t=1}^T$ of i.i.d. samples from P.

Algorithm 15.1 GD for Barycenters

procedure BARY-GD(Σ_0, η, P, T)
  for $t = 1, \dots, T$ **do**
    $S_t \leftarrow (1 - \eta)\, I_d + \eta \int \mathrm{GM}(\Sigma_{t-1}^{-1}, \Sigma) \, \mathrm{d}P(\Sigma)$
    $\Sigma_t \leftarrow S_t \Sigma_{t-1} S_t$
  return Σ_T

Algorithm 15.2 SGD for Barycenters

procedure BARY-SGD($\Sigma_0, (\eta_t)_{t=1}^T, (K_t)_{t=1}^T$)
  for $t = 1, \dots, T$ **do**
    $\hat{S}_t \leftarrow (1 - \eta_t)\, I_d + \eta_t \, \mathrm{GM}(\Sigma_{t-1}^{-1}, K_t)$
    $\Sigma_t \leftarrow \hat{S}_t \Sigma_{t-1} \hat{S}_t$
  return Σ_T

Note that whereas SGD requires choosing step sizes, for GD we can simply use step size 1 in practice, as justified in [ZP19]. However, for our theoretical results, we will require choosing a step size $\eta < 1$ for GD as well.

■ 15.4.2 Convergence guarantees

Denote the barycenter functional by $F(\Sigma) \coloneqq \frac{1}{2} \int W_2^2(\Sigma, \cdot) \, \mathrm{d}P$, and denote the *variance* of P by $\operatorname{var} P \coloneqq 2F(\Sigma^\star)$. We assume that P is supported on matrices whose eigenvalues lie in the range $[\lambda_{\min}, \lambda_{\max}]$, and we let $\kappa \coloneqq \lambda_{\max}/\lambda_{\min}$ denote the condition number.

Theorem 15.4.1. *Assume that P is supported on covariance matrices whose eigenvalues lie in the range $[\lambda_{\min}, \lambda_{\max}]$, $0 < \lambda_{\min} \leq \lambda_{\max} < \infty$. Let $\kappa \coloneqq \lambda_{\max}/\lambda_{\min}$ denote the condition number. Assume that we initialize at $\Sigma_0 \in \operatorname{supp} P$.*

1. *(GD) Let Σ_T^{GD} denote the T-th iterate of GD (Algorithm 15.1) with step size $\eta = \frac{1}{2\kappa}$. Then,*

$$\frac{1}{2\sqrt{\kappa}} W_2^2(\Sigma_T^{\mathrm{GD}}, \Sigma^\star) \leq F(\Sigma_T^{\mathrm{GD}}) - F(\Sigma^\star) \leq \exp\Big(-\frac{3T}{64\kappa^{5/2}}\Big) \{F(\Sigma_0) - F(\Sigma^\star)\}.$$

2. *(SGD) Let Σ_T^{SGD} denote the T-th iterate of SGD (Algorithm 15.2). Then, with appropriately chosen step sizes,*

$$\frac{1}{2\sqrt{\kappa}} \mathbb{E}\, W_2^2(\Sigma_T^{\mathrm{SGD}}, \Sigma^\star) \leq \mathbb{E}\, F(\Sigma_T^{\mathrm{SGD}}) - F(\Sigma^\star) \leq \frac{48\kappa^3 \operatorname{var} P}{T}.$$

We now elaborate on implications this theorem for both GD and for SGD.

For GD, Theorem 15.4.1 establishes a linear rate of convergence and answers a question left open in [Álv+16]. Moreover, when applied with P being the empirical measure and combined with the existing results of [ALP20; KSS21], it yields a procedure to estimate Wasserstein barycenters at the parametric rate after a number of iterations that is logarithmic in the sample size n.

For SGD, Theorem 15.4.1 shows that online SGD applied with P being the population measure yields an estimator Σ_n^{SGD} different from the empirical barycenter that also converges at the parametric rate to the true barycenter of P. When applied with P being the empirical measure, this leads to an alternative to gradient descent to estimate the empirical barycenter that exhibits a slower convergence but that has much cheaper iterations and better lends itself to parallelization.

As far as we are aware, these results provide the first non-asymptotic rates of convergence for first-order methods on the Bures–Wasserstein manifold.

In Figure 15.3, we present the results an experiment confirming these two results; see §15.5 for more details and further numerical results.

In fact, using Theorem 15.1.1 we can also relax the conditioning assumption to an *average-case* notion of conditioning. This is a significant improvement when the eigenvalue ranges differ significantly between matrices.

Theorem 15.4.2. *Define the quantities*

$$\|\lambda_{\min}\|_{1/2} \coloneqq \Big(\int \sqrt{\lambda_{\min}(\Sigma)}\, \mathrm{d}P(\Sigma)\Big)^2,$$
$$\|\lambda_{\max}\|_{1/2} \coloneqq \Big(\int \sqrt{\lambda_{\max}(\Sigma)}\, \mathrm{d}P(\Sigma)\Big)^2,$$

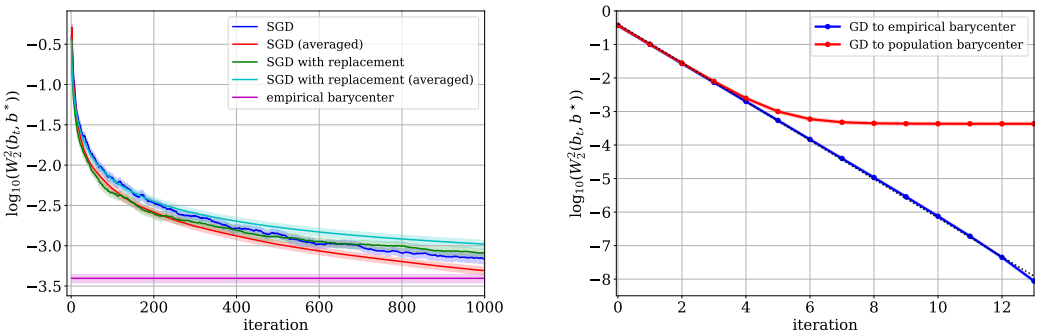


Figure 15.3: Left: convergence of SGD on Bures manifold for $n = 1000$, $d = 3$, and $b^\star = \gamma_{0,I_3}$. Right: linear convergence of GD on the same problem.

$$\|\lambda_{\max}\|_1 \coloneqq \int \lambda_{\max}(\Sigma)\, \mathrm{d}P(\Sigma)\,.$$

Then, the conclusions of Theorem 15.4.1 hold when replacing κ with the quantity $\|\lambda_{\max}\|_{1/2}/\|\lambda_{\min}\|_{1/2}$ everywhere for SGD, or when replacing κ with the quantity $\|\lambda_{\max}\|_1/\|\lambda_{\min}\|_{1/2}$ everywhere for GD. In particular, the conclusions for SGD hold when replacing κ with $\kappa^\star \coloneqq \sup_{\Sigma\in\mathrm{supp}(P)} \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$ everywhere.

We give the proof of this result in §15.9.3.

We also refer to [Alt+21] for results on computing *entropically-regularized Wasserstein barycenters* [Kro18; BCP19; CEK21] and geometric medians on the Bures–Wasserstein space.

■ 15.4.3 Outline of the proof

For simplicity, we assume that the Gaussians are centered (see previous discussion). While the centering assumption can be made without loss of generality, our results require that P is supported on well-conditioned matrices. Under this condition, it can be shown that the barycenter of P exists and is unique (Proposition 15.9.1).

We begin with a brief outline of the proof.

(i) If we initialize gradient descent (or stochastic gradient descent) at one of the elements of the support of P, then all of the iterates, all of the elements of $\mathrm{supp}\, P$, the barycenter of P, and all of elements of geodesics between these measures are well-conditioned Gaussians.

(ii) Using Lemma 15.3.5, we establish a PL inequality holds with a uniform constant for well-conditioned Gaussians.

(iii) The guarantees for GD and SGD on the Bures manifold follow immediately from the PL inequality and our general convergence results (Theorems 15.3.1, 15.3.2).

In the sequel, we use *geodesic convexity* as a key tool to control the iterates of the gradient descent algorithm. We note that this discussion is not about proving some sort of geodesic convexity for our objective, which cannot hold in general. Our main interest in geodesic convexity comes from the following fact: if all of the elements of the support of P lie in a geodesically convex set $\mathcal{S}$, and we initialize the algorithm at an element of $\mathcal{S}$, then all of the iterates of stochastic gradient descent are simply moving along geodesics within this set, and so remain in $\mathcal{S}$. The same is true for the iterates of gradient descent, provided that we replace geodesic convexity with *convexity along generalized geodesics*. Refer to §2.1 for definitions of these terms. We begin with the following fact.

Lemma 15.4.3. *For a measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, let $M(\mu) \coloneqq \int x \otimes x \,\mathrm{d}\mu(x)$. Then, the functional $\mu \mapsto \|M(\mu)\|_{\mathrm{op}} = \lambda_{\max}(M(\mu))$ is convex along generalized geodesics on $\mathcal{P}_2(\mathbb{R}^d)$.*

Proof. Let $\mathbb{S}^{d-1}$ denote the unit sphere of $\mathbb{R}^d$ and observe that for any $e \in \mathbb{S}^{d-1}$ the function $x \mapsto \langle x, e\rangle^2$ is convex on $\mathbb{R}^d$. By known results for geodesic convexity in Wasserstein space (see [AGS08, Proposition 9.3.2]), the functional $\mu \mapsto \int \langle \cdot, e\rangle^2 \,\mathrm{d}\mu = \langle e, M(\mu)\, e\rangle$ is convex along generalized geodesics in $\mathcal{P}_2(\mathbb{R}^d)$; hence, so is the functional $\mu \mapsto \max_{e \in \mathbb{S}^{D-1}} \langle e, M(\mu)\, e\rangle = \|M(\mu)\|_{\mathrm{op}}$. □

It follows readily from Lemma 15.4.3 that the set

$$\mathcal{S} \coloneqq \{\gamma_{0,\Sigma} \mid \Sigma \in \mathbf{S}^d_{++},\ \|\Sigma\|_{\mathrm{op}} \le \lambda_{\max}\}$$

is convex along generalized geodesics. Moreover since SGD moves along geodesics and is initialized at $b_0 \in \operatorname{supp} P \subset \mathcal{S}$, then all the iterates of SGD stay in $\mathcal{S}$. To show that the same holds for GD, observe that the set $\log_{b_t}(\mathcal{S})$ is convex. Therefore, $-\operatorname{grad} F(b_t) = \int (T_{b_t \to \mu} - \mathrm{id})\,\mathrm{d}P(\mu) \in \log_{b_t}(\mathcal{S})$ as a convex combination of elements in this set. This is equivalent to $b_{t+1} = \exp_{b_t}(-\operatorname{grad} F(b_t)) \in \mathcal{S}$. These observations control the maximum eigenvalue along GD and SGD.

To control the minimum eigenvalue, we can establish that

$$\mathcal{S}' \coloneqq \{\gamma_{0,\Sigma} \mid \Sigma \in \mathbf{S}^d_{++},\ \lambda_{\min}(\Sigma) \ge \lambda_{\min}\}$$

is geodesically convex (Theorem 15.8.1); however, this set is *not* convex along generalized geodesics. To analyze GD, we therefore appeal instead to Theorem 15.1.1.

This completes the first step (i) of the proof. Moving on to step (ii), we get from Theorem 15.9.4 that F satisfies a PL inequality at all well-conditioned Gaussians and in particular at all the iterates of both GD and SGD.

Combined with the general bounds in Theorems 15.3.1 and 15.3.2 and the variance inequality in Theorem 15.9.3, this completes the proof of Theorem 15.4.1.

15.5 Experiments

In this section, we demonstrate the linear convergence of GD, the fast rate of estimation for SGD, and some potential advantages of averaging SGD by way of numerical experiments. In evaluating SGD, we also include a variant that involves sampling with replacement from the empirical distribution.

15.5.1 Comparisons with averaging and SGD with replacement

First, we begin by illustrating how SGD indeed achieves the fast rate of convergence to the true barycenter on the Bures manifold, as indicated by Theorem 15.4.1.

To generate distributions with a known barycenter, we use the following fact. If the mean of the distribution $(\log_{b^\star})_{\#}P$ is 0, then $b^\star$ is a barycenter of P. This fact follows from our PL inequality (Theorem 15.9.4) or also from general arguments in [ZP19, Theorem 2]. We also use the fact that the tangent space of the Bures manifold is given by the set of all symmetric matrices [BJL19].

Figure 15.3 shows convergence of SGD for distributions on the Bures manifold. To generate a sample, we let A_i be a matrix with i.i.d. γ_{0,σ^2} entries. Our random sample on the Bures manifold is then given by

$$\Sigma_i = \exp_{\gamma_{0,I_d}}\Big(\frac{A_i + A_i^\mathsf{T}}{2}\Big)\,, \tag{15.15}$$

which has population barycenter $b^\star = \gamma_{0,I_d}$. An explicit form of this exponential map is derived in [MMP18]. We run two versions of SGD. The first variant uses each sample only once, and passes over the data once. The second variant samples from $\Sigma_1, \dots, \Sigma_n$ with replacement at each iteration and takes the stochastic gradient step towards the selected matrix. For the resulting sequences, we also show the results of averaging the iterates. Specifically, if $(b_t)_{t\in\mathbb{N}}$ is the sequence generated by SGD, then the averaged sequence is given by $\tilde{b}_0 = b_0$ and

$$\tilde{b}_{t+1} = \Big[\frac{t}{t+1}\,\mathrm{id} + \frac{1}{t+1}\,T_{\tilde{b}_t \to b_{t+1}}\Big]_{\#}\tilde{b}_t\,.$$

On Riemannian manifolds, averaged SGD is known to attain optimal statistical rates under smoothness and geodesic convexity assumptions [Tri+18].

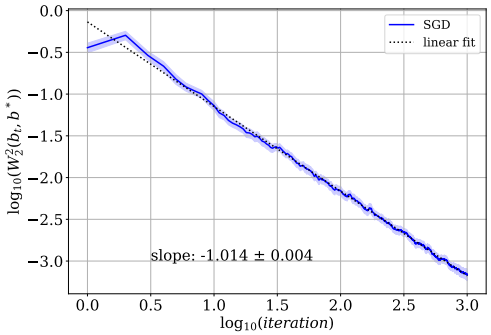


Figure 15.4: Log-log plot of convergence for SGD on Bures manifold for $n = 1000$, $d = 3$, and $b^\star = \gamma_{0,I_3}$. This corresponds to the experiment on the left in Figure 15.3.

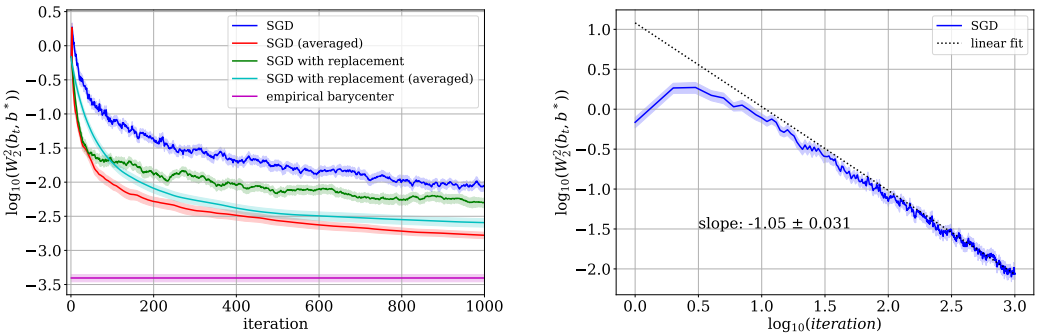


Figure 15.5: Convergence of SGD on Bures manifold. Here, $n = 1000$, $d = 3$, and barycenter given by $\mathrm{diag}(20, 1, 1)$. The result displays the average over 100 randomly generated datasets.

Here, we generate 100 datasets of size $n = 1000$ in the way specified above and set $\sigma^2 = 0.25$. In this experiment, the SGD step size is chosen to be $\eta_t = 2/[0.7 \cdot (t + 2/0.7 + 1)]$. The results from these 100 datasets are then averaged for each algorithm, and we also display 95% confidence bands for the resulting sequences. As is clear from the log-log plot in Figure 15.4, SGD achieves the fast $O(n^{-1})$ statistical rate on this dataset.

The right of Figure 15.3 shows convergence of GD to the empirical barycenter and true barycenter. We generate samples in the same way as before. This linear convergence was observed previously by [Álv+16].

In Figure 15.5, we repeat the same experiment, except this time the barycenter

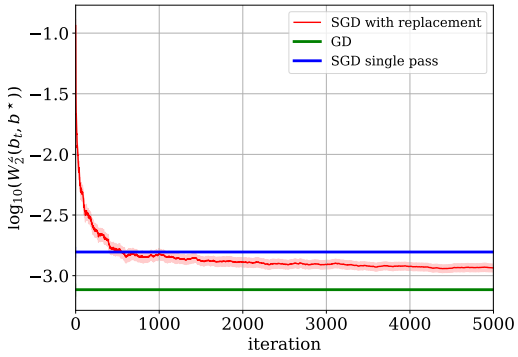


Figure 15.6: Convergence of SGD on Bures manifold. Here, $n = 500$, $d = 3$, and the distribution is given by (15.15) with $\Sigma^\star = I_3$ and $\sigma^2 = 0.25$. The result displays the average over 100 randomly generated datasets.

has covariance matrix

$$\Sigma^\star = \begin{bmatrix} 20 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and the entries of A_i are drawn i.i.d. from $\gamma_{0,1}$. In this situation, the condition numbers of the matrices generated according to this distribution are typically much larger than those centered around γ_{0,I_3}. To account for a potentially smaller PL constant, we chose $\eta_t = 2/[0.1 \cdot (t + 2/0.1 + 1)]$. It is again clear from the right pane in Figure 15.5 that SGD achieves the fast $O(n^{-1})$ statistical rate on this dataset. To account for the slow convergence initially, we only fit this line to the last 500 iterations. We also note that averaging yields drastically better performance in this case, which we are currently unable to theoretically justify.

Figure 15.6 shows convergence of SGD with replacement to the empirical barycenter. We generate $n = 500$ samples in the same way as in Figure 15.3, where the true barycenter is I_3 and $\sigma^2 = 0.25$. We calculate the error obtained by the empirical barycenter by running GD on this dataset until convergence, which is displayed with the green line. We also calculate the error obtained by a single pass of SGD, which is given by the blue line. SGD with replacement is then run for 5000 iterations, and we observe that it does indeed achieve better error than single pass SGD if run for long enough. SGD with replacement converges to the empirical barycenter, albeit at a slow rate.

15.5.2 Comparison with other algorithms

There are two natural competitors of Riemannian GD when minimizing the barycenter functional: (*i*) solving an SDP (§15.9.5), and (*ii*) Euclidean GD (see §15.9.4 for a description of the Euclidean gradient descent algorithm).

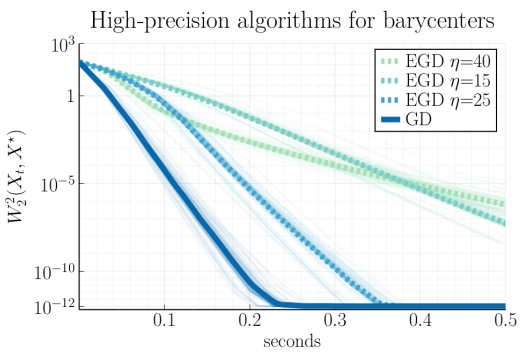


Figure 15.7: Riemannian vs. Euclidean GD.

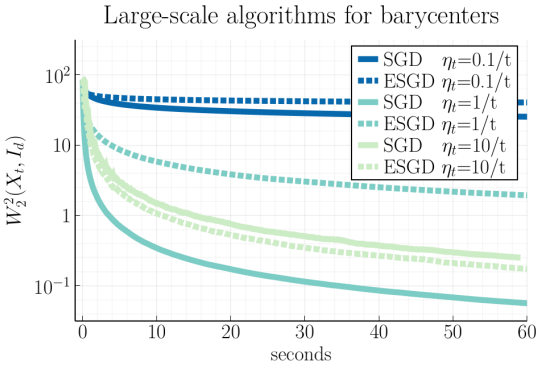


Figure 15.8: Riemannian vs. Euclidean SGD.

In Figure 15.7 we compare Riemannian and Euclidean GD on a random dataset consisting of $n = 50$ covariance matrices of dimension $d = 50$, each with condition number $\kappa = 1000$. The eigenspaces of the matrices are independent Haar distributed, and their eigenvalues are equally spaced in the interval $[\lambda_{\min}, \lambda_{\max}] = [0.03, 30]$. Qualitatively similar results are observed for other input distributions; see §15.5.3. We run 50 experiments and plot the average accuracy cut off at 10^{-12}; $X^\star$ denotes the best iterate. We omit SDP solvers from the plot because their runtime is orders of magnitude slower for this problem: using the Splitting Cone Solver (SCS) [ODo+16], the problem takes ∼15 seconds to solve, and MOSEK is even slower. We observe that Euclidean GD's rate of convergence is very sensitive to its step size, which depends heavily on the conditioning of the problem. Riemannian GD was the clear winner in our experiments, as its step size requires no tuning and it always performed no worse (in fact, often significantly better) than Euclidean GD.

In Figure 15.8 we compare Riemannian and Euclidean SGD. We average 300×300 covariance matrices drawn from a distribution whose barycenter is known to be the identity, see §15.5.3 for details. We observe that Riemannian SGD typically outperforms Euclidean SGD, sometimes substantially.

We comment on Figure 15.1, which illustrates the dimension independence of the two Riemannian algorithms, a main result of this work. It plots the number of passes until convergence $W_2^2(X_t, X^\star) \leq 10^{-r} \operatorname{var} P$ to the barycenter $X^\star$, for $r \in \{3, 5\}$. To compare the algorithms on equal footing, the y-axis measures "full

passes" over the $n = 50$ matrices: one pass constitutes one iteration of GD, or n iterations of SGD. We generate the input dataset just as in Figure 15.7. Observe also the tradeoff between GD and SGD: SGD converges rapidly to low-precision solutions, but takes longer to converge to high-precision solutions.

■ 15.5.3 Further experiments and details

Reproducibility details. Input generation details for Figures 15.1 and 15.7 are provided above. For Figure 15.8, recall that we generated matrices from a distribution whose barycenter is known to be the identity. By [ZP19, Theorem 2], if the mean of the distribution $(\log_{I_d})_{\#}P$ is 0, then I_d is the barycenter of P. In particular, if Q is a mean zero distribution supported on symmetric matrices that lie in the domain of the exponential map, then $P = (\exp_{I_d})_{\#}Q$ has I_d as its barycenter. In our experiments, we defined Q to be the law of a random matrix with Haar eigenbasis and uniform eigenvalues from the interval $[-(1-\delta), 1-\delta]$ for a parameter $\delta \in (0, 1)$. At the identity, the exponential map takes the simple form $\exp_{I_d} S = (I_d + S)^2$ and we see that P is then supported on covariance matrices with spectrum in $[\delta^2, (2-\delta)^2]$. The figure was generated with $\delta = 0.1$. All experiments were performed using Julia 1.5.1 on a desktop computer running Ubuntu 18.04 with an Intel i7-10700 CPU.

Further empirical comparisons. Here we further investigate the comparison of Riemannian and Euclidean GD done in Figure 15.7 by demonstrating qualitatively similar results for a variety of synthetic datasets. For each dataset, the measure P is the empirical measure of n matrices of dimension $d \times d$ that are drawn randomly as follows.

1. Haar eigenbasis and linearly spaced eigenvalues in $[\alpha, \beta]$.
2. Haar eigenbasis and i.i.d. $\mathsf{uniform}([\alpha, \beta])$ eigenvalues.
3. First split the matrices into 3 groups. Each matrix has Haar eigenbasis and i.i.d. $\mathsf{uniform}([\alpha, \beta])$ eigenvalues where $[\alpha, \beta] = 10^i \times [1, \kappa]$ for $i \in \{-2, 0, 2\}$ depending on its group.
4. Same as method 2 above, except all matrices have the same eigenbasis. (Note that GD converges in 1 step here since the matrices commute.)
5. Haar eigenbasis and eigenvalues uniform on a set of size $m \leq d$, whose elements are i.i.d. $\mathsf{uniform}([\alpha, \beta])$.
6. Same as method 5 above, except all matrices use the same eigenvalues.

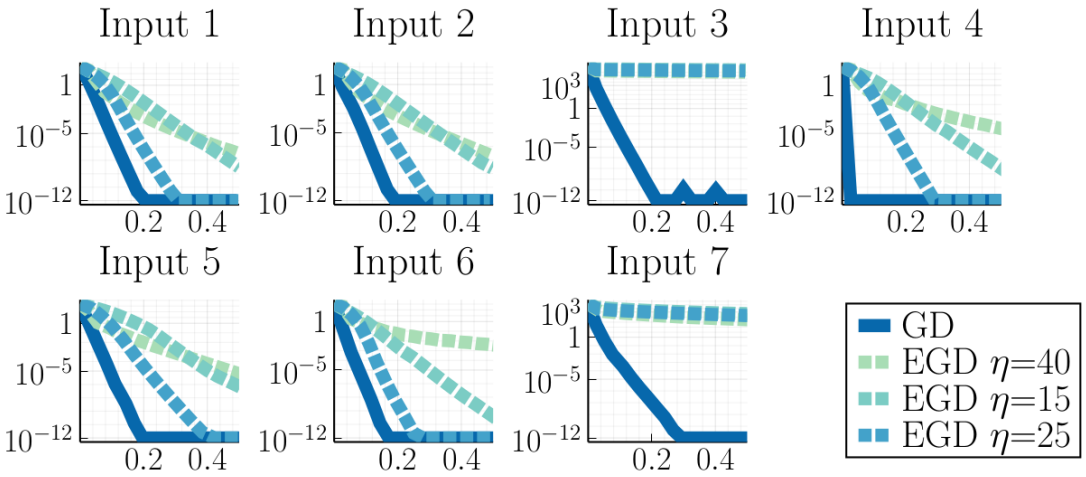


Figure 15.9: Comparison of high-precision barycenter algorithms for various types of synthetic data. Here, the matrices are poorly conditioned ($[\alpha, \beta] = [0.03, 30]$ whereby $\kappa = 1000$).

7. Mix of all methods above.

Figures 15.9 and 15.10 compare Euclidean and Riemannian GD on the barycenter problem as in Figure 15.7, but now with these 7 different input families. We average well-conditioned matrices in Figure 15.9, and ill-conditioned matrices in Figure 15.10. The plots are generated using $n = d = 50$ and $m = d/4$. For Method 7, the 50 matrices are divided into 6 groups of roughly equal size. The y-axis measures the W_2^2 distance to the best iterate; and the x-axis measures time in seconds.

In these figures we had to hand-tune the step size for Euclidean GD since the theoretical step size performs quite poorly. We used the same range of step sizes ($\eta \in \{15, 25, 40\}$) in all plots to demonstrate that the performance of Euclidean GD is quite sensitive to its step size. In contrast, GD performs well on all inputs with its (untuned) step size of 1.

15.5.4 Details of the non-convexity example

We consider the example of the Wasserstein metric restricted to centered Gaussian measures, which induces the Bures metric on positive definite matrices. Even restricted to such Gaussian measures, the Wasserstein barycenter objective is geodesically non-convex, despite the fact that it is Euclidean convex [WS22]. Figure 15.2 gives a simulated example of this fact. This figure plots the Bures distance squared between a positive definite matrix C and points along some

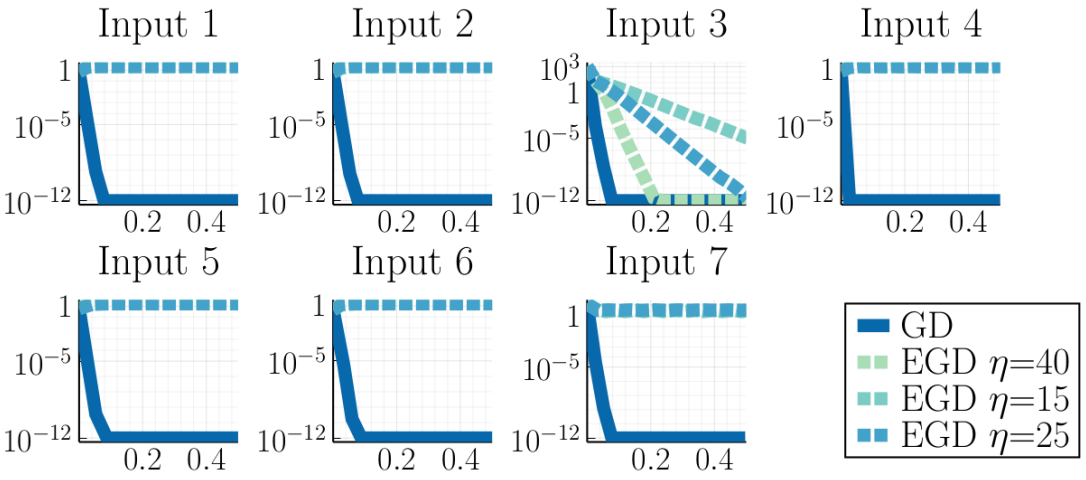


Figure 15.10: Comparison of high-precision barycenter algorithms for various types of synthetic data. Here, the matrices are well-conditioned ($[\alpha, \beta] = [1, 2]$ whereby $\kappa = 2$).

geodesic γ, which runs between two matrices A and B. The matrices used in this example are

$$A = \begin{bmatrix} 0.8 & -0.4 \\ -0.4 & 0.3 \end{bmatrix}, \qquad B = \begin{bmatrix} 0.3 & -0.5 \\ -0.5 & 1.0 \end{bmatrix}, \qquad C = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.6 \end{bmatrix},$$

and $\gamma(t)$, $t \in [0, 1]$, is taken to be the Bures or Euclidean geodesic from A to B (the Euclidean geodesic is given by $t \mapsto (1 - t)\, A + t\, B$). This function is clearly non-convex, and therefore we cannot assume that there is some underlying strong convexity (although the Bures distance is in fact strongly geodesically convex for sufficiently small balls [HGA15]).

■ 15.6 Curvature and the barycenter functional

One of the interesting features of the barycenter problem is that, because it is defined in terms of the squared distance function, it captures key geometric features of the underlying space; in fact, this is arguably the reason for the success of the barycenter for geometric applications. To further discuss this connection, it is insightful to abstract the situation to computing barycenters on a metric space.

Given a metric space (X, d) and a probability measure P on X, a barycenter of P is a solution of

$$\underset{b \in X}{\text{minimize}} \qquad F_P(b) := \frac{1}{2} \int d^2(b, \cdot)\, \mathrm{d}P\,.$$

The basic structure required on X in order to study first-order optimization methods is the presence of geodesics. This is formalized by the notion of a *geodesic space*, which is studied in metric geometry; see [BBI01]. Then, we may define a function $F : X \to \mathbb{R}$ to be *α-strongly convex* if for all geodesics $(x_t)_{t\in[0,1]}$ in X, it holds that

$$F(x_t) \leq (1-t)\, F(x_0) + t\, F(x_1) - \frac{\alpha\, t\,(1-t)}{2}\, d^2(x_0, x_1)\,, \qquad \text{for all } t \in [0,1]\,.$$

It is known that the convexity properties of the barycenter functional F_P are related to the *curvature* of the space. Here, curvature is interpreted as the *Alexandrov curvature*, which is the generalization of sectional curvature to geodesic spaces, see [BBI01]. Then, the result is that F_P is 1-strongly convex for every probability measure P on X if and only if X has *non-positive curvature*; see [Stu03] for precise statements. In fact, the 1-strong convexity of barycenter functionals is essentially the definition of non-positive curvature in this context.

Consequently, strong results are known for barycenters in non-positively curved spaces, ranging from basic properties such as existence and uniqueness, to statistical estimation and optimization; for details see the nice article [Stu03].

In contrast, it is well-known that Wasserstein space $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ (and hence, the Bures–Wasserstein space) is *non-negatively curved* [AGS08, Theorem 7.3.2]. This means, for instance, that convexity and properties related to convexity (such as the PL inequality employed in §15.9.3) are not automatic for the barycenter functional in Wasserstein space. On the other hand, we showed that this non-negative curvature is related to the *smoothness* of the barycenter functional.

15.7 Proofs for general Wasserstein barycenters

15.7.1 Convergence bounds for GD and SGD under a PL inequality

This subsection gives proofs of the general convergence theorems for GD and SGD in the present work. Both of these proofs use the non-negative curvature inequality (2.11). We note that the proof of Theorem 15.3.1 uses the non-negative curvature implicitly by invoking smoothness, while the use of non-negative curvature is explicit within the proof of Theorem 15.3.2.

15.7.1.1 Proof of Theorem 15.3.1 for GD

Using the smoothness (15.13) and the PL inequality (15.5), it holds that

$$G(b_{t+1}) - G(b_t) \leq -C_{\mathsf{PL}}\,[G(b_t) - G(\bar{b})]\,.$$

It yields $G(b_{t+1}) - G(\bar{b}) \leq (1 - C_{\mathsf{PL}})\,[G(b_t) - G(\bar{b})]$, which gives the result.

■ 15.7.1.2 Proof of Theorem 15.3.2 for SGD

Recall the SGD iterations on $n+1$ observations:

$$b_0 \coloneqq \mu_0\,, \qquad b_{t+1} \coloneqq [(1-\eta_t)\,\mathrm{id} + \eta_t\, T_{b_t \to \mu_{t+1}}]_{\#} b_t \quad \text{for } t = 0, \dots, n\,,$$

where the step size is given by

$$\eta_t = C_{\mathsf{PL}} \left(1 - \sqrt{1 - \frac{2\,(t+k)+1}{C_{\mathsf{PL}}^2\,(t+k+1)^2}}\right) \leq \frac{2}{C_{\mathsf{PL}}\,(t+k+1)}\,,$$

for some k such that $C_{\mathsf{PL}}^2\,(k+1)^2 \geq 2k+1$. We note that the step size η_t is chosen to solve the equation

$$1 - 2C_{\mathsf{PL}}\eta_t + \eta_t^2 = \left(\frac{t+k}{t+k+1}\right)^2.$$

Using the non-negative curvature (2.11), we get

$$\begin{aligned} W_2^2(b_{t+1}, \mu) &\leq \|\log_{b_t} b_{t+1} - \log_{b_t} \mu\|_{b_t}^2 = \|\eta_t \log_{b_t} \mu_{t+1} - \log_{b_t} \mu\|_{b_t}^2 \\ &= \|\log_{b_t} \mu\|_{b_t}^2 + \eta_t^2\, \|\log_{b_t} \mu_{t+1}\|_{b_t}^2 - 2\eta_t\, \langle \log_{b_t} \mu, \log_{b_t} \mu_{t+1} \rangle_{b_t}\,. \end{aligned}$$

Taking the expectation with respect to $(\mu, \mu_{t+1}) \sim Q^{\otimes 2}$ (conditioning appropriately on the increasing sequence of σ-fields), we have

$$\mathbb{E}\, G(b_{t+1}) \leq \mathbb{E}[(1+\eta_t^2)\, G(b_t) - \eta_t\, \|\nabla G(b_t)\|_{L^2(b_t)}^2]\,.$$

Using the PL inequality (15.5),

$$\mathbb{E}\, G(b_{t+1}) \leq \mathbb{E}\big[(1+\eta_t^2)\, G(b_t) - 2C_{\mathsf{PL}}\eta_t\, [G(b_t) - G(\bar{b})]\big]\,.$$

Subtracting $G(\bar{b})$ and rearranging,

$$\mathbb{E}\, G(b_{t+1}) - G(\bar{b}) \leq (1 - 2C_{\mathsf{PL}}\eta_t + \eta_t^2)\, [\mathbb{E}\, G(b_t) - G(\bar{b})] + \frac{\eta_t^2}{2}\, \mathrm{var}(Q)\,,$$

where we recall that $\mathrm{var}(Q) = 2G(\bar{b})$. With the chosen step size, we find

$$\mathbb{E}\, G(b_{t+1}) - G(\bar{b}) \leq \left(\frac{t+k}{t+k+1}\right)^2 [\mathbb{E}\, G(b_t) - G(\bar{b})] + \frac{2\,\mathrm{var}(Q)}{C_{\mathsf{PL}}^2\,(t+k+1)^2}\,.$$

Or equivalently,

$$(t+k+1)^2\, [\mathbb{E}\, G(b_{t+1}) - G(\bar{b})] \leq (t+k)^2\, [\mathbb{E}\, G(b_t) - G(\bar{b})] + \frac{2\,\mathrm{var}(Q)}{C_{\mathsf{PL}}^2}\,.$$

Unrolling over $t = 0, 1, \dots, n-1$ yields

$$(n+k)^2 \left[\mathbb{E}\, G(b_n) - G(\bar{b})\right] \leq k^2 \left[\mathbb{E}\, G(b_0) - G(\bar{b})\right] + \frac{2n \operatorname{var}(Q)}{C_{\mathsf{PL}}^2}\,,$$

or, equivalently,

$$\mathbb{E}\, G(b_n) - G(\bar{b}) \leq \frac{k^2}{(n+k)^2} \left[\mathbb{E}\, G(b_0) - G(\bar{b})\right] + \frac{2 \operatorname{var}(Q)}{C_{\mathsf{PL}}^2\,(n+k)}\,. \tag{15.16}$$

To conclude the proof, recall that from (15.12), we have

$$G(b_0) - G(\bar{b}) \leq \frac{1}{2}\, W_2^2(b_0, \bar{b})\,.$$

Taking the expectation over $b_0 \sim Q$ we find

$$\mathbb{E}\, G(b_0) - G(\bar{b}) \leq G(\bar{b}) = \frac{1}{2} \operatorname{var}(Q)\,,$$

as claimed. Together with (15.16), it yields

$$\mathbb{E}\, G(b_n) - G(\bar{b}) \leq \frac{\operatorname{var}(Q)}{n+k} \left(\frac{k^2}{2\,(n+k)} + \frac{2}{C_{\mathsf{PL}}^2}\right) \leq \frac{\operatorname{var}(Q)}{n} \left(\frac{k+1}{2} + \frac{2}{C_{\mathsf{PL}}^2}\right).$$

Plugging in the value of k completes the proof.

■ 15.7.2 Variance inequality: Theorem 15.3.3

We begin this section with a review of Kantorovich duality, which we use to discuss the dual of the barycenter problem. Then, we present the proof of Theorem 15.3.3.

Given two measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and maps $f \in L^1(\mu)$, $g \in L^1(\nu)$ such that $f(x) + g(y) \geq \langle x, y \rangle$ for μ-a.e. $x \in \mathbb{R}^d$ and ν-a.e. $y \in \mathbb{R}^d$, it is easy to see that

$$\frac{1}{2}\, W_2^2(\mu, \nu) \geq \int \left(\frac{\|\cdot\|^2}{2} - f\right) \mathrm{d}\mu + \int \left(\frac{\|\cdot\|^2}{2} - g\right) \mathrm{d}\nu\,.$$

Kantorovich duality (see e.g. [Vil03]) says that equality holds for some pair $f = \varphi$, $g = \varphi^*$ where φ is a proper LSC convex function and φ^* denotes its convex conjugate, i.e.,

$$\frac{1}{2}\, W_2^2(\mu, \nu) = \int \left(\frac{\|\cdot\|^2}{2} - \varphi\right) \mathrm{d}\mu + \int \left(\frac{\|\cdot\|^2}{2} - \varphi^*\right) \mathrm{d}\nu\,.$$

The map φ is called a Kantorovich potential for (μ, ν).

Accordingly, given $\bar{b} \in \mathcal{P}_2(\mathbb{R}^d)$, we call a measurable mapping $\varphi : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to L^1(\bar{b})$, $\mu \mapsto \varphi_\mu$, an *optimal dual solution* for the barycenter problem if the following two conditions are met: (1) for Q-a.e. μ, the mapping φ_μ is a Kantorovich potential for $(\bar{b}, \mu)$; (2) it holds that

$$\int \Big(\frac{\|\cdot\|^2}{2} - \varphi_\mu \Big) \, \mathrm{d}Q(\mu) = 0 \,. \tag{15.17}$$

It is easily seen that these conditions imply that $\bar{b}$ is the barycenter of Q:

$$\begin{aligned} G(b) = \frac{1}{2} \int W_2^2(b, \cdot) \, \mathrm{d}Q &\geq \int \Big[\int \Big(\frac{\|\cdot\|^2}{2} - \varphi_\mu \Big) \, \mathrm{d}b + \int \Big(\frac{\|\cdot\|^2}{2} - \varphi_\mu^* \Big) \, \mathrm{d}\mu \Big] \, \mathrm{d}Q(\mu) \\ &= \iint \Big(\frac{\|\cdot\|^2}{2} - \varphi_\mu^* \Big) \, \mathrm{d}\mu \, \mathrm{d}Q(\mu) = \frac{1}{2} \int W_2^2(\bar{b}, \cdot) \, \mathrm{d}Q = G(\bar{b}) \,. \end{aligned}$$

The existence of an optimal dual solution for the barycenter problem is known in the finitely supported case [AC11], and existence can be shown for the general case under mild conditions. For completeness, we give a self-contained proof of the existence of an optimal dual solution in the case where Q is supported on Gaussian measures in §15.9.1.

Proof of Theorem 15.3.3. By the strong convexity assumption, it holds for Q-a.e. $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ and a.e. $x \in \mathbb{R}^d$,

$$\varphi_\mu^*(x) + \varphi_\mu(y) \geq \langle x, y \rangle + \frac{\alpha(\mu)}{2} \, \|y - \nabla \varphi_\mu^*(x)\|^2 \,,$$

which can be rearranged into

$$\|x - y\|^2 - \alpha(\mu) \, \|y - \nabla \varphi_\mu^*(x)\|^2 \geq \frac{\|x\|^2}{2} - \varphi_\mu^*(x) + \frac{\|y\|^2}{2} - \varphi_\mu(y) \,.$$

Integrating this w.r.t. the optimal transport plan γ_μ between μ and $b \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\begin{aligned} &\frac{1}{2} \Big(W_2^2(\mu, b) - \alpha(\mu) \int \|T_{\mu \to b} - T_{\mu \to \bar{b}}\|^2 \, \mathrm{d}\mu \Big) \\ &\qquad \geq \int \Big(\frac{\|\cdot\|^2}{2} - \varphi_\mu^* \Big) \, \mathrm{d}\mu + \int \Big(\frac{\|\cdot\|^2}{2} - \varphi_\mu \Big) \, \mathrm{d}b \,. \end{aligned}$$

Observe also that (2.11) implies $\|T_{\mu \to b} - T_{\mu \to \bar{b}}\|^2_{L^2(\mu)} \geq W_2^2(b, \bar{b})$. Integrating these inequalities with respect to Q yields

$$G(b) - \frac{1}{2} \Big(\int \alpha \, \mathrm{d}Q \Big) \, W_2^2(b, \bar{b})$$

$$\geq \int \Big[\int \Big(\frac{\|\cdot\|^2}{2} - \varphi_\mu^* \Big) \,\mathrm{d}\mu + \int \Big(\frac{\|\cdot\|^2}{2} - \varphi_\mu \Big) \,\mathrm{d}b \Big] \,\mathrm{d}Q(\mu)$$
$$= \iint \Big(\frac{\|\cdot\|^2}{2} - \varphi_\mu^* \Big) \,\mathrm{d}\mu \,\mathrm{d}Q(\mu) = G(\bar{b}) \,,$$

where in the last two identities, we used (15.17). It finishes the proof. □

■ 15.7.3 Integrated PL inequality

The following lemma appears in [LV09, Lemma A.1] in the case of Lipschitz functions. A minor modification of their proof allows to handle locally Lipschitz rather than only Lipschitz functions. We include the modified proof for completeness.

Lemma 15.7.1. *Let $(b_s)_{s\in[0,1]}$ be a Wasserstein geodesic in $\mathcal{P}_2(\mathbb{R}^d)$. Let $\Omega \subseteq \mathbb{R}^d$ be a convex open subset for which $b_0(\Omega) = b_1(\Omega) = 1$. Then, for any function $f : \mathbb{R}^d \to \mathbb{R}$ which is locally Lipschitz on Ω, it holds that*

$$\Big| \int f \,\mathrm{d}b_0 - \int f \,\mathrm{d}b_1 \Big| \leq W_2(b_0, b_1) \int_0^1 \|\nabla f\|_{L^2(b_s)} \,\mathrm{d}s \,.$$

Proof. According to [Vil09b, Corollary 7.22], there exists a probability measure Π on the space of constant-speed geodesics in $\mathbb{R}^d$ such that $\gamma \sim \Pi$ and b_s is the law of $\gamma(s)$. In particular, it yields

$$\int f \,\mathrm{d}b_0 - \int f \,\mathrm{d}b_1 = \int \big[f\big(\gamma(0)\big) - f\big(\gamma(1)\big) \big] \,\mathrm{d}\Pi(\gamma) \,.$$

We can cover the geodesic $(\gamma(s))_{s\in[0,1]}$ by finitely many open neighborhoods contained in Ω so that f is Lipschitz on each such neighborhood; thus, the mapping $t \mapsto f(\gamma(s))$ is Lipschitz and we may apply the fundamental theorem of calculus, the Fubini–Tonelli theorem, and Cauchy–Schwarz:

$$\begin{aligned}
\int f \,\mathrm{d}b_0 - \int f \,\mathrm{d}b_1 &= \int \int_0^1 \big\langle \nabla f\big(\gamma(s)\big), \dot{\gamma}(s) \big\rangle \,\mathrm{d}s \,\mathrm{d}\Pi(\gamma) \\
&\leq \int_0^1 \int \mathrm{length}(\gamma) \, \big\| \nabla f\big(\gamma(s)\big) \big\| \,\mathrm{d}\Pi(\gamma) \,\mathrm{d}s \\
&\leq \int_0^1 \Big(\int \mathrm{length}(\gamma)^2 \,\mathrm{d}\Pi(\gamma) \Big)^{1/2} \Big(\int \big\| \nabla f\big(\gamma(s)\big) \big\|^2 \,\mathrm{d}\Pi(\gamma) \Big)^{1/2} \mathrm{d}s \\
&= W_2(b_0, b_1) \int_0^1 \|\nabla f\|_{L^2(b_s)} \,\mathrm{d}s \,.
\end{aligned}$$

It yields the result. □

Proof of Lemma 15.3.5. By Kantorovich duality [Vil03],

$$\frac{1}{2}\, W_2^2(b,\mu) = \int \Big(\frac{\|\cdot\|^2}{2} - \phi_{\mu\to b}\Big)\,\mathrm{d}\mu + \int \Big(\frac{\|\cdot\|^2}{2} - \phi_{b\to\mu}\Big)\,\mathrm{d}b\,,$$
$$\frac{1}{2}\, W_2^2(\bar{b},\mu) \geq \int \Big(\frac{\|\cdot\|^2}{2} - \phi_{\mu\to b}\Big)\,\mathrm{d}\mu + \int \Big(\frac{\|\cdot\|^2}{2} - \phi_{b\to\mu}\Big)\,\mathrm{d}\bar{b}\,.$$

This yields the inequality

$$G(b) - G(\bar{b}) \leq \int \Big(\frac{\|\cdot\|^2}{2} - \int \phi_{b\to\mu}\,\mathrm{d}Q(\mu)\Big)\,\mathrm{d}(b-\bar{b})\,.$$

Let $\bar{\phi} \coloneqq \int \phi_{b\to\mu}\,\mathrm{d}Q(\mu)$; this is a proper LSC convex function $\mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$. We apply Lemma 15.7.1 with $\Omega = \operatorname{int}\operatorname{dom}\bar{\phi}$. Since $\bar{\phi}$ is locally Lipschitz on the interior of its domain and $\bar{b} \ll b$, then $b(\Omega) = \bar{b}(\Omega) = 1$, whence

$$\begin{aligned} G(b) - G(\bar{b}) &\leq W_2(b,\bar{b}) \int_0^1 \|\nabla\bar{\phi} - \mathrm{id}\|_{L^2(b_s)}\,\mathrm{d}s \\ &\leq \sqrt{\frac{2\,[G(b) - G(\bar{b})]}{C_{\mathsf{var}}}} \int_0^1 \|\nabla\bar{\phi} - \mathrm{id}\|_{L^2(b_s)}\,\mathrm{d}s\,. \end{aligned}$$

Square and rearrange to yield

$$G(b) - G(\bar{b}) \leq \frac{2}{C_{\mathsf{var}}} \Big(\int_0^1 \|\nabla\bar{\phi} - \mathrm{id}\|_{L^2(b_s)}\,\mathrm{d}s\Big)^2\,.$$

Recognizing that $\nabla G(b) = \mathrm{id} - \nabla\bar{\phi}$ yields the result. □

15.8 Proofs for the geodesic convexity results

15.8.1 Proof of Theorem 15.1.1

See §2.1 and 2.3 for background on the relevant geometric concepts.

