Upper and Lower Bounds for Sampling

by

Chen Lu

B.S., Stanford University, 2019

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

September 2023

Author . . . .
Department of Mathematics
August 18, 2023

Certified by . . . .
Philippe Rigollet
Professor of Mathematics
Thesis Supervisor

Accepted by . . . .
Jonathan Kelner
Graduate Chair, Applied Mathematics

Upper and Lower Bounds for Sampling

by

Chen Lu

Submitted to the Department of Mathematics
on August 18, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis studies the problem of drawing samples from a probability distribution. Despite the prevalence of sampling problems in applications, the quantitative behavior of sampling algorithms remains poorly understood. This thesis contributes to the theoretical understanding of sampling by giving upper bounds and more importantly lower bounds for various sampling algorithms and problem classes. On the upper bound side, we propose new sampling algorithms, motivated by the perspective of sampling as optimization [JKO98], and give convergence guarantees for them. We also obtain state-of-the-art convergence results for the popular Metopolis-Adjusted Langevin Algorithm. On the lower bound side, we establish the query complexity for strongly log-concave sampling in all constant dimensions. Our lower bounds rely on simple geometric constructions, which can hopefully be of aid to similar results in high dimensions.

Thesis Supervisor: Philippe Rigollet
Title: Professor of Mathematics

Acknowledgements

First I must thank my advisor Philippe Rigollet. I did not start graduate school at MIT, and I initially reached out with nothing other than a vague sense that this could be interesting. Thanks to Philippe and the environment he created, I had a terrific time. Philippe made a big impact on my taste: he showed me that the best way to go about things is to optimize for fun in a serious way. He has both been generous with his time and energy, and given me a lot of freedom. As a result, my experience has accidentally converged to what Paul Graham called "the perfect life" for a graduate student [Gra05, Note 5]. Thanks Philippe, for taking a chance on me, and for helping me find my way.

I thank the other members of my committee, Subhabrata Sen and Jonathan Kelner. I was lucky to have Subhabrata as a mentor during my first year. I am very grateful for his time and support, and I thank him for the help that led to an exciting and welcoming first year.

I thank Bill Minicozzi and Larry Guth for being on my qualifying exam committee, and for being exceptionally great teachers. Through our brief interactions, Larry affected the way I think about mathematical ideas and how I communicate.

I'm lucky to have had Sinho Chewi as my main collaborator in graduate school, who has been a great friend and a role model. I thank Sinho for his endless enthusiasm and clarity of thought.

I thank all my other friends in graduate school. I thank Patrik for the dry humor and and succinct insights. I thank Austin for being another closet systems enthusiast. I thank George Stepaniants for sharing an interest in mathematical applications. I thank Felipe Suarez for the blitz games. I thank Enric Boix-Adsera for being a great travel companion. I thank Jaume de Dois Pont and Shyam Narayanan who helped further the lower bound dream. I thank Jiaze Qiu and David Ham for the pool and the late night burgers. I thank Anish Athalye, Assel Ismoldayeva and Zain Ruan for tolerating my very basic questions about systems. I thank the other participants of the grupito, Tyler Manu, Thibaut Le Gouic, Kwangju Ann, and Xiang Cheng. Much of the work of this thesis grew from those early meetings. I will cherish the time we spend together, computing Wasserstein gradients and Hessians on the board, and getting by on Zoom through the pandemic.

During my third and fourth years, I took a detour to explore problems in biology, and I couldn't have done it without the help of many people. I thank the Eric and Wendy Schmidt center, Caroline Uhler and Anthony Philippakis, and the other EWSC fellows, for their support, and for being people who speak the same language. I thank Anthonimos Vardis Kandiros, the other theory person in our cohort, for the sessions we had toying with math problems. I thank Mehrtash Babadi for the fun ML discussions. I thank Arnav Mehta

for generously explaining biology, sharing his data, and letting me be part of a real biology project. I thank Jason Buenrostro, Yan Hu, and the rest of the Buenrostro lab for helping me learn about ATAC-seq and supplying me with data. I thank the entire Chen lab for hosting me for a rotation. I thank Ehsan Habibi for sharing his excitement about developmental biology, Dylan Cable for his generosity in time and ideas, Sophia Liu for teaching me about microscopes and improving work-life balance for everyone through the IM teams. I also thank Fei Chen, who has been incredibly generous with his time, endlessly excited about science, and has also influenced the factors I consider when choosing a problem.

I also thank a number of people from before graduate school who have influenced me. I thank Andrea Montanari and John Duchi for being inspiring lecturers and mentors. I thank Stefan Wager for helping me with my first research project, for teaching me what statistics is about, and for showing me what is good writing. I thank Xinkun Nie for her generous support and for being an inspirational graduate student. I thank Doris Mai and Mansur Tisaev for being my other friends who understand the highs and lows of grad school.

I thank my parents, for giving me everything, and for letting me have things my way, though sometimes somewhat reluctantly. I thank my sister, for being the one who gets it.

Contents

Chapter 1

Introduction

The standard setting of sampling is as follows. We have a target distribution π, often supported on $\mathbb{R}^d$. The density of the target is given by $\pi(x) = \exp(-V(x))$, where $V : \mathbb{R}^d \to \mathbb{R}$ is the potential function. We are given access to V and its derivatives, often only up to an additive constant, and we want to generate samples in $\mathbb{R}^d$ whose distribution is close to π in some measure, for instance in total variation distance.

Sampling algorithms are widely used in many different areas, such as scientific computing, statistics, machine learning, and generative modelling. Yet our understanding of sampling algorithms remains limited. There are many algorithms that run faster in practice than what theory guarantees, and many heuristics among practitioners that we cannot justify rigorously (e.g. "increase the movement of the Markov chain"). There is also a complete lack of lower bounds, which means that we can't tell for a given class of target distributions which algorithm will perform the best. These are the questions that motivated this thesis.

Another motivation comes from a beautiful connection, first introduced in the seminal paper of Jordan, Kinderlehrer and Otto [JKO98], that the marginal distributions of a Langevin diffusion (LD) evolves as a gradient flow of the Kullback-Leibler (KL) divergence over the Wasserstein space of probability measures (a review of the theory is given in 2.1). This perspective of sampling as optimization naturally leads one to wonder whether we can use the extensive optimization toolkit to develop new sampling algorithms and fill in the aforementioned gaps in our understanding of sampling.

This thesis is organized in two parts. The first part is focused on new sampling algorithms and upper bound guarantees, and the second part is focused on sampling lower bounds. We now give a summary of the thesis that provides context and motivation and outlines the main contributions.

1.1 Summary of contributions

In Part I of the thesis, consisting of Chapters 3 to 5, we start with the connection between sampling and optimization and ask: can we borrow ideas from optimization to propose new sampling algorithms? Can we use optimization techniques to obtain sharper convergence guarantees for existing algorithms? Since Langevin dynamics is the analogue of gradient descent in sampling, are there sampling analogues of the other optimization algorithms as well, such as proximal/splitting methods, mirror descent, Nesterov's accelerated gradient descent, and Newton methods?

Chapter 3 deals with mirror-Langevin diffusions, or the sampling analogue of mirror descent. These stochastic processes were introduced in [Zha+20b]. We give a clean non-asymptotic convergence analysis of the continuous time mirror-Langevin diffusion. As a special case of such processes, we propose the Newton-Langevin diffusion and prove that it converges to stationarity exponentially with a rate which has no dependence on the target distribution. We give an application of this result to the problem of sampling from the uniform distribution on a convex body, using a strategy inspired by interior-point methods. Our techniques also yield new results on the convergence of the Langevin diffusion in Wasserstein distance.

The investigation of Langevin and mirror-Langevin diffusions lead to two natural questions. First, both these diffusion processes are flows that minimize the KL divergence, but are there sampling algorithms that minimize different functionals in Wasserstein space? Second, sampling algorithms based on diffusions are random, but there is no randomness in the evolution of the marginal distributions along a Wasserstein gradient flow. Is there a way to discretize Wasserstein gradient flows that avoids the randomness by directly evolving approximations of the marginal distributions?

In Chapter 4, we study Stein Variational Gradient Descent (SVGD), proposed by [LW16], which is a major advance that gives a deterministic sampling algorithm. SVGD is often described as the kernelized gradient flow for the KL divergence. We introduce a new perspective on SVGD that instead views it as a kernelized gradient flow of a different functional: the chi-squared divergence. We show that the gradient flow of the chi-squared divergence converges exponentially under conditions as weak as a Poincaré inequality. This perspective leads us to propose an alternative to SVGD, called Laplacian Adjusted Wasserstein Gradient Descent (LAWGD), that can be implemented from the spectral decomposition of the Laplacian operator of the target distribution. We obtain strong convergence guarantees for the continuous version of LAWGD, and demonstrate that it has good practical performance when the Laplacian operator is computationally tractable.

The prior two chapters only analyse the continuous Wasserstein gradient flows without getting into the details of discretization. In Chapter 5, we tackle the problem of discretization head on, and analyze the most practical and widely used discretization of the Langevin diffusion, the Metropolis-Adjusted Langevin Algorithm (MALA). Conventional wisdom in the sampling literature, backed by a popular diffusion scaling limit [RS02], suggests that the mixing time of MALA scales as $O(d^{1/3})$, where d is the dimension. However, the scaling limit requires stringent assumptions on the target distribution and is asymptotic in nature. In contrast, prior to our work the best known non-asymptotic mixing time bound for MALA on the class of log-smooth and strongly log-concave distributions is $O(d)$. We establish that the mixing time of MALA on this class is $\widetilde{\Theta}(d^{1/2})$ under a warm start. Our upper bound proof introduces a new technique based on a projection characterization of the Metropolis adjustment [BD01], which reduces the analysis of MALA to the well-studied unadjusted Langevin discretization, and bypasses direct computation of the acceptance probability.

Our lower bound for MALA is obtained by constructing a specific class log-smooth and strongly log-concave distributions, and upper bounding the spectral gap for MALA on them. While it shows that our analysis of MALA is tight, it does not tell us anything about what the lower bound should be for general sampling algorithms. It is then natural to wonder about the more general questions: for some given class of distributions, what is the fundamental limit on the sampling performance that can be achieved by any sampling algorithm? What is the best algorithm that matches that limit for that given class? Such questions will be the focus for Part II. There has been no work done on the question of general sampling lower bounds prior to the works in this thesis. Chapter 6 gives an overview of the context and ideas that motivate the assumptions made and the techniques used in obtaining the lower bounds.

In Chapter 7, we explore the simplest idea for proving a sampling lower bound: reduce sampling to optimization. As we will discuss in Chapter 6, this reduction to optimization does not give very interesting results for log-concave sampling. But for non log-concave sampling, when we measure the sampling progress by the Fisher information to the target distribution, we are able to obtain non-trivial results. When the desired accuracy level is low, we prove a lower bound that is matched by the performance of averaged Langevin Monte Carlo, proposed by [Bal+22]. When the desired accuracy level is high, we show that sampling from Fisher information at most ε 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.

Chapter 8 investigates the lower bound question in a toy setting, where we prove sampling lower bounds for rejection sampling algorithms only, and only on discrete distributions. We show that for discrete distributions with structural constraints, such as being monotone or discrete log-concave, the rejection sampling complexity is sublinear in the alphabet size of the support.

The results we obtain are simple, but the main insights, that we should use an information theoretic argument, and that we should construct distributions that are hard to distinguish with queries but easy to distinguish with samples, turn out to be fruitful for proving stronger sampling lower bounds.

In Chapter 9, using the ideas mentioned, we give the first general sampling lower bound of $\Omega(\log\log\kappa)$ on the query complexity of sampling from strongly log-concave and log-smooth distributions with condition number κ in one dimension. The lower bound is tight and achieved by rejection sampling. In Chapter 10, we use the same framework to prove that sampling from strongly log-concave and log-smooth distributions in dimension $d = 2$ requires $\Omega(\log\kappa)$ queries. This lower bound is tight in all fixed $d \geq 2$ dimensions, and it is also achieved by rejection sampling. Compared to the one dimensional case, the two dimensional construction is much more involved, and is inspired by work on the Kakeya conjecture in harmonic analysis.

The goal of general sampling lower bounds in high dimensions remains out of reach in this thesis. But for the subclass of Gaussian distributions, we are able to obtain dimension dependent lower bounds, which are presented in Chapter 11. We show that sampling from Gaussians with condition number κ in dimension d requires $\widetilde{\Omega}(\min(\sqrt{\kappa}\log d, d))$ queries. We also give an upper bound algorithm that achieves a rate of $O(\min(\sqrt{\kappa}\log d/\varepsilon, d))$, which shows that the lower bounds are nearly tight. The algorithm samples by taking linear combinations of a standard Gaussian vector with past query points and gradient queries. Since Gaussians are strongly log-concave, this implies that sampling from strongly log-concave distributions will have dimension dependence, which is notable in contrast to optimization, where optimizing strongly convex functions have dimension free rates. Our results suggest that similar to optimization, log-concave sampling separates into two regimes: the low dimensional and high dimensional. In the low dimensional regime, the optimal log-concave sampling algorithm is rejection sampling, which can be thought of as the analogue of the ellipsoid method in optimization. In the high dimensional regime, our results for Gaussian sampling suggests that the optimal algorithm is likely to be a gradient-based algorithm, similar in form to the Langevin algorithm or Hamiltonian Monte Carlo, which is analogous to gradient descent and accelerated gradient descent in optimization.

Chapter 2

Background

2.1 Sampling and optimization

Throughout this thesis, we will consider the following sampling problem: Let π be a log-concave probability measure over $\mathbb{R}^d$ so that π has density equal to e^{-V}. Often we will consider the case where the potential $V : \mathbb{R}^d \to \mathbb{R}$ is convex. In practical sampling scenarios, the case where the potential V is non-convex is the more challenging and prevalent one. But the difficulty in non-convex sampling is inherently tied to the difficulty of non-convex optimization, specifically the difficulty of locating the modes of the distribution, which makes analysis complicated. By studying the cases where the potential is convex, we are often able to isolate the difficulty that comes from sampling alone.

There is a deep connection between sampling and optimization that can be seen by considering 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 $\mathbb{R}^d$. This gives rise to a sampling algorithm for π, because 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.

Looking at (LD), we see that if we drop the Brownian motion term on the right hand side, we are left with the dynamics $\dot{x}_t = -\nabla V(x_t)$, known as the *gradient flow*. When the function V is convex, the gradient flow will converge to the unique minimizer of V, and a suitable time discretization of this dynamics yields the well-known *gradient descent* (GD) algorithm for optimization.

But the connection with optimization goes deeper. If we consider the marginal distribution μ_t of X_t, then it turns out that μ_t evolves according to a *gradient flow* over the Wasserstein space of probability measures that minimizes the Kullback-Leibler (KL) divergence $\mathsf{KL}(\cdot \,\|\, \pi)$. This perspective was introduced in the seminal paper of Jordan, Kinderlehrer and Otto[JKO98], and it is now commonly known as the *JKO scheme*. This point of view has not only led to a better theoretical understanding of sampling algorithms based on the Langevin diffusion [Ber18; CB18; Wib18; DMM19; VW19], but also fueled the discovery of new MCMC algorithms inspired by the diverse and powerful optimization toolbox [**wibisonowibisono2019proximal**; Mar+12; Sim+16; Che+18b; Ber18; Hsi+18; Wib18; Ma+19; Che+20b; DR20; Zha+20b]. We will now give some background on the theory of Wasserstein gradient flows and explain what we mean more precisely.

Wasserstein gradient flows

Let $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ of probability measures absolutely continuous w.r.t. Lebesgue measure and possessing a finite second moment, equipped with the 2-Wasserstein metric W_2. We refer readers to [Vil03; San15; San17] for introductory treatments of optimal transport, and to [AGS08; Vil09] for detailed treatments of Wasserstein gradient flows.

Let $F : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ be a functional defined on Wasserstein space. We say that a curve $(\mu_t)_{t\ge 0}$ of probability measures is a *Wasserstein gradient flow* for the functional F if it satisfies

$$\partial_t \mu_t = \operatorname{div}\bigl(\mu_t \nabla_{W_2} F(\mu_t)\bigr) \tag{2.1}$$

in a weak sense. Here, $\nabla_{W_2} F(\mu) := \nabla \delta F(\mu)$ is the Wasserstein gradient of the functional F at μ, where $\delta F(\mu) : \mathbb{R}^d \to \mathbb{R}$ is the *first variation* of F at μ, defined by

$$\lim_{\varepsilon \to 0} \frac{F(\mu + \varepsilon \xi) - F(\mu)}{\varepsilon} = \int \delta F(\mu) \, \mathrm{d}\xi, \qquad \text{for all } \xi \text{ with } \int \mathrm{d}\xi = 0,$$

and ∇ denotes the usual (Euclidean) gradient. Hence, the Wasserstein gradient, at each $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, is a map from $\mathbb{R}^d$ to $\mathbb{R}^d$.

Using the *continuity equation*, we can give an Eulerian interpretation to the evolution equation (2.1) (see [San15, §4] and [AGS08, §8]). Given a family of vector fields $(v_t)_{t\ge 0}$, let $(X_t)_{t\ge 0}$ be a curve in $\mathbb{R}^d$ with random initial point $X_0 \sim \mu_0$, and such that $(X_t)_{t\ge 0}$ is an integral curve of the vector fields $(v_t)_{t\ge 0}$, that is, $\dot{X}_t = v_t(X_t)$. If we let μ_t denote the law of X_t, then $(\mu_t)_{t\ge 0}$ evolves according to the *continuity equation*

$$\partial_t \mu_t = -\operatorname{div}(\mu_t v_t). \tag{2.2}$$

Comparing (2.1) and (2.2), we see that (2.1) describes the evolution of the marginal law $(\mu_t)_{t\ge 0}$ of the curve $(X_t)_{t\ge 0}$ with $X_0 \sim \mu_0$ and $\dot{X}_t = -[\nabla_{W_2} F(\mu_t)](X_t)$.

Wasserstein calculus provides the following (formal) calculation rule: the Wasserstein gradient flow $(\mu_t)_{t\ge 0}$ for the functional F dissipates F at the rate $\partial_t F(\mu_t) = -\mathbb{E}_{\mu_t}[\|\nabla_{W_2} F(\mu_t)\|^2]$. More generally, for a curve $(\mu_t)_{t\ge 0}$ evolving according to the continuity equation (2.2), the time-derivative of F is given by $\partial_t F(\mu_t) = \mathbb{E}_{\mu_t}\langle \nabla_{W_2} F(\mu_t), v_t\rangle$.

It turns out that Langevin dynamics is precisely the gradient flow where the functional is the Kullback-Leibler (KL) divergence $\mathsf{KL}(\cdot \,\|\, \pi)$ [JKO98; AGS08; Vil09]. This can be checked by showing that the Wasserstein gradient of the KL divergence [AGS08; San15] is given by

$$\big(\nabla_{W_2} D_{\mathrm{KL}}(\cdot \,\|\, \pi)\big)(\mu) = \nabla \ln \frac{\mathrm{d}\mu}{\mathrm{d}\pi}. \tag{2.3}$$

Substituting the gradient expression into the continuity equation (2.2), and the resulting PDE will be the distributional form of the Langevin dynamics.

Upper bound techniques

The techniques for obtaining convergence guarantees in sampling also closely resemble those used in optimization. Consider first the gradient flow for f: $\dot{x}_t = -\nabla f(x_t)$. We get $\partial_t[f(x_t) - f(x^*)] = -\|\nabla f(x_t)\|^2$ from a straightforward computation. 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 [Che+20c]. 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: $f(x_t) \le f(x_0)\, e^{-2t/C_{\mathsf{PL}}}$.

If we consider the Langevin dynamics, since it is the gradient flow of the KL divergence, using the expression for the gradient in (2.3) we obtain [Vil03, §9.1.5]

$$\partial_t D_{\mathrm{KL}}(\mu_t \,\|\, \pi) = -\int \Big\|\nabla \ln \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\Big\|^2 \,\mathrm{d}\mu_t = -4\int \Big\|\nabla \sqrt{\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}}\Big\|^2 \,\mathrm{d}\pi. \tag{2.4}$$

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 $D_{\mathrm{KL}}(\mu_t \parallel \pi) \le 2C_{\mathsf{LSI}} \int \big\|\nabla \sqrt{\mathrm{d}\mu_t/\mathrm{d}\pi}\big\|^2 \,\mathrm{d}\pi\,,$ which implies exponentially fast convergence: $D_{\mathrm{KL}}(\mu_t\|\pi) \le D_{\mathrm{KL}}(\mu_0\|\pi)\, e^{-2t/C_{\mathsf{LSI}}}$ by Grönwall's inequality.

2.2 Information theory and lower bounds

The connections mentioned between sampling and optimization leads one to wonder whether techniques for proving optimization lower bounds can also translate to sampling lower bounds, but as we will discuss in Chapter 6, this approach did not yeild results for us. Instead, the main technique we use in our lower bounds will be the well known Fano's inequality from information theory, which we review below. For background on information theory, see [CT06, §2].

The *mutual information* between two discrete random variables X and Y is defined as

$$I(X;Y) = \mathsf{KL}(P_{X,Y} \parallel P_X \otimes P_Y)\,,$$

where $P_{X,Y}$ is the joint distribution of X and P_X and P_Y are the marginal distributions. The mutual information is also the difference between the marginal entropy and the conditional entropy:

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)\,.$$

Fano's inequality is used in the setting where we have a discrete random variable X, we observe some random variable ξ that is correlated to X, and we come up with an estimator $\widehat{X}$ for X using only the data ξ. Specifically, we assume that $X \to \xi \to \widehat{X}$ forms a Markov chain. Then we have the following.

Theorem 1 (Fano's inequality, [CT06, Theorem 2.10.1])**.** *Suppose that $\mathcal{S}$ is a finite set and $X \sim \mathsf{uniform}(\mathcal{S})$. Suppose that $\widehat{X}$ is any estimator which is based on some data ξ, such that $X \to \xi \to \widehat{X}$. Then,*

$$\mathbb{P}\{\widehat{X} \neq X\} \ge 1 - \frac{I(\xi;X) + \log 2}{\log|\mathcal{S}|}\,.$$

Part I

Sampling Upper Bounds

Chapter 3

Mirror Langevin

3.1 Introduction

This chapter focuses on the sampling analogue of mirror descent. Specifically, we analyse the class of stochastic processes called *mirror-Langevin diffusions* (MLD) introduced in [Zha+20b]. In contrast to the Langevin diffusion(LD), 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.

A surprising aspect of our analysis is that we track the progress of these schemes not by measuring the objective function itself, the KL divergence, 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 1) and can be used for approximate sampling even when the second derivative of V vanishes (Corollary 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 and is the first result of this kind for sampling.

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 [LTV20]), to output a sample which is approximately uniformly distributed on $\mathcal{C}$; see Corollary 3.

Throughout this chapter, we work exclusively in the setting of continuous-time diffusions such as (LD). We refer to the works [DM15; Dal17a; Dal17b; RRT17; CB18; Wib18; DK19; DMM19; DRK19; Mou+19; VW19] for discretization error bounds for Langevin diffusion, and to [AC21] for discretization bounds for the mirror-Langevin diffusion.

This chapter is based on the joint work [Che+20b], with Sinho Chewi, Thibaut Le Gouic, Tyler Maunu, Philippe Rigollet, and Austin J. Stromme.

Related works. 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; Che+19; DK19; DM+19; Dwi+19; Mou+19; VW19]. Moreover, various ways of adapting Langevin diffusion to sample from bounded domains have been proposed [BEL18; Hsi+18; Zha+20b]; in particular, [Zha+20b] studied the discretized mirror-Langevin diffusion.

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 $\operatorname{var}_\mu g := \int (g - \mathbb{E}_\mu\, g)^2\,\mathrm{d}\mu$. When clear from context, we sometimes abuse notation by identifying a measure μ with its Lebesgue density.

3.2 Mirror-Langevin diffusions

Before introducing mirror-Langevin diffusions, our main objects of interest, we provide some intuition for their construction by drawing a parallel with convex optimization.

Gradient flows, mirror flows, and Newton's method

We briefly recall some background on 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 Nemirovskii and Yudin [NJ79] 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 chapter, 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].

Mirror-Langevin diffusions

We now introduce the *mirror-Langevin diffusion* (MLD) of [Zha+20b]. 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

[1] This ensures that $\nabla \phi$ is invertible, c.f. [Roc97, §26].

section. Then, the mirror-Langevin diffusion satisfies the SDE

$$X_t = \nabla\phi^\star(Y_t), \qquad \mathrm{d}Y_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,[\nabla^2\phi(X_t)]^{1/2}\,\mathrm{d}B_t\,, \quad \text{(MLD)}$$

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),

$$X_t = \nabla V^\star(Y_t), \qquad \mathrm{d}Y_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,[\nabla^2 V(X_t)]^{1/2}\,\mathrm{d}B_t. \quad \text{(NLD)}$$

From our intuition gained from optimization, we expect that NLD has special properties, such as affine invariance and faster convergence. We validate this intuition in Corollary 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 end this section by comparing 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+20b] for further comparison of the two algorithms.

An earlier draft of [Hsi+18] also introduced MLD, along with a continuous-time analysis of the diffusion. Their convergence analysis is based on the classical Bakry-Émery criterion (see [BGL14]), which is generally harder to check than the mirror Poincaré inequality (MP) that we introduce below; in particular, when $\phi = V$, we show that the mirror Poincaré inequality holds automatically.

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 Section 3.6. Moreover, since we analyse 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.

3.3 Convergence analysis

Convergence of gradient flows and mirror flows

We provide a brief reminder about the convergence analysis of gradient flows and mirror flows defined in Section 3.2 to provide intuition for the next section. Throughout, let f be a differentiable function with minimizer x^*.

Recall from Section 2.1 that the gradient flow for f, $\dot{x}_t = -\nabla f(x_t)$, is analysed as follows: we start from the identity $\partial_t[f(x_t) - f(x^*)] = -\|\nabla f(x_t)\|^2$, then apply the PL inequality (PL) to control the norm of the gradient on the right.

A similar analysis may be carried out for the mirror flow. Fix a mirror map ϕ and consider the mirror flow: $\dot{x}_t = -[\nabla^2\phi(x_t)]^{-1}\nabla f(x_t)$. It holds $\partial_t[f(x_t) - f(x^*)] = -\langle \nabla f(x_t), [\nabla^2\phi(x_t)]^{-1}\nabla f(x_t)\rangle$. 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.

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 $\phi = \|\cdot\|^2/2$.

As discussion in Section 2.1, the Langevin dynamics (LD) is a gradient from the the KL divergence with respect to the 2-Wasserstein distance. Hence the decay in the KL divergence along (LD) is also given by the norm of the

Wasserstein gradient, as in (2.4). In this setup, the role of the PL inequality (PL) is played by a *log-Sobolev inequality* (LSI).

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*

$$F(\mu) = \chi^2(\mu \parallel \pi) := \operatorname{var}_\pi \frac{\mathrm{d}\mu}{\mathrm{d}\pi} = \int \Bigl(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\Bigr)^2 \mathrm{d}\pi - 1, \qquad \text{if } \mu \ll \pi\,,$$

and $F(\mu) = \infty$ otherwise. It is well-known that the law $(\mu_t)_{t\geq 0}$ of LD satisfies the Fokker-Planck equation in the weak sense [KS91, §5.7]:

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

Using this, we can compute the derivative of the chi-squared divergence:

$$\frac{1}{2}\partial_t F(\mu_t) = \int \frac{\mu_t}{\pi}\,\partial_t\mu_t = \int \frac{\mu_t}{\pi} \operatorname{div}\bigl(\mu_t \nabla \ln \frac{\mu_t}{\pi}\bigr) = -\int \bigl\langle \nabla \ln \frac{\mu_t}{\pi}, \nabla \frac{\mu_t}{\pi}\bigr\rangle\, \mu_t = -\int \bigl\|\nabla \frac{\mu_t}{\pi}\bigr\|^2\, \pi\,,$$

and exponential convergence of the chi-squared divergence follows if π satisfies a Poincaré inequality:

$$\operatorname{var}_\pi g \leq C_{\mathsf{P}}\, \mathbb{E}_\pi[\|\nabla g\|^2] \qquad \text{for all locally Lipschitz } g \in L^2(\pi). \tag{P}$$

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 Sections 3.4 and 3.4 below, the Poincaré inequality is significantly weaker than the log-Sobolev inequality.

A similar analysis may be carried out for MLD using an appropriate variation of Poincaré inequalities.

Definition 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*

$$\operatorname{var}_\pi g \leq C_{\mathsf{MP}}\, \mathbb{E}_\pi \langle \nabla g, (\nabla^2\phi)^{-1} \nabla g\rangle \qquad \textit{for all locally Lipschitz } g \in L^2(\pi). \tag{MP}$$

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 $\chi^2(\cdot \,\|\, \pi)$ 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 $\|\cdot - \pi\|_{\mathrm{TV}}$, the Hellinger distance $H(\cdot, \pi)$, and the KL divergence $D_{\mathrm{KL}}(\cdot \,\|\, \pi)$.

Theorem 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*

$$2\|\mu_t - \pi\|_{\mathrm{TV}}^2,\ H^2(\mu_t, \pi),\ D_{\mathrm{KL}}(\mu_t \,\|\, \pi),\ \chi^2(\mu_t \,\|\, \pi) \leq e^{-\frac{2t}{C_{\mathsf{MP}}}} \chi^2(\mu_0 \,\|\, \pi), \quad \forall\, t \geq 0\,.$$

We give two proofs of this result below. Recall the law $(\mu_t)_{t\geq 0}$ of MLD satisfies the Fokker-Planck equation

$$\partial_t \mu_t = \operatorname{div}\bigl(\mu_t\, (\nabla^2\phi)^{-1} \nabla \ln \frac{\mu_t}{\pi}\bigr). \tag{3.1}$$

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].

Proof of Theorem 2. Using the Fokker-Planck equation (3.1), we may compute

$$\begin{aligned}\partial_t \chi^2(\mu_t \,\|\, \pi) &= \partial_t \int \frac{\mu_t^2}{\pi} = 2 \int \frac{\mu_t}{\pi}\, \partial_t \mu_t = 2 \int \frac{\mu_t}{\pi}\, \operatorname{div}\bigl(\mu_t\, (\nabla^2\phi)^{-1} \nabla \ln \frac{\mu_t}{\pi}\bigr) \\ &= -2 \int \bigl\langle \nabla \frac{\mu_t}{\pi}, (\nabla^2\phi)^{-1} \nabla \ln \frac{\mu_t}{\pi} \bigr\rangle\, \mu_t = -2 \int \bigl\langle \nabla \frac{\mu_t}{\pi}, (\nabla^2\phi)^{-1} \nabla \frac{\mu_t}{\pi} \bigr\rangle\, \pi.\end{aligned}$$

The mirror Poincaré inequality (MP) implies that this quantity is at most $-2C_{\mathsf{MP}}^{-1} \chi^2(\mu_t \,\|\, \pi)$, which completes the proof via Grönwall's inequality. □

We may reinterpret this proof within Markov semigroup theory.

Proof of Theorem 2 from a Markov semigroup perspective. We denote by $(P_t)_{t\geq 0}$ the semigroup of MLD; 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 $f \in L^2(\pi)$,

$$\int P_t\bigl(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}\bigr) f \,\mathrm{d}\pi = \int \bigl(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}\bigr) P_t f \,\mathrm{d}\pi = \int P_t f \,\mathrm{d}\mu_0 = \int f \,\mathrm{d}\mu_t = \int \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} f \,\mathrm{d}\pi\ ,$$

so that

$$P_t(\frac{\mu_0}{\pi}) = \frac{\mu_t}{\pi}.$$

Therefore,

$$\chi^2(\mu_t \,\|\, \pi) := \mathrm{var}_\pi(\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}) = \mathrm{var}_\pi\, P_t(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}).$$

Using a classical result of Markov semigroup theory (see for instance [CG09, Theorem 2.1] or [BGL14, Theorem 4.2.5])

$$\chi^2(\mu_t \,\|\, \pi) = \mathrm{var}_\pi\, P_t(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}) \le e^{-\frac{2t}{C}}\,\mathrm{var}_\pi(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}) = e^{-\frac{2t}{C}}\chi^2(\mu_0 \,\|\, \pi)$$

if and only if the semigroup $(P_t)_{t\ge 0}$ satisfies

$$\mathrm{var}_\pi(f) \le C\mathcal{E}(g,g), \qquad \text{for all} \;\; g \in D(\mathcal{E}), \tag{3.2}$$

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 (3.2) is precisely our assumption (MP) with $C = C_{\mathsf{MP}}$. □

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 transportation-cost inequality

$$W_2^2(\mu,\pi) \le 2C_{\mathsf{P}}\chi^2(\mu \,\|\, \pi), \qquad \forall \mu \ll \pi. \tag{3.3}$$

We thank Jon Niles-Weed for bringing the above to our attention.

The inequality (3.3) implies that if π has a finite Poincaré constant C_{P} then Theorem 2 also yields exponential convergence in Wasserstein distance. In the rest of the paper, we write this result as

$$\frac{1}{2C_{\mathsf{P}}}W_2^2(\mu_t,\pi) \le e^{-\frac{2t}{C_{\mathsf{MP}}}}\chi^2(\mu_0 \,\|\, \pi)\,,$$

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.

3.4 Applications

We specialize Theorem 2 to the following important applications.

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 1. *Suppose that V is strictly convex and twice continuously differentiable. Then, the law $(\mu_t)_{t\geq 0}$ of NLD satisfies*

$$2\|\mu_t - \pi\|_{\mathrm{TV}}^2,\ H^2(\mu_t, \pi),\ D_{\mathrm{KL}}(\mu_t \,\|\, \pi),\ \chi^2(\mu_t \,\|\, \mu),\ \frac{1}{2C_{\mathsf{P}}} W_2^2(\mu_t, \pi) \leq e^{-2t} \chi^2(\mu_0 \,\|\, \pi).$$

If π is log-concave, then it satisfies a Poincaré inequality [AB15; LV17] so that the result in Wasserstein distance holds. In fact, contingent on the famous *Kannan-Lovász-Simonovitz* (KLS) conjecture ([KLS95]), the Poincaré constant of any log-concave distribution π is upper bounded by a constant, independent of the dimension, times the largest eigenvalue of the covariance matrix of π.

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], thus 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 2. *Fix a target accuracy $\varepsilon > 0$. Suppose $\pi = e^{-V}$ is log-concave and set $\pi_\beta \propto e^{-V-\beta\|\cdot\|^2}$, 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 $\|\mu_t - \pi\|_{\mathrm{TV}} \leq \varepsilon$ by time $t = \frac{1}{2}\ln[2\chi^2(\mu_0 \,\|\, \pi_\beta)] + \ln(1/\varepsilon)$.*

Proof. From our assumption, it holds

$$D_{\mathrm{KL}}(\pi \,\|\, \pi_\beta) = \int \ln \frac{\mathrm{d}\pi}{\mathrm{d}\pi_\beta} \,\mathrm{d}\pi = \beta \int \|\cdot\|^2 \,\mathrm{d}\pi + \ln \int e^{-\beta\|\cdot\|^2} \,\mathrm{d}\pi \leq \beta \int \|\cdot\|^2 \,\mathrm{d}\pi \leq \frac{\varepsilon^2}{2}.$$

Moreover, Theorem 2 with the above choice of t yields $D_{\mathrm{KL}}(\mu_t \,\|\, \pi_\beta) \leq \varepsilon^2/2$. To conclude, we use Pinsker's inequality and the triangle inequality for $\|\cdot\|_{\mathrm{TV}}$. □

Extensions for the other cases where ϕ is only a *proxy* for V can be found in the original paper **??**.

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 $\pi_\beta \propto e^{-\beta\widetilde{V}}$, 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 [Nes04, §4]).

We now formulate our sampling result. In our continuous framework, it does not require self-concordance of the barrier function.

Corollary 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 $(\mu_t)_{t\ge 0}$ of NLD with target density $\pi_\beta \propto e^{-\beta\widetilde{V}}$ for $\beta \le \varepsilon^2/(2\nu)$ satisfies $\|\mu_t - \pi\|_{\mathrm{TV}} \le \varepsilon$ by time $t = \frac{1}{2}\ln[2\chi^2(\mu_0 \,\|\, \pi_\beta)] + \ln(1/\varepsilon)$.*

Proof. Lemma 1 in Section 3.5 ensures that $D_{\mathrm{KL}}(\pi_\beta \,\|\, \pi) \le \varepsilon^2/2$. We conclude as in the proof of Corollary 2, by using Theorem 2, Pinsker's inequality, and the triangle inequality for $\|\cdot\|_{\mathrm{TV}}$. □

Langevin diffusion under a Poincaré inequality

We conclude this section by giving some implications of Theorem 2 to the classical 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 Section 3.3.

Corollary 4. *Suppose that π satisfies a Poincaré inequality (P) with constant $C_{\mathsf{P}} > 0$. Then, the law $(\mu_t)_{t\ge 0}$ of the Langevin diffusion (LD) satisfies*

$$2\|\mu_t - \pi\|_{\mathrm{TV}}^2,\ H^2(\mu_t, \pi),\ D_{\mathrm{KL}}(\mu_t \,\|\, \pi),\ \chi^2(\mu_t \,\|\, \mu),\ \frac{1}{2C_{\mathsf{P}}} W_2^2(\mu_t, \pi) \le e^{-\frac{2t}{C_{\mathsf{P}}}} \chi^2(\mu_0 \,\|\, \pi).$$

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.

Moreover, it is classical that if π satisfies a log-Sobolev inequality (LSI) with constant C_{LSI} then it has Poincaré constant $C_{\mathsf{P}} \le C_{\mathsf{LSI}}$. Thus, the choice of the chi-squared divergence as a surrogate for the KL divergence when tracking progress indeed requires weaker assumptions on π.

3.5 Stability in KL with respect to exp-concave perturbations

The following lemma quantifies the approximation error of replacing π by π_β in Section 3.4 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 [Nes04, (4.2.17)].

Recall that b is *ν-exp-concave* if the mapping $\exp(-\nu^{-1}b)$ is concave.

Lemma 1. *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*

$$D_{\mathrm{KL}}(\widetilde{\pi} \,\|\, \pi) \le \nu\,.$$

Proof. On $\operatorname{int}\mathcal{K}$, we have

$$-\nabla \ln \frac{\mathrm{d}\widetilde{\pi}}{\mathrm{d}\pi} = \nabla b. \tag{3.4}$$

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

$$0 = D_{\mathrm{KL}}(\pi \,\|\, \pi) \ge D_{\mathrm{KL}}(\widetilde{\pi} \,\|\, \pi) + \mathbb{E}\Big\langle \nabla \ln \frac{\mathrm{d}\widetilde{\pi}}{\mathrm{d}\pi}(\widetilde{X}), X - \widetilde{X}\Big\rangle,$$

where $(X, \widetilde{X})$ are optimally coupled for π and $\widetilde{\pi}$. If we rearrange this inequality and use the identities in (3.4), we get

$$D_{\mathrm{KL}}(\widetilde{\pi} \,\|\, \pi) \le -\,\mathbb{E}\Big\langle \nabla \ln \frac{\mathrm{d}\widetilde{\pi}}{\mathrm{d}\pi}(\widetilde{X}), X - \widetilde{X}\Big\rangle = \mathbb{E}\langle \nabla b(\widetilde{X}), X - \widetilde{X}\rangle\,. \tag{3.5}$$

We now use the fact that b is ν-exp-concave. To that end, define the convex function

$$\varphi(t) = -\exp\bigl(-\frac{1}{\nu} b(\widetilde{X} + t\,(X - \widetilde{X}))\bigr), \qquad t \in [0,1]\,.$$

By convexity, we have

$$\varphi'(0) \cdot (1-0) \le \varphi(1) - \varphi(0) \le -\varphi(0) = \exp\bigl(-\frac{1}{\nu} b(\widetilde{X})\bigr).$$

Since

$$\varphi'(0) = \frac{1}{\nu} \exp\bigl(-\frac{1}{\nu} b(\widetilde{X})\bigr) \, \langle \nabla b(\widetilde{X}), X - \widetilde{X} \rangle\,,$$

the above inequality reads $\langle \nabla b(\widetilde{X}), X - \widetilde{X} \rangle \le \nu$, which completes the proof together with (3.5). □

Remark 1. 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 $\nu \le d$ [NN94; BE15; TY18].

3.6 Numerical experiments

In this section, we examine the numerical performance of the *Newton-Langevin Algorithm* (NLA), which is given by the following Euler discretization of NLD:

$$\nabla V(X_{k+1}) = (1-h)\nabla V(X_k) + \sqrt{2h}\,[\nabla^2 V(X_k)]^{1/2} \xi_k, \tag{NLA}$$

where $(\xi_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. $\mathcal{N}(0, I_d)$ variables.

Generalized Gaussian distribution

We first demonstrate the scale invariance of NLD established in Corollary 1, by sampling from an ill-conditioned generalized Gaussian distribution on $\mathbb{R}^{100}$ with $V(x) = \langle x, \Sigma^{-1} x\rangle^\gamma / 2$ for $\gamma = 3/4$.

Figure 3-1 compares the performance of NLA to that of the Unadjusted Langevin Algorithm (ULA) [DM+19] and of the Tamed Unadjusted Langevin Algorithm (TULA) [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 rapidly approximate the true distribution and also displays stability to the choice of the step size h.

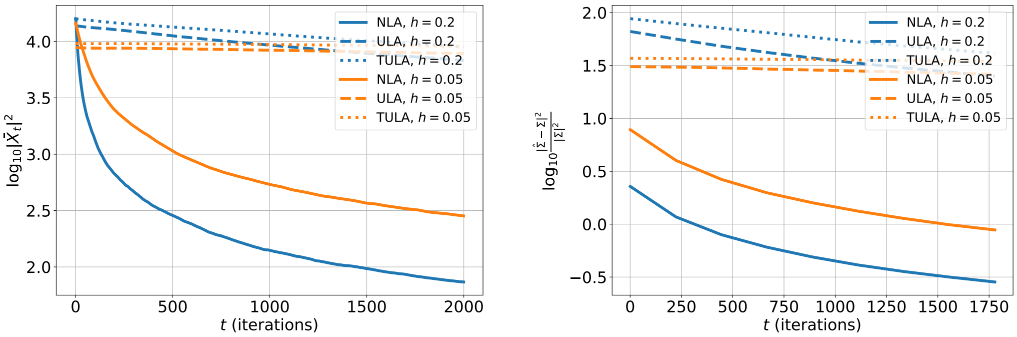

Figure 3-1: $V(x) = \langle x, \Sigma^{-1}x\rangle^{3/4}/2$, $\Sigma = \text{diag}(1, 2, \ldots, 100)$. Left: absolute squared error of the mean 0. Right: relative squared error for the scatter matrix Σ.

Gaussian distribution

We repeat the example in Figure 3-1 for the simpler case of the Gaussian distribution ($\gamma = 1$) on $\mathbb{R}^{100}$ with the same scatter matrix $\Sigma = \text{diag}(1, 2, \ldots, 100)$ in Figure 3-2. We again see the superiority of NLA over the Unadjusted Langevin Algorithm (ULA) [DM+19] and the Tamed Unadjusted Langevin Algorithm (TULA) [Bro+19]. Here and in Section 3.6 the additional parameter of TULA (denoted γ in [Bro+19]) is chosen equal to .1.

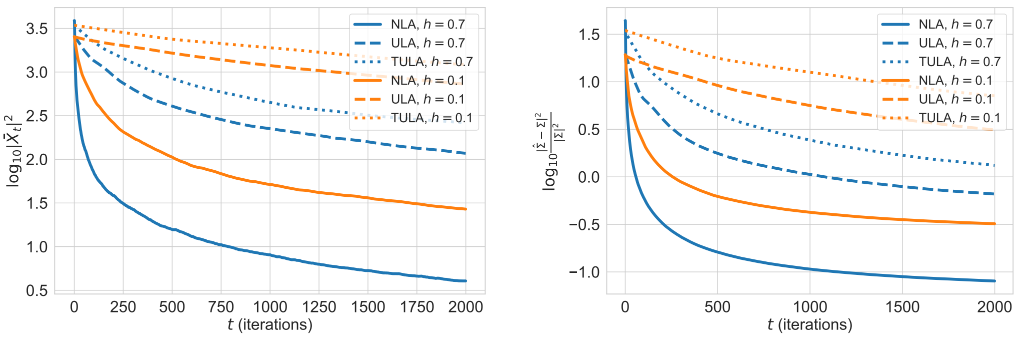

Figure 3-2: We display convergence of the various algorithms for an ill-conditioned Gaussian distribution, with $d = 100$ and $\Sigma = \text{diag}(1, 2, \ldots, 100)$. Left: error is the squared distance from 0. Right: error is the relative distance between scatter matrices. As in the experiment displayed in Figure 3-1, 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 3-2 in Figure 3-3. NLA maintains good performance for a wide range of step-size choices, 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.

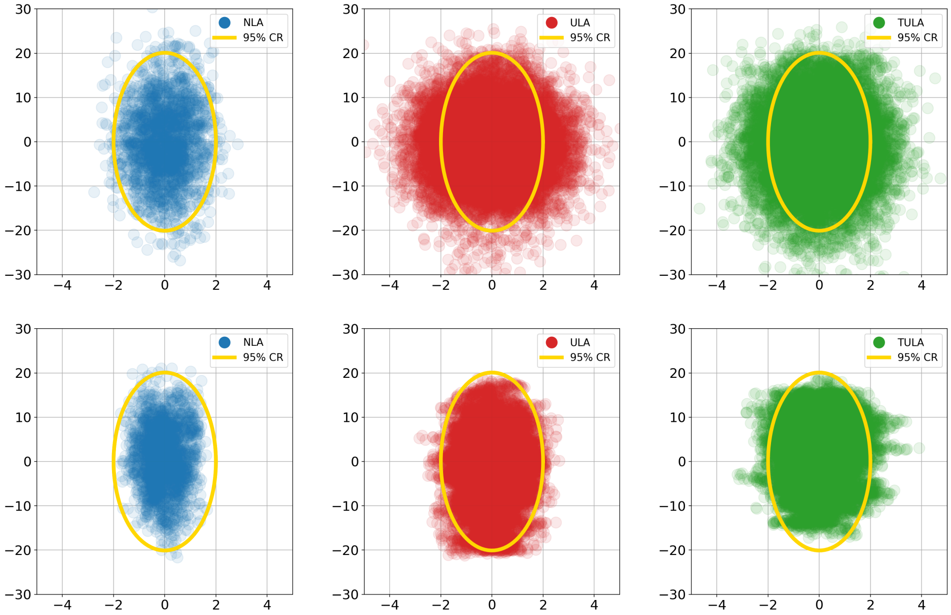

Figure 3-3: Samples from NLA, ULA, and TULA for the ill-conditioned Gaussian example of Figure 3-2, with $\Sigma = \text{diag}(1, 2, \ldots, 100)$. We display the projection onto the first (least spread) and last (most spread) population principal components, along with the projection of a 95% confidence region. Top: the step size for all algorithms is $h = 0.7$, Bottom: the step size for all algorithms is $h = 0.05$.

Uniform sampling on a convex body

We demonstrate the efficacy of NLD established in Corollary 3 in a simple simulation: sampling uniformly from the ill-conditioned rectangle $[-a, a] \times [-1, 1]$ with $a = 0.01$ (Figure 3-4). We compare NLA with the Projected Langevin Algorithm (PLA) [BEL18], both with 200 iterations and $h = 10^{-4}$. For NLA, we take $\widetilde{V}(x) = -\log(1 - x_1^2) - \log(a^2 - x_2^2)$ and $\beta = 10^{-4}$. PLA and MALA target the uniform distribution directly. NLA samples from an approximate distribution, given in Section 3.4. 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

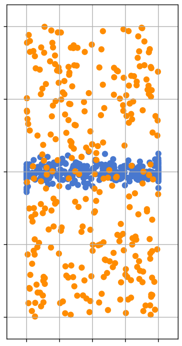

Figure 3-4: Uniform sampling from the set $[-0.01, 0.01] \times [-1, 1]$: PLA (blue) vs. NLA (orange). See Section 3.6.

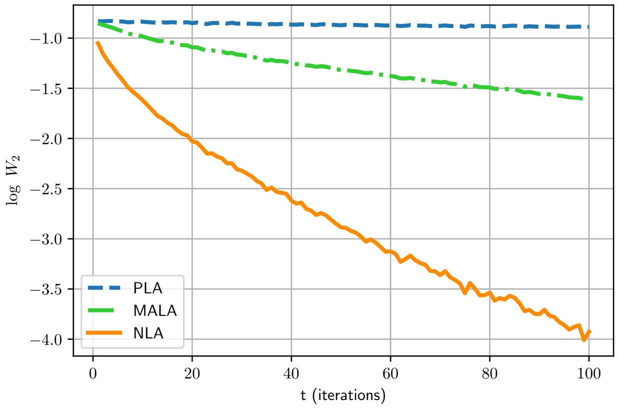

Figure 3-5: W_2 distance (on logarithmic scale) between the uniform distribution on the rectangle $[-0.01, 0.01] \times [-1, 1]$, and samples produced by NLA, PLA, and MALA.

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 3-5. We use the Sinkhorn distance ($\varepsilon = 0.01$) as an approximation for the 2-Wasserstein distance [Cut13; AWR17]. 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.

3.7 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 we leave open the problem of obtaining discretization error bounds. 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+19] or Hamiltonian Monte Carlo [Nea12], which also enjoy the scale invariance of NLD?

Chapter 4

Particle based methods: SVGD

4.1 Introduction

As we have seen in Chapter 3, the perspective of sampling as a Wasserstein gradient flow is fruitful both for analysing and proposing sampling algorithms. In order to arrive at a practical sampling algorithm, one must discretize the Wasserstein gradient flow. This chapter focuses on a novel discretized sampling algorithm, Stein Variational Gradient Descent.

The most common discretization is to discretize the Langevin diffusion, resulting in the Unadjusted Langevin Algorithm (ULA) [Dal17b; DM17]. However, it is unclear whether this diffusion based discretization is the most effective one. In fact, ULA is asymptotically biased, which results in slow convergence and often requires ad-hoc adjustments [Dwi+19]. To overcome this limitation, various methods that track the Wasserstein gradient flow more closely have been recently developed [Ber18; Wib18; SKL20].

An alternative sampling approach that avoids diffusions is to construct a sequence of *deterministic* mappings that approximately pushes forward an initial distribution to the target distribution. Let F denote a functional over the Wasserstein space of distributions. The Wasserstein gradient flow of F may be described as the deterministic and time-inhomogeneous Markov process $(X_t)_{t\geq 0}$ started at a random variable $X_0 \sim \mu_0$ and evolving according to $\dot{X}_t = -[\nabla_{W_2} F(\mu_t)](X_t)$, where μ_t denotes the distribution of X_t. Here $[\nabla_{W_2} F(\mu)](\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ is the Wasserstein gradient of F at μ. If $F(\mu) = \mathsf{KL}(\cdot\|\mu)$, where $\pi \propto e^{-V}$ is a given target distribution on $\mathbb{R}^d$, it is known [AGS08; Vil09; San17] that $\nabla_{W_2} F(\mu) = \nabla \ln(\mathrm{d}\mu/\mathrm{d}\pi)$. Therefore, a natural discretization of the Wasserstein gradient flow with step size $h > 0$, albeit one that cannot be implemented since it depends on the distribution μ_t of X_t, is:

$$X_{t+1} = X_t - h\nabla \ln\big(\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(X_t)\big), \qquad t = 0, 1, 2, \dots .$$

While μ_t can, in principle, be estimated by evolving a large number of particles $X_t^{[1]}, \dots, X_t^{[N]}$, estimation of μ_t is hindered by the curse of dimensionality and this approach still faces significant computational challenges despite attempts to improve the original JKO scheme [SKL20; WL20].

A major advance in this direction was achieved by allowing for *approximate* Wasserstein gradients, which makes the push forward maps tractable. More specifically, Stein Variational Gradient Descent (SVGD), recently proposed by [LW16] (see Section 4.2 for more details), consists in replacing $\nabla_{W_2} F(\mu)$ by its image $\mathcal{K}_\mu \nabla_{W_2} F(\mu)$ under the integral operator $\mathcal{K}_\mu : L^2(\mu) \to L^2(\mu)$ associated to a chosen kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and defined by $\mathcal{K}_\mu f(x) := \int K(x,y) f(y) \, \mathrm{d}\mu(y)$ for $f \in L^2(\mu)$. This leads to the following process:

$$X_{t+1} = X_t - h[\mathcal{K}_{\mu_t} \nabla_{W_2} F(\mu_t)](X_t) \,, \qquad t = 0, 1, 2, \dots \,. \tag{$\mathsf{SVGD_p}$}$$

where we apply the integral operator $\mathcal{K}_{\mu_t}$ individually to each coordinate of the Wasserstein gradient. In turn, this *kernelization trick* overcomes most of the above computational bottleneck. Building on this perspective, [DNS19] introduced a new geometry, different from the Wasserstein geometry and which they call the *Stein geometry*, in which the continuous limit of ($\mathsf{SVGD_p}$) becomes the gradient flow of the KL divergence.

However, despite this recent advance, the theoretical properties of SVGD are still largely unexplored, resulting in little understanding of SVGD's known problems, such as mode collapse or a lack of guidance on how to choose an appropriate kernel K. Consequently diffusion-based algorithms remain the dominant choice for applications. In this chapter, we we provide a new and stronger theoretical footing for the development of such deterministic mappings.

This chapter is based on the joint work [Che+20d], with Sinho Chewi, Thibaut Le Gouic, Tyler Maunu, and Philippe Rigollet.

Our contributions . We introduce, in Section 4.2, a new perspective on SVGD by viewing it as kernelized gradient flow of the chi-squared divergence rather than the KL divergence. This perspective is fruitful in two ways. First, it uses a single integral operator $\mathcal{K}_\pi$—as opposed to ($\mathsf{SVGD_p}$), which requires a family of integral operators $\mathcal{K}_\mu$, $\mu \ll \pi$—providing a conceptually clear guideline for choosing K, namely: K should be chosen to make $\mathcal{K}_\pi$ approximately equal to the identity operator. Second, under the idealized choice $\mathcal{K}_\pi = \mathrm{id}$, we show that this gradient flow converges exponentially fast in KL divergence as soon as the target distribution π satisfies a Poincaré inequality. In fact, our results are stronger than exponential convergence and they highlight *strong uniform ergodicity*: the gradient flow forgets the initial distribution after a finite time that is at most half of the Poincaré constant. To establish this exponential convergence under a relatively weak condition (Poincaré inequality), we employ the following technique. While the gradient flow aims at minimizing the chi-squared divergence by following the curve in Wasserstein space with steepest descent, we do not track its progress with the objective function itself, the chi-squared divergence, but instead we track it with the KL divergence. This is in a sense dual to argument employed in [Che+20b], where the chi-squared divergence is used to track the progress of a gradient flow on the KL divergence. A more standard analysis relying on Łojasiewicz inequalities also yields rates of convergence on the chi-squared divergence under stronger assumptions such as a log-Sobolev inequality, and log-concavity. These results establish the first finite-time theoretical guarantees for SVGD in an idealized setting.