We begin by proving that the functionals $-\sqrt{\lambda_{\min}}$ and $\sqrt{\lambda_{\max}}$ are geodesically convex. The following argument is implicit in the proofs of [AC11, Theorem 6.1] and [BJL19, Theorem 8], and we include it for completeness.

Theorem 15.8.1. *The functionals $-\sqrt{\lambda_{\min}} : \mathbf{S}^d_{++} \to \mathbb{R}$ and $\sqrt{\lambda_{\max}} : \mathbf{S}^d_{++} \to \mathbb{R}$ are convex along barycenters.*

Proof. If Q is a probability measure on $\mathbf{S}^d_{++}$ with barycenter $\Sigma^\star$, then

$$\Sigma^\star = \int \big(\Sigma^{\star\,1/2}\Sigma\Sigma^{\star\,1/2}\big)^{1/2}\,\mathrm{d}Q(\Sigma)\,,$$

see [AC11, Theorem 6.1]. This implies

$$\lambda_{\min}(\Sigma^\star) \geq \int \sqrt{\lambda_{\min}(\Sigma^{\star\,1/2}\Sigma\Sigma^{\star\,1/2})}\,\mathrm{d}Q(\Sigma) \geq \sqrt{\lambda_{\min}(\Sigma^\star)} \int \sqrt{\lambda_{\min}(\Sigma)}\,\mathrm{d}Q(\Sigma)\,,$$

whence

$$\sqrt{\lambda_{\min}(\Sigma^\star)} \geq \int \sqrt{\lambda_{\min}(\Sigma)}\,\mathrm{d}Q(\Sigma)\,.$$

A similar argument applies for $\sqrt{\lambda_{\max}}$. □

Remark 15.8.2. *This result implies for instance that the set of PSD matrices with eigenvalues lying in a certain range is geodesically convex.*

Since the update for Bures–Wasserstein SGD only involves moving along geodesics, the above result already suffices to control the eigenvalues of the SGD iterates. However, the update for Bures–Wasserstein GD entails movement along *generalized* geodesics, for which we need the control in Theorem 15.1.1.

Before proving Theorem 15.1.1, however, we provide some intuition for the proof. Denote by F the barycenter functional $F(\Sigma) \coloneqq \frac{1}{2}\int W_2^2(\Sigma, \cdot)\,\mathrm{d}Q$ corresponding to the measure Q and for the sake of intuition, pretend that $\sqrt{\lambda_{\min}}$ is differentiable everywhere. Let $(\Sigma_t)_{t\geq 0}$ denote the gradient flow of F, i.e., $\dot{\Sigma}_t = -\nabla F(\Sigma_t)$. We observe that the gradient of F can be written as an *average*, hence

$$\begin{aligned}\partial_t \sqrt{\lambda_{\min}(\Sigma_t)} &= -\langle \nabla \sqrt{\lambda_{\min}}(\Sigma_t), \nabla F(\Sigma_t)\rangle_{\Sigma_t} \\ &= \int \langle \nabla \sqrt{\lambda_{\min}}(\Sigma_t), \log_{\Sigma_t} \Sigma'\rangle_{\Sigma_t}\,\mathrm{d}Q(\Sigma')\,,\end{aligned}$$

see Fact 2 in §2.3.2. However, the geodesic concavity of $\sqrt{\lambda_{\min}}$ implies that

$$\sqrt{\lambda_{\min}(\Sigma')} \leq \sqrt{\lambda_{\min}(\Sigma_t)} + \langle \nabla \sqrt{\lambda_{\min}}(\Sigma_t), \log_{\Sigma_t} \Sigma'\rangle_{\Sigma_t}$$

and therefore

$$\partial_t \sqrt{\lambda_{\min}(\Sigma_t)} \geq \underbrace{\int \sqrt{\lambda_{\min}(\Sigma')}\,\mathrm{d}Q(\Sigma')}_{=:\sqrt{\alpha}} - \sqrt{\lambda_{\min}(\Sigma_t)}\,.$$

This shows that as soon as $\lambda_{\min}(\Sigma_t)$ hits α, then $\sqrt{\lambda_{\min}(\Sigma_t)}$ is increasing. Thus, the continuous-time gradient flow for F always has eigenvalues at least α provided that it is initialized appropriately and $\sqrt{\lambda_{\min}}$ is differentiable throughout its trajectory.

To summarize, the geodesic concavity of $\sqrt{\lambda_{\min}}$, together with the expression for the gradient of F as an average of tangent vectors pointing towards matrices

in the support of Q, yields eigenvalue control for the continuous-time gradient flow of F. This argument does not apply directly to the discrete-time GD updates, but nevertheless we show that the eigenvalues of the GD iterates can be controlled provided that the step size is taken sufficiently small; this is the content of Theorem 15.1.1.

Proof of Theorem 15.1.1. For $0 \leq \eta \leq 1$, let Σ_η denote the generalized barycenter of the distribution $Q_\eta \coloneqq (1-\eta)\,\delta_{\Sigma_0} + \eta\, Q$. For the average transport map

$$\bar{T} \coloneqq \int T_{\Sigma_0 \to \Sigma}\, \mathrm{d}Q(\Sigma)\,,$$

we have

$$\begin{aligned}\Sigma_\eta &= \big((1-\eta)\, I_d + \eta\, \bar{T}\big)\, \Sigma_0\, \big((1-\eta)\, I_d + \eta\, \bar{T}\big) \\ &= (1-\eta)^2\, \Sigma_0 + \eta^2\, \bar{T}\Sigma_0\bar{T} + \eta\,(1-\eta)\,(\bar{T}\Sigma_0 + \Sigma_0\bar{T})\,.\end{aligned}$$

On the other hand, let $\gamma_\Sigma(\eta)$ denote the geodesic joining Σ_0 to Σ at time η. Then,

$$\begin{aligned}\gamma_\Sigma(\eta) &= \big((1-\eta)\, I_d + \eta\, T_{\Sigma_0 \to \Sigma}\big)\, \Sigma_0\, \big((1-\eta)\, I_d + \eta\, T_{\Sigma_0 \to \Sigma}\big) \\ &= (1-\eta)^2\, \Sigma_0 + \eta^2\, \Sigma + \eta\,(1-\eta)\,(T_{\Sigma_0 \to \Sigma}\Sigma_0 + \Sigma_0 T_{\Sigma_0 \to \Sigma})\,.\end{aligned}$$

Upon integrating w.r.t. $\mathrm{d}Q(\Sigma)$ and comparing the two expressions, we find that

$$\Sigma_\eta = \int \gamma_\Sigma(\eta)\, \mathrm{d}Q(\Sigma) + \eta^2 \left(\bar{T}\Sigma_0\bar{T} - \int \Sigma\, \mathrm{d}Q(\Sigma)\right) \succeq \int \gamma_\Sigma(\eta)\, \mathrm{d}Q(\Sigma) - \beta\eta^2\, I_d\,.$$

Next, using the geodesic concavity of $\sqrt{\lambda_{\min}}$ and Jensen's inequality,

$$\begin{aligned}\lambda_{\min}\Big(\int \gamma_\Sigma(\eta)\, \mathrm{d}Q(\Sigma)\Big) &\geq \int \lambda_{\min}\big(\gamma_\Sigma(\eta)\big)\, \mathrm{d}Q(\Sigma) \\ &\geq \int \big((1-\eta)\, \sqrt{\lambda_{\min}(\Sigma_0)} + \eta\, \sqrt{\lambda_{\min}(\Sigma)}\big)^2\, \mathrm{d}Q(\Sigma) \\ &\geq \Big(\int \big((1-\eta)\, \sqrt{\lambda_{\min}(\Sigma_0)} + \eta\, \sqrt{\lambda_{\min}(\Sigma)}\big)\, \mathrm{d}Q(\Sigma)\Big)^2 \\ &\geq \big((1-\eta)\, \sqrt{\lambda_{\min}(\Sigma_0)} + \eta\, \sqrt{\alpha}\big)^2\,.\end{aligned}$$

We have established the inequality

$$\lambda_{\min}(\Sigma_\eta) \geq \big((1-\eta)\, \sqrt{\lambda_{\min}(\Sigma_0)} + \eta\, \sqrt{\alpha}\big)^2 - \beta\eta^2\,.$$

We now search for a value of $\lambda \geq 0$ such that if $\lambda_{\min}(\Sigma_0) \geq \lambda$, then $\lambda_{\min}(\Sigma_\eta) \geq \lambda$. From the above inequality, it suffices to have

$$\left((1-\eta)\sqrt{\lambda} + \eta\sqrt{\alpha}\right)^2 - \beta\eta^2 \overset{!}{\geq} \lambda\,.$$

Rearranging this expression, we want

$$(\sqrt{\alpha} - \sqrt{\lambda})\,\eta \overset{!}{\geq} \sqrt{\lambda + \beta\eta^2} - \sqrt{\lambda} = \sqrt{\lambda}\left(\sqrt{1 + \frac{\beta\eta^2}{\lambda}} - 1\right).$$

Applying the inequality $\sqrt{1+x} \leq 1 + x/2$, valid for $x \geq 0$, it suffices to have

$$(\sqrt{\alpha} - \sqrt{\lambda})\,\eta \overset{!}{\geq} \frac{\beta\eta^2}{2\sqrt{\lambda}}\,.$$

We now choose $\lambda = \alpha/4$, for which it can be verified that the above inequality holds for $\eta \leq \frac{\alpha}{2\beta}$. This concludes the proof. □

Remark 15.8.3. *In an earlier version of [Alt+21], we claimed that $-\sqrt{\lambda_{\min}}$ and $\sqrt{\lambda_{\max}}$ are convex along generalized geodesics, which is stronger than the statement of Theorem 15.1.1. Unfortunately, our proof of this claim was incorrect, as it relied upon [LL01, Corollary 3.5] which is false as written.*[2] *In fact, we have discovered a counterexample to our original claim: set*

$$\Sigma_0 \coloneqq \begin{bmatrix} 0.16 & 0.2 \\ 0.2 & 0.82 \end{bmatrix}, \qquad \Delta \coloneqq \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 0.2 \end{bmatrix}, \qquad \Sigma \coloneqq I_2 + \Sigma_0^{-1/2}\Delta\Sigma_0^{-1/2}\,.$$

Let $Q \coloneqq \frac{1}{2}\,\delta_{I_2} + \frac{1}{2}\,\delta_{\Sigma}$ and note that $\Sigma \succeq I_2$, i.e., Q is supported on matrices with eigenvalues at least 1. We can compute

$$\begin{aligned} T_{\Sigma_0 \to I_2} &= \Sigma_0^{-1/2}\,, \\ T_{\Sigma_0 \to \Sigma} &= \Sigma_0^{-1/2}\left(\Sigma_0^{1/2}\left(I_d + \Sigma_0^{-1/2}\Delta\Sigma_0^{-1/2}\right)\Sigma_0^{1/2}\right)^{1/2}\Sigma_0^{-1/2} \\ &= \Sigma_0^{-1/2}\left(\Sigma_0 + \Delta\right)^{1/2}\Sigma_0^{-1/2}\,, \\ \bar{T} &= \frac{1}{2}\,\Sigma_0^{-1/2}\left(\Sigma_0^{1/2} + \left(\Sigma_0 + \Delta\right)^{1/2}\right)\Sigma_0^{-1/2}\,, \end{aligned}$$

so that the generalized barycenter $\bar{\Sigma}$ of Q at Σ_0 is

$$\bar{\Sigma} = \Sigma_0^{-1/2}\left(\frac{\Sigma_0^{1/2} + \left(\Sigma_0 + \Delta\right)^{1/2}}{2}\right)^2\Sigma_0^{-1/2}\,.$$

[2] The "if" direction of the corollary is incorrect: upon taking $B = I_d$, it says that $X \preceq B^{1/2}$ implies $X^2 \preceq B$, which contradicts the well-known fact that the square function is not operator monotone.

However, it can be numerically verified that $\lambda_{\min}(\bar{\Sigma}) \leq 0.993 < 1$. *This shows that the set of PSD matrices with eigenvalues at least* 1 *is not closed under generalized geodesics. In particular,* $-\sqrt{\lambda_{\min}}$ *is an example of a functional which is convex along barycenters but not along generalized geodesics, which may be of interest in its own right. The revised statement of Theorem 15.1.1 fixes this issue, at the cost of worsening our quantitative results.*

We also remark that the above counterexample was obtained as follows. One can show that the statement

the set of PSD matrices with eigenvalues at least 1 *is closed under generalized geodesics*

is equivalent to the statement

$$\text{for all } \Sigma_0 \succ 0 \text{ and all } A, B \succeq \Sigma_0, \text{ it holds that}$$

$$\Big(\frac{A^{1/2} + B^{1/2}}{2}\Big)^2 \succeq \Sigma_0 \,.$$

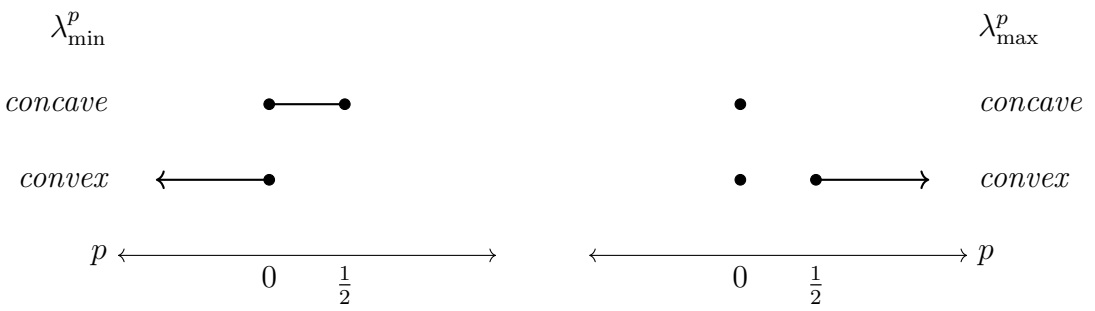
The equivalence between the two statements is obtained by considering the generalized barycenter of the distribution $P \coloneqq \frac{1}{2}\,\delta_{\Sigma_0^{-1/2} A \Sigma_0^{-1/2}} + \frac{1}{2}\,\delta_{\Sigma_0^{-1/2} B \Sigma_0^{-1/2}}$ *at* Σ_0. *Therefore, we discovered our counterexample by finding a counterexample to the latter statement. Note also the similarity of the second statement with the last conjecture in [CK85]. In contrast, it was shown in Lemma 15.4.3 that the set of matrices with eigenvalues* at most β *is convex along generalized geodesics.*

■ 15.8.2 Sharpness of Theorem 15.8.1

We investigate the sharpness of this result in the following sense: for what exponents $p \in \mathbb{R}$ is it true that the functionals $-\lambda_{\min}^p$, $\lambda_{\max}^p$ are geodesically convex? For instance, the functional $\lambda_{\max}$ was shown to be geodesically convex in Lemma 15.4.3.

In the following theorem, we show that the exponent $p = 1/2$ in Theorem 15.8.1 is optimal, in the sense that all possible geodesic convexity statements involving powers of $\lambda_{\min}$ and $\lambda_{\max}$ (except the trivial case $p = 0$) can be deduced from the result for $p = 1/2$.

Theorem 15.8.4. *The following diagrams depict the exponents* $p \in \mathbb{R}$ *for which* $\lambda_{\min}^p$ *and* $\lambda_{\max}^p$ *are concave or convex.*



The diagram is to be interpreted as follows. If part of the diagram is filled in with a solid black line, then the corresponding functional is geodesically concave/convex. If part of the diagram is not *filled in, then there exist counterexamples showing that the functional is* not *geodesically concave/convex.*

Proof. First, we establish the positive results, which follow from composition rules:

- For $0 \le p \le 1/2$, $\lambda^p_{\min}$ is the composition of the increasing concave function $(\cdot)^{2p}$ with the concave function $\sqrt{\lambda_{\min}}$, so it is concave.
- For $p \le 0$, $\lambda^p_{\min}$ is the composition of the decreasing convex function $(\cdot)^{2p}$ with the concave function $\sqrt{\lambda_{\min}}$, so it is convex.
- For $p \ge 1/2$, $\lambda^p_{\max}$ is the composition of the increasing convex function $(\cdot)^{2p}$ with the convex function $\sqrt{\lambda_{\max}}$, so it is convex.

Next, we turn towards the negative results. First, recall from Fact 4 in §2.3.2 that if Σ_0 and Σ_1 are one-dimensional, i.e., they are positive numbers, then the Bures–Wasserstein geodesic is

$$\Sigma_t = \left((1-t)\,\Sigma_0^{1/2} + t\,\Sigma_1^{1/2}\right)^2, \qquad t \in [0,1]\,.$$

Also, in this case, $\lambda_{\min}$ and $\lambda_{\max}$ coincide and equal the identity; we thus abuse notation slightly in this paragraph by writing λ for both to handle the two cases simultaneously. Once we reparametrize by the square roots, it is seen that asking for concavity/convexity of λ^p is equivalent to asking for usual convexity of $(\cdot)^{2p}$ on $\mathbb{R}_+$. This example rules out: (1) the concavity of λ^p for $p < 0$; (2) the convexity of λ^p for $0 < p < 1/2$; and (3) the concavity of λ^p for $p > 1/2$.

To rule out convexity of $\lambda^p_{\min}$ for $p > 0$, consider $\Sigma = \mathrm{diag}(\varepsilon, 1/\varepsilon)$ for small $\varepsilon > 0$. The transport map from Σ^{-1} to Σ is Σ, so from (2.26) the midpoint of this geodesic is $M := (\Sigma + \Sigma^{-1} + 2I_2)/4 = (\varepsilon + \varepsilon^{-1} + 2)I_2/4$. In particular, this implies that $\lambda_{\min}(M) \ge 1/(4\varepsilon) \gg \varepsilon = \max\{\lambda_{\min}(\Sigma), \lambda_{\min}(\Sigma^{-1})\}$. Thus $\lambda^p_{\min}$ is not convex for any $p > 0$.

To rule out concavity of $\lambda_{\max}^p$ for $p > 0$, note that for ε sufficiently small, in the previous example $\lambda_{\max}(M) \approx 1/(4\varepsilon) \ll 1/\varepsilon = \max\{\lambda_{\max}(\Sigma), \lambda_{\max}(\Sigma^{-1})\}$. Also, for any $p < 0$, the convexity of $\lambda_{\max}^p$ would imply the concavity of $\lambda_{\max}^{-p}$ due to the composition rules, hence $\lambda_{\max}^p$ is not convex.

This covers all cases. □

■ 15.8.3 Eigenvalue clipping is a Bures–Wasserstein contraction

Convex sets play an important role in Euclidean optimization because projection onto a convex set is a contraction (c.f. [Bub15, Lemma 3.1]), and hence projected gradient descent can be used to solve constrained optimization. Unfortunately, as the Bures–Wasserstein space is positively curved, we cannot automatically conclude that projection onto a geodesically convex set is a projection. Nevertheless, we can verify by hand the following result. In what follows, define for $0 < \beta < \infty$ the operator $\operatorname{clip}^\beta : \mathbf{S}_{++}^d \to \mathbf{S}_{++}^d$ in the following way: if $\Sigma = \sum_{i=1}^d \lambda_i u_i u_i^\mathsf{T}$ is an eigenvalue decomposition of Σ, then

$$\operatorname{clip}^\beta \Sigma \coloneqq \sum_{i=1}^d (\lambda_i \wedge \beta)\, u_i u_i^\mathsf{T} .$$

Proposition 15.8.5. *The operator* $\operatorname{clip}^\beta$ *is a contraction with respect to the Bures–Wasserstein metric, i.e.,* $W_2(\operatorname{clip}^\beta \Sigma, \operatorname{clip}^\beta \Sigma') \le W_2(\Sigma, \Sigma')$.

To prove this proposition, we first extend the clipping operation to an operator $\mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ via the singular values; namely, given a singular value decomposition $A = \sum_{i=1}^d s_i u_i v_i^\mathsf{T}$, we let $\operatorname{clip}^\beta A \coloneqq \sum_{i=1}^d (s_i \wedge \beta)\, u_i v_i^\mathsf{T}$.

Proof of Proposition 15.8.5. Fix $X, Y \in \mathbf{S}_{++}^d$. It is known (see, e.g., [BJL19]) that

$$W_2(X, Y) = \min_{\substack{A,B\in\mathbb{R}^{d\times d} \\ AA^\mathsf{T}=X \\ BB^\mathsf{T}=Y}} \|A - B\|_{\mathrm{HS}} .$$

Let $(\bar{A}, \bar{B})$ be a minimizing pair in the above expression. We aim to show

$$W_2(\operatorname{clip}^\beta X, \operatorname{clip}^\beta Y) \le \|\operatorname{clip}^{\sqrt{\beta}} \bar{A} - \operatorname{clip}^{\sqrt{\beta}} \bar{B}\|_{\mathrm{HS}} \overset{?}{\le} \|\bar{A} - \bar{B}\|_{\mathrm{HS}} = W_2(X, Y) .$$

We only have to show the second inequality, and we do so by showing that the operator $\operatorname{clip}^M : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ satisfies

$$\operatorname{clip}^M A = \operatorname*{arg\,min}_{\tilde{A}\in\mathbb{R}^{d\times d},\, \|\tilde{A}\|\le M} \|A - \tilde{A}\|_{\mathrm{HS}} , \qquad A \in \mathbb{R}^{d\times d} . \tag{15.18}$$

This will prove that clip^M is the Euclidean *projection* onto the closed convex set $\{\|\cdot\| \le M\}$, and such a projection is automatically 1-Lipschitz.

Indeed, showing (15.18) is standard. Write $A = U\Sigma V^\mathsf{T}$ for its singular value decomposition. Then,

$$\begin{aligned}\operatorname*{arg\,min}_{\tilde{A}\in\mathbb{R}^{d\times d},\ \|\tilde{A}\|\le M} \|\tilde{A}-A\|_{\mathrm{HS}}^2 &= \operatorname*{arg\,min}_{\tilde{A}\in\mathbb{R}^{d\times d},\ \|\tilde{A}\|\le M} \|\tilde{A}-U\Sigma V^\mathsf{T}\|_{\mathrm{HS}}^2 \\ &= \operatorname*{arg\,min}_{\tilde{A}\in\mathbb{R}^{d\times d},\ \|\tilde{A}\|\le M} \|U^\mathsf{T}\tilde{A}V-\Sigma\|_{\mathrm{HS}}^2 \\ &= \operatorname*{arg\,min}_{\tilde{A}\in\mathbb{R}^{d\times d},\ \|\tilde{A}\|\le M} \Big\{\sum_{i=1}^d \{\Sigma[i,i]-(U^\mathsf{T}\tilde{A}V)[i,i]\}^2 + \sum_{\substack{i,j\in[d]\\ i\neq j}} (U^\mathsf{T}\tilde{A}V)[i,j]^2\Big\}.\end{aligned}$$

On the other hand,

$$\begin{aligned}&\min_{\tilde{A}\in\mathbb{R}^{d\times d},\ \|\tilde{A}\|\le M} \Big\{\sum_{i=1}^d \{\Sigma[i,i]-(U^\mathsf{T}\tilde{A}V)[i,i]\}^2 + \sum_{i,j\in[d],\ i\neq j} (U^\mathsf{T}\tilde{A}V)[i,j]^2\Big\} \\ &\quad\ge \sum_{i=1}^d \{(\Sigma[i,i]-M)_+\}^2,\end{aligned}$$

with equality attained at the unique minimizer $\tilde{A}$ satisfying $U^\mathsf{T}\tilde{A}V = \mathrm{clip}^M \Sigma$, i.e., $\tilde{A} = \mathrm{clip}^M A$. □

15.9 Proofs for Bures–Wasserstein barycenters

15.9.1 Properties of the Bures–Wasserstein barycenter

Existence and uniqueness of the barycenter in the case where P is finitely supported follows from the seminal work of Agueh and Carlier [AC11]. We extend this result to the case where P is not finitely supported.

Proposition 15.9.1 (Gaussian barycenter)**.** *Fix $0 < \lambda_{\min} \le \lambda_{\max} < \infty$. Let $P \in \mathcal{P}_2(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d))$ be such that for all $\mu \in \operatorname{supp} P$, $\mu = \gamma_{m(\mu),\Sigma(\mu)}$ is a Gaussian with $\lambda_{\min} I_d \preceq \Sigma(\mu) \preceq \lambda_{\max} I_d$. Let $\gamma_{\bar{m},\bar{\Sigma}}$ be the Gaussian measure with mean $\bar{m} := \int m(\mu)\,\mathrm{d}P(\mu)$ and covariance matrix $\bar{\Sigma}$ which is a fixed point of the mapping $S \mapsto G(S) := \int (S^{1/2}\,\Sigma(\cdot)\,S^{1/2})^{1/2}\,\mathrm{d}P$. Then, $\gamma_{\bar{m},\bar{\Sigma}}$ is the unique barycenter of P.*

Proof. To show that there exists a fixed point for the mapping G, apply Brouwer's fixed-point theorem as in [AC11, Theorem 6.1]. To see that $\gamma_{\bar{m},\bar{\Sigma}}$ is indeed a

barycenter, observe the mapping

$$\varphi : (\mu, x) \mapsto \varphi_\mu(x) \coloneqq \langle x, m(\mu)\rangle + \frac{1}{2}\,\langle x - \bar{m}, \bar{\Sigma}^{-1/2}[\bar{\Sigma}^{1/2}\Sigma(\mu)\bar{\Sigma}^{1/2}]^{1/2}\bar{\Sigma}^{-1/2}(x-\bar{m})\rangle$$

satisfies the characterization (15.17) (so that φ is an optimal dual solution for the barycenter problem w.r.t. $\gamma_{\bar{m},\bar{\Sigma}}$) using the explicit form of the transport map (15.14), so $\gamma_{\bar{m},\bar{\Sigma}}$ is a barycenter of P. Uniqueness follows from the variance inequality (Theorem 15.3.3) once we establish regularity of the optimal transport maps in Lemma 15.9.2. □

Now that we know the barycenter of P is a centered non-degenerate Gaussian, we abuse notation and treat P as an element of $\mathcal{P}(\mathbf{S}^d_{++})$.

Lemma 15.9.2. *Suppose that $\Sigma, \Sigma' \in \mathbf{S}^d_{++}$ have eigenvalues which lie in the range $[\lambda_{\min}, \lambda_{\max}]$, and let $\kappa \coloneqq \lambda_{\max}/\lambda_{\min}$ denote the condition number. Then, the eigenvalues of the transport map $T_{\Sigma\to\Sigma'}$ lie in the range $[1/\sqrt{\kappa}, \sqrt{\kappa}]$.*

Proof. The transport map $T_{\Sigma\to\Sigma'}$ is explicitly given in (15.3), and it can be recognized as the matrix geometric mean of Σ^{-1} and Σ'. Applying a norm bound for the matrix geometric mean [BG12, Theorem 3], we deduce that

$$\lambda_{\max}(T_{\Sigma\to\Sigma'}) \le \lambda_{\max}(\Sigma'^{1/4}\Sigma^{-1/2}\Sigma'^{1/4}) \le \sqrt{\kappa}\,.$$

The symmetry of Σ and Σ' together with Fact 3 in §2.3.2 yields the opposite inequality $\lambda_{\min}(T_{\Sigma\to\Sigma'}) \ge 1/\sqrt{\kappa}$. □

Lemma 15.9.2 can also be recovered by applying Caffarelli's contraction theorem, see §10.2 for a proof.

Theorem 15.3.3 readily yields the following variance inequality. Recall that $\Sigma^\star$ denotes the covariance matrix of the barycenter of P; in an abuse of terminology, we refer to $\Sigma^\star$ itself as the barycenter of P.

Theorem 15.9.3. *Assume that the covariance matrices in the support of P have eigenvalues in the range $[\lambda_{\min}, \lambda_{\max}]$. Then, F satisfies a* variance inequality*,*

$$F(\Sigma) - F(\Sigma^\star) \ge \frac{1}{2\sqrt{\kappa}}\, W_2^2(\Sigma, \Sigma^\star)\,, \qquad \text{for all } \Sigma \in \mathbf{S}^d_{++}\,.$$

■ 15.9.2 A PL inequality on the Bures–Wasserstein manifold

Theorem 15.9.4. *Assume that the covariance matrices in the support of P have eigenvalues in the range $[\lambda_{\min}, \lambda_{\max}]$. Then, F satisfies a PL inequality at the matrix Σ:*

$$F(\Sigma) - F(\Sigma^\star) \le 2\sqrt{\kappa}\,\frac{\lambda_{\max}}{\lambda_{\min}(\Sigma)}\,\|\nabla F(\Sigma)\|_\Sigma^2\,.$$

Proof. Let $(\tilde{b}_s)_{s\in[0,1]}$ be the constant-speed geodesic between $\tilde{b}_0 \coloneqq b \coloneqq \gamma_{0,\Sigma}$ and $\tilde{b}_1 \coloneqq b^\star \coloneqq \gamma_{0,\Sigma^\star}$. Combining Lemma 15.3.5 (with an additional use of the Cauchy–Schwarz inequality) and Theorem 15.9.3, we get

$$F(b) - F(b^\star) \le 2\sqrt{\kappa} \int_0^1 \int \|\nabla F(b)\|^2 \,\mathrm{d}\tilde{b}_s \,\mathrm{d}s\,. \tag{15.19}$$

Define a random variable $X_s \sim \tilde{b}_s$ and observe that

$$\int \|\nabla F(b)\|^2 \,\mathrm{d}\tilde{b}_s = \mathbb{E}[\|(\tilde{M} - I_d)\, X_s\|^2]\,,$$

where

$$\tilde{M} = \int \Sigma^{-1/2}(\Sigma^{1/2} S \Sigma^{1/2})^{1/2} \Sigma^{-1/2} \,\mathrm{d}P(S)\,.$$

Moreover, recall that $X_s = (1-s)\, X_0 + s\, X_1$ where $X_0 \sim \tilde{b}_0$ and $X_1 \sim \tilde{b}_1$ are optimally coupled. Therefore, by Jensen's inequality, we have for all $s \in [0,1]$,

$$\begin{aligned}\mathbb{E}[\|(\tilde{M} - I_d)\, X_s\|^2] &\le (1-s)\, \mathbb{E}[\|(\tilde{M} - I_d)\, X_0\|^2] + s\, \mathbb{E}[\|(\tilde{M} - I_d)\, X_1\|^2] \\ &\le \frac{\lambda_{\max}}{\lambda_{\min}(\Sigma)}\, \mathbb{E}[\|(\tilde{M} - I_d)\, X_0\|^2]\,,\end{aligned}$$

where in the second inequality, we used the fact that

$$\begin{aligned}\mathbb{E}[\|(\tilde{M} - I_d)\, X_1\|^2] = \operatorname{tr}\big(\Sigma^\star\, (\tilde{M} - I_d)^2\big) &\le \|\Sigma^\star \Sigma^{-1}\|_{\mathrm{op}} \operatorname{tr}\big(\Sigma\, (\tilde{M} - I_d)^2\big) \\ &\le \frac{\lambda_{\max}}{\lambda_{\min}(\Sigma)}\, \mathbb{E}[\|(\tilde{M} - I_d)\, X_0\|^2]\,.\end{aligned}$$

Together with (15.19), it yields

$$F(b) - F(b^\star) \le 2\sqrt{\kappa}\, \frac{\lambda_{\max}}{\lambda_{\min}(\Sigma)} \underbrace{\mathbb{E}[\|(\tilde{M} - I_d)\, X_0\|^2]}_{=\|\nabla F(b)\|_b^2}\,. \qquad \square$$

■ 15.9.3 Riemannian gradient descent

In this section, we review the strategy of the proof and establish the dimension-free rates in Theorems 15.4.1 and 15.4.2.

Let F denote the barycenter functional,

$$F(\Sigma) \coloneqq \frac{1}{2} \int W_2^2(\Sigma, \cdot) \,\mathrm{d}P\,. \tag{15.20}$$

Standard optimization guarantees are often proven under the assumption that the objective function F is smooth and convex. Since we are considering Riemannian descent, this should be interpreted as convex and smooth along geodesics, as in [ZS16]. Unfortunately, the functional F is not geodesically convex (see Figure 15.2), and so we must look for weaker conditions which still imply convergence of GD/SGD. A gradient domination condition known as the *Polyak–Łojasiewicz inequality* (henceforth *PL inequality*) was introduced in the non-convex optimization literature as an appropriate substitute for strong convexity [KNS16], and it plays a key role in the analysis.

We have established the following properties of the barycenter functional.

Theorem 15.9.5. *Let* $0 < \lambda_{\min} \leq \lambda_{\max} < \infty$ *and write* $\kappa \coloneqq \lambda_{\max}/\lambda_{\min}$.

1. *The barycenter functional* F *is* 1*-geodesically smooth.*

2. *Assume that the covariance matrices in the support of* P *have eigenvalues in the range* $[\lambda_{\min}, \lambda_{\max}]$. *Then,* F *satisfies a* variance inequality*,*

$$F(\Sigma) - F(\Sigma^\star) \geq \frac{1}{2\sqrt{\kappa}} W_2^2(\Sigma, \Sigma^\star)\,, \qquad \text{for all } \Sigma \in \mathbf{S}_{++}^d\,.$$

3. *Assume that the covariance matrices in the support of* P *have eigenvalues in the range* $[\lambda_{\min}, \lambda_{\max}]$. *Then,* F *satisfies a PL inequality at the matrix* Σ*:*

$$F(\Sigma) - F(\Sigma^\star) \leq 2\sqrt{\kappa}\, \frac{\lambda_{\max}}{\lambda_{\min}(\Sigma)} \|\nabla F(\Sigma)\|_\Sigma^2\,.$$

Geodesic smoothness together with a PL inequality at every iterate are enough to obtain convergence guarantees for GD/SGD in objective value (i.e., the quantity $F(\Sigma) - F(\Sigma^\star)$), c.f. Theorems 15.3.1 and 15.3.2. The variance inequality is then used to deduce convergence of the iterate to $\Sigma^\star$.

The main difficulty when applying these results is the assumption required for the third point: it requires *a priori* control over the eigenvalues of the iterates of GD/SGD. This difficulty is addressed via the following strategy: identify a geodesically convex subset $\mathcal{S}$ of the Bures–Wasserstein manifold for which we can prove uniform bounds on the eigenvalues of matrices in $\mathcal{S}$. Since the iterates of SGD travel along geodesics, if P is supported in $\mathcal{S}$ and the algorithm is initialized in $\mathcal{S}$, it follows that all iterates of SGD will remain in $\mathcal{S}$. The situation is similar for GD, except that "geodesics" must be replaced by "generalized geodesics".

To obtain this control over the eigenvalues, we apply our geometric result (Theorem 15.1.1) to prove the following result.

Lemma 15.9.6. *Suppose that the covariance matrices in the support of P have eigenvalues in the range $[\lambda_{\min}, \lambda_{\max}]$, and that we initialize GD (respectively SGD) at a point in $\operatorname{supp} P$. Then, the iterates of GD with step size at most $\frac{1}{2\kappa}$ (respectively SGD) also have eigenvalues in the range $[\lambda_{\min}/4, \lambda_{\max}]$ (respectively $[\lambda_{\min}, \lambda_{\max}]$).*

Proof. The result for SGD follows because SGD moves along geodesics and the set of matrices with eigenvalues in $[\lambda_{\min}, \lambda_{\max}]$ is geodesically convex (Theorem 15.8.1). For GD, we instead invoke the generalized geodesic convexity of $\lambda_{\max}$ (Lemma 15.4.3) together with Theorem 15.1.1. □

We can now prove Theorem 15.4.1.

Proof of Theorem 15.4.1. The proof for SGD follows from Theorem 15.3.2. For GD, from Theorem 15.9.5 and Lemma 15.9.6, we have the PL inequality

$$F(\Sigma_t^{\mathrm{GD}}) - F(\Sigma^\star) \le 8\kappa^{3/2} \|\nabla F(\Sigma_t^{\mathrm{GD}})\|^2_{\Sigma_t^{\mathrm{GD}}}$$

at any GD iterate Σ_t^{GD}. Also, from the 1-smoothness of the barycenter functional, we obtain the descent lemma

$$F(\Sigma_{t+1}^{\mathrm{GD}}) - F(\Sigma_t^{\mathrm{GD}}) \le -\eta \left(1 - \frac{\eta}{2}\right) \|\nabla F(\Sigma_t^{\mathrm{GD}})\|^2_{\Sigma_t^{\mathrm{GD}}} .$$

With our step size choice $\eta = \frac{1}{2\kappa}$, this becomes

$$F(\Sigma_{t+1}^{\mathrm{GD}}) - F(\Sigma_t^{\mathrm{GD}}) \le -\frac{3}{8\kappa} \|\nabla F(\Sigma_t^{\mathrm{GD}})\|^2_{\Sigma_t^{\mathrm{GD}}} .$$

Combining these two inequalities and iterating yields the result for GD. □

We now sketch the modifications required to prove Theorem 15.4.2.

Proof of Theorem 15.4.2. We first note that

$$\frac{\|\lambda_{\max}\|_{1/2}}{\|\lambda_{\min}\|_{1/2}} = \left(\frac{\int \lambda_{\max}(\Sigma)^{1/2} \,\mathrm{d}P(\Sigma)}{\int \lambda_{\min}(\Sigma)^{1/2} \,\mathrm{d}P(\Sigma)}\right)^2 \le \sup_{\Sigma \in \operatorname{supp}(P)} \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} = \kappa^\star .$$

Above, the inequality follows from rearranging

$$\int \sqrt{\lambda_{\max}(\Sigma)} \,\mathrm{d}P(\Sigma) \le \sqrt{\kappa^\star} \int \sqrt{\lambda_{\min}(\Sigma)} \,\mathrm{d}P(\Sigma) .$$

We check that the variance inequality and PL inequality from Theorem 15.9.5 continue to hold under these assumptions.

Variance inequality. From the geodesic convexity of $-\sqrt{\lambda_{\min}}$ and $\sqrt{\lambda_{\max}}$, the barycenter $\Sigma^\star$ of P has eigenvalues in $[\|\lambda_{\min}\|_{1/2}, \|\lambda_{\max}\|_{1/2}]$. By modifying the proof of Lemma 15.9.2 and using Fact 3 in §2.3.2, the transport map $T_{\Sigma^\star \to \Sigma}$ has eigenvalues bounded below as

$$\lambda_{\min}(T_{\Sigma^\star \to \Sigma}) = \frac{1}{\lambda_{\max}(T_{\Sigma \to \Sigma^\star})} \geq \frac{1}{\lambda_{\max}(\Sigma^{\star\, 1/4} \Sigma^{-1/2} \Sigma^{\star\, 1/4})} \geq \frac{\lambda_{\min}(\Sigma)^{1/2}}{\|\lambda_{\max}\|_{1/2}^{1/2}}\,.$$

From Theorem 15.3.3, the variance inequality holds for P with constant

$$\int \lambda_{\min}(T_{\Sigma^\star \to \Sigma})\, \mathrm{d}P(\Sigma) \geq \Big(\frac{\|\lambda_{\min}\|_{1/2}}{\|\lambda_{\max}\|_{1/2}}\Big)^{1/2}\,.$$

PL inequality. Similarly, a modification of the proof of Theorem 15.9.4 shows that a PL inequality holds at Σ:

$$F(\Sigma) - F(\Sigma^\star) \leq 2 \Big(\frac{\|\lambda_{\max}\|_{1/2}}{\|\lambda_{\min}\|_{1/2}}\Big)^{1/2} \frac{\|\lambda_{\max}\|_{1/2}}{\lambda_{\min}(\Sigma)} \|\nabla F(\Sigma)\|_\Sigma^2\,.$$

Putting it together. The iterates of SGD all have eigenvalues in the range $[\|\lambda_{\min}\|_{1/2}, \|\lambda_{\max}\|_{1/2}]$, whereas from Theorem 15.1.1, the iterates of GD with step size at most $\frac{\|\lambda_{\min}\|_{1/2}}{2\,\|\lambda_{\max}\|_1}$ all have eigenvalues in the range $[\|\lambda_{\min}\|_{1/2}/4, \|\lambda_{\max}\|_1]$. □

■ 15.9.4 Euclidean gradient descent approach

In this section, we describe Euclidean projected gradient and projected stochastic gradient algorithms for computing Bures–Wasserstein barycenters.

Fix $0 < \alpha \leq \beta$ and denote by $\mathcal{K}_{\alpha,\beta}$ the subset of covariance matrices whose spectrum lies within $[\alpha, \beta]$. Let F denote the barycenter functional, defined in (15.20). Let $\Pi_{\alpha,\beta} : \mathbf{S}^d \to \mathcal{K}_{\alpha,\beta}$ denote the Euclidean projection onto $\mathcal{K}_{\alpha,\beta}$. Given a starting matrix Σ_0, the projected gradient descent scheme to minimize the barycenter functional of a measure P supported on $\mathcal{K}_{\lambda_{\min},\lambda_{\max}}$ is given by

$$\Sigma_{n+1}^{\mathrm{EGD}} := \Pi_{\lambda_{\min},\lambda_{\max}}\big(\Sigma_n - \eta\, \mathrm{D}\, F(\Sigma_n^{\mathrm{EGD}})\big)\,, \qquad n \geq 0\,. \tag{15.21}$$

Here, D denotes the Euclidean gradient of F. Also, suppose that $\Sigma_1, \ldots, \Sigma_n$ are i.i.d. samples from P. Then, the projected stochastic gradient scheme is

$$\Sigma_{n+1}^{\mathrm{ESGD}} := \Pi_{\lambda_{\min},\lambda_{\max}}\big(\Sigma_n^{\mathrm{ESGD}} - \eta_{n+1}\, \big\{I_d - \mathrm{GM}(\Sigma_{n+1}, (\Sigma_n^{\mathrm{ESGD}})^{-1})\big\}\big)\,, \qquad n \geq 0\,, \tag{15.22}$$

where for projected SGD we use time-varying step sizes. Convergence analysis for the iterations (15.21) and (15.22) are provided in [Alt+21].

For the iterations given by (15.21) and (15.22) to be practical, we need the projection step to be implementable. The following lemma takes care of this.

Lemma 15.9.7. *Let $\Pi_{\alpha,\beta} : \mathbf{S}^d \to \mathcal{K}_{\alpha,\beta}$ be the projection with respect to the Frobenius norm. Then*

$$\Pi_{\alpha,\beta}(Y) = \sum_{i=1}^{d} \left[(\lambda_i \wedge \beta) \vee \alpha\right] v_i v_i^{\mathsf{T}}$$

where $Y = \sum_{i=1}^{d} \lambda_i v_i v_i^{\mathsf{T}}$ is an orthogonal eigendecomposition of Y.

Proof. Let $Y = Q\Lambda Q^{\mathsf{T}}$ be an orthogonal eigendecomposition of Y. Since the Frobenius norm is unitarily invariant, we have

$$\begin{aligned}\Pi_{\alpha,\beta}(Y) &= \operatorname*{arg\,min}_{X \in \mathcal{K}_{\alpha,\beta}} \|X - Q\Lambda Q^{\mathsf{T}}\|_{\mathrm{F}}^2 = \operatorname*{arg\,min}_{X \in \mathcal{K}_{\alpha,\beta}} \|Q^{\mathsf{T}} X Q - \Lambda\|_{\mathrm{F}}^2 \\ &= Q \left(\operatorname*{arg\,min}_{X \in \mathcal{K}_{\alpha,\beta}} \|X - \Lambda\|_{\mathrm{F}}^2\right) Q^{\mathsf{T}}\end{aligned}$$

and the result follows. □

■ 15.9.5 SDP formulation

The SDP formulation of the Bures–Wasserstein barycenter is as follows. Suppose that P is a discrete distribution, $P = \sum_{i=1}^{k} p_i \delta_{\Sigma_i}$. The Wasserstein distance between $\Sigma_0, \Sigma_1 \in \mathbf{S}_{++}^d$ can be expressed as

$$W_2^2(\Sigma_0, \Sigma_1) = \min_{S \in \mathbb{R}^{d \times d}} \left\{ \operatorname{tr}(\Sigma_0 + \Sigma_1 - 2S) \quad \text{such that} \quad \begin{bmatrix} \Sigma_0 & S \\ S^{\mathsf{T}} & \Sigma_1 \end{bmatrix} \succeq 0 \right\}.$$

It follows that the barycenter $\Sigma^\star$ of P solves the optimization problem

$$\operatorname*{minimize}_{\substack{\Sigma^\star \in \mathbf{S}_{++}^d \\ S_1, \ldots, S_k \in \mathbb{R}^{d \times d}}} \left\{ \operatorname{tr}\Big(\Sigma^\star - 2 \sum_{i=1}^{k} p_i S_i\Big) \quad \text{such that} \quad \begin{bmatrix} \Sigma_i & S_i \\ S_i^{\mathsf{T}} & \Sigma^\star \end{bmatrix} \succeq 0\,, \ \forall i \in [k] \right\}.$$

■ 15.10 Conclusion

A question for future work is establishing more general conditions under which the PL inequality holds. In particular, one could examine general conditions where Lemma 15.3.5 implies a PL inequality. Another path involves studying the effectiveness of the averaging strategy used in §15.5, which empirically performs much better when the covariance matrices are poorly conditioned (see Figure 15.5). Previous results for averaging of stochastic gradient descent on manifolds have strong geodesic convexity and smoothness assumptions [Tri+18].

Chapter 16

Gaussian variational inference

Along with Markov chain Monte Carlo (MCMC) methods, variational inference (VI) has emerged as a central computational approach to large-scale Bayesian inference. Rather than sampling from the true posterior π, VI aims at producing a simple but effective approximation $\hat{\pi}$ to π for which summary statistics are easy to compute. However, unlike the well-studied MCMC methodology, algorithmic guarantees for VI are still relatively less well-understood. In this work, we propose principled methods for VI, in which $\hat{\pi}$ is taken to be a Gaussian or a mixture of Gaussians, which rest upon the theory of gradient flows on the Bures–Wasserstein space of Gaussian measures. Akin to MCMC, it comes with strong theoretical guarantees when π is log-concave.

This chapter is based on [Lam+22], joint with Marc Lambert, Francis Bach, Silvère Bonnabel, and Philippe Rigollet.

16.1 Introduction

This work brings together three active research areas: variational inference, variational Kalman filtering, and gradient flows on the Wasserstein space.

Variational inference. The development of large-scale Bayesian methods has fuelled the need for fast and scalable methods to approximate complex distributions. More specifically, Bayesian methodology typically generates a high-dimensional posterior distribution $\pi \propto \exp(-V)$ that is known only up to normalizing constants, making the computation even of simple summary statistics such as the mean and covariance a major computational hurdle. To overcome this limitation, two distinct computational approaches are largely favored. The first approach consists of Markov chain Monte Carlo (MCMC) methods that rely on carefully constructed Markov chains which (approximately) converge to π. For example, the *Langevin diffusion*

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \tag{16.1}$$

where $(B_t)_{t\geq 0}$ denotes standard Brownian motion on $\mathbb{R}^d$, admits π as a stationary distribution. Crucially, the Langevin diffusion can be discretized and implemented without knowledge of the normalizing constant of π, leading to practical algorithms for Bayesian inference. Recent theoretical efforts have produced sharp non-asymptotic convergence guarantees for algorithms based on the Langevin diffusion (or variants thereof), with many results known when π is strongly log-concave or satisfies isoperimetric assumptions (see, e.g., earlier chapters in this thesis and references therein).

More recently, Variational Inference (VI) has emerged as a viable alternative to MCMC [Jor+99; WJ08; BKM17]. The goal of VI is to approximate the posterior π by a more tractable distribution $\hat{\pi} \in \mathcal{P}$ such that

$$\hat{\pi} \in \operatorname*{arg\,min}_{p \in \mathcal{P}} \mathsf{KL}(p \,\|\, \pi)\,. \tag{16.2}$$

A common example arises when $\mathcal{P}$ is the class of product distributions, in which case $\hat{\pi}$ is called the *mean-field* approximation of $\mathcal{P}$. Unfortunately, by definition, mean-field approximations fail to capture important correlations present in the posterior π, and various remedies have been proposed, with varied levels of success. In this work, we largely focus on obtaining a Gaussian approximation to π, that is, we take $\mathcal{P}$ to be the class of non-degenerate Gaussian distributions on $\mathbb{R}^d$ [BB97; See99; HV04; OA09; Zha+18]. The expressive power of the variational model may be further increased by considering mixture distributions [LKS19a; DD21; DDP21].

Although the solution $\hat{\pi}$ of (16.2) is no longer equal to the true posterior, variational inference remains heavily used in practice because the problem (16.2) can be solved for simple models $\mathcal{P}$ via scalable optimization algorithms. In particular, VI avoids many of the practical hurdles associated with MCMC—such as the potentially long "burn-in" period of samplers and the lack of effective stopping criteria for the algorithm—while still producing informative summary statistics. In this regard, we highlight the fact that obtaining an approximation for the covariance matrix of π via MCMC methods requires drawing potentially many samples, whereas for many choices of $\mathcal{P}$ (e.g., the Gaussian approximation) the covariance matrix of $\hat{\pi}$ can be directly obtained from the solution to the VI problem (16.2).

However, in contrast with MCMC methods, to date there have not been many theoretical guarantees for VI, even when π is strongly log-concave and $\mathcal{P}$ is taken to be the class of Gaussians $\mathsf{normal}(m, \Sigma)$. The problem stems from the fact that the objective in (16.2) is typically non-convex in the pair (m, Σ). Obtaining such guarantees remains a pressing challenge for the field.

Variational Kalman filtering. There is also considerable interest in extending ideas behind variational inference to dynamical settings of Bayesian inference. Consider a general framework where $(\pi_t)_{t\in\mathcal{T}}$ represents the marginal laws of a stochastic process indexed by time t, which can be discrete or continuous. The goal is to recursively build a Gaussian approximation to $(\pi_t)_{t\in\mathcal{T}}$.

As a concrete example, suppose that $(\pi_t)_{t\geq 0}$ denotes the marginal law of the solution to the Langevin diffusion (16.1). In the context of Bayesian optimal filtering and smoothing, [Sär07] proposed the following heuristic. Let (m_t, Σ_t) denote the mean and covariance matrix of π_t. Then, it can be checked (see §16.6.2) that

$$\begin{aligned} \dot{m}_t &= -\,\mathbb{E}\,\nabla V(X_t) \\ \dot{\Sigma}_t &= 2I - \mathbb{E}[\nabla V(X_t) \otimes (X_t - m_t) + (X_t - m_t) \otimes \nabla V(X_t)] \end{aligned} \tag{16.3}$$

where $X_t \sim \pi_t$. These ordinary differential equations (ODEs) are intractable because they involve expectations under the law π_t of X_t, which is not available to the practitioner. However, if we replace X_t with a Gaussian $Y_t \sim p_t = \mathsf{normal}(m_t, \Sigma_t)$ with the same mean and covariance as X_t, then the system of ODEs

$$\boxed{\begin{aligned} \dot{m}_t &= -\,\mathbb{E}\,\nabla V(Y_t) \\ \dot{\Sigma}_t &= 2I - \mathbb{E}[\nabla V(Y_t) \otimes (Y_t - m_t) + (Y_t - m_t) \otimes \nabla V(Y_t)] \end{aligned}} \tag{16.4}$$

yields a well-defined evolution of Gaussian distributions $(p_t)_{t\geq 0}$, which we may optimistically believe to be a good approximation of $(\pi_t)_{t\geq 0}$. Moreover, the system of ODEs can be numerically approximated efficiently in practice using Gaussian quadrature rules to compute the above expectations. This is the principle behind the unscented Kalman filter [JUD00].

In the context of the Langevin diffusion, Särkkä's heuristic (16.4) provides a promising avenue towards computational VI. Indeed, since $\pi \propto \exp(-V)$ is the unique stationary distribution of the Langevin diffusion (16.1), an algorithm to approximate $(\pi_t)_{t\geq 0}$ is expected to furnish an algorithm to solve the VI problem (16.2). However, at present there is little theoretical understanding of how the system (16.4) approximates (16.3); moreover, Särkkä's heuristic only provides Gaussian approximations, and it is unclear how to extend the system (16.4) to more complex models (e.g., mixtures of Gaussians).

Our contributions: bridging the gap via Wasserstein gradient flows. We show that the approximation $(p_t)_{t\geq 0}$ in Särkkä's heuristic (16.4) arises precisely as the gradient flow of the Kullback–Leibler (KL) divergence $\mathsf{KL}(\cdot \,\|\, \pi)$ on the Bures–Wasserstein space of Gaussian distributions on $\mathbb{R}^d$ endowed with the 2-Wasserstein distance

from optimal transport [Vil03]. This perspective allows us to not only understand its convergence but also to extend it to the richer space of mixtures of Gaussian distributions, and propose an implementation as a novel system of interacting "Gaussian particles". Below, we describe our contributions in greater detail.

Our framework builds upon the seminal work of [JKO98], which introduced the celebrated *JKO scheme* in order to give meaning to the idea that the evolving marginal law of the Langevin diffusion (16.1) is a gradient flow of $\mathsf{KL}(\cdot \,\|\, \pi)$ on the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures with finite second moments. Subsequently, in order to emphasize the Riemannian geometry underlying this result, [Ott01] developed his eponymous calculus on $\mathcal{P}_2(\mathbb{R}^d)$, a framework which has had tremendous impact in analysis, geometry, PDE, probability, and statistics.

Inspired by this perspective, we show in Theorem 16.3.1 that Särkkä's approximation $(p_t)_{t\geq 0}$ is also a gradient flow of $\mathsf{KL}(\cdot \,\|\, \pi)$, with the main difference being that it is *constrained* to lie on the submanifold $\mathsf{BW}(\mathbb{R}^d)$ of $\mathcal{P}_2(\mathbb{R}^d)$ consisting of Gaussian distributions, known as the Bures–Wasserstein manifold. In turn, our result paves the way for new theoretical understanding via the powerful theory of gradient flows. As a first step, using well-known results about convex functionals on the Wasserstein space, we show in Corollary 16.3.3 that $(p_t)_{t\geq 0}$ converges rapidly to the solution of the VI problem (16.2) with $\mathcal{P} = \mathsf{BW}(\mathbb{R}^d)$ as soon as V is convex. Moreover, as discussed in §16.4.1, we can apply numerical integration based on cubature rules for Gaussian integrals to the system of ODEs (16.4), thus arriving at a fast method with robust empirical performance.

This combination of results brings VI closer to Langevin-based MCMC both on the practical and theoretical fronts, but still falls short of achieving non-asymptotic discretization guarantees as pioneered by [Dal17b] for MCMC. To further close the theoretical gap between VI and the state of the art for MCMC, we propose in §16.4.2 a stochastic gradient descent (SGD) algorithm as a time discretization of the Bures–Wasserstein gradient flow. This algorithm comes with convergence guarantees that establish VI as a solid competitor to MCMC not only from a practical standpoint but also from a theoretical one. Both have their relative merits; whereas MCMC targets the true posterior, VI leads to fast computation of summary statistics of the approximation $\hat{\pi}$ to π.

In §16.5, we consider an extension of these ideas to the substantially more flexible class of mixtures of Gaussians. Namely, the space of mixtures of Gaussians can be identified as a Wasserstein space over $\mathsf{BW}(\mathbb{R}^d)$ and hence inherits Otto's differential calculus. Leveraging this viewpoint, in Theorem 16.5.1 we derive the gradient flow of $\mathsf{KL}(\cdot \,\|\, \pi)$ over the space of mixtures of Gaussians and propose to implement it via a system of interacting particles. Unlike typical particle-based algorithms, here our particles correspond to Gaussian distributions, and the col-

lection thereof to a Gaussian mixture which is better equipped to approximate a continuous measure. Although we focus on the VI problem in this work, we anticipate that our notion of "Gaussian particles" may be a broadly useful extension of classical particle methods for PDEs.

Related work. Classical VI methods define a parametric family $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ and minimize $\theta \mapsto \mathsf{KL}(p_\theta \,\|\, \pi)$ over $\theta \in \Theta$ using off-the-shelf optimization algorithms [PBJ12; RGB14]. Since (16.2) is an optimization problem over the space of probability distributions, we argue for methods that respect a natural geometry over this space. In this regard, previous approaches to VI using natural gradients implicitly employ a different geometry [LKS19b; Hua+22; KR22], namely the reparameterization-invariant Fisher–Rao geometry [AN00]. The application of Wasserstein gradient flows to VI was introduced earlier in work on normalizing flows and Stein Variational Gradient Descent (SVGD) [LW16; Liu17].

Our work falls in line with a number of recent papers aiming to place VI on a solid theoretical footing [ARC16; WB19; Dom20; KJD22]. Some of these works in particular have obtained non-asymptotic algorithmic guarantees for specific examples, see, e.g., [CB13]. We also mention that the approach we take in this paper is closely related to the algorithms and analysis arrived at in [AR20; Dom20; GPO21]. In particular, [GPO21] derive an algorithm for low-rank Gaussian VI by seeking a descent condition for the KL divergence, yielding a method resembling Algorithm 16.1 albeit without quantitative convergence guarantees. Also, [AR20; Dom20] show that parametrizing the Gaussian by the *square root* of the covariance matrix yields convexity and smoothness properties for the Gaussian VI objective, which in turn allows for applying Euclidean gradient methods. This choice of parametrization is closely related to the Bures–Wasserstein geometry approach we take, see §2.3 for background. However, we note that these works do not analyze the effect of stochastic gradients, which is crucial for implementation.

The connection between VI and Kalman filtering was studied in the static case by [LBB22b; LBB23], and extended to the dynamical case by [LBB22a], providing a first justification of Särkkä's heuristic in terms of local variational Gaussian approximation. In particular, the closest linear process to the Langevin diffusion (16.1) is a Gaussian process governed by a McKean–Vlasov equation whose Gaussian marginals have parameters evolving according to Särkkä's ODEs.

Constrained gradient flows on the Wasserstein space have also been extensively studied [CG03; CPR09; TW11; ENS17], although our interpretation of Särkkä's heuristic is, to the best of our knowledge, new.

16.2 Background

In order to define gradient flows on the space of probability measures, we must first endow this space with a geometry; see §2.1 for more details. Given probability measures μ and ν on $\mathbb{R}^d$, define the 2-*Wasserstein distance*

$$W_2(\mu,\nu) = \left[\inf_{\gamma\in\mathcal{C}(\mu,\nu)} \int \|x-y\|^2 \,\mathrm{d}\gamma(x,y)\right]^{1/2},$$

where $\mathcal{C}(\mu,\nu)$ is the set of *couplings* of μ and ν, that is, joint distributions on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are μ and ν respectively. This quantity is finite as long as μ and ν belong to the space $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures over $\mathbb{R}^d$ with finite second moments. The 2-Wasserstein distance has the interpretation of measuring the smallest possible mean squared displacement of mass required to *transport* μ to ν; we refer to [Vil03; Vil09b; San15] for textbook treatments on optimal transport. Unlike other notions of distance between probability measures, such as the total variation distance, the 2-Wasserstein distance respects the geometry of the underlying space $\mathbb{R}^d$, leading to numerous applications in modern data science [see, e.g., PC19].

The space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is a metric space [Vil03, Theorem 7.3], and we refer to it as the *Wasserstein space*. However, as shown by Otto [Ott01], it has a far richer geometric structure: formally, $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ can be viewed as a Riemannian manifold, a fact which allows for considering gradient flows of functionals on $\mathcal{P}_2(\mathbb{R}^d)$. A fundamental example of such a functional is the KL divergence $\mathsf{KL}(\cdot\,\|\,\pi)$ to a target density $\pi \propto \exp(-V)$ on $\mathbb{R}^d$, for which [JKO98] showed that the Wasserstein gradient flow is the same as the evolution of the marginal law of the Langevin diffusion (16.1). This optimization perspective has had tremendous impact on our understanding and development of MCMC algorithms [Wib18].

16.3 Variational inference with Gaussians

In this section we describe our problem using two equivalent approaches: a variational approach based on a modified version of the JKO scheme of [JKO98] (§16.3.1), and a Wasserstein gradient flow approach based on Otto calculus (§16.3.2). Both lead to the same result (§16.3.3). While the former is more accessible to readers who are unfamiliar with gradient flows on the Wasserstein space, the latter leads to strong convergence guarantees (§16.3.4).

16.3.1 Variational approach: the Bures–JKO scheme

The space of non-degenerate Gaussian distributions on $\mathbb{R}^d$ equipped with the W_2 distance forms the *Bures–Wasserstein space* $\mathsf{BW}(\mathbb{R}^d) \subseteq \mathcal{P}_2(\mathbb{R}^d)$. On $\mathsf{BW}(\mathbb{R}^d)$, the

Wasserstein distance $W_2^2(p_0, p_1)$ between two Gaussians $p_0 = \mathsf{normal}(m_0, \Sigma_0)$ and $p_1 = \mathsf{normal}(m_1, \Sigma_1)$ admits the following closed form:

$$W_2^2(p_0, p_1) = \|m_0 - m_1\|^2 + \mathcal{B}^2(\Sigma_0, \Sigma_1)\,, \tag{16.5}$$

where $\mathcal{B}^2(\Sigma_0, \Sigma_1) = \mathrm{tr}(\Sigma_0 + \Sigma_1 - 2\,(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}})^{\frac{1}{2}})$ is the squared Bures metric [Bur69].

Given a target density $\pi \propto \exp(-V)$ on $\mathbb{R}^d$, and with a step size $h > 0$, we may define the iterates of the proximal point algorithm

$$p_{k+1,h} \coloneqq \operatorname*{arg\,min}_{p \in \mathsf{BW}(\mathbb{R}^d)} \left\{ \mathsf{KL}(p \,\|\, \pi) + \frac{1}{2h}\, W_2^2(p, p_{k,h}) \right\}. \tag{16.6}$$

Using (16.5), this is an explicit optimization problem involving the mean and covariance matrix of p. Although (16.6) is not solvable in closed form, by letting $h \searrow 0$ we obtain a limiting curve $(p_t)_{t\geq 0}$ via $p_t = \lim_{h\searrow 0} p_{\lfloor t/h \rfloor, h}$, which can be interpreted as the Bures–Wasserstein gradient flow of the KL divergence $\mathsf{KL}(\cdot \,\|\, \pi)$. This procedure mimics the JKO scheme [JKO98] with the additional constraint that the iterates lie in $\mathsf{BW}(\mathbb{R}^d)$, and we therefore call it the Bures–JKO scheme.

■ 16.3.2 Geometric approach: the Bures–Wasserstein gradient flow of the KL divergence

In the formal sense of Otto described above, $\mathsf{BW}(\mathbb{R}^d)$ is a submanifold of $\mathcal{P}_2(\mathbb{R}^d)$. Moreover, since Gaussians can be parameterized by their mean and covariance, $\mathsf{BW}(\mathbb{R}^d)$ can be identified with the manifold $\mathbb{R}^d \times \mathbf{S}^d_{++}$, where $\mathbf{S}^d_{++}$ is the cone of symmetric positive definite $d \times d$ matrices. Hence, $\mathsf{BW}(\mathbb{R}^d)$ is a genuine Riemannian manifold in its own right [see Mod17; MMP18; BJL19], and gradient flows can be defined using Riemannian geometry [Car92]. See §2.3 for more details. Since the functional $\mu \mapsto \mathcal{F}(\mu) = \mathsf{KL}(\mu \,\|\, \pi)$ defined over $\mathcal{P}_2(\mathbb{R}^d)$ restricts to a functional over $\mathsf{BW}(\mathbb{R}^d)$, we can also consider the gradient flow of $\mathcal{F}$ over the Bures–Wasserstein space; note that this latter gradient flow is necessarily a curve $(p_t)_{t\geq 0}$ such that each p_t is a Gaussian measure.

■ 16.3.3 Variational inference via the Bures–Wasserstein gradient flow

Using either approach, we can prove the following theorem.

Theorem 16.3.1. *Let $\pi \propto \exp(-V)$ be the target density on $\mathbb{R}^d$. Then, the limiting curve $(p_t)_{t\geq 0}$ where $p_t = \mathsf{normal}(m_t, \Sigma_t)$ is obtained via the Bures–JKO scheme (16.6), or equivalently, the Bures–Wasserstein gradient flow $(p_t)_{t\geq 0}$ of the KL divergence $\mathsf{KL}(\cdot \,\|\, \pi)$, satisfies Särkkä's system of ODEs (16.4).*

Proof. For the proof using the Bures–JKO scheme, see [Lam+22, Appendix A]. The proof using Otto calculus is presented in §16.7. □

This theorem shows that Särkkä's heuristic (16.4) precisely yields the Wasserstein gradient flow of the KL divergence over the submanifold $\mathsf{BW}(\mathbb{R}^d)$. Equipped with this interpretation, we are now able to obtain information about the asymptotic behavior of the approximation $(p_t)_{t\geq 0}$. Namely, we can hope that it converges to constrained minimizer $\hat{\pi} = \arg\min_{p\in\mathsf{BW}(\mathbb{R}^d)} \mathsf{KL}(p \,\|\, \pi)$, i.e., precisely the solution to the VI problem (16.2). In the next section, we show that this convergence in fact holds as soon as V is convex, and moreover with quantitative rates.

The solution $\hat{\pi}$ to (16.2), and consequently the limit point of Särkkä's approximation, is well-studied in the variational inference literature [see, e.g., OA09], and we recall standard facts about $\hat{\pi}$ here for completeness. It is known that $\hat{\pi}$ satisfies the equations

$$\mathbb{E}_{\hat{\pi}} \nabla V = 0 \qquad \text{and} \qquad \mathbb{E}_{\hat{\pi}} \nabla^2 V = \hat{\Sigma}^{-1}, \tag{16.7}$$

where $\hat{\Sigma}$ is the covariance matrix of $\hat{\pi}$ (these equations can also be derived as first-order necessary conditions by setting the Bures–Wasserstein gradient derived in §16.7 to zero). In particular, it follows from (16.7) that if $\nabla^2 V$ enjoys the bounds $\alpha I \preceq \nabla^2 V \preceq \beta I$ for some $-\infty \leq \alpha \leq \beta \leq \infty$, then any solution $\hat{\pi}$ to the constrained problem also satisfies $\beta^{-1}\, I \preceq \hat{\Sigma} \preceq (\alpha \vee 0)^{-1}\, I$.

■ 16.3.4 Continuous-time convergence

Besides providing an intuitive interpretation of Särkkä's heuristic, Theorem 16.3.1 readily yields convergence criteria for the system (16.4) which rest upon general principles for gradient flows. We begin with a key observation. For a functional $\mathcal{F} : \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ and $\alpha \in \mathbb{R}$, we say that $\mathcal{F}$ is *α-convex* if for all constant-speed geodesics $(p_t)_{t\in[0,1]}$ in $\mathsf{BW}(\mathbb{R}^d)$,

$$\mathcal{F}(p_t) \leq (1-t)\,\mathcal{F}(p_0) + t\,\mathcal{F}(p_1) - \frac{\alpha\, t\,(1-t)}{2}\, W_2^2(p_0, p_1)\,, \qquad t \in [0,1]\,.$$

Lemma 16.3.2. *For any $\alpha \in \mathbb{R}$, if $\nabla^2 V \succeq \alpha I$, then $\mathsf{KL}(\cdot \,\|\, \pi)$ is α-convex on $\mathsf{BW}(\mathbb{R}^d)$.*

Proof. The assumption that $\nabla^2 V \succeq \alpha I$ entails that the functional $\mathsf{KL}(\cdot \,\|\, \pi)$ is α-convex on the entire Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ [see, e.g., Vil09b, Theorem 17.15]. Since $\mathsf{BW}(\mathbb{R}^d)$ is a geodesically convex subset of $\mathcal{P}_2(\mathbb{R}^d)$ (see §2.3), then the geodesics in $\mathsf{BW}(\mathbb{R}^d)$ agree with the geodesics in $\mathcal{P}_2(\mathbb{R}^d)$, from which it follows that $\mathsf{KL}(\cdot \,\|\, \pi)$ is α-convex on $\mathsf{BW}(\mathbb{R}^d)$. □

Consequently, we obtain the following corollary; see §16.8 for the proof.

Corollary 16.3.3. *Suppose that* $\nabla^2 V \succeq \alpha I$ *for some* $\alpha \in \mathbb{R}$. *Then, for any* $p_0 \in \mathsf{BW}(\mathbb{R}^d)$, *there is a unique solution to the* $\mathsf{BW}(\mathbb{R}^d)$ *gradient flow of* $\mathsf{KL}(\cdot \,\|\, \pi)$ *started at* p_0. *Moreover:*
1. *If* $\alpha > 0$, *then for all* $t \geq 0$, $W_2^2(p_t, \hat{\pi}) \leq \exp(-2\alpha t)\, W_2^2(p_0, \hat{\pi})$.
2. *If* $\alpha > 0$, *then for all* $t \geq 0$, $\mathsf{KL}(p_t \,\|\, \pi) - \mathsf{KL}_\star \leq \exp(-2\alpha t)\, \{\mathsf{KL}(p_0 \,\|\, \pi) - \mathsf{KL}_\star\}$.
3. *If* $\alpha = 0$, *then for all* $t > 0$, $\mathsf{KL}(p_t \,\|\, \pi) - \mathsf{KL}_\star \leq \frac{1}{2t}\, W_2^2(p_0, \hat{\pi})$.

Here, $\mathsf{KL}_\star := \mathsf{KL}(\hat{\pi} \,\|\, \pi)$.

The assumption that $\nabla^2 V \succeq \alpha I$ for some $\alpha > 0$, i.e., that π is *strongly log-concave*, is a standard assumption in the MCMC literature. Under this same assumption, Corollary 16.3.3 yields convergence for the Bures–Wasserstein gradient flow of $\mathsf{KL}(\cdot \,\|\, \pi)$; however, the flow must first be discretized in time for implementation. If we assume additionally that the smoothness condition $\nabla^2 V \preceq \beta I$ holds, then a surge of recent research has succeeded in obtaining precise non-asymptotic guarantees for discretized MCMC algorithms. In §16.4.2 below, we will show how to do the same for VI.

16.4 Time discretization of the Bures–Wasserstein gradient flow

We are now equipped with dual perspectives on a dynamical solution to Gaussian VI: ODE and gradient flow. Each perspective leads to a different implementation. On the one hand, we discretize the system of ODEs defined in (16.4) using numerical integration. On the other, we discretize the gradient flow using stochastic gradient descent in the Bures–Wasserstein space.

16.4.1 Numerical integration of the ODEs

The system of ODEs (16.4) can be integrated in time using a classical Runge–Kutta scheme. The expectations under a Gaussian support are approximated by cubature rules used in Kalman filtering [AH09]. Moreover, a square root version of the ODE is also considered to ensure that covariance matrices remain symmetric and positive. See [Lam+22] for implementation and numerical experiments.

16.4.2 Bures–Wasserstein SGD and theoretical guarantees for VI

Although the ODE discretization proposed in the preceding section enjoys strong empirical performance, it is unclear how to quantify its impact on the convergence rates established in Corollary 16.3.3. Therefore, we now propose a stochastic gradient descent algorithm over the Bures–Wasserstein space, for which useful

analysis tools have been developed (see §15). This approach bypasses the use of the system of ODEs (16.4), and instead discretizes the Bures–Wasserstein gradient flow directly. Under the standard assumption of strong log-concavity and log-smoothness, it leads to an algorithm (Algorithm 16.1) for approximating $\hat{\pi}$ with provable convergence guarantees.

Algorithm 16.1 Bures–Wasserstein SGD

Require: strong convexity parameter $\alpha > 0$; step size $h > 0$; mean m_0 and covariance matrix Σ_0

for $k = 1, \dots, N$ **do**

draw a sample $\hat{X}_k \sim p_k$

set $m_{k+1} \leftarrow m_k - h \nabla V(\hat{X}_k)$

set $M_k \leftarrow I - h\,(\nabla^2 V(\hat{X}_k) - \Sigma_k^{-1})$

set $\Sigma_k^+ \leftarrow M_k \Sigma_k M_k$

set $\Sigma_{k+1} \leftarrow \mathrm{clip}^{1/\alpha}\, \Sigma_k^+$

Algorithm 16.1 maintains a sequence of Gaussian distributions $(p_k)_{k \in \mathbb{N}}$; here (m_k, Σ_k) denote the mean vector and covariance matrix at iteration k (see §16.9 for a derivation of the algorithm as SGD in the Bures–Wasserstein space). The clipping operator clip^τ, which is introduced purely for the purpose of theoretical analysis, simply truncates the eigenvalues from above; see §16.9. Our theoretical result for VI is given as the following theorem, whose proof is deferred to §16.9.

Theorem 16.4.1. *Assume that* $0 \prec \alpha I \preceq \nabla^2 V \preceq I$. *Also, assume that* $h \le \frac{\alpha^2}{60}$ *and that we initialize Algorithm 16.1 at a matrix satisfying* $\frac{\alpha}{9} I \preceq \Sigma_{\mu_0} \preceq \frac{1}{\alpha} I$. *Then, for all* $k \in \mathbb{N}$,

$$\mathbb{E}\, W_2^2(p_k, \hat{\pi}) \le \exp(-\alpha k h)\, W_2^2(p_0, \hat{\pi}) + \frac{36 d h}{\alpha^2}\,.$$

In particular, we obtain $\mathbb{E}\, W_2^2(p_k, \hat{\pi}) \le \varepsilon^2$ *provided we set* $h \asymp \frac{\alpha^2 \varepsilon^2}{d}$ *and the number of iterations to be* $k \gtrsim \frac{d}{\alpha^3 \varepsilon^2} \log(W_2(p_0, \hat{\pi})/\varepsilon)$.

The upper bound $\nabla^2 V \preceq I$ is notationally convenient for our proof but not necessary; in any case, any strongly log-concave and log-smooth density π can be rescaled so that the assumption holds.

Theorem 16.4.1 is similar in flavor to modern results for MCMC, both in terms of the assumptions (Hessian bounds and query access to the derivatives[1] of V) and

[1] A notable downside of Algorithm 16.1 is the requirement of a Hessian oracle for V, which results in a higher per-iteration cost than typical MCMC samplers.

the conclusion (a non-asymptotic polynomial-time algorithmic guarantee). We hope that such an encouraging result for VI will prompt more theoretical studies aimed at closing the gap between the two approaches.

16.5 Variational inference with mixtures of Gaussians

Thus far, we have shown that the tractability of Gaussians can be readily exploited in the context of Bures–Wasserstein gradient flows and translated into useful results for variational inference. Nevertheless, these results are limited by the lack of expressivity of Gaussians, namely their inability to capture complex features such as multimodality and, more generally, heterogeneity. To overcome this limitation, mixtures of Gaussians arise as a natural and powerful alternative; indeed, universal approximation of arbitrary probability measures by mixtures of Gaussians is well-known [see, e.g., DD20]. As we show next, the space of mixtures of Gaussians can also be equipped with a Wasserstein structure which gives rise to implementable gradient flows.

16.5.1 Geometry of the space of mixtures of Gaussians

We begin with the key observation already made by [CGT19], that any mixture of Gaussians can be canonically identified with a probability distribution (the mixing distribution) over the parameter space $\Theta = \mathbb{R}^d \times \mathbf{S}^d_{++}$ (the space of means and covariance matrices). Explicitly a probability measure $\mu \in \mathcal{P}(\Theta)$ corresponds to a Gaussian mixture as follows:

$$\mu \qquad \leftrightarrow \qquad \mathsf{p}_\mu := \int p_\theta \, \mathrm{d}\mu(\theta) \,, \tag{16.8}$$

where p_θ is the Gaussian distribution with parameters $\theta \in \Theta$. Equivalently, μ can be thought of as a probability measure over $\mathsf{BW}(\mathbb{R}^d)$, and hence the space of Gaussian mixtures on $\mathbb{R}^d$ can be identified with the Wasserstein space $\mathcal{P}_2(\mathsf{BW}(\mathbb{R}^d))$ over the Bures–Wasserstein space which is endowed with the distance (16.5) between Gaussian measures. Indeed, the theory of optimal transport can be developed with any Riemannian manifold (rather than $\mathbb{R}^d$) as the base space [Vil09b]. As before, the space $\mathcal{P}_2(\mathsf{BW}(\mathbb{R}^d))$ is endowed with a formal Riemannian structure, which respects the geometry of the base space $\mathsf{BW}(\mathbb{R}^d)$, and we can consider Wasserstein gradient flows over $\mathcal{P}_2(\mathsf{BW}(\mathbb{R}^d))$.

This framework encompasses both discrete mixtures of Gaussians (when μ is a discrete measure) and continuous mixtures of Gaussians. In the case when the mixing measure μ is discrete, the geometry of $\mathcal{P}_2(\mathsf{BW}(\mathbb{R}^d))$ was studied by [CGT19; DD20]. An important insight of our work, however, is that it is fruitful to consider

the full space $\mathcal{P}_2(\mathsf{BW}(\mathbb{R}^d))$ for deriving gradient flows, even if we eventually develop algorithms which propagate a finite number of mixture components.

■ 16.5.2 Gradient flow of the KL divergence and particle discretization

We consider the gradient flow of the KL divergence functional

$$\mu \mapsto \mathcal{F}(\mu) \coloneqq \mathsf{KL}(\mathsf{p}_\mu \,\|\, \pi) \tag{16.9}$$

over the space $\mathcal{P}_2(\mathsf{BW}(\mathbb{R}^d))$. The proof of the following theorem is given in §16.10.

Theorem 16.5.1. *The gradient flow $(\mu_t)_{t\geq 0}$ of the functional $\mathcal{F}$ defined in (16.9) over $\mathcal{P}_2(\mathsf{BW}(\mathbb{R}^d))$ can be described as follows. Let $\theta_0 = (m_0, \Sigma_0) \sim \mu_0$, and let $\theta_t = (m_t, \Sigma_t)$ evolve according to the ODE*

$$\boxed{\begin{aligned} \dot{m}_t &= -\,\mathbb{E}\,\nabla \ln \frac{\mathsf{p}_{\mu_t}}{\pi}(Y_t) \\ \dot{\Sigma}_t &= -\,\mathbb{E}\,\nabla^2 \ln \frac{\mathsf{p}_{\mu_t}}{\pi}(Y_t)\,\Sigma_t - \Sigma_t\,\mathbb{E}\,\nabla^2 \ln \frac{\mathsf{p}_{\mu_t}}{\pi}(Y_t) \end{aligned}} \tag{16.10}$$

where $Y_t \sim \mathsf{normal}(m_t, \Sigma_t)$. Then $\theta_t \sim \mu_t$.

The gradient flow in Theorem 16.5.1 describes the evolution of a particle θ_t which describes the parameters of a Gaussian measure, hence the name *Gaussian particle*. The intuition behind this evolution is as follows. Suppose we draw infinitely many initial particles (each being a Gaussian) from μ_0. By evolving all those particles through (16.10), which interact with each other via the term p_{μ_t}, they tend to aggregate in some parts of the space of Gaussian parameters and spread out in others. This distribution of Gaussian particles is precisely the mixing measure μ_t, which, in turn, corresponds to a Gaussian mixture. Since an infinite number of Gaussian particles is impractical, consider initializing this evolution at a finitely supported distribution μ_0, thus corresponding to a more familiar Gaussian mixture model with a finite number of components:

$$\mu_0 = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_0^{(i)}} = \frac{1}{N}\sum_{i=1}^N \delta_{(m_0^{(i)}, \Sigma_0^{(i)})} \qquad \leftrightarrow \qquad \mathsf{p}_{\mu_0} \coloneqq \frac{1}{N}\sum_{i=1}^N p_{(m_0^{(i)}, \Sigma_0^{(i)})}\,.$$

Interestingly, it can be readily checked that the system of ODEs (16.10) thus initialized maintains a finite mixture distribution:

$$\mu_t = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_t^{(i)}} = \frac{1}{N}\sum_{i=1}^N \delta_{(m_t^{(i)}, \Sigma_t^{(i)})}\,,$$

where the parameters $\theta_t^{(i)} = (m_t^{(i)}, \Sigma_t^{(i)})$ evolve according to the following interacting particle system, for $i \in [N]$

$$\dot{m}_t^{(i)} = -\mathbb{E}\, \nabla \ln \frac{\mathsf{p}_{\mu_t}}{\pi}(Y_t^{(i)})\,, \tag{16.11}$$

$$\dot{\Sigma}_t^{(i)} = -\mathbb{E}\, \nabla^2 \ln \frac{\mathsf{p}_{\mu_t}}{\pi}(Y_t^{(i)})\, \Sigma_t^{(i)} - \Sigma_t^{(i)}\, \mathbb{E}\, \nabla^2 \ln \frac{\mathsf{p}_{\mu_t}}{\pi}(Y_t^{(i)})\,, \tag{16.12}$$

where $Y_t^{(i)} \sim p_{\theta_t^{(i)}}$. This finite system of particles can now be implemented using the same numerical tools as for Gaussian VI. Note that due to this property of the dynamics, we can hope at best to converge to the best mixture of N Gaussians approximating π, but this approximation error is expected to vanish as $N \to \infty$. Also, similarly to (16.4), it is possible to write down Hessian-free updates using integration by parts.

The above system of particles may also be derived using a proximal point method similar to the Bures–JKO scheme, see §16.3.1. Indeed, infinitesimally, it has the variational interpretation

$$(\theta_{t+h}^{(1)}, \dots, \theta_{t+h}^{(N)}) \approx \operatorname*{arg\,min}_{\theta^{(1)},\dots,\theta^{(N)} \in \Theta} \left\{ \mathsf{KL}\Big(\frac{1}{N} \sum_{i=1}^N p_{\theta^{(i)}} \;\Big\|\; \pi \Big) + \frac{1}{2Nh} \sum_{i=1}^N W_2^2(p_{\theta^{(i)}}, p_{\theta_t^{(i)}}) \right\}.$$

Reassuringly, (16.11)–(16.12) reduce to (16.4) when $\mu_0 = \delta_{(m_0,\Sigma_0)}$ is a point mass, indicating that the theorem provides a natural extension of our previous results. However, although the model (16.8) is substantially more expressive than the Gaussian VI considered in §16.3, it has the downside that we lose many of the theoretical guarantees. For example, even when V is convex, the objective functional $\mathcal{F}$ considered here need not be convex; see §16.11. We nevertheless validate the practical utility of our approach in experiments in [Lam+22].

Unlike typical interacting particle systems which arise from discretizations of Wasserstein gradient flows, at each time t, the distribution p_{μ_t} is continuous. This extension provides considerably more flexibility—from a mixture of point masses to a mixture of Gaussians—compared to interacting particle-based algorithms hitherto considered for either sampling [LW16; Liu17; Che+20d; DNS23], or solving partial differential equations [Car+11; Car+12; Bon+15; CB16; CCP19; Cra+23].

■ 16.5.3 Time-varying weights with the Wasserstein–Fisher–Rao geometry

One notable shortcoming of the system (16.11)–(16.12) is that the weights of the mixture are held fixed, which can inhibit the Gaussian particles from quickly moving between separated modes of π. It is therefore desirable to design a principled algorithm which also allows for the mixture weights to be updated.

Towards this end, in §16.12 we derive the gradient flow of the KL divergence with respect to the Wasserstein–Fisher–Rao geometry [LMS16; Chi+18; LMS18], which yields an interacting system of Gaussian particles with changing weights. The equations are given as follows: at each time t, the mixing measure is the discrete measure

$$\mu_t = \sum_{i=1}^{N} w_t^{(i)} \delta_{(m_t^{(i)}, \Sigma_t^{(i)})} .$$

Let $Y_t^{(i)} \sim p_{m_t^{(i)}, \Sigma_t^{(i)}}$, and let $r_t^{(i)} = \sqrt{w_t^{(i)}}$. Then, the system of ODEs is given by

$$\begin{aligned}
\dot{m}_t^{(i)} &= -\mathbb{E}\, \nabla \ln \frac{\mathsf{p}_{\mu_t}}{\pi}(Y_t^{(i)}) , \\
\dot{\Sigma}_t^{(i)} &= -\mathbb{E}\, \nabla^2 \ln \frac{\mathsf{p}_{\mu_t}}{\pi}(Y_t^{(i)})\, \Sigma_t^{(i)} - \Sigma_t^{(i)}\, \mathbb{E}\, \nabla^2 \ln \frac{\mathsf{p}_{\mu_t}}{\pi}(Y_t^{(i)}) , \\
\dot{r}_t^{(i)} &= -\Big(\mathbb{E} \ln \frac{\mathsf{p}_{\mu_t}}{\pi}(Y_t^{(i)}) - \frac{1}{N} \sum_{j=1}^{N} \mathbb{E} \ln \frac{\mathsf{p}_{\mu_t}}{\pi}(Y_t^{(j)}) \Big)\, r_t^{(i)} .
\end{aligned}$$

We have implemented these equations and their empirical performance is encouraging. However, a fuller investigation of algorithms for VI with changing weights is beyond the scope of this work and we leave it for future research.

16.6 Background on Otto calculus

We refer to §2.1 and §2.3 for the main background.

16.6.1 The Bures–Wasserstein space

The space of non-degenerate Gaussian distributions equipped with the W_2 metric is known as the Bures–Wasserstein space, after [Bur69]. We denote this space as $\mathsf{BW}(\mathbb{R}^d)$. Background on the geometry of $\mathsf{BW}(\mathbb{R}^d)$ is given in §2.3; here, we recall some notions in order to fix notation when dealing with non-centered Gaussians.

Given $m \in \mathbb{R}^d$ and $\Sigma \succ 0$, we denote by $p_{m,\Sigma}$ the Gaussian on $\mathbb{R}^d$ with mean m and covariance Σ. Conversely, for a non-degenerate Gaussian p we write (m_p, Σ_p) for its mean and covariance. Via this correspondence, we can therefore identify the space of non-degenerate Gaussians with the manifold $\mathbb{R}^d \times \mathbf{S}^d_{++}$, where $\mathbf{S}^d_{++}$ denotes the cone of positive definite matrices. Abusing notation, we will do so whenever there is no danger of confusion.

Suppose that $p_{m_0,\Sigma_0}, p_{m_1,\Sigma_1} \in \mathsf{BW}(\mathbb{R}^d)$. Then, the optimal transport map from $p_0 \coloneqq p_{m_0,\Sigma_0}$ to $p_1 \coloneqq p_{m_1,\Sigma_1}$ is

$$\nabla\varphi(x) = m_1 + \Sigma_0^{-1/2} \left(\Sigma_0^{1/2} \Sigma_1 \Sigma_0^{1/2}\right)^{1/2} \Sigma_0^{-1/2} (x - m_0) .$$

Observe that $\nabla\varphi$ is an affine map. Since the pushforward of a Gaussian via an affine map is also Gaussian, it follows from Definition 2.1.3 that the constant speed geodesic $(p_t)_{t\in[0,1]}$ joining p_0 to p_1 also lies in $\mathsf{BW}(\mathbb{R}^d)$. In other words, $\mathsf{BW}(\mathbb{R}^d)$ is a *geodesically convex* subset of $\mathcal{P}_2(\mathbb{R}^d)$.

The tangent vector to the geodesic at time 0 is always an affine map of the form $x \mapsto a + S\,(x - m_{p_0})$, where $a \in \mathbb{R}^d$ and S is a symmetric matrix. The tangent space is

$$T_p\mathsf{BW}(\mathbb{R}^d) = \{x \mapsto a + S\,(x - m_p) \mid a \in \mathbb{R}^d,\ S \in \mathbf{S}^d\}\,,$$

which can therefore be identified with pairs $(a, S) \in \mathbb{R}^d \times \mathbf{S}^d$. With this abuse of notation, if $(a, S), (a', S') \in T_p\mathsf{BW}(\mathbb{R}^d)$, then

$$\begin{aligned}\langle (a,S),(a',S')\rangle_p &= \int \langle a + S\,(x - m_p), a' + S'\,(x - m_p)\rangle \,\mathrm{d}p(x)\\ &= \langle a, a'\rangle + \langle S, \Sigma_p S'\rangle\,. \end{aligned} \tag{16.13}$$

Specializing the notions from §2.1, we obtain

$$\begin{aligned}\log_p(q) &= \big(m_q - m_p,\ \Sigma_p^{-1/2}\,(\Sigma_p^{1/2}\Sigma_q\Sigma_p^{1/2})^{1/2}\,\Sigma_p^{-1/2} - I\big)\,,\\ \exp_p(a,S) &= \big(m_p + a + (S+I)\,(\cdot - m_p)\big)_{\#}p\\ &= \mathsf{normal}\big(m_p + a,\ (S+I)\,\Sigma_p\,(S+I)\big)\,.\end{aligned}$$

Here, $\exp_p(a, S)$ is defined if $S \succ -I$.

This definition of the tangent space is consistent with the Wasserstein space, in that we have the inclusion $T_p\mathsf{BW}(\mathbb{R}^d) \hookrightarrow T_p\mathcal{P}_2(\mathbb{R}^d)$, but the abuse of notation $T_p\mathsf{BW}(\mathbb{R}^d) = \mathbb{R}^d \times \mathbf{S}^d$ can sometimes cause confusion. Indeed, if $(p_t = p_{m_t,\Sigma_t})_{t\in[0,1]}$ is a constant-speed geodesic in $\mathsf{BW}(\mathbb{R}^d)$, and the tangent vector at time 0 is (a, S),

$$p_t = \exp_{p_0}\big(t\,(a,S)\big) = \mathsf{normal}\big(m_p + ta,\ (tS+I)\,\Sigma_p\,(tS+I)\big)\,.$$

In particular, $\Sigma_t \neq \Sigma_0 + t\,(S - I)$, and

$$\dot m_0 = a\,, \tag{16.14}$$

$$\dot\Sigma_0 = S\Sigma_0 + \Sigma_0 S\,. \tag{16.15}$$

Although we derived the equations (16.14) and (16.15) for geodesic curves, they also hold for any curve $(p_t)_{t\geq 0}$ with tangent vector equal to (a, S) at time 0. Using this, we can derive an expression for the Bures–Wasserstein gradient $\nabla_{\mathsf{BW}} f$ of a function $f : \mathbb{R}^d \times \mathbf{S}^d_{++} \to \mathbb{R}$. By definition, this satisfies, for any curve $(m_t, \Sigma_t)_{t\geq 0}$ with tangent vector (a, S) at time 0,

$$\langle \nabla_{\mathsf{BW}} f(m_0,\Sigma_0), (a,S)\rangle_{p_{m_0,\Sigma_0}} = \partial_t\big|_{t=0} f(m_t,\Sigma_t)\,.$$

Write $(\bar{a}, \bar{S}) = \nabla_{\mathsf{BW}} f(m_0, \Sigma_0)$. Then, we want

$$\begin{aligned}\langle \bar{a}, a\rangle + \langle \bar{S}, \Sigma_0 S\rangle &= \langle \nabla_m f(m_0,\Sigma_0), \dot{m}_0\rangle + \langle \nabla_\Sigma f(m_0,\Sigma_0), \dot{\Sigma}_0\rangle \\ &= \langle \nabla_m f(m_0,\Sigma_0), a\rangle + 2\,\langle \nabla_\Sigma f(m_0,\Sigma_0), \Sigma_0 S\rangle\,,\end{aligned}$$

where ∇_m, ∇_Σ denote the usual Euclidean gradients. Hence, by identification, we conclude that the Bures–Wasserstein gradient of f is related to the Euclidean gradient of f via

$$\nabla_{\mathsf{BW}} f(m,\Sigma) = \big(\nabla_m f(m,\Sigma),\ 2\,\nabla_\Sigma f(m,\Sigma)\big)\,. \tag{16.16}$$

■ 16.6.2 Evolution of the mean and covariance along the Fokker–Planck equation

It is known that the Wasserstein gradient of $\mathcal{F} \coloneqq \mathsf{KL}(\cdot \,\|\, \pi)$ is

$$\nabla_{W_2} \mathcal{F}(\mu) = \nabla \ln \frac{\mu}{\pi}\,. \tag{16.17}$$

See, e.g., [AGS08, Theorem 10.4.13]. Also, as shown by [JKO98], the Langevin diffusion is the gradient flow of $\mathsf{KL}(\cdot \,\|\, \pi)$. In Otto calculus, this means that the law $(\pi_t)_{t\geq 0}$ of the Langevin diffusion obeys the continuity equation (2.5) with velocity vector field $v_t = -\nabla_{W_2}\mathcal{F}(\pi_t) = -\nabla \ln(\pi_t/\pi)$, which is consistent with the Fokker–Planck equation (2.13).