Beyond providing a better understanding of SVGD, our novel perspective is instrumental in the development of a new sampling algorithm, which we call Laplacian Adjusted Wasserstein Gradient Descent (LAWGD) and present in Section 4.4. We show that it possesses a striking theoretical property: assuming that the target distribution π satisfies a Poincaré inequality, LAWGD converges exponentially fast, with *no* dependence on the Poincaré constant. This scale invariance has been recently demonstrated for the Newton-Langevin diffusion [Che+20b], but under the additional assumption that π is log-concave. A successful implementation of LAWGD hinges on the spectral decomposition of a certain differential operator which is within reach of modern PDE solvers. As a proof of concept, we show that LAWGD, implemented using a naïve finite differences method, performs well on mixtures of Gaussians in one or two dimensions, whereas SVGD fails. This is an indication that our novel perspective could be the correct one to further advance the state-of-the-art for sampling via deterministic mappings. Implementing LAWGD in high dimensions is challenging, and we are not advocating for it as the definitive solution of the sampling problem. Instead, LAWGD serves as the start of a family of interacting particle systems with an interacting potential that depends strongly and non-trivially on the target distribution and furthermore comes with strong theoretical guarantees. We hope this chapter can encourage further research in the application of numerical PDEs for sampling.

Related works . Since its introduction in [LW16], a number of variants of SVGD have been considered. They include a stochastic version [Li+20], a version that approximates the Newton direction in Wasserstein space [Det+18], a version that uses matrix kernels [Wan+19], an accelerated version [Liu+19], and a hybrid with Langevin [Zha+20a]. Several works have studied theoretical properties of SVGD, including its interpretation as a gradient flow under a modified geometry [Liu17; DNS19], and its asymptotic convergence [LLN19].

Notation . In this chapter, all probability measures are assumed to have densities w.r.t. Lebesgue measure; therefore, we frequently abuse notation by identifying a probability measure with its Lebesgue density. For a differentiable kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we denote by $\nabla_1 K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ (resp. $\nabla_2 K$) the gradient of the kernel w.r.t. the first (resp. second) argument. When describing particle algorithms, we use a subscript to denote the time index and brackets to denote the particle index, i.e., $X_t^{[i]}$ refers to the ith particle at time (or iteration number) t.

4.2 SVGD as a kernelized Wasserstein gradient flow

An introduction to the theory of gradient flows was given in Section 2.1. In this chapter, we are primarily concerned with two functionals on the 2-Wasserstein space: the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(\cdot \parallel \pi)$, and the chi-squared divergence $\chi^2(\cdot \parallel \pi)$ (see, e.g., [Tsy09]). We recall [AGS08; San15] that the gradients of these functionals are, respectively,

$$\big(\nabla_{W_2} D_{\mathrm{KL}}(\cdot \parallel \pi)\big)(\mu) = \nabla \ln \frac{\mathrm{d}\mu}{\mathrm{d}\pi}, \qquad \big(\nabla_{W_2} \chi^2(\cdot \parallel \pi)\big)(\mu) = 2\nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}. \tag{4.1}$$

SVGD as a kernelized gradient flow of the KL divergence

SVGD[1] is achieved by replacing the Wasserstein gradient $\nabla \ln(\mathrm{d}\mu_t/\mathrm{d}\pi)$ of the KL divergence with $\mathcal{K}_{\mu_t} \nabla \ln(\mathrm{d}\mu_t/\mathrm{d}\pi)$, leading to the particle evolution equation ($\mathsf{SVGD_p}$).

Recalling that $\pi \propto e^{-V}$, we get

$$\mathcal{K}_{\mu_t} \nabla \ln \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(x) := \int K(x,\cdot) \nabla \ln \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} \, \mathrm{d}\mu_t = \int K(x,\cdot) \nabla V \, \mathrm{d}\mu_t - \int \nabla_2 K(x,\cdot) \, \mathrm{d}\mu_t \,, \tag{4.2}$$

where, in the second identity, we used integration by parts. This expression shows that rather than having to estimate the distribution μ_t, it is sufficient to estimate the expectation $\int \nabla_2 K(x,\cdot) \, \mathrm{d}\mu_t$. This is the key to the computational tractability of SVGD. Indeed, the kernelized gradient flow can implemented by drawing N particles $X_0^{[1]}, \dots, X_0^{[N]} \overset{\text{i.i.d.}}{\sim} \mu_0$ and following the coupled dynamics

$$\dot X_t^{[i]} = -\mathcal{K}_{\mu_t} \nabla \ln \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(X_t^{[i]}) = -\int K(X_t^{[i]},\cdot) \nabla V \, \mathrm{d}\mu_t + \int \nabla_2 K(X_t^{[i]},\cdot) \, \mathrm{d}\mu_t, \qquad i \in [N].$$

With this, we can simply estimate the expectation with respect to μ_t with an

[1] Throughout this chapter, we call SVGD the generalization of the original method of [LW16; Liu17] that was introduced in [Wan+19].

average over all particles, which yeilds the SVGD algorithm:

$$X_{t+1}^{[i]} = X_t^{[i]} - \frac{h}{N}\sum_{j=1}^{N} K(X_t^{[i]}, X_t^{[j]})\nabla V(X_t^{[j]}) + \frac{h}{N}\sum_{j=1}^{N} \nabla_2 K(X_t^{[i]}, X_t^{[j]}), \qquad i \in [N]. \tag{4.3}$$

SVGD as a kernelized gradient flow of the chi-squared divergence

Recall from Section 4.2 that by the continuity equation, the continuous limit of the particle evolution equation ($\mathsf{SVGD_p}$) translates into the following PDE that describes the evolution of the distribution μ_t of X_t:

$$\partial_t \mu_t = \operatorname{div}\bigl(\mu_t \mathcal{K}_{\mu_t} \nabla \ln \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr). \tag{$\mathsf{SVGD_d}$}$$

We make the simple observation that

$$\mathcal{K}_{\mu_t}\nabla \ln \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(x) = \int K(x,y)\nabla \ln \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(y)\,\mathrm{d}\mu_t(y) = \int K(x,y)\nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(y)\,\mathrm{d}\pi(y) = \mathcal{K}_\pi \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(x).$$

Thus, the continuous-dynamics of SVGD, as given in ($\mathsf{SVGD_d}$), can equivalently be expressed as

$$\partial_t \mu_t = \operatorname{div}\bigl(\mu_t \mathcal{K}_\pi \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr). \tag{SVGD}$$

To interpret this equation, we recall that the Wasserstein gradient of the chi-squared divergence $\chi^2(\cdot \,\|\, \pi)$ at μ is $2\nabla(\mathrm{d}\mu/\mathrm{d}\pi)$ (by (4.1)), so the gradient flow for the chi-squared divergence is

$$\partial_t \mu_t = 2\operatorname{div}\bigl(\mu_t \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr). \tag{CSF}$$

Comparing (SVGD) and (CSF), we see that (up to a factor of 2), SVGD can be understood as the flow obtained by replacing the gradient of the chi-squared divergence, $\nabla(\mathrm{d}\mu/\mathrm{d}\pi)$, by $\mathcal{K}_\pi\nabla(\mathrm{d}\mu/\mathrm{d}\pi)$.

Although ($\mathsf{SVGD_d}$) and (SVGD) are equivalent ways of expressing the same dynamics, the formulation of (SVGD) presents a significant advantage: it involves a kernel integral operator $\mathcal{K}_\pi$ that does not change with time and depends only on the target distribution π.

4.3 Chi-squared gradient flow

In this section, we study the idealized case where $\mathcal{K}_\pi$ taken to be the identity operator. In this case, (SVGD) reduces to the gradient flow CSF. The existence, uniqueness, and regularity of this flow are studied in [OT11; OT13] and [AGS08, Theorem 11.2.1].

The rate of convergence of the gradient flow of the KL divergence is closely related to two functional inequalities: the Poincaré inequality controls the rate of exponential convergence in chi-squared divergence ([Pav14, Theorem 4.4], [Che+20b]) while a log-Sobolev inequality characterizes the rate of exponential convergence of the KL divergence [BGL14, Theorem 5.2.1]. In this section, we show that these inequalities also guarantee exponential rates of convergence of CSF.

Recall that π satisfies a *Poincaré inequality* with constant C_{P} if

$$\operatorname{var}_\pi f \le C_{\mathsf{P}}\, \mathbb{E}_\pi[\|\nabla f\|^2], \qquad \text{for all locally Lipschitz } f \in L^2(\pi), \tag{P}$$

while π satisfies a *log-Sobolev inequality* (LSI) with constant C_{LSI}

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

for all locally Lipschitz f for which $\operatorname{ent}_\pi(f^2) < \infty$.

We briefly review some facts regarding the strength of these assumptions. It is standard that the log-Sobolev inequality is stronger than the Poincaré inequality: (LSI) implies (P) with constant $C_{\mathsf{P}} \le C_{\mathsf{LSI}}$. In turn, if π is α-*strongly log-concave*, i.e. $\nabla^2 V \succeq \alpha I_d$, then it implies the validity of (LSI) with $C_{\mathsf{LSI}} \le 1/\alpha$, and thus a Poincaré inequality holds too. However, a Poincaré inequality is in general much weaker than strong log-concavity. For instance, if λ_π^2 denotes the largest eigenvalue of the covariance matrix of π, then it is currently known that π satisfies a Poincaré inequality as soon as it is log-concave, with $C_{\mathsf{P}} \le C(d)\lambda_\pi^2$, where $C(d)$ is a dimensional constant [Bob99; AB15; LV17], and the well-known *Kannan-Lovász-Simonovitz* (KLS) *conjecture* [KLS95] asserts that $C(d)$ does not actually depend on the dimension.

Our first result shows that a Poincaré inequality suffices to establish exponential decay of the KL divergence along CSF. In fact, we establish a remarkable property, which we call *strong uniform ergodicity*: under a Poincaré inequality, CSF forgets its initial distribution after a time of no more than $C_{\mathsf{P}}/2$. Uniform ergodicity is central in the theory of Markov processes [MT09, Ch. 16] but is often limited to compact state spaces. Moreover, this theory largely focuses on total variation, so the distance from the initial distribution to the target distribution is trivially bounded by 1.

Theorem 3. *Assume that π satisfies a Poincaré inequality* (P) *with constant $C_{\mathsf{P}} > 0$ and let $(\mu_t)_{t\ge0}$ denote the law of* CSF*. Assume that $\chi^2(\mu_0 \,\|\, \pi) < \infty$.*

Then,

$$D_{\mathrm{KL}}(\mu_t \,\|\, \pi) \le D_{\mathrm{KL}}(\mu_0 \,\|\, \pi)\, e^{-\frac{2t}{C_{\mathsf{P}}}}\,, \qquad \forall\, t \ge 0. \tag{4.4}$$

In fact, a stronger convergence result holds:

$$D_{\mathrm{KL}}(\mu_t \,\|\, \pi) \le \big(D_{\mathrm{KL}}(\mu_0 \,\|\, \pi) \wedge 2\big)\, e^{-\frac{2t}{C_{\mathsf{P}}}}\,, \qquad \forall\, t \ge \frac{C_{\mathsf{P}}}{2}. \tag{4.5}$$

Proof. Given the Wasserstein gradients (4.1) in Section 2.1, we get that $(\mu_t)_{t\ge 0}$ satisfies

$$\partial_t D_{\mathrm{KL}}(\mu_t \,\|\, \pi) = -2\, \mathbb{E}_{\mu_t}\Big\langle \nabla \ln \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}, \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\Big\rangle = -2\, \mathbb{E}_\pi\Big[\Big\|\nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\Big\|^2\Big].$$

Applying the Poincaré inequality (P) with $f = \mathrm{d}\mu_t/\mathrm{d}\pi - 1$, we get

$$\partial_t D_{\mathrm{KL}}(\mu_t \,\|\, \pi) \le -\frac{2}{C_{\mathsf{P}}} \chi^2(\mu_t \,\|\, \pi) \le -\frac{2}{C_{\mathsf{P}}} D_{\mathrm{KL}}(\mu_t \,\|\, \pi)\,,$$

where, in the last inequality, we use the fact that $D_{\mathrm{KL}}(\cdot \,\|\, \pi) \le \chi^2(\cdot \,\|\, \pi)$ (see [Tsy09, §2.4]). The bound (4.4) follows by applying Grönwall's inequality.

To prove (4.5), we use the stronger inequality $D_{\mathrm{KL}}(\cdot \,\|\, \pi) \le \ln[1 + \chi^2(\cdot \,\|\, \pi)]$ (see [Tsy09, §2.4]). Our differential inequality now reads:

$$\partial_t D_{\mathrm{KL}}(\mu_t \,\|\, \pi) \le -\frac{2}{C_{\mathsf{P}}}\big(e^{D_{\mathrm{KL}}(\mu_t\|\pi)} - 1\big) \iff \partial_t \psi\big(D_{\mathrm{KL}}(\mu_t \,\|\, \pi)\big) \le -\frac{2}{C_{\mathsf{P}}} \psi\big(D_{\mathrm{KL}}(\mu_t \,\|\, \pi)\big)\,,$$

where $\psi(x) = 1 - e^{-x} \le 1$. Grönwall's inequality now yields

$$\psi\big(D_{\mathrm{KL}}(\mu_t \,\|\, \pi)\big) \le e^{-\frac{2t}{C_{\mathsf{P}}}} \psi\big(D_{\mathrm{KL}}(\mu_0 \,\|\, \pi)\big) \le e^{-\frac{2t}{C_{\mathsf{P}}}}.$$

Note that $x \le 2\psi(x)$ whenever $\psi(x) \le 1/e$. Thus, if $t \ge C_{\mathsf{P}}/2$, we get $\psi\big(D_{\mathrm{KL}}(\mu_t \,\|\, \pi)\big) \le e^{-1}$ so

$$D_{\mathrm{KL}}(\mu_t \,\|\, \pi) \le 2\psi\big(D_{\mathrm{KL}}(\mu_t \,\|\, \pi)\big) \le e^{-\frac{2t}{C_{\mathsf{P}}}}\,,$$

which, together with (4.4), completes the proof of (4.5). □

Remark 2. In [Che+20b], it was observed that the chi-squared divergence decays exponentially fast along the gradient flow $(\mu_t)_{t\ge 0}$ for the KL divergence, provided that π satisfies a Poincaré inequality. This observation is made precise and more general in [MMS09] where it is noted that the *gradient flow of a functional* $\mathfrak{U}$ *dissipates a different functional* $\mathfrak{V}$ *at the same rate that the gradient flow of* $\mathfrak{V}$ *dissipates the functional* $\mathfrak{U}$. A similar method is used to study the thin film equation in [CT02] and [Car11, §5].

Since we are studying the gradient flow of the chi-squared divergence, it is natural to ask whether CSF converges to π in chi-squared divergence as well. In the next results, we show quantitative decay of the chi-squared divergence along the gradient flow under a Poincaré inequality (P), but we obtain only a polynomial rate of decay. However, if we additionally assume either that π is log-concave or that it satisfies a log-Sobolev inequality (LSI), then we obtain exponential decay of the chi-squared divergence along CSF.

Theorem 4. *Suppose that π satisfies a Poincaré inequality* (P)*. Then, provided $\chi^2(\mu_0 \,\|\, \pi) < \infty$, the law $(\mu_t)_{t\geq 0}$ of* CSF *satisfies*

$$\chi^2(\mu_t \,\|\, \pi) \leq \chi^2(\mu_0 \,\|\, \pi) \wedge \bigl(\frac{9C_{\mathsf{P}}}{8t}\bigr)^2 .$$

If we further assume that π is log-concave, then

$$\chi^2(\mu_t \,\|\, \pi) \leq \chi^2(\mu_0 \,\|\, \pi)\, e^{-\frac{t}{2C_{\mathsf{P}}}} .$$

Under the stronger assumption (LSI), we can show strong uniform ergodicity as in Theorem 3.

Theorem 5. *Assume that π satisfies a log-Sobolev inequality* (LSI)*. Let $(\mu_t)_{t\geq 0}$ denote the law of* CSF*, and assume that $\chi^2(\mu_0 \,\|\, \pi) < \infty$. Then, for all $t \geq 7C_{\mathsf{LSI}}$,*

$$\chi^2(\mu_t \,\|\, \pi) \leq \bigl(\chi^2(\mu_0 \,\|\, \pi) \wedge 2\bigr)\, e^{-\frac{t}{9C_{\mathsf{LSI}}}} .$$

Remark 3. Convergence in chi-squared divergence was studied in recent works such as [CLL19; VW19; Che+20b]. From standard comparisons between information divergences (see [Tsy09, §2.4]), it implies convergence in total variation distance, Hellinger distance, and KL divergence. Moreover, recent works have shown that the Poincaré inequality (P) yields transportation-cost inequalities which bound the 2-Wasserstein distance by powers of the chi-squared divergence [Din15; Led18; Che+20b; Liu20], so we obtain convergence in the 2-Wasserstein distance as well. In particular, we mention that [Che+20b] uses the chi-squared gradient flow (CSF) to prove a transportation-cost inequality.

Proof of Theorem 4 (non-log-concave case). According to [Che+20b, Proposition 1], the Poincaré inequality implies the following inequality for the chi-squared divergence:

$$\chi^2(\mu \,\|\, \pi)^{3/2} \leq \frac{9C_{\mathsf{P}}}{4}\, \mathbb{E}_\mu\bigl[\bigl\|\nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\bigr\|^2\bigr], \qquad \forall \mu \ll \pi. \tag{4.6}$$

Since the Wasserstein gradient of $\chi^2(\cdot \,\|\, \pi)$ at μ is given by $2\nabla(\mathrm{d}\mu/\mathrm{d}\pi)$ (see

Section 4.2), it yields

$$\partial_t \chi^2(\mu_t \parallel \pi) = -4\, \mathbb{E}_{\mu_t}\bigl[\bigl\|\nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr\|^2\bigr] \le -\frac{16}{9C_{\mathsf{P}}}\chi^2(\mu_t \parallel \pi)^{3/2}$$

Solving the above differential inequality yields

$$\chi^2(\mu_t \parallel \pi) \le \frac{\chi^2(\mu_0 \parallel \pi)}{\{1 + 8t\sqrt{\chi^2(\mu_0 \parallel \pi)}/(9C_{\mathsf{P}})\}^2},$$

which implies the desired result. □

We now prepare for the proof of exponentially fast convergence in chi-squared divergence for log-concave measures. The key to proving such results lies in differential inequalities of the form

$$\chi^2(\mu \parallel \pi) \le C_{\mathsf{PL}}\, \mathbb{E}_{\mu}\bigl[\bigl\|\nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\bigr\|^2\bigr], \qquad \forall \mu \ll \pi, \tag{4.7}$$

which may be interpreted as a *Polyak-Łojasiewicz* (PL) *inequality* [KNS16] for the functional $\chi^2(\cdot \parallel \pi)$. PL inequalities are well-known in the optimization literature, and can be even used when the objective is not convex [Che+20c]. In contrast, the preceding proof uses the weaker inequality (4.6), which may be interpreted as a *Łojasiewicz inequality* [Loj63].

To see that a PL inequality readily yields exponential convergence, observe that

$$\partial_t \chi^2(\mu_t \parallel \pi) = -4\, \mathbb{E}_{\mu_t}\bigl[\bigl\|\nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr\|^2\bigr] \le -\frac{4}{C_{\mathsf{PL}}}\chi^2(\mu_t \parallel \pi)\,.$$

Together with Grönwall's inequality, the differential inequality yields $\chi^2(\mu_t \parallel \pi) \le \chi^2(\mu_0 \parallel \pi)\, e^{-\frac{4t}{C_{\mathsf{PL}}}}$.

In order to prove a PL inequality of the type (4.7), we require two ingredients. The first one is a transportation-cost inequality for the chi-squared divergence proven in [Liu20], building on the works [Din15; Led18]. It asserts that if π satisfies a Poincaré inequality (P), then the following inequality holds:

$$W_2^2(\mu, \pi) \le 2C_{\mathsf{P}}\chi^2(\mu \parallel \pi), \qquad \forall \mu \ll \pi. \tag{4.8}$$

For the second ingredient, we use an argument of [OV00] to show that if π satisfies a chi-squared transportation-cost inequality such as (4.8), and in addition is log-concave, then it satisfies an inequality of the type (4.7). We remark that the converse statement, that is, if π satisfies a PL inequality (4.7) then it satisfies an appropriate chi-squared transportation-cost inequality, was proven in [Che+20b] without the additional assumption of log-concavity. It implies that for log-concave distributions, the PL inequality (4.7) and the chi-squared transportation-cost inequality (4.8) are, in fact, equivalent.

Theorem 6. *Let π be log-concave, and assume that for some $q \in (1, \infty)$ and a constant $\mathsf{C} > 0$,*

$$W_2^2(\mu, \pi) \le \mathsf{C}\chi^2(\mu \,\|\, \pi)^{2/q}, \qquad \forall\, \mu \ll \pi.$$

Then,

$$\chi^2(\mu \,\|\, \pi)^{2/p} \le 4\mathsf{C}\,\mathbb{E}_\mu\big[\big\|\nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\big\|^2\big], \qquad \forall\, \mu \ll \pi, \tag{4.9}$$

where p satisfies $1/p + 1/q = 1$.

Proof. Following [OV00], let T be the optimal transport map from μ to π. Since $\chi^2(\cdot \,\|\, \pi)$ is displacement convex [OT11; OT13] and has Wasserstein gradient $2\nabla(\mathrm{d}\mu/\mathrm{d}\pi)$ at μ (c.f. Section **??**), the "above-tangent" formulation of displacement convexity ([Vil03, Proposition 5.29]) yields

$$0 = \chi^2(\pi \,\|\, \pi) \ge \chi^2(\mu \,\|\, \pi) + 2\,\mathbb{E}_\mu\big\langle \nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}, T - \mathrm{id}\big\rangle \ge \chi^2(\mu \,\|\, \pi) - 2W_2(\mu, \pi)\sqrt{\mathbb{E}_\mu\big[\big\|\nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\big\|^2\big]},$$

where we used the Cauchy-Schwarz inequality for the last inequality. Rearranging the above display and using the transportation-cost inequality assumed in the statement of theorem, we get

$$\chi^2(\mu \,\|\, \pi) \le 2W_2(\mu, \pi)\sqrt{\mathbb{E}_\mu\big[\big\|\nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\big\|^2\big]} \le 2\sqrt{\mathsf{C}\,\mathbb{E}_\mu\big[\big\|\nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\big\|^2\big]\,\chi^2(\mu \,\|\, \pi)^{1/q}}.$$

The result follows by rearranging the terms. □

Proof of Theorem 4 (log-concave case). From the transportation-cost inequality (4.8) and Theorem 6 with $p = q = 2$, we obtain

$$\chi^2(\mu \,\|\, \pi) \le 8C_{\mathsf{P}}\,\mathbb{E}_\mu\big[\big\|\nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\big\|^2\big].$$

This PL inequality together with Grönwall's inequality readily yields the result. □

We conclude this section with the proof of Theorem 5, which shows exponential convergence of CSF in chi-squared divergence under the assumption of a log-Sobolev inequality (LSI) (but without the assumption of log-concavity).

Proof of Theorem 5. We first claim that

$$\partial_t \chi^2(\mu_t \,\|\, \pi) \le -\frac{4}{9C_{\mathsf{LSI}}}[\chi^2(\mu_t \,\|\, \pi) + 1]^{3/2} \ln[\chi^2(\mu_t \,\|\, \pi) + 1]. \tag{4.10}$$

Indeed, applying (LSI), we obtain

$$\partial_t \chi^2(\mu_t \,\|\, \pi) = -4 \int \bigl\|\nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr\|^2 \,\mathrm{d}\mu_t = -\frac{16}{9} \int \bigl\|\nabla \bigl|\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr|^{3/2}\bigr\|^2 \,\mathrm{d}\pi \le -\frac{8}{9C_{\mathsf{LSI}}} \operatorname{ent}_\pi\bigl(\bigl|\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr|^3\bigr).$$

Next, the variational formula for the entropy gives

$$\operatorname{ent}_\pi f = \sup\{\mathbb{E}_\pi(fg) : g \text{ satisfies } \mathbb{E}_\pi \exp g = 1\},$$

see [Han16, Lemma 3.15] or [BLM13, Theorem 4.13]. Choosing $g = \ln(\mathrm{d}\mu_t/\mathrm{d}\pi)$ yields

$$\begin{aligned}
\operatorname{ent}_\pi\bigl(\bigl|\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr|^3\bigr) &\ge \mathbb{E}_\pi\bigl[\bigl|\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr|^3 \ln \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr] = \frac{1}{3}\,\mathbb{E}_\pi\bigl[\bigl|\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr|^3 \ln\bigl(\bigl|\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr|^3\bigr)\bigr] \\
&\ge \frac{1}{3}\,\mathbb{E}_\pi\bigl[\bigl|\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr|^3\bigr] \ln \mathbb{E}_\pi\bigl[\bigl|\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr|^3\bigr] \\
&\ge \frac{1}{2}\,\mathbb{E}_\pi\bigl[\bigl|\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr|^2\bigr]^{3/2} \ln \mathbb{E}_\pi\bigl[\bigl|\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr|^2\bigr] \\
&= \frac{1}{2}[\chi^2(\mu_t \,\|\, \pi) + 1]^{3/2} \ln[\chi^2(\mu_t \,\|\, \pi) + 1],
\end{aligned}$$

where in the second inequality, we used that $x \mapsto x \ln x$ is convex on $\mathbb{R}_+$ and in the third, we used that it increasing when $x \ge 1$ together with

$$\mathbb{E}_\pi\bigl[\bigl|\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr|^2\bigr] = 1 + \chi^2(\mu_t \,\|\, \pi) \ge 1.$$

This proves (4.10).

To simplify the inequality (4.10), we use the crude bounds

$$\ln[\chi^2(\mu_t \,\|\, \pi) + 1] \ge \begin{cases} 1, & \text{if } \chi^2(\mu_t \,\|\, \pi) \ge e - 1 \\ \chi^2(\mu_t \,\|\, \pi)/2, & \text{otherwise.} \end{cases}$$

It yields respectively

$$\partial_t \chi^2(\mu_t \,\|\, \pi) \le -\frac{2}{9C_{\mathsf{LSI}}} \begin{cases} 2\chi^2(\mu_t \,\|\, \pi)^{3/2}, & \text{if } \chi^2(\mu_t \,\|\, \pi) \ge e - 1, \\ \chi^2(\mu_t \,\|\, \pi), & \text{otherwise.} \end{cases} \tag{4.11}$$

Solving the differential inequality in the first case yields

$$e - 1 \le \chi^2(\mu_t \,\|\, \pi) \le \Bigl[\frac{9C_{\mathsf{LSI}}\sqrt{\chi^2(\mu_0 \,\|\, \pi)}}{9C_{\mathsf{LSI}} + 2t\sqrt{\chi^2(\mu_0 \,\|\, \pi)}}\Bigr]^2 \le \Bigl[\frac{9C_{\mathsf{LSI}}}{2t}\Bigr]^2,$$

so that in this first case, it must holds that

$$t \leq \frac{9C_{\mathsf{LSI}}}{2\sqrt{e-1}} < 3.5C_{\mathsf{LSI}} =: t_0.$$

Therefore, if $t \geq t_0$, we are in the second case. In particular, $\chi^2(\mu_{t_0} \,\|\, \pi) \leq e - 1 \leq 2$ and integrating the differential inequality between t_0 and t we get

$$\chi^2(\mu_t \,\|\, \pi) \leq \chi^2(\mu_{t_0} \,\|\, \pi)\, e^{-\frac{2(t-t_0)}{9C_{\mathsf{LSI}}}} \leq \big(\chi^2(\mu_0 \,\|\, \pi) \wedge 2\big)\, e^{-\frac{2(t-t_0)}{9C_{\mathsf{LSI}}}}\,,$$

where in the last inequality, we used the fact that $t \mapsto \chi^2(\mu_t \,\|\, \pi)$ is decreasing so that it also holds $\chi^2(\mu_{t_0} \,\|\, \pi) \leq \chi^2(\mu_0 \,\|\, \pi)$. In particular, taking $t \geq 2t_0 = 7C_{\mathsf{LSI}}$ yields the desired result. □

4.4 Laplacian Adjusted Wasserstein Gradient Descent (LAWGD)

While the previous section leads to a better understanding of the convergence properties of SVGD in the case that $\mathcal{K}_\pi$ is the identity operator, it is still unclear how to choose the kernel K to approach this idealized setup. For SVGD with a general kernel K, the calculation rules of Section 4.2 together with the method of the previous section yield the formula

$$\partial_t D_{\mathrm{KL}}(\mu_t \,\|\, \pi) = -\,\mathbb{E}_\pi \big\langle \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}, \mathcal{K}_\pi \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} \big\rangle,$$

for the dissipation of the KL divergence along SVGD. From this, a natural way to proceed is to seek an inequality of the form

$$\mathbb{E}_\pi \langle f, \mathcal{K}_\pi f \rangle \gtrsim \mathbb{E}_\pi[f^2], \qquad \text{for all locally Lipschitz } f \in L^2(\pi). \tag{4.12}$$

Applying this inequality to each coordinate of $\nabla(\mathrm{d}\mu_t/\mathrm{d}\pi)$ separately and using a Poincaré inequality would then allow us to conclude as in the proof of Theorem 3. The inequality (4.12) can be interpreted as a positive lower bound on the smallest eigenvalue of the operator $\mathcal{K}_\pi$. However, this approach is doomed to fail; under mild conditions on the kernel K, it is a standard fact that the eigenvalues of $\mathcal{K}_\pi$ form a sequence converging to 0, so no such spectral gap can hold.[2]

This suggests that any approach which seeks to prove finite-time convergence results for SVGD in the spirit of Theorem 3 must exploit finer properties of the eigenspaces of the operator $\mathcal{K}_\pi$. Motivated by this observation, we develop a new

[2] It is enough that K is a symmetric kernel with $K \in L^2(\pi \otimes \pi)$, and that π is not discrete (so that $L^2(\pi)$ is infinite-dimensional); see [BGL14, Section A.6].

algorithm called Laplacian Adjusted Wasserstein Gradient Descent (LAWGD) in which the kernel K is chosen carefully so that $\mathcal{K}_\pi = \mathscr{L}^{-1}$ is the inverse of the generator of the Langevin diffusion that has π as invariant measure.

More precisely, the starting point for our approach is the following integration-by-parts formula, which is a crucial component of the theory of Markov semigroups [BGL14]:

$$\mathbb{E}_\pi\langle \nabla f, \nabla g\rangle = \mathbb{E}_\pi[f \mathscr{L} g], \qquad \text{for all locally Lipschitz } f, g \in L^2(\pi), \tag{4.13}$$

where $\mathscr{L} := -\Delta + \langle \nabla V, \nabla \cdot\rangle$. The operator $\mathscr{L}$ is the (negative) generator of the standard Langevin diffusion with stationary distribution π [Pav14, §4.5]. We refer readers to Section 4.5 for background on the spectral theory of $\mathscr{L}$.

In order to use (4.13), we replace the vector field $-\mathcal{K}_\pi \nabla(\mathrm{d}\mu_t/\mathrm{d}\pi)$ by the vector field $-\nabla \mathcal{K}_\pi(\mathrm{d}\mu_t/\mathrm{d}\pi)$. The new dynamics follow the evolution equation

$$\partial_t \mu_t = \operatorname{div}\bigl(\mu_t \nabla \mathcal{K}_\pi \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr). \tag{LAWGD}$$

The vector field in the above continuity equation may also be written

$$-\nabla \mathcal{K}_\pi \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(x) = -\int \nabla_1 K(x, \cdot)\, \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\, \mathrm{d}\pi = -\int \nabla_1 K(x, \cdot)\, \mathrm{d}\mu_t.$$

Replacing μ_t by an empirical average over particles and discretizing the process in time, we again obtain an implementable algorithm, which we give as Algorithm 1.

Algorithm 1 LAWGD $(\mathsf{K}_{\mathscr{L}}, \mu_0)$

```
draw $N$ particles $X_0^{[1]}, \dots, X_0^{[N]} \overset{\text{i.i.d.}}{\sim} \mu_0$
for $t = 1, \dots, T-1$ do
    for $i = 1, \dots, N$ do
        $X_{t+1}^{[i]} \leftarrow X_t^{[i]} - \frac{h}{N} \sum_{j=1}^N \nabla_1 \mathsf{K}_{\mathscr{L}}(X_t^{[i]}, X_t^{[j]})$
    end for
end for
return $X_T^{[1]}, \dots, X_T^{[N]}$
```

A careful inspection of Algorithm 1 reveals that the update equation for the particles in Algorithm 1 does not involve the potential V directly, unlike the SVGD algorithm (4.3); thus, the kernel for LAWGD must contain all the information about V.

Our choice for the kernel K is guided by the following observation (based on (4.13)):

$$\partial_t D_{\mathrm{KL}}(\mu_t \,\|\, \pi) = -\,\mathbb{E}_\pi\langle \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}, \nabla \mathcal{K}_\pi \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\rangle = -\,\mathbb{E}_\pi\big[\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} \mathscr{L} \mathcal{K}_\pi \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\big].$$

As a result, we choose K to ensure that $\mathcal{K}_\pi = \mathscr{L}^{-1}$. This choice yields

$$\partial_t D_{\mathrm{KL}}(\mu_t \,\|\, \pi) = -\,\mathbb{E}_\pi\big[\big(\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} - 1\big)^2\big] = -\chi^2(\mu_t \,\|\, \pi). \tag{4.14}$$

It remains to see which kernel K implements $\mathcal{K}_\pi = \mathscr{L}^{-1}$. To that end, assume that $\mathscr{L}$ has a discrete spectrum and let (λ_i, ϕ_i), $i = 0, 1, 2, \dots$ be its eigenvalue-eigenfunction pairs where λ_js are arranged in nondecreasing order. Assume further that $\lambda_1 > 0$ (which amounts to a Poincaré inequality; see Section 4.5) and define the following *spectral kernel*:

$$\mathsf{K}_{\mathscr{L}}(x, y) = \sum_{i=1}^{\infty} \frac{\phi_i(x)\phi_i(y)}{\lambda_i} \tag{4.15}$$

We now show that this choice of kernel endows LAWGD with a remarkable property: it converges to the target distribution exponentially fast, with a rate which has no dependence on the Poincaré constant. Moreover, akin to CSF—see (4.5)—it also also exhibit strong uniform ergodicity.

Theorem 7. *Assume that $\mathscr{L}$ has a discrete spectrum and that π satisfies a Poincaré inequality* (P) *with some finite constant. Let $(\mu_t)_{t\ge 0}$ be the law of* LAWGD *with the kernel described above. Then,*

$$D_{\mathrm{KL}}(\mu_t \,\|\, \pi) \le \big(D_{\mathrm{KL}}(\mu_0 \,\|\, \pi) \wedge 2\big)\, e^{-t}, \qquad \forall\, t \ge 1.$$

Proof. In light of (4.14), the proof is identical to that of Theorem 3. □

The convergence rate in Theorem 7 has no dependence on the target measure. This *scale-invariant convergence* also appears in [Che+20b], where it is shown for the Newton-Langevin diffusion with a strictly log-concave target measure π. In Theorem 7, we obtain similar guarantees under the much weaker assumption of a Poincaré inequality; indeed, there are many examples of non-log-concave distributions which satisfy a Poincaré inequality [VW19].

4.5 Review of spectral theory

In this chapter, we consider elliptic differential operators of the form $\mathscr{L} = -\Delta + \langle \nabla V, \nabla \cdot \rangle$, where V is a continuously differentiable potential. In this section, we provide a brief review of the spectral theory of these operators, and we refer to [Eva10, §6.5] for a standard treatment.

The operator $\mathscr{L}$ (when suitably interpreted) is a linear operator defined on a domain $\mathscr{D} \subset L^2(\pi)$. For any locally Lipschitz function $f \in L^2(\pi)$, integration by parts shows that

$$\mathbb{E}_\pi[f \mathscr{L} f] = \mathbb{E}_\pi[\|\nabla f\|^2].$$

Therefore, $\mathscr{L}$ has a non-negative spectrum. Also, we have $\mathscr{L}1 = 0$, so that 0 is always an eigenvalue of $\mathscr{L}$. We say that $\mathscr{L}$ has a *discrete spectrum* if it has a countable sequence of eigenvalues $0 = \lambda_0 \le \lambda_1 \le \lambda_2 \le \lambda_3 \le \cdots$ and corresponding eigenfunctions $(\phi_i)_{i=1}^\infty$ which form a basis of $\mathscr{D}$. The eigenfunctions can be chosen to be orthogonal and normalized such that $\|\phi_i\|_{L^2(\pi)} = 1$; we always assume this is the case. Then, $\mathscr{L}$ can be expressed as

$$\mathscr{L} = \sum_{i=1}^\infty \lambda_i \, \langle \phi_i, \cdot \rangle_{L^2(\pi)} \, \phi_i.$$

The operator $\mathscr{L}$ has a discrete spectrum under the following condition ([Fri34], [RS78, Theorem XIII.67], [BGL14, Corollary 4.10.9]):

$$V_{\mathsf{S}} \in L^1_{\text{loc}}(\mathbb{R}^d), \qquad \inf V_{\mathsf{S}} > -\infty, \qquad \text{and} \qquad \lim_{\|x\| \to \infty} V_{\mathsf{S}}(x) = +\infty,$$

where $V_{\mathsf{S}} := -\Delta V + \frac{1}{2}\|\nabla V\|^2$. Moreover, under this condition we also have $\lambda_i \to \infty$ as $i \to +\infty$. For example, this condition is satisfied for $V(x) = \|x\|^\alpha$ for $\alpha > 1$, but not for $\alpha = 1$. In fact, for $\alpha = 1$, the spectrum of $\mathscr{L}$ is not discrete [BGL14, §4.1.1].

The *Poincaré inequality* (P) is interpreted as a *spectral gap inequality*, since it asserts that $\lambda_1 = 1/C_{\mathsf{P}} > 0$. Thus, under a Poincaré inequality, $\mathscr{L} : \mathscr{D} \cap \{f \in L^2(\pi) \mid \mathbb{E}_\pi f = 0\} \to L^2(\pi)$ is bijective. Moreover, if it has a discrete spectrum, its inverse satisfies

$$\mathscr{L}^{-1} = \sum_{i=1}^\infty \lambda_i^{-1} \, \langle \phi_i, \cdot \rangle_{L^2(\pi)} \, \phi_i.$$

4.6 Numerical experiments

We compared the performances of LAWGD with SVGD in some experiments. All methods were implemented in `Python`. Since the Schrödinger operator requires the Laplacian and gradient of the potential V, we employ automatic differentiation to avoid laborious calculations of these derivatives.

To implement Algorithm 1, we numerically approximate the kernel $K = \mathsf{K}_{\mathscr{L}}$ given in (4.15). When π is the standard Gaussian distribution on $\mathbb{R}$, the eigendecomposition of the operator $\mathscr{L}$ in (4.13) is known explicitly [BGL14,

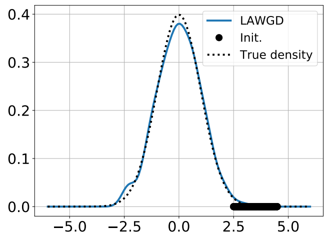

Figure 4-1: Samples from the standard Gaussian distribution generated by LAWGD, with kernel approximated by Hermite polynomials.

§2.7.1]. Specifically, the *probabilists'* Hermite polynomials are well-known to be eigenfunctions of the 1D Ornstein-Uhlenbeck operator $\mathscr{L}$ given by $\mathscr{L} f(x) := -f''(x)+xf'(x)$, and they satisfy the recursive relationship $H_{n+1}(x) = xH_n(x) - nH_{n-1}(x)$, with $H_0(x) = 1$ and $H_1(x) = x$. It also holds that $H_n'(x) = nH_{n-1}(x)$. With these equations, it is easy to check that the eigenvalue corresponding to H_n is $\lambda_n = n$. These are used as the eigenfunctions and eigenvalues, and we approximate the kernel via a truncated sum: $\hat{K}(x,y) = \sum_{i=1}^{k} \lambda_i^{-1} \phi_i(x)\phi_i(y)$ (Figure 4-1) involving the smallest eigenvalues of $\mathscr{L}$. In Figure 4-1 we use the first $k = 150$ Hermite polynomials, and we run LAWGD for 2000 iterations with a constant step size, with initial points drawn uniformly from the interval $[2.5, 4.5]$.

In the general case, we implement a basic finite difference (FD) method to approximate the eigenvalues and eigenfunctions of $\mathscr{L}$. We obtain better numerical results by first transforming the operator $\mathscr{L}$ into the Schrödinger operator $\mathscr{L}_{\mathsf{S}} := -\Delta + V_{\mathsf{S}}$, where $V_{\mathsf{S}} := \frac{1}{4}\|\nabla V\|^2 - \frac{1}{2}\Delta V$. If ϕ_{S} is an eigenfunction of $\mathscr{L}_{\mathsf{S}}$ with eigenvalue λ (normalized such that $\int \phi_{\mathsf{S}}^2 = 1$), then $\phi := e^{V/2}\phi_{\mathsf{S}}$ is an eigenfunction of L also with eigenvalue λ (and normalized such that $\int \phi^2 \, \mathrm{d}\pi = 1$); see [BGL14, §1.15.7].

On a grid of points (with spacing ε), if we replace the Laplacian with the FD operator $\Delta_\varepsilon f(x) := \{f(x-\varepsilon) + f(x+\varepsilon) - 2f(x)\}/\varepsilon^2$ (in 1D), then the FD Schrödinger operator $\mathscr{L}_{\mathsf{S},\varepsilon} := -\Delta_\varepsilon + V_{\mathsf{S}}$ can be represented as a sparse matrix, and its eigenvalues and (unit) eigenvectors are found with standard linear algebra solvers.

When the potential V is known only up to an additive constant, then the approximate eigenfunctions produced by this method are not normalized correctly; instead, they satisfy $\|\phi\|_{L^2(\pi)} = C$ for some constant C (which is the same for each eigenfunction). In turn, this causes the kernel K in LAWGD to be off by a multiplicative constant. For implementation purposes, however, this constant is absorbed in the step size of Algorithm 1. We also note that the eigenfunctions are differentiated using a FD approximation.

To demonstrate, we sample from a mixture of three Gaussians: $\frac{2}{5}\mathcal{N}(-3,1) + \frac{1}{5}\mathcal{N}(0,1) + \frac{2}{5}\mathcal{N}(4,2)$. We compare LAWGD with SVGD using the RBF kernel

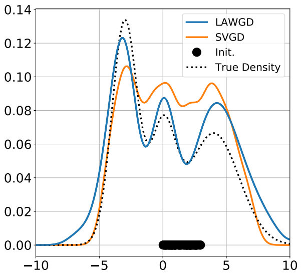

Figure 4-2: LAWGD and SVGD run with constant step size for a mixture of three Gaussians. Both kernel density estimators use the same bandwidth.

and median-based bandwidth as in [LW16]. We approximate the eigenfunctions and eigenvalues using a finite difference scheme, on 256 grid points evenly spaced between -14 and 14. Constant step sizes for LAWGD and SVGD are tuned and the algorithms are run for 5000 iterations, and the samples are initialized to be uniform on $[1, 4]$. The results are displayed in Figure 4-2. All 256 discrete eigenfunctions and eigenvalues are used.

In Figure 4-3, we display an example of sampling 50 particles from a mixture of two 2-dimensional Gaussian distributions given by $\pi = \frac{1}{2}\mathcal{N}((-1,-1)^\intercal, I_2) + \frac{1}{2}\mathcal{N}((1,1)^\intercal, I_2)$, using LAWGD and SVGD. To run this experiment, we use a 2-dimensional FD method, which approximates the Laplacian as

$$\Delta_\varepsilon f(x,y) := \frac{f(x-\varepsilon,y)+f(x+\varepsilon,y)+f(x,y-\varepsilon)+f(x,y+\varepsilon)-4f(x)}{\varepsilon^2}.$$

We again use the Schrödinger operator for stability and use FD again to compute the gradients of the eigenfunctions. We use a 128×128 grid of evenly spaced x and y values between -6 and 6. We calculate only the bottom 100 eigenvalues and eigenfunctions, since the other eigenfunctions incur additional computational cost without noticeably changing the result. Any negative eigenvalues (which arise from numerical errors) are discarded.

SVGD is run with the RBF kernel and median-based bandwith. True samples from π are displayed for comparison. Both LAWGD and SVGD are run for 20000 iterations with a constant step size. The samples from LAWGD tend to move very fast from their initial positions and then tend to settle into their final positions as seen in Figure 4-3. On the other hand, with constant step size, the samples of SVGD do not seem to converge, and one must use a decreasing step size scheme in order for the particles to stabilize. We also note that many of the samples generated by SVGD tend to blow up with a constant step size.

In Figure 4-4, we plot the particles of LAWGD and SVGD at iterations 100, 200, 1000, and 2000 to compare the speed of convergence.

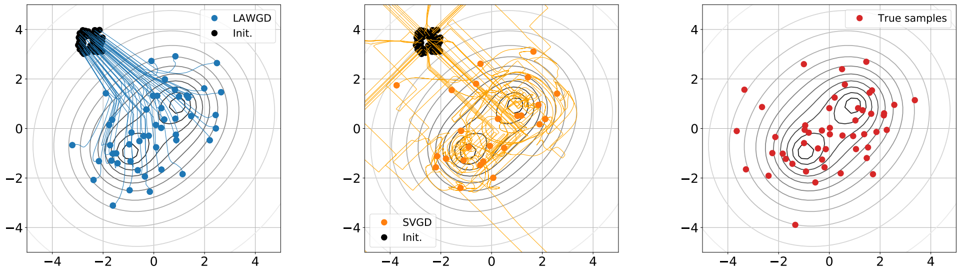

Figure 4-3: Left: 50 particles and trajectories generated from $\frac{1}{2}\mathcal{N}((-1,-1)^\intercal, I_2) + \frac{1}{2}\mathcal{N}((1,1)^\intercal, I_2)$ with LAWGD. Middle: 50 particles and trajectories generated by SVGD. Right: true samples from the distribution.

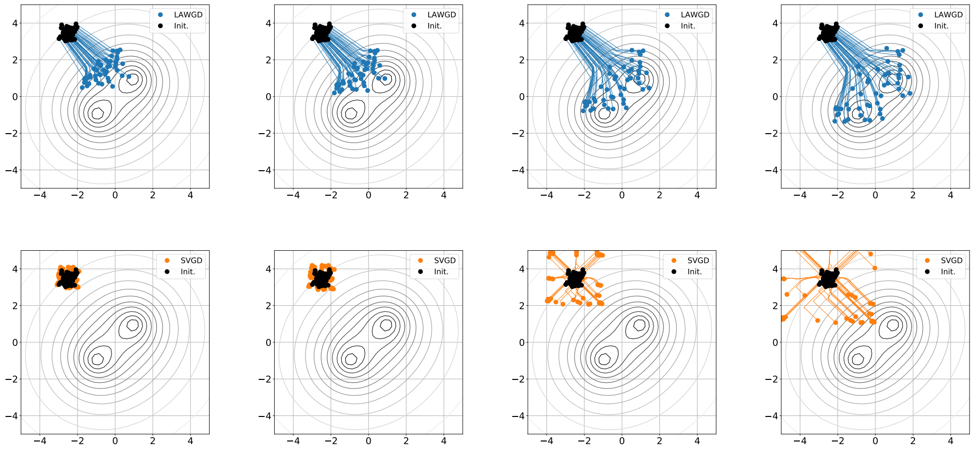

Figure 4-4: Top: LAWGD after 100, 200, 1000, and 2000 iterations. Bottom: SVGD after 100, 200, 1000, and 2000 iterations.

4.7 Conclusion

We conclude this chapter with some interesting open questions. The introduction of the chi-squared divergence as an objective function allows us to obtain both theoretical insights about SVGD and a new algorithm, LAWGD. This perspective opens the possibility of identifying other functionals defined over Wasserstein space and that yield gradient flows which are amenable to mathematical analysis and efficient computation. Towards this goal, an intriguing direction is to develop alternative methods, besides kernelization, which provide effective implementations of Wasserstein gradient flows. Finally, we note that LAWGD provides a hitherto unexplored connection between sampling and computing the spectral decomposition of the Schrödinger operator, the latter of which has been intensively studied in numerical PDEs. We hope our work further stimulates research at the intersection of these communities.

Chapter 5

MALA

5.1 Introduction

In this chapter we will analyse a widely used, practical implementation of the Langevin diffusion. A popular trick used to discretize continuous stochastic processes is Metropolis-Hastings (MH) adjustment [Met+53; Has70], which corrects for the bias in the naive discretization. The class of Metropolis-Hastings (MH) adjusted algorithms [Met+53; Has70], which includes the Random Walk Metropolis algorithm (RWM), the Metropolis-Adjusted Langevin Algorithm (MALA), and Hamiltonian Monte Carlo (HMC), is particularly popular in practice. Yet their convergence properties are still not well understood, especially their dependence on the problem dimension, a parameter of particular interest in modern applications. Using tools introduced in previous chapters, we will focus on the analysis of MALA and its dimension dependence.

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. [RGG+97] initiated the study of dimension dependence of RWM by means of an asymptotic framework: namely, when π is a product distribution, a scaling limit exists for RWM 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 RWM 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 implications, 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 derivatives; 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 chapter. 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.

This chapter is based on the joint work [**chewietal2020mala**], with Sinho Chewi, Kwangjun Ahn, Xiang Cheng, Thibaut Let Gouic, and Philippe Rigollet.

Related works. 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, RWM can draw samples with at most ε error in chi-squared divergence with $O(d \log \frac{1}{\varepsilon})$ steps, thereby providing a non-asymptotic affirmation of the scaling limit of [RGG+97]. 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 [Che+20a], which proves a complexity of $O(d \log \frac{1}{\varepsilon})$ steps to achieve ε error in chi-squared divergence. They also raise the question of whether there is a gap between the complexities of RWH and MALA.

[MV19] took a direct aim at improving the dimension dependence of mixing time bounds for MALA. They succeeded in obtaining a bound of $O(d^{2/3})$ 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 $O(d^{2/3})$; see the discussion in [Che+20a].

Our contributions. In this chapter, 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], 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 addition to establishing the optimal dimension dependence for MALA, our result is also one of the strongest guarantees for sampling with a warm start to-date, irrespective of the algorithm. Indeed, the algorithms which achieve similar or better dimension dependence compared to our result are: the underdamped Langevin algorithm [Che+18c, $O(d^{1/2})$], the higher-order Langevin algorithm [Mou+20, $O(d^{1/2})$], the randomized midpoint discretization of underdamped Langevin [SL19, $O(d^{1/3})$], and Hamiltonian Monte Carlo [MV18, $O(d^{1/4})$]. However, the dependence of these results on $1/\varepsilon$ is polynomial, whereas our dependence on $1/\varepsilon$ is polylogarithmic. Therefore, for a wide range of accuracy values which are inverse polynomial in the dimension (e.g., $\varepsilon = 1/d$), our result attains the best-known dependence on the dimension.

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; Dal17c]. In this way, our upper bound has the potential to connect the vast literature on discretization of SDEs with the more difficult analysis of Metropolised 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

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 assume that $\beta \geq 1 \geq \alpha$, and we denote by $\kappa := \beta/\alpha$ the *condition number.*

For the sake of normalization, we assume that $V(\mathbf{0}) = \min V = 0$, so that $\nabla V(\mathbf{0}) = \mathbf{0}$.

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

$$Q(\boldsymbol{x}, \cdot) := \frac{1}{(4\pi h)^{d/2}} \exp\Bigl(-\frac{\|\cdot - \boldsymbol{x} + h\nabla V(\boldsymbol{x})\|^2}{4h}\Bigr).$$

   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

$$A(\boldsymbol{x}, \boldsymbol{y}) := 1 \wedge a(\boldsymbol{x}, \boldsymbol{y}) \;, \qquad a(\boldsymbol{x}, \boldsymbol{y}) := \frac{\pi(\boldsymbol{y}) Q(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x}) Q(\boldsymbol{x}, \boldsymbol{y})} \,. \tag{5.1}$$

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

$$\begin{aligned} T(\boldsymbol{x},\boldsymbol{y}) &= [1 - A(\boldsymbol{x})]\,\delta_{\boldsymbol{x}}(\boldsymbol{y}) + Q(\boldsymbol{x},\boldsymbol{y})A(\boldsymbol{x},\boldsymbol{y}),\\ A(\boldsymbol{x}) &= \int Q(\boldsymbol{x},\boldsymbol{y})A(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \geq 0. \end{aligned} \tag{5.2}$$

For the rest of the paper, 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. In this chapter, 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:

$$\tau_{\mathrm{mix}}(\varepsilon,\mu_0;\mathsf{d}) := \inf\{n \in \mathbb{N} \,:\, \boldsymbol{x}_0 \sim \mu_0,\ \mathsf{d}(\mu_n,\pi) \leq \varepsilon\}\,.$$

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 $(Tg)(\boldsymbol{x}) := \int g(\boldsymbol{y})\,T(\boldsymbol{x},\mathrm{d}\boldsymbol{y})$. The *spectral gap* is defined as

$$\lambda := \inf\Bigl\{\frac{\mathcal{E}(f,f)}{\operatorname{var} f} \,:\, f \in L^2(\pi),\ \operatorname{var} f > 0\Bigr\}\,. \tag{λ}$$

Since it is often difficult to control the spectral gap directly, it is also convenient to introduce the *conductance*, defined as

$$\mathsf{C} := \inf\Bigl\{\frac{\int_S T(\boldsymbol{x},S^{\mathrm{c}})\,\pi(\mathrm{d}\boldsymbol{x})}{\pi(S)} \,:\, S \subseteq \mathbb{R}^d,\ \pi(S) \leq \frac{1}{2}\Bigr\}\,. \tag{C}$$

By Cheeger's inequality [LS88], it holds that

$$\mathsf{C}^2 \lesssim \lambda \lesssim \mathsf{C}. \tag{5.3}$$

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 Section 5.8. Though it is expected, this result appears to be new.

Theorem 8 (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 Section 5.5, we show that a tighter lower bound may be achieved using a carefully chosen perturbation of the Gaussian distribution.

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 2 (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 $\mu_0(E) \le M_0 \pi(E)$. 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 9. *Fix $\varepsilon > 0$ and consider a target distribution π satisfying the assumptions of Section 5.2. Then MALA with a M_0-warm start and step size*

$$h = \frac{c\alpha^{1/2}}{\beta^{4/3} d^{1/2} \log(d\kappa M_0/\varepsilon)}$$

for a sufficiently small absolute constant $c > 0$, has mixing time given by

$$\tau_{\text{mix}}(\varepsilon, \mu_0; \mathsf{d}) \lesssim \frac{\beta^{4/3} d^{1/2}}{\alpha^{3/2}} \log\Bigl(\frac{M_0}{\varepsilon}\Bigr) \log\Bigl(d\kappa + \frac{M_0}{\varepsilon}\Bigr)\,.$$

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 10. 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 chapter.

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 leveraging 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, Chapter 2] and ultimately to the Wasserstein distance using Talagrand's transportation inequality for strongly log-concave distributions.

The bound above is likely not sharp in terms of the accuracy parameter ε and the warm start parameter M_0. Indeed, we expect the dependency on the accuracy parameter to be $\log(1/\varepsilon)$, and the paper [Che+20a] develops a method, based on the conductance profile, to reduce the warm start dependence to $\log\log M_0$. Since the quantity $\log M_0$ can introduce additional dimensional factors under a feasible start [Dwi+19], it is important to improve the dependency on M_0. We leave open the question of refining our techniques to achieve these improvements.

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 1. *Suppose that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ with $\|\boldsymbol{x} - \boldsymbol{y}\| \le r$, it holds that $\|T_{\boldsymbol{x}} - T_{\boldsymbol{y}}\|_{\mathrm{TV}} \le 3/4$. 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:

$$\|T_{\boldsymbol{x}} - T_{\boldsymbol{y}}\|_{\mathrm{TV}} \le \|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} + \|Q_{\boldsymbol{x}} - Q_{\boldsymbol{y}}\|_{\mathrm{TV}} + \|T_{\boldsymbol{y}} - Q_{\boldsymbol{y}}\|_{\mathrm{TV}}. \tag{5.4}$$

The middle term is the TV distance between two Gaussian distributions, and using Pinsker's inequality it is straightforward to show that

$$\|Q_{\boldsymbol{x}} - Q_{\boldsymbol{y}}\|_{\mathrm{TV}} \le \frac{\|\boldsymbol{x} - \boldsymbol{y}\|}{\sqrt{2h}}, \qquad \text{provided } h \le \frac{2}{\beta},$$

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. [Dwi+19] show that these terms can be controlled when the step size is of order $h \approx 1/d$. An application of Lemma 1 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 is $O(d)$.

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 2. *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*

$$\mathbb{E}\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \le 2\,\mathbb{E}\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}\,.$$

Proof. See Section 5.6. □

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 (LD), which we restate here for convenience:

$$\mathrm{d}\bar{\boldsymbol{X}}_t = -\nabla V(\bar{\boldsymbol{X}}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t, \qquad \bar{\boldsymbol{X}}_0 = \boldsymbol{x}, \tag{5.5}$$

and $(\boldsymbol{B}_t)_{t\geq 0}$ is a standard Brownian motion. Using standard arguments from stochastic calculus (see (5.11)), we show that $\mathbb{E}\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} = O(h\sqrt{d})$ (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 1. To remedy this, we prove a new pointwise version of the projection characterization of Metropolis adjustment.

Theorem 10. *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 $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \leq 2\,\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} + \int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\,\Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|\,\mathrm{d}\boldsymbol{y}. \tag{5.6}$$

Consequently, for any convex increasing function $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ and $\boldsymbol{x} \sim \pi$, $\boldsymbol{y} \sim \bar{Q}(\boldsymbol{x},\cdot)$,

$$\mathbb{E}\,\Phi(\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}) \leq \frac{1}{2}\,\mathbb{E}\,\Phi(4\,\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}) + \frac{1}{2}\,\mathbb{E}\,\Phi\big(2\,\Big|\frac{Q(\boldsymbol{x},\boldsymbol{y})}{\bar{Q}(\boldsymbol{x},\boldsymbol{y})} - 1\Big|\big). \tag{5.7}$$

Proof. See Section 5.6. □

Remark 4. If we take the expectation of (5.6) when $\boldsymbol{x} \sim \pi$, we obtain

$$\mathbb{E}\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \leq 4\,\mathbb{E}\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}\,,$$

which qualitatively recovers Lemma 2.

The second inequality in Theorem 10 can be used in the usual way to deduce concentration bounds for $\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$ 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 $\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$, it suffices to apply well-established techniques for studying the discretization of SDEs.

Once we show that $\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$ is controlled with high probability, we are then able to apply a conductance argument, similar to Lemma 1, in order to prove our mixing time bound. We give an in-depth overview of the proof and provide proofs of technical details in Section 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 11. *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 $d^{-1/2}$.

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

[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 $\tau_{\text{mix}}(\varepsilon, \mu_0; \sqrt{\chi^2}) = \widetilde{\Omega}(\lambda^{-1})$. See, e.g., [BGL14, Chapter 4] for a rigorous treatment of spectral theory.

hence, cannot be improved uniformly over the class of distributions with smooth and strongly convex potentials.

Theorem 12. *The spectral gap λ of MALA with target distribution π_η and step size $0 < h \le 1$ satisfies $\lambda \lesssim h$.*

We give the proofs of these theorems in Section 5.7.

5.6 Proof of the upper bound

This section presents the proof of Theorem 9.

High-level overview of the proof

The bulk of the proof controls the mixing time in total variation and we use results from Section 5.6 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

$$\mathsf{C}_s := \inf\Bigl\{\frac{\int_S T(\boldsymbol{x}, S^{\mathrm{c}})\,\pi(\mathrm{d}\boldsymbol{x})}{\pi(S) - s} \Bigm| S \subseteq \mathbb{R}^d,\ s < \pi(S) \le \frac{1}{2}\Bigr\}\,. \tag{5.9}$$

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 3 ([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*

$$\|\mu_n - \pi\|_{\mathrm{TV}} \le M_0 s + M_0 \exp\bigl(-\frac{\mathsf{C}_s^2 n}{2}\bigr),$$

where M_0 is the warm start parameter of μ_0.

Corollary 5. *Taking $s = \varepsilon/(2M_0)$, it follows that*

$$\|\mu_n - \pi\|_{\mathrm{TV}} \le \varepsilon \qquad \textit{provided that } n \ge \frac{2}{\mathsf{C}_s^2} \ln \frac{2M_0}{\varepsilon}.$$

Motivated by the standard conductance lemma (Lemma 1) and the decomposition (5.4), in order to bound the s-conductance from below we will first bound $\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$, as in Section 5.4. The outline of the proof is as follows:

1. In Section 5.6, we prove the projection properties of MALA (Lemma 2 and Theorem 10).

2. In Section 5.6, we use the projection property (Lemma 2) along with stochastic calculus to bound the expectation $\mathbb{E}\,\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$ when $\boldsymbol{x} \sim \pi$.

3. In Section 5.6, we use the pointwise projection property, together with more stochastic calculus, in order to prove a concentration inequality for $\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$ when $\boldsymbol{x} \sim \pi$.

4. In Section 5.6, we use the concentration bound of Section 5.6, together with ideas from the proof of the standard conductance lemma (Lemma 1), in order to lower bound the s-conductance. Together with Corollary 5, it yields the mixing time bound of Theorem 9 in total variation distance.

5. Finally in Section 5.6, we explain how the mixing time bound in total variation distance implies mixing time bounds in other distances between probability measures.

Proof of the projection properties

We start with a basic fact about MALA.

Proposition 1. *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

$$\|Q_{\boldsymbol{x}} - T_{\boldsymbol{x}}\|_{\mathsf{TV}} = \frac{1}{2} \Bigl(T_{\boldsymbol{x}}(\{\boldsymbol{x}\}) + \int_{\mathbb{R}^d \setminus \{\boldsymbol{x}\}} |T(\boldsymbol{x}, \boldsymbol{y}) - Q(\boldsymbol{x}, \boldsymbol{y})| \, \mathrm{d}\boldsymbol{y} \Bigr) .$$

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 2 and Theorem 10).

Proof of Lemma 2. Since the transition kernel $\bar{Q}$ corresponding to the continuous-time Langevin diffusion is reversible with stationary distribution π, it follows from [BD01] that

$$\iint_{(\mathbb{R}^d\times\mathbb{R}^d)\setminus\Delta} |T(\boldsymbol{x},\boldsymbol{y}) - Q(\boldsymbol{x},\boldsymbol{y})|\,\pi(\mathrm{d}\boldsymbol{x})\,\mathrm{d}\boldsymbol{y} \le \iint_{(\mathbb{R}^d\times\mathbb{R}^d)\setminus\Delta} |\bar{Q}(\boldsymbol{x},\boldsymbol{y}) - Q(\boldsymbol{x},\boldsymbol{y})|\,\pi(\mathrm{d}\boldsymbol{x})\,\mathrm{d}\boldsymbol{y}\,,$$

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 $2\,\mathbb{E}_{\boldsymbol{x}\sim\pi}\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$. On the other hand, the left-hand side is equal to $\mathbb{E}_{\boldsymbol{x}\sim\pi}\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$ due to Proposition 1. □

Proof of Theorem 10. For any $\boldsymbol{x}$, we have

$$\begin{aligned}
\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} &= \int \{1 - A(\boldsymbol{x},\boldsymbol{y})\}\,Q(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \\
&= \int \Big[1 - \Big(1 \wedge \frac{\pi(\boldsymbol{y})Q(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x},\boldsymbol{y})}\Big)\Big]\,Q(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \\
&\le \int \Big|1 - \frac{\pi(\boldsymbol{y})Q(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x},\boldsymbol{y})}\Big|\,Q(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \\
&\le \int \Big|1 - \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x},\boldsymbol{y})}\Big|\,Q(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} + \int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\,\Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|\,\mathrm{d}\boldsymbol{y}.
\end{aligned}$$

Observe that the first term is given by

$$\int \Big|1 - \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x},\boldsymbol{y})}\Big|\,Q(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = \int \Big|Q(\boldsymbol{x},\boldsymbol{y}) - \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\Big|\,\mathrm{d}\boldsymbol{y} = 2\,\|Q_{\boldsymbol{x}} - \bar{Q}_{\boldsymbol{x}}\|_{\mathrm{TV}}\,,$$

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,

$$\begin{aligned}
\mathbb{E}\,\Phi(\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}) &\le \mathbb{E}\,\Phi\Big(2\,\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} + \int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\,\Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|\,\mathrm{d}\boldsymbol{y}\Big) \\
&\le \frac{1}{2}\,\mathbb{E}\,\Phi(4\,\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}) + \frac{1}{2}\,\mathbb{E}\,\Phi\Big(2\int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\,\Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|\,\mathrm{d}\boldsymbol{y}\Big)\,,
\end{aligned}$$

where we take expectation with respect to $\boldsymbol{x} \sim \pi$. Next, nothing 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

$$\begin{aligned}
&\mathbb{E}\,\Phi\Big(2\int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\,\Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|\,\mathrm{d}\boldsymbol{y}\Big) \\
&\qquad = \int \Phi\Big(2\int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\,\Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|\,\mathrm{d}\boldsymbol{y}\Big)\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}
\end{aligned}$$

$$\leq \iint \Phi\big(2\,\Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})}-1\Big|\big)\,\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}$$
$$= \iint \Phi\big(2\,\Big|\frac{Q(\boldsymbol{x},\boldsymbol{y})}{\bar{Q}(\boldsymbol{x},\boldsymbol{y})}-1\Big|\big)\,\pi(\boldsymbol{x})\bar{Q}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}\,,$$

where we switched $\boldsymbol{x}$ and $\boldsymbol{y}$ in the notation of the last line. □

Expectation of the total variation

We now bound $\mathbb{E}\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$ when $\boldsymbol{x} \sim \pi$ using the projection property (Lemma 2). Akin to prior work such as [DT12], our primary tool to analyse the discretization of the Langevin diffusion is the Girsanov theorem from stochastic calculus [see, e.g. Le 16; SV06, for classical treatments].

Lemma 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:

$$\frac{\mathrm{d}\mathbf{Q}_{\boldsymbol{x}}}{\mathrm{d}\bar{\mathbf{Q}}_{\boldsymbol{x}}}\big((\bar{\boldsymbol{X}}_t)_t\big) = \exp\Big[\frac{1}{\sqrt{2}}\int_0^h \langle\nabla V(\bar{\boldsymbol{X}}_t) - \nabla V(\boldsymbol{x}), \mathrm{d}\boldsymbol{B}_t\rangle - \frac{1}{4}\int_0^h \|\nabla V(\bar{\boldsymbol{X}}_t) - \nabla V(\boldsymbol{x})\|^2\,\mathrm{d}t\Big].$$

Proof. See the proof of Proposition 2 in [DT12]. □

In the following lemma, we use Lemma 11.

Lemma 5. *Assume $h \leq 1/(3\beta^{4/3})$. For any $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \leq \frac{1}{2}\beta h\sqrt{d + \beta^{2/3}\,\|\boldsymbol{x}\|^2}\,.$$

Proof. Let end denote the function that maps a continuous curve $(y_t)_{t\in[0,h]}$ in $\mathbb{R}^d$ to its endpoint: $\mathsf{end}((y_t)_{t\in[0,h]}) := y_h$. 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

$$\mathsf{KL}(\bar{Q}_{\boldsymbol{x}} \parallel Q_{\boldsymbol{x}}) = \mathsf{KL}(\mathsf{end}_{\#}\bar{\mathbf{Q}}_{\boldsymbol{x}} \parallel \mathsf{end}_{\#}\mathbf{Q}_{\boldsymbol{x}}) \le \mathsf{KL}(\bar{\mathbf{Q}}_{\boldsymbol{x}} \parallel \mathbf{Q}_{\boldsymbol{x}})\,.$$

On the other hand, the Girsanov theorem (in the form of Lemma 4) implies that

$$\begin{aligned}\mathsf{KL}(\bar{\mathbf{Q}}_{\boldsymbol{x}} \parallel \mathbf{Q}_{\boldsymbol{x}}) &= -\mathbb{E}\ln\frac{\mathrm{d}\mathbf{Q}_{\boldsymbol{x}}}{\mathrm{d}\bar{\mathbf{Q}}_{\boldsymbol{x}}}(\bar{\boldsymbol{X}}_t) = \frac{1}{4}\int_0^h \mathbb{E}[\|\nabla V(\bar{\boldsymbol{X}}_t) - \nabla V(\boldsymbol{x})\|^2]\,\mathrm{d}t \\ &\le \frac{\beta^2}{4}\int_0^h \mathbb{E}[\|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2]\,\mathrm{d}t \le \frac{3\beta^2 h^2\,(d + \beta^{2/3}\,\|\boldsymbol{x}\|^2)}{8}\,,\end{aligned}$$

where we used the β-smoothness of V and Lemma 11. Now applying Pinsker's inequality, we obtain the desired inequality. □

It follows from Lemma 5 that when $\boldsymbol{x} \sim \pi$, we get

$$\mathbb{E}\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \le \frac{1}{2}\beta h\,\mathbb{E}\sqrt{d + \beta^{2/3}\,\|\boldsymbol{x}\|^2} \le \frac{1}{2}\beta h\sqrt{d + \beta^{2/3}\,\mathbb{E}[\|\boldsymbol{x}\|^2]} \lesssim \beta^{4/3} h\sqrt{\frac{d}{\alpha}}\,, \tag{5.10}$$

where we used the second moment bound of Lemma 10. Together with Lemma 2, it yields

$$\mathbb{E}\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \le 2\,\mathbb{E}\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \lesssim \beta^{4/3} h\sqrt{\frac{d}{\alpha}}\,. \tag{5.11}$$

We conclude this section with a concentration inequality which we use later in the argument.

Lemma 6. *Assume $h \le 1/(3\beta^{4/3})$ and let $\boldsymbol{x} \sim \pi$. For any $\delta > 0$, with probability at least $1 - \delta$,*

$$\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \lesssim \beta^{4/3} h\sqrt{\frac{d + \log(1/\delta)}{\alpha}}.$$

Proof. Let $f(\boldsymbol{x}) := \frac{1}{2}\beta^{4/3} h\sqrt{d + \|\boldsymbol{x}\|^2}$. Then,

$$\|\nabla f(\boldsymbol{x})\| = \frac{\beta^{4/3} h\,\|\boldsymbol{x}\|}{2\sqrt{d + \|\boldsymbol{x}\|^2}} \le \frac{1}{2}\beta^{4/3} h.$$

Thus, $f(\boldsymbol{x})$ is $\frac{1}{2}\beta^{4/3} h$-Lipschitz, and it follows from sub-Gaussian concentration (Lemma 10) that with probability at least $1 - \delta$,

$$f(\boldsymbol{x}) \le \mathbb{E}\,f(\boldsymbol{x}) + \beta^{4/3} h\sqrt{\frac{1}{2\alpha}\ln\frac{1}{\delta}}.$$

We have calculated $\mathbb{E}\, f(\boldsymbol{x}) \lesssim \beta^{4/3} h \sqrt{d/\alpha}$ in (5.10), and the result now follows from the pointwise bound in Lemma 5. $\square$

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 2. *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*

$$h = \frac{c_1 \alpha^{1/2}}{\beta^{4/3} d^{1/2} \log(d\kappa/s)}\,,$$

the following holds with probability at least $1 - c_0 s \sqrt{h}$,

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathsf{TV}} \le \frac{1}{6}\,.$$

The idea of the proof is to use the pointwise projection of Theorem 10, 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 6. To control the second term, we will first obtain a bound on its moments.

Lemma 7. *Let $k \ge 1$ be any integer. Suppose that*

$$h \le \frac{\alpha^{1/2}}{C \beta^{4/3} d^{1/2} k}\,, \qquad \textit{for a sufficiently large absolute constant } C > 0\,.$$

Then, it holds that

$$\Bigl\{\mathbb{E}_{\boldsymbol{x}\sim\pi}\Bigl[\Bigl|\int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\,\Bigl|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Bigr|\,\mathrm{d}\boldsymbol{y}\Bigr|^k\Bigr]\Bigr\}^{1/k} \lesssim \alpha^{-1/4}\beta h\sqrt{k}\,(\sqrt{d} + \sqrt{k})\,.$$

The proof, given in Section 5.6, uses extensively tools from stochastic calculus. We remark that the quantity in Lemma 7 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].

We are now in a position to prove Proposition 2.

Proof of Proposition 2. Assume that the step size h is small enough so that Lemmas 6 and 7 both hold. More specifically, since the requirement of Lemma 7 is more stringent than that of Lemma 6, so we can simply impose $h \le \frac{\alpha^{1/2}}{C\beta^{4/3} d^{1/2} k}$ for a sufficiently large absolute constant $C > 0$.

From Lemma 6 with $\delta = c_0 s\sqrt{h}/2$, there exists a constant $C_1 > 0$ such that with probability at least $1 - c_0 s\sqrt{h}/2$,

$$\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \le \frac{C_1 \beta^{4/3} h}{2\sqrt{\alpha}} \sqrt{d + \ln \frac{2}{c_0 s \sqrt{h}}}\,.$$

From Lemma 7 and Markov's inequality, there exists a constant $C_2 > 0$ such that for any $\delta > 0$, with probability at least $1 - \delta$,

$$\int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})} \Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big| \,\mathrm{d}\boldsymbol{y} \le C_2 \alpha^{-1/4} \beta h \sqrt{k}\,(\sqrt{d} + \sqrt{k})\,\delta^{-1/k}\,.$$

Taking $k \sim \ln \frac{2}{c_0 s\sqrt{h}}$ and $\delta = c_0 s\sqrt{h}/2$, we have $\delta^{-1/k} = \Theta(1)$ and hence

$$\int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})} \Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big| \,\mathrm{d}\boldsymbol{y} \le C_2 \alpha^{-1/4} \beta h \sqrt{\ln \frac{2}{c_0 s\sqrt{h}}}\,\Big(\sqrt{d} + \sqrt{\ln \frac{2}{c_0 s\sqrt{h}}}\Big)\,.$$

Combining these two inequalities with the pointwise projection property (Theorem 10), it follows that with probability at least $1 - c_0 s\sqrt{h}$,

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \le \frac{C_1 \beta^{4/3} h}{\sqrt{\alpha}} \sqrt{d + \ln \frac{2}{c_0 s\sqrt{h}}} + C_2 \alpha^{-1/4} \beta h \sqrt{\ln \frac{2}{c_0 s\sqrt{h}}}\,\Big(\sqrt{d} + \sqrt{\ln \frac{2}{c_0 s\sqrt{h}}}\Big)\,. \tag{5.12}$$

If we choose the constant $c_1 > 0$ small enough, then choosing the step size as in the statement of Proposition 2, i.e., $h = \frac{c_1 \alpha^{1/2}}{\beta^{4/3} d^{1/2} \log(d\kappa/s)}$, makes the both terms in the left-hand side of (5.12) less than $1/12$. This completes the proof of Proposition 2. □

Proof of Lemma 7

We now prove the moment upper bound (Lemma 7). 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

$$\begin{aligned}\int \pi(\boldsymbol{x}) \Big|\int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})} \Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big| \,\mathrm{d}\boldsymbol{y}\Big|^k \,\mathrm{d}\boldsymbol{x} &\le \iint \pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x}) \Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|^k \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} \\ &= \int \Big(\int \Big|\frac{Q(\boldsymbol{x},\boldsymbol{y})}{\bar{Q}(\boldsymbol{x},\boldsymbol{y})} - 1\Big|^k \bar{Q}(\boldsymbol{x},\mathrm{d}\boldsymbol{y})\Big)\,\pi(\mathrm{d}\boldsymbol{x})\,,\end{aligned}$$

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 $f(\boldsymbol{x}) := |x-1|^k$. Recall the definitions of $\bar{\mathbf{Q}}_{\boldsymbol{x}}$ and $\mathbf{Q}_{\boldsymbol{x}}$ in Lemma 4. Hence we may apply the data processing inequality and bound the above by

$$F_k := \int \Big(\int \Big|\frac{\mathrm{d}\mathbf{Q}_{\boldsymbol{x}}}{\mathrm{d}\bar{\mathbf{Q}}_{\boldsymbol{x}}} - 1\Big|^k \, \mathrm{d}\bar{\mathbf{Q}}_{\boldsymbol{x}}\Big)\, \pi(\mathrm{d}\boldsymbol{x})\,. \tag{5.13}$$

Recall from Lemma 4 that

$$\frac{\mathrm{d}\mathbf{Q}_{\boldsymbol{x}}}{\mathrm{d}\bar{\mathbf{Q}}_{\boldsymbol{x}}}(\bar{\boldsymbol{X}}) = \exp H_h\,,$$

where for $t \ge 0$,

$$H_t := \frac{1}{\sqrt{2}} \int_0^t \langle \nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x}), \mathrm{d}\boldsymbol{B}_s\rangle - \frac{1}{4}\int_0^t \|\nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x})\|^2 \, \mathrm{d}s\,.$$

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 (5.13) 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

$$\begin{aligned} F_k &\le (C\beta^2 k)^{k/2}\, \bar{\mathbf{E}}_{\boldsymbol{x}}\Big[\Big|\int_0^h \exp(4H_t)\, \mathrm{d}t\Big|^{k/4} \Big|\int_0^h \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^4 \, \mathrm{d}t\Big|^{k/4}\Big] \\ &\le (C\beta^2 k)^{k/2} \sqrt{\bar{\mathbf{E}}_{\boldsymbol{x}}\Big[\Big|\int_0^h \exp(4H_t)\, \mathrm{d}t\Big|^{k/2}\Big]\, \bar{\mathbf{E}}_{\boldsymbol{x}}\Big[\Big|\int_0^h \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^4 \, \mathrm{d}t\Big|^{k/2}\Big]} \\ &\le (C\beta^2 k)^{k/2}\, h^{k/2-1} \underbrace{\sqrt{\Big(\bar{\mathbf{E}}_{\boldsymbol{x}} \int_0^h \exp(2kH_t)\, \mathrm{d}t\Big)}}_{\textcircled{A}}\, \underbrace{\sqrt{\Big(\bar{\mathbf{E}}_{\boldsymbol{x}} \int_0^h \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^{2k} \, \mathrm{d}t\Big)}}_{\textcircled{B}}\,. \end{aligned}$$

We will control the two terms separately, starting with the first term $\textcircled{A}$.

Lemma 8. *Let $0 \le t \le h \le 1/(20\beta k)$. Then,*

$$\bar{\mathbf{E}}_{\boldsymbol{x}} \exp(2kH_t) \le \exp(96\beta^4 h^3 k^2 \|\boldsymbol{x}\|^2 + 576\beta^2 d h^2 k^2).$$

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

$$\bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Bigl(\int_0^t \langle \boldsymbol{Z}_s, \mathrm{d}\boldsymbol{B}_s\rangle - \frac{1}{2}\int_0^t \|\boldsymbol{Z}_s\|^2 \,\mathrm{d}s\Bigr) = 1\,.$$

Together with the Cauchy-Schwarz inequality, it yields

$$\begin{aligned}
&\bar{\mathbf{E}}_{\boldsymbol{x}} \exp(2kH_t)\\
&\quad= \bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Bigl[\sqrt{2}k\int_0^t \langle \nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x}), \mathrm{d}\boldsymbol{B}_s\rangle - \frac{k}{2}\int_0^t \|\nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x})\|^2\,\mathrm{d}s\Bigr]\\
&\quad= \bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Bigl[\sqrt{2}k\int_0^t \langle \nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x}), \mathrm{d}\boldsymbol{B}_s\rangle\\
&\qquad\qquad\qquad + \bigl(-4k^2 + 4k^2 - \frac{k}{2}\bigr)\int_0^t \|\nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x})\|^2\,\mathrm{d}s\Bigr]\\
&\quad\le \sqrt{\bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Bigl[8k^2\int_0^t \|\nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x})\|^2\,\mathrm{d}s\Bigr]}\\
&\quad\le \sqrt{\bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Bigl[8\beta^2 k^2\int_0^t \|\bar{\boldsymbol{X}}_s - \boldsymbol{x}\|^2\,\mathrm{d}s\Bigr]} \le \sqrt{\bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Bigl[8\beta^2 h k^2 \sup_{s\in[0,h]} \|\bar{\boldsymbol{X}}_s - \boldsymbol{x}\|^2\Bigr]}\,.
\end{aligned}$$

In order to upper bound the above quantity, we develop the following bound on the moment generating function of $\sup_{s\in[0,h]}\|\bar{\boldsymbol{X}}_s - \boldsymbol{x}\|^2$.

Lemma 9. *Assume $h \le 1/(2\beta)$. For $0 < \lambda < 1/(24h)$,*

$$\bar{\mathbf{E}}_{\boldsymbol{x}} \exp\bigl(\lambda \sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2\bigr) \le \exp\bigl(12\beta^2\lambda h^2 \|\boldsymbol{x}\|^2 + d\ln\frac{1+24h\lambda}{1-24h\lambda}\bigr).$$

Proof. The proof is deferred to §5.6. □

We use Lemma 9 with $\lambda := 8\beta^2 h k^2$. In order to satisfy the preconditions of Lemma 9, we impose the restriction $h \le \frac{1}{14\beta k}$. Then, it follows that

$$\begin{aligned}
\bar{\mathbf{E}}_{\boldsymbol{x}} \exp(2kH_t) &\le \exp\bigl(96\beta^4 h^3 k^2 \|\boldsymbol{x}\|^2 + d\ln\frac{1+192\beta^2 h^2 k^2}{1-192\beta^2 h^2 k^2}\bigr)\\
&\le \exp(96\beta^4 h^3 k^2 \|\boldsymbol{x}\|^2 + 576\beta^2 d h^2 k^2)\,,
\end{aligned}$$

where the last inequality is $\ln\frac{1+x}{1-x} \le 3x$, which holds provided $x \le 1/2$; this is valid provided $h \le \frac{1}{20\beta k}$. This is our desired bound. $\square$

Hence, from Lemma 8, we obtain

$$\textcircled{A} \le \sqrt{h\exp(96\beta^4 h^3 k^2 \left\|\boldsymbol{x}\right\|^2 + 576\beta^2 d h^2 k^2)}\,.$$

Next, we estimate $\textcircled{B}$. In fact, Lemma 9 together with standard moment bounds under sub-exponential concentration (e.g. [Ver18, Proposition 2.7.1]) gives

$$\bar{\mathbf{E}}_{\boldsymbol{x}} \sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^{2k} \le C^k\,(\beta^k h^{2k}\left\|\boldsymbol{x}\right\|^{2k} + d^k h^k + h^k k^k)\,,$$

where $C > 0$ is a numerical constant. See Corollary 6 in §5.6 for details. Hence, it holds that

$$\textcircled{B} = \int_0^h \bar{\mathbf{E}}_{\boldsymbol{x}}[\|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^{2k}]\,\mathrm{d}t \le C^k h\,(\beta^k h^{2k}\left\|\boldsymbol{x}\right\|^{2k} + d^k h^k + h^k k^k).$$

Hence,

$$\begin{aligned}
(5.13) &\le (C\beta^2 k)^{k/2} h^{k/2-1} \times \textcircled{A} \times \textcircled{B} \\
&\le (C\beta^2 k)^{k/2} h^{k/2-1} \times h^{1/2}\exp(48\beta^4 h^3 k^2 \left\|\boldsymbol{x}\right\|^2 + 288\beta^2 d h^2 k^2) \\
&\qquad\qquad \times \sqrt{C^k h\,(\beta^k h^{2k}\left\|\boldsymbol{x}\right\|^{2k} + d^k h^k + h^k k^k)} \\
&\le (C^2\beta^2 hk)^{k/2}\exp(288\beta^2 d h^2 k^2) \\
&\qquad\qquad \times \exp(48\beta^4 h^3 k^2\left\|\boldsymbol{x}\right\|^2)\sqrt{C^k h\,(\beta^k h^{2k}\left\|\boldsymbol{x}\right\|^{2k} + d^k h^k + h^k k^k)}.
\end{aligned}$$

Next, we take the expectation w.r.t. $\boldsymbol{x} \sim \pi$ and use Cauchy-Schwarz:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\pi}\,\bar{\mathbf{E}}_{\boldsymbol{x}}[|\exp H_h - 1|^k] \\
&\quad\le (C\beta^2 hk)^{k/2}\exp(288\beta^2 d h^2 k^2) \\
&\qquad\qquad \times \sqrt{\mathbb{E}_{\boldsymbol{x}\sim\pi}\exp(96\beta^4 h^3 k^2\left\|\boldsymbol{x}\right\|^2)\,\mathbb{E}_{\boldsymbol{x}\sim\pi}[\beta^k h^{2k}\left\|\boldsymbol{x}\right\|^{2k} + d^k h^k + h^k k^k]}\,.
\end{aligned}$$

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 10), 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 10 now imply the bound

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\pi}\,\bar{\mathbf{E}}_{\boldsymbol{x}}[|\exp H_h - 1|^k] \\
&\quad\le (C\beta^2 hk)^{k/2} \times (\alpha^{-k/2}\beta^{k/2} d^{k/2} h^k + \alpha^{-k/2}\beta^{k/2} h^k k^{k/2} + d^{k/2} h^{k/2} + h^{k/2} k^{k/2})\,.
\end{aligned}$$

Taking k-th roots,

$$\begin{aligned}&(\mathbb{E}_{\boldsymbol{x}\sim\pi}\,\bar{\mathbf{E}}_{\boldsymbol{x}}[|\exp H_h - 1|^k])^{1/k}\\&\qquad\lesssim \beta\sqrt{hk}\times(\alpha^{-1/2}\beta^{1/2}d^{1/2}h + \alpha^{-1/2}\beta^{1/2}hk^{1/2} + d^{1/2}h^{1/2} + h^{1/2}k^{1/2})\\&\qquad\lesssim \alpha^{-1/4}\beta h\sqrt{k}\,(\sqrt{d}+\sqrt{k}),\end{aligned}$$

provided that $h \le \alpha^{1/2}/\beta$. This concludes the proof.

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 1).

Towards the goal of applying the bound on the mixing time via s-conductance given in Corollary 5, we take $s := \varepsilon/(2M_0)$, and we choose the step size

$$h = \frac{c_1\alpha^{1/2}}{\beta^{4/3}d^{1/2}\log(d\kappa/s)} \tag{5.14}$$

as in Proposition 2. Then, Proposition 2 guarantees the existence of an event E with probability $\pi(E) \ge 1 - c_0 s\sqrt{h}$ 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:

$$\begin{aligned}S_1 &:= \bigl\{\boldsymbol{x}\in S \bigm| T(\boldsymbol{x},S^{\mathsf{c}}) \le \frac{1}{4}\bigr\}\,, &&\text{bad set 1}\\S_2 &:= \bigl\{\boldsymbol{x}\in S^{\mathsf{c}} \bigm| T(\boldsymbol{x},S) \le \frac{1}{4}\bigr\}\,, &&\text{bad set 2}\\S_3 &:= (S_1\cup S_2)^{\mathsf{c}}. &&\text{good set}\end{aligned}$$

If $\pi(S_1) < \pi(S)/2$ or $\pi(S_2) < \pi(S^{\mathsf{c}})/2$, then may conclude from reversibility of the MALA kernel T that

$$\int_S T(\boldsymbol{x},S^{\mathsf{c}})\,\pi(\mathrm{d}\boldsymbol{x}) = \frac{1}{2}\Bigl(\int_S T(\boldsymbol{x},S^{\mathsf{c}})\,\pi(\mathrm{d}\boldsymbol{x}) + \int_{S^{\mathsf{c}}} T(\boldsymbol{x},S)\,\pi(\mathrm{d}\boldsymbol{x})\Bigr) \ge \frac{1}{2}\cdot\frac{\pi(S)}{2}\cdot\frac{1}{4} = \frac{\pi(S)}{16}\,.$$

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

$$\|T_{\boldsymbol{x}} - T_{\boldsymbol{y}}\|_{\mathrm{TV}} \ge \frac{1}{2}\,.$$

Since $\boldsymbol{x}, \boldsymbol{y} \in E$, we also have

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \wedge \|T_{\boldsymbol{y}} - Q_{\boldsymbol{y}}\|_{\mathrm{TV}} \le \frac{1}{6}\,.$$

Thus, using the decomposition (5.4),

$$\begin{aligned}
\frac{1}{2} \le \|T_{\boldsymbol{x}} - T_{\boldsymbol{y}}\|_{\mathrm{TV}} &\le \|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} + \|Q_{\boldsymbol{x}} - Q_{\boldsymbol{y}}\|_{\mathrm{TV}} + \|T_{\boldsymbol{y}} - Q_{\boldsymbol{y}}\|_{\mathrm{TV}} \\
&\le \frac{1}{6} + \frac{\|\boldsymbol{x} - \boldsymbol{y}\|}{\sqrt{2h}} + \frac{1}{6}\,,
\end{aligned}$$

where the middle term is controlled via

$$\|Q_{\boldsymbol{x}} - Q_{\boldsymbol{y}}\|_{\mathrm{TV}} \le \frac{\|\boldsymbol{x} - \boldsymbol{y}\|}{\sqrt{2h}}\,, \qquad \text{if } h \le \frac{2}{\beta}\,,$$

see [Dwi+19, Lemma 3]. Hence, we obtain:

$$\frac{\sqrt{2h}}{6} \le \|\boldsymbol{x} - \boldsymbol{y}\|\,,$$

which implies that $\operatorname{dist}(E \cap S_1, E \cap S_2) \ge \sqrt{2h}/6$. By the isoperimetric inequality (see Lemma 10), there is an absolute constant $c > 0$ such that

$$\pi\bigl([(E \cap S_1) \cup (E \cap S_2)]^{\mathsf{c}}\bigr) \ge \frac{c\sqrt{2}}{6} \sqrt{\alpha h}\, \pi(E \cap S_1)\,.$$

Since S_1, S_2, and S_3 partition $\mathbb{R}^d$, we see that $((E \cap S_1) \cup (E \cap S_2))^{\mathsf{c}} = E^{\mathsf{c}} \cap S_3$. As a result,

$$\begin{aligned}
\pi(S_3) + c_0 s \sqrt{\alpha h} \ge \pi(S_3) + \pi(E^{\mathsf{c}}) &\ge \frac{c\sqrt{2}}{6} \sqrt{\alpha h}\, \pi(E \cap S_1) \\
&\ge \frac{c\sqrt{2}}{6} \sqrt{\alpha h}\, \{\pi(S_1) - \pi(E^{\mathsf{c}})\} \\
&\ge \frac{c\sqrt{2}}{6} \sqrt{\alpha h}\, \Bigl\{\frac{\pi(S)}{2} - \pi(E^{\mathsf{c}})\Bigr\} \\
&\ge \frac{c\sqrt{2}}{12} \sqrt{\alpha h}\, \pi(S)\,, \qquad (5.15)
\end{aligned}$$

where (5.15) follows since $\pi(S)/2 \ge s/2 \ge 2 c_0 s \sqrt{h} \ge 2\pi(E^{\mathsf{c}})$ provided that $c_0 \sqrt{h} \le 1/4$.

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 (5.14) small enough), we obtain

$$\pi(S_3) \geq \frac{c\sqrt{2}}{24} \sqrt{\alpha h}\, \pi(S)\,.$$

From this,

$$\begin{aligned}\int_S T(\boldsymbol{x}, S^{\mathsf{c}})\, \pi(\mathrm{d}\boldsymbol{x}) &= \frac{1}{2}\Bigl(\int_S T(\boldsymbol{x}, S^{\mathsf{c}})\, \pi(\mathrm{d}\boldsymbol{x}) + \int_{S^{\mathsf{c}}} T(\boldsymbol{x}, S)\, \pi(\mathrm{d}\boldsymbol{x})\Bigr) \\ &\geq \frac{1}{2} \cdot \frac{1}{4} \cdot \pi(S_3) \geq \frac{c\sqrt{2}}{192} \sqrt{\alpha h}\, \pi(S)\,.\end{aligned}$$

Collecting the arguments, we obtain a lower bound on the s-conductance.

Proposition 3. *If the step size h is chosen as (5.14) 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, we have proven Theorem 9.

Auxiliary lemmas

Standard facts about strongly log-concave measures

The following properties of strongly log-concave measures are well-known.

Lemma 10. *The α-strong convexity of V implies the following properties:*

1. *(moment and tail bounds) For $\boldsymbol{x} \sim \pi$, it holds that $\mathbb{E}\|\boldsymbol{x}\|^2 \leq d/\alpha$.*

   In fact, for all $k \geq 2$,

   $$\mathbb{E}\|\boldsymbol{x}\|^k \leq \frac{3^k\, (d^{k/2} + k^{k/2})}{\alpha^{k/2}}\,.$$

   Consequently, $\mathbb{E}\exp(\lambda\, \|\boldsymbol{x}\|^2)$ is bounded above by a universal constant, provided that $0 \leq \lambda \leq \alpha/(40d)$.

2. *(isoperimetry) For any $S \subseteq \mathbb{R}^d$ with $\pi(A) \leq 1/2$, it holds that $\pi(S^\varepsilon \setminus S) \gtrsim \varepsilon\sqrt{\alpha}\, \pi(S)$, where*

   $$S^\varepsilon := \{\boldsymbol{x} \in \mathbb{R}^d \mid \exists y \in S \text{ with } \|\boldsymbol{x} - \boldsymbol{y}\| \leq \varepsilon\}.$$

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 [DKR19, 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 paper [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 e.g. [BGL14, §5.4.2, Corollary 5.7.2].

Alternatively, these facts all follow from the corresponding facts about standard Gaussians, as a consequence of Caffarelli's contraction theorem [Caf00; FGP20]; see also the discussion in [Vil03, §9.2.3]. □

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 start with a bound on the mean squared displacement $\mathbb{E}[\|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2]$ of the Langevin diffusion.

Lemma 11. *If* $(\bar{\boldsymbol{X}}_t)_{t \ge 0}$ *denotes the continuous-time Langevin process* (5.5) *started at* $\boldsymbol{x}$*, then for all* $t \le 1/(3\beta^{4/3})$*, we have*

$$\mathbb{E}[\|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2] \le 3t\,(d + \beta^{2/3}\,\|\boldsymbol{x}\|^2)\,.$$

Proof. Fix $s \in [0, t]$. From Itô's lemma [Le 16, Theorem 5.10], we have

$$\begin{aligned} \mathbb{E}[\|\bar{\boldsymbol{X}}_s - \boldsymbol{x}\|^2] &= \mathbb{E} \int_0^s \Bigl\{-2\,\langle \nabla V(\bar{\boldsymbol{X}}_u), \bar{\boldsymbol{X}}_u - \boldsymbol{x}\rangle + \frac{1}{2} \cdot 2d\Bigr\}\,\mathrm{d}u \\ &= \mathbb{E} \int_0^s \{-2\,\langle \nabla V(\bar{\boldsymbol{X}}_u), \bar{\boldsymbol{X}}_u - \boldsymbol{x}\rangle\}\,\mathrm{d}u + sd\,. \end{aligned}$$

To upper bound the first term on the right-hand side, we could conclude easily using a convexity of V with slightly different dependence on β in the final result. Instead, we take somewhat of a detour to show that this results hinges solely on the smoothness of V and can therefore be extended beyond the log-concave case.

Note that

$$\begin{aligned}|\langle \nabla V(\bar{\boldsymbol{X}}_u), \bar{\boldsymbol{X}}_u - \boldsymbol{x}\rangle| &\le |\langle \nabla V(\bar{\boldsymbol{X}}_u) - \nabla V(\boldsymbol{x}), \bar{\boldsymbol{X}}_u - \boldsymbol{x}\rangle| + |\langle \nabla V(\boldsymbol{x}), \bar{\boldsymbol{X}}_u - \boldsymbol{x}\rangle| \\ &\le \beta \, \|\bar{\boldsymbol{X}}_u - \boldsymbol{x}\|^2 + \frac{1}{2\beta^{4/3}} \, \|\nabla V(\boldsymbol{x})\|^2 + \frac{\beta^{4/3}}{2} \, \|\bar{\boldsymbol{X}}_u - \boldsymbol{x}\|^2 \\ &\le \frac{3\beta^{4/3}}{2} \, \|\bar{\boldsymbol{X}}_u - \boldsymbol{x}\|^2 + \frac{\beta^{2/3}}{2} \, \|\boldsymbol{x}\|^2 \,,\end{aligned}$$

where the last two inequalities follow from β-smoothness of V (see e.g. [Nes18, Theorem 2.1.5]), and our assumption $\arg\min V = \mathbf{0}$. Thus, letting $a(u) := \mathbb{E}[\|\bar{\boldsymbol{X}}_u - \boldsymbol{x}\|^2]$, we obtain the following integral inequality:

$$a(s) \le (d + \beta^{2/3} \, \|\boldsymbol{x}\|^2) \, s + 3\beta^{4/3} \int_0^s a(u) \, \mathrm{d}u \,, \qquad \forall s \in [0,t] \,.$$

Applying a version of Grönwall's inequality (e.g. [Str18, Lemma 1.2.4]), we obtain:

$$a(t) \le t \, (d + \beta^{2/3} \, \|\boldsymbol{x}\|^2) \exp(3\beta^{4/3} t) \le 3t \, (d + \beta^{2/3} \, \|\boldsymbol{x}\|^2) \,,$$

where the last line uses the hypothesis $t \le 1/(3\beta^{4/3})$. □

In addition, we will also need a concentration inequality for $\|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2$. We first present a bound on the moment generating function of the supremum of a one-dimensional Brownian motion using the reflection principle.

Lemma 12. *Let $(B_s)_{s\ge 0}$ be a standard one-dimensional Brownian motion. For $h, \lambda > 0$, such that $\lambda < \frac{1}{2h}$ the following holds:*

$$\mathbb{E} \exp\Bigl(\lambda \sup_{s\in[0,h]} |B_s|^2\Bigr) \le \frac{1 + 2h\lambda}{1 - 2h\lambda}.$$

Proof. The reflection principle [KS98, Proposition 6.19, 2.2.6] states that for every $t > 0$,

$$\mathbb{P}\bigl(\sup_{s\in[0,h]} B_s > t\bigr) = 2\,\mathbb{P}(B_h > t).$$

As a result, we have that

$$\begin{aligned}\mathbb{P}\bigl(\sup_{s\in[0,h]} |B_s|^2 > t\bigr) &= \mathbb{P}\bigl(\sup_{s\in[0,h]} |B_s| > \sqrt{t}\bigr) \\ &\le \mathbb{P}\bigl(\sup_{s\in[0,h]} B_s > \sqrt{t}\bigr) + \mathbb{P}\bigl(\inf_{s\in[0,h]} B_s < -\sqrt{t}\bigr) \\ &= 4\,\mathbb{P}(B_h > \sqrt{t}) \le 2\exp\bigl(-\frac{t}{2h}\bigr).\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}\exp\Bigl(\lambda \sup_{s\in[0,h]} |B_s|^2\Bigr) &= 1 + \lambda \int_0^\infty \exp(\lambda t)\,\mathbb{P}\bigl(\sup_{s\in[0,h]} |B_s|^2 > t\bigr)\,\mathrm{d}t \\ &\le 1 + 2\lambda \int_0^\infty \exp\bigl(-\frac{1-2h\lambda}{2h}\,t\bigr)\,\mathrm{d}t = 1 + \frac{4h\lambda}{1-2h\lambda}\,.\end{aligned}$$

□

The above argument is relevant for Lemma 9, which is restated and proved below.

Lemma 9. *Assume $h \le 1/(2\beta)$. For $0 < \lambda < 1/(24h)$,*

$$\bar{\mathbf{E}}_{\boldsymbol{x}} \exp\bigl(\lambda \sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2\bigr) \le \exp\bigl(12\beta^2\lambda h^2 \|\boldsymbol{x}\|^2 + d\ln\frac{1+24h\lambda}{1-24h\lambda}\bigr).$$

Proof. For a fixed realization of the sample path $(\bar{\boldsymbol{X}}_t)_{t\in[0,h]}$ and $0 \le t \le h$, define the function $f(t) := \sup_{s\in[0,t]} \|\bar{\boldsymbol{X}}_s - \boldsymbol{x}\|^2$. Then, for all $s \in [0,t]$,

$$\begin{aligned}\|\bar{\boldsymbol{X}}_s - \boldsymbol{x}\|^2 &= \Bigl\|-\int_0^s \nabla V(\bar{\boldsymbol{X}}_r)\,\mathrm{d}r + \sqrt{2}\,\boldsymbol{B}_s\Bigr\|^2 \le 2\Bigl\|-\int_0^s \nabla V(\bar{\boldsymbol{X}}_r)\,\mathrm{d}r\Bigr\|^2 + 4\|\boldsymbol{B}_s\|^2 \\ &\le 2h\int_0^s \|\nabla V(\bar{\boldsymbol{X}}_r)\|^2\,\mathrm{d}r + 4\|\boldsymbol{B}_s\|^2 \le 2\beta^2 h \int_0^s \|\bar{\boldsymbol{X}}_r\|^2\,\mathrm{d}r + 4\|\boldsymbol{B}_s\|^2 \\ &\le 4\beta^2 h \int_0^s \|\bar{\boldsymbol{X}}_r - \boldsymbol{x}\|^2\,\mathrm{d}r + 4\beta^2 h^2 \|\boldsymbol{x}\|^2 + 4\|\boldsymbol{B}_s\|^2 \\ &\le 4\beta^2 h \int_0^s f(r)\,\mathrm{d}r + 4\beta^2 h^2 \|\boldsymbol{x}\|^2 + 4\|\boldsymbol{B}_s\|^2\end{aligned}$$

which yields

$$f(t) = \sup_{s\in[0,t]} \|\bar{\boldsymbol{X}}_s - \boldsymbol{x}\|^2 \le 4\beta^2 h \int_0^t f(r)\,\mathrm{d}r + 4\beta^2 h^2 \|\boldsymbol{x}\|^2 + 4\sup_{s\in[0,h]} \|\boldsymbol{B}_s\|^2\,.$$

Applying Grönwall's inequality [Str18, Lemma 1.2.4], we see that

$$\begin{aligned}f(h) = \sup_{s\in[0,h]} \|\bar{\boldsymbol{X}}_s - \boldsymbol{x}\|^2 &\le \bigl(4\beta^2 h^2 \|\boldsymbol{x}\|^2 + 4\sup_{s\in[0,h]} \|\boldsymbol{B}_s\|^2\bigr)\exp(4\beta^2 h^2) \\ &\le 12\beta^2 h^2 \|\boldsymbol{x}\|^2 + 12 \sup_{s\in[0,h]} \|\boldsymbol{B}_s\|^2\,.\end{aligned}$$

Hence,

$$\mathbb{E}\exp\bigl(\lambda \sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2\bigr) \le \exp(12\beta^2\lambda h^2 \|\boldsymbol{x}\|^2)\,\mathbb{E}\exp\bigl(12\lambda \sup_{s\in[0,h]} \|\boldsymbol{B}_s\|^2\bigr)$$

$$\le \exp(12\beta^2\lambda h^2 \|\boldsymbol{x}\|^2) \big\{\mathbb{E}\exp\big(12\lambda \sup_{s\in[0,h]} |B_s|^2\big)\big\}^d$$
$$\le \exp(12\beta^2\lambda h^2 \|\boldsymbol{x}\|^2) \big(\frac{1+24h\lambda}{1-24h\lambda}\big)^d,$$

by Lemma 12 and the assumption $\lambda < 1/(24h)$. □

Corollary 6. *Assume $h \le 1/(2\beta)$. There exists a numerical constant $C > 0$ such that for all $k \ge 1$,*

$$\mathbb{E}\sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^{2k} \le C^k\,(\beta^k h^{2k}\,\|\boldsymbol{x}\|^{2k} + d^k h^k + h^k k^k).$$

Proof. In Lemma 9, take $\lambda := 1/(48h)$ to yield

$$\mathbb{E}\exp\big(\lambda \sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2\big) \le \exp\big(\frac{1}{4}\beta^2 h\,\|\boldsymbol{x}\|^2 + d\ln 3\big)\,.$$

It follows from Markov's inequality that for all $x \ge 0$,

$$\mathbb{P}\big(\sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2 \ge 12h^2\beta\,\|\boldsymbol{x}\|^2 + (48\ln 3)hd + x\big) \le \exp\big(-\frac{x}{48h}\big).$$

The result now follows from standard moment bounds under sub-exponential concentration [see, e.g., Ver18, Proposition 2.7.1]. □

Remark 5. Bounds such as the one in Corollary 6 are standard and have appeared in the literature before, e.g., [Mou+19, Lemma 11].

From total variation to other distances

In this section, we deduce the mixing time results of Theorem 9 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 2) is preserved by the iterations of MALA. In fact, this is true for all reversible Markov chains.

Lemma 13. *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 $\boldsymbol{y} \in \mathbb{R}^d$,

$$\frac{\mu_{n+1}(\boldsymbol{y})}{\pi(\boldsymbol{y})} = \int \frac{\mu_n(\boldsymbol{x})}{\pi(\boldsymbol{y})}\,T(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x} = \int \frac{\mu_n(\boldsymbol{x})}{\pi(\boldsymbol{x})}\,\frac{\pi(\boldsymbol{x})T(\boldsymbol{x},\boldsymbol{y})}{\pi(\boldsymbol{y})}\,\mathrm{d}\boldsymbol{x} \le M_0 \int T(\boldsymbol{y},\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = M_0\,,$$

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 14. *Let μ be M_0-warm with respect to π. Then,*

$$\chi^2(\mu \,\|\, \pi) \le 2M_0 \,\|\mu - \pi\|_{\mathrm{TV}}\,.$$

Proof. From the definition of the chi-squared divergence,

$$\chi^2(\mu \,\|\, \pi) = \int \bigl|\frac{\mu}{\pi} - 1\bigr|^2 \,\mathrm{d}\pi \le M_0 \int \bigl|\frac{\mu}{\pi} - 1\bigr| \,\mathrm{d}\pi = 2M_0 \,\|\mu - \pi\|_{\mathsf{TV}}\,.$$

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 7. *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:*

$$\tau_{\mathrm{mix}}(\varepsilon, \mu_0; \mathsf{d}) \le \tau_{\mathrm{mix}}\bigl(\frac{\varepsilon^2}{2M_0}, \mu_0; \mathsf{TV}\bigr)$$

for each of the distances

$$\mathsf{d} \in \Bigl\{\sqrt{\mathsf{KL}},\ \sqrt{\chi^2},\ \sqrt{\frac{\alpha}{2}}\,W_2\Bigr\}\,.$$

Proof. The mixing time in the chi-squared distance is a straightforward consequence of Lemmas 13 and 14. 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 transportation 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 7 implies the remaining mixing time results in Theorem 9.

5.7 Proof of the lower bound

This section presents the proofs of Theorems 11 and 12. The majority of this section is devoted to the proof of the upper bound on the conductance when $h \gg d^{-1/2}$ (Theorem 11). The proof of the upper bound on the spectral gap (Theorem 12) is given in Section 5.7.

High-level overview of the proof

Recall that we take $\eta = 1/4 - \delta$, where $\delta > 0$ is fixed throughout. As mentioned in Section 5.5, we consider the potential

$$\begin{aligned} V(\boldsymbol{x}) &= \frac{\|\boldsymbol{x}\|^2}{2} - \frac{1}{2d^{2\eta}} \sum_{i=1}^{d} \cos(d^\eta x_i) && (5.16) \\ &=: V_{\mathsf{G}}(\boldsymbol{x}) + V_{\mathsf{P}}(\boldsymbol{x}). && (5.17) \end{aligned}$$

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_{P} 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 20) shows that $\mathsf{KL}(\mathcal{N}(0,1)\|\pi) = O(d^{1-4\eta})$. 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_{P} is an oscillatory perturbation. Hence, MALA would not see the contribution from V_{P} as long as its movement due to the Langevin proposal is at least as long as the length scale of the fluctuations of V_{P}.