According to the particle interpretation (2.4) of dynamics in the Wasserstein space, if $x_0 \sim \pi_0$ and

$$\dot{x}_t = v_t(x_t) = -\nabla \ln \frac{\pi_t}{\pi}(x_t)\,,$$

then $x_t \sim \pi_t$. Note that $(x_t)_{t\geq 0}$ is *not* the Langevin diffusion (16.1) as it is the solution to a deterministic ODE (albeit with random initial condition), but the marginal law of $(x_t)_{t\geq 0}$ agrees with that of the Langevin diffusion. This provides a convenient tool for calculating the evolution of the mean and covariance along the Fokker–Planck equation, as we now demonstrate.

The evolution of the mean is

$$\dot{m}_t = \partial_t\, \mathbb{E}\, x_t = \mathbb{E}\, \dot{x}_t = -\,\mathbb{E}\, \nabla \ln \frac{\pi_t}{\pi}(x_t)\,.$$

Since $\mathbb{E}\, \nabla \ln \pi_t(x_t) = 0$ (which is verified via integration by parts), and $\pi \propto \exp(-V)$, this can also be written as

$$\dot{m}_t = -\,\mathbb{E}_{\pi_t}\, \nabla V\,.$$

Next, for the evolution of the covariance,

$$\begin{aligned}\partial_t \mathbb{E}(x_t \otimes x_t) &= \mathbb{E}(x_t \otimes \dot{x}_t + \dot{x}_t \otimes x_t) \\ &= -\mathbb{E}\big(x_t \otimes \nabla \ln \frac{\pi_t}{\pi}(x_t) + \nabla \ln \frac{\pi_t}{\pi}(x_t) \otimes x_t\big) \\ \partial_t \mathbb{E}(x_t) \otimes \mathbb{E}(x_t) &= m_t \otimes \mathbb{E}(\dot{x}_t) + \mathbb{E}(\dot{x}_t) \otimes m_t \\ &= -\mathbb{E}\big(m_t \otimes \nabla \ln \frac{\pi_t}{\pi}(x_t) + \nabla \ln \frac{\pi_t}{\pi}(x_t) \otimes m_t\big)\end{aligned}$$

which yields

$$\dot{\Sigma}_t = -\mathbb{E}\big((x_t - m_t) \otimes \nabla \ln \frac{\pi_t}{\pi}(x_t) + \nabla \ln \frac{\pi_t}{\pi}(x_t) \otimes (x_t - m_t)\big)\,.$$

Integration by parts yields

$$\begin{aligned}&\int (\bullet - m_t) \otimes \nabla \ln \pi_t \,\mathrm{d}\pi_t + \int \nabla \ln \pi_t \otimes (\bullet - m_t)\,\mathrm{d}\pi_t \\ &\quad = \int (\bullet - m_t) \otimes \nabla \pi_t + \int \nabla \pi_t \otimes (\bullet - m_t) = -2I\,.\end{aligned}$$

Hence,

$$\dot{\Sigma}_t = 2I - \mathbb{E}_{\pi_t}[\nabla V \otimes (\bullet - m_t) + (\bullet - m_t) \otimes \nabla V]\,.$$

This verifies equation (16.3). The equations in this section can also be derived using Itô calculus.

■ 16.7 Proofs via Otto calculus

Our aim in this section is to derive the Wasserstein gradient flow of the KL divergence $\mathsf{KL}(\cdot \,\|\, \pi)$ *constrained* to lie in the Bures–Wasserstein space of non-degenerate Gaussian measures.

Since the Bures–Wasserstein space can be formally viewed as a submanifold of the Wasserstein space, it leads to two natural approaches for computing the constrained gradient flow. In the first approach, we take the Wasserstein gradient of $\mathsf{KL}(\cdot\|\pi)$ and we compute the orthogonal projection onto the tangent space of the Bures–Wasserstein space. In the second approach, we note that the geometry of the Bures–Wasserstein space has been studied in its own right [see, e.g., BJL19] and in particular, the explicit expression (16.16) for the Bures–Wasserstein gradient is known. We can therefore view $\mathsf{KL}(\cdot \,\|\, \pi)$ as a functional over $\mathsf{BW}(\mathbb{R}^d)$ and compute its gradient directly using (16.16).

16.7.1 Orthogonal projection approach

First, we justify why computing the orthogonal projection of the $\mathcal{P}_2(\mathbb{R}^d)$ gradient gives the same result as computing the intrinsic gradient on $\mathsf{BW}(\mathbb{R}^d)$. Let $\mathcal{F}$ be any functional on $\mathcal{P}_2(\mathbb{R}^d)$. By definition, the BW gradient $\nabla_{\mathsf{BW}}\mathcal{F}$ satisfies

$$\partial_t \mathcal{F}(p_t) = \langle \nabla_{\mathsf{BW}} \mathcal{F}(p_t), v_t \rangle_{p_t} \tag{16.18}$$

for any curve $(p_t)_{t\in\mathbb{R}}$ in $\mathsf{BW}(\mathbb{R}^d)$ with tangent vectors $(v_t)_{t\in\mathbb{R}}$. Here, $\nabla_{\mathsf{BW}}\mathcal{F}(p_t) \in T_{p_t}\mathsf{BW}(\mathbb{R}^d)$. On the other hand, since $(p_t)_{t\geq 0}$ is also a curve in $\mathcal{P}_2(\mathbb{R}^d)$ and the Riemannian structure of $\mathsf{BW}(\mathbb{R}^d)$ is consistent with that of $\mathcal{P}_2(\mathbb{R}^d)$, the definition of the gradient in $\mathcal{P}_2(\mathbb{R}^d)$ yields

$$\partial_t \mathcal{F}(p_t) = \langle \nabla_{W_2} \mathcal{F}(p_t), v_t \rangle_{p_t} .$$

Note that the orthogonal projection

$$\mathrm{proj}_{T_{p_t}\mathsf{BW}(\mathbb{R}^d)} \nabla_{W_2}\mathcal{F}(p_t) = \operatorname*{arg\,min}_{w\in T_{p_t}\mathsf{BW}(\mathbb{R}^d)} \|w - \nabla_{W_2}\mathcal{F}(p_t)\|_{p_t}^2$$

is characterized as the unique element of $T_{p_t}\mathsf{BW}(\mathbb{R}^d)$ satisfying

$$\langle \mathrm{proj}_{T_{p_t}\mathsf{BW}(\mathbb{R}^d)} \nabla_{W_2}\mathcal{F}(p_t), v\rangle_{p_t} = \langle \nabla_{W_2}\mathcal{F}(p_t), v\rangle_{p_t}$$

for all $v \in T_{p_t}\mathsf{BW}(\mathbb{R}^d)$. Thus, (16.18) holds with

$$\nabla_{\mathsf{BW}}\mathcal{F}(p) = \mathrm{proj}_{T_p\mathsf{BW}(\mathbb{R}^d)} \nabla_{W_2}\mathcal{F}(p) .$$

This argument clearly works for arbitrary Riemannian submanifolds.

Next, we compute the projection of the $\mathcal{P}_2(\mathbb{R}^d)$ gradient of the KL divergence.

Using the formula (16.17) for the $\mathcal{P}_2(\mathbb{R}^d)$ gradient of the KL divergence and the description of the tangent space to $\mathsf{BW}(\mathbb{R}^d)$ in §16.6.1 and (16.13), the projected gradient $(\bar{a}, \bar{S}) \in \mathbb{R}^d \times \mathbf{S}^d$ is such that for all $(a, S) \in \mathbb{R}^d \times \mathbf{S}^d$,

$$\int \left\langle \nabla \ln \frac{p}{\pi}(x), a + S\,(x - m_p) \right\rangle \mathrm{d}p(x) = \langle (\bar{a}, \bar{S}), (a, S)\rangle_p = \langle \bar{a}, a\rangle + \langle \bar{S}, \Sigma_p S\rangle .$$

Using $\nabla p(x) = -\Sigma_p^{-1}\,(x - m_p)\,p(x)$ and integration by parts,

$$\begin{aligned}
&\int \left\langle \nabla \ln \frac{p}{\pi}(x), a + S\,(x - m_p) \right\rangle \mathrm{d}p(x) \\
&\qquad = \left\langle \mathbb{E}_p \nabla \ln \frac{p}{\pi}, a \right\rangle + \int \left\langle \Sigma_p S \nabla \ln \frac{p}{\pi}(x), \Sigma_p^{-1}\,(x - m_p) \right\rangle \mathrm{d}p(x)
\end{aligned}$$

$$
\begin{aligned}
&= \big\langle \mathbb{E}_p \nabla \ln \frac{p}{\pi}, a \big\rangle - \int \big\langle \Sigma_p S \nabla \ln \frac{p}{\pi}(x), \nabla p(x) \big\rangle \, \mathrm{d}x \\
&= \big\langle \mathbb{E}_p \nabla \ln \frac{p}{\pi}, a \big\rangle + \int \operatorname{div}\big(\Sigma_p S \nabla \ln \frac{p}{\pi}\big)(x) \, \mathrm{d}p(x) \\
&= \big\langle \mathbb{E}_p \nabla \ln \frac{p}{\pi}, a \big\rangle + \big\langle \mathbb{E}_p \nabla^2 \ln \frac{p}{\pi}, \Sigma_p S \big\rangle \, .
\end{aligned}
$$

Hence,

$$
(\bar{a}, \bar{S}) = \big(\mathbb{E}_p \nabla \ln \frac{p}{\pi}, \ \mathbb{E}_p \nabla^2 \ln \frac{p}{\pi} \big) \, . \tag{16.19}
$$

Using the fact that $\mathbb{E}_p \nabla \ln p = 0$, this can also be written

$$
(\bar{a}, \bar{S}) = (\mathbb{E}_p \nabla V, \ \mathbb{E}_p \nabla^2 V - \Sigma_p^{-1})
$$

which corresponds to the affine map

$$
x \mapsto \mathbb{E}_p \nabla V + (\mathbb{E}_p \nabla^2 V - \Sigma_p^{-1}) \, (x - m_p) \, . \tag{16.20}
$$

If $(p_t = p_{m_t, \Sigma_t})_{t \geq 0}$ evolves according to the constrained gradient flow, then using the expression for the projected Wasserstein gradient together with (16.14) and (16.15),

$$
\boxed{
\begin{aligned}
\dot{m}_t &= - \, \mathbb{E}_{p_t} \nabla V \, , \\
\dot{\Sigma}_t &= 2I - \Sigma_t \, \mathbb{E}_{p_t} \nabla^2 V - \mathbb{E}_{p_t} \nabla^2 V \, \Sigma_t \, .
\end{aligned}
}
$$

The sign in the above equations comes from the fact that we perform steepest *descent* in Bures–Wasserstein descent, i.e., the tangent vector to the curve at time t is $- \operatorname{proj}_{T_{p_t} \mathsf{BW}(\mathbb{R}^d)} \nabla_{W_2} \mathcal{F}(p_t)$.

The system of equations we have derived here differs from the system (16.4), but we can check that they agree using integration by parts. Indeed,

$$
\begin{aligned}
\dot{\Sigma}_t &= 2I - \Sigma_t \int \nabla^2 V \, \mathrm{d}p_t - \int \nabla^2 V \, \mathrm{d}p_t \, \Sigma_t \\
&= 2I + \Sigma_t \int \nabla p_t \otimes \nabla V + \int \nabla V \otimes \nabla p_t \, \Sigma_t \\
&= 2I + \Sigma_t \int \nabla \ln p_t \otimes \nabla V \, \mathrm{d}p_t + \int \nabla V \otimes \nabla \ln p_t \, \mathrm{d}p_t \, \Sigma_t \\
&= 2I - \mathbb{E}_{p_t}[(\bullet - m_t) \otimes \nabla V + \nabla V \otimes (\bullet - m_t)] \, .
\end{aligned}
$$

16.7.2 Alternate proof using direct Bures–Wasserstein calculation

In the second approach, we view $\mathcal{F}$ as a functional on $\mathsf{BW}(\mathbb{R}^d)$. Explicitly,

$$\mathcal{F}(m,\Sigma) = \int p_{m,\Sigma} \ln \frac{p_{m,\Sigma}}{\pi} \,.$$

Using (16.16),

$$\begin{aligned} \nabla_{\mathsf{BW}} \mathcal{F}(m,\Sigma) &= \big(\nabla_m \mathcal{F}(m,\Sigma),\ 2\,\nabla_\Sigma \mathcal{F}(m,\Sigma)\big) \\ &= \Big(\int \nabla_m p_{m,\Sigma} \ln \frac{p_{m,\Sigma}}{\pi},\ 2\int \nabla_\Sigma p_{m,\Sigma} \ln \frac{p_{m,\Sigma}}{\pi}\Big) \,. \end{aligned} \tag{16.21}$$

Furthermore, using the identities

$$\nabla_m p_{m,\Sigma}(x) = -\nabla_x p_{m,\Sigma}(x) \qquad \text{and} \qquad \nabla_\Sigma p_{m,\Sigma}(x) = \frac{1}{2}\nabla_x^2 p_{m,\Sigma}(x) \tag{16.22}$$

for the Gaussian distribution, integration by parts verifies that (16.21) agrees with (16.19).

16.8 Proof of Corollary 16.3.3

Corollary 16.3.3 is a consequence of general and well-known principles for gradient flows. To emphasize this generality, we will consider an abstract α-convex differentiable functional $\mathcal{F}$ defined over a geodesically convex subset of a Riemannian manifold; this ensures that the logarithmic map is well-defined in the following calculations. We assume that $\mathcal{F}$ is minimized at $p^\star$; by adding a constant to $\mathcal{F}$, we can assume $\inf \mathcal{F} = 0$. Let d denote the distance function on the manifold. If $(p_t)_{t\geq 0}$, $(q_t)_{t\geq 0}$ are two solutions to the gradient flow for $\mathcal{F}$, then

$$\partial_t \mathsf{d}^2(p_t,q_t) = 2\,\langle \log_{p_t}(q_t), \nabla\mathcal{F}(p_t)\rangle_{p_t} + 2\,\langle \log_{q_t}(p_t), \nabla\mathcal{F}(q_t)\rangle_{q_t} \,.$$

(The reader who is unfamiliar with Riemannian geometry should keep in mind that in Euclidean space, $\log_p(q) = q - p$.) Next, the α-convexity of $\mathcal{F}$ implies

$$\begin{aligned} \mathcal{F}(p_t) &\geq \mathcal{F}(q_t) + \langle \nabla\mathcal{F}(q_t), \log_{q_t}(p_t)\rangle_{q_t} + \frac{\alpha}{2}\,\mathsf{d}^2(p_t,q_t) \,, \\ \mathcal{F}(q_t) &\geq \mathcal{F}(p_t) + \langle \nabla\mathcal{F}(p_t), \log_{p_t}(q_t)\rangle_{p_t} + \frac{\alpha}{2}\,\mathsf{d}^2(p_t,q_t) \,. \end{aligned}$$

Adding these equations and rearranging yields

$$\partial_t \mathsf{d}^2(p_t,q_t) \leq -2\alpha\,\mathsf{d}^2(p_t,q_t) \,.$$

By Grönwall's inequality, it implies

$$\mathsf{d}^2(p_t, q_t) \leq \exp(-2\alpha t)\, \mathsf{d}^2(p_0, q_0)\,.$$

This inequality has two consequences. First, for any $\alpha \in \mathbb{R}$, $p_0 = q_0$ implies $p_t = q_t$: the solution to the gradient flow is unique. Second, if $\alpha > 0$, then we can set $q_t = p^\star$ for all $t \geq 0$ to deduce exponential contraction of the gradient flow to the minimizer $p^\star$, which is the first statement of Corollary 16.3.3.

To obtain convergence in functional values, observe that by definition of the gradient flow, we have on the one hand that

$$\partial_t \mathcal{F}(p_t) = -\|\nabla \mathcal{F}(p_t)\|_{p_t}^2\,. \tag{16.23}$$

On the other hand, if $\alpha > 0$, the convexity inequality and Young's inequality respectively, yield

$$\begin{aligned} 0 = \mathcal{F}(p^\star) &\geq \mathcal{F}(p) + \langle \nabla \mathcal{F}(p), \log_p(p^\star)\rangle_p + \frac{\alpha}{2}\,\mathsf{d}^2(p, p^\star) \\ &\geq \mathcal{F}(p) - \frac{1}{2\alpha}\,\|\nabla \mathcal{F}(p)\|_p^2 - \frac{\alpha}{2}\underbrace{\|\log_p(p^\star)\|_p^2}_{=\mathsf{d}^2(p,p^\star)} + \frac{\alpha}{2}\,\mathsf{d}^2(p, p^\star) \end{aligned} \tag{16.24}$$

and hence $\|\nabla \mathcal{F}(p)\|^2 \geq 2\alpha\, \mathcal{F}(p)$. Substituting this into (16.23) and applying Grönwall's inequality again, we deduce

$$\mathcal{F}(p_t) \leq \exp(-2\alpha t)\, \mathcal{F}(p_0)\,.$$

Finally, suppose $\alpha = 0$. We consider the Lyapunov functional

$$\mathcal{L}_t := t\, \mathcal{F}(p_t) + \frac{1}{2}\,\mathsf{d}^2(p_t, p^\star)\,.$$

Differentiating in time,

$$\partial_t \mathcal{L}_t = \mathcal{F}(p_t) - t\,\|\nabla \mathcal{F}(p_t)\|_{p_t}^2 + \langle \log_{p_t}(p^\star), \nabla \mathcal{F}(p_t)\rangle_{p_t}\,.$$

On the other hand, applying the convexity inequality in (16.24) with $\alpha = 0$ yields $\partial_t \mathcal{L}_t \leq 0$. Hence, $\mathcal{L}_t \leq \mathcal{L}_0$, and

$$\mathcal{F}(p_t) \leq \frac{\mathsf{d}^2(p_0, p^\star)}{2t}\,.$$

■ 16.9 Proof of Theorem 16.4.1

In this section, we use the Riemannian exponential and logarithmic maps, as discussed in §16.6.1. Also, let $\mathcal{F} \coloneqq \mathsf{KL}(\cdot \,\|\, \pi)$ denote the KL divergence.

For $\tau > 0$, the eigenvalue clipping operation is defined as

$$\mathrm{clip}^\tau : \qquad \Sigma = \sum_{i=1}^d \lambda_i u_i u_i^\mathsf{T} \quad \mapsto \quad \mathrm{clip}^\tau \Sigma \coloneqq \sum_{i=1}^d (\lambda_i \wedge \tau)\, u_i u_i^\mathsf{T} \,. \tag{16.25}$$

In the proof of Theorem 16.3.1 in §16.7, we showed that the BW gradient is

$$g_p \coloneqq \nabla_{\mathsf{BW}} \mathcal{F}(p) = \big(\mathbb{E}_p \nabla V,\ \mathbb{E}_p \nabla^2 V - \Sigma^{-1}\big) \tag{16.26}$$

where Σ is the covariance matrix of p. Here, the first component of the gradient governs the evolution of the mean, whereas the second component governs the evolution of the covariance; see §16.6.1. We propose to estimate the gradient in (16.26) via a sample,

$$\hat{g}_p \coloneqq \big(\nabla V(\hat{X}),\ \nabla^2 V(\hat{X}) - \Sigma^{-1}\big)\,, \qquad \hat{X} \sim p\,.$$

By comparing Algorithm 16.1 and the definition of the exponential map in §16.6.1, one can check that for $p_k^+ \coloneqq p_{m_{k+1}, \Sigma_k^+}$ and[2] $h \le 1$

$$p_k^+ = \exp_{p_k}(-h \hat{g}_k)\,,$$

where $\hat{g}_k \in T_{p_k}\mathsf{BW}(\mathbb{R}^d)$ is the stochastic gradient

$$\hat{g}_k(x) = \nabla V(\hat{X}_k) + (\nabla^2 V(\hat{X}_k) - \Sigma_k^{-1})\,(x - m_k)\,.$$

Thus, aside from the eigenvalue clipping operation (which is harmless, due to Proposition 15.8.5), Algorithm 16.1 is exactly a stochastic gradient descent scheme on $\mathsf{BW}(\mathbb{R}^d)$. Note also that from the definition of the exponential map in §2.1, the update can also be written at the particle level: if $X_k \sim p_k$ is independent of $\hat{g}_k$, then

$$X_k^+ \coloneqq X_k - h\, \hat{g}_k(X_k) \sim p_k^+\,. \tag{16.27}$$

In the next lemma, we obtain a uniform control on the smallest eigenvalues of the covariance matrices of the iterates.

Lemma 16.9.1. *Assume that $0 \prec \alpha I \preceq \nabla^2 V \preceq I$ holds and $h \le \alpha^2/60$. Also, in Algorithm 16.1, assume that $\Sigma_k \succeq \frac{\alpha}{9} I$. Then, $\Sigma_k^+ \succeq \frac{\alpha}{9} I$.*

[2] This latter requirement is needed because $\mathsf{BW}(\mathbb{R}^d)$ has a finite injectivity radius.

Proof. Since the statement of the lemma only involves the covariance matrices, we can suppose that all of the mean vectors are zero.

The key is to write Σ_k^+ as a generalized Bures–Wasserstein barycenter at Σ_k for an appropriate distribution. Recall that

$$\Sigma_k^+ = \big(I + h\,\Sigma_k^{-1} - h\,\nabla^2 V(\hat{X}_k)\big)\,\Sigma_k\,\big(I + h\,\Sigma_k^{-1} - h\,\nabla^2 V(\hat{X}_k)\big)\,. \tag{16.28}$$

Note that Σ_k^{-1} is the optimal transport map from the Gaussian p_{0,Σ_k} to $p_{0,\Sigma_k^{-1}}$.[3] Hence, we write

$$\begin{aligned} h\,\Sigma_k^{-1} - h\,\nabla^2 V(\hat{X}_k) &= h\,(\Sigma_k^{-1} - I) + h\,(I - \nabla^2 V(\hat{X}_k)) \\ &= h\log_{\Sigma_k}(\Sigma_k^{-1}) + h\log_{\Sigma_k}(\tilde{\Sigma}) \end{aligned}$$

where we defined the matrix $\tilde{\Sigma} = (2I - \nabla^2 V(\hat{X}_k))\,\Sigma_k\,(2I - \nabla^2 V(\hat{X}_k))$. To check that this is valid, we need $2I - \nabla^2 V(\hat{X}_k) \succeq 0$, i.e., $\nabla^2 V(\hat{X}_k) \preceq 2I$, which follows from $\nabla^2 V \preceq I$.

We have shown that

$$\Sigma_k^+ = \exp_{\Sigma_k}\Big(\int \log_{\Sigma_k}(\Sigma)\,\mathrm{d}P(\Sigma)\Big)$$

where

$$P = (1-2h)\,\delta_{\Sigma_k} + h\,\delta_{\Sigma_k^{-1}} + h\,\delta_{\tilde{\Sigma}} = (1-2h)\,\delta_{\Sigma_k} + 2h\,\Big(\frac{1}{2}\,\delta_{\Sigma_k^{-1}} + \frac{1}{2}\,\delta_{\tilde{\Sigma}}\Big)\,.$$

This is precisely the definition of a generalized Bures–Wasserstein barycenter.

Next, suppose that $\Sigma_k \succeq \lambda I$ for some $\lambda > 0$. Since $\Sigma_k \preceq \alpha^{-1} I$, and $I \preceq 2I - \nabla^2 V(\hat{X}_k) \preceq 2I$,

$$\alpha\, I \preceq \Sigma_k^{-1} \preceq \frac{1}{\lambda}\, I\,, \qquad \text{and} \qquad \lambda\, I \preceq \tilde{\Sigma} \preceq \frac{4}{\alpha}\, I\,.$$

Then, Theorem 15.1.1 implies the following. If we define the quantities

$$\lambda_- := \Big(\frac{1}{2}\sqrt{\alpha} + \frac{1}{2}\sqrt{\lambda}\Big)^2\,, \qquad \lambda_+ := \frac{1}{2}\,\frac{1}{\lambda} + \frac{1}{2}\,\frac{4}{\alpha}\,,$$

then for step sizes $2h \le \frac{\lambda_-}{2\lambda_+}$ and if $\Sigma_k \succeq \frac{\lambda_-}{4}\, I$, we also have $\Sigma_k^+ \succeq \frac{\lambda_-}{4}\, I$. To use this result, let us choose λ such that $\frac{\lambda_-}{4} = \lambda$; it can be seen that this holds with $\lambda = \frac{\alpha}{9}$. Since $\lambda_+ = \frac{13}{2\alpha}$, the step size condition then translates into $h \le \frac{2\alpha^2}{117}$, for which it suffices to have $h \le \frac{\alpha^2}{60}$. □

[3] This observation was also used in the analysis of Bures–Wasserstein gradient descent for entropically regularized barycenters in [Alt+21].

We also recall that the eigenvalue clipping operation is a Bures–Wasserstein contraction (see Proposition 15.8.5).

We now turn towards the proof of Theorem 16.4.1. In the proof, we let

$$\mathscr{F}_k \coloneqq \sigma(\hat{X}_0, \hat{X}_1, \hat{X}_2, \dots, \hat{X}_{k-1})$$

be the σ-algebra generated by the random samples up until iteration k.

Proof of Theorem 16.4.1. Conditioned on $\mathscr{F}_k$, and independently of $\hat{X}_k$, let $X_k \sim p_k$ and $Z \sim \hat{\pi}$ be optimally coupled; let $\bar{\mathbb{E}}$ denote the expectation taken w.r.t. (X_k, Z). Using Proposition 15.8.5, the fact that $\hat{\Sigma} \preceq \frac{1}{\alpha} I$ (see discussion in §16.3.3), and (16.27), we have

$$\begin{aligned}
\mathbb{E}[W_2^2(p_{k+1}, \hat{\pi}) \mid \mathscr{F}_k] &\le \mathbb{E}[W_2^2(p_k^+, \hat{\pi}) \mid \mathscr{F}_k] \\
&\le \mathbb{E}\big[\bar{\mathbb{E}}[\|X_k - h\,\hat{g}_k(X_k) - Z\|^2] \,\big|\, \mathscr{F}_k\big] \\
&= \mathbb{E}\big[\bar{\mathbb{E}}[\|X_k - Z\|^2 - 2h\,\langle \hat{g}_k(X_k), X_k - Z\rangle + h^2\,\|\hat{g}_k(X_k)\|^2] \,\big|\, \mathscr{F}_k\big] \\
&= W_2^2(p_k, \hat{\pi}) - 2h\,\bar{\mathbb{E}}\langle g_k(X_k), X_k - Z\rangle + h^2\,\mathbb{E}\big[\bar{\mathbb{E}}[\|\hat{g}_k(X_k)\|^2] \,\big|\, \mathscr{F}_k\big]\,,
\end{aligned}$$

where we abbreviated $g_k \coloneqq g_{p_k}$. From strong convexity of $\mathsf{KL}(\cdot \,\|\, \pi)$ on $\mathsf{BW}(\mathbb{R}^d)$ (Lemma 16.3.2),

$$\begin{aligned}
\bar{\mathbb{E}}\langle g_k(X_k), X_k - Z\rangle &\ge \mathsf{KL}(p_k \,\|\, \pi) - \mathsf{KL}(\hat{\pi} \,\|\, \pi) + \frac{\alpha}{2}\, W_2^2(p_k, \hat{\pi}) \\
&\ge \alpha\, W_2^2(p_k, \hat{\pi})\,.
\end{aligned}$$

Thus,

$$\mathbb{E}[W_2^2(p_{k+1}, \hat{\pi}) \mid \mathscr{F}_k] \le (1 - 2\alpha h)\, W_2^2(p_k, \hat{\pi}) + h^2 \underbrace{\mathbb{E}\big[\bar{\mathbb{E}}[\|\hat{g}_k(X_k)\|^2] \,\big|\, \mathscr{F}_k\big]}_{=:\mathsf{err}}\,.$$

It remains to bound the error term.

Recall that

$$\hat{g}_k(X_k) = (\nabla^2 V(\hat{X}_k) - \Sigma_k^{-1})\,(X_k - m_k) + \nabla V(\hat{X}_k)\,.$$

We bound the terms one by one. First,

$$\bar{\mathbb{E}}[\|\Sigma_k^{-1}\,(X_k - m_k)\|^2] = \mathrm{tr}(\Sigma_k^{-1}) \le \frac{9d}{\alpha}$$

where we used Lemma 16.9.1. Next, since $\nabla^2 V \preceq I$ by assumption,

$$\bar{\mathbb{E}}[\|\nabla^2 V(\hat{X}_k)\,(X_k - m_k)\|^2] \le \bar{\mathbb{E}}[\|X_k - m_k\|^2] = \mathrm{tr}(\Sigma_k) \le \frac{d}{\alpha}\,.$$

Lastly, let $\hat{Z} \sim \hat{\pi}$ be optimally coupled with $\hat{X}_k$. By the optimality condition for $\hat{\pi}$ (§16.3.3), we know that $\mathbb{E}\,\nabla V(\hat{Z}) = 0$. Applying the Poincaré inequality for $\hat{\pi}$ (which holds because $\hat{\pi}$ is strongly log-concave, see Lemma 2.2.8)

$$\begin{aligned}\bar{\mathbb{E}}[\|\nabla V(\hat{X}_k)\|^2] &\le 2\,\bar{\mathbb{E}}[\|\nabla V(\hat{Z})\|^2] + 2\,\bar{\mathbb{E}}[\|\hat{X}_k - \hat{Z}\|^2] \\ &\le \frac{2}{\alpha}\,\mathbb{E}_{\hat{\pi}}[\|\nabla^2 V\|_{\mathrm{HS}}^2] + 2\,W_2^2(p_k, \hat{\pi}) \\ &\le \frac{2d}{\alpha} + 2\,W_2^2(p_k, \hat{\pi})\,.\end{aligned}$$

Collecting the terms,

$$\mathsf{err} \le \frac{36d}{\alpha} + 6\,W_2^2(p_k, \hat{\pi})\,.$$

From the assumption $h \le \frac{\alpha^2}{60}$.

$$\mathbb{E}[W_2^2(p_{k+1}, \hat{\pi}) \mid \mathscr{F}_k] \le (1 - \alpha h)\,W_2^2(p_k, \hat{\pi}) + \frac{36dh^2}{\alpha}\,.$$

Iterating this bound proves the result. □

■ 16.10 Proof of Theorem 16.5.1

In order to present the proof of Theorem 16.5.1, we first review relevant facts about the Wasserstein space over a Riemannian manifold $(\mathcal{M}, \mathfrak{g})$. We refer readers to [Vil09b] for an in-depth treatment.

Similarly to the Euclidean setting, we can define the space of probability measures over $\mathcal{M}$ with finite second moment,

$$\mathcal{P}_2(\mathcal{M}) := \left\{\mu \in \mathcal{P}(\mathcal{M}) \;\middle|\; \int \mathsf{d}^2(p_0, \cdot)\,\mathrm{d}\mu < \infty \text{ for some } p_0 \in \mathcal{M}\right\},$$

where d denotes the induced distance on $\mathcal{M}$. We equip $\mathcal{P}_2(\mathcal{M})$ with the 2-Wasserstein metric

$$W_2^2(\mu, \nu) := \left[\inf_{\gamma \in \mathcal{C}(\mu,\nu)} \int \mathsf{d}^2(x, y)\,\mathrm{d}\gamma(x, y)\right]^{1/2},$$

which makes $(\mathcal{P}_2(\mathcal{M}), W_2)$ into a metric space. Moreover, at each regular measure $\mu \in \mathcal{P}_2(\mathcal{M})$, we can define the tangent space

$$T_\mu \mathcal{P}_2(\mathcal{M}) := \overline{\{\nabla\psi \mid \psi \in \mathcal{C}_c^\infty(\mathcal{M})\}}^{L^2(\mu)}$$

equipped with the inner product

$$\langle v, w \rangle_\mu := \int \mathfrak{g}_p\big(v(p), w(p)\big)\, \mathrm{d}\mu(p)\,,$$

which endows $(\mathcal{P}_2(\mathcal{M}), W_2)$ with the structure of a formal Riemannian manifold. Curves $(\mu_t)_{t\geq 0}$ in $\mathcal{P}_2(\mathcal{M})$ are still described by the continuity equation

$$\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0 \tag{16.29}$$

where now v_t is an element of the tangent bundle $T\mathcal{M}$ and div denotes the divergence operator on the Riemannian manifold. Equation (16.29) is to be interpreted in the weak sense, i.e., for any test function $\varphi : \mathcal{M} \to \mathbb{R}$,

$$\partial_t \int \varphi \, \mathrm{d}\mu_t = \int \mathfrak{g}(\nabla\varphi, v_t)\, \mathrm{d}\mu_t\,. \tag{16.30}$$

If $(\mu_t)_{t\geq 0}$ is a smooth curve such that μ_t admits a density ρ_t w.r.t. the Riemannian volume measure, then this is equivalent to the partial differential equation (PDE)

$$\partial_t \rho_t = \operatorname{div}(\rho_t v_t)\,.$$

As before, the continuity equation admits a particle interpretation: if $p_0 \sim \mu_0$ and $(p_t)_{t\geq 0}$ evolves via the ODE

$$\dot{p}_t = v_t(p_t)\,, \tag{16.31}$$

then $p_t \sim \mu_t$ for all $t \geq 0$.

Given a functional $\mathcal{F} : \mathcal{P}_2(\mathcal{M}) \to \mathbb{R} \cup \{\infty\}$ defined over the Wasserstein space, its gradient at μ is, by definition, the element $\nabla_{W_2}\mathcal{F}(\mu) \in T_\mu\mathcal{P}_2(\mathcal{M})$ such that: for all curves $(\mu_t)_{t\in\mathbb{R}}$ satisfying the continuity equation (16.29) with $\mu_0 = \mu$, it holds that

$$\partial_t\big|_{t=0}\mathcal{F}(\mu_t) = \langle \nabla_{W_2}\mathcal{F}(\mu), v_0\rangle_\mu = \int \mathfrak{g}\big(\nabla_{W_2}\mathcal{F}(\mu), v_0\big)\, \mathrm{d}\mu\,.$$

Using the continuity equation (16.30), it follows by direct identification that

$$\nabla_{W_2}\mathcal{F}(\mu) = \nabla\delta\mathcal{F}(\mu)\,,$$

where $\delta\mathcal{F}(\mu) : \mathcal{M} \to \mathbb{R}$, the first variation of $\mathcal{F}$ at μ, is defined up to an additive constant and satisfies

$$\partial_t\big|_{t=0}\mathcal{F}(\mu) = \int \delta\mathcal{F}(\mu)\, \partial_t\big|_{t=0}\mu_t\,.$$

A gradient flow of $\mathcal{F}$ is a curve $(\mu_t)_{t\geq 0}$ which satisfies the continuity equation (16.29) with velocity vector field $v_t = -\nabla_{W_2}\mathcal{F}(\mu_t)$, which in turn admits the particle interpretation (16.31).

We now consider the functional

$$\mathcal{F}(\mu) \coloneqq \mathsf{KL}(\mathsf{p}_\mu \,\|\, \pi)$$

and compute its first variation. Let $\mathfrak{m}$ denote the Riemannian volume measure; let $(\rho_t)_{t\in\mathbb{R}}$ be a smooth curve of densities $\rho_t = \frac{\mathrm{d}\mu_t}{\mathrm{d}\mathfrak{m}}$. Since

$$\begin{aligned}\mathcal{F}(\mu) &= \int V \,\mathrm{d}\mathsf{p}_\mu + \int \mathsf{p}_\mu \ln \mathsf{p}_\mu \\ &= \iint V \,\mathrm{d}p_\theta\, \rho(\theta)\,\mathrm{d}\mathfrak{m}(\theta) + \iint \ln\Big(\int p_{\theta'}\,\rho(\theta')\,\mathrm{d}\mathfrak{m}(\theta')\Big)\,\mathrm{d}p_\theta\,\rho(\theta)\,\mathrm{d}\mathfrak{m}(\theta)\end{aligned}$$

then

$$\begin{aligned}\partial_t\mathcal{F}(\mu_t) &= \iint V\,\mathrm{d}p_\theta\,\dot\rho_t(\theta)\,\mathrm{d}\mathfrak{m}(\theta) + \iint \frac{\int p_{\theta'}\,\dot\rho_t(\theta')\,\mathrm{d}\mathfrak{m}(\theta')}{\int p_{\theta'}\,\rho_t(\theta')\,\mathrm{d}\mathfrak{m}(\theta')}\,\mathrm{d}p_\theta\,\rho_t(\theta)\,\mathrm{d}\mathfrak{m}(\theta) \\ &\quad + \iint \ln\Big(\int p_{\theta'}\,\rho(\theta')\,\mathrm{d}\mathfrak{m}(\theta')\Big)\,\mathrm{d}p_\theta\,\dot\rho_t(\theta)\,\mathrm{d}\mathfrak{m}(\theta) \\ &= \iint (V + \ln \mathsf{p}_{\mu_t} + 1)\,\mathrm{d}p_\theta\,\dot\rho_t(\theta)\,\mathrm{d}\mathfrak{m}(\theta)\,.\end{aligned}$$

From this,

$$\delta\mathcal{F}(\mu) : \theta \mapsto \int (V + \ln \mathsf{p}_\mu + 1)\,\mathrm{d}p_\theta = \int \ln\frac{\mathsf{p}_\mu}{\pi}\,\mathrm{d}p_\theta + 1\,.$$

Next, we compute the Bures–Wasserstein gradient using (16.16) and (16.22):

$$\begin{aligned}\nabla_{\mathsf{BW}}\delta\mathcal{F}(\mu)(m,\Sigma) &= \Big(\int \ln\frac{\mathsf{p}_\mu}{\pi}\,\nabla_m p_{m,\Sigma},\; 2\int \ln\frac{\mathsf{p}_\mu}{\pi}\,\nabla_\Sigma p_{m,\Sigma}\Big) \\ &= \Big(\int \nabla \ln\frac{\mathsf{p}_\mu}{\pi}\,\mathrm{d}p_{m,\Sigma},\; \int \nabla^2 \ln\frac{\mathsf{p}_\mu}{\pi}\,\mathrm{d}p_{m,\Sigma}\Big)\,.\end{aligned}$$

Finally, to derive the system of ODEs (16.10), we combine the above expression for the Wasserstein gradient of $\mathcal{F}$ together with the particle interpretation (16.31) and the equations (16.14) and (16.15) for dynamics on the BW space.

■ 16.11 Lack of convexity of the KL divergence for mixtures of Gaussians

In this section, we provide counterexamples for the lack of convexity of the objective functional $\mu \mapsto \mathcal{F}(\mu) = \mathsf{KL}(\mathsf{p}_\mu \,\|\, \pi)$ on the space $\mathcal{P}_2(\mathsf{BW}(\mathbb{R}^d))$.

First, we point out that even when π is strongly log-concave, the functional $\mathcal{F}$ can be badly behaved. For example, if $\pi = p_{0,1} = \mathcal{N}(0,1)$ is a Gaussian of variance 1, then we can write it as a Gaussian mixture in many ways: $\pi = \int \mathcal{N}(m,a)\,\mathrm{d}\nu_{1-a}(m)$ for any $a \in [0,1]$, where $\nu_a = \mathcal{N}(0,a)$. In particular, the set of minimizers of $\mathcal{F}$ is not a singleton, and includes all of the measures $\nu_a \otimes \delta_a$ ((m,σ^2) is a random pair with independent components, where $m \sim \mathsf{normal}(0,1-a)$ and $\sigma^2 = a$ almost surely) for $a \in [0,1]$ (as well as all convex combinations—i.e., mixtures—thereof).

Next, we give an explicit example which demonstrates the lack of convexity of the entropy functional $\mu \mapsto \mathcal{H}(\mathsf{p}_\mu) \coloneqq \int \mathsf{p}_\mu \ln \mathsf{p}_\mu$. This can be understood as the KL divergence with zero potential ($V = 0$). Note that the entropy functional $\mathcal{H}$ is convex on $\mathcal{P}_2(\mathbb{R}^d)$ [AGS08, Section 9.4], but our claim is that its composition with the map $\mu \mapsto \mathsf{p}_\mu$ is not convex on $\mathcal{P}_2(\mathsf{BW}(\mathbb{R}^d))$.

In one dimension let $\mu_0 = \mathcal{N}(0,1) \otimes \delta_1$ and $\mu_1 = \mathcal{N}(0,\tau^2) \otimes \delta_1$. In words, a random pair (m_0, σ_0^2) drawn from μ_0 satisfies $m_0 \sim \mathcal{N}(0,1)$ and $\sigma_0^2 = 1$, and similarly for μ_1. What is the optimal coupling of μ_0 and μ_1? Clearly $\sigma_0^2 = \sigma_1^2 = 1$ is the trivial coupling, and since the Bures–Wasserstein distance over the means is the same as the Euclidean distance between the means, we want the usual W_2 optimal coupling between $\mathsf{normal}(0,1)$ and $\mathsf{normal}(0,\tau^2)$; it follows that $m_1 = \tau m_0$. Hence, the Bures geodesic between is $\{(m_t, \sigma_t^2) = ((1-t+t\tau)\,m_0, 1)\}_{t\in[0,1]}$; equivalently the (Bures–)Wasserstein geodesic between μ_0 and μ_1 is

$$\left\{\mu_t = \mathsf{normal}\big(0, (1-t+t\tau)^2\big) \otimes \delta_1\right\}_{t\in[0,1]}.$$

Next, recall that the Gaussian mixture p_{μ_t} is the law of X drawn in the two-stage procedure: first we draw $(m_t, \sigma_t^2) \sim \mu_t$, and given (m_t, σ_t^2) we draw $X \sim p_{m_t,\sigma_t^2}$. Thus,

$$\mathsf{p}_{\mu_t} = \int p_{m,\sigma^2}\,\mathrm{d}\mu_t(m,\sigma^2) = \int p_{m,1}\,\mathrm{d}\nu_{(1-t+t\tau)^2}(m) = p_{0,1+(1-t+t\tau)^2}\,.$$

Hence,

$$\mathcal{H}(\mathsf{p}_{\mu_t}) = \int \mathsf{p}_{\mu_t} \ln \mathsf{p}_{\mu_t} = -\frac{1}{2}\ln(2\pi\mathrm{e}) - \frac{1}{2}\ln\big(1+(1-t+t\tau)^2\big)\,.$$

Then, the convexity of $t \mapsto \mathcal{H}(\mathsf{p}_{\mu_t})$ is equivalent to the convexity of the function $t \mapsto -\ln(1+(1-t+t\tau)^2)$, which fails when, e.g., $\tau = 1/2$; in that case, the function is, in fact, concave on the interval $[0,1]$.

■ 16.12 The Wasserstein–Fisher–Rao gradient flow

Similarly to the setting in §16.5, here we identify probability measures μ over the Bures–Wasserstein space with the corresponding Gaussian mixture p_μ. The aim

of this section is to derive the gradient flow of the KL divergence $\mu \mapsto \mathsf{KL}(\mathsf{p}_\mu \,\|\, \pi)$, except we now equip the space $\mathcal{P}_2(\mathsf{BW}(\mathbb{R}^d))$ with the Wasserstein–Fisher–Rao geometry [LMS16; Chi+18; LMS18]. Deriving the gradient flow with respect to this geometry leads to dynamics for a system of interacting Gaussian particles in which the weight of each particle is also updated at each iteration.

16.12.1 Background on Wasserstein–Fisher–Rao geometry

Here we briefly summarize the relevant background on the Wasserstein–Fisher–Rao (WFR) geometry. The WFR metric is also called the *Hellinger–Kantorovich* metric by some authors.

The Fisher–Rao metric. The Fisher–Rao metric is a metric on the space $\mathcal{M}_+(\mathbb{R}^d)$ of positive measures (not necessarily probability measures). It is the induced metric on $\mathcal{M}_+(\mathbb{R}^d)$ if we enforce that the mapping $\mu \mapsto \sqrt{\mu}$ (defined for smooth probability densities μ) is an isometry into $L^2(\mathbb{R}^d)$. This means that

$$\mathsf{d}^2_{\mathsf{FR}}(\mu_0, \mu_1) = \int (\sqrt{\mu_0} - \sqrt{\mu_1})^2 \,,$$

and if μ_0 and μ_1 are probability measures then this is known to statisticians (up to a constant factor) as the squared Hellinger distance. (If we apply the analogous procedure to discrete probability measures, then this amounts to identifying the simplex with a subset of the unit sphere.) The Fisher–Rao metric is well-studied in the field of information geometry [AN00; Ay+17].