   With this in mind, note that the fluctuations of V_{P} 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_{P} 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 11, we use the following proposition.

Proposition 4. *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),

$$\mathsf{C} = \inf_{\substack{S\subseteq\mathbb{R}^d \\ \pi(S)\le 1/2}} \frac{\int_S T(\boldsymbol{x}, S^{\mathsf{c}})\,\pi(\mathrm{d}\boldsymbol{x})}{\pi(S)} \le 2\int_{E_0} T(\boldsymbol{x}, E_0^{\mathsf{c}})\,\pi(\mathrm{d}\boldsymbol{x})$$

$$\leq 2 \int_{E_0} \Big(\int_{E_0^\mathsf{c}} Q(\boldsymbol{x},\boldsymbol{y}) A(\boldsymbol{x},\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} \Big)\, \pi(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \leq 2 \int_{E_0} \Big(\int_{\mathbb{R}^d} Q(\boldsymbol{x},\boldsymbol{y}) A(\boldsymbol{x},\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} \Big)\, \pi(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}$$

$$\leq 2 \sup_{\boldsymbol{x}\in E_0} \int_{\mathbb{R}^d} Q(\boldsymbol{x},\boldsymbol{y}) A(\boldsymbol{x},\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} \leq 2 \sup_{\boldsymbol{x}\in E} \int_{\mathbb{R}^d} Q(\boldsymbol{x},\boldsymbol{y}) A(\boldsymbol{x},\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y}.$$

□

From Proposition 4, 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

$$\begin{aligned} Q(\boldsymbol{x},\boldsymbol{y}) A(\boldsymbol{x},\boldsymbol{y}) &= Q(\boldsymbol{x},\boldsymbol{y}) \min\Big\{1, \frac{\pi(\boldsymbol{y}) Q(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x}) Q(\boldsymbol{x},\boldsymbol{y})}\Big\} \\ &\leq \frac{\pi(\boldsymbol{y}) Q(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})} \\ &= \frac{1}{(4\pi h)^{d/2}} \exp\Big[V(\boldsymbol{x}) - V(\boldsymbol{y}) - \frac{\|\boldsymbol{y} - \boldsymbol{x} - h\nabla V(\boldsymbol{y})\|^2}{4h}\Big]\,. \end{aligned} \tag{5.18}$$

We substitute in the definition of our potential (5.16) and expand out the terms in (5.18), grouping them according to whether they involve V_{P} or not:

$$(5.18) = \frac{1}{(4\pi h)^{d/2}} \exp\Big[\frac{1}{2}\|\boldsymbol{x}\|^2 - \frac{1}{2}\|\boldsymbol{y}\|^2 - \frac{1}{4h}\|(1-h)\boldsymbol{y} - \boldsymbol{x}\|^2\Big] \tag{5.19}$$

$$\times \exp\Big[V_{\mathsf{P}}(\boldsymbol{x}) - V_{\mathsf{P}}(\boldsymbol{y}) + \frac{1}{2}\,\langle (1-h)\boldsymbol{y} - \boldsymbol{x}, \nabla V_{\mathsf{P}}(\boldsymbol{y})\rangle - \frac{h}{4}\,\|\nabla V_{\mathsf{P}}(\boldsymbol{y})\|^2\Big]\,. \tag{5.20}$$

Some algebra yields that (5.19) 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 $\mathcal{N}(\frac{1-h}{1+h^2}\,\boldsymbol{x}, \frac{2h}{1+h^2} I_d)$ 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

$$\underbrace{\frac{\exp\Bigl[\frac{h^2\,\|\boldsymbol{x}\|^2}{2\,(1+h^2)}+V_{\mathsf{P}}(\boldsymbol{x})\Bigr]}{(1+h^2)^{d/2}}}_{①}\times\underbrace{\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mu_{\boldsymbol{x}}}\exp\Bigl[-V_{\mathsf{P}}(\boldsymbol{y})+\frac{1}{2}\,\langle(1-h)\boldsymbol{y}-\boldsymbol{x},\nabla V_{\mathsf{P}}(\boldsymbol{y})\rangle-\frac{h}{4}\,\|\nabla V_{\mathsf{P}}(\boldsymbol{y})\|^2\Bigr]}_{②}\,.$$

Having this upper bound, we will prove that there is a set $E\subseteq\mathbb{R}^d$ with $\pi(E)\ge 1/2$ such that the following bounds hold for all $\boldsymbol{x}\in E$:

1. (Lemma 18)
$$① \le \exp\bigl[-\frac{1}{8}d^{1-4\eta}+o(d^{1-4\eta})\bigr]\,.$$

2. (Lemma 19)
$$② \le \exp\bigl[\frac{1}{16}d^{1-4\eta}+o(d^{1-4\eta})\bigr]\,.$$

From these bounds and the preceding calculations, we have

$$\sup_{\boldsymbol{x}\in E}\int Q(\boldsymbol{x},\boldsymbol{y})A(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\le\exp\bigl[-\frac{1}{8}d^{1-4\eta}+o(d^{1-4\eta})\bigr]\,.$$

This completes the proof of Theorem 11.

The next section is devoted to proving the two main bounds (Lemmas 18 and 19).

Proofs of technical statements

Notation and technical lemmas

We use the following notation:

$$\begin{cases} V_1(x) := \frac{1}{2}x^2-\frac{1}{2}d^{-2\eta}\cos(d^\eta x)\,,\\ V(\boldsymbol{x}) := \sum_{i=1}^d V_1(x_i)=\frac{1}{2}\|\boldsymbol{x}\|^2-\frac{1}{2}d^{-2\eta}\sum_{i=1}^d\cos(d^\eta x_i)\,,\\ \pi_1(x)\propto\exp(-V_1(x))\,,\\ \pi(\boldsymbol{x})\propto\exp(-V(\boldsymbol{x}))\,. \end{cases} \tag{5.21}$$

Thus, π_1 is the marginal distribution of π. We first list useful technical lemmas for proving Lemmas 18 and 19. First, the following trigonometric inequality will be used several times.

Lemma 15. *Let $\xi\sim\mathcal{N}(0,1)$, let p be a polynomial, and let $a,b\in\mathbb{R}$, $\gamma>0$ be constants. Then, there exists a constant C (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 $\mathbb{E}[e^{it\xi}]$ of a Gaussian is equal to $e^{-\frac{1}{2}t^2}$. First consider the case $p \equiv 1$. Let $\mathsf{im}(\cdot)$ denote the imaginary part. Then, we have

$$\begin{aligned}\mathbb{E}[\sin(a+bd^\gamma\xi)] &= \mathbb{E}[\mathsf{im}(e^{i\,(a+bd^\gamma\xi)})] \\ &= \mathsf{im}(e^{ia}\,\mathbb{E}[e^{ibd^\gamma\xi}]) \\ &= \mathsf{im}\Big(\exp\big(ia - \frac{b^2d^{2\gamma}}{2}\big)\Big) \\ &= \sin(a)\exp\big(-\frac{b^2d^{2\gamma}}{2}\big)\,.\end{aligned}$$

It is then clear that the result holds for $p=1$. Next, when $p(x)=x^\ell$ for some $\ell\in\mathbb{N}^+$,

$$\begin{aligned}\mathbb{E}[\xi^\ell\sin(a+bd^\gamma\xi)] &= \mathsf{im}(e^{ia}\,\mathbb{E}[\xi^\ell e^{ibd^\gamma\xi}]) \\ &= \mathsf{im}\Big(e^{ia}\,i^{-\ell}\,\mathbb{E}\Big[\frac{\mathrm{d}^\ell}{\mathrm{d}t^\ell}e^{it\xi}\Big|_{t=bd^\gamma}\Big]\Big) \\ &= \mathsf{im}\Big(e^{ia}\,i^{-\ell}\,\frac{\mathrm{d}^\ell}{\mathrm{d}t^\ell}e^{-\frac{t^2}{2}}\Big|_{t=bd^\gamma}\Big).\end{aligned}$$

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 16. *Assume $\eta < 1/4$. The following hold for π_1 and π defined in (5.21):*

(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 $\xi\sim\mathcal{N}(0,1)$, then

$$Z - \sqrt{2\pi} = \int_{\mathbb{R}}\exp\Big(-\frac{1}{2}\,x^2 + \frac{1}{2d^{2\eta}}\cos(d^\eta x)\Big)\,\mathrm{d}x - \sqrt{2\pi}$$

$$
\begin{aligned}
&= \sqrt{2\pi} \int_{\mathbb{R}} \exp\Big(\frac{1}{2d^{2\eta}} \cos(d^\eta x)\Big) \, \frac{\exp(-\frac{1}{2}x^2)}{\sqrt{2\pi}} \, \mathrm{d}x - \sqrt{2\pi} \\
&= \sqrt{2\pi} \, \Big(\mathbb{E} \exp\big(\frac{1}{2d^{2\eta}} \, \cos(d^\eta \xi)\big) - 1\Big) \\
&= \frac{\sqrt{2\pi}}{2d^{2\eta}} \, \mathbb{E} \cos(d^\eta \xi) + O(d^{-4\eta}).
\end{aligned}
$$

By Lemma 15, 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 $\xi \sim \mathcal{N}(0, 1)$,

$$
\begin{aligned}
\mathbb{E}_{x\sim\pi_1}[x^2] &= \int x^2 \, \frac{\exp(-V_1(\boldsymbol{x}))}{Z} \, \mathrm{d}x \\
&= \frac{\sqrt{2\pi}}{Z} \, \mathbb{E}\big[\xi^2 \exp\big(\frac{1}{2d^{2\eta}} \cos(d^\eta \xi)\big)\big] \\
&= \big(1 + O(d^{-4\eta})\big) \, \mathbb{E}\big[\xi^2 \exp\big(\frac{1}{2d^{2\eta}} \cos(d^\eta \xi)\big)\big].
\end{aligned}
$$

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 15, 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 15, the first term is $\mathbb{E} \cos(d^\eta \xi) = o(d^{-4\eta})$. Next, the second term is

$$
\frac{1}{2d^{2\eta}} \, \mathbb{E} \cos^2(d^\eta \xi) = \frac{1}{4d^{2\eta}} + \frac{1}{4d^{2\eta}} \, \mathbb{E} \cos(2d^\eta \xi).
$$

From Lemma 15, $\mathbb{E} \cos(2d^\eta \xi) = o(d^{-4\eta})$. Therefore, the result follows.

□

Lemma 17. *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 $\xi \sim \mathcal{N}(0,2)$. 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

$$\mathbb{P}\bigl(\max_{i\in[d]} \xi_i > \mathbb{E}\max_{i\in[d]} \xi_i + t\bigr) \leq \exp\bigl(-\frac{t^2}{4}\bigr).$$

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.

Proof of Lemma 18

Lemma 18. *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 $\boldsymbol{x} \in E_1$,*

$$\frac{\exp\Bigl[\frac{h^2\,\|\boldsymbol{x}\|^2}{2\,(1+h^2)} + V_{\mathsf{P}}(\boldsymbol{x})\Bigr]}{(1+h^2)^{d/2}} \leq \exp\bigl[-\frac{1}{8}d^{1-4\eta} + o(d^{1-4\eta})\bigr]\,.$$

Proof. We decompose the left-hand side as

$$\frac{\exp\Bigl[\frac{h^2\,\|\boldsymbol{x}\|^2}{2\,(1+h^2)} + V_{\mathsf{P}}(\boldsymbol{x})\Bigr]}{(1+h^2)^{d/2}} = \frac{1}{(1+h^2)^{d/2}}\exp\Bigl[\frac{h^2\,\|\boldsymbol{x}\|^2}{2\,(1+h^2)}\Bigr] \times \exp[V_{\mathsf{P}}(\boldsymbol{x})]$$

and bound each term separately.

We begin with the first term. By Lemma 16-(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 10) 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 $x \geq 0$,

$$\begin{aligned}
\frac{1}{(1+h^2)^{d/2}} \exp\Bigl[\frac{h^2\,\|\boldsymbol{x}\|^2}{2\,(1+h^2)}\Bigr] &\leq \exp\Bigl[\frac{h^2\,(d+O(d^{1-4\eta})+O(d^{1/2}))}{2\,(1+h^2)} - \frac{d}{2}\ln(1+h^2)\Bigr] \\
&\leq \exp\Bigl[\frac{h^2\,(d+O(d^{1-4\eta})+O(d^{1/2}))}{2\,(1+h^2)} - \frac{dh^2}{2} + \frac{dh^4}{4}\Bigr] \\
&= \exp\Bigl[\frac{h^2\,(O(d^{1-4\eta})+O(d^{1/2}))}{2\,(1+h^2)} - \frac{dh^4}{2\,(1+h^2)} + \frac{dh^4}{4}\Bigr] \\
&= \exp\Bigl[\frac{h^2\,(O(d^{1-4\eta})+O(d^{1/2}))}{2\,(1+h^2)} + \frac{-dh^4+2dh^6}{4\,(1+h^2)}\Bigr] \\
&\leq \exp[O(d^{1-4\eta}h^2) + O(d^{1/2}h^2)]\,.
\end{aligned}$$

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 move on to the second term. Recall from the calculation in Lemma 16-(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

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\pi}[V_{\mathsf{P}}(\boldsymbol{x})] = -\frac{1}{2d^{2\eta}}\sum_{i=1}^{d}\mathbb{E}_{x_i\sim\pi_1}\cos(d^\eta x_i) = -\frac{1}{8}d^{1-4\eta} + O(d^{1-8\eta}).$$

Since π is 1/2-strongly log-concave, another sub-Gaussian concentration argument (Lemma 10) shows that there exists a subset E_1'' with $\pi(E_1'') \geq 7/8$ such that for $\boldsymbol{x} \in E_1''$,

$$\exp[V_{\mathsf{P}}(\boldsymbol{x})] \leq \exp\bigl[-\frac{1}{8}d^{1-4\eta} + O(d^{1-8\eta}) + O(d^{1/2-2\eta})\bigr] \leq \exp\bigl[-\frac{1}{8}d^{1-4\eta} + o(d^{1-4\eta})\bigr]\,,$$

since $1-4\eta > 0$ by the hypothesis.

Now taking $E_1 := E_1' \cap E_1''$, the above calculations show that for $\boldsymbol{x} \in E_1$,

$$\frac{\exp\Bigl[\frac{h^2\,\|\boldsymbol{x}\|^2}{2\,(1+h^2)} + V_{\mathsf{P}}(\boldsymbol{x})\Bigr]}{(1+h^2)^{d/2}} \leq \exp\bigl[-\frac{1}{8}d^{1-4\eta} + o(d^{1-4\eta})\bigr]\,,$$

which completes the proof. □

Proof of Lemma 19

Lemma 19. *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 $\boldsymbol{x} \in E_2$,*

$$\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mu_{\boldsymbol{x}}} \exp\Bigl[-V_{\mathsf{P}}(\boldsymbol{y}) + \frac{1}{2}\langle(1-h)\boldsymbol{y}-\boldsymbol{x}, \nabla V_{\mathsf{P}}(\boldsymbol{y})\rangle - \frac{h}{4}\|\nabla V_{\mathsf{P}}(\boldsymbol{y})\|^2\Bigr] \\ \leq \exp\bigl[\frac{1}{16}d^{1-4\eta} + o(d^{1-4\eta})\bigr]\,.$$

Proof. Recall the definition $V_{\mathsf{P}}(\boldsymbol{x}) = -\frac{1}{2}d^{-2\eta}\sum_{i=1}^d \cos(d^\eta x_i)$. Since V_{P} is separable, it suffices to consider the following quantity: for $\mu_{x_i} := \mathcal{N}(\frac{1-h}{1+h^2}\,x_i, \frac{2h}{1+h^2})$,

$$\max_{i\in[d]} \mathop{\mathbb{E}}_{y_i\sim\mu_{x_i}} \exp\Bigl(\frac{\cos(d^\eta y_i)}{2d^{2\eta}} + \frac{((1-h)y_i - x_i)\sin(d^\eta y_i)}{4d^\eta} - \frac{h\sin^2(d^\eta y_i)}{16d^{2\eta}}\Bigr)\,. \tag{5.22}$$

Indeed, the lemma is proved as soon as we show

$$(5.22) \leq \exp\bigl[\frac{1}{16}d^{-4\eta} + o(d^{-4\eta})\bigr]\,. \tag{5.23}$$

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 (5.23), let us first simplify the expression (5.22). Letting $\xi \sim \mathcal{N}(0,1)$, 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

$$\bar{h} := \frac{h}{1+h^2} \qquad \text{and} \qquad \tilde{h} := \frac{(1-h)^2}{1+h^2}\,h\,,$$

and treat them as being on the same order as h. Using these simplifying notations and rearranging, we are left to consider

$$\mathbb{E}\exp\Bigl(\underbrace{\frac{\cos(d^\eta y_1)}{2d^{2\eta}}}_{=:\Delta_1} - \underbrace{\frac{h\sin^2(d^\eta y_1)}{16d^{2\eta}}}_{=:\Delta_2} - \underbrace{\frac{2\bar{h}x_1\sin(d^\eta y_1)}{4d^\eta}}_{=:\Delta_3} + \underbrace{\frac{\sqrt{2\tilde{h}}\xi\sin(d^\eta y_1)}{4d^\eta}}_{=:\Delta_4}\Bigr)\,, \tag{5.24}$$

where $y_1 = \frac{1-h}{1+h^2}\,x_1 + \sqrt{\frac{2h}{1+h^2}}\,\xi$. Now we will estimate (5.24) by a Taylor expansion.

Throughout, we will assume $\|\boldsymbol{x}\|_\infty \leq 4\sqrt{\ln(8d)}$. By Lemma 17, this holds on an event E_2 of probability $\pi(E_2) \geq 3/4$. From this, we note the immediate bounds

$$|\Delta_1| = O(d^{-2\eta}), \qquad |\Delta_2| = O(d^{-2\eta}h), \qquad |\Delta_3| = \widetilde{O}(d^{-\eta}h), \qquad |\Delta_4| = O_{\mathsf{p}}(d^{-\eta}\sqrt{h}).$$

Here, O_{p} denotes probabilistic big-O notation. Using $h = O(d^{-1/3}) = o(d^{-4\eta/3})$, we have

$$|\Delta_1| = O(d^{-2\eta}), \quad |\Delta_2| = o(d^{-(3+1/3)\eta}), \quad |\Delta_3| = o(d^{-(2+1/3)\eta}), \quad |\Delta_4| = o_{\mathsf{p}}(d^{-(1+2/3)\eta}). \tag{5.25}$$

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 $\mathbb{E}[|\Delta_4|^3] = o(d^{-5\eta})$.

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 $h \geq d^{-\frac{1}{2}+3\delta}$. Thus, the trigonometric terms precisely fit into the setting of Lemma 15, and we will apply Lemma 15 to estimate these terms.

Now let us estimate the terms of order one and two.

- *First- and lower-order terms.* We have

  $$(\leq 1\text{st order}) = 1 + \mathbb{E}\,\Delta_1 - \mathbb{E}\,\Delta_2 - \mathbb{E}\,\Delta_3 + \mathbb{E}\,\Delta_4\,.$$

  By Lemma 15, 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 15 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 15 and 17. 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 15. Collecting together the terms, we have

  $$(\leq 1\text{st order}) = 1 - \frac{h}{32d^{2\eta}} + o(d^{-5\eta}). \tag{5.26}$$

- *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.

	$O(d^{-2\eta})$	$o(d^{-(3+1/3)\eta})$	$o(d^{-(2+1/3)\eta})$	$o_{\mathsf{p}}(d^{-(1+2/3)\eta})$
$O(d^{-2\eta})$	(5.27)	$o(d^{-4\eta})$	$o(d^{-4\eta})$	(5.28)
$o(d^{-(3+1/3)\eta})$		$o(d^{-4\eta})$	$o(d^{-4\eta})$	$o_{\mathsf{p}}(d^{-4\eta})$
$o(d^{-(2+1/3)\eta})$			$o(d^{-4\eta})$	$o_{\mathsf{p}}(d^{-4\eta})$
$o_{\mathsf{p}}(d^{-(1+2/3)\eta})$				(5.29)

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 (5.25).

We now estimate the terms which are not covered by the table. Let us estimate the remaining terms one by one. First, by Lemma 15,

$$\frac{1}{2}\,\mathbb{E}[\Delta_1^2] = \mathbb{E}\,\frac{\cos^2(d^\eta y_1)}{8d^{4\eta}} = \frac{1}{16d^{4\eta}} + \mathbb{E}\,\frac{\cos(2d^\eta y_1)}{16d^{4\eta}} = \frac{1}{16d^{4\eta}} + o(d^{-8\eta})\,. \tag{5.27}$$

Next, by Lemma 15,

$$\mathbb{E}[\Delta_1\Delta_4] = \mathbb{E}\Big[\frac{\sqrt{2\tilde{h}}\xi}{8d^{3\eta}}\cos(d^\eta y_1)\sin(d^\eta y_1)\Big] = \frac{\sqrt{2\tilde{h}}}{16d^{3\eta}}\,\mathbb{E}[\xi\sin(2d^\eta y_1)] = o(d^{-7\eta}). \tag{5.28}$$

Lastly, invoking Lemma 15 yet again,

$$\frac{1}{2}\,\mathbb{E}[\Delta_4^2] = \mathbb{E}\,\frac{\tilde{h}\xi^2\sin^2(d^\eta y_1)}{16d^{2\eta}} = \mathbb{E}\,\frac{\tilde{h}\xi^2}{32d^{2\eta}} - \mathbb{E}\,\frac{\tilde{h}\xi^2\cos(2d^\eta y_1)}{32d^{2\eta}} = \frac{\tilde{h}}{32d^{2\eta}} + o(d^{-6\eta}). \tag{5.29}$$

Combining all together, we obtain,

$$(\text{2nd order}) = \frac{1}{16d^{4\eta}} + \frac{\tilde{h}}{32d^{2\eta}} + o(d^{-4\eta})\,. \tag{5.30}$$

Therefore, we combine (5.26) and (5.30) to conclude

$$\begin{aligned}(5.24) &\le \exp\Big[\frac{1}{16}d^{-4\eta} - \frac{h}{32d^{2\eta}} + \frac{\tilde{h}}{32d^{2\eta}} + o(d^{-4\eta})\Big]\\ &= \exp\Big[\frac{1}{16}d^{-4\eta} + o(d^{-4\eta})\Big]\,,\end{aligned}$$

where the last line follows from the fact $\tilde{h} - h = \frac{(1-h)^2}{1+h^2}\,h - h \le 0$. This implies

(5.23), and hence the proof is complete. □

Upper bound on the spectral gap

Note that when $\eta < 1/4$, the adversarial potential defined in (5.21) satisfies the assumptions of the following theorem, as a consequence of our computation in Lemma 16.

Theorem 13. *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 $f(\boldsymbol{x}) := x_1$. 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. □

Auxiliary lemmas

Lemma 20. *Let $\gamma := \mathcal{N}(0, I_d)$ and let π be the adversarial target distribution defined in (5.21). 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

$$\mathsf{KL}(\gamma \,\|\, \pi) = \int \gamma(\boldsymbol{x}) \ln\bigl(\frac{Z^d}{(2\pi)^{d/2}} \exp V_{\mathsf{P}}(\boldsymbol{x})\bigr) \, \mathrm{d}\boldsymbol{x} = d \ln \frac{Z}{\sqrt{2\pi}} - \frac{1}{2d^{2\eta}} \sum_{i=1}^{d} \mathbb{E} \cos(d^\eta \xi_i).$$

From our estimate of the normalizing constant in Lemma 16,

$$d \ln \frac{Z}{\sqrt{2\pi}} = d \ln\bigl(1 + O(d^{-4\eta})\bigr) = O(d^{1-4\eta}).$$

On the other hand, from the proof of Lemma 15,

$$-\frac{1}{2d^{2\eta}} \sum_{i=1}^{d} \mathbb{E} \cos(d^\eta \xi_i) = o(d^{1-4\eta}).$$

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.

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. *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*

$$\tau_{\mathrm{mix}}(\varepsilon, \mu_0; \mathrm{TV}) \lesssim d^{1/3} \log\Bigl(\frac{M_0}{\varepsilon}\Bigr).$$

Using the results of Section 5.6, 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 21. *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 $A(\boldsymbol{x}) \ge 5/6$.*

Proof of Proposition 5. We sketch the proof, following the s-conductance mixing time strategy outlined in Section 5.6. Let $E := \{\boldsymbol{x} \in \mathbb{R}^d \mid A(\boldsymbol{x}) \ge 5/6\}$. Lemma 21 guarantees that $\pi(E) \ge 1 - 2\exp(-c_1 d^{1/3})$. By our assumption, we have $\log(\varepsilon h/M_0) = \Omega(d^{-1/3})$, so $\pi(E) \ge 1 - c'\sqrt{hs}$ for some constant $c' > 0$, where $s := \varepsilon/(2M_0)$. Moreover, on the event E we have (by Proposition 1) that

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{TV} = 1 - A(\boldsymbol{x}) \le \frac{1}{6}\,.$$

Then the argument in the proof of Proposition 3 implies that the s-conductance, defined in (5.9), is lower bounded by $\mathsf{C}_s \gtrsim \sqrt{h}$, and Corollary 5 gives the desired mixing time bound. □

Proof of Lemma 21. 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 $t > 0$,

$$\mathbb{P}\big(\big|\,\|\boldsymbol{x}\| - \sqrt{d}\,\big| > t\big) \le 2\exp(-ct^2)\,.$$

As a result, the event

$$E_1 := \big\{\big|\,\|\boldsymbol{x}\| - \sqrt{d}\,\big| \le t_1\big\}$$

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

$$E_2 := \{|\xi_1| \le t_2\}$$

holds with probability at least $1 - 2\exp(-ct_2^2)$, and the event

$$E_3 := \big\{\big|\,\|\boldsymbol{\xi}_{-1}\| - \sqrt{d}\,\big| \le t_3\big\}$$

hold with probability at least $1 - 2\exp(-ct_3^2)$. Define the quantities

$$\varepsilon_1 := \|\boldsymbol{x}\| - \sqrt{d}\,, \qquad \varepsilon_2 := \xi_1\,, \qquad \varepsilon_3 := \|\boldsymbol{\xi}_{-1}\| - \sqrt{d}\,.$$

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

$$\begin{aligned}
\frac{h}{4}\,\bigl|\,\|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2\,\bigr| &= \frac{h}{4}\,|x_1^2 - [(1-h)x_1 + \sqrt{2h}\,\xi_1]^2 - 2h\,\|\boldsymbol{\xi}_{-1}\|^2| \\
&= \frac{h}{4}\,|(\sqrt{d}+\varepsilon_1)^2 - [(1-h)\,(\sqrt{d}+\varepsilon_1) + \sqrt{2h}\,\varepsilon_2]^2 - 2h\,(\sqrt{d}+\varepsilon_3)^2| \\
&= O(dh^3 + d^{1/2}h^2 t_1 + h^{3/2}d^{1/2}t_2 + d^{1/2}h^2 t_3)\,,
\end{aligned}$$

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,

$$\frac{h}{4}\,\bigl|\,\|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2\,\bigr| = O(dh^3 + d^{2/3}h^2 + d^{1/2}h^{3/2})\,.$$

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. □

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 6. *For every $\theta < 1/3$, if we take step size $h = d^{-\theta}$, then the conductance of the MALA chain is exponentially small:*

$$\exists \delta > 0 \qquad \text{such that} \qquad \mathsf{C} \lesssim \exp[-\Omega(d^\delta)]\,.$$

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 $\pi(E) \ge 1/2$ 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 4. Specifically, we will take $E := \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\| \le \sqrt{d}\}$; 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 $A(\boldsymbol{x},\boldsymbol{y}) = a(\boldsymbol{x},\boldsymbol{y}) \wedge 1 \le \sqrt{a(\boldsymbol{x},\boldsymbol{y})}$.[†] Since $V(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$, a little algebra using the definition (5.1) shows that

$$a(\boldsymbol{x},\boldsymbol{y}) = \exp\bigl(\frac{h}{4}\,(\|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2)\bigr)\,.$$

Further calculations show that

$$\begin{aligned}
\int_{\mathbb{R}^d} Q(\boldsymbol{x},\boldsymbol{y}) A(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} &\le \int_{\mathbb{R}^d} Q(\boldsymbol{x},\boldsymbol{y}) a(\boldsymbol{x},\boldsymbol{y})^{1/2}\,\mathrm{d}\boldsymbol{y} \\
&= \int_{\mathbb{R}^d} \frac{1}{(4\pi h)^{d/2}} \exp\bigl(-\frac{1}{4h}\,\|\boldsymbol{y} - (1-h)\boldsymbol{x}\|^2\bigr) \exp\bigl(\frac{h}{2}\,(\|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2)\bigr)\,\mathrm{d}\boldsymbol{y} \\
&= \frac{1}{(4\pi h)^{d/2}} \int_{\mathbb{R}^d} \exp\Bigl(-\frac{1+h^2/2}{4h}\,\Bigl\|\boldsymbol{y} - \frac{1-h}{1+h^2/2}\,\boldsymbol{x}\Bigr\|^2\Bigr)\,\mathrm{d}\boldsymbol{y} \\
&\qquad\qquad \times \exp\bigl(\frac{h^2\,(1-h/4)}{1+h^2/2}\,\|\boldsymbol{x}\|^2\bigr) \\
&= \exp\Bigl(\frac{h^2\,(1-h/4)}{4\,(1+h^2/2)}\,\|\boldsymbol{x}\|^2 - \frac{d}{2}\ln\bigl(1+\frac{h^2}{2}\bigr)\Bigr).
\end{aligned}$$

For $\boldsymbol{x} \in E$, we can bound this via

$$\int_{\mathbb{R}^d} Q(\boldsymbol{x},\boldsymbol{y}) A(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \le \exp\Bigl(\frac{h^2\,(1-h/4)d}{4\,(1+h^2/2)} - \frac{d}{2}\ln\bigl(1+\frac{h^2}{2}\bigr)\Bigr) = \exp\Bigl(-\frac{h^3 d}{16}\,\bigl(1+O(h)\bigr)\Bigr)$$

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 $O(d^{1/3})$.

Proposition 7. *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 13. □

[†] One can check that the simple bound $A(\boldsymbol{x},\boldsymbol{y}) \le a(\boldsymbol{x},\boldsymbol{y})$ is not enough for the proof to go through. A similar argument to upper bound the acceptance probability is made in [HSV14].

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.

Since the work in this chapter appeared, there have been newer works that improve our understanding of MALA. In [LST21a], Lee, Shen and Tian show that there exist initializations with the warm start parameter $M_0 \sim \exp(d)$, such that the mixing time of MALA is lower bounded by $\Omega(d)$, which shows that the warm start assumption in our upper bound in Theorem 9 is actually neccessary. In [WSC22], Wu, Schmidler and Chen refined our analysis, and showed that the mixing time of MALA under a warm start is $\widetilde{\Theta}(\kappa d^{1/2})$. In [AC23], Altschuler and Chewi show that one can obtain a warm start from a feasible start in $\widetilde{O}(d^{1/2})$ using the discretized underdamped Langevin diffusion, which combining with our results on MALA gives a high accuracy sampling algorithm for log-concave distributions.

An interesting future direction is to extend our analysis to other Metropolis-Hastings Markov chains. An feature of Theorem 9 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 10). As an 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 Section 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+18c] 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.

Part II

Lower bounds

Chapter 6

Overview of lower bound ideas

Our results in Chapter 5, together with the subsequent papers mentioned in Section 5.9, completely characterizes the behavior of MALA on log-smooth and strongly log-concave distributions. But it has long been empirically observed that there are algorithms, such as the Hamiltonian Monte Carlo, that are expected to outperform MALA. So if we fix a class of distributions, can we determine which sampling algorithm is optimal over that class? To answer such a question, we would need to prove a general sampling lower bound, and that is the focus of the second part of this thesis.

Examples of what we might hope for are lower bound results in convex optimization, where there are lower bounds for entire classes of convex functions. There are two regimes in convex optimization. The first is the low dimensional regime, where the ellipsoid algorithm has linear dependence on the dimension, but has logarithmic dependence on the error ε. The second is the high dimensional regime, where algorithms such as gradient descent and accelerated gradient descent have dimension independent rates. There are corresponding lower bounds for the two regimes. In the low dimensional case, Nemirovski [Nem94] showed general convex optimization problems in a bounded domain in $\mathbb{R}^d$ requires $\Omega(d \log \frac{1}{\varepsilon})$ iterations, which is achieved by the ellipsoid method. In the high dimensional case, Nesterov [Nes18] proved that convex optimization requires at least $\Omega(\sqrt{\frac{L}{\varepsilon}})$ iterations for smooth functions, and at least $\Omega(\sqrt{\kappa} \log \frac{1}{\varepsilon})$ iterations for smooth and strongly convex functions, and these rates are achieved by the accelerated gradient descent algorithm.

In contrast, for sampling there has essentially been no work on general lower bounds. It is natural to wonder what such lower bounds might look like for sampling, and whether sampling also exhibit two separate regimes depending on dimension. Such questions are the focus of the remainder of this thesis.

Two choices have to be made before we prove any lower bound: first we need to choose the computational model of the algorithms; second we need to choose the class of distributions. For the first choice, we decided to focus on query complexity, meaning that an algorithm will be allowed to make queries to the target distribution, then it is allowed to do whatever computations it chooses based on the queries, and we only lower bound the total number of queries needed. We thought about more fine-grained computational models as well. For example, some of Nesterov's lower bounds assume that the algorithm will only take linear combinations of its previous queries. It turns out that many sampling algorithms (Unadjusted Langevin dynamics and Hamiltonian Monte Carlo) only output points that are linear combinations of past queries and standard Gaussian vectors, so such computational models also make sense for sampling. But more restricted computational model turned out to introduce further complexity, so query complexity was easier to work with.

For the query model, we allow algorithms to make point-wise queries to the potential $V(x)$ up to an additive constant (which is equivalent to queries to the unnormalized density values of $\pi(x)$), as well as one or more derivatives of the potential, $\nabla V(x), \nabla^2 V(x), \ldots$. This is the most common query framework in continuous space sampling, where it is often hard to know the normalizing constant of the target distribution.

For the choice of target distributions, it quickly becomes obvious that the distributions have to have some smoothness (otherwise the mass of the distribution can be concentrated arbitrarily closely around anywhere in space, so sampling will be trivially hard), and they have to be bounded (otherwise the mass can be arbitrarily far from the origin, so again sampling is trivially hard). If we allow for the target distributions π to be non log-concave, in most cases sampling will be trivially hard. We are able to obtain some non-trivial sampling lower bounds for non log-concave distributions, which are given in Chapter 7. But in the most part, to isolate out the sampling difficulty, we will mainly be focused on lower bounds specifically for log-concave distributions.

So what methods might one try to prove a sampling lower bound? The simplest idea is to reduce the sampling problem to an optimization problem. We can construct distributions whose mass are concentrated around their modes, which are the minimizers of their potential functions. Then if we can sample well from these distributions, we can locate the minimizer of the potential, so a lower bound on optimization directly leads to a sampling lower bound. Such an approach was taken in in [GLL22] But this approach is unsatisfying for log-concave sampling for two reasons. First, this reduction only captures the difficulty of finding *where* the mass of the target is, but it does not capture the difficulty of finding *how* the mass is spread out near the minimizer of the potential. Specifically, we could consider classes of distributions whose potentials are all globally minimized at the origin. The optimization difficulty for this class of potentials would be trivial, but the sampling difficulty remains non-trivial. Second, this approach is likely give very loose lower bound for sampling. For instance, in the high dimensional setting, the convex optimization lower bounds are dimension independent, so a sampling lower bound that is directly deduced from optimization will be dimension independent as well. However, this is contrary to existing evidence, where all known sampling algorithms seem to depend polynomially on the dimension.

Another idea is to mimic the proof technique of optimization lower bounds for sampling. Most optimization lower bounds hold against deterministic algorithms, and they construct specific adversarial functions [Bub15; Nes18] such that each step of the algorithm only learns a limited amount of information about the location of the minimizer. In contrast, lower bounds for randomized algorithms are relatively recent and still not fully understood [WS17], which poses a major challenge for sampling algorithms, since they are inherently randomized. Second, whereas optimization constructions can employ local perturbations to hide the minima, sampling constructions need to hide the bulk of the mass of the target distribution, making them surprisingly delicate.

A different approach for lower bounds is to use an information theoretic argument. For sampling, it would go something like this. Say we want to prove a lower bound for sampling over some class of distributions. We would then pick a subset of distributions in that class, such that if we are given samples from any distribution in this subset, we would be able to identify which one in the subset it is. This means that the sampling task is easier than the identification task, because if we can sample then we can use those samples to identify the target. Then if we could prove a lower bound of on the difficulty of identifying a distribution that is randomly chosen from the subset, that would imply a lower bound for the sampling task. It turns out that this is the approach that we will use. The key step is to construct distributions that can be distinguished only with samples. One might think that we can use samples to estimate some statistic of the distribution, which is then used to identify the target, but if we need too many samples to estimate the statistic, then the resulting lower bound would be vacuous. For example, [GLL20] proves a lower bound of $\tilde{\Omega}(d)$ for the number of queries needed to estimate the normalizing constant of log-smooth and strongly-log-concave distributions, but this does not translate into a meaningful sampling lower bound because we cannot estimate the normalizing constant accurately with $o(d)$ number of samples. Constructing the right distributions that can be distinguished with samples turns out to be the main difficulty and contribution of many of our lower bounds.

Chapter 7

Fisher information lower bounds

In this chapter, we explore the simplest idea for proving sampling lower bounds: by reducing sampling to optimization. To obtain non-trivial results, we will allow the target distribution to be non log-concave, and hence the results in this chapter are orthogonal to the log-concave lower bounds that we pursue in the rest of Part II. We will also measure the progress of sampling by the Fisher information, which is an unusual measure of divergence that behaves quite differently from standard measures such as KL or total variation. Our results have two interesting implications. The first is that in the low accuracy (large Fisher information from the target) sampling regime, an averaged version of unadjusted Langevin algorithm is optimal. The second is that in the high accuracy (small Fisher information from the target) regime, sampling algorithms must have polynomial dependence on the accuracy ε.

This chapter is based on the joint work [Che+23a], with Sinho Chewi, Patrik Gerber, and Holden Lee.

7.1 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 [VW19; Wib19; Ma+21; Che+22b], distributions with tail decay conditions [DM17; Che+18a; Xu+18; Li+19; MMS20; EH21; ZXG21; HBE22], or mixtures of log-concave distributions [LRG18].

Recently [Bal+22] 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], Balasubramanian et al. proposed to call any measure μ satisfying $\sqrt{\mathsf{FI}(\mu \parallel \pi)} \le \varepsilon$ an ε-stationary point for sampling, where $\mathsf{FI}(\mu \,\|\, \pi) := \mathbb{E}_\mu[\|\nabla \ln \frac{\mu}{\pi}\|^2]$ denotes the *relative Fisher information* of μ from π. They 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. They further showed that averaged Langevin Monte Carlo (LMC) can find an ε-stationary point in $\mathcal{O}(\beta^2 d K_0/\varepsilon^4)$ iterations, where $K_0 := \mathsf{KL}(\mu_0 \,\|\, \pi)$ is the initial Kullback–Leibler (KL) divergence to the target π.

In this chapter, we establish the first lower bounds for Fisher information guarantees for sampling, resolving an open question posed in [Bal+22]. As we discuss further 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 query complexity for sampling (Definition 4) are given in Section 7.2. Precise statements of our results are given in Sections 7.3 and 7.4. 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 (equivalence, informal)**.** *The following two 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 15 (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 $K_0 := \mathsf{KL}(\mu_0 \,\|\, \pi_\beta)$, is at* *least $\Omega(K_0/d)$. The lower bound is attained by averaged LMC (Langevin Monte Carlo) as given in [Bal+22].*

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, $\mathsf{FI}(\mu \,\|\, \pi_\beta) \lesssim \beta d$. 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 16 (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 $K_0 := \mathsf{KL}(\mu_0 \,\|\, \pi_\beta) \le 1$, is at least $(\sqrt{\beta}/\varepsilon)^{2d/(d+2)-o(1)}$ as $\varepsilon \to 0$.*

We give a more precise form of our lower bound in Section 7.4. In infinite dimension (actually, $d \ge \widetilde{\Omega}(\sqrt{\log(\beta/\varepsilon^2)})$ suffices, see Section 7.4), the lower bound reads $\widetilde{\Omega}(\beta/\varepsilon^2)$, which can be compared to the averaged LMC upper bound of $\mathcal{O}(\beta^2 d/\varepsilon^4)$. 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+18c; GLL20; CBL22; LST21b]. So far, unconditional lower bounds have only been obtained in restricted settings such as in dimension 1; see [Che+22d] 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 work takes 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 provides compelling evidence that Fisher information guarantees are the correct analogue of stationary point guarantees in non-convex optimization, thereby supporting the framework of [Bal+22].

- **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 Section 7.5 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, the recent work of [CBS22] exhibits an algorithm with $\mathcal{O}(\log(1/\varepsilon))$ complexity. To our knowledge, this is one of the first instances in which sampling is provably harder than optimization.

7.2 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 3 (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 4 (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 [Bal+22] shows that using averaged LMC,

$$\mathscr{C}(d, K_0, \varepsilon; \beta) \lesssim 1 \vee \frac{\beta^2 d K_0}{\varepsilon^4}\,. \tag{7.1}$$

We also note the following rescaling lemma.

Lemma 22 (rescaling)**.** *It holds that*

$$\mathscr{C}(d, K_0, \varepsilon; \beta) = \mathscr{C}\bigl(d, K_0, \frac{\varepsilon}{\sqrt{\beta}}; 1\bigr)\,.$$

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 Section 7.1.) 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 $\mu := \operatorname{law}(\sqrt{\beta}\, X_\beta)$. 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 $\hat\mu_\beta := \operatorname{law}(\hat X/\sqrt{\beta})$ 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 $\mathscr{C}(d, K_0, \varepsilon) := \mathscr{C}(d, K_0, \varepsilon; 1)$.

7.3 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 17 (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 $\mu_\beta := \mathcal{N}(x, \beta^{-1} I_d)$, it holds that $\mathsf{FI}(\mu_\beta \,\|\, \pi_\beta) \le 10\beta d$.*

2. *Conversely, suppose that μ is a probability measure on $\mathbb{R}^d$ such that $\mathsf{FI}(\mu \,\|\, \pi_\beta) \le \beta d$. Let $X \sim \mu$ be a sample. Then, $\|\nabla V(X)\| \le 3\varepsilon$ with probability at least $1/2$.*

Proof. See Section 7.6. □

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

$$X_t = X_{kh} - (t - kh)\, \nabla U(X_{kh}) + \sqrt{2}\, (B_t - B_{kh})\,, \tag{7.2}$$

where $(B_t)_{t \ge 0}$ is a standard Brownian motion in $\mathbb{R}^d$. Let $\mu_t := \operatorname{law}(X_t)$ denote the law of the algorithm at time t. Then, the *averaged* LMC algorithm at iteration N outputs a sample from the law of $\bar\mu_{Nh} := (Nh)^{-1} \int_0^{Nh} \mu_t \, \mathrm{d}t$. 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 $X_0, X_h, X_{2h}, \dots, X_{kh}$ using the LMC recursion, and then output X_t which is obtained via the partial LMC update (7.2).

Corollary 8 (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 $\int \|\cdot\|^2 \, \mathrm{d}\pi_\beta \le \operatorname{poly}(\Delta, d, 1/\varepsilon)$. Consider running averaged LMC with step size $h = \widetilde\Theta(1/\beta)$, initial distribution $\mu_0 = \mathcal{N}(0, \beta^{-1} I_d)$, 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 17 with the analysis of averaged LMC in [Bal+22]; see Section 7.6. □

This matches the usual $\mathcal{O}(\Delta/\varepsilon^2)$ 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 (7.2) recovers the gradient descent algorithm. However, it is remarkable that the analysis of [Bal+22] 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 18 (first lower bound)**.** *Suppose that the dimension d satisfies $\widetilde{\mathcal{O}}(K_0) \ge d \ge \widetilde{\Omega}(K_0^{2/3})$. 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 $f(0) - \inf f \le \Delta$. 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 17 completes the proof. Details are given in Section 7.6. □

The lower bound of Theorem 18 is matched by averaged LMC, see (7.1). In the theorem, the restriction $d \ge \widetilde{\Omega}(K_0^{2/3})$ arises because the lower bound construction of [Car+20] for finding a ε-stationary point of a smooth function requires a large dimension $d \ge \widetilde{\Omega}(1/\varepsilon^4)$. If, as conjectured in [BM20] and [CBS22], the lower bound construction can be embedded in dimension $d \gtrsim \log(1/\varepsilon)$, then the restriction in Theorem 18 would instead become $d \gtrsim \log K_0$.

7.4 Bump construction and the second lower bound

The main drawback of the first lower bound (Theorem 18) 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 22).

Theorem 19 (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*

$$\mathscr{C}(d, K_0 = 1, \varepsilon) \gtrsim \Big(\frac{cd}{\log(1/\varepsilon)}\Big)^{d/2} \frac{1}{\varepsilon^{2d/(d+2)}}\,. \tag{7.3}$$

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

$$\tilde{\pi}_\omega(x) \coloneqq \exp\Big(r^2 \phi\Big(\frac{\|x-\omega\|}{r}\Big) - \frac{1}{2}\,(\|x\| - R)_+^2\Big) \eqqcolon \exp\big(-V_\omega(x)\big)\,, \tag{7.4}$$

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,

$$\mathsf{TV}(\mu, \pi_\omega)^2 \le \frac{1}{4} C_{\mathsf{PI}}(\pi_\omega)\, \mathsf{FI}(\mu \,\|\, \pi_\omega)\,,$$

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 actual proof requires careful balancing of the parameters r, R, d and ε and some delicate calculations to satisfy all of the desired properties. The full details are given in Section 7.7. $\square$

The lower bound in Theorem 19 deteriorates in high dimension; note that due to the restriction $\varepsilon \le \exp(-c'd)$, the first factor in (7.3) 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

$$\mathscr{C}(d, K_0 = 1, \varepsilon) \gtrsim \max_{d_\star \le d} \Big[\Big(\frac{c d_\star}{\log(1/\varepsilon)} \Big)^{d_\star/2} \varepsilon^{4/(d_\star+2)} \Big] \frac{1}{\varepsilon^2}\,.$$

By optimizing over $d_\star$, we show (Section 7.7) that if $\varepsilon \le 1/C$, then

$$\begin{aligned}\mathscr{C}(d,1,\varepsilon) &\gtrsim \begin{cases} \dfrac{1}{\varepsilon^{2d/(d+2)} \exp(C\sqrt{\log(1/\varepsilon)\log\log(1/\varepsilon)})}\,, & \text{for all } d \lesssim \sqrt{\dfrac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}}\,, \\ \dfrac{1}{\varepsilon^{2} \exp(C\sqrt{\log(1/\varepsilon)\log\log(1/\varepsilon)})}\,, & \text{for all } d \gtrsim \sqrt{\dfrac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}}\,, \end{cases} \\ &= \frac{1}{\varepsilon^{\min\{2d/(d+2),2\}-o(1)}}\,, \qquad \text{for all } d \ge 1\,,\end{aligned}$$

as $\varepsilon \to 0$, where $C > 0$ is universal. Noting $2d/(d+2) < 2$, this yields the simplified bound in Theorem 16.

For $d = 1$, the lower bound of Theorem 19 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 20 (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 Section 7.7. □

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 Section 7.8. We highlight a few interesting conclusions of the discussion here.

- Although rejection sampling can indeed obtain Fisher information guarantees with complexity $\mathcal{O}(\log(1/\varepsilon))$ (Proposition 9), 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 $\mathcal{O}(\exp(3M_0)\log(1/\varepsilon))$.

- In the model with the stronger initialization oracle (i.e., bounded M_0), any algorithm which has $\operatorname{polylog}(1/\varepsilon)$ dependence on the accuracy ε necessarily incurs exponential dependence on M_0 (Corollary 11). 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 Che+21; LST21b; 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 $\mathcal{O}(\log(1/\varepsilon))$ complexity (Proposition 10).

7.5 Separation between log-concave and non log-concave sampling

We show that $\mathcal{O}(\log \frac{1}{\varepsilon})$ 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.

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 23 (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*

$$\mathsf{FI}(\mu Q_t \,\|\, \pi) \lesssim \frac{\varepsilon_\chi\,(d + \log(1/\varepsilon_\chi))}{t} + \beta^2 d t\,.$$

To prove Lemma 23, we start with

$$\begin{aligned}\mathsf{FI}(\mu Q_t \,\|\, \pi) &\coloneqq \int \|\nabla \ln(\mu Q_t)(x) - \nabla \ln \pi(x)\|^2 \,\mu Q_t(\mathrm{d}x) \\ &\le 2\,\mathsf{FI}(\mu Q_t \,\|\, \pi Q_t) + 2\int_{\mathbb{R}^d} \|\nabla \log(\pi Q_t)(x) - \nabla \log \pi(x)\|^2 \,\mu Q_t(\mathrm{d}x)\,. \end{aligned} \tag{7.5}$$

For the first term in (7.5), we use the following lemma on error in the score function (gradient of the log-density).

Lemma 24 (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 $\chi^2(\mu \,\|\, \pi) \le \varepsilon_\chi^2 \le 1$. Then,*

$$\mathsf{FI}(\mu Q_t \,\|\, \pi Q_t) = \int_{\mathbb{R}^d} \|\nabla \ln(\mu Q_t)(x) - \nabla \ln(\pi Q_t)(x)\|^2 \,\mu Q_t(\mathrm{d}x) \lesssim \frac{\varepsilon_\chi \left(d + \ln \frac{1}{\varepsilon_\chi}\right)}{t}\,.$$

For the second term in (7.5), we use the following score perturbation lemma.

Lemma 25 ([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 23.

Proof of Lemma 23. For the second term in (7.5), Lemma 25 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 26 below yields

$$\mathbb{E}_{\mu Q_t}[\|\nabla V\|^2] \lesssim \mathsf{FI}(\mu Q_t \,\|\, \pi) + \beta d\,.$$

Hence, from (7.5) and Lemma 24,

$$\begin{aligned}\mathsf{FI}(\mu Q_t \,\|\, \pi) &\lesssim \mathsf{FI}(\mu Q_t \,\|\, \pi Q_t) + \mathbb{E}_{\mu Q_t}[\|\nabla \ln(\pi Q_t) - \nabla \ln \pi\|^2] \\ &\lesssim \frac{\varepsilon_\chi\,(d + \log(1/\varepsilon_\chi))}{t} + \beta^2 dt + \beta^2 t^2\, \mathsf{FI}(\mu Q_t \,\|\, \pi)\,.\end{aligned}$$

If $t \lesssim 1/\beta$ for a small enough implied constant, it implies

$$\mathsf{FI}(\mu Q_t \,\|\, \pi) \lesssim \frac{\varepsilon_\chi\,(d + \log(1/\varepsilon_\chi))}{t} + \beta^2 dt$$

as desired. □

High-accuracy Fisher information guarantees for log-concave targets

We now apply the post-processing lemma (Lemma 23). We recall the following high-accuracy guarantee for sampling from log-concave targets in chi-squared divergence, based on the proximal sampler.

Theorem 21 ([Che+22a, Corollary 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 (RGO) 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*

$$N \le \widetilde{\mathcal{O}}\Big(C_{\mathsf{PI}}\beta d\,\big(\log(1+\chi^2(\mu_0 \,\|\, \pi)) \vee \log\frac{1}{\varepsilon_\chi}\big)\Big)\,.$$

We now briefly justify why this morally leads to an $\mathcal{O}(\log(1/\varepsilon))$ 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 23, 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 $\varepsilon_\chi \le \widetilde{\mathcal{O}}(\varepsilon^4/d^2)$. Applying Theorem 21, this can be achieved using

$$N = \widetilde{\mathcal{O}}\Big(C_{\mathsf{PI}} d\,\big(\log(1+\chi^2(\mu_0 \,\|\, \pi)) \vee \log\frac{\sqrt{d}}{\varepsilon}\big)\Big)$$

queries in expectation. Let us give crude bounds for these terms. First, let $\mathfrak{m}_2^2 := \mathbb{E}_\pi[\|\cdot\|^2]$ 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 $\log(1+\chi^2(\mu_0 \,\|\, \pi)) \le \widetilde{\mathcal{O}}(d)$ [see Che+22b, Lemma 29]. 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 19, 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 Section 7.2; 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 19, the effective radius R of the distributions depends on $1/\varepsilon$. We claim, however, that if we set $d, R = \operatorname{polylog}(1/\varepsilon)$, then the upper bound for log-concave targets is $\operatorname{polylog}(1/\varepsilon)$ (with the caveats just discussed) and the lower bound for non-log-concave targets is $\operatorname{poly}(1/\varepsilon)$. 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 $\operatorname{polylog}(1/\varepsilon)$ query complexity is possible for distributions satisfying a Poincaré inequality, which form a strict superclass of log-concave distributions.

7.6 Proofs for the first lower bound

Proof of the equivalence

In order to prove the equivalence in Theorem 17, we recall the following useful lemma from [Che+22b].

Lemma 26 ([Che+22b, Lemma 16]). *Let $\pi \propto \exp(-V)$ be a β-log-smooth density on $\mathbb{R}^d$. Then, for any probability measure μ,*

$$\mathbb{E}_\mu[\|\nabla V\|^2] \le \mathsf{FI}(\mu \,\|\, \pi) + 2\beta d\,.$$

With the lemma in hand, we are ready to prove Theorem 17.

Proof of Theorem 17. 1. We can explicitly compute

$$\begin{aligned}
\mathsf{FI}(\mu_\beta \,\|\, \pi_\beta) &= \int \|\nabla \ln \mu_\beta - \nabla \ln \pi_\beta\|^2 \,\mathrm{d}\mu_\beta \\
&= \int \|\beta\,(z-x) - \beta\,\nabla V(z)\|^2 \,\mathrm{d}\mu_\beta(z) \\
&\le 2\beta^2 \int \|z-x\|^2 \,\mathrm{d}\mu_\beta(z) + 2\beta^2 \int \|\nabla V(z)\|^2 \,\mathrm{d}\mu_\beta(z) \\
&\le 2\beta^2 \int \|z-x\|^2 \,\mathrm{d}\mu_\beta(z) + 4\beta^2 \int \{\|z-x\|^2 + \|\nabla V(x)\|^2\} \,\mathrm{d}\mu_\beta(z) \\
&\le 6\beta^2 \int \|z-x\|^2 \,\mathrm{d}\mu_\beta(z) + 4\beta^2 \underbrace{\|\nabla V(x)\|^2}_{\le \varepsilon^2}\,,
\end{aligned}$$

where we used that ∇V is Lipschitz. Also, $\int \|z - x\|^2 \,\mathrm{d}\mu_\beta(z) = d/\beta$. Hence,

$$\mathsf{FI}(\mu_\beta \,\|\, \pi_\beta) \le 6\beta d + 4\beta^2 \varepsilon^2 = 10\beta d\,,$$

provided $\beta = d/\varepsilon^2$.

2. Conversely, since $\nabla \ln(1/\pi_\beta) = \beta\,\nabla V$ is β-Lipschitz, then Lemma 26 yields

$$\mathbb{E}_\mu[\|\nabla V\|^2] = \frac{1}{\beta^2}\,\mathbb{E}_\mu[\|\nabla(\beta V)\|^2] \le \frac{1}{\beta^2}\,\{\mathsf{FI}(\mu \,\|\, \pi_\beta) + 2\beta d\} \le \frac{3d}{\beta}\,.$$

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 $1/2$.

□

Proof of the averaged LMC guarantee

In order to apply [Bal+22, Theorem 4], we need a bound on the KL divergence at initialization. Such bounds are standard; however, since [Che+22b, Lemma 30] assumes that we start at a stationary point of V (contrary to the present setting), we present an adapted version.

Lemma 27 (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 $\mu_0 = \mathcal{N}(0, \beta^{-1} I_d)$, we have the bound*

$$\mathsf{KL}(\mu_0 \,\|\, \pi) \lesssim \Delta + d\,\bigl(1 \vee \ln(\beta \mathfrak{m}^2)\bigr)\,.$$

Proof. Write

$$\frac{\mu_0}{\pi} = \exp\Bigl(U - \frac{\beta}{2}\,\|\cdot\|^2\Bigr)\,\frac{\int \exp(-U)}{\int \exp(-U - \delta\,\|\cdot\|^2)}\,\frac{\int \exp(-U - \delta\,\|\cdot\|^2)}{(2\pi/\beta)^{d/2}}\,,$$

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 $x = -\frac{1}{\beta}\,\nabla U(0)$,

$$U\Bigl(-\frac{1}{\beta}\,\nabla U(0)\Bigr) - U(0) \le -\frac{1}{2\beta}\,\|\nabla U(0)\|^2$$

or

$$\|\nabla U(0)\|^2 \le 2\beta \left(U(0) - U\bigl(-\frac{1}{\beta}\,\nabla U(0)\bigr)\right) \le 2\beta \left(U(0) - \inf U\right) \le 2\beta\Delta\,.$$

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,

$$\frac{\int \exp(-U - \delta\,\|\cdot\|^2)}{(2\pi/\beta)^{d/2}} \le \frac{\exp(-\inf U)\int \exp(-\delta\,\|\cdot\|^2)}{(2\pi/\beta)^{d/2}} = \exp(-\inf U)\,\bigl(\frac{\beta}{2\delta}\bigr)^{d/2}\,.$$

Combining these bounds,

$$\mathsf{KL}(\mu_0 \,\|\, \pi) = \mathbb{E}_{\mu_0}\ln\frac{\mu_0}{\pi} \le U(0) - \inf U + \Delta + \frac{\beta}{2}\,\mathbb{E}_{\mu_0}[\|\cdot\|^2] + \ln 2 + 4\delta\mathfrak{m}^2 + \frac{d}{2}\ln\frac{\beta}{2\delta}\,.$$

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 8. 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

$$K_0 \coloneqq \mathsf{KL}(\mu_0 \,\|\, \pi_\beta) \lesssim \beta\Delta + d\left(1 \vee \ln(\beta\,\mathbb{E}_{\pi_\beta}[\|\cdot\|^2])\right) = \widetilde{\mathcal{O}}(\beta\Delta + d)\,. \tag{7.6}$$

The main result of [Bal+22] 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 (7.6), 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 17. □

Proof of the first lower bound

Proof of Theorem 18. Let $\mathscr{F}$ be the family of functions 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 and satisfies $f(0) - \inf f \le \Delta$, 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 $\delta := 4\sqrt{\beta d}$. From the Fisher information lemma (Lemma 26), 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 $\mu_0 := \mathcal{N}(0, \beta^{-1} I_d)$. We need to compute the value of $K_0 := \sup_{f\in\mathscr{F}} \mathsf{KL}(\mu_0 \,\|\, \pi_f)$, and for this we use Lemma 27. 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 $\lambda > 0$,

$$\mathbb{E}_{\pi_f}[\|\cdot\|^2] \le \frac{1}{\lambda}\,\{\mathsf{KL}(\pi_f \,\|\, \nu) + \ln \mathbb{E}_\nu \exp(\lambda\,\|\cdot\|^2)\}\,,$$

where $\nu := \mathcal{N}(0, \tau I_d)$. 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

$$\mathbb{E}_{\pi_f}[\|\cdot\|^2] \le \mathsf{poly} \cdot \big(1 + \mathsf{FI}(\pi_f \,\|\, \nu)\big)\,.$$

The Fisher information is computed to be

$$\mathsf{FI}(\pi_f \,\|\, \nu) = \mathsf{poly} \cdot \mathbb{E}_{\pi_f}\big[\big\|\nabla\big(\tilde{f}_{T,U}\big(\rho(\cdot/\mathsf{poly})\big)\big)\big\|^2\big]\,.$$

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 27, we can take $K_0 \lesssim \Delta + \widetilde{\mathcal{O}}(d)$. 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 is satisfied provided $d \ge \widetilde{\Omega}(K_0^{2/3})$. □

7.7 Proofs for the second lower bound

Proof of Theorem 19

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

$$\tilde{\pi}_\omega(x) \coloneqq \exp\Bigl(r^2\phi\Bigl(\frac{\|x-\omega\|}{r}\Bigr) - \frac{1}{2}\,(\|x\| - R)_+^2\Bigr) \eqqcolon \exp\bigl(-V_\omega(x)\bigr)\,, \tag{7.7}$$

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 $\alpha > 0$.

[1] One such function is $\phi(x) = \begin{cases} \frac{11}{64} - \frac{1}{2}x^2\,, & \text{for } x \in [0,1/4]\,,\\ \frac{1}{27}\left(4 + 8x - 48x^2 + 56x^3 - 20x^4\right), & \text{for } x \in [1/4,1]\,,\\ 0\,, & \text{otherwise}\,. \end{cases}$

(ϕ**.3**) $\sup_{x>0} |\phi''(x)| \le 1$.

The above implies that on $\mathbb{R}^d$, $x \mapsto \phi(\|x\|)$ is 1-smooth (see Lemma 32), 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

$$\tilde\pi_{\mathsf{init}}(x) \coloneqq \exp\bigl(-\frac{1}{2}\,(\|x\| - R)_+^2\bigr)\,,$$

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 28 (lower bound construction)**.** *There exist universal constants $c_\varepsilon, c > 0$ such that for every $d \ge 1$ and $\varepsilon \le \exp(-c_\varepsilon d)$ 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,*

$$\sqrt{\mathsf{FI}(\mu \,\|\, \pi_\omega)} \le \varepsilon \implies \mathsf{TV}(\mu, \pi_\omega) \le \frac{1}{3}\,.$$

(P.3) *(initial KL divergence) There exists a probability measure π_{init} that satisfies*

$$\max_{\omega \in \mathscr{P}_{2r,R}} \mathsf{KL}(\pi_{\mathsf{init}} \,\|\, \pi_\omega) \le \log 2\,.$$

(P.4) *(lower bound on the packing number) It holds that*

$$|\mathscr{P}_{2r,R}| \ge \Bigl(\frac{cd}{\log(1/\varepsilon)}\Bigr)^{d/2} \frac{1}{\varepsilon^{2d/(d+2)}}\,.$$

Proof. Property **(P.1)** holds by the definition of the parameters r and R, see (7.15) and Lemma 30. We prove Property **(P.2)** in Section 7.7, Property **(P.3)** in Section 7.7, and Property **(P.4)** in Section 7.7. □

Remark 6. In order for the bound in Property **(P.4)** to be non-trivial, i.e., $|\mathscr{P}_{2r,R}| \gtrsim 1$, we require $\varepsilon^{-2d/(d+2)} \gtrsim (\sqrt{\log(1/\varepsilon)/(cd)})^d$. Taking logarithms, we want

$$\frac{2d}{d+2}\log\frac{1}{\varepsilon} \overset{!}{\geq} \frac{d}{2}\log\log\frac{1}{\varepsilon} - \frac{d}{2}\log d + \Omega(d)\,.$$

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 prove Theorem 19 by applying a standard information theoretic argument with Fano's inequality (Theorem 1).

Proof of Theorem 19. 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 Section 7.2, 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

$$\widehat{\omega} := \arg\min_{\omega\in\mathscr{P}_{2r,R}} \|X - \omega\|\,.$$

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 $\mathsf{TV}(\mu_N, \pi_\omega) \leq 1/3$. Let $X \sim \mu_N$; then,

$$\mathbb{P}\{X \in \omega + B_r\} = \mu_N(\omega + B_r) \geq \pi_\omega(\omega + B_r) - \mathsf{TV}(\mu_N, \pi_\omega) \geq \frac{1}{2} - \frac{1}{3} = \frac{1}{6}\,,$$

where we used Property **(P.1)**. Hence we see that

$$\mathbb{P}\{\widehat{\omega} = \omega\} \geq \frac{1}{6}\,. \tag{7.8}$$

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 1):, we therefore have

$$\mathbb{P}\{\widehat{\omega} \neq \omega\} \ge 1 - \frac{I(\xi_N; \omega) + \ln 2}{\ln M}. \tag{7.9}$$

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

$$H\big(V_\omega(x_i), \nabla V_\omega(x_i) \bigm| \xi_{i-1}\big) \le h\Big(\frac{1}{|\mathscr{P}_{2r,R}(i)|}\Big) \le h\big(\frac{2}{M}\big)\,,$$

where $h(p) \coloneqq p \ln \frac{1}{p} + (1-p) \ln \frac{1}{1-p}$ is the binary entropy function. Assuming that $M \geq 4$ (which can be ensured thanks to Remark 6),

$$h\Big(\frac{2}{M}\Big) \leq \frac{4}{M} \ln \frac{M}{2}\,.$$

Substituting this into (7.9),

$$\mathbb{P}\{\widehat{\omega} \neq \omega\} \geq 1 - \frac{\frac{4N}{M}\ln(M/2) + \ln 2}{\ln M}\,. \tag{7.10}$$

If $M \geq 4$, and $N \leq \frac{1}{12}\,M$, we would obtain $\mathbb{P}\{\widehat{\omega} \neq \omega\} > 5/6$, contradicting (7.8). 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 (7.10) conditional on ζ, and obtain

$$\mathbb{P}\{\widehat{\omega} \neq \omega \mid \zeta\} \geq 1 - \frac{\frac{4N}{M}\ln(M/2) + \ln 2}{\ln M}\,.$$

Taking expectation over ζ, we see that the lower bound (7.10), and hence $N \gtrsim M$, holds for randomized algorithms as well.

The proof of Theorem 19 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 28.

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 $B_R \setminus (\omega + B_r)$,

$$\begin{aligned} Z_\omega &= \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + \tilde{\pi}_\omega(B_R) \\ &= \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + (R^d - r^d)\, V_d + \int_{B_r} \exp\Big(r^2 \phi\big(\frac{\|x\|}{r}\big)\Big)\, \mathrm{d}x \\ &= \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + (R^d - r^d)\, V_d + r^d\, I_r\,, \end{aligned}$$

where we define $I_r := \int_{B_1} \exp(r^2 \phi(\|x\|))\, \mathrm{d}x$. We record some useful properties of the quantities defined thus far that will be used throughout the proof of Lemma 28.

Lemma 29 (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)*
$$\frac{1}{2} \le \frac{r^d\, I_r}{(2\pi)^{d/2} \exp(r^2 \phi(0))} \le 2\,. \tag{7.11}$$

2. *(mass outside B_R) There is a universal constant $c_0 > 2$, independent of c, such that*
$$\sqrt{\frac{\pi}{2}}\, V_d d R^{d-1} \le \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) \le V_d c_0^d \left(dR^{d-1} + d^{(d+1)/2}\right). \tag{7.12}$$

3. *(mass on the bump)*
$$\ln \frac{I_r}{V_d} \ge cd\,. \tag{7.13}$$

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:

$$\begin{aligned} I_r &:= \int_{B_1} \exp\bigl(r^2 \phi(\|x\|)\bigr)\, \mathrm{d}x \\ &= \underbrace{\int_{B_1 \setminus B_\alpha} \exp\bigl(r^2 \phi(\|x\|)\bigr)\, \mathrm{d}x}_{\mathsf{A}} + \underbrace{\exp\bigl(r^2\phi(0)\bigr) \int_{B_\alpha} \exp\Bigl(-\frac{r^2\,\|x\|^2}{2}\Bigr)\, \mathrm{d}x}_{\mathsf{B}}\,. \end{aligned}$$

As ϕ is decreasing by (ϕ.**1**), clearly $\mathsf{A} \le V_d \exp(r^2 \phi(\alpha))$, and the second term is given by

$$\mathsf{B} = \frac{(2\pi)^{d/2}}{r^d} \exp\bigl(r^2\phi(0)\bigr)\, \mathbb{P}(\|X\| \le \alpha r)\,,$$

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 $r \ge c_1\sqrt{d}$. Recall that $\log\Gamma(\frac{d}{2}+1) = \frac{d}{2}\log d + \mathcal{O}(d)$. Thus, for $r \ge c_1\sqrt{d}$ we have

$$\begin{aligned} \log\frac{\mathsf{A}}{\mathsf{B}} &\le \log\frac{2V_d \exp(r^2\phi(\alpha))}{(2\pi)^{d/2}\, r^{-d} \exp(r^2\phi(0))} = \log\frac{2\exp(r^2\phi(\alpha))}{2^{d/2}\, r^{-d}\exp(r^2\phi(0))\,\Gamma(\frac{d}{2}+1)} \\ &= \mathcal{O}(d) - r^2\left(\phi(0) - \phi(\alpha)\right) + d\log r - \frac{d}{2}\log d \\ &= \mathcal{O}(d) - d\left(\bigl(\frac{r}{\sqrt{d}}\bigr)^2 \left(\phi(0)-\phi(\alpha)\right) - \log\bigl(\frac{r}{\sqrt{d}}\bigr)\right). \end{aligned}$$

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:

$$\mathsf{B} \leq I_r \leq 2\mathsf{B}\,, \tag{7.14}$$

proving (7.11). We now turn to the proof of (7.12). 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 (7.13). We again use the fact $\log\Gamma(\frac{d}{2}+1) = \frac{d}{2}\log d + \mathcal{O}(d)$. Therefore, for $r \geq (c_1 \vee c_2)\sqrt{d}$ and using (7.11) we obtain

$$\begin{aligned}
\log\frac{I_r}{V_d} &\geq \log\frac{(2\pi)^{d/2}\,r^{-d}\exp(r^2\phi(0))/2}{\pi^{d/2}/\Gamma(\frac{d}{2}+1)} \\
&= d\,\Bigl(\bigl(\frac{r}{\sqrt{d}}\bigr)^2\,\phi(0) - \log\bigl(\frac{r}{\sqrt{d}}\bigr)\Bigr) + \mathcal{O}(d)\,.
\end{aligned}$$

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. □

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 provided that

$$f(r) \coloneqq (I_r + V_d)\,r^d \overset{!}{=} \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + V_dR^d \eqqcolon g(R)\,. \tag{7.15}$$

Lemma 30 (choice of r, R)**.** *For any value of $d \geq 1$ and $R \geq 0$, there exists a corresponding value of r such that (7.15) 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 (7.15) satisfies*

$$r \geq c_r\big(\log(6c_0)\big)\sqrt{d}\,, \tag{7.16}$$

$$R/r \geq 2\,, \tag{7.17}$$

where $c_r(\cdot)$ is the function defined in Lemma 29.

The argument $\log(6c_0)$ to $c_r(\cdot)$ in Lemma 30 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 (7.16). Note that since (7.16) is a hypothesis of Lemma 29, we cannot invoke Lemma 29 during the proof of (7.16) 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

$$d\log\frac{R}{r} \leq \log\big(1 + \frac{I_r}{V_d}\big) = \log\Big(1 + \frac{\int_{B_1}\exp(r^2\phi(\|x\|))\,\mathrm{d}x}{V_d}\Big) \leq \log\big(1 + \exp\big(r^2\phi(0)\big)\big)\,.$$

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 (7.16).

Next, we prove (7.17). 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 (7.12) from Lemma 29,

$$(I_r + V_d)\,r^d \leq V_d\,(c_0^d dR^{d-1} + c_0^d d^{(d+1)/2} + R^d) \leq V_d\,(c_0^d\sqrt{d}R^d + c_0^d\sqrt{d}R^d + R^d)$$

$$\leq V_d \cdot 3c_0^d \sqrt{d}\, R^d\,.$$

Taking logarithms, rearranging, and using (7.13) from Lemma 29,

$$d \log \frac{R}{r} \geq \log\bigl(1 + \frac{I_r}{V_d}\bigr) - d \log c_0 - \log(3\sqrt{d}) \geq \bigl(c - \log c_0 - \underbrace{\frac{\log(3\sqrt{d})}{d}}_{\leq \log 3}\bigr)\, d\,.$$

Taking $c = \log c_0 + \log 3 + \log 2 = \log(6c_0)$, this implies $R/r \geq 2$ as desired. □

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

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

We begin with a Poincaré inequality for π_{init}.

Lemma 31 (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

$$\begin{aligned}
\mathbb{E}_{\pi_{\mathsf{init}}}[\|\cdot\|^2] &= \frac{\int_{B_R} \|\cdot\|^2}{Z_{\mathsf{init}}} + \frac{\int_{\mathbb{R}^d \setminus B_R} \|\cdot\|^2 \exp(-\frac{1}{2}\,(\|\cdot\| - R)^2)}{Z_{\mathsf{init}}} \\
&\leq \frac{\int_{B_R} \|\cdot\|^2}{V_d R^d} + \frac{A_{d-1} \int_0^\infty (r+R)^{d+1} \exp(-r^2/2)\, \mathrm{d}r}{A_{d-1} \int_0^\infty (r+R)^{d-1} \exp(-r^2/2)\, \mathrm{d}r} \\
&\leq R^2 + \int (r+R)^2\, \nu(\mathrm{d}r)\,,
\end{aligned}$$

where ν is the probability measure on $\mathbb{R}_+$ with density

$$\nu(r) \propto (r+R)^{d-1} \exp\bigl(-\frac{r^2}{2}\bigr)\,. \tag{7.18}$$

Note that ν is 1-strongly-log-concave. Hence, by [DM+19, Proposition 1],

$$\int (r+R)^2\,\nu(\mathrm{d}r) \lesssim R^2 + \int r^2\,\nu(\mathrm{d}r) \lesssim R^2 + r_\star^2 + \int (r-r_\star)^2\,\nu(\mathrm{d}r) \le R^2 + r_\star^2 + 1\,,$$

where $r_\star$ is the mode of ν. To find the mode, (7.18) 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 22 (Holley–Stroock perturbation principle, [HS86])**.** *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 9 (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 ($\boldsymbol{\phi}$**.1**), we know that $\tilde\pi_\omega \ge \tilde\pi_{\mathsf{init}}$ and hence $Z_\omega \ge Z_{\mathsf{init}}$. It follows that

$$\frac{Z_{\mathsf{init}}}{Z_\omega} \le \frac{\pi_\omega}{\pi_{\mathsf{init}}} = \frac{\tilde\pi_\omega}{\tilde\pi_{\mathsf{init}}}\,\frac{Z_{\mathsf{init}}}{Z_\omega} \le \frac{\tilde\pi_\omega}{\tilde\pi_{\mathsf{init}}} \le \exp\bigl(r^2\phi(0)\bigr)\,.$$

Also, by (7.15),

$$\begin{aligned} Z_\omega = \tilde\pi_\omega(\mathbb{R}^d\setminus B_R) + V_d\,R^d + (I_r - V_d)\,r^d &\le \tilde\pi_\omega(\mathbb{R}^d\setminus B_R) + V_d\,R^d + (I_r + V_d)\,r^d \\ &= 2\,\bigl(\tilde\pi_\omega(\mathbb{R}^d\setminus B_R) + V_d\,R^d\bigr) = 2\,Z_{\mathsf{init}}\,. \end{aligned}$$

Hence, $Z_{\mathsf{init}}/Z_\omega \ge 1/2$. The result now follows from Lemma 31 and the Holley–Stroock perturbation principle (Theorem 22). □

To translate Fisher information guarantees into total variation guarantees, we use the following consequence of the Poincaré inequality.

Proposition 8 (Fisher information controls total variation). *Suppose that a probability measure π satisfies a Poincaré inequality. Then, for any probability measure μ,*

$$\mathsf{TV}(\mu, \pi)^2 \le \frac{C_{\mathsf{PI}}(\pi)}{4}\, \mathsf{FI}(\mu \,\|\, \pi)\,.$$

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 $\varepsilon \le \exp(-c_\varepsilon d)$, then we can choose r and R (depending on ε) such that: (i) r and R are related according to (7.15), which is necessary for Property (**P.1**); (ii) $R \ge c_R\sqrt{d}$, which is necessary for Lemma 30; and (iii) Property (**P.2**) holds.

*Proof of Property (**P.2**).* For any $\omega \in \mathscr{P}_{2r,R}$, suppose that μ satisfies $\sqrt{\mathsf{FI}(\mu \,\|\, \pi_\omega)} \le \varepsilon$. Then, by Corollary 9 and Proposition 8, we have

$$\mathsf{TV}^2(\mu, \pi_\omega) \le \frac{C_{\mathsf{PI}}(\pi_\omega)}{4}\, \mathsf{FI}(\mu \,\|\, \pi) \le \frac{c_{\mathsf{PI}} R^2 \exp(r^2\phi(0))}{2d}\, \varepsilon^2\,. \tag{7.19}$$

Hence, if we choose

$$R^2 = \frac{2d}{9c_{\mathsf{PI}}\varepsilon^2 \exp(r^2\phi(0))} \tag{7.20}$$

then $\sqrt{\mathsf{FI}(\mu \,\|\, \pi_\omega)} \le \varepsilon$ implies $\mathsf{TV}(\mu, \pi_\omega) \le 1/3$, i.e., Property (**P.2**) holds.

To justify (7.20), 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 (7.15) holds. Also, $r(\cdot)$ is an increasing function. In order for (7.20) to hold, it is equivalent to require

$$R^2 \exp\bigl(r(R)^2\, \phi(0)\bigr) = \frac{2d}{9c_{\mathsf{PI}}\varepsilon^2} \tag{7.21}$$

where the left-hand side is an increasing function of R. We also want R to satisfy $R \ge c_R\sqrt{d}$, where c_R is the universal constant in Lemma 30. From Lemma 30, for the choice of $R = c_R\sqrt{d}$,

$$r(c_R\sqrt{d}) \le \frac{c_R\sqrt{d}}{2}\,.$$

Therefore, for this choice of R, the left-hand side of (7.21) is bounded by

$$c_R^2 d \exp\bigl(\frac{c_R^2 d}{4}\, \phi(0)\bigr)\,.$$

If it holds that

$$\varepsilon^2 \le \frac{2}{9c_{\mathsf{PI}} c_R^2 \exp(c_R^2 d\phi(0)/4)} \tag{7.22}$$

then the R satisfying (7.20) necessarily satisfies $R \ge c_R\sqrt{d}$. In turn, (7.22) holds if $\varepsilon \le \exp(-c_\varepsilon d)$ for a universal constant $c_\varepsilon > 0$. □

Proof of Property (P.3)

*Proof of Property **(P.3)**.* In the proof of Corollary 9, we showed that $Z_\omega \le 2\,Z_{\mathsf{init}}$. The KL divergence is bounded by

$$\mathsf{KL}(\pi_{\mathsf{init}} \,\|\, \pi_\omega) = \mathbb{E}_{\pi_{\mathsf{init}}} \ln\Bigl(\underbrace{\frac{\tilde{\pi}_{\mathsf{init}}}{\tilde{\pi}_\omega}}_{\le 1}\,\underbrace{\frac{Z_\omega}{Z_{\mathsf{init}}}}_{\le 2}\Bigr) \le \log 2\,,$$

which is what we wanted to show. □

Proof of Property (P.4)

*Proof of Property **(P.4)**.* We choose r and R to satisfy (7.15) and (7.20). If $\varepsilon \le \exp(-c_\varepsilon d)$, then we showed in the proof of Property **(P.2)** that $R \ge c_R\sqrt{d}$ and hence Lemmas 29 and 30 apply.

As in the proof of (7.17) in Lemma 30, $R \ge \sqrt{d}$ implies

$$(I_r + V_d)\,r^d \le V_d \cdot 3c_0^d\sqrt{d}\,R^d\,.$$

Taking logarithms in (7.11) from Lemma 29 and using the above inequality, we obtain

$$r^2\phi(0) \le \log\frac{2r^d I_r}{(2\pi)^{d/2}} \le \mathcal{O}(d) + \log V_d + d\log R\,.$$

From (7.20), we have

$$\log R = \frac{1}{2}\log d + \log\frac{1}{\varepsilon} - \frac{1}{2}r^2\phi(0) + \mathcal{O}(1)\,.$$

Substituting this in and using $\log V_d = -\frac{d}{2}\log d + \mathcal{O}(d)$,

$$r^2\phi(0) \le d\log\frac{1}{\varepsilon} - \frac{d}{2}r^2\phi(0) + \mathcal{O}(d)$$

which is rearranged to yield

$$r^2\phi(0) \le \frac{2d}{d+2}\log\frac{1}{\varepsilon} + \mathcal{O}(1)\,.$$

Then, the packing number is lower bounded by

$$\begin{aligned}|\mathscr{P}_{2r,R}| \ge \Bigl(\frac{R-r}{2r}\Bigr)^d &\ge \Bigl(\frac{R}{4r}\Bigr)^d \\ &\ge \Bigl(c\,\frac{\sqrt{d}\exp(-\frac{d}{d+2}\log(1/\varepsilon))}{\varepsilon\sqrt{\log(1/\varepsilon)}}\Bigr)^d \\ &\ge \Bigl(c\sqrt{\frac{d}{\log(1/\varepsilon)}}\Bigr)^d\,\frac{1}{\varepsilon^{2d/(d+2)}}\,,\end{aligned}$$

for some universal constant c. □

Auxiliary lemmas

Lemma 32. *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 $x \ne 0$

$$\begin{aligned}\frac{\partial^2\phi(\|x\|)}{\partial x_j\,\partial x_i} &= \frac{\partial}{\partial x_j}\,\phi'(\|x\|)\,\frac{x_i}{\|x\|} \\ &= \phi''(\|x\|)\,\frac{x_i x_j}{\|x\|^2} - \phi'(\|x\|)\,\frac{x_i x_j}{\|x\|^3} + \delta_{i,j}\,\phi'(\|x\|)\,\frac{1}{\|x\|}\,.\end{aligned}$$

Thus, in matrix form we have

$$\nabla_x^2\phi(\|x\|) = \frac{\phi'(\|x\|)}{\|x\|}\,I_d + \Bigl(\frac{\phi''(\|x\|)}{\|x\|^2} - \frac{\phi'(\|x\|)}{\|x\|^3}\Bigr)\,xx^{\mathsf{T}}.$$

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. □

Optimization of the bound

We wish to find d which maximizes

$$\Bigl(\frac{cd}{\log(1/\varepsilon)}\Bigr)^{d/2}\,\varepsilon^{4/(d+2)}\,,$$

or after taking logarithms, we wish to maximize

$$f(d) := \frac{d}{2}\log d - \frac{4}{d+2}\log\frac{1}{\varepsilon} - \frac{d}{2}\log\log\frac{1}{\varepsilon} - \frac{d}{2}\log\frac{1}{c}\,.$$

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.,

$$\frac{d\,(d+2)\log d}{8} \ge \log\frac{1}{\varepsilon}\,.$$

It suffices to find d such that $g(d) := d^2\log d \ge 8\log(1/\varepsilon)$. 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

$$\begin{aligned}\log g\Bigl(\alpha\sqrt{\frac{y}{\log y}}\Bigr) &= 2\log\alpha + \log y - \log\log y + \log\log\Bigl(\alpha\sqrt{\frac{y}{\log y}}\Bigr)\\ &= 2\log\alpha + \log y + \underbrace{\log\frac{\frac{1}{2}\log y - \frac{1}{2}\log\log y + \log\alpha}{\log y}}_{\to\log(1/2)\ \text{as}\ y\to\infty}\,.\end{aligned}$$

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 19 holds. Then,

$$f(d) \ge -\frac{d}{2}\log\log\frac{1}{\varepsilon} - \frac{d}{2}\log\frac{1}{c} \asymp -\sqrt{\bigl(\log\frac{1}{\varepsilon}\bigr)\,\bigl(\log\log\frac{1}{\varepsilon}\bigr)}\,.$$

This verifies the expression in Section 7.4.

To justify the simplified expression of the bound that we gave in the informal statement of Theorem 16, note that in dimension

$$d \lesssim \sqrt{\frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}} \tag{7.23}$$

we have

$$\log\Bigl(\bigl(\frac{cd}{\log(1/\varepsilon)}\bigr)^{d/2}\Bigr) = \frac{d}{2}\underbrace{\Bigl(\log d - \log\log\frac{1}{\varepsilon} - \log\frac{1}{c}\Bigr)}_{\text{negative as } \varepsilon\searrow 0} \gtrsim -\sqrt{\bigl(\log\frac{1}{\varepsilon}\bigr)\bigl(\log\log\frac{1}{\varepsilon}\bigr)}\,.$$

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 (7.23) 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 (7.23) fails, then we instead have the bound

$$\mathscr{C}(d,1,\varepsilon) \gtrsim \frac{1}{\varepsilon^{2}\exp(C\sqrt{\log(1/\varepsilon)\log\log(1/\varepsilon)})} \ge \frac{1}{\varepsilon^{2d/(d+2)}\exp(C\sqrt{\log(1/\varepsilon)\log\log(1/\varepsilon)})}\,.$$

In either case, we have $\mathscr{C}(d,1,\varepsilon) \ge (1/\varepsilon)^{2d/(d+2)-o(1)}$. Together with Theorem 20 on the univarate case and Lemma 22 on rescaling, it yields Theorem 16.

Proof of Theorem 20

In the univarate case, we can sharpen Theorem 19 by obtaining a better bound on the Poincaré constant of π_ω. We use the following result.

Theorem 23 (Muckenhoupt's criterion)**.** *Let π be a probability density on $\mathbb{R}$ and let m be a median of π. Then,*

$$C_{\mathsf{PI}}(\pi) \asymp \max\Bigl\{\sup_{x<m}\pi\bigl((-\infty,x]\bigr)\int_x^m\frac{1}{\pi},\ \sup_{x>m}\pi\bigl([x,+\infty)\bigr)\int_m^x\frac{1}{\pi}\Bigr\}\,.$$

Proof. See [BGL14, Theorem 4.5.1]. □

Lemma 33 (improved Poincaré inequality for π_ω)**.** *Suppose that $d = 1$ and $R \ge 1$. Then, for all $\omega \in \mathscr{P}_{2r,R}$,*

$$C_{\mathsf{PI}}(\pi_\omega) \lesssim R^2\,.$$

Proof. We use Muckenhoupt's criterion (Theorem 23). First, we note that by Property **(P.1)**, it holds that $\pi_\omega(\omega + B_r) = \frac{1}{2}$ which implies that $\omega - r \le m \le \omega + r$. We proceed to check that

$$\sup_{x>m}\pi_\omega\bigl([x,+\infty)\bigr)\int_m^x\frac{1}{\pi_\omega} \lesssim R^2\,.$$

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 9, we have $Z_\omega \le 2\,Z_{\mathsf{init}} = 2\,\tilde\pi_{\mathsf{init}}(\mathbb{R}\setminus B_R) + 4R \le 2\sqrt{2\pi} + 4R \lesssim R$. Then,

$$\pi_\omega\bigl([x,+\infty)\bigr)\int_m^x \frac{1}{\pi_\omega} \le Z_\omega\,(x-m) \lesssim Rr \le R^2\,.$$

Next, suppose that $\omega + r < x < R$. Then,

$$\pi_\omega\bigl([x,+\infty)\bigr)\int_m^x \frac{1}{\pi_\omega} = \tilde\pi_\omega\bigl([x,+\infty)\bigr)\int_m^x \frac{1}{\tilde\pi_\omega} \le \Bigl(R - x + \sqrt{\frac{\pi}{2}}\Bigr)\,(x-m) \lesssim R^2\,.$$

Finally, suppose that $x > R$. Then, using standard Gaussian tail bounds,

$$\begin{aligned}\pi_\omega\bigl([x,+\infty)\bigr)\int_m^x \frac{1}{\pi_\omega} &= \tilde\pi_\omega\bigl([x,+\infty)\bigr)\int_m^x \frac{1}{\tilde\pi_\omega}\\ &\le \Bigl[\sqrt{2\pi}\,\bigl(\frac{1}{2}\wedge\frac{1}{x-R}\bigr)\exp\Bigl(-\frac{(x-R)^2}{2}\Bigr)\Bigr]\,\Bigl[R - m + (x-R)\exp\Bigl(\frac{(x-R)^2}{2}\Bigr)\Bigr]\,.\end{aligned}$$

If $x - R \le 1$, then this yields

$$\pi_\omega\bigl([x,+\infty)\bigr)\int_m^x \frac{1}{\pi_\omega} \lesssim R\,.$$

Otherwise, if $x - R \ge 1$, then we obtain

$$\begin{aligned}\pi_\omega\bigl([x,+\infty)\bigr)\int_m^x \frac{1}{\pi_\omega} &\lesssim \frac{1}{x-R}\exp\Bigl(-\frac{(x-R)^2}{2}\Bigr)\,\Bigl[R + (x-R)\exp\Bigl(\frac{(x-R)^2}{2}\Bigr)\Bigr]\\ &\lesssim R\,.\end{aligned}$$

This completes the proof. □

We now use the improved Poincaré inequality in order to establish Theorem 20.

Proof of Theorem 20. We follow the proof of Theorem 19. The proofs of Properties **(P.1)** and **(P.3)** remain unchanged.

In the proof of Property **(P.2)**, the equation (7.19) is replaced by

$$\mathsf{TV}^2(\mu,\pi_\omega) \le c_{\mathsf{PI}} R^2 \varepsilon^2$$

for a different universal constant $c_{\mathsf{PI}} > 0$, using Lemma 33. Hence, we choose $R^2 = 1/(9c_{\mathsf{PI}}\varepsilon^2)$ in order to verify Property **(P.2)**. Since we require $R \ge c_R$ for a universal constant $c_R \ge 1$, this requires $\varepsilon \le \exp(-c_\varepsilon)$ for a universal constant $c_\varepsilon > 0$.

Next, we turn towards the sharpened statement of Property (**P.4**). From (7.15), r is chosen so that

$$(I_r + 2)\, r = \tilde{\pi}_\omega(\mathbb{R} \setminus B_R) + 2R\,.$$

Using (7.11) from Lemma 29, we have

$$r I_r \asymp \exp\bigl(r^2 \phi(0)\bigr) \gtrsim r\,.$$

This implies that

$$\exp\bigl(r^2 \phi(0)\bigr) \gtrsim (I_r + 2)\, r \gtrsim R\,,$$

or $r \gtrsim \sqrt{\log R} \asymp \sqrt{\log(1/\varepsilon)}$. Hence,

$$|\mathscr{P}_{2r,R}| \ge \frac{R}{4r} \gtrsim \frac{1}{\varepsilon \sqrt{\log(1/\varepsilon)}}\,.$$

By substituting this new bound on the packing number into the information theoretic argument of Theorem 19 (see (7.10), where $M = |\mathscr{P}_{2r,R}|$), we obtain Theorem 20. □

7.8 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 poly$(1/\varepsilon)$ lower bounds of Theorems 19 and 20 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 $Z_\pi := \int \tilde{\pi}$ denote the normalizing constant, and let $\pi := \tilde{\pi}/Z$ denote the target distribution. Rejection sampling requires knowledge of an upper envelope $\tilde{\mu}$ for $\tilde{\pi}$, i.e., a function $\tilde{\mu}$ satisfying $\tilde{\mu} \ge \tilde{\pi}$ pointwise. The algorithm proceeds by repeatedly drawing samples from the density $\mu := \tilde{\mu}/Z_\mu$, where $Z_\mu := \int \tilde{\mu}$; each sample X is accepted with probability $\tilde{\pi}(X)/\tilde{\mu}(X)$.

It is standard to show [see, e.g., Che+22d] 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 Section 7.5) that $\log(1/\varepsilon)$ rates are attainable using rejection sampling. The reason why this does not contradict our lower bounds in Theorems 19 and 20 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

$$\max\Bigl\{\sup \ln \frac{\mu_0}{\pi},\ \sup \ln \frac{\pi}{\mu_0}\Bigr\} \leq M_0 < \infty\,.$$

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 9 (Fisher information guarantees via rejection sampling)**.** *It holds that*

$$\mathscr{C}_\infty(d, M_0, \varepsilon) \leq \widetilde{\mathcal{O}}\Bigl(\exp(3M_0) \log \frac{\sqrt{d}}{\varepsilon}\Bigr)\,.$$

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,

$$\tilde{\pi} = \frac{\tilde{\pi}}{\tilde{\pi}(0)} = \frac{\pi}{\pi(0)} \leq \frac{\exp(M_0)\,\mu_0}{\exp(-M_0)\,\mu_0(0)} = \underbrace{\frac{\exp(2M_0)}{\mu_0(0)}}_{\coloneqq Z_{\mu_0}}\,\mu_0\,.$$

This shows that $\tilde{\mu}_0 \coloneqq Z_{\mu_0}\,\mu_0$ is an upper envelope for $\tilde{\pi}$. Also, using $\pi(0) = 1/Z_\pi$,

$$\frac{Z_{\mu_0}}{Z_\pi} = \exp(2M_0)\,\frac{\pi(0)}{\mu_0(0)} \leq \exp(3M_0)\,.$$

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 $p = (1 - Z_\pi/Z_{\mu_0})^N \leq \exp(-N Z_\pi/Z_{\mu_0})$ is the probability of failure. We calculate

$$1 + \chi^2(\mu \,\|\, \pi) = \mathbb{E}_\mu\Bigl(\frac{\mu}{\pi}\Bigr) = 1 - p + p\,\mathbb{E}_\mu\Bigl(\frac{\mu_0}{\pi}\Bigr) \leq 1 + p\exp(M_0)\,.$$

Applying Lemma 23 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 20, 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 20, 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 10 (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*

$$\mathscr{C}_\infty\bigl(1, c\log(1/\varepsilon), \varepsilon\bigr) \gtrsim \frac{1}{\varepsilon\sqrt{\log(1/\varepsilon)}}\,.$$

Note also the following corollary.

Corollary 11 (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 $\sqrt{\mathsf{FI}(\mu \,\|\, \pi)} \le \varepsilon$ using N queries, where the query complexity satisfies*

$$N \le f(M_0)\,\mathrm{polylog}\bigl(\frac{1}{\varepsilon}\bigr)$$

for some increasing function $f : [1,\infty) \to \mathbb{R}_+$. Then, there is a universal *constant $c' > 0$ such that*

$$f(M_0) \geq \widetilde{\Omega}\bigl(\exp(c' M_0)\bigr)\,.$$

Proof. Using Corollary 10 with $M_0 = c\log(1/\varepsilon)$, we have

$$f\bigl(c\log\frac{1}{\varepsilon}\bigr)\operatorname{polylog}\bigl(\frac{1}{\varepsilon}\bigr) \geq N \gtrsim \frac{1}{\varepsilon\sqrt{\log(1/\varepsilon)}}\,,$$

or

$$f\bigl(c\log\frac{1}{\varepsilon}\bigr) \geq \frac{1}{\varepsilon\operatorname{polylog}(1/\varepsilon)}\,.$$

Writing this in terms of $M_0 = c\log(1/\varepsilon)$, or $\varepsilon = \exp(-M_0/c)$,

$$f(M_0) \geq \frac{\exp(M_0/c)}{(M_0/c)^{\mathcal{O}(1)}} = \widetilde{\Omega}\Bigl(\exp\bigl(\frac{M_0}{c}\bigr)\Bigr)$$

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 9).

The stronger oracle model is strictly stronger. We also observe the following consequence of these observations. On one hand, our lower bound in Thoerem 20 shows that

$$\mathscr{C}(1, K_0 = 1, \varepsilon) \geq \Omega\Bigl(\frac{1}{\varepsilon\sqrt{\log(1/\varepsilon)}}\Bigr)\,.$$

On the other hand, for constant M_0, rejection sampling (Proposition 9) yields

$$\mathscr{C}_\infty(1, M_0, \varepsilon) \leq \widetilde{\mathcal{O}}\Bigl(\exp(3M_0)\log\frac{1}{\varepsilon}\Bigr)\,.$$

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 $\mathcal{O}(\log(1/\varepsilon))$ queries. The algorithm involves uses a simple grid search.

Proposition 10 (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. $V(x) = \frac{1}{2}\,(\|x\| - R)_+^2$, *for* $\|x\| \ge R$.
3. V *is* 1*-smooth.*

Then, there is an algorithm which outputs μ with $\sqrt{\mathsf{FI}(\mu \,\|\, \pi)} \le \varepsilon$ 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 $\tilde\pi := \exp(-V)$, 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

$$\widehat{V}(x) := \begin{cases} \frac{1}{2}\,(\|x\| - R)_+^2\,, & \|x\| \ge R\,, \\ V(x_{\mathscr{N}}) + \langle \nabla V(x_{\mathscr{N}}), x - x_{\mathscr{N}}\rangle - \frac{1}{2}\,\|x - x_{\mathscr{N}}\|^2\,, & \|x\| < R\,. \end{cases}$$

By 1-smoothness of V, we have $V \ge \widehat{V}$, so that if we let $\tilde\mu_0 := \exp(-\widehat{V})$, then $\tilde\mu_0 \ge \tilde\pi$. Also, for $\|x\| < R$, we have the bound

$$\begin{aligned} \tilde\mu_0(x) &= \exp\Bigl(-V(x_{\mathscr{N}}) - \langle \nabla V(x_{\mathscr{N}}), x - x_{\mathscr{N}}\rangle + \frac{1}{2}\,\|x - x_{\mathscr{N}}\|^2\Bigr) \\ &\le \exp\bigl(-V(x) + \|x - x_{\mathscr{N}}\|^2\bigr) = \tilde\pi(x)\exp(\|x - x_{\mathscr{N}}\|^2) \le \exp(1)\,\tilde\pi(x)\,, \end{aligned}$$

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 9, one can show that for the output μ of rejection sampling, it holds that $\mathsf{FI}(\mu Q_t \,\|\, \pi) \le \varepsilon^2$ 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 $|\mathscr{N}| \le (cR)^d$ for a universal constant $c > 0$, it requires $\mathcal{O}((cR)^d)$ 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 $\mathcal{O}(\log(1/\varepsilon))$ complexity. However, if there is no a priori upper bound on the radius R, then the lower bounds of Theorem 20 and Corollary 10 apply.

7.9 Conclusion

In this chapter, we have provided the first lower bounds for the query complexity of obtaining Fisher information guarantees 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 19 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 19 and 20 do not capture the dependence of K_0, and this is also left for future work.

Chapter 8

Rejection sampling lower bounds

8.1 Introduction

We begin our general sampling lower bound investigations by looking at a toy example: discrete distributions supported on a finite alphabet. Formally, let p be a probability distribution on the set of integers $[N] := \{1, \dots, N\}$. Our algorithms will be given query access to $\tilde{p} := Zp$, with an unknown constant Z, and we wish to lower bound the number of queries needed to draw a sample that is close in total variation distance (TV) to p.

It is clear that if we impose no assumptions on the distribution, then the mass of the target can be concentrated on any point in the support, and hence the lower bound would be $\Omega(n)$, which will be tight [BP17]. Instead, we consider various classes of shape-constrained discrete distributions that exploit the ordering of the set $[N]$ (monotone, strictly unimodal, discrete log-concave). We also consider a class of distributions on the complete binary tree of size N, where only a partial ordering of the alphabet is required.

Despite the simplicity of the distribution class, we were unable to obtain lower bounds that hold for all possible sampling algorithms, so we restricted the class of algorithms to rejection sampling algorithms only. Rejection sampling, along with Monte Carlo simulation and importance sampling, can be traced back to the work of Stan Ulam and John von Neumann [Neu51; Eck87]. As the name suggests, rejection sampling is an algorithm which proposes candidate samples, which are then accepted with a probability carefully chosen to ensure that accepted samples have distribution p. Specifically, we seek to characterize the number of queries needed to obtain rejection sampling algorithms with constant acceptance probability (e.g. at least 1/2), uniformly over the classes of target distributions. This still gives the algorithm total freedom in how to construct its proposal distribution.

For the shape-constrained classes of discrete distributions, we show that the rejection sampling complexity scales sublinearly in the alphabet size N, and we provide rejection sampling algorithms with matching upper bounds that show our bounds are tight. The results can be compared with the literature on sublinear algorithms [Gol10; Gol17]. This body of work is largely focused on statistical questions such as estimation or testing and the present paper extends it in another statistical direction, namely sampling from a distribution known only up to normalizing constant, which is a standard step of Bayesian inference.

Our results in this chapter are fairly simple, but the key idea in the construction, that of hiding probability mass at different length scales away from the origin, turns out to be useful for the general sampling lower bound that we obtain in Chapter 9.

This chapter is based on the joint work [Che+22c], with Sinho Chewi, Patrik Gerber, Thibaut Le Gouic, and Philippe Rigollet.

8.2 Background on rejection sampling complexity

Classical setting with exact density queries

To illustrate the idea of rejection sampling, we first consider the classical setting where we can make queries to the *exact* target distribution p. Given a *proposal distribution* q and an upper bound M on the ratio $\max_{x\in[N]} p(x)/q(x)$, rejection sampling proceeds by drawing a sample $X \sim q$ and a uniform random variable $U \sim \mathsf{uniform}(0,1)$. If $U \leq p(X)/(Mq(X))$, the sample X is returned; otherwise, the whole process is repeated. Note that the rejection step is equivalent to flipping a biased coin: conditionally on X, the sample X is accepted with probability $p(X)/(Mq(X))$ and rejected otherwise. We refer to this procedure as *rejection sampling with acceptance probability* $p/(Mq)$.

It is easy to check that the output of this algorithm is indeed distributed according to p. Since Mq forms an upper bound on p, the region $G_q = \{(x,y) : x \in [N],\, y \in [0, Mq(x)]\}$ is a superset of $G_p = \{(x,y) : x \in [N],\, y \in [0, p(x)]\}$. Then, a uniformly random point from G_q conditioned on lying in G_p is in turn uniform on G_p, and so its x-coordinate has distribution p. A good rejection sampling scheme hinges on the design of a good proposal q that leads to few rejections.

If $q = p$, then the first sample X is accepted. More generally, the number of iterations required before a variable is accepted follows a geometric distribution with parameter $1/M$ (and thus has expectation M). In other words, the bound M characterizes the quality of the rejection sampling proposal q, and the task of designing an efficient rejection sampling algorithm is equivalent to determining a strategy for building the proposal q which guarantees a small value of the ratio M using few queries.

Density queries up to normalization

In this chapter, we instead work in the setting where we can only query the target distribution up to normalization, which is natural for Bayesian statistics, randomized algorithms, and online learning. Formally, let $\mathcal{P}$ be a class of probability distributions over a finite alphabet $\mathscr{X}$, and consider a target distribution $p \in \mathcal{P}$. We assume that the algorithm $\mathcal{A}$ has access to an oracle which, given $x \in \mathscr{X}$, outputs the value $Zp(x)$, where Z is an unknown constant. The value of Z does not change between queries. Equivalently, we can think of the oracle as returning the value $p(x)/p(x_0)$, where $x_0 \in \mathscr{X}$ is a fixed point with $p(x_0) > 0$.

To implement rejection sampling in this query model, the algorithm must construct an *upper envelope* for $\tilde{p}$, i.e., a function $\tilde{q}$ satisfying $\tilde{q} \ge \tilde{p}$. We can then normalize $\tilde{q}$ to obtain a probability distribution q. To draw new samples from p, we first draw samples $X \sim q$, which are then accepted with probability $\tilde{p}(X)/\tilde{q}(X)$. The following theorem shows that the well-known guarantees for rejection sampling also extend to our query model.

Theorem 24. *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 let Z_q denote 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

$$\begin{aligned}
\mathbb{P}(X \in A) &= \sum_{n=0}^{\infty} \mathbb{P}\Bigl(X_{n+1} \in A,\ U_i > \frac{\tilde{p}(X_i)}{\tilde{q}(X_i)}\,\forall\, i \in [n],\ U_{n+1} \le \frac{\tilde{p}(X_{n+1})}{\tilde{q}(X_{n+1})}\Bigr) \\
&= \sum_{n=0}^{\infty} \mathbb{P}\Bigl(X_{n+1} \in A,\ U_{n+1} \le \frac{\tilde{p}(X_{n+1})}{\tilde{q}(X_{n+1})}\Bigr)\, \mathbb{P}\Bigl(U_1 > \frac{\tilde{p}(X_1)}{\tilde{q}(X_1)}\Bigr)^n \\
&= \sum_{n=0}^{\infty} \Bigl(\int_A \frac{\tilde{p}(x)}{\tilde{q}(x)}\, q(\mathrm{d}x)\Bigr) \Bigl(\int_{\mathscr{X}} \bigl(1 - \frac{\tilde{p}(x)}{\tilde{q}(x)}\bigr)\, q(\mathrm{d}x)\Bigr)^n
\end{aligned}$$

$$= p(A)\,\frac{Z_p}{Z_q}\sum_{n=0}^{\infty}\big(1-\frac{Z_p}{Z_q}\big)^n = p(A)\,. \qquad \square$$

After n queries to the oracle for p (up to normalization), the output $\mathcal{A}(n,\tilde{p})$ of the algorithm is an upper envelope $\tilde{q} \geq \tilde{p}$, and in light of the above theorem it is natural to define the *ratio*

$$r(\mathcal{A}, n, \tilde{p}) := \frac{Z_q}{Z_p} = \frac{\sum_{x\in\mathscr{X}} \tilde{q}(x)}{\sum_{x\in\mathscr{X}} \tilde{p}(x)}\,.$$

The ratio achieved by the algorithm determines the expected number of queries to $\tilde{p}$ needed to generate each new additional sample from p.

As discussed in the introduction, our goal when designing a rejection sampling algorithm is to minimize this ratio uniformly over the choice of target $p \in \mathcal{P}$. We therefore define the rejection sampling complexity of the class $\mathcal{P}$ as follows.

Definition 5. *For a class of distributions $\mathcal{P}$,* the rejection sampling complexity of $\mathcal{P}$ *is the minimum number $n \in \mathbb{N}$ of queries needed, such that there exists and algorithm $\mathcal{A}$ that satisfies*

$$\sup_{\tilde{p}\in\tilde{\mathcal{P}}} r(\mathcal{A}, n, \tilde{p}) \leq 2\,,$$

where $\tilde{\mathcal{P}} := \{\tilde{p} = Zp \,:\, Z > 0\}$ is the set of all positive rescalings of distributions in $\mathcal{P}$.

The constant 2 in Definition 5 is arbitrary and could be replaced by any number strictly greater than 1, but we fix this choice at 2 for simplicity. With this choice of constant, and once the upper envelope is constructed, new samples from the target can be generated with a constant (≤ 2) expected number of queries per sample.

Note that when the alphabet $\mathscr{X}$ is finite and of size N, then N is a trivial upper bound for the complexity of $\mathcal{P}$, simply by querying all of the values of $\tilde{p}$ and then returning the exact upper envelope $\mathcal{A}(N,\tilde{p}) = \tilde{p}$. Therefore, for the discrete setting, our interest lies in exhibiting natural classes of distributions whose complexity scales *sublinearly* in N.

In this chapter, we specifically focus on *deterministic* algorithms $\mathcal{A}$. In fact, we believe that adding internal randomness to the algorithm does not significantly reduce the query complexity. Using Yao's minimax principle [Yao77], it seems likely that our lower bounds can extended to hold for randomized algorithms. We leave this extension for future work.

Table 8.1: Rejection sampling complexities for classes of discrete distributions. Here, N always denotes the alphabet size, $\mathscr{X} = \{1, \ldots, N\}$.

Class	Definition	Complexity	Theorem	Algorithm
monotone	Definition 6	$\Theta(\log N)$	Theorem 25	Algorithm 2
strictly unimodal	Definition 7	$\Theta(\log N)$	Theorem 26	Algorithm 3
cliff-like	Definition 8	$\Theta(\log\log N)$	Theorem 27	Algorithm 4
discrete log-concave	Definition 9	$\Theta(\log\log N)$	Theorem 28	Algorithm 4
monotone on a binary tree	Definition 10	$\Theta(N/\log N)$	Theorem 29	Algorithm 5

8.3 Results for shape-constrained discrete distributions

In order to improve on the trivial rate of $\mathcal{O}(N)$ on an alphabet of size N, we need to assume some structure of the target distributions. A well-known set of structural assumptions are shape constraints [GJ14; SS11], which have been extensively studied in the setting of estimation and inference. When the alphabet is $[N]$, shape constraints are built on top of the linear ordering of the support. We show that such assumptions indeed significantly reduce the complexity of the restricted classes of distributions to sublinear rates. We also consider the setting where the linear ordering of the support is relaxed to a partial ordering, and show it also results in sublinear complexity

Our complexity results for various classes of discrete distributions are summarized in Table 8.1. We define the various classes below, and give the sublinear complexity algorithms that construct the upper envelopes in Figure 8-1.

Structured distributions on a linearly ordered set

A natural class of discrete distributions which exploits the linear ordering of the set $[N]$ is the class of monotone distributions, defined below.

Definition 6. *The class of* monotone *distributions on* $[N]$ *is the class of probability distributions* p *on* $[N]$ *with* $p(1) \geq p(2) \geq p(3) \geq \cdots \geq p(N)$.