Next, we describe the Riemannian geometry underlying the Fisher–Rao metric. Consider a curve $t \mapsto \mu_t$ of positive measures with time derivative $\dot{\mu}$. Since the Fisher–Rao metric endows the square root of the density with a Hilbert metric, we place endow the time derivative of the square root, $\dot{\sqrt{\mu}} = \dot{\mu}/(2\sqrt{\mu})$, with the Hilbert norm $\|\dot{\mu}/(2\sqrt{\mu})\|_{L^2(\mathbb{R}^d)}$. Thus, the norm at the tangent space $T_\mu \mathcal{M}_+(\mathbb{R}^d)$ is given by

$$\|\dot{\mu}\|^2_\mu = \int \frac{\dot{\mu}^2}{4\mu} \,.$$

Actually, because we are working with positive measures (called *unbalanced* measures to distinguish from the usual optimal transport problem which requires the measures to have the same total mass), this kind of geometry is useful for studying problems in which the total mass changes over time. For example, PDEs of the form $\partial_t \mu_t = \alpha_t \mu_t$ are called reaction equations because they describe, e.g., how the concentration of a chemical changes over time in reaction to the

environment. Motivated by this application, we parameterize $\dot{\mu}$ via $\dot{\mu} = \alpha\mu$, in which case the norm is

$$\|\alpha\|_\mu^2 = \frac{1}{4}\int \alpha^2 \,\mathrm{d}\mu\,. \tag{16.32}$$

Wasserstein geometry. We recall from §2.1 that Wasserstein geometry is motivated by a completely different class of PDEs, namely *transport equations* encoded by the continuity equation

$$\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0\,,$$

which describe the evolving law of a particle x_t tracing out an integral curve of the family of vector fields: $\dot{x}_t = v_t(x_t)$. The Riemannian structure is obtained by equipping the tangent space $T_\mu \mathcal{P}_2(\mathbb{R}^d)$ with the norm

$$\|v\|_\mu^2 = \int \|v\|^2 \,\mathrm{d}\mu\,.$$

Wasserstein–Fisher–Rao geometry. Next we combine the two geometric structures, which can model transport-reaction equations such as

$$\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = \alpha_t \mu_t\,. \tag{16.33}$$

The tangent space norm is then given by the combination combination

$$\|(\alpha, v)\|_\mu^2 = \int (\alpha^2 + \|v\|^2)\,\mathrm{d}\mu\,.$$

(At this point some authors add a factor $\frac{1}{4}$ in front of the α^2, which is natural in view of (16.32). This is convenient for studying geometric properties of the space, but it is not necessary for our purposes.) As in the pure Fisher–Rao case, this is a metric on the space of positive measures $\mathcal{M}_+(\mathbb{R}^d)$.

It induces the distance

$$\mathsf{WFR}^2(\mu_0, \mu_1) \coloneqq \inf\Big\{ \int_0^1 \|(\alpha_t, v_t)\|_{\mu_t}^2 \,\mathrm{d}t \;\Big|\; (\mu_t, \alpha_t, v_t)_{t\in[0,1]} \text{ solves (16.33)} \Big\}\,.$$

One can show that the tangent space to $\mathcal{M}_+(\mathbb{R}^d)$ consists of pairs (α, v) for which $\alpha = u$ and $v = \nabla u$ for some function $u : \mathbb{R}^d \to \mathbb{R}$. Thus, compared to the Wasserstein metric in which the tangent space norm is the $\dot{H}^1(\mu)$ norm $\|u\|_{\dot{H}^1(u)} = \|\nabla u\|_{L^2(\mu)}$, the Wasserstein–Fisher–Rao metric has the interpretation of completing the tangent space norm to the full Sobolev norm $H^1(\mu)$.

Constraining the dynamics to lie within probability measures. In order to have our dynamics stay on the space of probability measures, we follow [LLN19] and consider instead the equation

$$\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = \Big(\alpha_t - \int \alpha_t \, \mathrm{d}\mu_t\Big)\, \mu_t \,,$$

which now conserves mass. The tangent space norm is modified to read

$$\|(\alpha, v)\|_\mu^2 = \int \Big[\Big(\alpha - \int \alpha \, \mathrm{d}\mu\Big)^2 + \|v\|^2\Big] \, \mathrm{d}\mu \,.$$

Particle interpretation. The particle interpretation of the WFR geometry is more complicated to state than for the Wasserstein geometry, but it can be done. Instead of considering a particle x, we consider a pair (x, r) consisting of a particle $x \in \mathbb{R}^d$ and a number $r > 0$ (this number is actually interpreted as the *square root* of the mass of the particle). The pair (x, r) should be thought of as an element of the cone space $\mathfrak{C}(\mathbb{R}^d) \coloneqq (\mathbb{R}^d \times \mathbb{R}_+)/(\mathbb{R}^d \times \{0\})$ (in other words, we take the space $\mathbb{R}^d \times \mathbb{R}_+$ and identify all of the points with zero mass which sit at the "tip of the cone"). The cone space is the natural setting for WFR geometry; for example, one can introduce a metric on $\mathfrak{C}(\mathbb{R}^d)$ and show that the WFR distance is an optimal transport problem w.r.t. this metric. We will not go into such detail, but nevertheless we introduce the cone space because is important for the particle interpretation of WFR dynamics.

Curves of measures $(\mu_t)_{t\in[0,1]}$ in the WFR geometry admit a particle interpretation in terms of trajectories on $\mathfrak{C}(\mathbb{R}^d)$. Namely, the equation (16.33) can be interpreted as follows. There exists a curve of measures $t \mapsto \widetilde{\mu}_t$ over the cone space $\mathfrak{C}(\mathbb{R}^d)$, such that if $r : \mathfrak{C}(\mathbb{R}^d) \to \mathbb{R}_+$ denotes the mapping $(x, r) \mapsto r$, and $x : \mathfrak{C}(\mathbb{R}^d) \to \mathbb{R}^d$ maps $(x, r) \mapsto x$, then

$$\mu_t = x_\#(r^2 \widetilde{\mu}_t) \,.$$

Moreover, if we draw $(x_0, r_0) \sim \widetilde{\mu}_0$ and follow the ODEs

$$\begin{aligned} \dot{x}_t &= v_t(x_t) \,, \\ \dot{r}_t &= \Big(\alpha_t(x_t) - \int \alpha_t \, \mathrm{d}\mu_t\Big)\, r_t \,, \end{aligned}$$

then $(x_t, r_t) \sim \widetilde{\mu}_t$. Here the notation $\sim$ is an (egregious) abuse of notation because $\widetilde{\mu}_t$ is not a probability measure; by $(x, r) \sim \widetilde{\mu}$ more precisely we mean that $\widetilde{\mu}_t = (\mathsf{ODE}_t)_\# \widetilde{\mu}_0$ where ODE_t is the solution mapping $(x_0, r_0) \mapsto (x_t, r_t)$ to the above system of ODEs at time t.

To make this interpretation more concrete, we specialize to the case of discrete measures. Suppose that we start at a probability measure

$$\mu_0 = \sum_{i=1}^{N} w_0^{(i)} \delta_{x_0^{(i)}} .$$

Then, we lift to the cone space:

$$\widetilde{\mu}_0 = \sum_{i=1}^{N} \delta_{(x_0^{(i)}, \sqrt{w_0^{(i)}})} = \sum_{i=1}^{N} \delta_{(x_0^{(i)}, r_0^{(i)})}$$

where we set $r_t^{(i)} = \sqrt{w_t^{(i)}}$. Next, we follow the ODEs

$$\begin{aligned} \dot{x}_t^{(i)} &= v_t(x_t^{(i)}) , \\ \dot{r}_t^{(i)} &= \Big(\alpha_t(x_t^{(i)}) - \sum_{j=1}^{N} w_t^{(j)} \alpha_t(x_t^{(j)}) \Big) \, r_t^{(i)} . \end{aligned}$$

Upon projecting back to the base space, we obtain another discrete measure

$$\mu_t = \sum_{i=1}^{N} w_t^{(i)} \delta_{x_t^{(i)}} = \sum_{i=1}^{N} (r_t^{(i)})^2 \, \delta_{x_t^{(i)}} .$$

As a sanity check, we check that these dynamics ensure that μ_t is a probability measure for all t. The time derivative of the sum of the weights is

$$\begin{aligned} \partial_t \sum_{i=1}^{N} w_t^{(i)} &= 2 \sum_{i=1}^{N} r_t^{(i)} \, \partial_t r_t^{(i)} = 2 \sum_{i=1}^{N} (r_t^{(i)})^2 \left(\alpha_t(x_i^{(t)}) - \mathbb{E}_{\mu_t} \alpha_t \right) \\ &= 2 \Big(\sum_{i=1}^{N} w_t^{(i)} \alpha_t(x_i^{(t)}) - \mathbb{E}_{\mu_t} \alpha_t \Big) = 0 . \end{aligned}$$

■ 16.12.2 Derivation of the gradient flow

Next, we derive the Wasserstein–Fisher–Rao gradient flow of the functional $\mu \mapsto \mathcal{F}(\mu) \coloneqq \mathsf{KL}(\mathsf{p}_\mu \,\|\, \pi)$ on the space $(\mathcal{P}_2(\mathsf{BW}(\mathbb{R}^d)), \mathsf{WFR})$ of Gaussian mixtures equipped with the Wasserstein–Fisher–Rao metric (over the Bures–Wasserstein space). The WFR gradient of $\mathcal{F}$, $\nabla_{\mathsf{WFR}} \mathcal{F}(\mu)$, is the pair

$$\nabla_{\mathsf{WFR}} \mathcal{F}(\mu) = \Big(\nabla_{\mathsf{BW}} \delta\mathcal{F}(\mu), \; \delta\mathcal{F}(\mu) - \int \delta\mathcal{F}(\mu) \, \mathrm{d}\mu \Big) .$$

This result is essentially stated as [LLN19, Proposition A.1], although we have generalized the formula to hold when the base space is no longer $\mathbb{R}^d$. Note also that we have already calculated the first variation of $\mathcal{F}$, as well as the BW gradient, in §16.10.

The interpretation of the formula is that in the gradient flow of $\mathcal{F}$, we have a particle (m, Σ) associated with some mass w evolving according to

$$\begin{aligned}
\dot{m} &= -\mathbb{E}_{p_{m,\Sigma}} \nabla \ln \frac{\mathsf{p}_\mu}{\pi}, \\
\dot{\Sigma} &= -\Sigma\, \mathbb{E}_{p_{m,\Sigma}} \nabla^2 \ln \frac{\mathsf{p}_\mu}{\pi} - \mathbb{E}_{p_{m,\Sigma}} \nabla^2 \ln \frac{\mathsf{p}_\mu}{\pi}\, \Sigma, \\
\dot{r} &= -\left(\mathbb{E}_{p_{m,\Sigma}} \ln \frac{\mathsf{p}_\mu}{\pi} - \mathbb{E}_{\mathsf{p}_\mu} \ln \frac{\mathsf{p}_\mu}{\pi} \right) r,
\end{aligned}$$

where $r = \sqrt{w}$. The interpretation may be clearer in the discrete case, so suppose that we initialize the dynamics at a discrete measure

$$\mu_0 = \sum_{i=1}^N w_0^{(i)} \delta_{(m_0^{(i)}, \Sigma_0^{(i)})}.$$

Next we solve the coupled system of ODEs, for $i \in [N]$,

$$\begin{aligned}
\dot{m}^{(i)} &= -\mathbb{E}_{p_{m^{(i)},\Sigma^{(i)}}} \nabla \ln \frac{\mathsf{p}_\mu}{\pi}, \\
\dot{\Sigma}^{(i)} &= -\Sigma^{(i)}\, \mathbb{E}_{p_{m^{(i)},\Sigma^{(i)}}} \nabla^2 \ln \frac{\mathsf{p}_\mu}{\pi} - \mathbb{E}_{p_{m^{(i)},\Sigma^{(i)}}} \nabla^2 \ln \frac{\mathsf{p}_\mu}{\pi}\, \Sigma^{(i)}, \\
\dot{r}^{(i)} &= -\left(\mathbb{E}_{p_{m^{(i)},\Sigma^{(i)}}} \ln \frac{\mathsf{p}_\mu}{\pi} - \mathbb{E}_{\mathsf{p}_\mu} \ln \frac{\mathsf{p}_\mu}{\pi} \right) r^{(i)},
\end{aligned}$$

where $r^{(i)} = \sqrt{w^{(i)}}$ and

$$\mu_t = \sum_{i=1}^N w_t^{(i)} \delta_{(m_t^{(i)}, \Sigma_t^{(i)})}.$$

Since the normalization constant of π cancels out in the above equations, they are implementable without this knowledge.

16.13 Conclusion

Using the powerful theory of Wasserstein gradient flows, we derived new algorithms for VI using either Gaussians or mixtures of Gaussians as approximating distributions. The consequences are twofold. On the one hand, strong convergence

guarantees under classical conditions contribute markedly to closing the theoretical gap between MCMC and Gaussian VI. On the other hand, discretization of the Wasserstein gradient flow for mixtures of Gaussians yields a new *Gaussian particle method* for time discretization which, unlike classical particle methods, maintains a continuous probability distribution at each time.

We conclude by briefly listing some possible directions for future study. For Gaussian variational inference, our theoretical result (Theorem 16.4.1) can be strengthened by weakening the assumption that π is strongly log-concave, or by developing algorithms which do not require Hessian information for V.

Towards the first question, we remark that (similarly to the analysis in §9), the KL divergence objective functional can be viewed as composite optimization consisting of the sum of a *potential energy* term that is convex and smooth (provided V is) and a convex but non-smooth *entropy* term. This observation has motivated the study of proximal gradient methods over the Wasserstein space, which are generally not implementable [SKL20]. However, the proximal operator for the entropy turns out to be implementable in closed form over the Bures–Wasserstein space, motivating the development of proximal gradient methods for Gaussian VI; we pursue this further in the work [Dia+23].

Code for the experiments is available at `https://github.com/marc-h-lambert/W-VI`.

Chapter 17

Theory for diffusion models

We provide theoretical convergence guarantees for score-based generative models (SGMs) such as denoising diffusion probabilistic models (DDPMs), which constitute the backbone of large-scale real-world generative models such as DALL·E 2. Our main result is that, assuming accurate score estimates, such SGMs can efficiently sample from essentially any realistic data distribution. In contrast to prior works, our results (1) hold for an L^2-accurate score estimate (rather than L^∞-accurate); (2) do not require restrictive functional inequality conditions that preclude substantial non-log-concavity; (3) scale polynomially in all relevant problem parameters; and (4) match state-of-the-art complexity guarantees for discretization of the Langevin diffusion, provided that the score error is sufficiently small. We view this as strong theoretical justification for the empirical success of SGMs. We also examine SGMs based on the critically damped Langevin diffusion (CLD). Contrary to conventional wisdom, we provide evidence that the use of the CLD does *not* reduce the complexity of SGMs.

This chapter is based on [Che+23a], joint with Sitan Chen, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang.

17.1 Introduction

Score-based generative models (SGMs) are a family of generative models which achieve state-of-the-art performance for generating audio and image data [Soh+15; HJA20; DN21; Kin+21; Son+21a; Son+21b; VKK21]. One notable example of an SGM are denoising diffusion probabilistic models (DDPMs) [Soh+15; HJA20], which are a key component in large-scale generative models such as DALL·E 2 [Ram+22]. As the importance of SGMs continues to grow due to newfound applications in commercial domains, it is a pressing question of both practical and theoretical concern to understand the mathematical underpinnings which explain their startling empirical successes.

As we explain in more detail in §17.2, at their mathematical core, SGMs consist

of two stochastic processes, which we call the forward process and the reverse process. The forward process transforms samples from a data distribution q (e.g., natural images) into pure noise, whereas the reverse process transforms pure noise into samples from q, hence performing generative modelling. Implementation of the reverse process requires estimation of the *score function* of the law of the forward process, which is typically accomplished by training neural networks on a score matching objective [Hyv05; Vin11; SE19].

Providing precise guarantees for estimation of the score function is difficult, as it requires an understanding of the non-convex training dynamics of neural network optimization that is currently out of reach. However, given the empirical success of neural networks on the score estimation task, a natural and important question is whether or not accurate score estimation implies that SGMs provably converge to the true data distribution in realistic settings. This is a surprisingly delicate question, as even with accurate score estimates, as we explain in §17.2.1, there are several other sources of error which could cause the SGM to fail to converge. Indeed, despite a flurry of recent work on this question [De +21; BMR22; De 22; Liu+22; LLT22; Pid22], prior analyses fall short of answering this question, for (at least) one of three main reasons:

1. **Super-polynomial convergence.** The bounds obtained are not quantitative (e.g., [De +21; Liu+22; Pid22]), or scale exponentially in the dimension and other problem parameters [BMR22; De 22], and hence are typically vacuous for the high-dimensional settings of interest in practice.
2. **Strong assumptions on the data distribution.** The bounds require strong assumptions on the true data distribution, such as a log-Sobelev inequality (LSI) (see, e.g., [LLT22]). While the LSI is slightly weaker than log-concavity, it ultimately precludes the presence of substantial non-convexity, which impedes the application of these results to complex and highly multi-modal real-world data distributions. Indeed, obtaining a polynomial-time convergence analysis for SGMs that holds for multi-modal distributions was posed as an open question in [LLT22].
3. **Strong assumptions on the score estimation error.** The bounds require that the score estimate is L^∞-accurate (i.e., *uniformly* accurate), as opposed to L^2-accurate (see, e.g., [De +21]). This is particularly problematic because the score matching objective is an L^2 loss (see §17.2 for details), and there are empirical studies suggesting that in practice, the score estimate is not in fact L^∞-accurate (e.g., [ZC23]). Intuitively, this is because we cannot expect that the score estimate we obtain in practice will be accurate in regions of space where the true density is very low, simply because we do not expect to see many (or indeed, any) samples from such regions.

Providing an analysis which goes beyond these limitations is a pressing first step towards theoretically understanding why SGMs actually work in practice.

Concurrent work. The concurrent and independent work of [LLT23] also obtains similar guarantees to our Corollary 17.3.5.

17.1.1 Our contributions

In this work, we take a step towards bridging theory and practice by providing a convergence guarantee for SGMs, under realistic (in fact, quite minimal) assumptions, which scales polynomially in all relevant problem parameters. Namely, our main result (Theorem 17.3.4) only requires the following assumptions on the data distribution q, which we make more quantitative in §17.3:

A1 The score function of the forward process is L-Lipschitz.

A2 The second moment of q is finite.

A3 The data distribution q has finite KL divergence w.r.t. the standard Gaussian.

We note that all of these assumptions are either standard or, in the case of **A2**, far weaker than what is needed in prior work. Crucially, unlike prior works, we do *not* assume log-concavity, an LSI, or dissipativity; hence, our assumptions cover *arbitrarily non-log-concave* data distributions. Our main result is summarized informally as follows.

Theorem 17.1.1 (Informal, see Theorem 17.3.4)**.** *Under assumptions* **A1-A3***, and if the score estimation error in L^2 is at most $\widetilde{O}(\varepsilon)$, then with an appropriate choice of step size, the SGM outputs a measure which is ε-close in total variation (TV) distance to q in $\widetilde{O}(L^2 d/\varepsilon^2)$ iterations.*

We remark that our iteration complexity is actually quite tight: in fact, this matches state-of-the-art discretization guarantees for the Langevin diffusion, see [VW19] and §3.

We find Theorem 17.1.1 to be quite surprising, because it shows that SGMs can sample from the data distribution q with polynomial complexity, even when q is highly non-log-concave (a task that is usually intractable), *provided that one has access to an accurate score estimator*. This answers the open question of [LLT22] regarding whether or not SGMs can sample from multimodal distributions, e.g., mixtures of distributions with bounded log-Sobolev constant. In the context of neural networks, our result implies that so long as the neural network succeeds at the learning task, the remaining part of the SGM algorithm based on the diffusion model is principled, in that it admits a strong theoretical justification.

In general, learning the score function is also a difficult task. Nevertheless, our result opens the door to further investigations, such as: do score functions for real-life data have intrinsic (e.g., low-dimensional) structure which can be exploited by neural networks? A positive answer to this question, combined with our sampling result, would then provide an end-to-end guarantee for SGMs.

More generally, our result can be viewed as a black-box reduction of the task of sampling to the task of learning the score function of the forward process, at least for distributions satisfying our mild assumptions. As a simple consequence, existing computational hardness results for learning natural high-dimensional distributions like mixtures of Gaussians [DKS17; Bru+21; GVV22] and pushforwards of Gaussians by shallow ReLU networks [DV21; Che+22a; CLL22] immediately imply hardness of score estimation for these distributions. To our knowledge this yields the first known information-computation gaps for this task.

Arbitrary distributions with bounded support. The assumption that the score function is Lipschitz entails in particular that the data distribution has a density w.r.t. Lebesgue measure; in particular, our theorem fails when q satisfies the manifold hypothesis, i.e., is supported on a lower-dimensional submanifold of $\mathbb{R}^d$. But this is for good reason: it is not possible to obtain non-trivial TV guarantees, because the output distribution of the SGM has full support. Instead, we show in §17.3.2 that we can obtain polynomial convergence guarantees in the Wasserstein metric by stopping the SGM algorithm early, under the *sole* assumption that that data distribution q has bounded support. Since any data distribution encountered in real life satisfies this condition, our results yield the following compelling takeaway:

Given an L^2-accurate score estimate, SGMs can sample from (essentially) any data distribution.

This constitutes a powerful theoretical justification for the use of SGMs in practice.

Critically damped Langevin diffusion (CLD). Using our techniques, we also investigate the use of the critically damped Langevin diffusion (CLD) for SGMs, which was proposed in [DVK22]. Although numerical experiments and intuition from the log-concave sampling literature suggest that the CLD could potentially speed up sampling via SGMs, we provide theoretical evidence to the contrary: in §17.3.3, we conjecture that SGMs based on the CLD do not exhibit improved dimension dependence compared to the original DDPM algorithm.

■ 17.1.2 Prior work

We now provide a more detailed comparison to prior work, in addition to the previous discussion above.

By now, there is a vast literature on providing precise complexity estimates for log-concave sampling; see, e.g., the book draft [Che23] for an exposition to recent developments. The proofs in this work build upon the techniques developed in this literature. However, our work addresses the significantly more challenging setting of *non-log-concave* sampling.

The paper [De +21] provides sampling guarantees for the diffusion Schrödinger bridge [Son+21b]. However, as previously mentioned their result is not quantitative, and they require an L^∞-accurate score estimate. The works [BMR22; LLT22] instead analyze SGMs under the more realistic assumption of an L^2-accurate score estimate. However, the bounds of [BMR22] suffer from the curse of dimensionality, whereas the bounds of [LLT22] require q to satisfy an LSI.

The recent work of [De 22], motivated by the *manifold hypothesis*, considers a different pointwise assumption on the score estimation error which allows the error to blow up at time 0 and at spatial ∞. We discuss the manifold setting in more detail in §17.3.2. Unfortunately, the bounds of [De 22] also scale exponentially in problem parameters such as the manifold diameter.

After the first version of this work appeared online, we became aware of two concurrent and independent works [Liu+22; LLT23] which share similarities with our work. Namely, [Liu+22] uses a similar proof technique as our Theorem 17.3.4 (albeit without explicit quantitative bounds), whereas [LLT23] obtains similar guarantees to our Corollary 17.3.5 below. The follow-up work of [CLL23] further improves upon the results in this chapter.

We also mention that the use of reversed SDEs for sampling is also implicit in the interpretation of the proximal sampler algorithm [LST21c] given in §4, and the present work can be viewed as expanding upon the theory of §4 using a different forward channel (the OU process).

17.2 Background on SGMs

Throughout this chapter, given a probability measure p which admits a density w.r.t. Lebesgue measure, we abuse notation and identify it with its density function. Additionally, we will let q denote the data distribution from which we want to generate new samples. We assume that q is a probability measure on $\mathbb{R}^d$ with full support, and that it admits a smooth density. (See, however, §17.3.2 on applications of our results to the case when q does not admit a density, such as the case when q is supported on a lower-dimensional submanifold of $\mathbb{R}^d$.) In this case, we can write the density of q in the form $q = \exp(-U)$, where $U : \mathbb{R}^d \to \mathbb{R}$ is the *potential*.

In this section, we provide a brief exposition to SGMs, following [Son+21b].

17.2.1 Background on denoising diffusion probabilistic models (DDPM)

Forward process. In denoising diffusion probabilistic modelling (DDPM), we start with a forward process, which is a stochastic differential equation (SDE). For clarity, we consider the simplest possible choice, which is the Ornstein–Uhlenbeck (OU) process

$$\mathrm{d}\bar{X}_t = -\bar{X}_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \qquad \bar{X}_0 \sim q\,, \tag{17.1}$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$. The OU process is the unique time-homogeneous Markov process which is also a Gaussian process, with stationary distribution equal to the standard Gaussian distribution γ^d on $\mathbb{R}^d$. In practice, it is also common to introduce a positive smooth function $g : \mathbb{R}_+ \to \mathbb{R}$ and consider the time-rescaled OU process

$$\mathrm{d}\bar{X}_t = -g(t)^2\,\bar{X}_t\,\mathrm{d}t + \sqrt{2}\,g(t)\,\mathrm{d}B_t\,, \qquad X_0 \sim q\,, \tag{17.2}$$

but in this work we stick with the choice $g \equiv 1$.

The forward process has the interpretation of transforming samples from the data distribution q into pure noise. From the well-developed theory of Markov diffusions, it is known that if $q_t := \mathrm{law}(X_t)$ denotes the law of the OU process at time t, then $q_t \to \gamma^d$ exponentially fast in various divergences and metrics such as the 2-Wasserstein metric W_2; see [BGL14].

Reverse process. If we reverse the forward process (17.1) in time, then we obtain a process that transforms noise into samples from q, which is the aim of generative modelling. In general, suppose that we have an SDE of the form

$$\mathrm{d}\bar{X}_t = b_t(\bar{X}_t)\,\mathrm{d}t + \sigma_t\,\mathrm{d}B_t\,,$$

where $(\sigma_t)_{t\geq 0}$ is a deterministic matrix-valued process. Then, under mild conditions on the process (e.g., [Föl85; Cat+22]), which are satisfied for all processes under consideration in this work, the reverse process also admits an SDE description. Namely, if we fix the terminal time $T > 0$ and set

$$\bar{X}^{\leftarrow}_t := \bar{X}_{T-t}\,, \qquad \text{for } t \in [0,T]\,, \tag{17.3}$$

then the process $(\bar{X}^{\leftarrow}_t)_{t\in[0,T]}$ satisfies the SDE

$$\mathrm{d}\bar{X}^{\leftarrow}_t = b^{\leftarrow}_t(\bar{X}^{\leftarrow}_t)\,\mathrm{d}t + \sigma_{T-t}\,\mathrm{d}B_t\,,$$

where the backwards drift satisfies the relation

$$b_t + b^{\leftarrow}_{T-t} = \sigma_t\sigma_t^{\mathsf{T}}\nabla \ln q_t\,, \qquad q_t := \mathrm{law}(\bar{X}_t)\,. \tag{17.4}$$

Applying this to the forward process (17.1), we obtain the reverse process

$$\mathrm{d}\bar{X}_t^{\leftarrow} = \{\bar{X}_t^{\leftarrow} + 2\,\nabla \ln q_{T-t}(\bar{X}_t^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \qquad \bar{X}_0^{\leftarrow} \sim q_T\,, \tag{17.5}$$

where now $(B_t)_{t\in[0,T]}$ is the reversed Brownian motion.[1] Here, $\nabla \ln q_t$ is called the *score function* for q_t. Since q (and hence q_t for $t \geq 0$) is not explicitly known, in order to implement the reverse process the score function must be estimated on the basis of samples.

Score matching. In order to estimate the score function $\nabla \ln q_t$, consider minimizing the $L^2(q_t)$ loss over a function class $\mathscr{F}$,

$$\underset{s_t\in\mathscr{F}}{\text{minimize}} \quad \mathbb{E}_{q_t}[\|s_t - \nabla \ln q_t\|^2]\,, \tag{17.6}$$

where $\mathscr{F}$ could be, e.g., a class of neural networks. The idea of score matching, which goes back to [Hyv05; Vin11], is that after applying integration by parts for the Gaussian measure, the problem (17.6) is *equivalent* to the following problem:

$$\underset{s_t\in\mathscr{F}}{\text{minimize}} \quad \mathbb{E}\Big[\Big\|s_t(\bar{X}_t) + \frac{1}{\sqrt{1-\exp(-2t)}}\,Z_t\Big\|^2\Big]\,, \tag{17.7}$$

where $Z_t \sim \mathsf{normal}(0, I_d)$ is independent of $\bar{X}_0$ and we set $\bar{X}_t = \exp(-t)\,\bar{X}_0 + \sqrt{1-\exp(-2t)}\,Z_t$, in the sense that (17.6) and (17.7) share the same minimizers. We give a self-contained derivation in §17.2.3 for the sake of completeness. Unlike (17.6), however, the objective in (17.7) can be replaced with an empirical version and estimated on the basis of samples $\bar{X}_0^{(1)}, \ldots, \bar{X}_0^{(n)}$ from q, leading to the finite-sample problem

$$\underset{s_t\in\mathscr{F}}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n}\Big\|s_t(\bar{X}_t^{(i)}) + \frac{1}{\sqrt{1-\exp(-2t)}}\,Z_t^{(i)}\Big\|^2\,, \tag{17.8}$$

where $(Z_t^{(i)})_{i\in[n]}$ are i.i.d. standard Gaussians independent of the data $(\bar{X}_0^{(i)})_{i\in[n]}$. Moreover, if we parametrize the score function as $s_t = -\frac{1}{\sqrt{1-\exp(-2t)}}\,\widehat{z}_t$, then the empirical problem is equivalent to

$$\underset{\widehat{z}_t\in-\sqrt{1-\exp(-2t)}\,\mathscr{F}}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n}\big\|\widehat{z}_t(\bar{X}_t^{(i)}) - Z_t^{(i)}\big\|^2\,,$$

[1] For ease of notation, we do not distinguish between the forward and the reverse Brownian motions.

which has the illuminating interpretation of predicting the added noise $Z_t^{(i)}$ from the noised data $\bar{X}_t^{(i)}$.

We remark that given the objective function (17.6), it is most natural to assume an $L^2(q_t)$ error bound $\mathbb{E}_{q_t}[\|s_t - \nabla \ln q_t\|^2] \le \varepsilon_{\text{score}}^2$ for the score estimator. If s_t is taken to be the empirical risk minimizer for an appropriate function class, then guarantees for the $L^2(q_t)$ error can be obtained via standard statistical analysis, as was done in [BMR22].

Discretization and implementation. We now discuss the final steps required to obtain an implementable algorithm. First, in the learning phase, given samples $\bar{X}_0^{(1)}, \dots, \bar{X}_0^{(n)}$ from q (e.g., a database of natural images), we train a neural network on the empirical score matching objective (17.8), see [SE19]. Let $h > 0$ be the step size of the discretization; we assume that we have obtained a score estimate s_{kh} of $\nabla \ln q_{kh}$ for each time $k = 0, 1, \dots, N$, where $T = Nh$.

In order to approximately implement the reverse SDE (17.5), we first replace the score function $\nabla \ln q_{T-t}$ with the estimate s_{T-t}. Then, for $t \in [kh, (k+1)h]$ we freeze this coefficient in the SDE at time kh. It yields the new SDE

$$\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + 2\, s_{T-kh}(X_{kh}^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \qquad t \in [kh, (k+1)h]\,. \tag{17.9}$$

Since this is a linear SDE, it can be integrated in closed form; in particular, conditionally on $X_{kh}^{\leftarrow}$, the next iterate $X_{(k+1)h}^{\leftarrow}$ has an explicit Gaussian distribution.

There is one final detail: although the reverse SDE (17.5) should be started at q_T, we do not have access to q_T directly. Instead, taking advantage of the fact that $q_T \approx \gamma^d$, we instead initialize the algorithm at $X_0^{\leftarrow} \sim \gamma^d$, i.e., from pure noise.

Let $p_t := \mathrm{law}(X_t^{\leftarrow})$ denote the law of the algorithm at time t. The goal of this work is to bound $\mathsf{TV}(p_T, q)$, taking into account three sources of error: (1) the estimation of the score function; (2) the discretization of the SDE with step size $h > 0$; and (3) the initialization of the algorithm at γ^d rather than at q_T.

■ 17.2.2 Background on the critically damped Langevin diffusion (CLD)

The critically damped Langevin diffusion (CLD) is based on the forward process

$$\begin{aligned} \mathrm{d}\bar{X}_t &= -\bar{V}_t\,\mathrm{d}t\,, \\ \mathrm{d}\bar{V}_t &= -(\bar{X}_t + 2\,\bar{V}_t)\,\mathrm{d}t + 2\,\mathrm{d}B_t\,. \end{aligned} \tag{17.10}$$

Compared to the OU process (17.1), this is now a coupled system of SDEs, where we have introduced a new variable $\bar{V}$ representing the velocity process. The stationary distribution of the process is $\boldsymbol{\gamma}^{2d}$, the standard Gaussian measure on phase space $\mathbb{R}^d \times \mathbb{R}^d$, and we initialize at $\bar{X}_0 \sim q$ and $\bar{V}_0 \sim \gamma^d$.

More generally, the CLD (17.10) is an instance of what is referred to as the *kinetic Langevin* or the *underdamped Langevin* process in the sampling literature. In the context of log-concave sampling, the smoother paths of $\bar{X}$ leads to smaller discretization error, thereby furnishing an algorithm with $\widetilde{O}(\sqrt{d}/\varepsilon)$ gradient complexity (as opposed to sampling based on the overdamped Langevin process, which has complexity $\widetilde{O}(d/\varepsilon^2)$), see [Che+18b; SL19; DR20; Ma+21] and §6. In the recent paper [DVK22], Dockhorn, Vahdat, and Kreis proposed to use the CLD as the basis for an SGM and they empirically observed improvements over DDPM.

Applying (17.4), the corresponding reverse process is

$$\begin{aligned} \mathrm{d}\bar{X}_t^{\leftarrow} &= -\bar{V}_t^{\leftarrow}\,\mathrm{d}t\,, \\ \mathrm{d}\bar{V}_t^{\leftarrow} &= \big(\bar{X}_t^{\leftarrow} + 2\,\bar{V}_t^{\leftarrow} + 4\,\nabla_v \ln \boldsymbol{q}_{T-t}(\bar{X}_t^{\leftarrow}, \bar{V}_t^{\leftarrow})\big)\,\mathrm{d}t + 2\,\mathrm{d}B_t\,, \end{aligned} \tag{17.11}$$

where $\boldsymbol{q}_t \coloneqq \mathrm{law}(\bar{X}_t, \bar{V}_t)$ is the law of the forward process at time t. Note that the gradient in the score function is only taken w.r.t. the velocity coordinate. Upon replacing the score function with an estimate $\boldsymbol{s}$, we arrive at the algorithm

$$\begin{aligned} \mathrm{d}X_t^{\leftarrow} &= -V_t^{\leftarrow}\,\mathrm{d}t\,, \\ \mathrm{d}V_t^{\leftarrow} &= \big(X_t^{\leftarrow} + 2\,V_t^{\leftarrow} + 4\,\boldsymbol{s}_{T-kh}(X_{kh}^{\leftarrow}, V_{kh}^{\leftarrow})\big)\,\mathrm{d}t + 2\,\mathrm{d}B_t\,, \end{aligned}$$

for $t \in [kh, (k+1)h]$.

■ 17.2.3 Derivation of the score matching objective

In this section, we present a self-contained derivation of the score matching objective (17.7) for the reader's convenience. See also [Hyv05; Vin11; SE19].

Recall that the problem is to solve

$$\underset{s_t \in \mathscr{F}}{\text{minimize}} \quad \mathbb{E}_{q_t}[\|s_t - \nabla \ln q_t\|^2]\,.$$

This objective cannot be evaluated, even if we replace the expectation over q_t with an empirical average over samples from q_t. The trick is to use an integration by parts identity to reformulate the objective. Here, C will denote any constant that does not depend on the optimization variable s_t. Expanding the square,

$$\mathbb{E}_{q_t}[\|s_t - \nabla \ln q_t\|^2] = \mathbb{E}_{q_t}[\|s_t\|^2 - 2\,\langle s_t, \nabla \ln q_t\rangle] + C\,.$$

We can rewrite the second term using integration by parts:

$$\int \langle s_t, \nabla \ln q_t\rangle\,\mathrm{d}q_t = \int \langle s_t, \nabla q_t\rangle = -\int (\operatorname{div} s_t)\,\mathrm{d}q_t$$

$$= -\iint (\operatorname{div} s_t)\big(\exp(-t)\, x_0 + \sqrt{1-\exp(-2t)}\, z_t\big)\, \mathrm{d}q(x_0)\, \mathrm{d}\gamma^d(z_t)\,,$$

where $\gamma^d = \mathsf{normal}(0, I_d)$ and we used the explicit form of the law of the OU process at time t. Recall the Gaussian integration by parts identity: for any vector field $v : \mathbb{R}^d \to \mathbb{R}^d$,

$$\int (\operatorname{div} v)\, \mathrm{d}\gamma^d = \int \langle x, v(x)\rangle\, \mathrm{d}\gamma^d(x)\,.$$

Applying this identity,

$$\int \langle s_t, \nabla \ln q_t\rangle\, \mathrm{d}q_t = -\frac{1}{\sqrt{1-\exp(-2t)}} \int \langle z_t, s_t(x_t)\rangle\, \mathrm{d}q(x_0)\, \mathrm{d}\gamma^d(z_t)$$

where $x_t = \exp(-t)\, x_0 + \sqrt{1-\exp(-2t)}\, z_t$. Substituting this in,

$$\begin{aligned}\mathbb{E}_{q_t}[\|s_t - \nabla \ln q_t\|^2] &= \mathbb{E}\Big[\|s_t(X_t)\|^2 + \frac{2}{\sqrt{1-\exp(-2t)}}\, \langle Z_t, s_t(X_t)\rangle\Big] + C \\ &= \mathbb{E}\Big[\Big\|s(X_t) + \frac{1}{\sqrt{1-\exp(-2t)}}\, Z_t\Big\|^2\Big] + C\,,\end{aligned}$$

where $X_0 \sim q$ and $Z_t \sim \gamma^d$ are independent, and we set $X_t \coloneqq \exp(-t)\, X_0 + \sqrt{1-\exp(-2t)}\, Z_t$.

17.3 Results

We now state our assumptions and our main results.

17.3.1 Results for DDPM

For DDPM, we make the following mild assumptions on the data distribution q.

Assumption 17.3.1 (Lipschitz score)**.** *For all $t \geq 0$, the score $\nabla \ln q_t$ is L-Lipschitz.*

Assumption 17.3.2 (Second moment bound)**.** *We assume $\mathfrak{m}_2^2 \coloneqq \mathbb{E}_q[\|\cdot\|^2] < \infty$.*

Assumption 17.3.1 is standard and has been used in the prior works [BMR22; LLT22]. However, unlike [LLT22], we do not assume Lipschitzness of the score estimate. Moreover, unlike [De +21; BMR22], we do not assume any convexity or dissipativity assumptions on the potential U, and unlike [LLT22] we do not assume that q satisfies a log-Sobolev inequality. Hence, our assumptions cover a wide

range of highly non-log-concave data distributions. Our proof technique is fairly robust and even Assumption 17.3.1 could be relaxed (as well as other extensions, such as considering the time-changed forward process (17.2)), although we focus on the simplest setting in order to better illustrate the conceptual significance of our results.

We also assume a bound on the score estimation error.

Assumption 17.3.3 (Score estimation error)**.** *For all* $k = 1, \dots, N$,

$$\mathbb{E}_{q_{kh}}[\|s_{kh} - \nabla \ln q_{kh}\|^2] \le \varepsilon_{\text{score}}^2 .$$

This is the same assumption as in [LLT22], and as discussed in §17.2.1, it is a natural and realistic assumption in light of the derivation of score matching.

Our main result for DDPM is the following theorem.

Theorem 17.3.4 (DDPM)**.** *Suppose that Assumptions 17.3.1, 17.3.2, and 17.3.3 hold. Let* p_T *be the output of the DDPM algorithm (§17.2.1) at time* T*, and suppose that the step size* $h := T/N$ *satisfies* $h \lesssim 1/L$*, where* $L \ge 1$*. Then, it holds that*

$$\mathsf{TV}(p_T, q) \lesssim \underbrace{\sqrt{\mathsf{KL}(q \parallel \gamma^d)}\exp(-T)}_{\text{convergence of forward process}} + \underbrace{(L\sqrt{dh} + L\mathfrak{m}_2 h)\sqrt{T}}_{\text{discretization error}} + \underbrace{\varepsilon_{\text{score}}\sqrt{T}}_{\text{score estimation}} .$$

Proof. See §17.5. □

To interpret this result, suppose that $\mathsf{KL}(q \parallel \gamma^d) \le \text{poly}(d)$ and $\mathfrak{m}_2 \le d$. Choosing $T \asymp \log(\mathsf{KL}(q \parallel \gamma^d)/\varepsilon)$ and $h \asymp \frac{\varepsilon^2}{L^2 d}$, and hiding logarithmic factors,

$$\mathsf{TV}(p_T, q) \le \widetilde{O}(\varepsilon + \varepsilon_{\text{score}}) , \qquad \text{for } N = \widetilde{\Theta}\Big(\frac{L^2 d}{\varepsilon^2}\Big) .$$

In particular, in order to have $\mathsf{TV}(p_T, q) \le \varepsilon$, it suffices to have score error $\varepsilon_{\text{score}} \le \widetilde{O}(\varepsilon)$.

We remark that the iteration complexity of $N = \widetilde{\Theta}(\frac{L^2 d}{\varepsilon^2})$ matches state-of-the-art complexity bounds for the Langevin Monte Carlo (LMC) algorithm for sampling under a log-Sobolev inequality (LSI), see [VW19] and §3. This provides some evidence that our discretization bounds are of the correct order, at least with respect to the dimension and accuracy parameters, and without higher-order smoothness assumptions.

17.3.2 Consequences for arbitrary data distributions with bounded support

We now elaborate upon the implications of our results under the *sole* assumption that the data distribution q is compactly supported, $\operatorname{supp} q \subseteq \mathsf{B}(0, R)$. In particular, we do not assume that q has a smooth density w.r.t. Lebesgue measure, which allows for studying the case when q is supported on a lower-dimensional submanifold of $\mathbb{R}^d$ as in the *manifold hypothesis*. This setting was investigated recently in [De 22].

For this setting, our results do not apply directly because the score function of q is not well-defined and hence Assumption 17.3.1 fails to hold. Also, the bound in Theorem 17.3.4 has a term involving $\mathsf{KL}(q \parallel \gamma^d)$ which is infinite if q is not absolutely continuous w.r.t. γ^d. As pointed out by [De 22], in general we cannot obtain non-trivial guarantees for $\mathsf{TV}(p_T, q)$, because p_T has full support and therefore $\mathsf{TV}(p_T, q) = 1$ under the manifold hypothesis. Nevertheless, we show that we can apply our results using an early stopping technique.

Namely, consider q_t the law of the OU process at a time $t > 0$, initialized at q. Then, we show in Lemma 17.5.7 that, if $t \asymp \varepsilon_{W_2}^2/(\sqrt{d}\,(R \vee \sqrt{d}))$ where $0 < \varepsilon_{W_2} \ll \sqrt{d}$, then q_t satisfies Assumption 17.3.1 with $L \lesssim dR^2\,(R \vee \sqrt{d})^2/\varepsilon_{W_2}^4$, $\mathsf{KL}(q_t \parallel \gamma^d) \le \operatorname{poly}(R, d, 1/\varepsilon)$, and $W_2(q_t, q) \le \varepsilon_{W_2}$. By substituting q by q_t into the result of Theorem 17.3.4, we obtain Corollary 17.3.5 below.

Taking q_t as the new target corresponds to stopping the algorithm early: instead of running the algorithm backward for a time T, we run the algorithm backward for a time $T - t$ (here, $T - t$ should be a multiple of the step size h).

Corollary 17.3.5 (Compactly supported data)**.** *Suppose that q is supported on the ball of radius $R \ge 1$. Let $t \asymp \varepsilon_{W_2}^2/(\sqrt{d}\,(R \vee \sqrt{d}))$. Then, the output p_{T-t} of DDPM is $\varepsilon_{\mathrm{TV}}$-close in TV to the distribution q_t, which is ε_{W_2}-close in W_2 to q, provided that h is chosen appropriately according to Theorem 17.3.4 and*

$$N = \widetilde{\Theta}\Big(\frac{d^3 R^4\,(R \vee \sqrt{d})^4}{\varepsilon_{\mathrm{TV}}^2\,\varepsilon_{W_2}^8}\Big) \qquad \text{and} \qquad \varepsilon_{\mathrm{score}} \le \widetilde{O}(\varepsilon_{\mathrm{TV}})\,.$$

Observing that both the TV and W_1 metrics are upper bounds for the bounded Lipschitz metric $\mathsf{d}_{\mathrm{BL}}(\mu, \nu) \coloneqq \sup\{\int f \,\mathrm{d}(\mu - \nu) \;\big|\; f : \mathbb{R}^d \to [-1, 1] \text{ is 1-Lipschitz}\}$, we immediately obtain the following corollary.

Corollary 17.3.6 (Compactly supported data, BL metric)**.** *Suppose that q is supported on the ball of radius $R \ge 1$. Let $t \asymp \varepsilon^2/(\sqrt{d}\,(R \vee \sqrt{d}))$. Then, the output p_{T-t} of the DDPM algorithm satisfies $\mathsf{d}_{\mathrm{BL}}(p_{T-t}, q) \le \varepsilon$, provided that the step size h is chosen appropriately according to Theorem 17.3.4 and $N = \widetilde{\Theta}(d^3 R^4\,(R \vee \sqrt{d})^4/\varepsilon^{10})$ and $\varepsilon_{\mathrm{score}} \le \widetilde{O}(\varepsilon)$.*

Finally, if the output p_{T-t} of DDPM at time $T-t$ is projected onto $\mathsf{B}(0, R_0)$ for an appropriate choice of R_0, then we can also translate our guarantees to the standard W_2 metric, which we state as the following corollary.

Corollary 17.3.7 (Compactly supported data, W_2 metric)**.** *Suppose that q is supported on the ball of radius $R \geq 1$. Let $t \asymp \varepsilon^2/(\sqrt{d}\,(R \vee \sqrt{d}))$, and let p_{T-t,R_0} denote the output of DDPM at time $T-t$ projected onto $\mathsf{B}(0, R_0)$ for $R_0 = \widetilde{\Theta}(R)$. Then, it holds that $W_2(p_{T-t,R_0}, q) \leq \varepsilon$, provided that the step size h is chosen appropriately according to Theorem 17.3.4, $N = \widetilde{\Theta}(d^3 R^8\,(R \vee \sqrt{d})^4/\varepsilon^{12})$, and $\varepsilon_{\text{score}} \leq \widetilde{O}(\varepsilon)$.*

Note that the dependencies in the three corollaries above are polynomial in all of the relevant problem parameters. In particular, since the last corollary holds in the W_2 metric, it is directly comparable to [De 22] and vastly improves upon the exponential dependencies therein.

■ 17.3.3 Results for CLD

In order to state our results for score-based generative modelling based on the CLD, we must first modify Assumptions 17.3.1 and 17.3.3 accordingly.