We show in Theorem 25 that the rejection sampling complexity of the class of monotone distributions is $\Theta(\log N)$, achieved via Algorithm 2. It is also straightforward to extend Algorithm 2 to handle the class of strictly unimodal distributions defined next (see Theorem 26 and Algorithm 3).

Definition 7. *The class of* strictly unimodal *distributions on* $[N]$ *is the class of probability distributions* p *on* $[N]$ *such that: there exists a point* $x \in [N]$ *with* $p(1) < p(2) < \cdots < p(x)$ *and* $p(x) > p(x+1) > \cdots > p(N)$.

Algorithm 2 Construct upper envelope for monotone distributions on $[N]$

```
Query the values $\tilde{p}(2^i)$, $0 \le i \le \lceil \log_2 N \rceil - 1$.
Construct the upper envelope $\tilde{q}$ as follows: set $\tilde{q}(1) := \tilde{p}(1)$, and
    $\tilde{q}(x) := \tilde{p}(2^i)\,, \qquad \text{for } x \in (2^i, 2^{i+1}]\,.$
```

Algorithm 3 Construct upper envelope for strictly unimodal distributions on $[N]$

```
Use binary search to find the mode of $\tilde{p}$.
Use Algorithm 2 to construct an upper envelope on each side of the mode.
```

Algorithm 4 Construct upper envelope for discrete log-concave distributions on $[N]$

```
Use binary search to find the first index $1 \le i \le \lceil \log_2 N \rceil$ such that
    $\tilde{p}(2^i) \le \tilde{p}(1)/2$, or else determine that $i$ does not exist.
If $i$ does not exist, output the constant upper envelope $\tilde{q} \equiv \tilde{p}(1)$.
Otherwise, output
    $\tilde{q}(x) := \begin{cases} \tilde{p}(1)\,, & x < 2^i\,, \\ \tilde{p}(2^i)\exp\bigl[-\frac{\log(\tilde{p}(1)/\tilde{p}(2^i))}{2^i - 1}\,(x - 2^i)\bigr]\,, & x \ge 2^i\,. \end{cases}$
```

Algorithm 5 Construct upper envelope for monotone distributions on binary trees of size $[N]$

```
Query $\tilde{p}(x)$ for all vertices $x$ which are at depth at most $\ell_0 := \ell - \lfloor \log_2 \ell \rfloor + 1$,
    where $\ell$ is the maximum depth of the tree.
Output
    $\tilde{q}(x) := \begin{cases} \tilde{p}(x)\,, & \text{if } \operatorname{depth}(x) \le \ell_0\,, \\ \tilde{p}(y)\,, & \text{if } \operatorname{depth}(x) > \ell_0\,,\ \operatorname{depth}(y) = \ell_0\,, \\ & \text{and } x \text{ is a descendant of } y\,. \end{cases}$
```

Figure 8-1: Algorithms for constructing rejection sampling upper envelopes which attain the minimax rates described in Table 8.1.

It is natural to ask whether further structural properties can yield even faster algorithms for sampling. This is indeed the case, and we start by illustrating this on a simple toy class of distributions.

Definition 8. *The class of* cliff-like *distributions on* $[N]$ *is the class of probability distributions* $\mathsf{uniform}([N_0])$ *for* $N_0 \in [N]$.

Since the class of cliff-like distributions is contained in the class of monotone distributions, Algorithm 2 yields a simple upper bound of $\mathcal{O}(\log N)$ for this class. However, we can do better by observing that in order to construct a good rejection sampling upper envelope for this class, we do not need to locate the index N_0 of the cliff exactly; it suffices to find it approximately, which in this context means finding an index N_0' such that $N_0' \le N_0 \le 2N_0'$. Since we only need to search over $\mathcal{O}(\log N)$ possible values for N_0', binary search can accomplish this using only $\mathcal{O}(\log\log N)$ queries. We prove in Theorem 27 that this rate is tight.

Remark 7. The class of cliff-like distributions provides a simple example of a class for which obtaining queries to the exact distribution is *not* equivalent to obtaining queries for the distribution up to a normalizing constant. Indeed, in the former model, the value of $p(1) = 1/N_0$ reveals the distribution in one query, implying a complexity of $\Theta(1)$, whereas we prove in Theorem 27 that the complexity under the second model is $\Theta(\log\log N)$.

Instead of formally describing the algorithm for sampling from cliff-like distributions, we generalize the algorithm to cover a larger class of structured distributions: the class of *discrete log-concave distributions* (see [SW14, §4]).

Definition 9. *The class of* discrete log-concave *distributions on* $[N]$ *is the class of probability distributions* p *on* $[N]$ *such that for all* $x \in \{2, \dots, N-1\}$, *we have* $p(x)^2 \ge p(x-1)p(x+1)$. *Equivalently it is the class of distributions* p *on* $[N]$ *for which there exists a convex function* $V : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ *such that* $p(x) = \exp(-V(x))$ *for all* $x \in [N]$. *In addition, we assume that the common mode of all of the distributions is at* 1.[1]

We prove in Theorem 28 that the rejection sampling complexity of discrete log-concave distributions is $\Theta(\log\log N)$, achieved by Algorithm 4 (note that this algorithm also applies for cliff-like distributions, since cliff-like distributions are discrete log-concave).

Remark 8. The class of discrete log-concave distributions is another case for which rejection sampling with exact density queries is much easier than with queries up to a normalizing constant. In the former model, [Dev87] requires only a single query to construct a rejection sampling upper envelope with ratio ≤ 5. In contrast, we show in Theorem 28 that the complexity under the second model is $\Theta(\log\log N)$.

[1] Without this condition, the class of discrete log-concave distributions includes the family of all Dirac measures on $[N]$, and the rejection sampling complexity is then trivially $\Theta(N)$.

Monotone on a binary tree

The previous examples of structured classes all rely on the linear ordering of $[N]$. We now show that it is possible to develop sublinear algorithms when the linear ordering is relaxed to a partial ordering. Specifically, we consider a structured class of distributions on balanced binary trees (note that the previously considered distributions can be viewed as distributions on a path graph).

Definition 10. *The class of* monotone distributions on a binary tree *with N vertices is the class of probability distributions p on a binary tree with N vertices, with maximum depth $\lceil \log_2(N+1) \rceil$, such that for every non-leaf vertex x with children x_1 and x_2, one has $p(x) \ge p(x_1) + p(x_2)$.*

We prove in Theorem 29 that the rejection sampling complexity of this class is $\Theta(N/\log N)$; the corresponding algorithm is given as Algorithm 5.

In a sense, Definition 10 reduces to the class of monotone distributions when the underlying graph is a path, since each vertex in the (rooted) path graph has one "child".The reader may wonder whether replacing the condition $p(x) \ge p(x_1) + p(x_2)$ with $p(x) \ge p(x_1) \vee p(x_2)$ is more natural. In Theorem 30, we show that rejection sampling cannot achieve sublinear complexity under the latter definition.

8.4 Proofs of the complexity bounds

We begin with a few general comments on the lower bounds. Recall that the rejection sampling task, given query access to the unnormalized distribution $\tilde{p}$, is to construct an upper envelope $\tilde{q} \ge \tilde{p}$ satisfying $Z_q \le 2Z_p$. We in fact prove lower bounds for an easier task, namely, the task of constructing a proposal distribution q such that $\|p/q\|_\infty \le 2$. Note that if we have an upper envelope $\tilde{q} \ge \tilde{p}$ with $Z_q \le 2Z_p$, then the corresponding normalized distribution q satisfies

$$\Big\|\frac{p}{q}\Big\|_\infty := \sup_{x \in \mathscr{X}} \frac{p(x)}{q(x)} = \sup_{x \in \mathscr{X}} \underbrace{\frac{\tilde{p}(x)}{\tilde{q}(x)}}_{\le 1} \underbrace{\frac{Z_q}{Z_p}}_{\le 2} \le 2\,, \tag{8.1}$$

so the latter task is indeed easier.

The proofs of the lower bounds are to an extent situational, but we outline here a fairly generic strategy that seems useful for many (but not all) classes of distributions. First, we fix a reference distribution $p^\star \in \mathcal{P}$ and assume that the algorithm has access to queries to an oracle for $p^\star$ (up to normalization). Also, suppose that the algorithm makes queries at the points $x_1, \dots, x_n$. Since we assume that the algorithm is deterministic, if $p \in \mathcal{P}$ is another distribution which agrees with $p^\star$ at the queries $x_1, \dots, x_n$ (up to normalization), then the algorithm produces the same output regardless of whether it is run on p or $p^\star$. In particular, the output q of the algorithm must satisfy both $\|p/q\|_\infty \le 2$ and $\|p^\star/q\|_\infty \le 2$.

More generally, for each $y \in [N]$ we can construct an adversarial perturbation $p_y \in \mathcal{P}$ of $p^\star$ which maximizes the probability of y, subject to being consistent with the queried values. Then the rejection sampling guarantee of the algorithm ensures that

$$2 \ge \Big\|\frac{p_y}{q}\Big\|_\infty \ge \frac{p_y(y)}{q(y)} = \frac{1}{q(y)} \sup\Big\{p(y) : p \in \mathcal{P},\ \frac{p(x_i)}{p(x_j)} = \frac{p^\star(x_i)}{p^\star(x_j)} \text{ for all } i, j \in [n]\Big\}\,.$$

Since q is a probability distribution, this yields the inequality

$$1 = \sum_{y\in[N]} q(y) \ge \frac{1}{2} \sum_{y \in [N]} \sup\Big\{p(y) : p \in \mathcal{P},\ \frac{p(x_i)}{p(x_j)} = \frac{p^\star(x_i)}{p^\star(x_j)} \text{ for all } i, j \in [n]\Big\}\,. \tag{8.2}$$

By analyzing this inequality for the various classes of interest, it is seen to furnish a lower bound on the number of queries n. Thus, the lower bound strategy consists of choosing a judicious reference distribution $p^\star$, constructing the adversarial perturbations p_y, and using the inequality (8.2) to produce a lower bound on n.

Monotone distributions

Theorem 25. *Let $\mathcal{P}$ be the class of monotone distributions supported on $[N]$, as given in Definition 6. Then the rejection sampling complexity of $\mathcal{P}$ is $\Theta(\log N)$.*

Upper bound

In the proof, let p denote the target distribution and assume that we can query the values of $\tilde p = pZ_p$. Also, by rounding N up to the nearest power of 2, and considering p to be supported on this larger alphabet, we can assume that N is a power of 2; this will not affect the complexity bound.

Proof. We construct the upper envelope $\tilde q$ as follows: first query the values of $\tilde p(2^i)$, $0 \le i \le \log_2 N - 1$, which requires $\mathcal{O}(\log N)$ queries; then $\tilde q$ is given as follows: set $\tilde q(1) := \tilde p(1)$ and

$$\tilde q(x) := \tilde p(2^i)\,, \qquad \text{for } x \in (2^i, 2^{i+1}]\,.$$

Note that $\tilde q$ is an upper envelope of $\tilde p$ because p is assumed to be monotone.

To complete the proof of the upper bound in Theorem 25, we just have to check that $Z_q/Z_p \le 2$. We use the definitions of the normalizing constants:

$$
\begin{aligned}
Z_p &= \tilde p(1) + \sum_{i=0}^{\log_2 N-1} \sum_{x=2^i+1}^{2^{i+1}} \tilde p(x) \ge \tilde p(1) + \sum_{i=0}^{\log_2 N-1} 2^i \tilde p(2^{i+1}) \\
&\ge \tilde p(1) + \frac{1}{2} \sum_{i=0}^{\log_2 N-2} \sum_{x=2^{i+1}+1}^{2^{i+2}} \tilde q(x) = \underbrace{\tilde p(1)}_{=(\tilde q(1)+\tilde q(2))/2} + \frac{1}{2}\sum_{x=3}^{N} \tilde q(x) = \frac{1}{2}\sum_{x=1}^{N} \tilde q(x) = \frac{1}{2} Z_q\,.
\end{aligned}
$$

The bound above shows that $Z_q/Z_p \le 2$, which concludes the proof. □

Lower bound

In this proof, we follow the lower bound strategy encapsulated in (8.2).

Proof. Let $x_1 < \ldots < x_n$ denote the queries; to simplify the proof, we will also assume that 1 and N are part of the queries. This can be interpreted as giving the algorithm two free queries, and the rest of the proof can be understood as a lower bound on the number of queries that the algorithm made, minus two.

We choose our reference distribution to be $p^\star(x) \propto 1/x$, i.e., we take

$$
p^\star(x) = \frac{c_N}{x \log N}\,, \qquad \text{for } x \in [N]\,,
$$

where c_N is used to normalize the distribution, and it satisfies $c_N \asymp 1$. To construct the adversarial perturbation p_y, suppose that y lies strictly between the queries x_i and x_{i+1}. Let $\alpha := (y - x_i)^{-1} \sum_{x_i < x \le y} p^\star(x)$ denote the average of $p^\star$ on $(x_i, y]$. Then, we define

$$
p_y(x) := \begin{cases} \alpha\,, & x_i < x \le y\,, \\ p^\star(x)\,, & \text{otherwise}\,. \end{cases}
$$

Since we replace the part of $p^\star$ on $(x_i, y]$ with its average value on this interval, then p_y is also a probability distribution:

$$
\sum_{x\in[N]} p_y(x) = \sum_{x\in[N]} p^\star(x) + \sum_{x_i < x \le y} \{\alpha - p^\star(x)\} = 1\,.
$$

Since $p^\star$ is decreasing, it is clear that p_y is too. Also, we can lower bound α via

$$
\alpha = \frac{1}{y - x_i} \sum_{x_i < x \le y} \frac{c_N}{x \log N} \ge \frac{c_N}{\log N} \frac{\log \frac{y+1}{x_i+1}}{y - x_i}\,.
$$

Since p_y agrees with the queries, we can substitute this into (8.2) to obtain

$$\frac{2\log N}{c_N} \ge \sum_{i=1}^{n-1} \sum_{x_i<y<x_{i+1}} \frac{\log\frac{y+1}{x_i+1}}{y-x_i}\,.$$

In what follows, let $\Delta_i := x_{i+1}/x_i$. We will only focus on the terms with $\Delta_i \ge 8$, so assume now that $\Delta_i \ge 8$. Let us evaluate the inner term via dyadic summation:

$$\begin{aligned}\sum_{x_i<y<x_{i+1}} \frac{\log\frac{y+1}{x_i+1}}{y-x_i} &= \sum_{0<y<x_{i+1}-x_i} \frac{\log(1+\frac{y}{x_i+1})}{y} \\ &\ge \sum_{0\le j\le \log_2\frac{x_{i+1}-x_i-1}{x_i+1}-1} \; \sum_{2^j(x_i+1)\le y<2^{j+1}(x_i+1)} \frac{\log(1+\frac{y}{x_i+1})}{y} \\ &\gtrsim \sum_{0\le j\le \log_2\frac{x_{i+1}-x_i-1}{x_i+1}-1} \; \sum_{2^j(x_i+1)\le y<2^{j+1}(x_i+1)} \frac{j}{2^{j+1}\,(x_i+1)} \\ &\gtrsim \sum_{0\le j\le \log_2(\Delta_i/4)} j \gtrsim (\log\Delta_i)^2\,.\end{aligned}$$

Let $A := \{i \in [n-1] : \Delta_i \ge 8\}$. Our calculations above yield

$$\log N \gtrsim \sum_{i\in A} (\log\Delta_i)^2\,.$$

Observe now that $\prod_{i=1}^{n-1}\Delta_i = N$ and $\prod_{i\in A^c}\Delta_i \le 8^{|A^c|} \le 8^n$, so that $\prod_{i\in A}\Delta_i \ge N/8^n$. Hence, applying the Cauchy-Schwarz inequality,

$$\log N \gtrsim \frac{1}{|A|}\Bigl|\sum_{i\in A}\log\Delta_i\Bigr|^2 \ge \frac{[(\log N - n\log 8)_+]^2}{|A|}\,.$$

We can now conclude as follows: either $n \ge (\log N)/(2\log 8)$, in which case we are done, or else $n \le (\log N)/(2\log 8)$. In the latter case, the above inequality can be rearranged to yield $n-1 \ge |A| \gtrsim \log N$, which proves the desired statement in this case as well. □

Strictly unimodal distributions

Theorem 26. *Let $\mathcal{P}$ be the class of strictly unimodal distributions supported on $[N]$, as given in Definition 7. Then the rejection sampling complexity of $\mathcal{P}$ is $\Theta(\log N)$.*

Upper bound

Proof. Since the strategy is very similar to the upper bound for the class of monotone distributions (Theorem 25), we briefly outline the procedure here. Using binary search, we can locate the mode of the distribution using $\mathcal{O}(\log N)$ queries. Once the mode is located, the strategy for constructing an upper envelope for monotone distributions can be employed on each side of the mode. □

Lower bound

Proof. We again refer to the class of monotone distributions (Theorem 25), for which the lower bound is given in 8.4. Essentially the same proof goes through for this setting as well, and we make two brief remarks on the modifications. First, the reference distribution $p^\star$ in that proof is also strictly unimodal. Second, although the adversarial perturbations p_y constructed in that proof are not strictly unimodal, they can be made strictly unimodal via infinitesimal perturbations, so it is clear that the proof continues to hold. □

Cliff-like distributions

Theorem 27. *Let $\mathcal{P}$ be the class of cliff-like distributions supported on $[N]$, as given in Definition 8. Then the rejection sampling complexity of $\mathcal{P}$ is $\Theta(\log\log N)$.*

Upper bound

Proof. Since the class of cliff-like distributions is contained in the class of discrete log-concave distributions, the upper bound for the former class is subsumed by Theorem 28 on the latter class. □

Lower bound

In this proof, we reduce the task of building a rejection sampling proposal q for the class of cliff-like distributions to the computational task of finding the cliff in an array. Formally, the latter task is defined as follows.

Task 1 (finding the cliff in an array)**.** There is an unknown array of the form $a = [1, \ldots, 1, 0, \ldots, 0]$, of size N. Let k be the largest index such that $a[i] = 1$. Given query access to the array, what is the minimum number of queries needed to determine the value of k?

The number of queries needed to solve Task 1 is $\Theta(\log N)$ (achieved via binary search). We now give the reduction.

Proof. Suppose that the algorithm makes queries to $\tilde{p}$. Let x_- be the largest query point with $\tilde{p}(x_-) > 0$, and let x_+ be the smallest query point with $\tilde{p}(x_+) = 0$. Given $x_- \le y < x_+$, the adversarial perturbation p_y is the uniform distribution on $[y]$. Substituting this into (8.2), and replacing ratios between $p^\star$ with ratios between $\tilde{p}$, we obtain

$$2 \ge \sum_{x_- \le y < x_+} p_y(y) = \sum_{x_- \le y < x_+} \frac{1}{y} \ge \log \frac{x_+}{x_-}\,.$$

Hence, an algorithm which can achieve the desired rejection sampling guarantee can guarantee that $x_+ \le c x_-$, where $c = e^2$ is a constant.

This reduces the lower bound for the rejection sampling complexity to the following question: what is the minimum number of queries to ensure that $x_+ \le c x_-$?

At this point we can reduce to Task 1. Suppose after n queries we can indeed ensure that $x_+ \le c x_-$. Consider an array a of size $\log_c N$, which has a cliff at index k. (We may round c up to the nearest integer, and N up to the nearest multiple of c in order to avoid ceilings and floors.) From this array we construct the unnormalized distribution $\tilde{p}$ on $[N]$ via

$$\tilde{p}(x) := \mathbb{1}\{x \le c^k\}\,, \qquad x \in [N]\,.$$

The rejection sampling algorithm provides us with $x_+ \le c x_-$ such that $\tilde{p}(x_-) = 1$ and $\tilde{p}(x_+) = 0$, i.e., $x_- \le c^k < x_+ \le c x_-$. Taking logarithms, we see that

$$\log_c x_- \le k < \log_c x_- + 1\,.$$

Hence, taking $\log_c x_-$ and rounding to the nearest integer (possibly doing a constant number of extra queries to the array afterwards for verification) locates the cliff k in n queries. Using the lower bound for Task 1, we see that $n = \Omega(\log \log N)$ as claimed. □

Discrete log-concave distributions

Theorem 28. *Let $\mathcal{P}$ be the class of discrete log-concave distributions on $[N]$, as in Definition 9, and recall that the modes of the distributions are assumed to be* 1*. Then the rejection sampling complexity of $\mathcal{P}$ is $\Theta(\log \log N)$.*

Upper bound

We make a few simplifying assumptions just as in the upper bound proof for Theorem 25. Let p denote the target distribution, assume that the queries are made to $\tilde{p} = p Z_p$, and let $V : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ be a convex function such that

$\tilde{p}(x) = \exp(-V(x))$ for $x \in [N]$. Also, we round N up to the nearest power of 2, which does not change the complexity bound.

Proof. First we make one query to obtain the value of $\tilde{p}(1)$. Then we find the integer $0 \le i_0 \le \log_2 N - 1$ (if it exists) such that

$$2\tilde{p}(2^{i_0}) \ge \tilde{p}(1)\,, \qquad 2\tilde{p}(2^{i_0+1}) \le \tilde{p}(1)\,.$$

To do this, observe that the values $\tilde{p}(2^i)$, $0 \le i \le \log_2 N$ are decreasing, and by performing binary search over these $\mathcal{O}(\log N)$ values we can find the integer i_0 or else conclude that it does not exist using $\mathcal{O}(\log\log N)$ queries.

If i_0 does not exist, then the target satisfies $2\tilde{p}(x) \ge \tilde{p}(1)$ for all $x \in [N]$, so the constant upper envelope $\tilde{q} = \tilde{p}(1)$ suffices.

If i_0 exists, denote $x_0 = 2^{i_0+1}$, and construct the upper envelope $\tilde{q}$ as follows: query $\tilde{p}(x_0)$, and let

$$\tilde{q}(x) = \begin{cases} \tilde{p}(1)\,, & x < x_0\,, \\ \tilde{p}(x_0)\, e^{-\lambda\,(x-x_0)}\,, & x \ge x_0\,, \end{cases}$$
$$\lambda = \frac{\log \frac{\tilde{p}(1)}{\tilde{p}(x_0)}}{x_0 - 1} = \frac{V(x_0) - V(1)}{x_0 - 1}\,.$$

We check that $\tilde{q}$ is a valid upper envelope of $\tilde{p}$. If we take logarithms and denote $V_q(x) = -\log \tilde{q}(x)$, then we see that

$$V_q(x) = \begin{cases} V(1)\,, & x < x_0\,, \\ V(x_0) + \lambda\,(x - x_0)\,, & x \ge x_0\,. \end{cases}$$

Because V is convex, we see that V_q is a lower bound of V, so $\tilde{q}$ is an upper bound of $\tilde{p}$.

To finish the proof, we just have to bound Z_q/Z_p. Let $Z_{q,1} = \sum_{x<x_0} \tilde{q}(x)$, and $Z_{q,2} = \sum_{x \ge x_0} \tilde{q}(x)$, so $Z_q = Z_{q,1} + Z_{q,2}$. We will bound these two terms separately. For the first term, by the definition of x_0 we can bound

$$Z_{q,1} = \sum_{x<x_0} \tilde{p}(1) \le 2 \sum_{x<x_0/2} \tilde{p}(1) \le 4 \sum_{x<x_0/2} \tilde{p}(x)\,.$$

For the second term,

$$Z_{q,2} \le \tilde{p}(x_0) \sum_{z=0}^{\infty} e^{-\lambda z} = \tilde{p}(x_0)\,(1 - e^{-\lambda})^{-1} = \tilde{p}(x_0)\, \Bigl(1 - \bigl(\frac{\tilde{p}(x_0)}{\tilde{p}(1)}\bigr)^{x_0 - 1}\Bigr)^{-1} \le 2\tilde{p}(x_0)\,.$$

Putting this together,

$$Z_q = Z_{q,1} + Z_{q,2} \le 4Z_p\,.$$

For clarity, we have presented the proof with the bound $Z_q/Z_p \le 4$. At the cost of more cumbersome proof, the above strategy can be modified to yield the guarantee $Z_q/Z_p \le 2$. □

Lower bound

Proof. Since the class of cliff-like distributions is contained in the class of discrete log-concave distributions, the lower bound for the latter class is subsumed by Theorem 27 on the former class. □

Monotone on a binary tree

Theorem 29. *Let $\mathcal{P}$ be the class of monotone distributions on a binary tree with N vertices, as in Definition 10. Then the rejection sampling complexity of $\mathcal{P}$ is $\Theta(N/(\log N))$.*

Let $\mathcal{T}$ denote the binary tree. For the upper bound, we may embed $\mathcal{T}$ into a slightly larger tree, and for the lower bound we can perform the construction on a slightly smaller tree. In this way, we may assume that $\mathcal{T}$ is a complete binary tree of depth ℓ, and hence $N = \sum_{j=0}^{\ell} 2^j = 2^{\ell+1} - 1$; this does not affect the complexity results. Throughout the proofs, we write $|x|$ for the depth of the vertex x in the tree, where the root is considered to be at depth 0.

Upper bound

Proof. Let c be a constant to be chosen later. The algorithm is to query the value of $\tilde{p}$ at all vertices at depth at most $\ell_0 := \ell - \log_2 \ell + c$. Then the upper envelope is constructed as follows,

$$\tilde{q}(x) := \begin{cases} \tilde{p}(x)\,, & \text{if } |x| \le \ell_0\,, \\ \tilde{p}(y)\,, & \text{if } |x| > \ell_0\,,\ |y| = \ell_0\,,\ \text{and } x \text{ is a descendant of } y\,. \end{cases}$$

Clearly $\tilde{q} \ge \tilde{p}$. Also, the number of queries we made is

$$\sum_{j=0}^{\ell_0} 2^j = 2^{\ell_0+1} - 1 \lesssim \frac{2^\ell}{\ell} \lesssim \frac{N}{\log N}\,.$$

Finally, we bound the ratio Z_q/Z_p. By definition,

$$Z_q = \sum_{x \in \mathfrak{T}} \tilde{q}(x) = \sum_{x \in \mathfrak{T},\, |x| \le \ell_0} \tilde{p}(x) + \sum_{x \in \mathfrak{T},\, |x| > \ell_0} \tilde{q}(x)\,.$$

For the second sum, we can write

$$\sum_{x \in \mathfrak{T},\, |x| > \ell_0} \tilde{q}(x) = \sum_{y \in \mathfrak{T},\, |y| = \ell_0} \tilde{p}(y)\,(2^{\ell - \ell_0 + 1} - 1) = \sum_{y \in \mathfrak{T},\, |y| = \ell_0} \tilde{p}(y)\,(2^{\log_2 \ell - c + 1} - 1)$$
$$\le \ell\, 2^{-c+1} \sum_{y \in \mathfrak{T},\, |y| = \ell_0} \tilde{p}(y)\,.$$

On the other hand, if x denotes any vertex, let x_1, x_2 denote its two children; then, for any level j,

$$\sum_{x \in \mathfrak{T},\, |x| = j+1} \tilde{p}(x) = \sum_{x \in \mathfrak{T},\, |x| = j} \{\tilde{p}(x_1) + \tilde{p}(x_2)\} \le \sum_{x \in \mathfrak{T},\, |x| = j} \tilde{p}(x)\,.$$

Hence,

$$\sum_{x \in \mathfrak{T},\, |x| \le \ell_0} \tilde{p}(x) \ge (\ell_0 + 1) \sum_{x \in \mathfrak{T},\, |x| = \ell_0} \tilde{p}(x)$$

which yields

$$Z_q \le \bigl(1 + \frac{\ell\, 2^{-c+1}}{\ell_0 + 1}\bigr) \sum_{x \in \mathfrak{T},\, |x| \le \ell_0} \tilde{p}(x) \le 2 Z_p\,,$$

if ℓ and c are sufficiently large. □

Lower bound

The proof of the lower bound follows the strategy encapsulated in (8.2).

Proof. Suppose that an algorithm achieves rejection sampling ratio 2 with n queries. Again let $\ell_0 := \ell - \log_2 \ell + c$, where the constant c will possibly be different from the one in the upper bound. The reference distribution will be

$$\tilde{p}(x) := \begin{cases} 2^{-|x|}\,, & |x| \le \ell_0\,, \\ 0\,, & |x| > \ell_0\,. \end{cases}$$

Note that $p \in \mathcal{P}$. The normalizing constant is $Z_p = \ell_0 + 1$, since there are 2^j vertices at level j. For each $|y| > \ell_0$, we will create a perturbation distribution

p_y in the following way:

$$\tilde{p}_y(x) := \begin{cases} 2^{-|x|}, & |x| \le \ell_0 , \\ 2^{-\ell_0}, & |x| > \ell_0 \text{ and } y \text{ is a descendant of } x \text{ (or equal to } x) , \\ 0, & \text{otherwise} . \end{cases}$$

Thus, $\tilde{p}_y$ places extra mass on the path leading to y; note also that $p_y \in \mathcal{P}$. The normalizing constant for p_y is

$$Z_{p_y} = Z_p + \sum_{j=\ell_0+1}^{|y|} 2^{-\ell_0} \le \ell_0 + 1 + (\ell - \ell_0)\, 2^{-\ell_0} = \ell_0 \, \{1 + o(1)\} ,$$

where $o(1)$ tends to 0 as $\ell \to \infty$.

Next, let $\mathcal{Q}$ denote the set of vertices x at level ℓ_0 for which at least one of the descendants of x (not including x itself) is queried by the algorithm, and let $\mathcal{Q}^{\mathsf{c}}$ denote the vertices at level ℓ_0 which do not belong to $\mathcal{Q}$. Note if $x \in \mathcal{Q}^{\mathsf{c}}$ and y is a descendant of x, then p_y is consistent with the queries made by the algorithm. Let $\mathcal{D}(x)$ denote the descendants of x. Now, applying (8.2) with $p^\star = p$,

$$2 \ge \sum_{x \in \mathcal{T},\, |x| \le \ell_0} p(x) + \sum_{x \in \mathcal{Q}^{\mathsf{c}}} \sum_{y \in \mathcal{D}(x)} p_y(y) = 1 + \sum_{x \in \mathcal{Q}^{\mathsf{c}}} \sum_{y \in \mathcal{D}(x)} p_y(y)$$

which yields

$$1 \ge \sum_{x \in \mathcal{Q}^{\mathsf{c}}} \sum_{y \in \mathcal{D}(x)} p_y(y) \ge \sum_{x \in \mathcal{Q}^{\mathsf{c}}} \frac{2^{-\ell_0}}{\ell_0\,(1 + o(1))}\,(2^{\ell - \ell_0 + 1} - 2) \gtrsim \frac{2^{\ell - 2\ell_0 + 1}}{\ell_0\,(1 + o(1))}\,\{2^{\ell_0} - |\mathcal{Q}|\} .$$

It then yields

$$\begin{aligned} n \ge |\mathcal{Q}| &\gtrsim 2^{\ell_0} - \frac{\ell_0\,(1 + o(1))}{2^{\ell - 2\ell_0 + 1}} = 2^{\ell_0} \Bigl(1 - \frac{\ell_0\,(1 + o(1))}{2^{\ell - \ell_0 + 1}}\Bigr) \\ &= 2^{\ell_0} \Bigl(1 - \frac{\ell_0\,(1 + o(1))}{2^{\log_2 \ell - c + 1}}\Bigr) = 2^{\ell_0} \Bigl(1 - \frac{\ell_0\,(1 + o(1))}{\ell\, 2^{-c+1}}\Bigr) . \end{aligned}$$

If we now choose $c \ll 0$ to be a *negative* constant, we can verify

$$n \gtrsim 2^{\ell_0} = 2^{\ell - \log_2 \ell + c} \gtrsim \frac{N}{\log N} ,$$

completing the proof. □

An alternate definition of monotone

In this section, we show that if we adopt an alternative definition of monotone on a binary tree, then the rejection sampling complexity is trivial.

Theorem 30. *Let $\mathcal{P}$ be the class of probability distributions p on a binary tree with N vertices, with maximum depth $\lceil \log_2(N+1)\rceil$, such that if for every non-leaf vertex x, if the children of x are x_1 and x_2, then $p(x) \ge p(x_1) \vee p(x_2)$. Then, the rejection sampling complexity of $\mathcal{P}$ is $\Theta(N)$.*

Proof. It suffices to show the lower bound, and the proof will be similar to the one in Section 8.4. We may assume that the binary tree is a complete binary tree with depth ℓ. Suppose that an algorithm achieves a rejection sampling ratio 2 after n queries. We define the reference distribution $p^\star$ via

$$\tilde{p}^\star(x) := \begin{cases} 1\,, & |x| \le \ell - 2\,, \\ 0\,, & |x| > \ell - 2\,. \end{cases}$$

The normalizing constant is $Z_{p^\star} = 2^{\ell-1} - 1$. For each leaf vertex y, we define the perturbation distribution p_y via

$$\tilde{p}_y(x) := \begin{cases} 1\,, & |x| \le \ell - 2 \text{ or } x \text{ is an ancestor of } y \text{ (including if } x = y)\,, \\ 0\,, & |x| > \ell - 2\,. \end{cases}$$

The normalizing constant of p_y is $Z_{p_y} = 2^{\ell-1} + 1$.

Let $\mathcal{Q}$ denote the set of leaf vertices which are queried by the algorithm, and let $\mathcal{Q}^{\mathsf{c}}$ denote the set of leaf vertices not in $\mathcal{Q}$. Then, from (8.2),

$$2 \ge \sum_{x \in \mathcal{T},\, |x| \le \ell-2} p^\star(x) + \sum_{y \in \mathcal{Q}^{\mathsf{c}}} p_y(y) = 1 + \sum_{y \in \mathcal{Q}^{\mathsf{c}}} p_y(y)$$

and rearranging this yields

$$1 \ge \sum_{y \in \mathcal{Q}^{\mathsf{c}}} \frac{1}{2^{\ell-1}+1} = \frac{1}{2^{\ell-1}+1}\,\{2^\ell - |\mathcal{Q}|\}\,.$$

This is further rearranged to yield

$$n \ge |\mathcal{Q}| \ge 2^{\ell-1}\,\bigl(2 - 1 - \frac{1}{2^{\ell-1}}\bigr) \gtrsim 2^\ell = N\,,$$

where the last inequality holds if $\ell > 1$. □

8.5 Conclusion

We studied the query complexity of rejection sampling within a minimax framework, and we showed that for various natural classes of discrete distributions, rejection sampling can obtain exact samples with an expected number of queries which is sublinear in the size of the support of the distribution. Our algorithms can also be run in sublinear time, which make them substantially faster than the baseline of multinomial sampling.

A natural direction for future work is to investigate the complexity of rejection sampling on other structured classes of distributions, such as distributions on graphs, or distributions on continuous spaces. In many of these other settings, the complexity of algorithms based on Markov chains has been studied extensively, but the complexity of rejection sampling remains to be understood.

For the goal of general sampling lower bounds, the results in this chapter suggests the following insight: we want to construct distributions that are hard to distinguish from queries, but whose mass are supported at different places. The clearest example of this idea is our lower bound for the cliff-like distributions in Seciton 8.4.

Chapter 9

General lower bound in dimension one

9.1 Introduction

In this chapter, we build on the ideas from Chapter 8, and prove the first general sampling lower bound for log-concave distributions in one dimension. Specifically, the class of target distributions will be the strongly log-concave and log-smooth distributions on $\mathbb{R}$, i.e., the class of distributions with a density $p \propto \exp(-V)$, where the potential $V : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable, α-strongly convex, and β-smooth. The relevant parameter of this class is the *condition number* $\kappa \coloneqq \beta/\alpha$, 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 p.

This chapter is based on the joint work [Che+22d], with Sinho Chewi, Patrik Gerber, Thibaut Le Gouic, and Philippe Rigollet.

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 [Dal17c; DMM19], as mentioned in Chapted 6 it is unclear how to use lower bound techniques for optimization [Nes13] 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.

Other prior work on sampling lower bounds has fallen largely into one of several categories. One line of work studies lower bounds against a specific class of algorithm such as underdamped Langevin [CLW21] or MALA [Che+21;

LST21a; WSC22]. However, these lower bounds techniques are tailored to the restricted class of algorithms that they consider and are not suitable for proving general query lower bounds. Another line of work considers lower bounds against computing normalizing constants [RV08; GLL20]. The work [Tal19] also investigates the computational complexity of sampling.

We mention two further lower bounds in different settings. The work of [CBL22] proves a lower bound against stochastic gradient oracles, and the work of [GLL22] proves a lower bound on the number of individual function value (i.e., zeroth-order) queries needed to sample from a density of the form $\exp(-\sum_{i\in I} f_i + \mu \|\cdot\|^2)$, where each f_i is convex, Lipschitz, and whose domain is the unit ball. In contrast, we consider deterministic, first-order oracle access. Moreover, their considerations are somewhat orthogonal to ours: [CBL22] focuses more on the role of noise, whereas we consider exact gradient access; and the lower bound of [GLL22] applies a direct reduction from optimization, which is also not in the spirit of the present work (in particular, we explicitly set the mode of the target distribution to zero).

A lower complexity bound in one dimension. Recall that for convex optimization, there are two relevant regimes [see, e.g., Bub15]: (1) the low-dimension regime, in which algorithms such as the cutting plane method achieve the rate $\mathcal{O}(d\log(1/\varepsilon))$ (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 chapter, 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 p (uniformly over p 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 Stan Ulam and John von Neumann [Neu51; Eck87]. We develop an algorithm which uses $\mathcal{O}(\log\log\kappa)$ queries in order to build a proposal distribution. Once the proposal distribution is constructed, new samples which are ε-close to p in total variation distance can be generated using $\mathcal{O}(\log(1/\varepsilon))$ additional queries per sample.

Although our algorithm is tailored to distributions in one dimension, the task of sampling from a one-dimensional log-concave distribution is an important subroutine for higher dimensional algorithms, such as the Hit-and-Run algorithm. We describe the application of our algorithm to Hit-and-Run in Section **??**.

9.2 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 chapter.

Definition 11. *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* p *supported on* $\mathbb{R}$*, whose density is of the form* $p(x) = \exp(-V(x))$*, for a potential function* $V : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ *which is twice continuously differentiable and satisfies*

$$\alpha \le V''(x) \le \beta\,, \qquad \forall x \in \mathbb{R}\,. \tag{9.1}$$

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 $p = \exp(-V)$. 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.

We now state our lower bound.

Theorem 31. *Consider the class* $\mathcal{P}$ *of univariate* α*-strongly log-concave and* β*-log-smooth distributions as defined in Definition 11, and let* $\kappa := \beta/\alpha$ *denote the*

[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.

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 p, and outputs a random variable whose law is at most $\frac{1}{64}$ away from p 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 Section 9.4. The strategy is to construct a family of distributions $\{p_1 \ldots, p_m\}$ 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 p_i, it can also *identify* p_i. We construct the family $\{p_i\}_{i=1}^m$ in such a way that identifying p_i from queries to low-order 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 $\{p_i\}_{i=1}^m$. Suppose that we have a distribution $p \propto \exp(-V)$ which is rescaled to satisfy $1 \le V'' \le \kappa$. The bound $V'' \ge 1$ implies that a substantial fraction of the mass of p is supported on the interval $[-1, 1]$. On the other hand, the bound $V'' \le \kappa$ allows for the density p 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 p_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 $\{p_i\}_{i=1}^m$ of $m \asymp \sqrt{\kappa}$ distributions, where p_i is the uniform distribution on $[-i/\sqrt{\kappa}, i/\sqrt{\kappa}]$; see Figure 9-1.

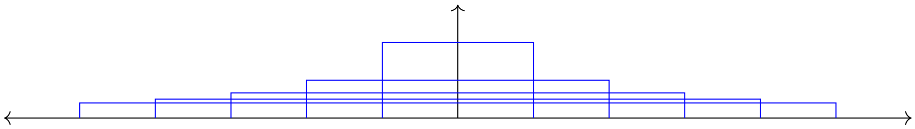

Figure 9-1: A family of uniform distributions.

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 p_i and p_j to be appreciable, we require $j \ge 2i$. This motivates us to consider the subfamily $\{p_{2^i}, 1 \le i \le \log_2 \sqrt{\kappa}\}$, of which there are $\mathcal{O}(\log\kappa)$ elements. For this subfamily, we can hope to reduce the task of sampling to that of identifying p_{2^i} via queries, and binary search for this problem requires only $\mathcal{O}(\log\log\kappa)$ 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 the appendix.

9.3 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 p, 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 6 Envelope

```
Use binary search to find the first index $i_+ \in \{0, 1, \dots, \lceil \frac{1}{2}\log_2\kappa \rceil\}$ with $V(2^{i_+}/\sqrt{\kappa}) \ge \frac{1}{2}$.
Use binary search to find the first index $i_- \in \{0, 1, \dots, \lceil \frac{1}{2}\log_2\kappa \rceil\}$ with $V(-2^{i_-}/\sqrt{\kappa}) \ge \frac{1}{2}$.
Set $x_- \coloneqq -2^{i_-}/\sqrt{\kappa}$ and $x_+ \coloneqq 2^{i_+}/\sqrt{\kappa}$.
return
```

$$\tilde{q}(x) \coloneqq \begin{cases} \exp\Big[-\dfrac{x-x_-}{2x_-} - \dfrac{(x-x_-)^2}{2}\Big]\,, & x \le x_-\,, \\ 1\,, & x_- \le x \le x_+\,, \\ \exp\Big[-\dfrac{x-x_+}{2x_+} - \dfrac{(x-x_+)^2}{2}\Big]\,, & x \ge x_+\,. \end{cases}$$

We give the high-level pseudocode for building an upper envelope in Algorithm 6, and for generating new samples in Algorithm 7. 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 Sample

```
Normalize $\tilde{q}$ to form $q$.
**while** sample is not accepted **do**
    Sample $X \sim q$.
    Accept $X$ w.p. $\tilde{p}(X)/\tilde{q}(X)$.
**end while**
**return** $X$
```

Before implementing Algorithm 6, we first perform several preprocessing steps. Recall that the mode of the distribution p is assumed to be at 0, and that $p \propto \exp(-V)$. We also assume that $1 \le V'' \le \kappa$. To reduce to this case, say we start with $\alpha \le V'' \le \beta$, and the bounds α, β are known. Then, observe that the rescaled potential $\bar{V}(x) \coloneqq V(x/\sqrt{\alpha})$ satisfies $1 \le \bar{V}'' \le \kappa = \beta/\alpha$. 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 $\bar{p} \propto \exp(-\bar{V})$; it can be checked that $\bar{X}/\sqrt{\alpha}$ is a sample from p. 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 6 requires performing binary search over an array of size $\mathcal{O}(\log \kappa)$, which requires only $\mathcal{O}(\log\log\kappa)$ queries; similar comments apply to the second step. We prove in Section 9.5 that the indices i_- and i_+ always exist under our assumptions. We prove in Section 9.5 that the output $\tilde{q}$ of Algorithm 6 is an *upper envelope* for the oracle, i.e., $\tilde{q} \ge \exp(-V)$. The upper envelope $\tilde{q}$ constructed in Algorithm 6 is the input to Algorithm 7; see Figure 9-2.

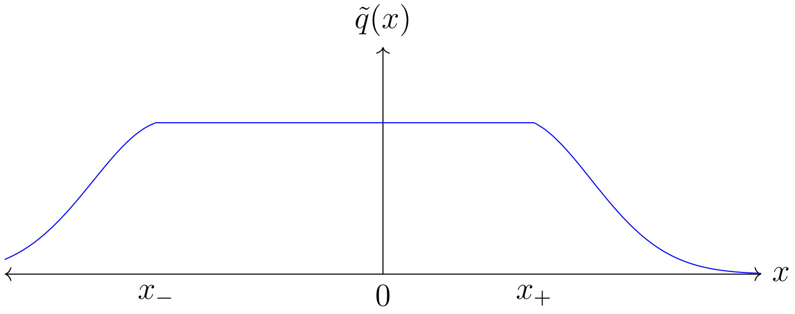

Figure 9-2: The upper envelope $\tilde{q}$ constructed in Algorithm 6.

In Algorithm 7, 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 chapter. 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 p, 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 p 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 32. *Suppose that the target distribution p belongs to the class of univariate strongly log-concave and log-smooth distributions (Definition 11). Algorithm 6 uses $\mathcal{O}(\log\log\kappa)$ 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 returns an exact sample from p after an additional $\mathcal{O}(1)$ expected queries to the oracle.*

2. *(approximate sampling) Fix an accuracy parameter $0 < \varepsilon < 1$. If we limit Algorithm 7 to use at most $\mathcal{O}(\log(1/\varepsilon))$ queries, then the output of Algorithm 7 (or 'FAILURE', if Algorithm 7 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 p.*

We give the proof in Section 9.5.

9.4 Proof of the lower bound

The construction

Let m be the largest integer such that

$$\exp\bigl(-\frac{2^{2m-2}}{2\kappa}\bigr) \geq \frac{1}{2}\,. \tag{9.2}$$

Define two auxiliary functions

$$\phi(x) := \begin{cases} \kappa\,, & 1/2 \leq x < 1\,, \\ 1\,, & 1 \leq x < 2\,, \\ \kappa\,, & 2 \leq x < 5/2\,, \\ 0 & \text{otherwise}\,, \end{cases} \qquad \psi(x) := \begin{cases} 1\,, & 5/2 \leq x < 4\,, \\ \kappa\,, & 4 \leq x < 5\,, \\ 0\,, & \text{otherwise}\,. \end{cases}$$

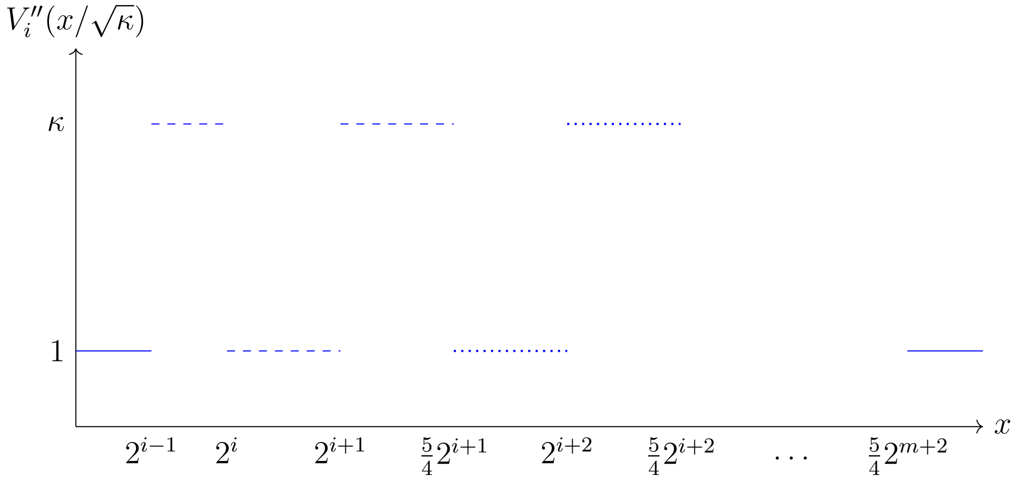

Figure 9-3: The dashed lines correspond to ϕ and the dotted lines correspond to ψ.

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 $\mathbb{R}_+$. The second derivative is given by

$$V_i''(x) \coloneqq \mathbb{1}\{x \leq \kappa^{-\frac{1}{2}}2^{i-1}\} + \phi\Big(\frac{x}{\kappa^{-\frac{1}{2}}2^i}\Big) + \sum_{j=i}^{m-1} \psi\Big(\frac{x}{\kappa^{-\frac{1}{2}}2^j}\Big) + \mathbb{1}\{x \geq 5\kappa^{-\frac{1}{2}}2^{m-1}\}\,, \; x \geq 0\,.$$

Observe that all of the terms in the above summation have disjoint supports, see Figure 9-3.

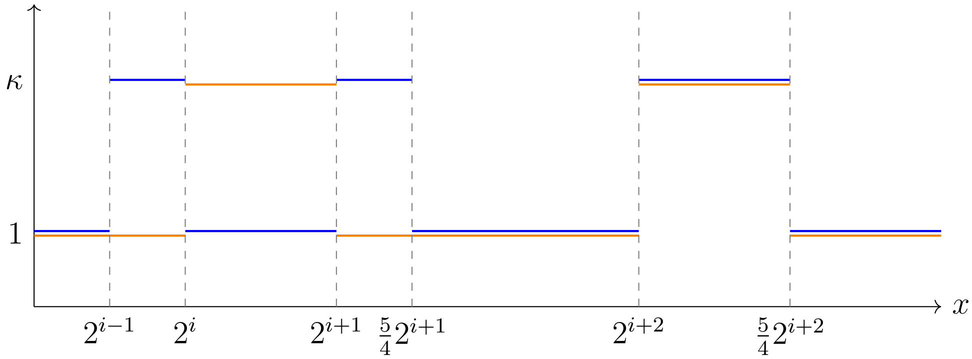

Figure 9-4: We plot V_i'' (in blue) and V_{i+1}'' (in orange). In this figure, we do not distort the horizontal axis lengths to make it easier to visually compare the relative lengths of intervals on which the second derivatives are constant.

The following lemma provides intuition for the construction.

Lemma 34. *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 9-4 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 $x \ge 0$,

$$\begin{aligned} V_{i+1}''(x) - V_i''(x) &= \mathbb{1}\{\kappa^{-\frac{1}{2}}2^{i-1} < x \le \kappa^{-\frac{1}{2}}2^{i}\} - \phi\big(\frac{x}{\kappa^{-\frac{1}{2}}2^{i}}\big) + \phi\big(\frac{x}{\kappa^{-\frac{1}{2}}2^{i+1}}\big) - \psi\big(\frac{x}{\kappa^{-\frac{1}{2}}2^{i}}\big) \\ &= \begin{cases} -(\kappa-1)\,, & \kappa^{-\frac{1}{2}}2^{i-1} \le x \le \kappa^{-\frac{1}{2}}2^{i}\,, \\ +(\kappa-1)\,, & \kappa^{-\frac{1}{2}}2^{i} \le x \le \kappa^{-\frac{1}{2}}2^{i+1}\,, \\ -(\kappa-1)\,, & \kappa^{-\frac{1}{2}}2^{i+1} \le x \le \frac{5}{4}\kappa^{-\frac{1}{2}}2^{i+1}\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \end{aligned}$$

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

$$\begin{aligned} V_{i+1}\big(\frac{5}{4}\kappa^{-\frac{1}{2}}2^{i+1}\big) - V_i\big(\frac{5}{4}\kappa^{-\frac{1}{2}}2^{i+1}\big) = {} & \underbrace{-\frac{\kappa-1}{2}\,\big(\kappa^{-\frac{1}{2}}2^{i-1}\big)^2}_{\text{integral on } [\kappa^{-\frac{1}{2}}2^{i-1},\kappa^{-\frac{1}{2}}2^{i}]} \\ & \underbrace{-(\kappa-1)\kappa^{-\frac{1}{2}}2^{i-1}\,\kappa^{-\frac{1}{2}}2^{i} + \frac{\kappa-1}{2}\,\big(\kappa^{-\frac{1}{2}}2^{i}\big)^2}_{\text{integral on } [\kappa^{-\frac{1}{2}}2^{i},\kappa^{-\frac{1}{2}}2^{i+1}]} \\ & \underbrace{+(\kappa-1)\kappa^{-\frac{1}{2}}2^{i-1}\,\frac{1}{4}\kappa^{-\frac{1}{2}}2^{i+1} - \frac{\kappa-1}{2}\,\big(\frac{1}{4}\kappa^{-\frac{1}{2}}2^{i+1}\big)^2}_{\text{integral on } [\kappa^{-\frac{1}{2}}2^{i+1},\frac{5}{4}\kappa^{-\frac{1}{2}}2^{i+1}]} \\ = {} & \frac{\kappa-1}{\kappa}\,\{-2^{2i-3} - 2^{2i-1} + 2^{2i-1} + 2^{2i-2} - 2^{2i-3}\} \\ = {} & 0\,, \end{aligned}$$

as desired. □

We also need a lemma showing that each probability distribution $p_i \propto \exp(-V_i)$ places a substantial amount of mass on the interval $(\kappa^{-\frac{1}{2}}2^{i-2}, \kappa^{-\frac{1}{2}}2^{i-1}]$.

Lemma 35. *For each $i \in [m]$,*

$$p_i\big((\kappa^{-\frac{1}{2}}2^{i-2}, \kappa^{-\frac{1}{2}}2^{i-1}]\big) \ge \frac{1}{32}\,.$$

Proof. According to the definition of p_i, we have

$$p_i\big((\kappa^{-\frac{1}{2}}2^{i-2}, \kappa^{-\frac{1}{2}}2^{i-1}]\big) = \frac{\int_{\kappa^{-\frac{1}{2}}2^{i-2}}^{\kappa^{-\frac{1}{2}}2^{i-1}} \exp(-x^2/2)\,\mathrm{d}x}{Z_{p_i}}\,, \qquad Z_{p_i} := \int_{\mathbb{R}} \exp(-V_i)\,.$$

Recalling that m is chosen so that $\exp(-x^2/2) \ge 1/2$ whenever $|x| \le \kappa^{-\frac{1}{2}}2^{m-1}$ (see (9.2)), we can conclude that

$$\int_{\kappa^{-\frac{1}{2}}2^{i-2}}^{\kappa^{-\frac{1}{2}}2^{i-1}} \exp\big(-\frac{x^2}{2}\big)\,\mathrm{d}x \ge \frac{1}{2}\,\kappa^{-\frac{1}{2}}2^{i-2}\,.$$

For the normalizing constant, observe that

$$\int_0^\infty \exp(-V_i) = \int_0^{\kappa^{-\frac{1}{2}}2^i} \exp(-V_i) + \int_{\kappa^{-\frac{1}{2}}2^i}^\infty \exp(-V_i) \le \kappa^{-\frac{1}{2}}2^i + \int_{\kappa^{-\frac{1}{2}}2^i}^\infty \exp(-V_i)\,.$$

Since $V_i'' = \kappa$ on $[\kappa^{-\frac{1}{2}}2^{i-1}, \kappa^{-\frac{1}{2}}2^i]$, it follows that $V_i'(\kappa^{-\frac{1}{2}}2^i) \ge \kappa^{\frac{1}{2}}2^{i-1}$, and so

$$V_i(x) \ge \kappa^{\frac{1}{2}}2^{i-1}\,(x - \kappa^{-\frac{1}{2}}2^i) + \frac{(x - \kappa^{-\frac{1}{2}}2^i)^2}{2}\,, \qquad x \ge \kappa^{-\frac{1}{2}}2^i\,.$$

Therefore,

$$\begin{aligned}\int_{\kappa^{-\frac{1}{2}}2^i}^\infty \exp(-V_i) &\le \int_{\kappa^{-\frac{1}{2}}2^i}^\infty \exp\Big(-\kappa^{\frac{1}{2}}2^{i-1}\,(x - \kappa^{-\frac{1}{2}}2^i) - \frac{(x - \kappa^{-\frac{1}{2}}2^i)^2}{2}\Big)\,\mathrm{d}x \\ &\le \frac{1}{\kappa^{\frac{1}{2}}2^{i-1}} \le \frac{1}{\sqrt{\kappa}}\,,\end{aligned}$$

where we applied a standard tail estimate for Gaussian densities (Lemma 36). Putting it together,

$$p_i\big((\kappa^{-\frac{1}{2}}2^{i-2}, \kappa^{-\frac{1}{2}}2^{i-1}]\big) \ge \frac{2^{i-3}}{2\,(2^i + 1)} \ge \frac{1}{32}\,,$$

which proves the result. □

Lower bound via Fano's inequality

In this section, we use the densities $\{p_i\}_{i=1}^m$ constructed in the previous section together with Fano's inequality from information theory in order to prove the lower bound.

Proof of Theorem 31. Let $Z \sim \mathsf{uniform}([m])$ be an index chosen uniformly at random. Suppose that an algorithm makes n queries to the oracle for p_Z, and given $Z = i$, outputs a sample Y whose law q_i is at total variation distance at most $\frac{1}{64}$ from p_i. In light of Lemma 35, a good candidate estimator for Z from the observation of Y is given by

$$\widehat{Z} \coloneqq \{k \in \mathbb{N} : Y \in (\kappa^{-\frac{1}{2}} 2^{k-2}, \kappa^{-\frac{1}{2}} 2^{k-1}]\} .$$

On the one hand, the probability that the estimator is correct is bounded by

$$\begin{aligned}
\mathbb{P}\{\widehat{Z} = Z\} &= \frac{1}{m} \sum_{i=1}^m \mathbb{P}\{\widehat{Z} = i \mid Z = i\} \\
&= \frac{1}{m} \sum_{i=1}^m \mathbb{P}\{Y \in (\kappa^{-\frac{1}{2}} 2^{i-2}, \kappa^{-\frac{1}{2}} 2^{i-1}] \mid Z = i\} \\
&= \frac{1}{m} \sum_{i=1}^m q_i\big((\kappa^{-\frac{1}{2}} 2^{i-2}, \kappa^{-\frac{1}{2}} 2^{i-1}]\big) \\
&\ge \frac{1}{m} \sum_{i=1}^m p_i\big((\kappa^{-\frac{1}{2}} 2^{i-2}, \kappa^{-\frac{1}{2}} 2^{i-1}]\big) - \frac{1}{64} \\
&\ge \frac{1}{64} ,
\end{aligned}$$

where the last inequality uses Lemma 35.

On the other hand, we can lower bound $\mathbb{P}\{\widehat{Z} \ne Z\}$ using Fano's inequality. Let $x_1, \dots, x_n$ 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

$$\mathbb{P}\{\widehat{Z} \ne Z\} \ge 1 - \frac{I(\{x_j, W_Z(x_j)\}_{j \in [n]}; Z) + \log 2}{\log m} ,$$

where I denotes the mutual information. By the chain rule for mutual information [CT06],

$$I\big(\{x_j, W_Z(x_j)\}_{j\in[n]}; Z\big) = \sum_{j=1}^{n} I\big(x_j, W_Z(x_j); Z \mid x_1, W_Z(x_1), \dots, x_{j-1}, W_Z(x_{j-1})\big)\,.$$

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 $x_j \le \kappa^{-\frac{1}{2}} 2^{Z-1}$, and $W_Z(x_j) = W_m(x_j)$ if $x_j \ge \frac{5}{4}\kappa^{-\frac{1}{2}} 2^{Z+1}$. It yields that:

- for $Z \le \log_2(\frac{4}{5}\sqrt{\kappa}x_j) - 1$, $W_Z(x_j)$ takes a unique value given by $W_m(x_j)$,
- for $Z \ge \log_2(\sqrt{\kappa}x_j) + 1$, $W_Z(x_j)$ takes a unique value given by $W_1(x_j)$,

and otherwise, Z lives in an interval of size at most $\log_2(\sqrt{\kappa}x_j)+1-(\log_2(\frac{4}{5}\sqrt{\kappa}x_j)-1) \le 2 + \log_2(5/4)$ 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

$$I\big(\{x_j, W_Z(x_j)\}_{j\in[n]}; Z\big) \le n \log 5\,.$$

Substituting this into Fano's inequality yields

$$\mathbb{P}\{\widehat{Z} \ne Z\} \ge 1 - \frac{n\log 5 + \log 2}{\log m}\,. \tag{9.3}$$

In general, if the algorithm is randomized, then we can apply the inequality (9.3) conditioned on the random seed ξ of the algorithm, since ξ is independent of Z. It yields

$$\mathbb{P}\{\widehat{Z} \ne Z \mid \xi\} \ge 1 - \frac{n\log 5 + \log 2}{\log m}\,,$$

and upon taking expectations we see that (9.3) holds for randomized algorithms as well.

Combined with (**??**), we obtain $n \gtrsim \log m \gtrsim \log\log\kappa$ as desired. □

9.5 Proof of the upper bound

Let p be the target distribution and let $\tilde{p} = pZ_p$ denote the unnormalized distribution which we access via oracle queries. We recall our preprocessing steps: we assume that the query values take the form $\tilde{p}(x) = \exp(-V(x))$, with $V(0) = V'(0) = 0$ and V satisfying (9.1). This is without loss of generality because we can query $\tilde{p}(0)$ and replace subsequent queries $\tilde{p}(x)$ with $\tilde{p}(x)/\tilde{p}(0)$, 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 32. Our goal is to use the oracle queries to construct an upper envelope $\tilde{q}$ that satisfies $\tilde{q} \ge \tilde{p}$, and $Z_q \lesssim Z_p$, where

$$Z_p \coloneqq \int_{\mathbb{R}} \tilde{p}\,, \qquad Z_q \coloneqq \int_{\mathbb{R}} \tilde{q}$$

are the normalizing constants. The guarantees of Theorem 32 will then follow from standard result on rejection sampling, which we had proved in Theorem 24.

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 $\mathcal{O}(\log \kappa)$, we can find i_0 using only $\mathcal{O}(\log\log\kappa)$ queries to $\tilde{p}$.

Let $x_0 \coloneqq 2^{i_0}/\sqrt{\kappa}$. We first claim that

$$\int_0^{x_0} \tilde{p} \gtrsim x_0\,. \tag{9.4}$$

When $i_0 = 0$, this holds because

$$\int_0^{x_0} \tilde{p} = \int_0^{1/\sqrt{\kappa}} \exp(-V) \ge \int_0^{1/\sqrt{\kappa}} \exp\bigl(-\frac{\kappa x^2}{2}\bigr)\,\mathrm{d}x \ge \frac{1}{3\sqrt{\kappa}} = \frac{x_0}{3}\,.$$

When $i_0 > 0$, this holds because, by definition of i_0, we have $V(x_0/2) \le 1/2$, and so

$$\int_0^{x_0} \tilde{p} \ge \int_0^{x_0/2} \exp(-V) \gtrsim x_0\,.$$

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 $\tilde{q} \ge \tilde{p}$ and hence that $\tilde{q}$ is a valid upper envelope, observe first that since $\tilde{p}(0) = 1$, and $\tilde{p}$ is decreasing, we get that $\tilde{p}(x) \le 1 = \tilde{q}(x)$ 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 $\tilde{p}(x) \le \tilde{q}(x)$ also for the tail $x > x_0$.

To complete the proof, we show that $Z_q \lesssim Z_p$. In light of (9.4) it is sufficient to show that $Z_q \lesssim x_0$. To see this, observe that by Lemma 36, we have

$$Z_q = \int_0^{x_0} \tilde{q} + \int_{x_0}^\infty \tilde{q} \le x_0 + \int_{x_0}^\infty \exp\bigl(-\frac{1}{2x_0}\,(x - x_0) - \frac{1}{2}\,(x - x_0)^2\bigr)\,\mathrm{d}x \le 3x_0\,.$$

This completes the proof. □

We finish by proving an elementary lemma about Gaussian integrals.

Lemma 36. *Let* $a, x_0 > 0$. *Then,*

$$\int_{x_0}^\infty \exp\bigl(-a\,(x - x_0) - \frac{1}{2}\,(x - x_0)^2\bigr)\,\mathrm{d}x \le \frac{1}{a}\,.$$

Proof. Completing the square,

$$\begin{aligned} \int_{x_0}^\infty \exp\bigl(-a\,(x - x_0) - \frac{1}{2}\,(x - x_0)^2\bigr)\,\mathrm{d}x &= \int_0^\infty \exp\bigl(-ax - \frac{1}{2}\,x^2\bigr)\,\mathrm{d}x \\ &= \sqrt{2\pi}\exp\bigl(\frac{a^2}{2}\bigr)\mathbb{P}(Z > a)\,, \end{aligned}$$

where $Z \sim \mathcal{N}(0, 1)$. The result follows from the Mills ratio inequality [Gor41]. □

9.6 Conclusion

In this chapter, 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 Nemirovsky and Yudin [NY83]. A clear future direction suggested by this chapter is to extend this result to higher dimensions, and to ultimately develop a theory of lower complexity bounds and optimal algorithms for sampling. The high dimensional question is particularly interesting because sampling complexity is expected to scale polynomially with the dimension, so determining the optimal dimension dependence of sampling from log-concave distributions is of great theoretical and practical importance. We make partial progress towards this question in Chapters 10 and 11.

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 chapter 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.

Chapter 10

General lower bound in fixed dimension

10.1 Introduction

This chapter extends the result of Chapter 9 to a tight lower bound that holds in all constant dimensions. To recap, the class of target distributions are π on $\mathbb{R}^d$ is α-strongly log-concave and β-log-smooth, with its mode located at the origin. Since the dimension d is considered fixed, the the main parameter of interest is the *condition number*, given by $\kappa \coloneqq \beta/\alpha$. Sampling algorithms are given query access to V and ∇V, and the goal is to produce a sample whose law is close to π in total variation distance. The complexity of the algorithm is measured by the number of queries made.

This chapter is based on the joint work [Che+23b], with Sinho Chewi, Jaume de Dois Pont, Jerry Li, and Shyam Narayanan.

Our contributions. We give a tight characterization of the complexity of log-sampling in any constant dimension $d \geq 2$:

Theorem 33 (informal, see Theorem 34)**.** *For any dimension $d \geq 2$, any sampler for d-dimensional log-concave distributions with condition number κ requires $\Omega(\log \kappa)$ queries.*

Note that this result is exponentially stronger than the $\Omega(\log \log \kappa)$ lower bound in the univariate case [Che+22d]. Moreover, when the dimension d is held fixed, we obtain a matching $O(\log \kappa)$ algorithmic upper bound, based on folklore ideas from the classical literature on sampling from convex bodies (Theorem 35). Together with the result of [Che+22d] for $d = 1$, this settles the complexity of log-concave sampling in constant dimension.

On a technical level, the lower bound is based on a novel construction inspired by work on the Kakeya conjecture in harmonic analysis, which we believe may be of independent interest. We give a detailed description of the construction in Section 10.3.

10.2 Technical overview

Here we summarize the main technical ideas used to prove our lower bounds. For details, see Section 10.3 for Theorem 33, Section 11.3 for Theorem 36, and Section 11.4 for Theorem 37.

Theorem 33 is proved with a construction in dimension two. For convenience, in this section we use radial coordinates to denote points in $\mathbb{R}^2$, so $\omega \coloneqq (x, y) = (r, \theta)$, where $r \in \mathbb{R}_+$ and $\theta \in [0, 2\pi)$. We denote sectors of $\mathbb{R}^2$ enclosed by angles θ_1 and θ_2 as $S(\theta_1, \theta_2) \coloneqq \{(r, \theta) \in \mathbb{R}^2 : \theta \in [\theta_1, \theta_2]\}$, and denote bounded sectors as $S_{\mathsf{bdd}}(\theta_1, \theta_2, r) \coloneqq \{(r', \theta) \in \mathbb{R}^2 : \theta \in [\theta_1, \theta_2],\ r' \le r\}$.

The argument is information-theoretic in nature. We will construct a family of strongly log-concave and log-smooth distributions $\{\pi_1, \dots, \pi_m\}$, where each $\pi_b \propto \exp(-V_b)$, which satisfies two key properties. First, different distributions π_b and $\pi_{b'}$ are well separated in total variation distance; and second, if b is chosen uniformly at random from $[m]$, then querying the potential $(V_b(\omega), \nabla V_b(\omega))$ at any $\omega \in \mathbb{R}^2$ will reveal $O(1)$ bits of information about b. The lower bound in Theorem 33 follows readily from the existence of such a family, provided that m and κ are polynomially related. On the one hand, because the distributions are well-separated in total variation, if we can sample well from the distribution π_b using queries, we can identify the index b with high probability. On the other hand, because there are m distributions and every query reveals $O(1)$ bits of information about b, we need at least $\Omega(\log m) = \Omega(\log \kappa)$ queries to identify b, which results in a $\Omega(\log \kappa)$ query lower bound for log-concave sampling.

How do we construct such a family? A first attempt is to consider distributions supported on thin convex sets that have no overlap. For $b = \frac{1}{\kappa}, \frac{2}{\kappa}, \dots, 1$, let $\pi_b = \mathsf{uniform}(\mathcal{Z}_b)$, where $\mathcal{Z}_b = S_{\mathsf{bdd}}(\frac{\pi}{2}\, b, \frac{\pi}{2}\,(b + \frac{1}{2\kappa}), 1)$, and the size of the family is $m = \lfloor \kappa \rfloor$. The potential V_b is the convex indicator of $\mathcal{Z}_b$, i.e., it is 0 on $\mathcal{Z}_b$ and $+\infty$ outside. Morally, the distributions π_b can be thought of as having condition number κ.

This family does satisfy the two properties needed for the lower bound: different distributions are certainly well-separated because they have disjoint supports; and when we query any potential V_b at a point $\omega \in \mathbb{R}^2$, we always receive one bit of information: whether or not ω lies in the support of π_b. However, the distributions in this family are neither strongly log-concave nor log-smooth. It is easy to make them strongly log-concave while still satisfying the desired properties: we can adjust the distributions by adding the same quadratic function $\frac{\|\cdot\|^2}{2}$ to all of the potentials V_b. But it is much harder to make this family log-smooth.

One way to make this construction log-smooth is to let the potentials V_b

grow slowly (linearly) to infinity outside of the their zero sets $\mathcal{Z}_b$, which leads to a modified second attempt: for $m = \kappa^{\Omega(1)}$, $b = \frac{1}{m}, \ldots, 1$, let π_b have potential $V_b = \tilde{V}_b + \frac{\|\cdot\|^2}{2\kappa^{O(1)}}$, where $\mathcal{Z}_b = S(\frac{\pi}{2}\, b, \frac{\pi}{2}\,(b + \frac{1}{2m}))$, and $\tilde{V}_b(\omega) = \kappa \operatorname{dist}(\omega, \mathcal{Z}_b)$. Note that the potentials V_b are in fact still not smooth at the boundaries of the sets $\mathcal{Z}_b$, but this can be fixed by mollifying V_b. The distributions in this family will be well-separated, because an $\Omega(1)$ fraction of the mass of π_b will lie in $\mathcal{Z}_b$, and the sets $\mathcal{Z}_b$ are disjoint for different b. Unfortunately, this family no longer reveals $O(1)$ bits per query: for any $\omega \in \mathbb{R}^2$, we can identify b with a single query to $(V_b(\omega), \nabla V_b(\omega))$, because either $\omega \in \mathcal{Z}_b$, or $\nabla V_b(\omega)$ reveals the direction of $\mathcal{Z}_b$, and in both cases the index b itself is identified.

We can reduce the information revealed by queries by more carefully controlling the growth of $\tilde{V}_b$, so that the further away a point ω lies from $\mathcal{Z}_b$, the fewer the number of bits will be revealed by $(\tilde{V}_b(\omega), \nabla \tilde{V}_b(\omega))$. This motivates a third attempt at the construction. For $m = 2^N = \kappa^{\Omega(1)}$, $b = \frac{1}{m}, \ldots, 1 - \frac{1}{m}$, let $b = 0.b_1 \ldots b_N$ be the binary expansion of b, and let $[b]_k = 0.b_1 \ldots b_k$ be the truncation of b up to the k-th bit. For $k = 1, \ldots, N$, let $\mathcal{Z}^{\text{radial}}_{k,b} = S(\frac{\pi}{2}\,[b]_k, \frac{\pi}{2}\,([b]_k + 2^{-k}))$, and let $\phi^{\text{radial}}_{k,b}(x) = \kappa^{O(1)}\, 2^{-k} \operatorname{dist}(x, \mathcal{Z}^{\text{radial}}_{k,b})$. Finally, let $V^{\text{radial}}_b = \frac{\|\cdot\|^2}{2\kappa^{O(1)}} + \tilde{V}^{\text{radial}}_b$, where

$$\tilde{V}^{\text{radial}}_b = \max_{k=1,\ldots,N} \phi^{\text{radial}}_{k,b}\,.$$

The potentials V^{radial}_b will again have to be mollified to be made smooth. It turns out that the potentials $\tilde{V}^{\text{radial}}_b$ will grow fast enough outside $\mathcal{Z}^{\text{radial}}_{N,b}$ such that the distributions will be well-separated. It also turns out that queries indeed reveal $O(1)$ bits of information on average. This can be seen as follows: note that the sets $\mathcal{Z}^{\text{radial}}_{k,b}$ are sectors such that $\mathcal{Z}^{\text{radial}}_{k,b} \supset \mathcal{Z}^{\text{radial}}_{k+1,b}$, and as k increases, $\mathcal{Z}^{\text{radial}}_{k,b}$ becomes thinner around the ray $\{\theta = \frac{\pi}{2}\, b\}$; also note that as k increases, the growth rate of $\phi^{\text{radial}}_{k,b}$ outside its zero set $\mathcal{Z}^{\text{radial}}_{k,b}$ is decreasing; these two properties imply that if we query a point $\omega = (r, \theta)$ that is far from the sector $\mathcal{Z}^{\text{radial}}_{i,b}$ (in the sense that $\theta \notin [\frac{\pi}{2}\,[b]_i - 100 \cdot 2^{-i}, \frac{\pi}{2}\,[b]_i + 100 \cdot 2^{-i}]$), then the value of $\tilde{V}^{\text{radial}}_b(\omega)$ will not depend on any $\phi^{\text{radial}}_{k,b}$ for $k > i$, and hence querying $\tilde{V}^{\text{radial}}_b(\omega)$ will only reveal b up to the i-th bit. As a result, if b is chosen uniformly, then for a fixed query ω with high probability we will have $\omega \notin \mathcal{Z}^{\text{radial}}_{k,b}$ for any $k = O(1)$, so the query will only reveal $O(1)$ bits of information about b.

Yet this construction fails because of the mollification step, which we have so far ignored. To make the potentials V_b smooth, we will instead take $V_b = \chi_\delta * \tilde{V}^{\text{radial}}_b + \frac{\|\cdot\|^2}{2\kappa^{O(1)}}$, where χ_δ is supported on a ball of radius $\delta < 2^{-2N}$. We would hope that the potential $\chi_\delta * \tilde{V}^{\text{radial}}_b$ still satisfies the property that querying a point $\omega = (r, \theta)$ that is far from $\mathcal{Z}^{\text{radial}}_{i,b}$ only reveals b up to the i-th bit. When r is not too close to the origin (say $r > 100 \cdot 2^{-i}$), this is indeed still true: if ω satisfies $\theta \notin [\frac{\pi}{2}[b]_i - 200 \cdot 2^{-i}, \frac{\pi}{2}[b]_i + 200 \cdot 2^{-i}]$, then the entire δ-neighbourhood of ω will satisfy $\theta \notin [\frac{\pi}{2}[b]_i - 100 \cdot 2^{-i}, \frac{\pi}{2}[b]_i + 100 \cdot 2^{-i}]$, so the value of $\tilde{V}_b^{\text{radial}}$ on the δ-neighbourhood of ω will not depend on any $\phi_{k,b}^{\text{radial}}$ for $k > i$, hence the value of $(\chi_\delta * \tilde{V}_b^{\text{radial}})(\omega)$ will also not reveal any information of b beyond the i-th bit. But when ω is very close to the origin $(r < \delta)$, the δ-neighbourhood of ω will intersect $\mathcal{Z}_{N,b}^{\text{radial}}$, which means that the value of $(\chi_\delta * \tilde{V}_b^{\text{radial}})(\omega)$ will depend on $\phi_{k,b}^{\text{radial}}$ for all k and hence on all bits of b. In other words, mollification leaks information around the origin. As a result, if we query points δ-close to the origin, we will again identify b in a single query.

The way to resolve the leakage at the origin is to create a branching structure, such that all V_b are equal near the origin so that no information is leaked at small scales, and such that far away from the origin V_b is small around the ray $\{\theta = \frac{\pi}{2}b\}$ so that π_b still concentrates around different sectors. We keep the choices of m and b from the previous construction. The potentials will be $V_b = \chi_\delta * \tilde{V}_b + \frac{\|\cdot\|^2}{2\kappa^{O(1)}}$, where $\tilde{V}_b = \max_{k=1,\ldots,N} \phi_{k,b}$, and $\phi_{k,b}(\omega) = \kappa^{O(1)}\, 2^{-k}\, \text{dist}(\omega, \mathcal{Z}_{k,b})$. The zero set $\mathcal{Z}_{k,b}$, instead of being a radial sector like $\mathcal{Z}_{k,b}^{\text{radial}}$, is now thickened adaptively.

We intuitively describe how to generate $\mathcal{Z}_{k,b}$. Each $\mathcal{Z}_{k,b}$ will be a thickening of $\mathcal{Z}_{k,b}^{\text{radial}}$, by simply including all points within some distance d_k of $\mathcal{Z}_{k,b}^{\text{radial}}$. We define $\mathcal{Z}_{\le k,b} := \bigcap_{k' \le k} \mathcal{Z}_{k',b}$: note that each $\mathcal{Z}_{\le k,b}$ is getting smaller as k increases, and $\mathcal{Z}_{\le N,b}$ is the zero set of $\tilde{V}_b$.

Consider some radii $r_0 < r_1 < r_2 < \ldots$. To generate $\mathcal{Z}_{1,b}$, we thicken $\mathcal{Z}_{1,b}^{\text{radial}}$ (corresponding to the radial sector matching on the first bit), so that it contains $S_{\mathsf{bdd}}(0, \pi/2, r_0)$ (corresponding to the quarter-circle near the origin). This avoids leaking information near the origin, as every x within radius r will be in $\mathcal{Z}_{1,b}$, which means $\phi_{1,b}$ will also be 0. Indeed, we can thicken $\mathcal{Z}_{1,b}^{\text{radial}}$ just the right amount so that it contains $S_{\mathsf{bdd}}(0, \pi/2, r_0)$. For the concrete example where $N = 4$, and $b = 0.1010$, we show a description of $\mathcal{Z}_{1,b}$ in Figure 10-1a: we shade $S_{\mathsf{bdd}}(0, \pi/2, r_0)$ in dark blue, $Z_{1,b}^{\text{radial}} = S(\pi/4, \pi/2)$ in medium blue, and the additional thickening required in light blue.

To generate $\mathcal{Z}_{k,b}$ for $k \ge 2$, we thicken a much thinner angular sector. This ensures that at large radii, the arc of $\mathcal{Z}_{k,b}$ is not too big. We will inductively thicken $\mathcal{Z}_{k,b}$ by some amount d_k just enough to contain $\mathcal{Z}_{k-1,b} \cap S_{\mathsf{bdd}}(0, \pi/2, r_{k-1})$. Consider one more example for $k = 2$ (again for $N = 4$, and $b = 0.1010$), in Figure 10-1b. Note that $\mathcal{Z}_{2,b}^{\text{radial}}$ is the sector $S(\frac{\pi}{4}, \frac{3\pi}{8})$ (shaded in medium blue), and the thickened region (in light blue emanating from both sides of the sector) is just enough to capture all of $\mathcal{Z}_{1,b}$ that was within radius r_1. However, for larger radii, $\mathcal{Z}_{2,b}$ is much thinner than $\mathcal{Z}_{1,b}$. In addition, if we know the first bit $b_1 = 1$, then querying V_b anywhere in $\{r \le r_1\}$ will not reveal any information about the second bit b_2. This is because either we were in $\mathcal{Z}_{1,b}$ which only depends on b_1 (in which case $\phi_{1,b} = \phi_{2,b} = 0$ as we thickened to make sure

$\mathcal{Z}_{2,b} \supset Z_{1,b} \cap S_{\mathsf{bdd}}(0, \pi/2, r_1))$, or we weren't, in which case $\phi_{1,b}$ grows much more quickly than $\phi_{2,b}$.

We can also continue this process inductively for $k = 3, 4$ (Figures 10-1c and 10-1d): we show $\mathcal{Z}_{\leq k,b}$. The intuition for why this prevents leaking of information near the origin is that even if k is large, $\mathcal{Z}_{k,b}$ in the smaller-radius regions is decided by $\mathcal{Z}_{k',b}$ for $k' \ll k$, so we cannot learn any later bits.

The comparisons of $\mathcal{Z}_{k,b}^{\text{radial}}$ and $\mathcal{Z}_{\leq k,b}$ for $b = 0.1010$ and for all $k \leq 4$ are shown together in Figure 10-1. The picture is not to scale, and the radial arcs represent the radii $r_i = 2^i r_0$, for $i = 0, \ldots, 4$.

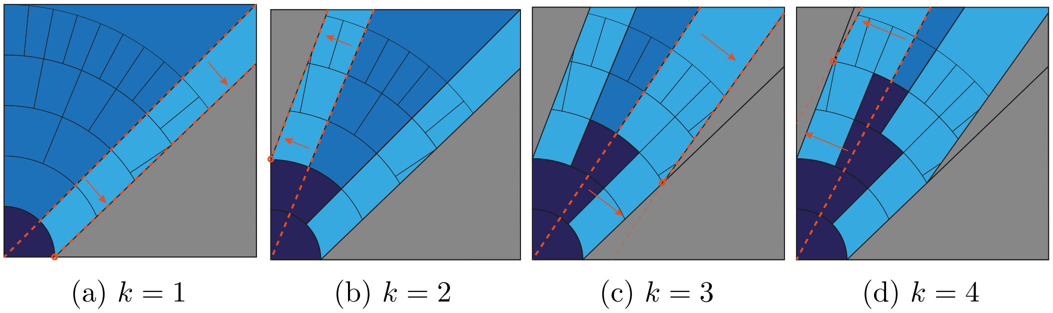
(a) $k = 1$ (b) $k = 2$ (c) $k = 3$ (d) $k = 4$

Figure 10-1: Comparison of $\mathcal{Z}_{\leq k,b}^{\text{radial}}$ (the sector in medium blue) with $\mathcal{Z}_{k,b}$ (union of dark, medium, and light blue), for $k = 1, 2, 3, 4$, and $b = 0.1010$. Dark blue represents the larger angular sectors closer to the origin, and light blue represents the additional fattening from taking sumsets. Each $\mathcal{Z}_{k,b}$ is constructed by thickening $\mathcal{Z}_{k,b}^{\text{radial}}$ enough (illustrated by the red arrows) such that no information about the k-th bit is revealed close to the origin, but $\mathcal{Z}_{k,b}$ continues to get thinner at large radii.

The construction of $\mathcal{Z}_{\leq k,b}$ means that for $k > 1$, querying $\phi_{k,b}$ within $\{r \leq 2^{k-1} r_0\}$ will not reveal the k-th bit, and so even querying the mollified $\chi_\delta * \phi_{k,b}$ within $\{r \leq 2^{k-2} r_0\}$ will not reveal the k-th bit, which stops information leaking near the origin.

Since $\tilde{V}_b = \max_{k=1,\ldots,N} \phi_{k,b}$, the zero set of $\tilde{V}_b$ coincides with $\mathcal{Z}_{\leq N,b}$, and for the choice of $b = 0.1010$, this is shown in the first panel of Figure 10-2. It turns out that each π_b will concentrate around the zero set of $\tilde{V}_b$, and the other panels of Figure 10-2 show these zero sets for seven different values of b in the set $\{\frac{1}{16}, \ldots, \frac{15}{16}\}$ at larger scales. We can see that far out from the origin the zero sets become well-separated, and hence the distributions are well-separated in total variation.

We already discussed how the thickening of $\mathcal{Z}_{k,b}$ means that querying $\phi_{k,b}$, and hence $\tilde{V}_b$, near the origin will not reveal the higher bits of b. For query points $\omega = (r, \theta)$ where r is large, the same analysis on $\tilde{V}_b^{\text{radial}}$ tells us that $\tilde{V}_b(x)$ (even after mollification) will reveal $O(1)$ bits of information about b

when b is chosen uniformly. As mentioned earlier, such a family of distributions readily leads to a sampling lower bound of $\Omega(\log m)$, where m is the size of the family. Since we can choose $m = \kappa^{\Omega(1)}$, this leads to the $\Omega(\log \kappa)$ lower bound. Details of the proof can be found in Section 10.3.

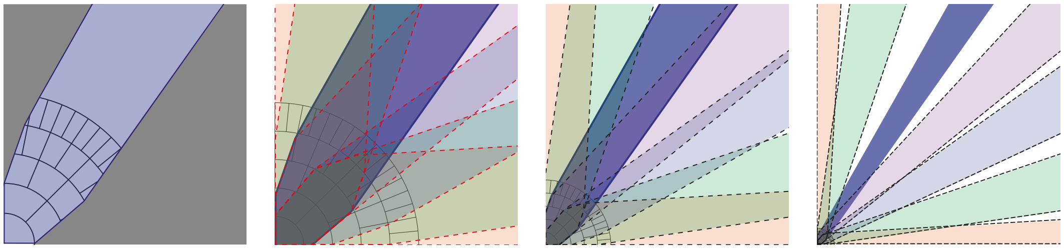

Figure 10-2: Zeros sets of $\tilde{V}_b$. The first panel shows the zero set for $b = 0.1010$. The other panels show the zeros sets for different values of b at different scales. Note that far away from the origin the zero sets become well-separated, which leads to the distributions being well-separated in total variation. Note that if b, b' match in the first ℓ bits, then they will agree up to the ℓ-th circle, as those circles only depend on $\mathcal{Z}_{\leq \ell, b}$ even for ℓ much less than K.

Connections to Kakeya constructions. The construction outlined above is related to Perron's construction [Per28] of Besicovich (Kakeya) sets known as *Perron trees*. Kakeya sets are sets with area zero that contain the translation of a unit segment in any direction. While Kakeya sets over finite fields have been investigated before in theoretical computer science, e.g., [SS08; Dvi09; Juk11], our construction is inspired by Kakeya sets over continuous domains, namely $\mathbb{R}^2$. To our knowledge, this is one of the first applications of these geometric ideas to theoretical computer science.

There are many similarities between our construction and that of Perron. Perron's construction proceeds by the method of sprouting. Sprouting is an iterative process in which, at each step, one adds further and further smaller triangles to the pre-existing construction. The figure is then rescaled in order to have height 1. The construction after n steps contains 2^n triangles of small aperture $\Omega(2^{-n})$, and has area $O(n^{-1})$. We do a similar process in the definition of our sets $\mathcal{Z}_{k,b}$, and indeed, ultimately our hard instance has a very similar tree-like structure.

While we were inspired by the construction of Perron trees, there are also key differences between our hard instance and Perron's construction. Indeed, in our setting, we need to minimize overlap (so that the resulting distributions are well-separated) while simultaneously ensuring that information is not leaked by queries. In contrast, Kakeya sets are explicitly designed to maximize overlap. Secondly, the iterates of Perron trees are convex sets, not convex functions. One must turn these convex sets into convex functions somehow. This is additionally complicated by the fact that these iterates are not nested. In our construction, we must take great care to create nested convex sets, so that the resulting functions are convex and still maintain the structure of the sets.

10.3 Proof of the lower bound

Overview

Our goal is to show the following theorem:

Theorem 34 (lower bound in dimension two)**.** *There is a universal constant $\varepsilon_0 > 0$ such that the following holds. The query complexity of sampling from the class of distributions $\pi \propto \exp(-V)$ on $\mathbb{R}^2$ such that V is 1-strongly convex, κ-smooth, and minimized at 0, with accuracy ε_0 in total variation distance, is at least $\Omega(\log \kappa)$.*

The strategy to do so will be to construct a finite family $\mathcal{S}$ of potentials in the given class which satisfies the following two properties:

- The potentials are *hard to identify via queries* (in the sense of Definition 16 below), and therefore any algorithm must query V at $\Omega(\log \kappa)$ points in order to identify which $V \in \mathcal{S}$ the algorithm is querying.
- The potentials are *well-separated* (in the sense of Definition 17 below), which loosely means that they have mostly non-overlapping support and hence (by Proposition 11) a single sample from $\pi \propto \exp(-V)$ suffices to identify $V \in \mathcal{S}$ with constant probability.

Before describing the potentials $\mathcal{S}$ in more detail, we note some basic definitions.

Definition 12. *Given two functions $f, g : \mathbb{R}^d \to \mathbb{R}$, the* convolution *$f * g$ is the function defined as $(f * g)(x) := \int_{\mathbb{R}^d} f(y)g(x-y)dy$, for all $x \in \mathbb{R}^d$.*

Definition 13. *For $\delta > 0$, we define χ_δ to be the* indicator function of the ball *B_δ of radius δ around the origin. By this, we mean $\chi_\delta(x) = 1$ if $\|x\|_2 \le \delta$, and $\chi_\delta(x) = 0$ otherwise.*

The family $\mathcal{S}$ of potentials will have cardinality $\kappa^{\Omega(1)}$, so that identification of the potential requires $\Omega(\log \kappa)$ bits of information. Actually, by rescaling the potentials, it suffices for each potential V to be $\kappa^{-O(1)}$-convex and $\kappa^{O(1)}$-smooth. Our eventual construction also satisfies the following properties.

- Each $V \in \mathcal{S}$ is of the form $V = \tilde{V} * \chi_\delta + \|\cdot\|^2/(2\kappa^{O(1)})$, where $\tilde{V} : \mathbb{R}^2 \to \mathbb{R}$ is a convex, non-negative, and piecewise linear potential, and δ will have scale $\delta = \kappa^{-\Theta(1)}$.

- Each $V \in \mathcal{S}$ is zero in a small neighbourhood of a ray ℓ emanating from the origin, and grows fast outside of this ray; hence, the potentials are well-separated.

- Suppose that ℓ, ℓ' are the rays corresponding to two potentials $V, V' \in \mathcal{S}$. At distances from ℓ and ℓ' that are much larger than the angle $\angle(\ell, \ell')$, the potentials V, V' are exactly equal. This is the property makes the potentials hard to identify via queries.

Throughout the proof, we assume that κ is sufficiently large, $\kappa \geq \Omega(1)$.

Definitions and the information-theoretic argument

Definition 14 (density and normalizing constant). *Given a strictly convex function $V : \mathbb{R}^d \to \mathbb{R}$, we denote by P_V the probability distribution with density $Z^{-1} \exp(-V)$ w.r.t. Lebesgue measure, where $Z := \int \exp(-V)$ is the normalizing constant. In an abuse of notation, we also use P_V to refer to the density itself.*

Definition 15 (queries and extended oracle). *For a fixed potential V, and given a query $x \in \mathbb{R}^d$, the extended oracle responds with $V(B_\delta(x_1))$, which consists of the value of V for all points in the ball of radius δ centered at x. For a sequence of (possibly adaptive and randomized) queries $x_1, \dots, x_n$ and observations $V(B_\delta(x_1)), \dots, V(B_\delta(x_n))$, we denote the information from the i-th query by $\xi_i := \{x_i, V(B_\delta(x_i))\}$, and the information from all the queries by*

$$\xi_{1:n} := \{\xi_1, \dots, \xi_n\} .$$

Note that the extended oracle in Definition 15 provides more information (the set of values of the potential in some ball around the query point x) to the algorithm than our original first-order query model, from which the algorithm only observes $(V(x), \nabla V(x))$ at the query x. A lower bound for sampling in this stronger query model clearly implies a lower bound in the original query model. We consider the stronger model out of technical convenience, as this notion is robust to the mollification in the construction of the potentials.

Definition 16 (hard to identify via queries). *A finite set $\mathcal{S}$ of potentials in $\mathbb{R}^d$ is called $\mathcal{I}$*-hard to identify with queries at scale δ *if the following holds: for $V \sim \mathsf{uniform}(\mathcal{S})$, any sequence of queries $x_1, \dots, x_n$ to the extended oracle made by a deterministic adaptive algorithm satisfies*

$$I(\xi_{1:n}; V) \leq \mathcal{I} n ,$$

where I denotes the mutual information.

Definition 17 (well-separated set)**.** *A set $\mathcal{S}$ of potentials is* well-separated *if there is a family of measurable sets $(\Omega_V)_{V\in\mathcal{S}}$ where the sets Ω_V are disjoint, and a universal constant $c > 0$ such that*

$$P_V(\Omega_V) \ge c\,, \qquad \text{for all } V \in \mathcal{S}\,.$$

The motivation for this definition is the following fact:

Proposition 11 (one sample identifies well-separated distributions)**.** *Let $\mathcal{S}$ be a well-separated set of potentials and conditionally on $V \sim \mathsf{uniform}(\mathcal{S})$, suppose that X is a sample from a probability measure $\widehat{P}_V$ which is at most $\frac{c}{2}$ away from P_V in total variation distance. Then,*

$$\mathbb{P}\{X \in \Omega_V\} \ge \frac{c}{2}\,.$$

Proof. By conditioning on V,

$$\mathbb{P}\{X \in \Omega_V\} = \mathbb{E}\,\mathbb{P}\{X \in \Omega_V \mid V\} = \mathbb{E}\,\widehat{P}_V(\Omega_V) \ge \mathbb{E}\big[P_V(\Omega_V) - \|P_V - \widehat{P}_V\|_{\mathrm{TV}}\big] \ge \frac{c}{2}\,,$$

which is what we wanted to show. □

This shows that the minimum-distance estimator

$$\widehat{V} \coloneqq \operatorname*{arg\,min}_{V\in\mathcal{S}} \inf_{z\in\Omega_V} \|X - z\| \tag{10.1}$$

succeeds at estimating the randomly drawn V with constant probability. On the other hand, using Fano's inequality (Theorem 1) we can reduce Theorem 34 to the following proposition:

Proposition 12 (well-separated set which is hard to identify via queries)**.** *Let $\kappa \ge \Omega(1)$. Then, there is a set $\mathcal{S}$ of potentials such that:*

1. *All elements of $\mathcal{S}$ are $\kappa^{-O(1)}$-convex and $\kappa^{O(1)}$-smooth, and have their minimum at zero.*

2. *$\mathcal{S}$ has cardinality $\kappa^{\Omega(1)}$.*

3. *$\mathcal{S}$ is well-separated with $c = \Omega(1)$.*

4. *$\mathcal{S}$ is hard to identify via queries at scale $\delta = \kappa^{-\Theta(1)}$, and with $\mathcal{I} = O(1)$.*

Proof of Theorem 34. Suppose that there is a sampling algorithm which, given any target distribution $\pi \propto \exp(-V)$ on $\mathbb{R}^2$ such that V is 1-strongly convex, $\bar{\kappa}$-smooth, and minimized at 0, outputs a sample X whose law is ε_0 close in total variation distance to π using $n(\bar{\kappa})$ queries to the extended oracle. Let $\mathcal{S}$ be the family in Proposition 12. By choosing $\varepsilon_0 = c/2 = \Omega(1)$ and rescaling the potentials accordingly, then Proposition 11 implies that the sampling algorithm can identify $V \sim \mathsf{uniform}(\mathcal{S})$ using $n(\bar{\kappa})$ queries with constant probability, where $\bar{\kappa} = \kappa^{O(1)}$. Namely, for the estimator $\widehat{V}$ in (10.1),

$$\mathbb{P}\{\widehat{V} = V\} \geq \frac{c}{2} = \Omega(1)\,. \tag{10.2}$$

On the other hand, we can prove a lower bound for the error probability of any estimator $\widehat{V}$ constructed using adaptive queries. First we assume that the estimator is deterministic given previous queries. Because the set $\mathcal{S}$ is hard to identify, by Fano's inequality (Theorem 1) we have

$$\mathbb{P}\{\widehat{V} \neq V\} \geq 1 - \frac{I(\xi_{1:n(\bar{\kappa})}; V) + \log 2}{\log |\mathcal{S}|} \geq 1 - \frac{\mathcal{I}n(\bar{\kappa}) + \log 2}{\log |\mathcal{S}|} = 1 - \Omega\Bigl(\frac{n(\bar{\kappa})}{\log \kappa}\Bigr)\,, \tag{10.3}$$

for all $n(\bar{\kappa}) \leq c\,|\mathcal{S}| = O(\log \kappa)$. If the estimator is instead randomized, it depend on a random seed ζ that is independent of V. In this case, the same argument as above conditional on ζ gives

$$\mathbb{P}\{\widehat{V} \neq V \mid \zeta\} \geq 1 - \Omega\Bigl(\frac{n(\bar{\kappa})}{\log \kappa}\Bigr)\,.$$

Taking expectation over ζ, we see that (10.3) holds also for randomized algorithms. Combined with (10.2), we see that $n(\bar{\kappa}) \geq \Omega(\log \kappa) = \Omega(\log \bar{\kappa})$. □

Reductions and properties of the construction

Recall from Section 10.3 that each $V \in \mathcal{S}$ is of the form $V = \tilde{V} * \chi_\delta + \|\cdot\|^2/(2\kappa^{O(1)})$. In this section, we reduce the desired properties of $\mathcal{S}$, namely that $\mathcal{S}$ is well-separated and hard to identify via queries, to geometric properties of the potentials summarized in Proposition 13 below.

By increasing κ by a factor of at most two, which will not harm the final lower bound, we can assume that $\kappa = 2^N$ for some positive integer N. We also set $\delta \coloneqq \kappa^{-5}$. Let B_N denote the set of binary strengths of length N. For each $b \in B_N$ and $\ell \in [N]$, we let $[b]_\ell \coloneqq 0.00b_1 \dots b_\ell$ in binary representation, and set $[b] \coloneqq [b]_N$.

Proposition 13 (geometric properties)**.** *There are functions $\tilde{V}_b$, for $b \in B_N$, such that:*

(P0) *$\tilde{V}_b$ is convex and $\kappa^{O(1)}$-smooth on average at scale $\delta = \kappa^{-5}$, i.e., $\tilde{V}_b * \chi_\delta$ is $\kappa^{O(1)}$-smooth, and attains its minimum $V_b(0) = 0$ at zero.*

(P1) *The zero set* $\mathcal{Z}_b \coloneqq \{\tilde{V}_b = 0\}$ *contains the* $10^3\delta$*-neighborhood of the set*

$$\tilde{\mathcal{Z}}_b \coloneqq \{(x, \beta x) \in \mathbb{R}^2 \mid x \geq 0,\ [b] - 2^{-N} \leq \beta \leq [b] + 2^{-N}\}\,, \tag{10.4}$$

and is contained in the 1*-neighbourhood of* $\tilde{\mathcal{Z}}_b$*.*

(P2) *Moreover, for all* $x, y \in \mathbb{R}^2$*,*

$$\tilde{V}_b(x, y) \geq \kappa^4 \left(\operatorname{dist}((x, y), \tilde{\mathcal{Z}}_b) - 1\right)_+ .$$

(P3) *If* b*,* b' *coincide in the first* ℓ *bits then* $\tilde{V}_b$ *and* $\tilde{V}_{b'}$ *coincide in the set*

$$\left\{(x, y) \in \mathbb{R}^2 \;\middle|\; x < \frac{1}{4}\, 2^{-3N} \text{ or } |y - [b]_\ell\, x| > 100 \cdot 2^{-\ell} x\right\}.$$

We check that these properties imply that Proposition 12 holds.

Proof of Proposition 12. Let $\mathcal{S}$ be the collection of potentials $V_b \coloneqq \tilde{V}_b * \chi_\delta + \|\cdot\|^2/(2\kappa^{16})$ for $b \in B_N$, where $\{\tilde{V}_b : b \in B_N\}$ are the functions from Proposition 13. We now verify the four properties of Proposition 12.

Proof of 1. By (P0), we know that $\tilde{V}_b$ is convex, which implies that $\tilde{V}_b * \chi_\delta$ is also convex. Therefore, V_b is κ^{-16}-strongly convex. In addition, by (P0), $\tilde{V}_b * \chi_\delta$ is $\kappa^{O(1)}$-smooth, which means that V_b is $\kappa^{O(1)} + \kappa^{-16} \leq \kappa^{O(1)}$-smooth.

Proof of 2. By construction, $|\mathcal{S}| = \kappa$.

Proof of 3. We now show that $\mathcal{S}$ is c-separated. For any string b, recall the definition of $\tilde{\mathcal{Z}}_b$ from (10.4). Define the set

$$\Omega_b \coloneqq \{(x, \beta x) \in \mathbb{R}^2 \mid x \geq 2^{-3N},\ [b] - 0.4 \cdot 2^{-N} \leq \beta \leq [b] + 0.4 \cdot 2^{-N}\}\,.$$

It is clear that $\{\Omega_b : b \in B_N\}$ is a family of disjoint sets. By (P1) we know that the zero set $\mathcal{Z}_b$ of $\tilde{V}_b$ contains a $10^3\delta$-neighborhood of $\tilde{\mathcal{Z}}_b$. Since $\Omega_b \subset \tilde{\mathcal{Z}}_b$, it follows that $\tilde{V}_b * \chi_\delta = 0$ on Ω_b.

Let $\tilde{\Omega}_b \coloneqq \{(x, y) \in \Omega_b : \|(x, y)\| \leq \kappa^8\}$. Note that the full set of points (x, y) with $\|(x, y)\| \leq \kappa^8$ has volume $\pi\kappa^{16}$, and Ω_b is a sector of the plane with arc $\Theta(2^{-N})$, minus a small set of points (specifically, the points in the sector with $x \leq 2^{-3N}$, which also means $y \leq O(2^{-3N})$). Therefore, the volume of $\tilde{\Omega}_b$ is $\Theta(\kappa^{16} \cdot 2^{-N}) = \Theta(\kappa^{15})$. In addition, all points $(x, y) \in \tilde{\Omega}$ have $V_b(x, y) = -\|(x, y)\|^2/(2\kappa^{16}) \geq -1/2$. Hence,

$$\int_{\Omega_b} \exp(-V_b) \geq \int_{\tilde{\Omega}_b} \exp(-V_b) \geq \Omega(\kappa^{15})\,. \tag{10.5}$$

Next, we bound the full integral of $\exp(-V_b)$ across $\mathbb{R}^d$ by splitting $\mathbb{R}^d$ into four regions $\mathbb{R}^d = \tilde{\mathcal{Z}}_b \cup \Psi_{1,b} \cup \Psi_{2,b} \cup \Psi_{3,b}$, defined as follows:

- $\Psi_{1,b} \coloneqq \{(x,y) \in \mathbb{R}^2 \setminus \tilde{\mathcal{Z}}_b : \operatorname{dist}((x,y), \tilde{\mathcal{Z}}_b) \le 2,\ \|(x,y)\| \le \kappa^9\}$.
- $\Psi_{2,b} \coloneqq \{(x,y) \in \mathbb{R}^2 \setminus (\tilde{\mathcal{Z}}_b \cup \Psi_{1,b}) : \|(x,y)\| \le \kappa^9\}$.
- $\Psi_{3,b} = \mathbb{R}^2 \setminus (\tilde{\mathcal{Z}}_b \cup \Psi_{1,b} \cup \Psi_{2,b})$.

Note that all points $\Psi_{3,b}$ have norm at least κ^9. To show that most of the mass of P_{V_b} is concentrated on $\tilde{\mathcal{Z}}_b$, we must show that the integrals over $\Psi_{1,b}$, $\Psi_{2,b}$, and $\Psi_{3,b}$ are small. In a nutshell, the integral over $\Psi_{1,b}$ is small because the 2-neighborhood of $\tilde{\mathcal{Z}}_b$ is small (relative to the size of $\tilde{\mathcal{Z}}_b$ itself); the integral over $\Psi_{2,b}$ is small because $\tilde{V}_b$ increases rapidly outside $\tilde{\mathcal{Z}}_b$; and the integral over $\Psi_{3,b}$ is small because the Gaussian part of V_b is small over this region.

On these four regions, we have the following bounds. First, $\int_{\mathbb{R}^2} \exp(-\|\cdot\|^2/(2\kappa^{16})) = 2\pi\kappa^{16}$. Therefore, since the sector $\tilde{\mathcal{Z}}_b$ has arc $\Theta(2^{-N})$, by rotational symmetry

$$\int_{\tilde{\mathcal{Z}}_b} \exp(-V_b) \le \int_{\tilde{\mathcal{Z}}_b} \exp\bigl(-\frac{\|\cdot\|^2}{2\kappa^{16}}\bigr) \le O(2^{-N}) \int_{\mathbb{R}^2} \exp\bigl(-\frac{\|\cdot\|^2}{2\kappa^{16}}\bigr) \le O(\kappa^{15})\,.$$

Note that $\Psi_{1,b}$ consists of two strips adjacent to $\tilde{\mathcal{Z}}_b$, where each strip has width 2 and length $O(\kappa^9)$, together with a piece of area $O(1)$ near the origin. Thus, $\operatorname{vol}(\Psi_{1,b}) \le O(\kappa^9)$, yielding

$$\int_{\Psi_{1,b}} \exp(-V_b) \le \operatorname{vol}(\Psi_{1,b}) \le O(\kappa^9)\,.$$

Next, for $(x,y) \in \mathbb{R}^2$ such that $\operatorname{dist}((x,y), \tilde{\mathcal{Z}}_b) \ge 3/2$, by (P2) we have $\tilde{V}_b(x,y) \ge \kappa^4$. After mollification at scale $\delta \le 1/2$, we conclude that $\tilde{V}_b * \chi_\delta \ge \kappa^4$ on $\Psi_{2,b}$. In addition, $\Psi_{2,b}$ is contained in the ball of radius κ^9, so the volume of $\Psi_{2,b}$ is at most $\pi\kappa^{18}$. Therefore,

$$\int_{\Psi_{2,b}} \exp(-V_b) \le \pi\kappa^{18} \exp(-\kappa^4)\,.$$

Finally, all points in $\Psi_{3,b}$ have ℓ_2 norm at least κ^9, so

$$\int_{\Psi_{3,b}} \exp(-V_b) \le \iint_{\|(x,y)\| \ge \kappa^9} \exp\bigl(-\frac{\|(x,y)\|^2}{2\kappa^{16}}\bigr) \le O(\kappa^8) \exp\bigl(-\Omega(\kappa^2)\bigr)\,,$$

by standard Gaussian tail estimates. Therefore,

$$\int_{\mathbb{R}^2} \exp(-V_b) \le O\Bigl(\kappa^{15} + \kappa^9 + \exp\bigl(-\Omega(\kappa^4)\bigr) + \exp\bigl(-\Omega(\kappa^2)\bigr)\Bigr) \le O(\kappa^{15})\,. \tag{10.6}$$

Overall, (10.5) and (10.6) together imply that $P_{V_b}(\Omega_b) \ge \Omega(1)$, i.e., $\mathcal{S}$ is

$\Omega(1)$-well-separated.

Proof of 4. Finally, we show that $\mathcal{S}$ is hard to identify via queries at scale $\delta = \kappa^{-\Theta(1)}$ with $\mathcal{I} = O(1)$. We consider b drawn uniformly at random from B_N.

First, however, we need to extend (P3) to V_b (i.e., taking into account the mollification at scale δ). We claim that if b, b' coincide in the first ℓ bits, then V_b and $V_{b'}$ coincide in the set

$$\bigl\{(x,y) \in \mathbb{R}^2 \bigm| x < \frac{1}{8}\, 2^{-3N} \text{ or } |y - [b]_\ell\, x| > 200 \cdot 2^{-\ell} x\bigr\}\,. \tag{10.7}$$

In light of (P3), it suffices to show that if (x,y) lies in this set and $\|(x',y') - (x,y)\| \le \delta$, then $x' < \frac{1}{4}\, 2^{-3N}$ or $|y' - [b]_\ell\, x'| > 100 \cdot 2^{-\ell}\, x'$. In other words, the δ-neighborhood of (10.7) is contained in the set in (P3). In the first case, $x' < \frac{1}{4}\, 2^{-3N}$ follows if $\delta < \frac{1}{8}\, 2^{-3N}$, but since $\delta = \kappa^{-5} = 2^{-5N}$ this holds for large κ. In the second case,

$$|y' - [b]_\ell\, x'| \ge |y - [b]_\ell\, x| - \delta - [b]_\ell\, \delta \ge 200 \cdot 2^{-\ell}\, x - 2\delta\,.$$

This is greater than $100 \cdot 2^{-\ell}\, x$ provided that $2\delta \le 100 \cdot 2^{-\ell}\, x$, but this follows because $\delta = 2^{-5N}$ and $x \ge \frac{1}{8}\, 2^{-3N}$ (as we are in the negation of the first case). In fact, by replacing δ with 2δ, the same argument shows that for all (x,y) lying in the set (10.7), we have $V_b(B_\delta(x,y)) = V_{b'}(B_\delta(x,y))$. Note also that (10.7) shows that it is useless to query any points (x,y) with $x < \frac{1}{8}\, 2^{-3N}$, so for the remainder of the proof we assume that the algorithm does not do so.