Assumption 17.3.8. *For all $t \geq 0$, the score $\nabla_v \ln \boldsymbol{q}_t$ is L-Lipschitz.*

Assumption 17.3.9. *For all $k = 1, \ldots, N$,*

$$\mathbb{E}_{\boldsymbol{q}_{kh}}[\|\boldsymbol{s}_{kh} - \nabla_v \ln \boldsymbol{q}_{kh}\|^2] \leq \varepsilon_{\text{score}}^2\,.$$

If we ignore the dependence on L and assume that the score estimate is sufficiently accurate, then the iteration complexity guarantee of Theorem 17.3.4 is $N = \widetilde{\Theta}(d/\varepsilon^2)$. On the other hand, recall from §17.2.2 that based on intuition from the literature on log-concave sampling and from empirical findings in [DVK22], we might expect that SGMs based on the CLD have a smaller iteration complexity than DDPM. We establish the following theorem.

Theorem 17.3.10 (CLD)**.** *Suppose that Assumptions 17.3.2, 17.3.8, and 17.3.9 hold. Let $\boldsymbol{p}_T$ be the output of the SGM algorithm based on the CLD (§17.2.2) at time T, and suppose that the step size $h := T/N$ satisfies $h \lesssim 1/L$, where $L \geq 1$. Then, there is a universal constant $c > 0$ such that*

$$\mathsf{TV}(\boldsymbol{p}_T, q \otimes \gamma^d) \lesssim \underbrace{\sqrt{\mathsf{KL}(q \parallel \gamma^d) + \mathsf{FI}(q \parallel \gamma^d)}\,\exp(-cT)}_{\textit{convergence of forward process}} + \underbrace{(L\sqrt{dh} + L\mathfrak{m}_2 h)\,\sqrt{T}}_{\textit{discretization error}} + \underbrace{\varepsilon_{\text{score}}\sqrt{T}}_{\textit{score estimation error}}$$

where $\mathsf{FI}(q \parallel \gamma^d)$ is the relative Fisher information $\mathsf{FI}(q \parallel \gamma^d) := \mathbb{E}_q[\|\nabla \ln(q/\gamma^d)\|^2]$.

Note that the result of Theorem 17.3.10 is in fact no better than our guarantee for DDPM in Theorem 17.3.4. Although this is possibly an artefact of our analysis, we believe that it is in fact fundamental. From the form of the reverse process (17.11), the SGM based on CLD lacks a certain property (that the discretization error should only depend on the size of the increment of the X process, not the increments of both the X and V processes) which is crucial for the improved dimension dependence of the CLD over Langevin in log-concave sampling. Hence, in general, we conjecture that under our assumptions, SGMs based on the CLD do not achieve a better dimension dependence than DDPM.

We provide evidence for our conjecture via a lower bound. In our proofs of Theorems 17.3.4 and 17.3.10, we rely on bounding the KL divergence between certain measures on the path space $\mathcal{C}([0,T];\mathbb{R}^d)$ via Girsanov's theorem. The following result lower bounds this KL divergence, even for the setting in which the score estimate is perfect ($\varepsilon_{\text{score}} = 0$) and the data distribution q is Gaussian.

Theorem 17.3.11. *Let $\boldsymbol{p}_T$ be the output of the SGM algorithm based on the CLD (§17.2.2) at time T, where the data distribution q is the standard Gaussian γ^d, and the score estimate is exact ($\varepsilon_{\text{score}} = 0$). Suppose that the step size h satisfies $h \leq \frac{1}{10}$. Then, for the path measures $\boldsymbol{P}_T$ and $\boldsymbol{Q}_T^{\leftarrow}$ of the algorithm and the continuous-time process (17.11) respectively, it holds that*

$$\mathsf{KL}(\boldsymbol{Q}_T^{\leftarrow} \parallel \boldsymbol{P}_T) \geq dhT\,.$$

Theorem 17.3.11 shows that in order to make the KL divergence between the path measures small, we must take $h \lesssim 1/d$, which leads to an iteration complexity that scales linearly in the dimension d. Theorem 17.3.11 is not a proof that SGMs based on the CLD cannot achieve better than linear dimension dependence, as it is possible that the output $\boldsymbol{p}_T$ of the SGM is close to $q \otimes \gamma^d$ even if the path measures are not close, but it rules out the possibility of obtaining a better dimension dependence via our Girsanov-based proof technique. We believe that it provides compelling evidence for our conjecture, i.e., that under our assumptions, the CLD does not improve the complexity of SGMs over DDPM.

We remark that in this section, we have only considered the error arising from discretization of the SDE. It is possible that the score function for the SGM with the CLD is easier to estimate than the score function for DDPM, providing a *statistical* benefit of using the CLD. Indeed, under the manifold hypothesis, the score $\nabla \ln q_t$ for DDPM blows up at $t = 0$, but the score $\nabla_v \ln \boldsymbol{q}_t$ for CLD is well-defined at $t = 0$, and hence may lead to improvements over DDPM. We do not investigate this question here and leave it as future work.

We omit the proofs of the results for CLD from this thesis and we refer to the paper [Che+23a] for the details and proofs.

17.4 Technical overview

We now give a technical overview for the proof for DDPM (Theorem 17.3.4). The proof for CLD (Theorem 17.3.10) follows along similar lines.

Recall that we must deal with three sources of error: (1) the estimation of the score function; (2) the discretization of the SDE; and (3) the initialization of the reverse process at γ^d rather than at q_T.

First, we ignore the errors (1) and (2), and focus on the error (3). Hence, we consider the continuous-time reverse SDE (17.5), initialized from either γ^d or from q_T. Let the law of the two processes at time t be denoted $\tilde{p}_t$ and q_{T-t} respectively; how fast do these laws diverge away from each other?

The two main ways to study Markov diffusions is via the 2-Wasserstein distance W_2, or via information divergences such as the KL divergence or the χ^2 divergence. In order for the reverse process to be contractive in the W_2 distance, one typically needs some form of log-concavity assumption for the data distribution q. For example, if $\nabla \ln q(x) = -x/\sigma^2$ (i.e., $q \sim \mathsf{normal}(0, \sigma^2 I_d)$), then for the reverse process (17.5) we have

$$\mathrm{d}\bar{X}_T^{\leftarrow} = \{\bar{X}_T^{\leftarrow} + 2\nabla \ln q(\bar{X}_T^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t = \big(1 - \frac{2}{\sigma^2}\big)\,\bar{X}_T^{\leftarrow}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,.$$

For $\sigma^2 \gg 1$, the coefficient in front of $\bar{X}_T^{\leftarrow}$ is positive; this shows that for times near T, the reverse process is actually *expansive*, rather than contractive. This poses an obstacle for an analysis in W_2. Although it is possible to perform a W_2 analysis using a weaker condition, such as a dissipativity condition, it typically leads to exponential dependence on the problem parameters (e.g., [De 22]).

On the other hand, the situation is different for an information divergence d. By the data-processing inequality, we always have

$$\mathsf{d}(q_{T-t}, \tilde{p}_t) \le \mathsf{d}(q_T, \tilde{p}_0) = \mathsf{d}(q_T, \gamma^d)\,.$$

This motivates studying the processes via information divergences. We remark that the convergence of reversed SDEs has been studied in the context of log-concave sampling in §4 for the proximal sampler algorithm [LST21c], providing the intuition behind these observations.

Next, we consider the score estimation error (1) and the discretization error (2). In order to perform a discretization analysis in KL or χ^2, there are two salient proof techniques. The first is the interpolation method of [VW19] (originally for KL divergence, but extended to χ^2 divergence in §3), which is the method used in [LLT22]. The interpolation method writes down a differential inequality for $\partial_t \mathsf{d}(q_{T-t}, p_t)$, which is used to bound $\mathsf{d}(q_{T-(k+1)h}, p_{(k+1)h})$ in terms of $\mathsf{d}(q_{T-kh}, p_{kh})$ and an additional error term. Unfortunately, the analysis of [LLT22] required

taking d to be the χ^2 divergence, for which the interpolation method is quite delicate. In particular, the error term is bounded using a log-Sobolev assumption on q, see §3 for further discussion. Instead, we pursue the second approach, which is to apply Girsanov's theorem from stochastic calculus and to instead bound the divergence between measures on path space; this turns out to be doable using standard techniques. This is because, as noted in §3, the Girsanov approach is more flexible as it requires less stringent assumptions.[2]

To elaborate, the main difficulty of using the interpolation method with an L^2-accurate score estimate (Assumption 17.3.3) is that the score estimation error is controlled by assumption under the law of the *true* process (17.5), but the interpolation analysis requires a control of the score estimation error under the law of the *algorithm* (17.9). Consequently, the work of [LLT22] required an involved change of measure argument in order to relate the errors under the two processes. In contrast, the Girsanov approach allows us to directly work with the score estimation error under the true process (17.5).

Notation

Stochastic processes and their laws.

- The data distribution is $q = q_0$.
- The forward process (17.1) is denoted $(\bar{X}_t)_{t\in[0,T]}$, and $\bar{X}_t \sim q_t$.
- The reverse process (17.5) is denoted $(\bar{X}_t^{\leftarrow})_{t\in[0,T]}$, where $\bar{X}_t^{\leftarrow} := \bar{X}_{T-t} \sim q_{T-t}$.
- The SGM algorithm (17.9) is denoted $(X_t^{\leftarrow})_{t\in[0,T]}$, and $X_t^{\leftarrow} \sim p_t$. Recall that we initialize at $p_0 = \gamma^d$, the standard Gaussian measure.
- The process $(X_t^{\leftarrow,\infty,q_T})_{t\in[0,T]}$ is the same as $(X_t^{\leftarrow})_{t\in[0,T]}$, except that we initialize this process at q_T rather than at γ^d. We write $X_t^{\leftarrow,\infty,q_T} \sim p_t^{\infty,q_T}$.

Conventions for Girsanov's theorem. When we apply Girsanov's theorem, it is convenient to instead think about a single stochastic process, which for ease of notation we denote simply via $(X_t)_{t\in[0,T]}$, and we consider different measures over the path space $\mathcal{C}([0,T];\mathbb{R}^d)$.

The three measures we consider over path space are:

- $Q_T^{\leftarrow}$, under which $(X_t)_{t\in[0,T]}$ has the law of the reverse process (17.5);

[2] After the first draft of this work was made available online, we became aware of the concurrent and independent work of [Liu+22] which also uses an approach based on Girsanov's theorem.

- P_T^{∞,q_T}, under which $(X_t)_{t\in[0,T]}$ has the law of the SGM algorithm initialized at q_T (corresponding to the process $(X_t^{\leftarrow,\infty,q_T})_{t\in[0,T]}$ defined above).

We also use the following notion from stochastic calculus [Le 16, Definition 4.6]:

- A local martingale $(L_t)_{t\in[0,T]}$ is a stochastic process s.t. there exists a sequence of non-decreasing stopping times $T_n \to T$ s.t. $L^n = (L_{t\wedge T_n})_{t\in[0,T]}$ is a martingale.

Other parameters. We recall that $T > 0$ denotes the total time for which we run the forward process; $h > 0$ is the step size of the discretization; $L \geq 1$ is the Lipschitz constant of the score function; $\mathfrak{m}_2^2 \coloneqq \mathbb{E}_q[\|\cdot\|^2]$ is the second moment under the data distribution; and $\varepsilon_{\text{score}}$ is the L^2 score estimation error.

17.5 Proofs

17.5.1 Preliminaries on Girsanov's theorem and a first attempt at applying Girsanov's theorem

First, we recall a consequence of Girsanov's theorem that can be obtained by combining Pages 136–139, Theorem 5.22, and Theorem 4.13 of [Le 16].

Theorem 17.5.1. *For $t \in [0,T]$, let $\mathcal{L}_t = \int_0^t b_s \,\mathrm{d}B_s$ where B is a Q-Brownian motion. Assume $\mathbb{E}_Q \int_0^T \|b_s\|^2 \,\mathrm{d}s < \infty$. Then, $\mathcal{L}$ is a Q-martingale in $L^2(Q)$. Moreover, if*

$$\mathbb{E}_Q \,\mathcal{E}(\mathcal{L})_T = 1\,, \quad \text{where} \quad \mathcal{E}(\mathcal{L})_t \coloneqq \exp\Big(\int_0^t b_s \,\mathrm{d}B_s - \frac{1}{2}\int_0^t \|b_s\|^2 \,\mathrm{d}s\Big)\,, \tag{17.12}$$

then $\mathcal{E}(\mathcal{L})$ is also a Q-martingale and the process

$$t \mapsto B_t - \int_0^t b_s \,\mathrm{d}s \tag{17.13}$$

is a Brownian motion under $P \coloneqq \mathcal{E}(\mathcal{L})_T\, Q$, the probability distribution with density $\mathcal{E}(\mathcal{L})_T$ w.r.t. Q.

If the assumptions of Girsanov's theorem are satisfied (i.e., the condition in (17.12)), we can apply Girsanov's theorem to $Q = Q_T^{\leftarrow}$ and

$$b_t = \sqrt{2}\,(s_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_t))\,, \tag{17.14}$$

where $t \in [kh, (k+1)h]$. This tells us that under $P = \mathcal{E}(\mathcal{L})_T \, Q_T^{\leftarrow}$, there exists a Brownian motion $(\beta_t)_{t\in[0,T]}$ s.t.

$$\mathrm{d}B_t = \sqrt{2}\,(s_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_t))\,\mathrm{d}t + \mathrm{d}\beta_t\,. \tag{17.15}$$

Recall that under $Q_T^{\leftarrow}$ we have a.s.

$$\mathrm{d}X_t = \{X_t + 2\,\nabla \ln q_{T-t}(X_t)\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \qquad X_0 \sim q_T\,. \tag{17.16}$$

The equation above still holds P-a.s. since $P \ll Q_T^{\leftarrow}$ (even if B is no longer a P-Brownian motion). Plugging (17.15) into (17.16) we have P-a.s.,[3]

$$\mathrm{d}X_t = \{X_t + 2\,s_{T-kh}(X_{kh})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\beta_t\,, \qquad X_0 \sim q_T\,.$$

In other words, under P, the distribution of X is the SGM algorithm started at q_T, i.e., $P = P_T^{\infty,q_T} = \mathcal{E}(\mathcal{L})_T \, Q_T^{\leftarrow}$. Therefore,

$$\mathsf{KL}(Q_T^{\leftarrow} \,\|\, P_T^{\infty,q_T}) = \mathbb{E}_{Q_T^{\leftarrow}} \ln \frac{\mathrm{d}Q_T^{\leftarrow}}{\mathrm{d}P_T^{\infty,q_T}} = \mathbb{E}_{Q_T^{\leftarrow}} \ln \mathcal{E}(\mathcal{L})_T^{-1} \tag{17.17}$$

$$= \sum_{k=0}^{N-1} \mathbb{E}_{Q_T^{\leftarrow}} \int_{kh}^{(k+1)h} \|s_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\|^2\,\mathrm{d}t\,, \tag{17.18}$$

where we used $\mathbb{E}_{Q_T^{\leftarrow}} \mathcal{L}_t = 0$ because $\mathcal{L}$ is a martingale.

The equality (17.17) allows us to bound the discrepancy between the SGM algorithm and the reverse process.

■ 17.5.2 Checking the assumptions of Girsanov's theorem and the Girsanov discretization argument

In most applications of Girsanov's theorem in sampling, a sufficient condition for (17.12) to hold, known as *Novikov's condition*, is satisfied. Here, Novikov's condition writes

$$\mathbb{E}_{Q_T^{\leftarrow}} \exp\Big(\sum_{k=0}^{N-1} \int_{kh}^{(k+1)h} \|s_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\|^2\,\mathrm{d}t\Big) < \infty\,, \tag{17.19}$$

and if Novikov's condition holds, we can apply Girsanov's theorem directly. However, under Assumptions 17.3.1, 17.3.2, and 17.3.3 alone, Novikov's condition need

[3] We still have $X_0 \sim q_T$ under P because the marginal at time $t = 0$ of P is equal to the marginal at time $t = 0$ of $Q_T^{\leftarrow}$. That is a consequence of the fact that $\mathcal{E}(\mathcal{L})$ is a (true) $Q_T^{\leftarrow}$-martingale.

not hold. Indeed, in order to check Novikov's condition, we would want X_0 to have sub-Gaussian tails for instance.

Furthermore, we also could not check that the condition (17.12), which is weaker than Novikov's condition, holds. Therefore, in the proof of the next Theorem, we use a approximation technique to show that

$$\mathsf{KL}(Q_T^{\leftarrow} \parallel P_T^{\infty,q_T}) = \mathbb{E}_{Q_T^{\leftarrow}} \ln \frac{\mathrm{d}Q_T^{\leftarrow}}{\mathrm{d}P_T^{\infty,q_T}} \leq \mathbb{E}_{Q_T^{\leftarrow}} \ln \mathcal{E}(\mathcal{L})_T^{-1} \tag{17.20}$$

$$= \sum_{k=0}^{N-1} \mathbb{E}_{Q_T^{\leftarrow}} \int_{kh}^{(k+1)h} \|s_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\|^2 \,\mathrm{d}t\,. \tag{17.21}$$

We then use a discretization argument based on stochastic calculus to further bound this quantity. The result is the following theorem.

Theorem 17.5.2 (Discretization error for DDPM)**.** *Suppose that the Assumptions 17.3.1, 17.3.2, and 17.3.3 hold. Let $Q_T^{\leftarrow}$ and P_T^{∞,q_T} denote the measures on path space corresponding to the reverse process (17.5) and the SGM algorithm with L^2-accurate score initialized at q_T. Assume that $L \geq 1$ and $h \lesssim 1/L$. Then,*

$$\mathsf{TV}(P_T^{\infty,q_T}, Q_T^{\leftarrow})^2 \leq \mathsf{KL}(Q_T^{\leftarrow} \parallel P_T^{\infty,q_T}) \lesssim (\varepsilon_{\text{score}}^2 + L^2 dh + L^2 \mathfrak{m}_2^2 h^2)\, T\,.$$

Proof. We start by proving

$$\sum_{k=0}^{N-1} \mathbb{E}_{Q_T^{\leftarrow}} \int_{kh}^{(k+1)h} \|s_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\|^2 \,\mathrm{d}t \lesssim (\varepsilon_{\text{score}}^2 + L^2 dh + L^2 \mathfrak{m}_2^2 h^2)\, T\,.$$

Then, we give the approximation argument to prove the inequality (17.20).

Bound on the discretization error. For $t \in [kh, (k+1)h]$, we decompose

$$\mathbb{E}_{Q_T^{\leftarrow}}[\|s_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\|^2] \tag{17.22}$$

$$\lesssim \mathbb{E}_{Q_T^{\leftarrow}}[\|s_{T-kh}(X_{kh}) - \nabla \ln q_{T-kh}(X_{kh})\|^2] \tag{17.23}$$

$$+ \mathbb{E}_{Q_T^{\leftarrow}}[\|\nabla \ln q_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_{kh})\|^2] \tag{17.24}$$

$$+ \mathbb{E}_{Q_T^{\leftarrow}}[\|\nabla \ln q_{T-t}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\|^2] \tag{17.25}$$

$$\lesssim \varepsilon_{\text{score}}^2 + \mathbb{E}_{Q_T^{\leftarrow}}\left[\left\|\nabla \ln \frac{q_{T-kh}}{q_{T-t}}(X_{kh})\right\|^2\right] + L^2\, \mathbb{E}_{Q_T^{\leftarrow}}[\|X_{kh} - X_t\|^2]\,. \tag{17.26}$$

We must bound the change in the score function along the forward process. If $S : \mathbb{R}^d \to \mathbb{R}^d$ is the mapping $S(x) \coloneqq \exp(-(t-kh))\, x$, then $q_{T-kh} = S_{\#} q_{T-t} * \mathsf{normal}(0, 1 - \exp(-2\,(t-kh)))$. We can then use [LLT22, Lemma C.12] with

$\alpha = \exp(t-kh) = 1 + O(h)$ and $\sigma^2 = 1 - \exp(-2\,(t-kh)) = O(h)$ to obtain

$$\left\|\nabla \ln \frac{q_{T-kh}}{q_{T-t}}(X_{kh})\right\|^2 \lesssim L^2 dh + L^2 h^2 \left\|X_{kh}\right\|^2 + (1+L^2)\, h^2 \left\|\nabla \ln q_{T-t}(X_{kh})\right\|^2 \tag{17.27}$$

$$\lesssim L^2 dh + L^2 h^2 \left\|X_{kh}\right\|^2 + L^2 h^2 \left\|\nabla \ln q_{T-t}(X_{kh})\right\|^2 \tag{17.28}$$

where the last line uses $L \geq 1$.

For the last term,

$$\left\|\nabla \ln q_{T-t}(X_{kh})\right\|^2 \lesssim \left\|\nabla \ln q_{T-t}(X_t)\right\|^2 + \left\|\nabla \ln q_{T-t}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\right\|^2 \tag{17.29}$$

$$\lesssim \left\|\nabla \ln q_{T-t}(X_t)\right\|^2 + L^2 \left\|X_{kh} - X_t\right\|^2, \tag{17.30}$$

where the second term above is absorbed into the third term of the decomposition (17.26). Hence,

$$\begin{aligned}
&\mathbb{E}_{Q_T^{\leftarrow}}[\|s_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\|^2] \\
&\quad\lesssim \varepsilon_{\text{score}}^2 + L^2 dh + L^2 h^2\, \mathbb{E}_{Q_T^{\leftarrow}}[\|X_{kh}\|^2] \\
&\qquad + L^2 h^2\, \mathbb{E}_{Q_T^{\leftarrow}}[\|\nabla \ln q_{T-t}(X_t)\|^2] + L^2\, \mathbb{E}_{Q_T^{\leftarrow}}[\|X_{kh} - X_t\|^2]\,.
\end{aligned}$$

Using the fact that under $Q_T^{\leftarrow}$, the process $(X_t)_{t\in[0,T]}$ is the time reversal of the forward process $(\bar{X}_t)_{t\in[0,T]}$, we can apply the moment bounds in Lemma 17.5.3 and the movement bound in Lemma 17.5.4 to obtain

$$\begin{aligned}
&\mathbb{E}_{Q_T^{\leftarrow}}[\|s_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\|^2] \\
&\quad\lesssim \varepsilon_{\text{score}}^2 + L^2 dh + L^2 h^2\,(d + \mathfrak{m}_2^2) + L^3 dh^2 + L^2\,(\mathfrak{m}_2^2 h^2 + dh) \\
&\quad\lesssim \varepsilon_{\text{score}}^2 + L^2 dh + L^2 \mathfrak{m}_2^2 h^2\,.
\end{aligned}$$

Approximation argument. For $t \in [0,T]$, let $\mathcal{L}_t = \int_0^t b_s \,\mathrm{d}B_s$ where B is a $Q_T^{\leftarrow}$-Brownian motion and we define

$$b_t = \sqrt{2}\,\{s_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\}\,, \tag{17.31}$$

for $t \in [kh, (k+1)h]$. We proved that $\mathbb{E}_{Q_T^{\leftarrow}} \int_0^T \|b_s\|^2 \,\mathrm{d}s \lesssim (\varepsilon_{\text{score}}^2 + L^2 dh + L^2 \mathfrak{m}_2^2 h^2)\, T < \infty$. Using [Le 16, Proposition 5.11], $(\mathcal{E}(\mathcal{L})_t)_{t\in[0,T]}$ is a local martingale. Therefore, there exists a non-decreasing sequence of stopping times $T_n \nearrow T$ s.t. $(\mathcal{E}(\mathcal{L})_{t\wedge T_n})_{t\in[0,t]}$ is a martingale. Note that $\mathcal{E}(\mathcal{L})_{t\wedge T_n} = \mathcal{E}(\mathcal{L}^n)_t$ where $\mathcal{L}_t^n = \mathcal{L}_{t\wedge T_n}$. Since $\mathcal{E}(\mathcal{L}^n)$ is a martingale, we have

$$\mathbb{E}_{Q_T^{\leftarrow}}\, \mathcal{E}(\mathcal{L}^n)_T = \mathbb{E}_{Q_T^{\leftarrow}}\, \mathcal{E}(\mathcal{L}^n)_0 = 1\,,$$

i.e., $\mathbb{E}_{Q_T^{\leftarrow}} \mathcal{E}(\mathcal{L})_{T_n} = 1$.

We apply Girsanov's theorem to $\mathcal{L}_t^n = \int_0^t b_s \, \mathbb{1}_{[0,T_n]}(s) \, \mathrm{d}B_s$, where B is a $Q_T^{\leftarrow}$-Brownian motion. Since $\mathbb{E}_{Q_T^{\leftarrow}} \int_0^T \|b_s \, \mathbb{1}_{[0,T_n]}(s)\|^2 \, \mathrm{d}s \le \mathbb{E}_{Q_T^{\leftarrow}} \int_0^T \|b_s\|^2 \, \mathrm{d}s < \infty$ (see the last paragraph) and $\mathbb{E}_{Q_T^{\leftarrow}} \mathcal{E}(\mathcal{L}^n)_T = 1$, we obtain that under $P^n := \mathcal{E}(\mathcal{L}^n)_T \, Q_T^{\leftarrow}$ there exists a Brownian motion β^n s.t. for $t \in [0,T]$,

$$\mathrm{d}B_t = \sqrt{2} \, \{ s_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_t) \} \, \mathbb{1}_{[0,T_n]}(t) \, \mathrm{d}t + \mathrm{d}\beta_t^n \, . \tag{17.32}$$

Recall that under $Q_T^{\leftarrow}$ we have a.s.

$$\mathrm{d}X_t = \{ X_t + 2 \nabla \ln q_{T-t}(X_t) \} \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}B_t \, , \qquad X_0 \sim q_T \, . \tag{17.33}$$

The equation above still holds P^n-a.s. since $P^n \ll Q_T^{\leftarrow}$. Combining the last two equations we then obtain P^n-a.s.,

$$\begin{aligned} \mathrm{d}X_t = & \{ X_t + 2 \, s_{T-kh}(X_{kh}) \} \, \mathbb{1}_{[0,T_n]}(t) \, \mathrm{d}t \\ & + \{ X_t + 2 \nabla \ln q_{T-t}(X_t) \} \, \mathbb{1}_{[T_n,T]}(t) \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}\beta_t^n \, , \end{aligned} \tag{17.34}$$

and $X_0 \sim q_T$. In other words, P^n is the law of the solution of the SDE (17.34). At this stage we have the bound

$$\begin{aligned} \mathsf{KL}(Q_T^{\leftarrow} \parallel P^n) &= \mathbb{E}_{Q_T^{\leftarrow}} \ln \mathcal{E}(\mathcal{L})_{T_n}^{-1} = \mathbb{E}_{Q_T^{\leftarrow}} \Big[-\mathcal{L}_{T_n} + \frac{1}{2} \int_0^{T_n} \|b_s\|^2 \, \mathrm{d}s \Big] \\ &= \mathbb{E}_{Q_T^{\leftarrow}} \frac{1}{2} \int_0^{T_n} \|b_s\|^2 \, \mathrm{d}s \le \mathbb{E}_{Q_T^{\leftarrow}} \frac{1}{2} \int_0^{T} \|b_s\|^2 \, \mathrm{d}s \\ &\lesssim (\varepsilon_{\text{score}}^2 + L^2 dh + L^2 \mathfrak{m}_2^2 h^2) \, T \, , \end{aligned} \tag{17.35}$$

where we used that $\mathbb{E}_{Q_T^{\leftarrow}} \mathcal{L}_{T_n} = 0$ because $\mathcal{L}$ is a $Q_T^{\leftarrow}$-martingale and T_n is a bounded stopping time [Le 16, Corollary 3.23]. Our goal is now to show that we can obtain the final result by an approximation argument.

We consider a coupling of $(P^n)_{n \in \mathbb{N}}, P_T^{\infty, q_T}$: a sequence of stochastic processes $(X^n)_{n \in \mathbb{N}}$ over the same probability space, a stochastic process X and a single Brownian motion W over that space s.t.[4]

$$\begin{aligned} \mathrm{d}X_t^n = & \{ X_t^n + 2 \, s_{T-kh}(X_{kh}^n) \} \, \mathbb{1}_{[0,T_n]}(t) \, \mathrm{d}t \\ & + \{ X_t^n + 2 \nabla \ln q_{T-t}(X_t^n) \} \, \mathbb{1}_{[T_n,T]}(t) \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}W_t \, , \end{aligned}$$

and

$$\mathrm{d}X_t = \{ X_t + 2 \, s_{T-kh}(X_{kh}^n) \} \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}W_t \, ,$$

[4] Such a coupling always exists, see [Le 16, Corollary 8.5].

with $X_0 = X_0^n$ a.s. and $X_0 \sim q_T$. Note that the distribution of X^n (resp. X) is P^n (resp. P_T^{∞,q_T}).

Let $\varepsilon > 0$ and consider the map $\pi_\varepsilon : \mathcal{C}([0,T];\mathbb{R}^d) \to \mathcal{C}([0,T];\mathbb{R}^d)$ defined by

$$\pi_\varepsilon(\omega)(t) \coloneqq \omega(t \wedge T - \varepsilon)\,.$$

Noting that $X_t^n = X_t$ for every $t \in [0, T_n]$ and using Lemma 17.5.5, we have $\pi_\varepsilon(X^n) \to \pi_\varepsilon(X)$ a.s., uniformly over $[0,T]$. Therefore, $\pi_{\varepsilon\#}P^n \to \pi_{\varepsilon\#}P_T^{\infty,q_T}$ weakly. Using the lower semicontinuity of the KL divergence and the data-processing inequality [AGS08, Lemma 9.4.3 and Lemma 9.4.5], we obtain

$$\begin{aligned}
\mathsf{KL}((\pi_\varepsilon)_\# Q_T^{\leftarrow} \parallel (\pi_\varepsilon)_\# P_T^{\infty,q_T}) &\le \liminf_{n\to\infty} \mathsf{KL}((\pi_\varepsilon)_\# Q_T^{\leftarrow} \parallel (\pi_\varepsilon)_\# P^n) && (17.36)\\
&\le \liminf_{n\to\infty} \mathsf{KL}(Q_T^{\leftarrow} \parallel P^n) && (17.37)\\
&\lesssim (\varepsilon_{\text{score}}^2 + L^2 dh + L^2 \mathfrak{m}_2^2 h^2)\, T\,. && (17.38)
\end{aligned}$$

Finally, using Lemma 17.5.6, $\pi_\varepsilon(\omega) \to \omega$ as $\varepsilon \to 0$, uniformly over $[0,T]$. Therefore, using [AGS08, Corollary 9.4.6], $\mathsf{KL}((\pi_\varepsilon)_\# Q_T^{\leftarrow} \parallel (\pi_\varepsilon)_\# P_T^{\infty,q_T}) \to \mathsf{KL}(Q_T^{\leftarrow} \parallel P_T^{\infty,q_T})$ as $\varepsilon \searrow 0$. Therefore,

$$\mathsf{KL}(Q_T^{\leftarrow} \parallel P_T^{\infty,q_T}) \lesssim (\varepsilon_{\text{score}}^2 + L^2 dh + L^2 \mathfrak{m}_2^2 h^2)\, T\,. \tag{17.39}$$

We conclude with Pinsker's inequality ($\mathsf{TV}^2 \le \mathsf{KL}$). □

■ 17.5.3 Proof of Theorem 17.3.4

We can now conclude our main result.

Proof of Theorem 17.3.4. We recall the notation from §17.4. By the data processing inequality,

$$\mathsf{TV}(p_T, q) \le \mathsf{TV}(P_T, P_T^{\infty,q_T}) + \mathsf{TV}(P_T^{\infty,q_T}, Q_T^{\leftarrow}) \le \mathsf{TV}(q_T, \gamma^d) + \mathsf{TV}(P_T^{\infty,q_T}, Q_T^{\leftarrow})\,.$$

Using the convergence of the OU process in KL divergence [see, e.g., BGL14, Theorem 5.2.1] and applying Theorem 17.5.2 for the second term,

$$\mathsf{TV}(p_T, q) \lesssim \sqrt{\mathsf{KL}(q \parallel \gamma^d)}\exp(-T) + (\varepsilon_{\text{score}} + L\sqrt{dh} + L\mathfrak{m}_2 h)\sqrt{T}\,,$$

which proves the result. □

17.5.4 Auxiliary lemmas

In this section, we prove some auxiliary lemmas which are used in the proof of Theorem 17.3.4.

Lemma 17.5.3 (Moment bounds for DDPM)**.** *Suppose that Assumptions 17.3.1 and 17.3.2 hold. Let $(\bar{X}_t)_{t\in[0,T]}$ denote the forward process (17.1).*

1. *(moment bound) For all $t \geq 0$,*

$$\mathbb{E}[\|\bar{X}_t\|^2] \leq d \vee \mathfrak{m}_2^2 \,.$$

2. *(score function bound) For all $t \geq 0$,*

$$\mathbb{E}[\|\nabla \ln q_t(\bar{X}_t)\|^2] \leq Ld \,.$$

Proof. 1. Along the OU process, we have $\bar{X}_t \overset{\mathsf{d}}{=} \exp(-t)\,\bar{X}_0 + \sqrt{1-\exp(-2t)}\,\xi$, where $\xi \sim \mathsf{normal}(0, I_d)$ is independent of $\bar{X}_0$. Hence,

$$\mathbb{E}[\|\bar{X}_t\|^2] = \exp(-2t)\,\mathbb{E}[\|X\|^2] + \{1-\exp(-2t)\}\,d \leq d \vee \mathfrak{m}_2^2 \,. \qquad \square$$

2. This follows from the L-smoothness of $\ln q_t$ [see, e.g., VW19, Lemma 9]. We give a short proof for the sake of completeness.

   If $\mathscr{L}_t f := \Delta f - \langle \nabla U_t, \nabla f\rangle$ is the generator associated with $q_t \propto \exp(-U_t)$,

$$0 = \mathbb{E}_{q_t} \mathscr{L}_t U_t = \mathbb{E}_{q_t} \Delta U_t - \mathbb{E}_{q_t}[\|\nabla U_t\|^2] \leq Ld - \mathbb{E}_{q_t}[\|\nabla U_t\|^2] \,.$$

Lemma 17.5.4 (Movement bound for DDPM)**.** *Suppose that Assumption 17.3.2 holds. Let $(\bar{X}_t)_{t\in[0,T]}$ denote the forward process (17.1). For $0 \leq s < t$ with $\delta := t - s$, if $\delta \leq 1$, then*

$$\mathbb{E}[\|\bar{X}_t - \bar{X}_s\|^2] \lesssim \delta^2 \mathfrak{m}_2^2 + \delta d \,.$$

Proof. We can write

$$\begin{aligned}
\mathbb{E}[\|\bar{X}_t - \bar{X}_s\|^2] &= \mathbb{E}\Big[\Big\|-\int_s^t \bar{X}_r \,\mathrm{d}r + \sqrt{2}\,(B_t - B_s)\Big\|^2\Big] \\
&\lesssim \delta \int_s^t \mathbb{E}[\|\bar{X}_r\|^2]\,\mathrm{d}r + \delta d \lesssim \delta^2\,(d + \mathfrak{m}_2^2) + \delta d \\
&\lesssim \delta^2 \mathfrak{m}_2^2 + \delta d \,,
\end{aligned}$$

where we used Lemma 17.5.3. $\square$

We omit the proofs of the two next lemmas as they are straightforward.

Lemma 17.5.5. *Consider $f_n, f : [0,T] \to \mathbb{R}^d$ s.t. there exists an increasing sequence $(T_n)_{n\in\mathbb{N}} \subseteq [0,T]$ satisfying the conditions*

- *$T_n \to T$ as $n \to \infty$,*
- *$f_n(t) = f(t)$ for every $t \le T_n$.*

Then, for every $\varepsilon > 0$, $f_n \to f$ uniformly over $[0, T-\varepsilon]$. In particular, it holds that $f_n(\cdot \wedge T - \varepsilon) \to f(\cdot \wedge T - \varepsilon)$ uniformly over $[0,T]$.

Lemma 17.5.6. *Consider $f : [0,T] \to \mathbb{R}^d$ continuous, and $f_\varepsilon : [0,T] \to \mathbb{R}^d$ s.t. $f_\varepsilon(t) = f(t \wedge (T-\varepsilon))$ for $\varepsilon > 0$. Then $f_\varepsilon \to f$ uniformly over $[0,T]$ as $\varepsilon \to 0$.*

■ 17.5.5 Proof of Corollary 17.3.7

Proof of Corollary 17.3.7. For $R_0 > 0$, let Π_{R_0} denote the projection onto $\mathsf{B}(0, R_0)$. We want to prove that $W_2((\Pi_{R_0})_\# p_{T-t}, q) \le \varepsilon$. We use the decomposition

$$W_2((\Pi_{R_0})_\# p_{T-t}, q) \le W_2((\Pi_{R_0})_\# p_{T-t}, (\Pi_{R_0})_\# q_t) + W_2((\Pi_{R_0})_\# q_t, q)\,.$$

For the first term, since $(\Pi_{R_0})_\# p_{T-t}$ and $(\Pi_{R_0})_\# q_t$ both have support contained in $\mathsf{B}(0,R_0)$, we can upper bound the Wasserstein distance by the total variation distance. Namely, [Rol22, Lemma 9] implies that

$$W_2((\Pi_{R_0})_\# p_{T-t}, (\Pi_{R_0})_\# q_t) \lesssim R_0 \sqrt{\mathsf{TV}((\Pi_{R_0})_\# p_{T-t}, (\Pi_{R_0})_\# q_t)} + R_0 \exp(-R_0)\,.$$

By the data-processing inequality,

$$\mathsf{TV}((\Pi_{R_0})_\# p_{T-t}, (\Pi_{R_0})_\# q_t) \le \mathsf{TV}(p_{T-t}, q_t) \le \varepsilon_{\mathrm{TV}}\,,$$

where $\varepsilon_{\mathrm{TV}}$ is from Corollary 17.3.5, yielding

$$W_2((\Pi_{R_0})_\# p_{T-t}, (\Pi_{R_0})_\# q_t) \lesssim R_0 \sqrt{\varepsilon_{\mathrm{TV}}} + R_0 \exp(-R_0)\,.$$

Next, we take $R_0 \ge R$ so that $(\Pi_{R_0})_\# q = q$. Since Π_{R_0} is 1-Lipschitz, we have

$$W_2((\Pi_{R_0})_\# q_t, q) = W_2((\Pi_{R_0})_\# q_t, (\Pi_{R_0})_\# q) \le W_2(q_t, q) \le \varepsilon_{W_2}\,,$$

where ε_{W_2} is from Corollary 17.3.5. Combining these bounds,

$$W_2((\Pi_{R_0})_\# p_{T-t}, q) \lesssim R_0 \sqrt{\varepsilon_{\mathrm{TV}}} + R_0 \exp(-R_0) + \varepsilon_{W_2}\,.$$

We now take $\varepsilon_{W_2} = \varepsilon/3$, $R_0 = \widetilde{\Theta}(R)$, and $\varepsilon_{\mathrm{TV}} = \widetilde{\Theta}(\varepsilon^2/R^2)$ to obtain the desired result. The iteration complexity follows from Corollary 17.3.5. □

17.5.6 Regularization

Lemma 17.5.7. *Suppose that* $\operatorname{supp} q \subseteq \mathsf{B}(0,R)$ *where* $R \geq 1$*, and let* q_t *denote the law of the OU process at time* t*, started at* q*. Let* $\varepsilon > 0$ *be such that* $\varepsilon \ll \sqrt{d}$ *and set* $t \asymp \varepsilon^2/(\sqrt{d}\,(R \vee \sqrt{d}))$*. Then,*

1. $W_2(q_t, q) \leq \varepsilon$.
2. q_t *satisfies*

$$\mathsf{KL}(q_t \,\|\, \gamma^d) \lesssim \frac{\sqrt{d}\,(R \vee \sqrt{d})^3}{\varepsilon^2}\,.$$

3. *For every* $t' \geq t$, $q_{t'}$ *satisfies Assumption 17.3.1 with*

$$L \lesssim \frac{dR^2\,(R \vee \sqrt{d})^2}{\varepsilon^4}\,.$$

Proof. 1. For the OU process (17.1), we can write

$$\bar{X}_t \coloneqq \exp(-t)\,\bar{X}_0 + \sqrt{1-\exp(-2t)}\,Z\,,$$

where $Z \sim \mathsf{normal}(0, I_d)$ is independent of $\bar{X}_0$. Hence, for $t \lesssim 1$,

$$\begin{aligned} W_2^2(q, q_t) &\leq \mathbb{E}\big[\big\|\big(1-\exp(-t)\big)\,\bar{X}_0 + \sqrt{1-\exp(-2t)}\,Z\big\|^2\big] \\ &= \big(1-\exp(-t)\big)^2\,\mathbb{E}[\|\bar{X}_0\|^2] + \big(1-\exp(-2t)\big)\,d \lesssim R^2t^2 + dt\,. \end{aligned}$$

We now take $t \lesssim \min\{\varepsilon/R, \varepsilon^2/d\}$ to ensure that $W_2^2(q, q_t) \leq \varepsilon^2$. Since $\varepsilon \ll \sqrt{d}$, it suffices to take $t \asymp \varepsilon^2/(\sqrt{d}\,(R \vee \sqrt{d}))$.

2. For this, we use the short-time regularization result in [OV01, Corollary 2], which implies that

$$\mathsf{KL}(q_t \,\|\, \gamma^d) \leq \frac{W_2^2(q, \gamma^d)}{4t} \lesssim \frac{W_2^2(q, \delta_0) + W_2^2(\gamma^d, \delta_0)}{t} \lesssim \frac{\sqrt{d}\,(R \vee \sqrt{d})^3}{\varepsilon^2}\,.$$

3. Using [MS22, Lemma 4], along the OU process,

$$\frac{1}{1-\exp(-2t)}\,I_d - \frac{\exp(-2t)\,R^2}{(1-\exp(-2t))^2}\,I_d \preccurlyeq -\nabla^2 \ln q_t(x) \preccurlyeq \frac{1}{1-\exp(-2t)}\,I_d\,. \tag{17.40}$$

With our choice of t, it implies

$$\begin{aligned}\|\nabla^2 \ln q_{t'}\|_{\mathrm{op}} &\lesssim \frac{1}{1-\exp(-2t')} \vee \frac{\exp(-2t')\, R^2}{(1-\exp(-2t'))^2} \lesssim \frac{1}{t} \vee \frac{R^2}{t^2} \\ &\lesssim \frac{dR^2\,(R \vee \sqrt{d})^2}{\varepsilon^4}\,.\end{aligned}$$

□

■ 17.6 Conclusion

In this work, we provided the first convergence guarantees for SGMs which hold under realistic assumptions (namely, L^2-accurate score estimation and arbitrarily non-log-concave data distributions) and which scale polynomially in the problem parameters. Our results take a step towards explaining the remarkable empirical success of SGMs, at least under the assumption that the score function is learned with small L^2 error.

The main limitation of this work is that we did not address the question of when the score function can be learned well. In general, studying the non-convex training dynamics of learning the score function via neural networks is challenging, but we believe that the resolution of this problem, even for simple learning tasks, would shed considerable light on SGMs. Together with the results in this paper, it would yield the first end-to-end guarantees for SGMs.

In another direction, and in light of the interpretation of our result as a reduction of the task of sampling to the task of score function estimation, we ask whether there are situations of interest in which it is easier to learn the score function (not necessarily via a neural network) than it is to (directly) sample.

Bibliography

[AB09] Hedy Attouch and Jérôme Bolte. "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features". In: *Math. Program.* 116.1-2, Ser. B (2009), pp. 5–16.

[AB15] David Alonso-Gutiérrez and Jesús Bastero. *Approaching the Kannan–Lovász–Simonovits and variance conjectures*. Vol. 2131. Lecture Notes in Mathematics. Springer, Cham, 2015, pp. x+148.

[AB21] Jason M. Altschuler and Enric Boix-Adserà. "Wasserstein barycenters can be computed in polynomial time in fixed dimension". In: *Journal of Machine Learning Research* 22.44 (2021), pp. 1–19.

[AB22] Jason M. Altschuler and Enric Boix-Adserà. "Wasserstein barycenters are NP-hard to compute". In: *SIAM J. Math. Data Sci.* 4.1 (2022), pp. 179–203.

[Aba+16a] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. "Deep learning with differential privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS '16. Vienna, Austria: Association for Computing Machinery, 2016, pp. 308–318.

[Aba+16b] Martin Abadi et al. "TensorFlow: a system for large-scale machine learning". In: *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*. OSDI'16. Savannah, GA, USA: USENIX Association, 2016, pp. 265–283.

[ABS21] Luigi Ambrosio, Elia Brué, and Daniele Semola. *Lectures on optimal transport*. Vol. 130. Unitext. La Matematica per il 3+2. Springer, Cham, [2021] ©2021, pp. ix+250.

[ABY13] Marc Arnaudon, Frédéric Barbaresco, and Le Yang. "Riemannian medians and means with applications to radar signal processing". In: *IEEE Journal of Selected Topics in Signal Processing* 7.4 (2013), pp. 595–604.

[AC11] Martial Agueh and Guillaume Carlier. "Barycenters in the Wasserstein space". In: *SIAM J. Math. Anal.* 43.2 (2011), pp. 904–924.

[AC17] Martial Agueh and Guillaume Carlier. "Vers un théorème de la limite centrale dans l'espace de Wasserstein?" In: *C. R. Math. Acad. Sci. Paris* 355.7 (2017), pp. 812–818.

[AC21] Kwangjun Ahn and Sinho Chewi. "Efficient constrained sampling via the mirror-Langevin algorithm". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, K. Nguyen, P. S. Liang, J. W. Vaughan, and Y. Dauphin. Vol. 34. Curran Associates, Inc., 2021, pp. 28405–28418.

[AC23] Jason M. Altschuler and Sinho Chewi. "Faster high-accuracy log-concave sampling via algorithmic warm starts". In: *arXiv preprint 2302.10249* (2023).

[ADC20] Shahab Asoodeh, Mario Diaz, and Flavio P. Calmon. "Privacy amplification of iterative algorithms via contraction coefficients". In: *2020 IEEE International Symposium on Information Theory (ISIT)*. Los Angeles, CA, USA: IEEE Press, 2020, pp. 896–901.

[Afs11] Bijan Afsari. "Riemannian L^p center of mass: existence, uniqueness, and convexity". In: *Proc. Amer. Math. Soc.* 139.2 (2011), pp. 655–673.

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Second. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2008, pp. x+334.

[AH09] Ienkaran Arasaratnam and Simon Haykin. "Cubature Kalman filters". In: *IEEE Trans. Automat. Control* 54.6 (2009), pp. 1254–1269.

[AHR08] Jacob Abernethy, Elad E. Hazan, and Alexander Rakhlin. "Competing in the dark: an efficient algorithm for bandit linear optimization". English (US). In: *21st Annual Conference on Learning Theory, COLT 2008*. 2008, pp. 263–273.

[Alb+21] Dallas Albritton, Scott Armstrong, Jean-Christophe Mourrat, and Matthew Novack. "Variational methods for the kinetic Fokker–Planck equation". In: *arXiv preprint 1902.04037* (2021).

[ALP20] Adil Ahidar-Coutrix, Thibaut Le Gouic, and Quentin Paris. "Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics". In: *Probab. Theory Related Fields* 177.1-2 (2020), pp. 323–368.

[Alt+21] Jason M. Altschuler, Sinho Chewi, Patrik R. Gerber, and Austin J. Stromme. "Averaging on the Bures–Wasserstein manifold: dimension-free convergence of gradient descent". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, K. Nguyen, P. S. Liang, J. W. Vaughan, and Y. Dauphin. Vol. 34. Curran Associates, Inc., 2021, pp. 22132–22145.

[Álv+16] Pedro C. Álvarez-Esteban, Eustasio del Barrio, Juan A. Cuesta-Albertos, and Carlos Matrán. "A fixed-point approach to barycenters in Wasserstein space". In: *J. Math. Anal. Appl.* 441.2 (2016), pp. 744–762.

[AMR05] Claude Auderset, Christian Mazza, and Ernst A. Ruh. "Angular Gaussian and Cauchy estimation". In: *J. Multivariate Anal.* 93.1 (2005), pp. 180–197.

[AN00] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*. Vol. 191. Translations of Mathematical Monographs. Translated from the 1993 Japanese original by Daishi Harada. American Mathematical Society, Providence, RI; Oxford University Press, Oxford, 2000, pp. x+206.

[ANR17] Jason M. Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 1964–1974.

[AR20] Pierre Alquier and James Ridgway. "Concentration of tempered posteriors and of their variational approximations". In: *Ann. Statist.* 48.3 (2020), pp. 1475–1497.

[ARC16] Pierre Alquier, James Ridgway, and Nicolas Chopin. "On the properties of variational approximations of Gibbs posteriors". In: *J. Mach. Learn. Res.* 17 (2016), Paper No. 239, 41.

[AT22a] Jason M. Altschuler and Kunal Talwar. "Privacy of noisy stochastic gradient descent: more iterations without more privacy loss". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho. 2022.

[AT22b] Jason M. Altschuler and Kunal Talwar. "Resolving the mixing time of the Langevin algorithm to its stationary distribution for log-concave sampling". In: *arXiv preprint 2210.08448* (2022).

[Ay+17] Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information geometry*. Vol. 64. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]. Springer, Cham, 2017, pp. xi+407.

[Bac+22] Julio Backhoff-Veraguas, Joaquin Fontbona, Gonzalo Rios, and Felipe Tobar. "Bayesian learning with Wasserstein barycenters". In: *ESAIM Probab. Stat.* 26 (2022), pp. 436–472.

[Bac+92] François Louis Baccelli, Guy Cohen, Geert Jan Olsder, and Jean-Pierre Quadrat. *Synchronization and linearity*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. An algebra for discrete event systems. John Wiley & Sons, Ltd., Chichester, 1992, pp. xx+489.

[Bač14a] Miroslav Bačák. "Computing medians and means in Hadamard spaces". In: *SIAM J. Optim.* 24.3 (2014), pp. 1542–1566.

[Bač14b] Miroslav Bačák. *Convex analysis and optimization in Hadamard spaces*. Vol. 22. De Gruyter Series in Nonlinear Analysis and Applications. De Gruyter, Berlin, 2014, pp. viii+185.

[Bal+19] Borja Balle, Gilles Barthe, Marco Gaboardi, and Joseph Geumlek. "Privacy amplification by mixing and diffusion mechanisms". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[Bal+22] Krishna Balasubramanian, Sinho Chewi, Murat A. Erdogdu, Adil Salim, and Matthew Zhang. "Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo". In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 2896–2923.

[Bar+18] Jean-Baptiste Bardet, Nathaël Gozlan, Florent Malrieu, and Pierre-André Zitt. "Functional inequalities for Gaussian convolutions of compactly supported measures: explicit bounds and dimension dependence". In: *Bernoulli* 24.1 (2018), pp. 333–353.

[Bar01] Franck Barthe. "Levels of concentration between exponential and Gaussian". In: *Ann. Fac. Sci. Toulouse Math. (6)* 10.3 (2001), pp. 393–404.

[Bau17] Fabrice Baudoin. "Bakry–Émery meet Villani". In: *Journal of Functional Analysis* 273.7 (2017), pp. 2275–2291.

[BB18] Adrien Blanchet and Jérôme Bolte. "A family of functional inequalities: Łojasiewicz inequalities and displacement convex functions". In: *J. Funct. Anal.* 275.7 (2018), pp. 1650–1673.

[BB97] David Barber and Christopher Bishop. "Ensemble learning for multi-layer networks". In: *Advances in Neural Information Processing Systems*. Ed. by M. Jordan, M. Kearns, and S. Solla. Vol. 10. MIT Press, 1997.

[BB99] Jean-David Benamou and Yann Brenier. "A numerical method for the optimal time-continuous mass transport problem and related problems". In: *Monge Ampère equation: applications to geometry and optimization (Deerfield Beach, FL, 1997)*. Vol. 226. Contemp. Math. Amer. Math. Soc., Providence, RI, 1999, pp. 1–11.

[BBI01] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*. Vol. 33. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2001, pp. xiv+415.

[BBT17] Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. "A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications". In: *Math. Oper. Res.* 42.2 (2017), pp. 330–348.

[BC12] Sébastien Bubeck and Nicolò Cesa-Bianchi. "Regret analysis of stochastic and nonstochastic multi-armed bandit problems". In: *Foundations and Trends® in Machine Learning* 5.1 (2012), pp. 1–122.

[BC13] Franck Barthe and Dario Cordero-Erausquin. "Invariances in variance estimates". In: *Proc. Lond. Math. Soc. (3)* 106.1 (2013), pp. 33–64.

[BCG08] Dominique Bakry, Patrick Cattiaux, and Arnaud Guillin. "Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré". In: *J. Funct. Anal.* 254.3 (2008), pp. 727–759.

[BCP19] Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. "Penalization of barycenters in the Wasserstein space". In: *SIAM J. Math. Anal.* 51.3 (2019), pp. 2261–2285.

[BCR06] Franck Barthe, Patrick Cattiaux, and Cyril Roberto. "Interpolated inequalities between exponential and Gaussian, Orlicz hypercontractivity and isoperimetry". In: *Rev. Mat. Iberoam.* 22.3 (2006), pp. 993–1067.

[BCR07] Franck Barthe, Patrick Cattiaux, and Cyril Roberto. "Isoperimetry between exponential and Gaussian". In: *Electron. J. Probab.* 12 (2007), no. 44, 1212–1237.

[BD01] Louis J. Billera and Persi Diaconis. "A geometric interpretation of the Metropolis–Hastings algorithm". In: *Statist. Sci.* 16.4 (2001), pp. 335–339.

[BE19] Sébastien Bubeck and Ronen Eldan. "The entropic barrier: exponential families, log-concave geometry, and self-concordance". In: *Math. Oper. Res.* 44.1 (2019), pp. 264–276.

[BEL18] Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. "Sampling from a log-concave distribution with projected Langevin Monte Carlo". In: *Discrete Comput. Geom.* 59.4 (2018), pp. 757–783.

[Ben+15] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. "Iterative Bregman projections for regularized transportation problems". In: *SIAM J. Sci. Comput.* 37.2 (2015), A1111–A1138.

[Ber18] Espen Bernton. "Langevin Monte Carlo and JKO splitting". In: *Proceedings of the 31st Conference on Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 1777–1798.

[Bes+95] Julian Besag, Peter Green, David Higdon, and Kerrie Mengersen. "Bayesian computation and stochastic systems". In: *Statistical Science* 10.1 (1995), pp. 3–66.

[BFR19] Joris Bierkens, Paul Fearnhead, and Gareth Roberts. "The zig-zag process and super-efficient sampling for Bayesian analysis of big data". In: *The Annals of Statistics* 47.3 (2019), pp. 1288–1320.

[BG12] Rajendra Bhatia and Priyanka Grover. "Norm inequalities related to the matrix geometric mean". In: *Linear Algebra Appl.* 437.2 (2012), pp. 726–733.

[BGG12] François Bolley, Ivan Gentil, and Arnaud Guillin. "Convergence to equilibrium in Wasserstein distance for Fokker–Planck equations". In: *J. Funct. Anal.* 263.8 (2012), pp. 2430–2457.

[BGG18] François Bolley, Ivan Gentil, and Arnaud Guillin. "Dimensional improvements of the logarithmic Sobolev, Talagrand and Brascamp–Lieb inequalities". In: *Ann. Probab.* 46.1 (2018), pp. 261–301.

[BGK05] Anton Bovier, Véronique Gayrard, and Markus Klein. "Metastability in reversible diffusion processes. II. Precise asymptotics for small eigenvalues". In: *J. Eur. Math. Soc. (JEMS)* 7.1 (2005), pp. 69–99.

[BGL14] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*. Vol. 348. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Cham, 2014, pp. xx+552.

[BGN22] Espen Bernton, Promit Ghosal, and Marcel Nutz. "Entropic optimal transport: geometry and large deviations". In: *Duke Math. J.* 171.16 (2022), pp. 3363–3400.

[BH97] Serguei G. Bobkov and Christian Houdré. "Some connections between isoperimetric and Sobolev-type inequalities". In: *Mem. Amer. Math. Soc.* 129.616 (1997), pp. viii+111.

[Bha07] Rajendra Bhatia. *Positive definite matrices*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2007, pp. x+254.

[BI13] Dario A. Bini and Bruno Iannazzo. "Computing the Karcher mean of symmetric positive definite matrices". In: *Linear Algebra Appl.* 438.4 (2013), pp. 1700–1710.

[Big+18] Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. "Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line". In: *Electron. J. Stat.* 12.2 (2018), pp. 2253–2289.

[BJL19] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. "On the Bures–Wasserstein distance between positive definite matrices". In: *Expo. Math.* 37.2 (2019), pp. 165–191.

[BKM17] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: a review for statisticians". In: *J. Amer. Statist. Assoc.* 112.518 (2017), pp. 859–877.

[BL00] Sergey G. Bobkov and Michel Ledoux. "From Brunn–Minkowski to Brascamp–Lieb and to logarithmic Sobolev inequalities". In: *Geom. Funct. Anal.* 10.5 (2000), pp. 1028–1052.

[BL06] Jonathan M. Borwein and Adrian S. Lewis. *Convex analysis and nonlinear optimization*. Second. Vol. 3. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Theory and examples. Springer, New York, 2006, pp. xii+310.

[BL76] Herm J. Brascamp and Elliott H. Lieb. "On extensions of the Brunn–Minkowski and Prékopa–Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation". In: *J. Functional Analysis* 22.4 (1976), pp. 366–389.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013, pp. x+481.

[BM20] Sébastien Bubeck and Dan Mikulincer. "How to trap a gradient flow". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 940–960.

[BM22] Nawaf Bou-Rabee and Milo Marsden. "Unadjusted Hamiltonian MC-MC with stratified Monte Carlo time integration". In: *arXiv preprint 2211.11003* (2022).

[BMR22] Adam Block, Youssef Mroueh, and Alexander Rakhlin. "Generative modeling with denoising auto-encoders and Langevin sampling". In: *arXiv preprint 2002.00107* (2022).

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet allocation". In: *J. Mach. Learn. Res.* 3 (Mar. 2003), pp. 993–1022.

[Bob03] Sergey G. Bobkov. "Spectral gap and concentration for some spherically symmetric probability measures". In: *Geometric aspects of functional analysis*. Vol. 1807. Lecture Notes in Math. Springer, Berlin, 2003, pp. 37–43.

[Bob99] Sergey G. Bobkov. "Isoperimetric and analytic inequalities for log-concave probability measures". In: *Ann. Probab.* 27.4 (1999), pp. 1903–1921.

[Bon+15] Giovanni A. Bonaschi, José A. Carrillo, Marco Di Francesco, and Mark A. Peletier. "Equivalence of gradient flows and entropy solutions for singular nonlocal interaction equations in 1D". In: *ESAIM Control Optim. Calc. Var.* 21.2 (2015), pp. 414–441.

[Bon13] Silvère Bonnabel. "Stochastic gradient descent on Riemannian manifolds". In: *IEEE Trans. Automat. Control* 58.9 (2013), pp. 2217–2229.

[Bor22] Steffen Borgwardt. "An LP-based, strongly-polynomial 2-approximation algorithm for sparse Wasserstein barycenters". In: *Operational Research* 22.2 (Apr. 2022), pp. 1511–1551.

[Bou+05] Stéphane Boucheron, Olivier Bousquet, Gábor Lugosi, and Pascal Massart. "Moment inequalities for functions of independent random variables". In: *Ann. Probab.* 33.2 (2005), pp. 514–560.

[Bou23] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, Cambridge, 2023, pp. xviii+338.

[Bov+02] Anton Bovier, Michael Eckhoff, Véronique Gayrard, and Markus Klein. "Metastability and low lying spectra in reversible Markov chains". In: *Comm. Math. Phys.* 228.2 (2002), pp. 219–255.

[Bov+04] Anton Bovier, Michael Eckhoff, Véronique Gayrard, and Markus Klein. "Metastability in reversible diffusion processes. I. Sharp asymptotics for capacities and exit times". In: *J. Eur. Math. Soc. (JEMS)* 6.4 (2004), pp. 399–424.

[BPC16] Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. "Wasserstein barycentric coordinates: histogram regression using optimal transport". In: *ACM Trans. Graph.* 35.4 (July 2016).

[BPP98] Itai Benjamini, Robin Pemantle, and Yuval Peres. "Unpredictable paths and percolation". In: *Ann. Probab.* 26.3 (1998), pp. 1198–1211.

[BR03] Franck Barthe and Cyril Roberto. "Sobolev inequalities for probability measures on the real line". In: vol. 159. 3. Dedicated to Professor Aleksander Pełczyński on the occasion of his 70th birthday (Polish). 2003, pp. 481–497.

[Bra+14] Silouanos Brazitikos, Apostolos Giannopoulos, Petros Valettas, and Beatrice-Helen Vritsiou. *Geometry of isotropic convex bodies*. Vol. 196. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2014, pp. xx+594.

[Brè67] Lev M. Brègman. "A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming". In: *Ž. Vyčisl. Mat i Mat. Fiz.* 7 (1967), pp. 620–631.

[Bro+17] Nicolas Brosse, Alain Durmus, Éric Moulines, and Marcelo Pereyra. "Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo". In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by Satyen Kale and Ohad Shamir. Vol. 65. Proceedings of Machine Learning Research. PMLR, July 2017, pp. 319–342.

[Bro+19] Nicolas Brosse, Alain Durmus, Éric Moulines, and Sotirios Sabanis. "The tamed unadjusted Langevin algorithm". In: *Stochastic Process. Appl.* 129.10 (2019), pp. 3638–3663.

[Bru+21] Joan Bruna, Oded Regev, Min Jae Song, and Yi Tang. "Continuous LWE". In: *STOC '21—Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, [2021] ©2021, pp. 694–707.

[Bub15] Sébastien Bubeck. "Convex optimization: algorithms and complexity". In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357.

[Bur69] Donald Bures. "An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w^*-algebras". In: *Trans. Amer. Math. Soc.* 135 (1969), pp. 199–212.

[Bur73] Donald L. Burkholder. "Distribution function inequalities for martingales". In: *Ann. Probability* 1 (1973), pp. 19–42.

[BV05] François Bolley and Cédric Villani. "Weighted Csiszár–Kullback–Pinsker inequalities and applications to transportation inequalities". In: *Ann. Fac. Sci. Toulouse Math. (6)* 14.3 (2005), pp. 331–352.

[BZ99] Lorenzo Bertini and Bogusław Zegarlinski. "Coercive inequalities for Gibbs measures". In: *J. Funct. Anal.* 162.2 (1999), pp. 257–286.

[CAD21] Samuel Cohen, Michael Arbel, and Marc P. Deisenroth. "Estimating barycenters of measures in high dimensions". In: *arXiv preprint 2007.07105* (2021).

[Caf00] Luis A. Caffarelli. "Monotonicity properties of optimal transportation and the FKG and related inequalities". In: *Comm. Math. Phys.* 214.3 (2000), pp. 547–563.

[Car+11] José A. Carrillo, Marco Di Francesco, Alessio Figalli, Thomas Laurent, and Dejan Slepčev. "Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations". In: *Duke Math. J.* 156.2 (2011), pp. 229–271.

[Car+12] José A. Carrillo, Marco Di Francesco, Alessio Figalli, Thomas Laurent, and Dejan Slepčev. "Confinement in nonlocal interaction equations". In: *Nonlinear Anal.* 75.2 (2012), pp. 550–558.

[Car+20] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. "Lower bounds for finding stationary points I". In: *Math. Program.* 184.1-2, Ser. A (2020), pp. 71–120.

[Car+21] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. "Lower bounds for finding stationary points II: first-order methods". In: *Math. Program.* 185.1-2, Ser. A (2021), pp. 315–355.

[Car92] Manfredo Perdigão do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Translated from the second Portuguese edition by Francis Flaherty. Birkhäuser Boston, Inc., Boston, MA, 1992, pp. xiv+300.

[Cat+22] Patrick Cattiaux, Giovanni Conforti, Ivan Gentil, and Christian Léonard. "Time reversal of diffusion processes under a finite entropy condition". In: *arXiv preprint 2104.07708* (2022).

[CB13] Edward Challis and David Barber. "Gaussian Kullback–Leibler approximate inference". In: *J. Mach. Learn. Res.* 14 (2013), pp. 2239–2286.

[CB16] Katy Craig and Andrea L. Bertozzi. "A blob method for the aggregation equation". In: *Math. Comp.* 85.300 (2016), pp. 1681–1717.

[CB18] Xiang Cheng and Peter Bartlett. "Convergence of Langevin MCMC in KL-divergence". In: *Proceedings of Algorithmic Learning Theory*. Ed. by Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan. Vol. 83. Proceedings of Machine Learning Research. PMLR, 2018, pp. 186–211.

[CBB17] Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. "Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review". In: *Brain-Computer Interfaces* 4.3 (2017), pp. 155–174.

[CBL22] Niladri S. Chatterji, Peter L. Bartlett, and Philip M. Long. "Oracle lower bounds for stochastic gradient sampling algorithms". In: *Bernoulli* 28.2 (2022), pp. 1074–1092.

[CBS23] Sinho Chewi, Sébastien Bubeck, and Adil Salim. "On the complexity of finding stationary points of smooth functions in one dimension". In: *Proceedings of the 34th International Conference on Algorithmic Learning Theory*. Ed. by Shipra Agrawal and Francesco Orabona. Vol. 201. Proceedings of Machine Learning Research. PMLR, Feb. 2023, pp. 358–374.

[CCN21] Hong-Bin Chen, Sinho Chewi, and Jonathan Niles-Weed. "Dimension-free log-Sobolev inequalities for mixture distributions". In: *Journal of Functional Analysis* 281.11 (2021), p. 109236.

[CCP19] José A. Carrillo, Katy Craig, and Francesco S. Patacchini. "A blob method for diffusion". In: *Calc. Var. Partial Differential Equations* 58.2 (2019), Paper No. 53, 53.

[CCS18] Sebastian Claici, Edward Chien, and Justin M. Solomon. "Stochastic Wasserstein barycenters". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 999–1008.

[CD14] Marco Cuturi and Arnaud Doucet. "Fast computation of Wasserstein barycenters". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, June 2014, pp. 685–693.

[CE10] Guillaume Carlier and Ivar Ekeland. "Matching for teams". In: *Econom. Theory* 42.2 (2010), pp. 397–418.

[CE22] Yuansi Chen and Ronen Eldan. "Localization schemes: a framework for proving mixing bounds for Markov chains". In: *arXiv preprint 2203.04163* (2022).

[CEK21] Guillaume Carlier, Katharina Eichinger, and Alexey Kroshnin. "Entropic-Wasserstein barycenters: PDE characterization, regularity, and CLT". In: *SIAM J. Math. Anal.* 53.5 (2021), pp. 5880–5914.

[Cel+12] Gilles Celeux, Mohammed El Anbari, Jean-Michel Marin, and Christian P. Robert. "Regularization in regression: comparing Bayesian and frequentist methods in a poorly informative situation". In: *Bayesian Anal.* 7.2 (2012), pp. 477–502.

[CFJ17] Maria Colombo, Alessio Figalli, and Yash Jhaveri. "Lipschitz changes of variables between perturbations of log-concave measures". In: *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* 17.4 (2017), pp. 1491–1519.

[CFM04] Dario Cordero-Erausquin, Matthieu Fradelizi, and Bernard Maurey. "The (B) conjecture for the Gaussian measure of dilates of symmetric convex sets and related problems". In: *J. Funct. Anal.* 214.2 (2004), pp. 410–427.

[CG03] Eric A. Carlen and Wilfrid Gangbo. "Constrained steepest descent in the 2-Wasserstein metric". In: *Ann. of Math. (2)* 157.3 (2003), pp. 807–846.

[CG09] Patrick Cattiaux and Arnaud Guillin. "Trends to equilibrium in total variation distance". In: *Ann. Inst. Henri Poincaré Probab. Stat.* 45.1 (2009), pp. 117–145.

[CGG07] Patrick Cattiaux, Ivan Gentil, and Arnaud Guillin. "Weak logarithmic Sobolev inequalities and entropic convergence". In: *Probab. Theory Related Fields* 139.3-4 (2007), pp. 563–603.

[CGT00] Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint. *Trust-region methods*. MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA, 2000, pp. xx+959.

[CGT19] Yongxin Chen, Tryphon T. Georgiou, and Allen Tannenbaum. "Optimal transport for Gaussian mixture models". In: *IEEE Access* 7 (2019), pp. 6269–6278.

[Cha+20] Niladri Chatterji, Jelena Diakonikolas, Michael I. Jordan, and Peter Bartlett. "Langevin Monte Carlo without smoothness". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, Aug. 2020, pp. 1716–1726.

[Cha04] Djalil Chafai. "Entropies, convexity, and functional inequalities: on Φ-entropies and Φ-Sobolev inequalities". In: *J. Math. Kyoto Univ.* 44.2 (2004), pp. 325–363.

[Che+18a] Yuansi Chen, Raaz Dwivedi, Martin J. Wainwright, and Bin Yu. "Fast MCMC sampling algorithms on polytopes". In: *J. Mach. Learn. Res.* 19 (2018), Paper No. 55, 86.

[Che+18b] Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. "Underdamped Langevin MCMC: a non-asymptotic analysis". In: *Proceedings of the 31st Conference on Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 300–323.

[Che+20a] Yuansi Chen, Raaz Dwivedi, Martin J. Wainwright, and Bin Yu. "Fast mixing of Metropolized Hamiltonian Monte Carlo: benefits of multi-step gradients". In: *J. Mach. Learn. Res.* 21 (2020), Paper No. 92, 71.

[Che+20b] Xiang Cheng, Niladri S. Chatterji, Yasin Abbasi-Yadkori, Peter L. Bartlett, and Michael I. Jordan. "Sharp convergence rates for Langevin dynamics in the nonconvex setting". In: *arXiv preprint 1805.01648* (2020).

[Che+20c] Xiang Cheng, Dong Yin, Peter Bartlett, and Michael Jordan. "Stochastic gradient and Langevin processes". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 1810–1819.

[Che+20d] Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, and Philippe Rigollet. "SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 2098–2109.

[Che+20e] Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet, and Austin J. Stromme. "Exponential ergodicity of mirror-Langevin diffusions". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 19573–19585.

[Che+20f] Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J. Stromme. "Gradient descent algorithms for Bures–Wasserstein barycenters". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 1276–1304.

[Che+21a] Sinho Chewi, Murat A. Erdogdu, Mufan (Bill) Li, Ruoqi Shen, and Matthew Zhang. "Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev". In: *arXiv preprint 2112.12662* (2021).

[Che+21b] Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. "Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, Aug. 2021, pp. 1260–1300.

[Che+22a] Sitan Chen, Aravind Gollakota, Adam Klivans, and Raghu Meka. "Hardness of noise-free learning for two-hidden-layer neural networks". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 10709–10724.

[Che+22b] Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. "Improved analysis for a proximal algorithm for sampling". In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 2984–3014.

[Che+22c] Sinho Chewi, Murat A. Erdogdu, Mufan (Bill) Li, Ruoqi Shen, and Matthew Zhang. "Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev". In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 1–2.

[Che+22d] Sinho Chewi, Patrik R. Gerber, Chen Lu, Thibaut Le Gouic, and Philippe Rigollet. "The query complexity of sampling from strongly log-concave distributions in one dimension". In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 2041–2059.

[Che+22e] Sinho Chewi, Patrik R. Gerber, Chen Lu, Thibaut Le Gouic, and Philippe Rigollet. "Rejection sampling from shape-constrained distributions in sublinear time". In: *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, Mar. 2022, pp. 2249–2265.

[Che+23a] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. "Sampling is as easy as learning the score: theory for

diffusion models with minimal data assumptions". In: *The Eleventh International Conference on Learning Representations*. 2023.

[Che+23b] Sinho Chewi, Jaume de Dios Pont, Jerry Li, Chen Lu, and Shyam Narayanan. "Query lower bounds for log-concave sampling". In: *arXiv preprint 2304.02599* (2023).

[Che+23c] Sinho Chewi, Patrik R. Gerber, Holden Lee, and Chen Lu. "Fisher information lower bounds for sampling". In: *Proceedings of the 34th International Conference on Algorithmic Learning Theory*. Ed. by Shipra Agrawal and Francesco Orabona. Vol. 201. Proceedings of Machine Learning Research. PMLR, Feb. 2023, pp. 375–410.

[Che21a] Yuansi Chen. "An almost constant lower bound of the isoperimetric coefficient in the KLS conjecture". In: *Geom. Funct. Anal.* 31.1 (2021), pp. 34–61.

[Che21b] Sinho Chewi. "The entropic barrier is n-self-concordant". In: *arXiv preprint 2112.10947* (2021).

[Che23] Sinho Chewi. *Log-concave sampling*. Available online at https://chewisinho.github.io/. Forthcoming, 2023.

[Chi+18] Lénaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. "An interpolating distance between optimal transport and Fisher–Rao metrics". In: *Found. Comput. Math.* 18.1 (2018), pp. 1–44.

[CK85] Nai N. Chan and Man Kam Kwong. "Hermitian matrix inequalities and a conjecture". In: *Amer. Math. Monthly* 92.8 (1985), pp. 533–541.

[CL89] Mu Fa Chen and Shao Fu Li. "Coupling methods for multidimensional diffusion processes". In: *Ann. Probab.* 17.1 (1989), pp. 151–177.

[CLL19] Yu Cao, Jianfeng Lu, and Yulong Lu. "Exponential decay of Rényi divergence under Fokker–Planck equations". In: *J. Stat. Phys.* 176.5 (2019), pp. 1172–1184.

[CLL22] Sitan Chen, Jerry Li, and Yuanzhi Li. "Learning (very) simple generative models is hard". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 35143–35155.

[CLL23] Hongrui Chen, Holden Lee, and Jianfeng Lu. "Improved analysis of score-based generative modeling: user-friendly bounds under minimal smoothness assumptions". In: *arXiv preprint 2211.01916* (2023).

[CLW20] Yu Cao, Jianfeng Lu, and Lihan Wang. "On explicit L^2-convergence rate estimate for underdamped Langevin dynamics". In: *arXiv preprint 1908.04746* (2020).

[CLW21] Yu Cao, Jianfeng Lu, and Lihan Wang. "Complexity of randomized algorithms for underdamped Langevin dynamics". In: *Commun. Math. Sci.* 19.7 (2021), pp. 1827–1853.

[CM10] Djalil Chafai and Florent Malrieu. "On fine properties of mixtures with respect to concentration of measure and Sobolev type inequalities". In: *Ann. Inst. Henri Poincaré Probab. Stat.* 46.1 (2010), pp. 72–96.

[CMT96] Juan A. Cuesta-Albertos, Carlos Matrán-Bea, and Araceli Tuero-Diaz. "On lower bounds for the L^2-Wasserstein metric in a Hilbert space". In: *J. Theoret. Probab.* 9.2 (1996), pp. 263–283.

[COO15] Guillaume Carlier, Adam Oberman, and Edouard Oudet. "Numerical methods for matching for teams and Wasserstein barycenters". In: *ESAIM Math. Model. Numer. Anal.* 49.6 (2015), pp. 1621–1642.

[Cor+09] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to algorithms*. Third. MIT Press, Cambridge, MA, 2009, pp. xx+1292.

[Cor17] Dario Cordero-Erausquin. "Transport inequalities for log-concave measures, quantitative forms, and applications". In: *Canad. J. Math.* 69.3 (2017), pp. 481–501.

[Cou20] Thomas A. Courtade. "Bounds on the Poincaré constant for convolution measures". In: *Ann. Inst. Henri Poincaré Probab. Stat.* 56.1 (2020), pp. 566–579.

[CP22] Sinho Chewi and Aram-Alexandre Pooladian. "An entropic generalization of Caffarelli's contraction theorem via covariance inequalities". In: *arXiv preprint 2203.04954* (2022).

[CPR09] Emanuele Caglioti, Mario Pulvirenti, and Frédéric Rousset. "On a constrained 2-D Navier–Stokes equation". In: *Comm. Math. Phys.* 290.2 (2009), pp. 651–677.

[Cra+23] Katy Craig, Karthik Elamvazhuthi, Matt Haberland, and Olga Turanova. "A blob method for inhomogeneous diffusion with applications to multi-agent control and sampling". In: *arXiv preprint 2202.12927* (2023).

[Csi75] Imre Csiszár. "I-divergence geometry of probability distributions and minimization problems". In: *Ann. Probability* 3 (1975), pp. 146–158.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Second. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006, pp. xxiv+748.

[CT93] Gong Chen and Marc Teboulle. "Convergence analysis of a proximal-like minimization algorithm using Bregman functions". In: *SIAM J. Optim.* 3.3 (1993), pp. 538–543.

[Cut13] Marco Cuturi. "Sinkhorn distances: lightspeed computation of optimal transport". In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger. Vol. 26. Curran Associates, Inc., 2013.

[CV19] Zongchen Chen and Santosh S. Vempala. "Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions". In: *Approximation, randomization, and combinatorial optimization. Algorithms and techniques*. Vol. 145. LIPIcs. Leibniz Int. Proc. Inform. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2019, Art. No. 64, 12.

[CYS21] Rishav Chourasia, Jiayuan Ye, and Reza Shokri. "Differential privacy dynamics of Langevin diffusion and noisy gradient descent". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 14771–14781.

[Dal17a] Arnak S. Dalalyan. "Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent". In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by Satyen Kale and Ohad Shamir. Vol. 65. Proceedings of Machine Learning Research. PMLR, July 2017, pp. 678–689.

[Dal17b] Arnak S. Dalalyan. "Theoretical guarantees for approximate sampling from smooth and log-concave densities". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 79.3 (2017), pp. 651–676.

[Dav76] Burgess Davis. "On the L^p norms of stochastic integrals and other martingales". In: *Duke Math. J.* 43.4 (1976), pp. 697–704.

[DD20] Julie Delon and Agnès Desolneux. "A Wasserstein-type distance in the space of Gaussian mixture models". In: *SIAM J. Imaging Sci.* 13.2 (2020), pp. 936–970.

[DD21] Kamélia Daudel and Randal Douc. "Mixture weights optimisation for alpha-divergence variational inference". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 4397–4408.

[DDP21] Kamélia Daudel, Randal Douc, and François Portier. "Infinite-dimensional gradient-based descent for alpha-divergence minimisation". In: *Ann. Statist.* 49.4 (2021), pp. 2250–2270.

[De +21] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. "Diffusion Schrödinger bridge with applications to score-based generative modeling". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 17695–17709.

[De 22] Valentin De Bortoli. "Convergence of denoising diffusion models under the manifold hypothesis". In: *Transactions on Machine Learning Research* (2022).

[DGN14] Olivier Devolder, François Glineur, and Yurii Nesterov. "First-order methods of smooth convex optimization with inexact oracle". In: *Math. Program.* 146.1-2, Ser. A (2014), pp. 37–75.

[Dia+23] Michael Diao, Krishnakumar Balasubramanian, Sinho Chewi, and Adil Salim. "Forward-backward Gaussian variational inference via JKO in the Bures–Wasserstein space". In: *arXiv preprint 2304.05398* (2023).

[Din15] Ying Ding. "A note on quadratic transportation and divergence inequality". In: *Statist. Probab. Lett.* 100 (2015), pp. 115–123.

[DK19] Arnak S. Dalalyan and Avetik Karagulyan. "User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient". In: *Stochastic Process. Appl.* 129.12 (2019), pp. 5278–5311.

[DKR22] Arnak S. Dalalyan, Avetik Karagulyan, and Lionel Riou-Durand. "Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets". In: *Journal of Machine Learning Research* 23.235 (2022), pp. 1–38.

[DKS17] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. "Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures (extended abstract)". In: *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*. IEEE Computer Soc., Los Alamitos, CA, 2017, pp. 73–84.

[DL21] Zhiyan Ding and Qin Li. "Langevin Monte Carlo: random coordinate descent and variance reduction". In: *J. Mach. Learn. Res.* 22 (2021), Paper No. 205, 51.

[DM17] Alain Durmus and Éric Moulines. "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm". In: *Ann. Appl. Probab.* 27.3 (2017), pp. 1551–1587.

[DM19] Alain Durmus and Éric Moulines. "High-dimensional Bayesian inference via the unadjusted Langevin algorithm". In: *Bernoulli* 25.4A (2019), pp. 2854–2882.

[DMM19] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. "Analysis of Langevin Monte Carlo via convex optimization". In: *J. Mach. Learn. Res.* 20 (2019), Paper No. 73, 46.

[DMS09] Jean Dolbeault, Clément Mouhot, and Christian Schmeiser. "Hypocoercivity for kinetic equations with linear relaxation terms". In: *Comptes Rendus Mathématique* 347.9-10 (2009), pp. 511–516.

[DN21] Prafulla Dhariwal and Alexander Nichol. "Diffusion models beat GANs on image synthesis". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794.

[DNS23] Andrew Duncan, Nikolas Nuesken, and Lukasz Szpruch. "On the geometry of Stein variational gradient descent". In: *arXiv preprint 1912.00894* (2023).

[Dog+19] Pierre Dognin, Igor Melnyk, Youssef Mroueh, Jarret Ross, Cicero Dos Santos, and Tom Sercu. "Wasserstein barycenter model ensembling". In: *International Conference on Learning Representations*. 2019.

[Dom20] Justin Domke. "Provable smoothness guarantees for black-box variational inference". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 2587–2596.

[Dou+18] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, 2018, pp. xviii+757.

[DR13] Cynthia Dwork and Aaron Roth. "The algorithmic foundations of differential privacy". In: *Found. Trends Theor. Comput. Sci.* 9.3-4 (2013), pp. 211–487.

[DR16] Cynthia Dwork and Guy N. Rothblum. "Concentrated differential privacy". In: *arXiv preprint 1603.01887* (2016).

[DR20] Arnak S. Dalalyan and Lionel Riou-Durand. "On sampling from a log-concave density using kinetic Langevin diffusions". In: *Bernoulli* 26.3 (2020), pp. 1956–1988.

[DT12] Arnak S. Dalalyan and Alexandre B. Tsybakov. "Sparse regression learning by aggregation and Langevin Monte-Carlo". In: *J. Comput. System Sci.* 78.5 (2012), pp. 1423–1443.

[DV21] Amit Daniely and Gal Vardi. "From local pseudorandom generators to hardness of learning". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, Aug. 2021, pp. 1358–1394.

[Dvi22] Darina Dvinskikh. "Stochastic approximation versus sample average approximation for Wasserstein barycenters". In: *Optim. Methods Softw.* 37.5 (2022), pp. 1603–1635.

[DVK22] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. "Score-based generative modeling with critically-damped Langevin diffusion". In: *International Conference on Learning Representations*. 2022.

[Dwi+19] Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. "Log-concave sampling: Metropolis–Hastings algorithms are fast". In: *Journal of Machine Learning Research* 20.183 (2019), pp. 1–42.

[DZ10] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*. Vol. 38. Stochastic Modelling and Applied Probability. Corrected reprint of the second (1998) edition. Springer-Verlag, Berlin, 2010, pp. xvi+396.

[EB80] Donald L. Ermak and Helen Buckholtz. "Numerical integration of the Langevin equation: Monte Carlo simulation". In: *J. Comput. Phys.* 35.2 (1980), pp. 169–182.

[Ebe16] Andreas Eberle. "Reflection couplings and contraction rates for diffusions". In: *Probab. Theory Related Fields* 166.3-4 (2016), pp. 851–886.

[Eck87] Roger Eckhardt. "Stan Ulam, John von Neumann, and the Monte Carlo method". In: 15, Special Issue. With contributions by Tony Warnock, Gary D. Doolen and John Hendricks, Stanislaw Ulam 1909–1984. 1987, pp. 131–137.

[EH14] Tim van Erven and Peter Harremoës. "Rényi divergence and Kullback–Leibler divergence". In: *IEEE Trans. Inform. Theory* 60.7 (2014), pp. 3797–3820.

[EH21] Murat A. Erdogdu and Rasa Hosseinzadeh. "On the convergence of Langevin Monte Carlo: the interplay between tail growth and smoothness". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, Aug. 2021, pp. 1776–1822.

[EHZ22] Murat A. Erdogdu, Rasa Hosseinzadeh, and Shunshi Zhang. "Convergence of Langevin Monte Carlo in chi-squared and Rényi divergence". In: *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, Mar. 2022, pp. 8151–8175.

[EL18] Ronen Eldan and James R. Lee. "Regularization under diffusion and anticoncentration of the information content". In: *Duke Math. J.* 167.5 (2018), pp. 969–993.

[Elv+20] Filip Elvander, Isabel Haasler, Andreas Jakobsson, and Johan Karlsson. "Multi-marginal optimal transport using partial information with applications in robust localization and sensor fusion". In: *Signal Processing* 171 (2020), p. 107474.

[EMS18] Murat A. Erdogdu, Lester Mackey, and Ohad Shamir. "Global nonconvex optimization with discretized diffusions". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018.

[ENS17] Simon Eberle, Barbara Niethammer, and André Schlichting. "Gradient flow formulation and longtime behaviour of a constrained Fokker–Planck equation". In: *Nonlinear Anal.* 158 (2017), pp. 142–167.

[Fel+18] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. "Privacy amplification by iteration". In: *59th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2018*. IEEE Computer Soc., Los Alamitos, CA, 2018, pp. 521–532.

[FGP20] Max Fathi, Nathael Gozlan, and Maxime Prod'homme. "A proof of the Caffarelli contraction theorem via entropic regularization". In: *Calc. Var. Partial Differential Equations* 59.3 (2020), Paper No. 96, 18.

[FKP94] Alan Frieze, Ravi Kannan, and Nick Polson. "Sampling from log-concave distributions". In: *Ann. Appl. Probab.* 4.3 (1994), pp. 812–837.

[FKT20] Vitaly Feldman, Tomer Koren, and Kunal Talwar. "Private stochastic convex optimization: optimal rates in linear time". In: *STOC '20—Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, [2020] ©2020, pp. 439–449.

[FLO21] James Foster, Terry Lyons, and Harald Oberhauser. "The shifted ODE method for underdamped Langevin MCMC". In: *arXiv preprint 2101.03446* (2021).

[FMW16] Matthieu Fradelizi, Mokshay Madiman, and Liyao Wang. "Optimal concentration of information content for log-concave densities". In: *High dimensional probability VII*. Vol. 71. Progr. Probab. Springer, [Cham], 2016, pp. 45–60.

[Föl85] Hans Föllmer. "An entropy approach to the time reversal of diffusion processes". In: *Stochastic differential systems (Marseille-Luminy, 1984)*. Vol. 69. Lect. Notes Control Inf. Sci. Springer, Berlin, 1985, pp. 156–163.

[Fox15] Daniel J. F. Fox. "A Schwarz lemma for Kähler affine metrics and the canonical potential of a proper convex cone". In: *Ann. Mat. Pura Appl. (4)* 194.1 (2015), pp. 1–42.

[Fro+15] Roy Frostig, Rong Ge, Sham Kakade, and Aaron Sidford. "Unregularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization". In: *International Conference on Machine Learning*. Vol. 37. Proceedings of Machine Learning Research. PMLR, July 2015, pp. 2540–2548.

[FTC21] Jiaojiao Fan, Amirhossein Taghvaei, and Yongxin Chen. "Scalable computations of Wasserstein barycenter via input convex neural networks". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 1571–1581.

[FYC23] Jiaojiao Fan, Bo Yuan, and Yongxin Chen. "Improved dimension dependence of a proximal algorithm for sampling". In: *arXiv preprint 2302.10081* (2023).

[GC11] Mark Girolami and Ben Calderhead. "Riemann manifold Langevin and Hamiltonian Monte Carlo methods". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 73.2 (2011). With discussion and a reply by the authors, pp. 123–214.

[Gel+14] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian data analysis*. Third. Texts in Statistical Science Series. CRC Press, Boca Raton, FL, 2014, pp. xiv+661.

[Gel90] Matthias Gelbrich. "On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces". In: *Math. Nachr.* 147 (1990), pp. 185–203.

[Gen+20] Ivan Gentil, Christian Léonard, Luigia Ripani, and Luca Tamanini. "An entropic interpolation proof of the HWI inequality". In: *Stoch. Process. Appl.* 130.2 (2020), pp. 907–923.

[Gen08] Ivan Gentil. "From the Prékopa–Leindler inequality to modified logarithmic Sobolev inequality". In: *Ann. Fac. Sci. Toulouse Math. (6)* 17.2 (2008), pp. 291–308.

[GHN23] Aritra Guha, Nhat Ho, and XuanLong Nguyen. "On excess mass behavior in Gaussian mixture models with Orlicz–Wasserstein distances". In: *arXiv preprint 2301.11496* (2023).

[GJ20] Nathael Gozlan and Nicolas Juillet. "On a mixture of Brenier and Strassen theorems". In: *Proc. Lond. Math. Soc. (3)* 120.3 (2020), pp. 434–463.

[GK96] István Gyöngy and Nicolai Krylov. "Existence of strong solutions for Itô's stochastic equations via approximations". In: *Probab. Theory Related Fields* 105.2 (1996), pp. 143–158.

[GLG15] Andrew Gelman, Daniel Lee, and Jiqiang Guo. "Stan: a probabilistic programming language for Bayesian inference and optimization". In: *Journal of Educational and Behavioral Statistics* 40.5 (2015), pp. 530–543.

[GLL20] Rong Ge, Holden Lee, and Jianfeng Lu. "Estimating normalizing constants for log-concave distributions: algorithms and lower bounds". In: *STOC '20—Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, [2020] ©2020, pp. 579–586.

[GLL22] Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. "Private convex optimization via exponential mechanism". In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 1948–1989.

[GN22] Adam Gustafson and Hariharan Narayanan. "John's walk". In: *Advances in Applied Probability* (2022), pp. 1–19.

[Goo+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger. Vol. 27. Curran Associates, Inc., 2014.

[Gop+23a] Sivakanth Gopi, Yin Tat Lee, Daogao Liu, Ruoqi Shen, and Kevin Tian. "Algorithmic aspects of the log-Laplace transform and a non-Euclidean proximal sampler". In: *arXiv preprint 2302.06085* (2023).

[Gop+23b] Sivakanth Gopi, Yin Tat Lee, Daogao Liu, Ruoqi Shen, and Kevin Tian. "Private convex optimization in general norms". In: *Symposium on Discrete Algorithms*. 2023, pp. 5068–5089.

[Gor41] Robert D. Gordon. "Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument". In: *Ann. Math. Statistics* 12 (1941), pp. 364–366.

[Goz10] Nathael Gozlan. "Poincaré inequalities and dimension free concentration of measure". In: *Ann. Inst. Henri Poincaré Probab. Stat.* 46.3 (2010), pp. 708–739.

[GPC15] Alexandre Gramfort, Gabriel Peyré, and Marco Cuturi. "Fast optimal transport averaging of neuroimaging data". In: *Information Processing in Medical Imaging*. Ed. by Sebastien Ourselin, Daniel C. Alexander, Carl-Fredrik Westin, and M. Jorge Cardoso. Cham: Springer International Publishing, 2015, pp. 261–272.

[GPO21] Théo Galy-Fajou, Valerio Perrone, and Manfred Opper. "Flexible and efficient inference with particles for the variational Gaussian approximation". In: *Entropy* 23.8 (2021), Paper No. 990, 34.

[Gro75] Leonard Gross. "Logarithmic Sobolev inequalities". In: *American Journal of Mathematics* 97.4 (1975), pp. 1061–1083.

[GS90] Alan E. Gelfand and Adrian F. M. Smith. "Sampling-based approaches to calculating marginal densities". In: *J. Amer. Statist. Assoc.* 85.410 (1990), pp. 398–409.

[GSL92] Alan E. Gelfand, Adrian F. M. Smith, and Tai-Ming Lee. "Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling". In: *J. Amer. Statist. Assoc.* 87.418 (1992), pp. 523–532.

[GT20] Arun Ganesh and Kunal Talwar. "Faster differentially private samplers via Rényi divergence analysis of discretized Langevin MCMC". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 7222–7233.

[Gui+09] Arnaud Guillin, Christian Léonard, Liming Wu, and Nian Yao. "Transportation-information inequalities for Markov processes". In: *Probab. Theory Related Fields* 144.3-4 (2009), pp. 669–695.

[Gum+21] Sergey Guminov, Pavel Dvurechensky, Nazarii Tupitsa, and Alexander Gasnikov. "On a combination of alternating minimization and Nesterov's momentum". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 3886–3898.

[GV22] Khashayar Gatmiry and Santosh S. Vempala. "Convergence of the Riemannian Langevin algorithm". In: *arXiv preprint 2204.10818* (2022).

[GVV22] Aparna Gupte, Neekon Vafa, and Vinod Vaikuntanathan. "Continuous LWE is as hard as LWE & applications to learning Gaussian mixtures". In: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science—FOCS 2022*. IEEE Computer Soc., Los Alamitos, CA, [2022] ©2022, pp. 1162–1173.

[GW01] Fu-Zhou Gong and Feng-Yu Wang. "Heat kernel estimates with application to compactness of manifolds". In: *Q. J. Math.* 52.2 (2001), pp. 171–180.

[GWS21] Suriya Gunasekar, Blake Woodworth, and Nathan Srebro. "Mirrorless mirror descent: a natural derivation of mirror descent". In: *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*. Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, Apr. 2021, pp. 2305–2313.

[Haa+21] Isabel Haasler, Axel Ringh, Yongxin Chen, and Johan Karlsson. "Multimarginal optimal transport with a tree-structured cost and the Schrödinger bridge problem". In: *SIAM J. Control Optim.* 59.4 (2021), pp. 2428–2453.

[Han16] Ramon van Handel. *Probability in high dimension*. 2016.

[Has70] W. Keith Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1 (1970), pp. 97–109.

[HBE20] Ye He, Krishnakumar Balasubramanian, and Murat A. Erdogdu. "On the ergodicity, bias and asymptotic normality of randomized midpoint sampling method". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 7366–7376.

[HBE22] Ye He, Krishnakumar Balasubramanian, and Murat A. Erdogdu. "Heavy-tailed sampling via transformed unadjusted Langevin algorithm". In: *arXiv preprint 2201.08349* (2022).

[HGA15] Wen Huang, Kyle A. Gallivan, and Pierre-Antoine Absil. "A Broyden class of quasi-Newton methods for Riemannian optimization". In: *SIAM J. Optim.* 25.3 (2015), pp. 1660–1685.

[Hil14] Roland Hildebrand. "Canonical barriers on convex cones". In: *Math. Oper. Res.* 39.3 (2014), pp. 841–850.

[Hin18] Oliver Hinder. "Cutting plane methods can be extended into nonconvex optimization". In: *Proceedings of the 31st Conference on Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 1451–1454.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851.

[Ho+17] Nhat Ho, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. "Multilevel clustering via Wasserstein means". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 1501–1509.

[Hör67] Lars Hörmander. "Hypoelliptic second order differential equations". In: *Acta Math.* 119 (1967), pp. 147–171.

[HR21] Jan-Christian Hütter and Philippe Rigollet. "Minimax estimation of smooth optimal transport maps". In: *Ann. Statist.* 49.2 (2021), pp. 1166–1194.

[HS87] Richard Holley and Daniel Stroock. "Logarithmic Sobolev inequalities and stochastic Ising models". In: *J. Statist. Phys.* 46.5-6 (1987), pp. 1159–1194.

[Hsi+18] Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. "Mirrored Langevin dynamics". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018.

[HSV14] Martin Hairer, Andrew M. Stuart, and Sebastian J. Vollmer. "Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions". In: *Ann. Appl. Probab.* 24.6 (2014), pp. 2455–2490.

[Hua+22] Daniel Z. Huang, Jiaoyang Huang, Sebastian Reich, and Andrew M. Stuart. "Efficient derivative-free Bayesian inference for large-scale inverse problems". In: *Inverse Problems* 38.12 (2022), Paper No. 125006, 40.

[HV04] Antti Honkela and Harri Valpola. "Unsupervised variational Bayesian kearning of nonlinear models". In: *Advances in Neural Information Processing Systems*. Ed. by L. Saul, Y. Weiss, and L. Bottou. Vol. 17. MIT Press, 2004.

[HW17] Christopher C. Holmes and Stephen G. Walker. "Assigning a value to a power likelihood in a general Bayesian model". In: *Biometrika* 104.2 (2017), pp. 497–503.

[Hyv05] Aapo Hyvärinen. "Estimation of non-normalized statistical models by score matching". In: *J. Mach. Learn. Res.* 6 (2005), pp. 695–709.

[JA99] Valen E. Johnson and James H. Albert. *Ordinal data modeling*. Statistics for Social Science and Public Policy. Springer-Verlag, New York, 1999, pp. x+258.

[Jan+20] Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. "Entropic optimal transport between unbalanced Gaussian measures has a closed form". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 10468–10479.

[Jia21] Qijia Jiang. "Mirror Langevin Monte Carlo: the case under isoperimetry". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 715–725.

[Jin+21] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. "On nonconvex optimization for machine learning: gradients, stochasticity, and saddle points". In: *J. ACM* 68.2 (2021), Art. 11, 29.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. "The variational formulation of the Fokker–Planck equation". In: *SIAM J. Math. Anal.* 29.1 (1998), pp. 1–17.

[Jor+99] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. "An introduction to variational methods for graphical models". In: *Learning in Graphical Models*. Cambridge, MA, USA: MIT Press, 1999, pp. 105–161.

[JUD00] Simon Julier, Jeffrey Uhlmann, and Hugh F. Durrant-Whyte. "A new method for the nonlinear transformation of means and covariances in filters and estimators". In: *IEEE Trans. Automat. Control* 45.3 (2000), pp. 477–482.

[Kel17] Martin Kell. "On interpolation and curvature via Wasserstein geodesics". In: *Advances in Calculus of Variations* 10.2 (2017), pp. 125–167.

[KHK23] Mohammad R. Karimi, Ya-Ping Hsieh, and Andreas Krause. "A dynamical system view of Langevin-based non-convex sampling". In: *arXiv preprint 2210.13867* (2023).

[Kin+21] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. "Variational diffusion models". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 21696–21707.

[KJD22] Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. "An optimization-centric view on Bayes' rule: reviewing and generalizing variational inference". In: *Journal of Machine Learning Research* 23.132 (2022), pp. 1–109.

[KL22] Bo'az Klartag and Joseph Lehec. "Bourgain's slicing problem and KLS isoperimetry up to polylog". In: *Geom. Funct. Anal.* 32.5 (2022), pp. 1134–1159.

[KLS95] Ravi Kannan, László Lovász, and Miklós Simonovits. "Isoperimetric problems for convex bodies and a localization lemma". In: *Discrete Comput. Geom.* 13.3-4 (1995), pp. 541–559.

[KM12] Young-Heon Kim and Emanuel Milman. "A generalization of Caffarelli's contraction theorem via (reverse) heat flow". In: *Math. Ann.* 354.3 (2012), pp. 827–862.

[KN12] Ravindran Kannan and Hariharan Narayanan. "Random walks on polytopes and an affine interior point method for linear programming". In: *Math. Oper. Res.* 37.1 (2012), pp. 1–20.

[KNS16] Hamed Karimi, Julie Nutini, and Mark Schmidt. "Linear convergence of gradient and proximal-gradient methods under the Polyak–Lojasiewicz condition". In: *European Conference on Machine Learning and Knowledge Discovery in Databases—Volume 9851*. ECML PKDD 2016. Riva del Garda, Italy: Springer-Verlag, 2016, pp. 795–811.

[Kol11] Alexander V. Kolesnikov. "Mass transportation and contractions". In: *arXiv preprint 1103.1479* (2011).

[Kol34] Andrei N. Kolmogorov. "Zufällige Bewegungen (zur Theorie der Brownschen Bewegung)". In: *Ann. of Math. (2)* 35.1 (1934), pp. 116–117.

[Kor+21] Alexander Korotin, Lingxiao Li, Justin M. Solomon, and Evgeny Burnaev. "Continuous Wasserstein-2 barycenter estimation without minimax optimization". In: *International Conference on Learning Representations*. 2021.

[KP21] Bo'az Klartag and Eli Putterman. "Spectral monotonicity under Gaussian convolution". In: *arXiv preprint 2107.09496* (2021).

[KR22] Mohammad E. Khan and Håvard Rue. "The Bayesian learning rule". In: *arXiv preprint 2107.04562* (2022).

[Kro+19] Alexey Kroshnin, Nazarii Tupitsa, Darina Dvinskikh, Pavel Dvurechensky, Alexander Gasnikov, and Cesar Uribe. "On the complexity of approximating Wasserstein barycenters". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 3530–3540.

[Kro18] Alexey Kroshnin. "Fréchet barycenters in the Monge-Kantorovich spaces". In: *J. Convex Anal.* 25.4 (2018), pp. 1371–1395.

[KS91] Ioannis Karatzas and Steven E. Shreve. *Brownian motion and stochastic calculus*. Second. Vol. 113. Graduate Texts in Mathematics. Springer-Verlag, New York, 1991, pp. xxiv+470.

[KS94] Martin Knott and Cyril S. Smith. "On a generalization of cyclic monotonicity and distances among random vectors". In: *Linear Algebra Appl.* 199 (1994), pp. 363–371.

[KSS21] Alexey Kroshnin, Vladimir Spokoiny, and Alexandra Suvorikova. "Statistical inference for Bures–Wasserstein barycenters". In: *Ann. Appl. Probab.* 31.3 (2021), pp. 1264–1298.

[KY03] Harold J. Kushner and G. George Yin. *Stochastic approximation and recursive algorithms and applications*. Second. Vol. 35. Applications of Mathematics (New York). Stochastic Modelling and Applied Probability. Springer-Verlag, New York, 2003, pp. xxii+474.

[Lam+22] Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. "Variational inference via Wasserstein gradient flows". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho. 2022.

[LBB22a] Marc Lambert, Silvère Bonnabel, and Francis Bach. "The continuous-discrete variational Kalman filter (CD-VKF)". In: *2022 IEEE 61st Conference on Decision and Control (CDC)*. 2022, pp. 6632–6639.

[LBB22b] Marc Lambert, Silvère Bonnabel, and Francis Bach. "The recursive variational Gaussian approximation (R-VGA)". In: *Stat. Comput.* 32.1 (2022), Paper No. 10, 24.

[LBB23] Marc Lambert, Silvère Bonnabel, and Francis Bach. "The limited-memory recursive variational Gaussian approximation (L-RVGA)". In: *arXiv preprint 2303.14195* (2023).

[LC22] Jiaming Liang and Yongxin Chen. "A proximal algorithm for sampling from non-smooth potentials". In: *arXiv preprint 2110.04597* (2022).

[LC23] Jiaming Liang and Yongxin Chen. "A proximal algorithm for sampling". In: *arXiv preprint 2202.13975* (2023).

[Le +22] Thibaut Le Gouic, Quentin Paris, Philippe Rigollet, and Austin J. Stromme. "Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space". In: *J. Eur. Math. Soc.* (2022).

[Le 16] Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*. French. Vol. 274. Graduate Texts in Mathematics. Springer, [Cham], 2016, pp. xiii+273.

[Le 86] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, 1986, pp. xxvi+742.

[LE23] Mufan (Bill) Li and Murat A. Erdogdu. "Riemannian Langevin algorithm for solving semidefinite programs". In: *arXiv preprint 2010.11176* (2023).

[Led00] Michel Ledoux. "The geometry of Markov diffusion generators". In: vol. 9. 2. Probability theory. 2000, pp. 305–366.

[Led18] Michel Ledoux. *Remarks on some transportation cost inequalities*. 2018.

[Leh22] Joseph Lehec. "The Langevin Monte Carlo algorithm in the non-smooth log-concave case". In: *arXiv preprint 2101.10695* (2022).

[Léo14] Christian Léonard. "A survey of the Schrödinger problem and some of its connections with optimal transport". In: *Discrete Contin. Dyn. Syst.* 34.4 (2014), pp. 1533–1574.

[LFN18] Haihao Lu, Robert M. Freund, and Yurii Nesterov. "Relatively smooth convex optimization by first-order methods, and applications". In: *SIAM J. Optim.* 28.1 (2018), pp. 333–354.

[Li+19] Xuechen Li, Yi Wu, Lester Mackey, and Murat A. Erdogdu. "Stochastic Runge–Kutta accelerates Langevin Monte Carlo and beyond". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[Li+20] Lingxiao Li, Aude Genevay, Mikhail Yurochkin, and Justin M. Solomon. "Continuous regularized Wasserstein barycenters". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 17755–17765.

[Li+21] Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtarik. "PAGE: a simple and optimal probabilistic gradient estimator for nonconvex optimization". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 6286–6295.

[Li+22] Ruilin Li, Molei Tao, Santosh S. Vempala, and Andre Wibisono. "The mirror Langevin algorithm converges with vanishing bias". In: *Proceedings of the 33rd International Conference on Algorithmic Learning Theory*. Ed. by Sanjoy Dasgupta and Nika Haghtalab. Vol. 167. Proceedings of Machine Learning Research. PMLR, Mar. 2022, pp. 718–742.

[Lin+20] Tianyi Lin, Nhat Ho, Xi Chen, Marco Cuturi, and Michael Jordan. "Fixed-support Wasserstein barycenters: computational hardness and fast algorithm". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 5368–5380.

[Lin+22] Tianyi Lin, Nhat Ho, Marco Cuturi, and Michael I. Jordan. "On the complexity of approximating multimarginal optimal transport". In: *Journal of Machine Learning Research* 23.65 (2022), pp. 1–43.

[Liu+22] Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. "Let us build bridges: understanding and extending diffusion generative models". In: *arXiv preprint 2208.14699* (2022).

[Liu08] Jun S. Liu. *Monte Carlo strategies in scientific computing*. Springer Series in Statistics. Springer, New York, 2008, pp. xvi+343.

[Liu17] Qiang Liu. "Stein variational gradient descent as gradient flow". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017.

[Liu20] Yuan Liu. "The Poincaré inequality and quadratic transportation-variance inequalities". In: *Electron. J. Probab.* 25 (2020), Paper No. 1, 16.

[LKS19a] Wu Lin, Mohammad E. Khan, and Mark Schmidt. "Fast and simple natural-gradient variational inference with mixture of exponential-family approximations". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 3992–4002.

[LKS19b] Wu Lin, Mohammad E. Khan, and Mark Schmidt. "Stein's lemma for the reparameterization trick with exponential family mixtures". In: *arXiv preprint 1910.13398* (2019).

[LL01] Jimmie D. Lawson and Yongdo Lim. "The geometric mean, matrices, metrics, and more". In: *Amer. Math. Monthly* 108.9 (2001), pp. 797–812.

[LL08] Claude Le Bris and Pierre-Louis Lions. "Existence and uniqueness of solutions to Fokker–Planck type equations with irregular coefficients". In: *Comm. Partial Differential Equations* 33.7-9 (2008), pp. 1272–1317.

[LL17] Thibaut Le Gouic and Jean-Michel Loubes. "Existence and consistency of Wasserstein barycenters". In: *Probab. Theory Related Fields* 168.3-4 (2017), pp. 901–917.

[LLN19] Yulong Lu, Jianfeng Lu, and James Nolen. "Accelerating Langevin sampling with birth-death". In: *arXiv preprint 1905.09863* (2019).

[LLR20] Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. "Projection to fairness in statistical learning". In: *arXiv preprint 2005.11720* (2020).

[LLT22] Holden Lee, Jianfeng Lu, and Yixin Tan. "Convergence for score-based generative modeling with polynomial complexity". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho. 2022.

[LLT23] Holden Lee, Jianfeng Lu, and Yixin Tan. "Convergence of score-based generative modeling for general data distributions". In: *Proceedings of the 34th International Conference on Algorithmic Learning Theory*. Ed. by Shipra Agrawal and Francesco Orabona. Vol. 201. Proceedings of Machine Learning Research. PMLR, Feb. 2023, pp. 946–985.

[LLV20] Aditi Laddha, Yin Tat Lee, and Santosh Vempala. "Strong self-concordance and sampling". In: *STOC '20—Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, [2020] ©2020, pp. 1212–1222.

[LMH15] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. "A universal catalyst for first-order optimization". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran Associates, Inc., 2015.

[LMS16] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. "Optimal transport in competition with reaction: the Hellinger–Kantorovich distance and geodesic curves". In: *SIAM J. Math. Anal.* 48.4 (2016), pp. 2869–2911.

[LMS18] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. "Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures". In: *Invent. Math.* 211.3 (2018), pp. 969–1117.

[LO00] Rafał Latała and Krzysztof Oleszkiewicz. "Between Sobolev and Poincaré". In: *Geometric aspects of functional analysis*. Vol. 1745. Lecture Notes in Math. Springer, Berlin, 2000, pp. 147–168.

[Łoj63] Stanisław Łojasiewicz. "Une propriété topologique des sous-ensembles analytiques réels". In: *Les Équations aux Dérivées Partielles (Paris, 1962)*. Éditions du Centre National de la Recherche Scientifique (CNRS), Paris, 1963, pp. 87–89.

[LP02] Damien Lamberton and Gilles Pagès. "Recursive computation of the invariant distribution of a diffusion". In: *Bernoulli* 8.3 (2002), pp. 367–405.

[LRG18] Holden Lee, Andrej Risteski, and Rong Ge. "Beyond log-concavity: provable guarantees for sampling multi-modal distributions using simulated tempering Langevin Monte Carlo". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018.

[LS16] Shiwei Lan and Babak Shahbaba. "Sampling constrained probability distributions using spherical augmentation". In: *Algorithmic advances in Riemannian geometry and applications*. Adv. Comput. Vis. Pattern Recognit. Springer, Cham, 2016, pp. 25–71.

[LS88] Gregory F. Lawler and Alan D. Sokal. "Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality". In: *Trans. Amer. Math. Soc.* 309.2 (1988), pp. 557–580.

[LS93] Lázló Lovász and Miklós Simonovits. "Random walks in a convex body and an improved volume algorithm". In: *Random Structures Algorithms* 4.4 (1993), pp. 359–412.

[LST20] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. "Logsmooth gradient concentration and tighter runtimes for Metropolized Hamiltonian Monte Carlo". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 2565–2597.

[LST21a] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. "Lower bounds on Metropolized sampling methods for well-conditioned distributions". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 18812–18824.

[LST21b] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. "Structured logconcave sampling with a restricted Gaussian oracle". In: *arXiv preprint 2010.03106* (2021).

[LST21c] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. "Structured logconcave sampling with a restricted Gaussian oracle". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, Aug. 2021, pp. 2993–3050.

[LV07] László Lovász and Santosh Vempala. "The geometry of logconcave functions and sampling algorithms". In: *Random Structures Algorithms* 30.3 (2007), pp. 307–358.

[LV09] John Lott and Cédric Villani. "Ricci curvature for metric-measure spaces via optimal transport". In: *Ann. of Math. (2)* 169.3 (2009), pp. 903–991.

[LV18a] Yin Tat Lee and Santosh S. Vempala. "Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation". In: *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2018, pp. 1115–1121.

[LV18b] Yin Tat Lee and Santosh S. Vempala. "Stochastic localization + Stieltjes barrier = tight bound for log-Sobolev". In: *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, 2018, pp. 1122–1129.

[LV22] Yin Tat Lee and Santosh Vempala. "Geodesic walks in polytopes". In: *SIAM J. Comput.* 51.2 (2022), STOC17-400–STOC17-488.

[LW16] Qiang Liu and Dilin Wang. "Stein variational gradient descent: a general purpose Bayesian inference algorithm". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc., 2016.

[LW22] Jianfeng Lu and Lihan Wang. "Complexity of zigzag sampling algorithm for strongly log-concave distributions". In: *Stat. Comput.* 32.3 (2022), Paper No. 48, 12.

[LY00] Lucien Le Cam and Grace Lo Yang. *Asymptotics in statistics*. Second. Springer Series in Statistics. Some basic concepts. Springer-Verlag, New York, 2000, pp. xiv+285.

[LY21] Yin Tat Lee and Man-Chung Yue. "Universal barrier is n-self-concordant". In: *Math. Oper. Res.* 46.3 (2021), pp. 1129–1148.

[Ma+21] Yi-An Ma, Niladri S. Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L. Bartlett, and Michael I. Jordan. "Is there an analog of Nesterov acceleration for gradient-based MCMC?" In: *Bernoulli* 27.3 (2021), pp. 1942–1992.

[Mac03] David J. C. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, New York, 2003, pp. xii+628.

[Mar+12] James Martin, Lucas C. Wilcox, Carsten Burstedde, and Omar Ghattas. "A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion". In: *SIAM J. Sci. Comput.* 34.3 (2012), A1460–A1487.

[Mar+16] Yosra Marnissi, Emilie Chouzenoux, Jean-Christophe Pesquei, and Amel Benazza-Benyahia. "An auxiliary variable method for Langevin based MCMC algorithms". In: *2016 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE. 2016, pp. 1–5.

[Mar70] Bernard Martinet. "Régularisation d'inéquations variationnelles par approximations successives". In: *Rev. Française Informat. Recherche Opérationnelle* 4.Sér. R-3 (1970), pp. 154–158.

[Mar99] Fabio Martinelli. "Lectures on Glauber dynamics for discrete spin models". In: *Lectures on probability theory and statistics (Saint-Flour, 1997)*. Vol. 1717. Lecture Notes in Math. Springer, Berlin, 1999, pp. 93–191.

[Met+53] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. "Equation of state calculations by fast computing machines". In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092.

[MGM22] Anton Mallasto, Augusto Gerolin, and Hà Quang Minh. "Entropy-regularized 2-Wasserstein distance between Gaussian measures". In: *Inf. Geom.* 5.1 (2022), pp. 289–323.

[Mir17] Ilya Mironov. "Rényi differential privacy". In: *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*. Los Alamitos, CA, USA: IEEE Computer Society, Aug. 2017, pp. 263–275.

[MMP18] Luigi Malagò, Luigi Montrucchio, and Giovanni Pistone. "Wasserstein Riemannian geometry of Gaussian densities". In: *Inf. Geom.* 1.2 (2018), pp. 137–179.

[MMS20] Mateusz B. Majka, Aleksandar Mijatović, and Łukasz Szpruch. "Non-asymptotic bounds for sampling algorithms without log-concavity". In: *Ann. Appl. Probab.* 30.4 (2020), pp. 1534–1581.

[MN19] Gonzalo Mena and Jonathan Niles-Weed. "Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[Mod17] Klas Modin. "Geometry of matrix decompositions seen through optimal transport and information geometry". In: *J. Geom. Mech.* 9.3 (2017), pp. 335–390.

[Mon21] Pierre Monmarché. "High-dimensional MCMC with a standard splitting scheme for the underdamped Langevin diffusion". In: *Electronic Journal of Statistics* 15.2 (2021), pp. 4117–4166.

[Mou+22] Wenlong Mou, Nicolas Flammarion, Martin J. Wainwright, and Peter L. Bartlett. "Improved bounds for discretization of Langevin diffusions: near-optimal rates without convexity". In: *Bernoulli* 28.3 (2022), pp. 1577–1601.

[MS21] Dan Mikulincer and Yair Shenfeld. "The Brownian transport map". In: *arXiv preprint 2111.11521* (2021).

[MS22] Dan Mikulincer and Yair Shenfeld. "On the Lipschitz properties of transportation along heat flows". In: *arXiv preprint 2201.01382* (2022).

[MT07] Frank McSherry and Kunal Talwar. "Mechanism design via differential privacy". In: *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. 2007.

[MT09] Sean Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Second. With a prologue by Peter W. Glynn. Cambridge University Press, Cambridge, 2009, pp. xxviii+594.

[MV19] Oren Mangoubi and Nisheeth K. Vishnoi. "Nonconvex sampling with the Metropolis-adjusted Langevin algorithm". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by Alina Beygelzimer and Daniel Hsu. Vol. 99. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 2259–2293.

[NDC21] Dao Nguyen, Xin Dang, and Yixin Chen. "Unadjusted Langevin algorithm for non-convex weakly smooth potentials". In: *arXiv preprint 2101.06369* (2021).

[Nea11] Radford M. Neal. "MCMC using Hamiltonian dynamics". In: *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC Handb. Mod. Stat. Methods. CRC Press, Boca Raton, FL, 2011, pp. 113–162.

[Nee22] Joe Neeman. "Lipschitz changes of variables via heat flow". In: *arXiv preprint 2201.03403* (2022).

[Nem94] Arkadi S. Nemirovski. *Efficient methods in convex programming*. Lecture Notes (Technion - Georgia Tech). 1994.

[Nes12] Yurii Nesterov. "How to make the gradients small". In: *Optima. Mathematical Optimization Society Newsletter* 88 (2012), pp. 10–11.

[Nes15] Yurii Nesterov. "Universal gradient methods for convex optimization problems". In: *Math. Program.* 152.1-2, Ser. A (2015), pp. 381–404.

[Nes18] Yurii Nesterov. *Lectures on convex optimization*. Vol. 137. Springer Optimization and Its Applications. Springer, 2018, pp. xxiii+589.

[Neu51] John von Neumann. "Various techniques used in connection with random digits". In: *Monte Carlo method*. Ed. by A. S. Householder, G. E. Forsythe, and H. H. Germond. Vol. 12. National Bureau of Standards Applied Mathematics Series. Washington, DC: US Government Printing Office, 1951. Chap. 13, pp. 36–38.

[Ngu14] Van Hoang Nguyen. "Dimensional variance inequalities of Brascamp–Lieb type and a local approach to dimensional Prékopa's theorem". In: *J. Funct. Anal.* 266.2 (2014), pp. 931–955.

[NN94] Yurii Nesterov and Arkadi S. Nemirovski. *Interior-point polynomial algorithms in convex programming*. Vol. 13. SIAM Studies in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994, pp. x+405.

[NP06] Yurii Nesterov and Boris T. Polyak. "Cubic regularization of Newton method and its global performance". In: *Math. Program.* 108.1, Ser. A (2006), pp. 177–205.

[NS17] Yurii Nesterov and Vladimir Spokoiny. "Random gradient-free minimization of convex functions". In: *Found. Comput. Math.* 17.2 (2017), pp. 527–566.

[NW22] Marcel Nutz and Johannes Wiesel. "Entropic optimal transport: convergence of potentials". In: *Probab. Theory Related Fields* 184.1-2 (2022), pp. 401–424.

[NY83] Arkadi S. Nemirovski and David B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. Translated from the Russian and with a preface by E. R. Dawson. John Wiley & Sons, Inc., New York, 1983, pp. xv+388.

[OA09] Manfred Opper and Cédric Archambeau. "The variational Gaussian approximation revisited". In: *Neural Comput.* 21.3 (2009), pp. 786–792.

[ODo+16] Brendan O'Donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. "Conic optimization via operator splitting and homogeneous self-dual embedding". In: *J. Optim. Theory Appl.* 169.3 (2016), pp. 1042–1068.

[Øks03] Bernt Øksendal. *Stochastic differential equations*. Sixth. Universitext. An introduction with applications. Springer-Verlag, Berlin, 2003, pp. xxiv+360.

[OP15] Shin-ichi Ohta and Miklós Pálfia. "Discrete-time gradient flows and law of large numbers in Alexandrov spaces". In: *Calc. Var. Partial Differential Equations* 54.2 (2015), pp. 1591–1610.

[OR93] Ingram Olkin and Svetlozar T. Rachev. "Maximum submatrix traces for positive definite matrices". In: *SIAM J. Matrix Anal. Appl.* 14.2 (1993), pp. 390–397.

[Ott01] Felix Otto. "The geometry of dissipative evolution equations: the porous medium equation". In: *Comm. Partial Differential Equations* 26.1-2 (2001), pp. 101–174.

[Ott98] Felix Otto. "Dynamics of labyrinthine pattern formation in magnetic fluids: a mean-field theory". In: *Arch. Rational Mech. Anal.* 141.1 (1998), pp. 63–103.

[OV00] Felix Otto and Cédric Villani. "Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality". In: *J. Funct. Anal.* 173.2 (2000), pp. 361–400.

[OV01] Felix Otto and Cédric Villani. "Comment on: "Hypercontractivity of Hamilton–Jacobi equations" [J. Math. Pures Appl. (9) **80** (2001), no. 7, 669–696] by S. G. Bobkov, I. Gentil and M. Ledoux". In: *J. Math. Pures Appl. (9)* 80.7 (2001), pp. 697–700.

[Pav14] Grigorios A. Pavliotis. *Stochastic processes and applications*. Vol. 60. Texts in Applied Mathematics. Diffusion processes, the Fokker–Planck and Langevin equations. Springer, New York, 2014, pp. xiv+339.

[PBJ12] John Paisley, David M. Blei, and Michael I. Jordan. "Variational Bayesian inference with stochastic search". In: *Proceedings of the 29th International Conference on Machine Learning*. ICML'12. Edinburgh, Scotland: Omnipress, 2012, pp. 1363–1370.

[PBJ15] John Paisley, David M. Blei, and Michael I. Jordan. "Bayesian nonnegative matrix factorization with stochastic variational inference". In: *Handbook of mixed membership models and their applications*. Chapman & Hall/CRC Handb. Mod. Stat. Methods. CRC Press, Boca Raton, FL, 2015, pp. 205–224.

[PC19] Gabriel Peyré and Marco Cuturi. *Computational optimal transport: with applications to data science*. Now, 2019.

[PCN22] Aram-Alexandre Pooladian, Marco Cuturi, and Jonathan Niles-Weed. "Debiaser beware: pitfalls of centering regularized transport maps". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 17830–17847.

[Per16] Marcelo Pereyra. "Proximal Markov chain Monte Carlo algorithms". In: *Stat. Comput.* 26.4 (2016), pp. 745–760.

[Pid22] Jakiw Pidstrigach. "Score-based generative models detect manifolds". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 35852–35865.

[PN22] Aram-Alexandre Pooladian and Jonathan Niles-Weed. "Entropic estimation of optimal transport maps". In: *arXiv preprint 2109.12004* (2022).

[Pol63] Boris T. Polyak. "Gradient methods for minimizing functionals". In: *Ž. Vyčisl. Mat i Mat. Fiz.* 3 (1963), pp. 643–653.

[PP12] Gilles Pagès and Fabien Panloup. "Ergodic approximation of the distribution of a stationary diffusion: rate of convergence". In: *Ann. Appl. Probab.* 22.3 (2012), pp. 1059–1100.

[PP14] Ari Pakman and Liam Paninski. "Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians". In: *J. Comput. Graph. Statist.* 23.2 (2014), pp. 518–542.

[Pro21] Maxime Prod'homme. "Contributions au problème du transport optimal et à sa régularité". Thèse de doctorat dirigée par Fathi, Max et Otto, Felix Mathématiques et Applications Toulouse 3 2021. PhD thesis. 2021.

[PST12] Natesh S. Pillai, Andrew M. Stuart, and Alexandre H. Thiéry. "Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions". In: *Ann. Appl. Probab.* 22.6 (2012), pp. 2320–2356.

[PZ16] Victor M. Panaretos and Yoav Zemel. "Amplitude and phase variation of point processes". In: *Ann. Statist.* 44.2 (2016), pp. 771–812.

[PZ19] Victor M. Panaretos and Yoav Zemel. "Statistical aspects of Wasserstein distances". In: *Annu. Rev. Stat. Appl.* 6 (2019), pp. 405–431.

[Rab+12] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. "Wasserstein barycenter and its application to texture mixing". In: *Scale Space and Variational Methods in Computer Vision*. Ed. by Alfred M. Bruckstein, Bart M. ter Haar Romeny, Alexander M. Bronstein, and Michael M. Bronstein. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 435–446.

[Ram+22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. "Hierarchical text-conditional image generation with CLIP latents". In: *arXiv preprint 2204.06125* (2022).

[RBP22] Théo Ryffel, Francis Bach, and David Pointcheval. "Differential privacy guarantees for stochastic gradient Langevin dynamics". In: *arXiv preprint 2201.11980* (2022).

[RC04] Christian P. Robert and George Casella. *Monte Carlo statistical methods*. Second. Springer Texts in Statistics. Springer-Verlag, New York, 2004, pp. xxx+645.

[RGB14] Rajesh Ranganath, Sean Gerrish, and David M. Blei. "Black box variational inference". In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*. Ed. by Samuel Kaski and Jukka Corander. Vol. 33. Proceedings of Machine Learning Research. Reykjavik, Iceland: PMLR, Apr. 2014, pp. 814–822.

[RGG97] Gareth O. Roberts, Andrew Gelman, and Walter R. Gilks. "Weak convergence and optimal scaling of random walk Metropolis algorithms". In: *Ann. Appl. Probab.* 7.1 (1997), pp. 110–120.

[Roc76] R. Tyrrell Rockafellar. "Monotone operators and the proximal point algorithm". In: *SIAM J. Control Optim.* 14.5 (1976), pp. 877–898.

[Roc97] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Reprint of the 1970 original, Princeton Paperbacks. Princeton University Press, Princeton, NJ, 1997, pp. xviii+451.

[Rol22] Paul T. V. Rolland. "Predicting in uncertain environments: methods for robust machine learning". PhD thesis. EPFL, 2022.

[RP15] Julien Rabin and Nicolas Papadakis. "Convex color image segmentation with optimal transport distances". In: *Scale space and variational methods in computer vision*. Vol. 9087. Lecture Notes in Comput. Sci. Springer, Cham, 2015, pp. 256–269.

[RR91] Malempati M. Rao and Zhong D. Ren. *Theory of Orlicz spaces*. Vol. 146. Monographs and Textbooks in Pure and Applied Mathematics. Marcel Dekker, Inc., New York, 1991, pp. xii+449.

[RR98] Gareth O. Roberts and Jeffrey S. Rosenthal. "Optimal scaling of discrete approximations to Langevin diffusions". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 60.1 (1998), pp. 255–268.

[RRT17] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. "Nonconvex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis". In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by Satyen Kale and Ohad Shamir. Vol. 65. Proceedings of Machine Learning Research. PMLR, July 2017, pp. 1674–1703.

[RS15] Firas Rassoul-Agha and Timo Seppäläinen. *A course on large deviations with an introduction to Gibbs measures*. Vol. 162. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2015, pp. xiv+318.

[RS18] Julien Roussel and Gabriel Stoltz. "Spectral methods for Langevin dynamics and associated error estimates". In: *ESAIM. Mathematical Modelling and Numerical Analysis* 52.3 (2018), pp. 1051–1083.

[RU02] Ludger Rüschendorf and Ludger Uckelmann. "On the n-coupling problem". In: *J. Multivariate Anal.* 81.2 (2002), pp. 242–258.

[San15] Filippo Santambrogio. *Optimal transport for applied mathematicians*. Vol. 87. Progress in Nonlinear Differential Equations and their Applications. Calculus of variations, PDEs, and modeling. Birkhäuser/Springer, Cham, 2015, pp. xxvii+353.

[San17] Filippo Santambrogio. "{Euclidean, metric, and Wasserstein} gradient flows: an overview". In: *Bull. Math. Sci.* 7.1 (2017), pp. 87–154.

[Sär07] Simo Särkkä. "On unscented Kalman filtering for state estimation of continuous-time nonlinear systems". In: *IEEE Trans. Automat. Control* 52.9 (2007), pp. 1631–1641.

[SBD21] Matteo Sordello, Zhiqi Bu, and Jinshuo Dong. "Privacy amplification via iteration for shuffled and online PNSGD". In: *Machine Learning and Knowledge Discovery in Databases. Research Track*. Ed. by Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read, and Jose A. Lozano. Cham: Springer International Publishing, 2021, pp. 796–813.

[Sch19] André Schlichting. "Poincaré and log–Sobolev inequalities for mixtures". In: *Entropy* 21.1 (2019).

[SE19] Yang Song and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[See99] Matthias Seeger. "Bayesian model selection for support vector machines, Gaussian processes and other kernel classifiers". In: *Advances in Neural Information Processing Systems*. Ed. by S. Solla, T. Leen, and K. Müller. Vol. 12. MIT Press, 1999.

[Seg+18] Vivien Seguy, Bharath B. Damodaran, Remi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. "Large scale optimal transport and mapping estimation". In: *International Conference on Learning Representations*. 2018.

[Sha+19] Stephane Shao, Pierre E. Jacob, Jie Ding, and Vahid Tarokh. "Bayesian model comparison with the Hyvärinen score: computation and consistency". In: *J. Amer. Statist. Assoc.* 114.528 (2019), pp. 1826–1837.

[Sim+16] Umut Simsekli, Roland Badeau, Taylan Cemgil, and Gaël Richard. "Stochastic quasi-Newton Langevin Monte Carlo". In: *Proceedings of the 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, June 2016, pp. 642–651.

[SKL20] Adil Salim, Anna Korba, and Giulia Luise. "The Wasserstein proximal gradient algorithm". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 12356–12366.

[SL19] Ruoqi Shen and Yin Tat Lee. "The randomized midpoint method for log-concave sampling". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[SLD18] Sanvesh Srivastava, Cheng Li, and David B. Dunson. "Scalable Bayes via barycenter in Wasserstein space". In: *Journal of Machine Learning Research* 19.8 (2018), pp. 1–35.

[Soh+15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. "Deep unsupervised learning using nonequilibrium thermodynamics". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 2256–2265.

[Sol+15] Justin M. Solomon, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. "Convolutional Wasserstein distances: efficient optimal transportation on geometric domains". In: *ACM Trans. Graph.* 34.4 (July 2015).

[Son+21a] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. "Maximum likelihood training of score-based diffusion models". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 1415–1428.

[Son+21b] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. "Score-based generative modeling through stochastic differential equations". In: *International Conference on Learning Representations*. 2021.

[SR20] Adil Salim and Peter Richtarik. "Primal dual interpretation of the proximal stochastic gradient Langevin algorithm". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 3786–3796.

[SS19] Simo Särkkä and Arno Solin. *Applied stochastic differential equations*. Vol. 10. Institute of Mathematical Statistics Textbooks. Cambridge University Press, Cambridge, 2019, pp. ix+316.

[Str18] Daniel W. Stroock. *Elements of stochastic calculus and analysis*. CRM Short Courses. Springer, Cham, 2018, pp. xiv+206.

[Stu03] Karl-Theodor Sturm. "Probability measures on metric spaces of nonpositive curvature". In: *Heat kernels and analysis on manifolds, graphs, and metric spaces (Paris, 2002)*. Vol. 338. Contemp. Math. Amer. Math. Soc., Providence, RI, 2003, pp. 357–390.

[Stu11] Karl-Theodor Sturm. "Generalized Orlicz spaces and Wasserstein distances for convex-concave scale functions". In: *Bulletin des Sciences Mathématiques* 135.6-7 (2011), pp. 795–802.

[SV06] Daniel W. Stroock and S. R. Srinivasa Varadhan. *Multidimensional diffusion processes*. Classics in Mathematics. Reprint of the 1997 edition. Springer-Verlag, Berlin, 2006, pp. xii+338.

[SW14] Adrien Saumard and Jon A. Wellner. "Log-concavity and strong log-concavity: a review". In: *Stat. Surv.* 8 (2014), pp. 45–114.

[TP18] Michalis K. Titsias and Omiros Papaspiliopoulos. "Auxiliary gradient-based sampling algorithms". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 80.4 (2018), pp. 749–767.

[Tri+18] Nilesh Tripuraneni, Nicolas Flammarion, Francis Bach, and Michael I. Jordan. "Averaging stochastic gradient descent on Riemannian manifolds". In: *Proceedings of the 31st Conference on Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 650–687.

[Tro77] M. M. Tropper. "Ergodic properties of infinite-dimensional stochastic systems". In: *J. Statist. Phys.* 17.6 (1977), pp. 511–528.

[Tsy09] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. Springer, New York, 2009, pp. xii+214.

[TV00] Giuseppe Toscani and Cédric Villani. "On the trend to equilibrium for some dissipative systems with slowly increasing a priori bounds". In: *J. Statist. Phys.* 98.5-6 (2000), pp. 1279–1309.

[TW11] Adrian Tudorascu and Marcus Wunsch. "On a nonlinear, nonlocal parabolic problem with conservation of mass, mean and variance". In: *Comm. Partial Differential Equations* 36.8 (2011), pp. 1426–1454.

[Val07] Stefán I. Valdimarsson. "On the Hessian of the optimal transport potential". In: *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* 6.3 (2007), pp. 441–456.

[Vav93] Stephen A. Vavasis. "Black-box complexity of local minimization". In: *SIAM J. Optim.* 3.1 (1993), pp. 60–80.

[Vaz09] Andrew Vazsonyi. "On the point for which the sum of the distances to n given points is minimum". In: *Ann. Oper. Res.* 167 (2009). Translated from the French original [Tohoku Math. J. **43** (1937), 355–386] and annotated by Frank Plastria, pp. 7–41.

[Ver18] Roman Vershynin. *High-dimensional probability*. Vol. 47. Cambridge Series in Statistical and Probabilistic Mathematics. An introduction with applications in data science, With a foreword by Sara van de Geer. Cambridge University Press, Cambridge, 2018, pp. xiv+284.

[Vil03] Cédric Villani. *Topics in optimal transportation*. Vol. 58. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2003, pp. xvi+370.

[Vil09a] Cédric Villani. "Hypocoercivity". In: *Mem. Amer. Math. Soc.* 202.950 (2009), pp. iv+141.

[Vil09b] Cédric Villani. *Optimal transport*. Vol. 338. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Old and new. Springer-Verlag, Berlin, 2009, pp. xxii+973.

[Vin11] Pascal Vincent. "A connection between score matching and denoising autoencoders". In: *Neural Comput.* 23.7 (2011), pp. 1661–1674.

[VKK21] Arash Vahdat, Karsten Kreis, and Jan Kautz. "Score-based generative modeling in latent space". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 11287–11302.

[VPD22] Maxime Vono, Daniel Paulin, and Arnaud Doucet. "Efficient MCMC sampling with dimension-free convergence rate using ADMM-type splitting". In: *J. Mach. Learn. Res.* 23 (2022), Paper No. [25], 69.

[VW19] Santosh Vempala and Andre Wibisono. "Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 8094–8106.

[VZ00] Yehuda Vardi and Cun-Hui Zhang. "The multivariate L_1-median and associated data depth". In: *Proc. Natl. Acad. Sci. USA* 97.4 (2000), pp. 1423–1426.

[Wai19] Martin J. Wainwright. *High-dimensional statistics*. Vol. 48. Cambridge Series in Statistical and Probabilistic Mathematics. A non-asymptotic viewpoint. Cambridge University Press, Cambridge, 2019, pp. xvii+552.

[Wal16] Stephen G. Walker. "Bayesian information in an experiment and the Fisher information distance". In: *Statist. Probab. Lett.* 112 (2016), pp. 5–9.

[Wan00] Feng-Yu Wang. "Functional inequalities for empty essential spectrum". In: *J. Funct. Anal.* 170.1 (2000), pp. 219–245.

[Wan05] Feng-Yu Wang. "A generalization of Poincaré and log-Sobolev inequalities". In: *Potential Anal.* 22.1 (2005), pp. 1–15.

[Wan14] Liyao Wang. *Heat capacity bound, energy fluctuations and convexity*. Thesis (Ph.D.)–Yale University. ProQuest LLC, Ann Arbor, MI, 2014, p. 114.

[WB19] Yixin Wang and David M. Blei. "Frequentist consistency of variational Bayes". In: *J. Amer. Statist. Assoc.* 114.527 (2019), pp. 1147–1161.

[Wei04] Dror Weitz. *Mixing in time and space for discrete spin systems*. Thesis (Ph.D.)–University of California, Berkeley. ProQuest LLC, Ann Arbor, MI, 2004, p. 175.

[Wib18] Andre Wibisono. "Sampling as optimization in the space of measures: the Langevin dynamics as a composite optimization problem". In: *Proceedings of the 31st Conference on Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 2093–3027.

[Wib19] Andre Wibisono. "Proximal Langevin algorithm: rapid convergence under isoperimetry". In: *arXiv preprint 1911.01469* (2019).

[Wie12] Ami Wiesel. "Geodesic convexity and covariance estimation". In: *IEEE Trans. Signal Process.* 60.12 (2012), pp. 6182–6189.

[WJ08] Martin J. Wainwright and Michael I. Jordan. "Graphical models, exponential families, and variational inference". In: *Foundations and Trends® in Machine Learning* 1.1-2 (2008), pp. 1–305.

[WL20] Yifei Wang and Wuchen Li. "Information Newton's flow: second-order optimization method in probability space". In: *arXiv preprint 2001.04341* (2020).

[WS17] Blake Woodworth and Nathan Srebro. "Lower bound for randomized first order convex optimization". In: *arXiv preprint 1709.03594* (2017).

[WS22] Melanie Weber and Suvrit Sra. "Projection-free nonconvex stochastic optimization on Riemannian manifolds". In: *IMA J. Numer. Anal.* 42.4 (2022), pp. 3241–3271.

[WSC22] Keru Wu, Scott Schmidler, and Yuansi Chen. "Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling". In: *Journal of Machine Learning Research* 23.270 (2022), pp. 1–63.

[Wu00] Liming Wu. "Uniformly integrable operators and large deviations for Markov processes". In: *J. Funct. Anal.* 172.2 (2000), pp. 301–376.

[WW16] Feng-Yu Wang and Jian Wang. "Functional inequalities for convolution probability measures". In: *Ann. Inst. Henri Poincaré Probab. Stat.* 52.2 (2016), pp. 898–914.

[Xu+18] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. "Global convergence of Langevin dynamics based algorithms for nonconvex optimization". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018.

[Yan10] Le Yang. "Riemannian median and its estimation". In: *LMS J. Comput. Math.* 13 (2010), pp. 461–479.

[YBL17] Florian Yger, Maxime Berar, and Fabien Lotte. "Riemannian approaches in brain-computer interfaces: a review". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.10 (2017), pp. 1753–1762.

[YS22] Jiayuan Ye and Reza Shokri. "Differentially private learning needs hidden state (or much faster convergence)". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho. 2022.

[ZC23] Qinsheng Zhang and Yongxin Chen. "Fast sampling of diffusion models with exponential integrator". In: *The Eleventh International Conference on Learning Representations*. 2023.

[Zeg01] Bogusław Zegarliński. "Entropy bounds for Gibbs measures with non-Gaussian tails". In: *J. Funct. Anal.* 187.2 (2001), pp. 368–395.

[Zha+18] Guodong Zhang, Shengyang Sun, David Duvenaud, and Roger Grosse. "Noisy natural gradient as variational inference". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 5852–5861.

[Zha+20] Kelvin S. Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. "Wasserstein control of mirror Langevin Monte Carlo". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 3814–3841.

[Zha+23] Matthew Zhang, Sinho Chewi, Mufan (Bill) Li, Krishnakumar Balasubramanian, and Murat A. Erdogdu. "Improved discretization analysis for underdamped Langevin Monte Carlo". In: *arXiv preprint 2302.08049* (2023).

[Zim13] David Zimmermann. "Logarithmic Sobolev inequalities for mollified compactly supported measures". In: *J. Funct. Anal.* 265.6 (2013), pp. 1064–1083.

[Zim16] David Zimmermann. "Elementary proof of logarithmic Sobolev inequalities for Gaussian convolutions on $\mathbb{R}$". In: *Ann. Math. Blaise Pascal* 23.1 (2016), pp. 129–140.

[ZP19] Yoav Zemel and Victor M. Panaretos. "Fréchet means and Procrustes analysis in Wasserstein space". In: *Bernoulli* 25.2 (2019), pp. 932–976.

[ZS16] Hongyi Zhang and Suvrit Sra. "First-order methods for geodesically convex optimization". In: *29th Annual Conference on Learning Theory*. Ed. by Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir. Vol. 49. Proceedings of Machine Learning Research. Columbia University, New York, New York, USA: PMLR, June 2016, pp. 1617–1638.

[ZWG13] Teng Zhang, Ami Wiesel, and Maria Sabrina Greco. "Multivariate generalized Gaussian distribution: convexity and graphical models". In: *IEEE Trans. Signal Process.* 61.16 (2013), pp. 4141–4148.

[ZXG21] Difan Zou, Pan Xu, and Quanquan Gu. "Faster convergence of stochastic gradient Langevin dynamics for non-log-concave sampling". In: *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*. Ed. by Cassio de Campos and Marloes H. Maathuis. Vol. 161. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 1152–1162.