We now move to a stronger oracle model. Namely, given a query point $(x,y) \in \mathbb{R}^2$, let ℓ be the largest integer such that $|y - [b]_\ell\, x| \le 200 \cdot 2^{-\ell}\, x$. Then, the oracle outputs $\hat{\xi} := [b]_{\ell+1}$, i.e., the oracle reveals the first $\ell + 1$ bits of b. To see that this new oracle is indeed stronger, observe that we can simulate the previous oracle using the revealed bits $[b]_{\ell+1}$; namely, pick any bit string b' which is consistent, in the sense that $[b']_{\ell+1} = [b]_{\ell+1}$. Then, by the choice of ℓ, we have $|y - [b]_{\ell+1}\, x| > 200 \cdot 2^{-(\ell+1)}\, x$, so that $V_b(B_\delta(x,y)) = V_{b'}(B_\delta(x,y))$, and hence we can output $V_b(B_\delta(x,y))$ given knowledge of $[b]_{\ell+1}$. It therefore suffices to bound the mutual information $I(\hat{\xi}_{1:n}; b)$ where $\hat{\xi}_{1:n}$ denotes the output of the stronger oracle on a sequence of adaptive but deterministic queries $(x_1,y_1),\dots,(x_n,y_n)$.

We can then write

$$\begin{aligned} I(\hat{\xi}_{1:n}; b) &= \sum_{i=1}^n I(\hat{\xi}_i; b \mid \hat{\xi}_{1:i-1}) &\quad (10.8)\\ &= \sum_{i=1}^n \{H(\hat{\xi}_i \mid \hat{\xi}_{1:i-1}) - H(\hat{\xi}_i, \mid \hat{\xi}_{1:i-1}, b)\} &\quad (10.9) \end{aligned}$$

$$\leq \sum_{i=1}^{n} H(\hat{\xi}_i \mid \hat{\xi}_{1:i-1})\,, \tag{10.10}$$

where $H(\cdot \mid \cdot)$ denotes the conditional entropy. The first line follows from the chain rule for mutual information, the second line follows from definition of mutual information, and third line follows from non-negativity of conditional entropy. Thus, we are done if we can show that $H(\hat{\xi}_i \mid \hat{\xi}_{1:i-1}) \leq O(1)$, for all $i \leq c\,|\mathcal{S}|$.

Conditionally on any particular realization of $\hat{\xi}_{1:i-1}$, let ℓ_0 denote the number of bits of b revealed thus far and let $[b_0]_{\ell_0}$ denote the revealed bits. Clearly the bit string b is uniformly distributed on the set B'_N of bit strings b' with $[b']_{\ell_0} = [b_0]_{\ell_0}$. Also, since we have assumed that the algorithm's queries are deterministic given the past history, the next query point (x_i, y_i) is deterministic. Then, the conditional probability that $\ell \geq \ell_0$ bits are revealed by the next query is

$$\begin{aligned}
&\mathbb{P}\{200 \cdot 2^{-\ell}\, x_i < |y_i - [b]_\ell\, x_i| \leq 200 \cdot 2^{-(\ell-1)}\, x_i \mid \hat{\xi}_{1:i-1}\} \\
&\qquad \leq \mathbb{P}\Big\{\frac{y_i}{x_i} - 200 \cdot 2^{-(\ell-1)} \leq [b]_\ell \leq \frac{y_i}{x_i} + 200 \cdot 2^{-(\ell-1)} \;\Big|\; \hat{\xi}_{1:i-1}\Big\}\,.
\end{aligned}$$

This is the probability that a uniformly chosen element of B'_N belongs to an interval of length $\Theta(2^{-\ell})$. Since there are $2^{N-\ell_0}$ elements of B'_N, and $\Theta(2^{N-\ell})$ of them belong to any fixed interval of length $\Theta(2^{-\ell})$, we conclude that the above probability is $O(2^{-(\ell-\ell_0)})$.

We then have

$$H(\hat{\xi}_i \mid \hat{\xi}_{1:i-1}) \leq \mathbb{E} \sum_{\ell \geq \ell_0} (\ell - \ell_0)\, O(2^{-(\ell-\ell_0)}) \leq O(1)\,, \tag{10.11}$$

where the expectation is taken over ℓ_0 (which depends on the realization of $\hat{\xi}_{1:i-1}$). Substituting the above bound into (10.10), we conclude that $I(\xi_{1:n}; b) = O(n)$, which implies that $\mathcal{S}$ is indeed hard to identify via queries. □

Construction of the distributions

This section contains the proof of Proposition 13.

For integers $1 \leq k \leq N$, let $[b]_k$ be the number $0.00b_1b_2\ldots b_k$ in binary representation, and let $[b]_k \coloneqq [b] \coloneqq [b]_N$ for $k \geq N$. Define

$$\phi_{k,b}(x, y) \coloneqq \big(|y - [b]_k\, x| - (2^{-k}\, x + 2^{-(3N-k)})\big)_+\,. \tag{10.12}$$

We also write $\phi_k \coloneqq \phi_{k,b}$ when b is clear from context. For $x \geq 0$, the function

ϕ_k essentially measures the distance to the set

$$\{(x, [b]_k\, x + \xi_k) \in \mathbb{R}^2 : x \ge 0,\ |\xi_k| \le 2^{-k}\, x + 2^{-(3N-k)}\}\,.$$

Finally, we define the potential

$$\tilde V_b(x,y) \coloneqq 2^{7N} \max_{k=1,\dots,N} 2^{-k} \phi_k(x,y)\,. \tag{10.13}$$

Proof of Proposition 13. We prove that the construction (10.13) satisfies each of the four properties in turn.

Proof of Property (P0). The convexity of $\tilde V_b$ follows because each ϕ_k is convex. To check that $\tilde V_b$ is $\kappa^{O(1)}$-smooth on average, using the compositionality of the maximum (i.e., $\max(a, \max(b,c)) = \max(a,b,c)$) we see that that $\tilde V_b$ can be written as a maximum of affine functions, each of slope $\kappa^{O(1)}$; hence, $\tilde V_b$ is $\kappa^{O(1)}$-Lipschitz. Differentiating under the integral,

$$\nabla(\tilde V_b * \chi_\delta)(x,y) = \iint_{B_\delta} \nabla \tilde V_b(x+u, y+v)\,\mathrm{d}u\,\mathrm{d}v = \iint \nabla \tilde V_b\, \mathbb{1}_{B_\delta(x,y)}\,,$$

where the expression makes sense because $\tilde V_b$ is Lipschitz and hence differentiable a.e. by Rademacher's theorem, and the absolute continuity of $\tilde V_b$ ensures the validity of the fundamental theorem of calculus. Then, by Hölder's inequality,

$$\begin{aligned}\|\nabla(\tilde V_b * \chi_\delta)(x,y) - \nabla(\tilde V_b * \chi_\delta)(x',y')\| &\le \bigl(\sup \|\nabla \tilde V_b\|\bigr)\, \|\mathbb{1}_{B_\delta(x,y)} - \mathbb{1}_{B_\delta(x',y')}\|_{L^1} \\ &\le \kappa^{O(1)}\, \mathrm{vol}\bigl(B_\delta(x,y) \mathbin{\triangle} B_\delta(x',y')\bigr)\,.\end{aligned}$$

By elementary considerations, the volume of the symmetric difference between the balls is bounded by $O(\kappa^{O(1)}\, \|(x,y) - (x',y')\|)$, and therefore $\nabla(\tilde V_b * \chi_\delta)$ is $\kappa^{O(1)}$-Lipschitz.

Finally, it is obvious that $\tilde V_b \ge 0$ and $\tilde V_b = 0$ at the origin.

Proof of Property (P1). We only need to verify that any point (x,y) which is $10^3\delta$-close to $\tilde{\mathcal{Z}}_b$ satisfies $\tilde V_b(x,y) = 0$, as the second part of Property (P1) is automatically implied by Property (P2). For such a point (x,y), there exists (x',y') such that

$$x' \ge 0\,, \qquad |x'-x| \wedge |y'-y| \le 10^3\delta\,, \qquad \text{and} \qquad |y' - [b]\, x'| \le 2^{-N} x'\,.$$

This also implies $|y' - [b]_k\, x'| \le 2^{-k}\, x'$ for all $1 \le k \le N$, since $|[b]_k - [b]| \le 2^{-k} - 2^{-N}$. Therefore, for all $1 \le k \le N$, $|y - [b]_k\, x| \le 2^{-k}\,(x + 10^3\delta) + 2\cdot 10^3 \delta \le 2^{-k} x + 2^{-(3N-k)}$, since $\delta = 2^{-5N}$. By the definition (10.13) of $\tilde V_b$ and the definition of ϕ_k in (10.12), it follows that $\tilde V_b(x,y) = 0$.

Proof of Property (P2). We just need to check that

$$2^{6N}\phi_N(x,y) \geq \kappa^4 \left(\mathrm{dist}((x,y),\tilde{\mathcal{Z}}_b) - 1\right)_+ ,$$

or equivalently, $2^{2N}\phi_N(x,y) \geq (\mathrm{dist}((x,y),\tilde{\mathcal{Z}}_b) - 1)_+$. We first consider the case when $x \geq 0$, and we may assume that $(x,y) \notin \tilde{\mathcal{Z}}_b$ as otherwise the claim is obvious. If (x,y) has distance Δ to its closest point in $\tilde{\mathcal{Z}}_b$, then any y' such that $(x,y') \in \tilde{\mathcal{Z}}_b$ must satisfy $|y-y'| \geq \Delta$. Applying this to $y' = [b]\,x \pm 2^{-N}\,x$, we obtain

$$\mathrm{dist}\big((x,y),\tilde{\mathcal{Z}}_b\big) \leq |y - [b]\,x + 2^{-N}x| \wedge |y - [b]\,x - 2^{-N}\,x| = |y - [b]\,x| - 2^{-N}\,x\,.$$

In turn, it implies that $\phi_N(x,y) \geq (\mathrm{dist}((x,y),\tilde{\mathcal{Z}}_b) - 2^{-(3N-k)})_+ \geq (\mathrm{dist}((x,y),\tilde{\mathcal{Z}}_b) - 1)_+$.

If $x < 0$, then $\mathrm{dist}((x,y),\tilde{\mathcal{Z}}_b) \leq \|(x,y)\| \leq \sqrt{2}\max(|x|,|y|)$. Then, for N large,

$$\begin{aligned}
2^{2N}\phi_N(x,y) &= 2^{2N}\left(|y - [b]\,x| - 2^{-N}\,x - 2^{-(3N-k)}\right)_+ \\
&= 2^{N-1/2}\left(2^{N+1/2}\,|y - [b]\,x| + \sqrt{2}\,|x| - 2^{-(2N-k)+1/2}\right)_+ \\
&\geq 2^{N-1/2}\left(2^{3/2}\,\max\big(0,|y| - \frac{1}{2}\,|x|\big) + \sqrt{2}\,|x| - 1\right)_+ \\
&\geq 2^{N-1/2}\left(\sqrt{2}\,\max(|x|,|y|) - 1\right)_+ \geq \left(\mathrm{dist}((x,y),\tilde{\mathcal{Z}}_b) - 1\right)_+ .
\end{aligned}$$

Proof of Property (P3). The last property follows from Proposition 14 below, because if b, b' agree on the first ℓ bits, then on the set in the statement of Property (P3),

$$\begin{aligned}
\tilde{V}_b &= 2^{7N}\max_{k=1,\dots,N} 2^{-k}\phi_{k,b} = 2^{7N}\max_{k=1,\dots,\ell} 2^{-k}\phi_{k,b} = 2^{7N}\max_{k=1,\dots,\ell} 2^{-k}\phi_{k,b'} \\
&= 2^{7N}\max_{k=1,\dots,N} 2^{-k}\phi_{k,b'} \\
&= \tilde{V}_{b'}\,.
\end{aligned}$$

The second and fourth equalities invoke Proposition 14, and the third equality uses the fact that $\phi_{k,b}$ only depends on b through $[b]_k$. This completes the proof. □

Proposition 14 (potentials agree if bits agree)**.** *Let $S_\ell(b)$ be the set*

$$S_\ell(b) := \left\{(x,y) \in \mathbb{R}^2 : x < \frac{1}{4}\,2^{-3N} \text{ or } |y - [b]_\ell\,x| \geq 100 \cdot 2^{-\ell}x\right\}.$$

Then, for $x, y \in S_\ell(b)$,

$$\max_{k=1,\dots,N} 2^{-k}\phi_k(x,y) = \max_{k=1,\dots,\ell} 2^{-k}\phi_k(x,y)\,.$$

In turn, Proposition 14 follows by induction from:

Proposition 15 (induction)**.** *If $(x,y) \in S_\ell(b)$, and for some $k > \ell$ we have $\phi_k(x,y) > 0$, then $\phi_k(x,y) \le 2\phi_{k-1}(x,y)$.*

Proof. First, we may assume that $x > 0$. This is because if $x \le 0$,

$$\begin{aligned}
\phi_{k-1}(x,y) &\ge |y - [b]_k\, x| - |[b]_{k-1} - [b]_k|\, |x| - 2^{-(k-1)}\, x - 2^{-(3N-k+1)} \\
&\ge |y - [b]_k\, x| + 2^{-k}x - 2^{-(k-1)}x - 2^{-(3N-k+1)} \\
&= |y - [b]_k\, x| - 2^{-k}x - 2^{-(3N-k+1)} \\
&\ge \phi_k(x,y)\,,
\end{aligned}$$

since we are assuming $\phi_k(x,y) > 0$.

Now, since $x > 0$, we start by estimating

$$\begin{aligned}
\phi_{k-1}(x,y) &\ge |y - [b]_k\, x| - |[b]_{k-1} - [b]_k|\, x - 2^{-(k-1)}x - 2^{-(3N-k+1)} \\
&\ge |y - [b]_k\, x| - 3 \cdot 2^{-k}x - 2^{-(3N-k+1)} \\
&= \phi_k(x,y) - 2 \cdot 2^{-k}x + 2^{-(3N-k+1)}
\end{aligned}$$

and

$$\phi_k(x,y) = |y - [b]_k\, x| - (2^{-k}x + 2^{-(3N-k)})\,.$$

First, suppose that $x \le \frac{1}{4}\, 2^{-3N}$. Then, $2^{-(3N-k+1)} \ge 2 \cdot 2^{-k}x$, so in fact $\phi_{k-1}(x,y) \ge \phi_k(x,y)$. Alternatively, if $x \ge \frac{1}{4}\, 2^{-3N}$ and $|y - [b]_\ell\, x| \ge 100 \cdot 2^{-\ell}x$, then

$$\begin{aligned}
2\phi_{k-1}(x,y) &\ge 2\,|y - [b]_\ell\, x| - 2\,|[b]_\ell - [b]_{k-1}|\, x - 4 \cdot 2^{-k}x - 2^{-(3N-k)} \\
&\ge 2\,|y - [b]_\ell\, x| - 6 \cdot 2^{-\ell}x - 2^{-(3N-k)}\,, \\
\phi_k(x,y) &\le |y - [b]_\ell\, x| + |[b]_\ell - [b]_k|\, x - 2^{-k}x - 2^{-(3N-k)} \\
&\le |y - [b]_\ell\, x| + 2^{-\ell}x - 2^{-(3N-k)}\,.
\end{aligned}$$

As a result, when $|y - [b]_\ell\, x| \ge 100 \cdot 2^{-\ell}x$, we see that $\phi_{k-1}(x,y) \ge \frac{1}{2}\,\phi_k(x,y)$. □

10.4 Upper bound in constant dimension

In this section we give a simple proof that in constant dimension, one can approximately generate a sample from a log-concave distribution with condition number κ, in $O(\log \kappa)$ queries. Our query dependence also has a polylogarithmic dependence on $\frac{1}{\varepsilon}$, if we wish to generate a sample that is ε-close in TV distance to the true distribution. (We do not attempt to optimize the dependence on dimension d or the polylogarithmic dependence on $\frac{1}{\varepsilon}$.)

Let V be a convex function that is 1-strongly convex and κ-smooth, such that V is minimized at the origin and $V(0) = 0$. For any real value $y \geq 0$, define $B_V(y)$ to be the set of points x such that $V(x) \leq y$.

First, we note the following basic facts that follow immediately from our convexity assumptions.

Proposition 16 (basic facts about log-concavity)**.** *1. $B_V(y)$ is a convex body for any $y > 0$, and contains 0.*

2. $B_V(y)$ is contained in the ball of radius $\sqrt{2y}$ and contains the ball of radius $\sqrt{2y/\kappa}$.

3. For any $0 < y < y'$, $B_V(y') \subset \frac{y'}{y} B_V(y)$.

Next, we show how to obtain a crude $d^{O(1)}$-approximation for $B_V(1)$ using $d^{O(1)} \log \kappa$ first-order queries. The proof is essentially folklore and follows from the ellipsoid method.

Proposition 17 (ellipsoid method)**.** *Let B be a convex body that contains $B(0, r)$ and is contained in $B(0, R)$, along with a membership and separation oracle. Using $d^{O(1)} \log \frac{R}{r}$ adaptive queries to the membership and separation oracle, we can find an ellipsoid E centered around some point z such that $E \subset B \subset E'$, where E' is E dilated by an $O(d^{3/2})$ factor about z.*

We can apply the above proposition to the convex body $B_V(1)$.

Corollary 12 (sublevel set approximation)**.** *Using $d^{O(1)} \log \kappa$ adaptive queries to V and ∇V, we can find an ellipsoid E centered around some point z such that $E \subset B_V(1) \subset E'$, where E' is E dilated by an $O(d^{3/2})$ factor about z.*

Proof. It suffices to show that from a single first-order query at a point x, we can generate a membership and separation oracle for $B_V(1)$. Indeed, the membership part is straightforward as we just check whether $V(x) \leq 1$ (which is equivalent to $x \in B_V(1)$). The separation oracle is also simple, and can be done using the gradient. Specifically, suppose that $V(x) > 1$; then, $V(x') \geq V(x) + \langle \nabla V(x), x' - x \rangle$, which means that every x' with $\langle \nabla V(x), x' \rangle \geq \langle \nabla V(x), x \rangle$ is such that $V(x') \geq V(x) > 1$, i.e., $\nabla V(x)$ is a separation oracle for $B_V(1)$ at x. □

We are able to prove our sampling upper bound, using a rejection sampling approach.

Theorem 35 (upper bound for log-concave sampling)**.** *For any constant $d \geq 2$ and any 1-strongly convex and κ-smooth function V with minimum at 0, we can approximately sample from $\pi \propto \exp(-V)$ to total variation error at most ε using $O(\log \kappa + \log^{O(1)}(1/\varepsilon))$ adaptive queries to V and ∇V (here we emphasize that the asymptotic notation treats d as constant).*

Proof. Given V and any integer $t \geq 1$, let p_t be the probability that a sample from π lies in $tB_V(1)$. The normalizing constant is $Z \geq \int_{B_V(1)} \exp(-V) \geq e^{-1} \operatorname{vol}(B_V(1))$, but integral over $(t+1)B_V(1) \backslash tB_V(1)$ is

$$\begin{aligned}\int_{(t+1)B_V(1)\backslash tB_V(1)} \exp(-V) &\leq \exp(-t) \operatorname{vol}\big((t+1)B_V(1)\big) \\ &= \exp(-t)\,(t+1)^d \operatorname{vol}\big(B_V(1)\big)\,,\end{aligned}$$

using Proposition 16 which implies that $V(x) \geq t$ for any $x \notin tB_V(1)$. Therefore, the probability of $(t+1)B_V(1)\backslash tB_V(1)$ under π is at most

$$\pi\big((t+1)B_V(1)\backslash tB_V(1)\big) \leq \frac{\exp(-t)\,(t+1)^d \operatorname{vol}(B_V(1))}{e^{-1}\operatorname{vol}(B_V(1))} = \exp(-(t-1))\,(t+1)^d\,.$$

By summing this quantity for all integers greater than t, the probability of the complement of $tB_V(1)$ is at most $\sum_{u \geq t} \exp(-(u-1))\,(u+1)^d = \sum_{u \geq t} \exp(-u + d\log(u+1)+1)$. Note that for $t \geq \Omega(d \log d)$, this quantity is at most $O(\exp(-t/2))$. Taking $t = C\,(d \log d + \log(1/\varepsilon))$ for a large constant C, we obtain $\pi(\mathbb{R}^d \backslash tB_V(1)) \leq \varepsilon/2$.

The algorithm now works as follows. We use Corollary 12 to find $E \subset B_V(1) \subset E'$. We pick a uniformly random point X in tE' for $t = C\,(d \log d + \log(1/\varepsilon))$. We then accept the point X with probability $\exp(-V(X))$, and if we reject we restart the procedure. First, note that this algorithm, upon termination, samples exactly from π conditioned on tE', which is at most $\frac{\varepsilon}{2}$ away from π in total variation distance. In addition, each rejection sampling step succeeds with probability at least $\operatorname{vol}(E)/(e \operatorname{vol}(tE'))$, since with probability $\operatorname{vol}(E)/\operatorname{vol}(tE')$ we choose a point in E in which case $V(X) \leq 1$ so we accept with probability at least e^{-1}. This is equal to $1/(t\,O(d^{3/2}))^d = d^{-O(d)}\,t^{-d} = d^{-O(d)}\,(\log \frac{1}{\varepsilon})^{-d}$. So, after $(d \log \frac{1}{\varepsilon})^{O(d)}$ rounds of rejection sampling, each of which only needs one query to V, we accept the sample with probability at least $1 - \frac{\varepsilon}{2}$, which means that overall we have generated a sample which is ε-close in distribution to π in total variation distance.

The overall query complexity is a combination of finding E, E' and then running the rejection sampling, for a total complexity of $d^{O(1)} \log \kappa + (d \log \frac{1}{\varepsilon})^{O(d)}$. So, for any fixed dimension d and error probability ε, the query complexity for log-concave sampling is $O(\log \kappa)$. In addition, the dependence on the error probability is polylogarithmic for any fixed d. □

Remark 9. We briefly note that the exponential dependence on d is not necessary: using more sophisticated tools developed for sampling from convex bodies one should be able to obtain a complexity of $\log(\kappa)\,(d \log \frac{1}{\varepsilon})^{O(1)}$. However, we choose to not optimize the dimension dependence in this result for the sake of simplicity, and since we are focused on the setting of $d = O(1)$.

10.5 Conclusion

The lower bound in Theorem 34 together with Theorem 31 in Chapter 9 completely characterize the complexity of log-concave sampling in constant dimensions. We see that in low dimensions, the optimal sampling algorithm is rejection sampling, which can thought of as the analogue of the ellipsoid method in optimization. In fact, the rejection sampling algorithm that achieves the rate from Theorem 34 even uses the ellipsoid method as a subroutine.

The next big question to answer is whether log-concave sampling in high dimensions behaves differently, and whether the high dimensional lower bound will be dimension dependent. We do not manage to obtain a complete answer to this question in this thesis, but in the next chapter we make some partial progress by studying lower bounds for Gaussian sampling.

Chapter 11

General lower bounds in high dimension: the Gaussian case

11.1 Introduction

In this final chapter, we give progress towards the goal of general sampling lower bounds in high dimension. We restrict the class of target distributions to the Gaussian distributions with mean 0, so the potential function is given by $V(x) = \langle x, \Sigma^{-1} x \rangle$, where Σ is the covariance matrix. The query oracle will return the value $V(x)$, as well as its gradient $\nabla V(x) = \Sigma^{-1} x$. We show that under such a model, any sampling algorithm would require at least $\Omega(\log d)$ queries. Since the class of Gaussian distributions are a subset of the log-concave and strongly log-smooth distributions, our results show that sampling on this extended class will also have dimension dependent rates. The true dimension dependence of log-concave sampling is more likely to be polynomial, and that remains an interesting open problem.

We give two lower bounds that cover two regimes where the condition number κ is large or small. We also give an upper bound of a sampling algorithm that samples from a Gaussian using $O(\min(\sqrt{\kappa} \log d, d))$ queries, for which our lower bounds are nearly tight. The techniques we use in this chapter are quite distinct from those in the previous Chapters 9 and 10: instead of explicitly constructing a set of hard-to-distinguish distributions, we reduce the Gaussian sampling task to trace estimation of the covariance matrix. It is interesting to see whether our explicit techniques from the previous two chapters can work for Gaussians or general log-concave distributions in high dimensions.

This chapter is based on the joint work [Che+23b], with Sinho Chewi, Jaume de Dois Pont, Jerry Li, and Shyam Narayanan.

Related works. Our upper bound for sampling from Gaussians (Theorem 41) is closely related to the use of the conjugate gradient algorithm for sampling from Gaussians [NS22]. Also, our $O(\log \kappa)$ upper bound algorithm is closely related to rounding procedures which have been previously used in the convex body sampling literature (see, e.g., [LV06]).

While matrix-vector queries have been studied in scientific computing for decades (e.g., [BFG96]), they have only been studied in the theoretical computer science literature recently, with a fully formalized model described in [Sun+19]. The most relevant works to ours are those that study the matrix-vector query complexity of spectral properties, such as estimating top eigenvectors [SAR18; Bra+20], trace and matrix norms [Hut90; WWZ14; RWZ20; DM21; Mey+21], the full eigenspectrum [Coh+18; BKM22], and low-rank approximation [MM15; BCW22]. We remark that the non-adaptive matrix-vector product model is closely related to sketching, which has enjoyed a large body of work (see, e.g., [Woo14] for a survey).

Our contributions. Our Gaussian lower bounds implies that when the dimension is sufficiently large, a polynomial dependence on the condition number κ is unavoidable (in contrast to Theorem 34, which only gives a logarithmic dependence on κ in low dimension). In fact, our lower bounds hold for the special case of sampling from Gaussians, for which they are nearly tight. We first prove the following theorem.

Theorem 36 (informal, see Corollary 13)**.** *Any sampler for centered d-dimensional Gaussians with condition number κ requires $\Omega(\min(\sqrt{\kappa}, d))$ queries.*

We emphasize the fact that in our setting, the Gaussians are centered. Note that if the Gaussians were allowed to have varying means, then one can deduce a sampling lower bound by reducing the optimization task of minimizing a convex quadratic function $x \mapsto \langle (x - x_\star), \Sigma^{-1} (x - x_\star) \rangle$ to the task of sampling from the corresponding Gaussian $\mathcal{N}(x_\star, \Sigma)$. However, as previously alluded to, this does not address the inherent difficulty of the sampling problem.

The proof of Theorem 36 rests upon an elegant technique developed in the literature on the matrix-vector query model in which the conditioning properties and sharp characterizations of the eigenvalue distribution of Wishart matrices are used to produce difficult lower bound instances for various tasks. We adapt this method to our context by reducing the task of inverse trace estimation to sampling (see Theorem 38).

As we show in Section 11.5, the lower bound is nearly tight over the class of Gaussians, as it is possible to sample from a Gaussian using $O(\min(\sqrt{\kappa} \log d, d))$ queries using the block Krylov method. However, note that the lower bound from Theorem 36 does not match the block Krylov upper bound, and the lower bound of Theorem 36 is vacuous when κ is constant. In particular, it leaves open the possibility that the complexity of sampling from well-conditioned Gaussians is dimension-free. While such dimension-free rates are possible in convex optimization, our next result shows that the same is in fact not possible for log-concave sampling:

Theorem 37. (informal, see Theorem 40) *Let d be sufficiently large, and let $\kappa \le d^{1/5-\delta}$. Then, any sampler for d-dimensional Gaussians with condition number κ requires $\Omega_\delta(\sqrt{\kappa}\log d)$ queries.*

In the regime for which Theorem 37 is valid, the lower bound matches the block Krylov upper bound up to constant factors, and hence we settle the complexity of sampling from Gaussians in this regime. Moreover, Theorems 36 and 37 together imply the *first* dimension-dependent lower bounds for general log-concave sampling. We conjecture that Theorem 37 holds for all κ for which $\sqrt{\kappa}\log d \le d$, and we leave this question for future work.

Although Theorem 37 may appear to only be a mild improvement over Theorem 36, analyzing this regime is quite delicate, and we believe that the tools based on Wishart matrices employed in the proof of Theorem 36 may be insufficient to reach Theorem 37. Instead, we prove Theorem 37 by first establishing sharp lower bounds on the performance of block Krylov algorithms for the sampling task, and then providing a novel reduction (Lemma 44) which shows that block Krylov algorithms are optimal for this task. This reduction is quite general, and as the block Krylov algorithm and the matrix-vector query model are of wide interest in scientific computing and numerical linear algebra, we believe that our reduction may be broadly useful for tackling other problems in this space.

We remark that a concise way of summarizing Theorems 36 and 37 if we do not care about lower order terms is that sampling from Gaussians requires $\widetilde{\Omega}(\min(\sqrt{\kappa}\log d, d))$ queries, where we write $f = \widetilde{\Omega}(g)$ to mean $f = \Omega(g\log^{-O(1)}(g))$.

11.2 Technical overview

We summarize the main technical ideas in the proofs in this section. For setails, see Section 11.3 for Theorem 36, and Section 11.4 for Theorem 37.

Recall that our goal is to provide a lower bound on sampling from a Gaussian $\mathcal{N}(0,\Sigma)$, where $\Lambda := \Sigma^{-1}$ has condition number κ. Note that the corresponding potential is $V(x) = \frac{1}{2}\langle x, \Lambda\, x\rangle$, and we are allowed zeroth-order and first-order queries, which means for a query x, we receive $x^\intercal \Lambda x$ and Λx. Hence, adaptive queries are equivalent to adaptive matrix-vector product computations with Λ.

The first observation we make is that we can reduce the problem of sampling from the Gaussian to estimating the trace of Σ. This is because if X is a sample from a distribution which is close in total variation distance to

$\mathcal{N}(0, \Sigma)$, then $\|X\|_2^2 \approx \operatorname{tr}(\Sigma)$ with high probability. Therefore, it suffices to demonstrate a lower bound for the following problem: given matrix-vector product computations with Λ, approximately compute $\operatorname{tr}(\Lambda^{-1})$.

Lower bound via Wishart matrices

For any d, let $W \in \mathbb{R}^{d\times d}$ have the $\mathsf{Wishart}(d)$ distribution. That is, $W = XX^\intercal$, where $X \in \mathbb{R}^{d\times d}$ has i.i.d. $\mathcal{N}(0, 1/d)$ entries. We take W to be the precision matrix, $\Lambda = W$. Our first lower bound shows that $\Omega(d)$ matrix-vector queries with W are necessary to estimate the trace of W^{-1} even to constant multiplicative accuracy, with constant success probability (Theorem 39). Since the condition number of W is $\Theta(d^2)$ with high probability, we obtain one extreme of the claimed lower bound $\Omega(\min(\sqrt{\kappa}, d))$. The general lower bound for all κ then follows from a padding argument.

This lower bound approach is inspired by [Bra+20], which proved a query lower bound for estimating the minimum eigenvalue of W. Their approach relies on the fact that if we condition on any sequence of $(1-\Omega(1))\,d$ adaptive queries, the posterior distribution of the remaining eigenvalues behaves similarly to the original distribution of the eigenvalues of W. In addition, while the smallest eigenvalue of W is usually about $1/d^2$, its distribution has heavy tails: with probability $\Theta(\sqrt{\varepsilon})$, the smallest eigenvalue of W is below ε/d^2. Consequently, even conditioned on $d/2$ adaptive queries, we are unable to learn the minimum eigenvalue up to a constant factor with high probability.

In our setting, we instead wish to show that learning the trace of W^{-1} is hard. However, the smallest eigenvalue of the Wishart matrix is so small that with high probability, $\operatorname{Tr}(W^{-1}) = \Theta(\lambda_{\min}(W)^{-1})$. While most of the time the trace is $O(d^2)$, with probability $\Theta(\sqrt{\varepsilon})$ the posterior distribution of the smallest eigenvalue of W after our adaptive queries may be ε/d^2. Hence, we will be unable to determine whether the trace is $\le O(d^2)$ or $\ge \Omega(d^2/\varepsilon)$ with high probability.

This lower bound technique is clean and nearly optimal, but as previously mentioned it is vacuous (of constant order) when $\kappa = O(1)$, whereas we expect the complexity of the problem to increase as $d \to \infty$. To tackle this setting, we introduce a second approach.

Lower bounds via reduction to block Krylov

Our second technique works in two parts. First, we show that for a specific hard distribution over instances, any block Krylov-style algorithm requires $\Omega(\min(\sqrt{\kappa}\log d, d))$ queries to estimate $\operatorname{tr}(\Sigma)$. Then, we show a general purpose reduction which demonstrates that for this hard instance (and indeed, any rotationally invariant instance), block Krylov methods are actually optimal.

Lower bound for block Krylov algorithms. Recall the block Krylov technique: the algorithm chooses K i.i.d. random vectors $v_1, \ldots, v_K \sim \mathcal{N}(0, I)$, and computes $\Lambda^j v_k$ for all $j \le T, k \le K$. This can be done using KT adaptive queries, by querying $\Lambda^j v_k$ to learn $\Lambda^{j+1} v_k$. For our purposes, it suffices to consider block Krylov algorithms with $K = T$ and to prove a lower bound on the smallest number K needed to successfully estimate $\mathrm{tr}(\Sigma)$, for $\Sigma = \Lambda^{-1}$.

We will construct two diagonal matrices D, D' with all eigenvalues between 1 and κ, such that $\mathrm{Tr}(D^{-1})$ and $\mathrm{Tr}((D')^{-1})$ are sufficiently different. In addition, if Λ, Λ' are random rotations of D, D', respectively, then $\{\Lambda^j v_k\}_{j,k \le K}$ and $\{(\Lambda')^j v_k\}_{j,k \le K}$ are hard to distinguish for $K \le c\sqrt{\kappa} \log d$ for a small constant c (Lemma 40). Thus, unless $K \ge \Omega(\sqrt{\kappa} \log d)$, we cannot estimate the trace.

To explain the intuition behind Lemma 40, we first consider what happens if we only have $\{\Lambda^j v\}_{j \le K}$ for a single random vector v (i.e., power method). Letting $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of Λ, we have $\Lambda^j v = \sum_{i=1}^d \lambda_i^j \alpha_i u_i$, where u_i is the i-th eigenvector of Λ and $v = \sum_{i=1}^d \alpha_i u_i$. Intuitively, the only information we obtain from these vectors are their pairwise inner products, since we could have randomly rotated Λ. Therefore, the only information we have is $\langle \Lambda^j v, \Lambda^{j'} v \rangle = \sum_{i=1}^d \lambda_i^{j+j'} \alpha_i^2$, which is the set $\{\sum_{i=1}^d \lambda_i^j \alpha_i^2\}_{j \le 2K}$. Since v is random, we may think of all of the α_i^2 as 1 for simplicity, and so we know $\{\sum_{i=1}^d \lambda_i^j\}_{j \le 2K}$. Our goal is to use this information to learn $\mathrm{Tr}(\Lambda^{-1}) = \sum_{i=1}^d \lambda_i^{-1}$.

We connect this to the problem of estimating $1/x$ as a linear combination of $1, x, x^2, \ldots, x^K$, a classic problem in approximation theory that is often tackled with *Chebyshev polynomials*. Indeed, this relation to Chebyshev polynomials is the main tool in the analysis of essentially all Krylov methods. In our setting, as we desire lower bounds, we apply the fact that Chebyshev polynomials are *optimal* in generating certain approximations. More concretely, suppose that there are only K distinct eigenvalues $\lambda_1, \ldots, \lambda_K$, with each λ_i having some multiplicity N_i. Since we want to show that estimating $\mathrm{tr}(\Lambda^{-1})$ is hard, this amounts to showing that knowing $\sum_{i=1}^K N_i \lambda_i^j$ for $0 \le j \le K$ is insufficient to learn $\sum_{i=1}^K N_i/\lambda_i$. We express this as a linear program (if we relax the N_i to be reals), the dual of which precisely captures whether $1/x$ can be approximated well by a degree-K polynomial at $\lambda_1, \ldots, \lambda_K$ (Proposition 24). If we choose the λ_i to be the local extrema of a degree-K Chebyshev polynomial, shifted so that $\lambda_1 = 1$ and $\lambda_K = \kappa$, then it is known that one cannot estimate $1/x$ up to error $d^{-\Omega(1)}$ at these points (which is needed for trace estimation), unless $K \ge \Omega(\sqrt{\kappa} \log d)$. At a high level, this is the reason why we need $\Omega(\sqrt{\kappa} \log d)$ iterations of power method.

For general block Krylov algorithms, the algorithm obtains $\langle v_\ell, \Lambda^j v_k \rangle$, for $0 \le j \le K$ and $1 \le k, \ell \le K$. Now, the information that the algorithm sees is captured by the matrices $\{\langle v_\ell, \Lambda^j v_k \rangle\}_{k,\ell \le K}$, for $j = 1, \ldots, K$. Here, we show that provided K is sufficiently small compared to d, we can find choices of multiplicities $N_1, \ldots, N_K$ and $N'_1, \ldots, N'_K$, such that the corresponding matrices D, D' have significantly different traces (i.e., $\sum_{i=1}^{K}(N_i - N'_i)/\lambda_i$ is large) but the information from queries is not enough to distinguish between Λ and Λ', which we establish via a coupling argument.

Reduction to block Krylov algorithms. The argument outlined above shows block Krylov algorithms with $K = o(\sqrt{\kappa} \log d)$ cannot distinguish between two families of randomly rotated matrices with difference traces (Λ coming from D and Λ' coming from D'), and hence cannot solve the trace estimation task. Our next technical contribution is a reduction which allows us to simulate the output of any *adaptive* algorithm with K queries on our hard instance, given only the responses to a block Krylov algorithm. Thus, a lower bound against block Krylov methods translates into a lower bound against any query algorithm. We now give a high-level description of the reduction.

Since we prove lower bounds based on randomized constructions, it suffices to consider adaptive deterministic algorithms, i.e., each query v_k is a deterministic function of the previous queries and oracle outputs. The difficulty of proving such a lower bound against such an algorithm is the adaptivity of the queries, which makes it difficult to reason about how much information the algorithm has learned. However, since our lower bound construction for block Krylov algorithms is rotationally invariant, intuitively the adaptivity does not help: the algorithm may as well query a random direction which it has not yet explored.

However, this intuition is not entirely correct: if the algorithm has previously queried a vector v and received the information Λv, then it may useful to query Λv in order to receive the information $\Lambda^2 v$, instead of querying a completely random new direction. Indeed, computing powers $v, \Lambda v, \Lambda^2 v, \ldots$ is precisely the essence of the power method, as discussed above. To account for this, we move to the following stronger oracle model: if the algorithm has selected vectors $v_1, \ldots, v_k$, then at iteration k it receives all of the information $(\Lambda^i v_j)_{i+j \le k}$ for free. Now, there is provably no benefit to querying vectors which lie in the span of the previous queries and oracle outputs.

Recall that our goal is to argue that an adaptive deterministic algorithm can be simulated by an algorithm which simply makes i.i.d. Gaussian queries $z_1, z_2, \ldots, z_K$, in the following sense. In the stronger oracle model, at iteration k, the adaptive algorithm has made queries $(v_1^{\mathsf{alg}}, \ldots, v_k^{\mathsf{alg}})$ and received information $(\Lambda^i v_j^{\mathsf{alg}})_{i+j \le k}$ and it picks a new vector v_{k+1} which lies orthogonal to its received information. Suppose that using only the Gaussian queries $z_1, z_2, \ldots, z_k$, we have simulated queries $v_1^{\mathsf{sim}}, v_2^{\mathsf{sim}}, \ldots, v_k^{\mathsf{sim}}$ which are equivalent to the execution of the adaptive algorithm in the sense that the law of the information $(\Lambda^i v_j^{\mathsf{sim}})_{i+j \le k}$ is precisely the same as the law of the algorithm's information $(\Lambda^i v_j^{\mathsf{alg}})_{i+j \le k}$. Since the algorithm is deterministic, v_k^{alg} is a function $v_k((\Lambda^i v_j^{\mathsf{alg}})_{i+j<k})$ of algorithm's accumulated information. Thus, in order

to simulate the adaptive algorithm for one more step, it is natural to consider taking $v_k^{\mathsf{sim}} := v_k((\Lambda^i v_j^{\mathsf{sim}})_{i+j<k})$. However, we will be unable to compute $\Lambda^i v_k^{\mathsf{sim}}$ for any $i \geq 1$, because the simulation must be based on the Gaussian queries $z_1, z_2, \dots, z_k$, whereas this definition of v_k^{sim} requires making queries at $v_1^{\mathsf{sim}}, v_2^{\mathsf{sim}}, \dots, v_{k-1}^{\mathsf{sim}}$.

Thus far, we have not invoked the rotational invariance of Λ, which is crucial to the argument. The key is that although we cannot directly take $v_k((\Lambda^i v_j^{\mathsf{sim}})_{i+j<k})$ to be our next simulated point, we can *rotate* $\tilde{v}_k$ into $v_k((\Lambda^i v_j^{\mathsf{sim}})_{i+j<k})$ via a unitary matrix U_k; moreover, we can arrange that U_k fixes all of the previous information $(\Lambda^i v_j^{\mathsf{sim}})_{i+j<k}$, because $v_k((\Lambda^i v_j^{\mathsf{sim}})_{i+j<k})$ lies orthogonal to this information (recall, we can assume that each deterministic function $v_k(\cdot)$ outputs a vector orthogonal to its inputs, due to our choice of oracle model). The intuition is that due to the rotational invariance of Λ, then conditioned on the data $(\Lambda^i v_j^{\mathsf{sim}})_{i+j<k}$, the distribution of Λ is still rotationally invariant on the orthogonal subspace of the data; hence, $U_k \tilde{v}_k = v_k((\Lambda^i v_j^{\mathsf{sim}})_{i+j<k})$ ought to have the same law as $\tilde{v}_k$, i.e., querying the completely random direction $\tilde{v}_k$ is just as good as querying according to what the adaptive algorithm specifies.

Unfortunately there are further difficulties to overcome with this approach. Namely, suppose that we define each simulated point v_k^{sim} to be the output $U_k \tilde{v}_k$ of a rotation matrix applied to $\tilde{v}_k$. We would like to take U_k such that $U_k \tilde{v}_k = v_k((\Lambda^i v_j^{\mathsf{sim}})_{i+j<k})$ but this is no longer computable based on $(\Lambda^i \tilde{v}_j)_{i+j<k}$. However, we note that $\Lambda^i v_j^{\mathsf{sim}} = \Lambda^i U_j \tilde{v}_j = U_j \tilde{\Lambda}^i \tilde{v}_j$ where $\tilde{\Lambda} := U_j^{\mathsf{T}} \Lambda U_j$. This shows that $\Lambda^i v_j^{\mathsf{sim}}$ is computed from the query of $\tilde{v}_j$, not on the original matrix Λ but on the modified matrix $\tilde{\Lambda}$, together with the matrix U_j. Since we hope that $\tilde{\Lambda}$ has the same law as Λ, then this is good enough for the purposes of simulating the adaptive algorithm. Actually, in order for the induction to work out, it becomes clear that we need to define a sequence of matrices $\Lambda_1, \Lambda_2, \dots, \Lambda_k$, where each Λ_k is related to the previous Λ_{k-1} via $\Lambda_k = U_k^{\mathsf{T}} \Lambda_{k-1} U_k$, and U_k is chosen such that $v_k^{\mathsf{sim}} = U_k \tilde{v}_k = v_k((\Lambda_{k-1}^i v_j^{\mathsf{sim}})_{i+j<k})$. Then, we must argue that the simulated sequence $v_1^{\mathsf{sim}}, v_2^{\mathsf{sim}}, \dots, v_k^{\mathsf{sim}}$ has the same law as the algorithm's sequence $v_1^{\mathsf{alg}}, v_2^{\mathsf{alg}}, \dots, v_k^{\mathsf{alg}}$.

This last step, however, turns out to be delicate. Indeed, although it is obvious that for a *fixed* orthogonal matrix U', the law of Λ is the same as the law of $(U')^{\mathsf{T}} \Lambda U'$, the rotation matrices U_k we choose in the above argument are dependent on the previous queries and oracle outputs, and are hence dependent on Λ itself. In the presence of such dependence, it is not obvious why the law of Λ_k should be the same as the law of Λ, and to address this we prove a conditioning lemma in Section 11.4. Once the conditioning lemma is proved, the remainder of the proof follows along the lines just described, and the details of the induction are carried out in Section 11.4.

11.3 A lower bound via Wishart matrices

We define $W \sim \mathsf{Wishart}(d)$ to mean $W = XX^\intercal$ where each entry of $X \in \mathbb{R}^{d\times d}$ is $\mathcal{N}(0, \frac{1}{d})$. We aim to prove the following two theorems, which together imply a query complexity lower bound for sampling from Gaussians.

Theorem 38 (reducing inverse trace estimation to sampling)**.** *Let $\delta > 0$. There is a universal constant $c > 0$ (depending only on δ) such that the following hold. Suppose that $d \ge c^{-1}$ and there exists a query algorithm such that, for any Gaussian target distribution $\pi := \mathcal{N}(0, \Sigma)$ in $\mathbb{R}^d$ with $cd^{-2}\, I_d \preceq \Sigma^{-1} \preceq c^{-1}\, I_d$, outputs a sample from a distribution $\widehat{\pi}$ such that either $\|\widehat{\pi} - \pi\|_{\mathrm{TV}} \le c$ or $\sqrt{cd^{-2}}\, W_2(\widehat{\pi}, \pi) \le c$, using n queries to π.*

Then, given $W \sim \mathsf{Wishart}(d)$, there exists an algorithm which makes at most $c^{-1}n$ matrix-vector queries to W and outputs an estimator $\widehat{\mathrm{tr}}$ such that $\frac{1}{2}\operatorname{tr}(W^{-1}) \le \widehat{\mathrm{tr}} \le 2\operatorname{tr}(W^{-1})$ with probability at least $1 - \delta$.

Theorem 39 (lower bound for inverse trace estimation)**.** *Let $W \sim \mathsf{Wishart}(d)$ for $d \ge 2$. For any $C > 0$, there exists $\delta > 0$ (depending only on C) such that any algorithm which makes n matrix-vector queries to W and outputs an estimator $\widehat{\mathrm{tr}}$ such that $C^{-1}\operatorname{tr}(W^{-1}) \le \widehat{\mathrm{tr}} \le C\operatorname{tr}(W^{-1})$ with probability at least $1 - \delta$ must use $n \ge \Omega(d)$ queries.*

Remark 10. Suppose that we want to sample from a target distribution π which is α-strongly log-concave. It is straightforward to check that total variation guarantees are invariant under rescaling the target (replacing π with $S_\# \pi$, where $S : \mathbb{R}^d \to \mathbb{R}^d$ is the scaling map $Sx := \zeta x$ for some $\zeta > 0$), whereas Wasserstein guarantees are not. Instead, the scale-invariant quantity is $\sqrt{\alpha}\, W_2$, which is what appears in Theorem 38.

Consider the class of centered Gaussian distributions on $\mathbb{R}^d$ which are α-strongly log-concave and β-log-smooth; let $\kappa := \beta/\alpha$ denote the condition number. Let $\mathscr{C}_{\mathsf{G},\mathsf{d}}(\kappa, d, \varepsilon)$ denote the query complexity of outputting a sample which is ε-close in the metric d to a target distribution in this class, where d is one of the scale-invariant distances $\mathsf{d} \in \{\mathrm{TV}, \sqrt{\alpha}\, W_2\}$. Then, Theorems 38 and 39 (with $C = 2$ and δ, c being universal constants) show that for $d \ge c^{-1}$,

$$\mathscr{C}_{\mathsf{G},\mathsf{d}}(c^{-2}d^2, d, c) \ge \Omega(d)\,. \tag{11.1}$$

By embedding the construction into higher dimensions, we obtain the following corollary.

Corollary 13 (query lower bound via Wishart matrices)**.** *For $\mathsf{d} \in \{\mathrm{TV}, \sqrt{\alpha}\, W_2\}$, there is a universal constant $c > 0$ such that*

$$\mathscr{C}_{\mathsf{G},\mathsf{d}}(\kappa, d, c) \ge \Omega\bigl(\sqrt{\kappa} \wedge d\bigr)\,.$$

Proof. If $\kappa \ge c^{-2}d^2$, then (11.1) yields

$$\mathscr{C}_{\mathsf{G},\mathsf{d}}(\kappa, d, c) \ge \Omega(d) \ge \Omega\bigl(\sqrt{\kappa} \wedge d\bigr)\,.$$

Otherwise, if $\kappa \le c^{-2}d^2$, let $d_\star$ be the largest integer such that $\kappa \ge c^{-2}d_\star^2$. Then, by embedding the $d_\star$-dimensional construction into dimension d,

$$\mathscr{C}_{\mathsf{G},\mathsf{d}}(\kappa, d, c) \ge \mathscr{C}_{\mathsf{G},\mathsf{d}}(\kappa, d_\star, c) \ge \Omega(d_\star) \ge \Omega\bigl(\sqrt{\kappa} \wedge d\bigr)\,,$$

which concludes the proof. □

Reducing inverse trace estimation to sampling

In this section, we prove Theorem 38, which is based on the concentration of the squared norm of a Gaussian. We recall the following identity:

Lemma 37 (concentration of the squared norm)**.** *Let* $Z \sim \mathcal{N}(0, \Sigma)$. *Then,*

$$\operatorname{var}(\|Z\|^2) = 2\,\|\Sigma\|_{\mathrm{HS}}^2\,.$$

Proof. Note that since all quantities are rotationally invariant, we may assume without loss of generality that Σ is diagonal. Then the equality claimed is just the variance of a non-homogenous chi-squared random variable. □

We now prove Theorem 38.

Proof of Theorem 38. Let $W \sim \mathsf{Wishart}(d)$ and let $\Sigma := W^{-1}$. By Proposition 20, there exists $c > 0$ (depending only on δ) such that with probability at least $1 - \delta/3$, it holds that

$$cd^{-2}\, I_d \preceq \Sigma^{-1} \preceq c^{-1}\, I_d\,.$$

We work on the event $\mathcal{E}$ that this holds.

Case 1: total variation distance. From Lemma 37 and Chebyshev's inequality, we deduce that if $Z_1, \dots, Z_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$ and $\widehat{\operatorname{tr}}_\star := m^{-1} \sum_{i=1}^m \|Z_i\|^2$,

$$\mathbb{P}\bigl\{\bigl|\widehat{\operatorname{tr}}_\star - \operatorname{tr}\Sigma\bigr| \ge \frac{1}{2}\operatorname{tr}\Sigma\bigr\} \le \frac{\operatorname{var}\widehat{\operatorname{tr}}_\star}{(\operatorname{tr}\Sigma)^2/4} = \frac{8}{m}\cdot\frac{\operatorname{tr}(\Sigma^2)}{\operatorname{tr}(\Sigma)^2} \le \frac{8}{m}\,.$$

Take $m \ge 48/\delta$ so that this probability is at most $\delta/3$. Conditionally on W, let $\widehat{\pi}_W$ denote the law of the sample X of the algorithm when run on the target $\mathcal{N}(0, \Sigma)$. By running the sampling algorithm m times, we can obtain i.i.d. samples $X_1, \dots, X_m \overset{\text{i.i.d.}}{\sim} \widehat{\pi}_W$. Then, for $\widehat{\operatorname{tr}} := m^{-1} \sum_{i=1}^m \|X_i\|^2$,

$$\mathbb{P}\bigl\{\bigl|\widehat{\operatorname{tr}} - \operatorname{tr}\Sigma\bigr| \ge \frac{1}{2}\operatorname{tr}\Sigma\bigr\} \le \mathbb{P}(\mathcal{E}^{\mathrm{c}}) + \mathbb{P}\bigl\{\bigl|\widehat{\operatorname{tr}} - \operatorname{tr}\Sigma\bigr| \ge \frac{1}{2}\operatorname{tr}\Sigma,\ \mathcal{E}\bigr\}$$

$$\begin{aligned}
&\le \frac{\delta}{3} + \mathbb{E}\bigl[\mathbb{P}\bigl\{\bigl|\widehat{\mathrm{tr}}_\star - \operatorname{tr}\Sigma\bigr| \ge \frac{1}{2}\operatorname{tr}\Sigma \bigm| W\bigr\}\,\mathbb{1}_{\mathcal{E}}\bigr] + \mathbb{E}\bigl[\|\widehat{\pi}_W^{\otimes m} - \mathcal{N}(0,\Sigma)^{\otimes m}\|_{\mathrm{TV}}\,\mathbb{1}_{\mathcal{E}}\bigr] \\
&\le \frac{\delta}{3} + \frac{\delta}{3} + cm\,.
\end{aligned}$$

If we choose $c \le \delta/(3m)$, then $\widehat{\mathrm{tr}}$ is an estimator of $\operatorname{tr}(W^{-1})$ with multiplicative error at most 2 which succeeds with probability at least $1-\delta$. Note that both c and m depend only on δ.

Case 2: Wasserstein distance. Consider a coupling of X and Z such that, conditionally on W, we have $\mathbb{E}[\|X-Z\|^2 \mid W] = \mathbb{E}[W_2^2(\widehat{\pi}_W, \mathcal{N}(0,\Sigma)) \mid W]$. Let $(X_1,Z_1),\dots,(X_m,Z_m)$ be i.i.d. copies of this coupling. Also, let $\mathcal{E}'$ denote the event that $\lambda_{\min}(W^{-1}) \ge \bar{c}d^2$, where $\bar{c}$ is a constant depending only on δ, chosen so that $\mathbb{P}(\mathcal{E}'^{\mathsf{c}}) \le \delta/3$ using Proposition 20. Then, conditionally on W in the event $\mathcal{E} \cap \mathcal{E}'$,

$$\begin{aligned}
\mathbb{E}\bigl[|\widehat{\mathrm{tr}} - \operatorname{tr}\Sigma| \bigm| W\bigr] &\le \mathbb{E}\bigl[|\widehat{\mathrm{tr}} - \widehat{\mathrm{tr}}_\star| \bigm| W\bigr] + \mathbb{E}\bigl[|\widehat{\mathrm{tr}}_\star - \operatorname{tr}\Sigma| \bigm| W\bigr] \\
&\le \mathbb{E}\bigl[|\widehat{\mathrm{tr}} - \widehat{\mathrm{tr}}_\star| \bigm| W\bigr] + \frac{2\operatorname{tr}\Sigma}{\sqrt{m}}\,,
\end{aligned}$$

where we used Lemma 37. Using $\|x\|^2 - \|y\|^2 = \langle x-y, x+y\rangle$, for any $\lambda > 0$,

$$\begin{aligned}
\mathbb{E}\bigl[|\widehat{\mathrm{tr}} - \widehat{\mathrm{tr}}_\star| \bigm| W\bigr] &\le \mathbb{E}\bigl[|\,\|X\|^2 - \|Z\|^2\,| \bigm| W\bigr] \\
&\le \mathbb{E}\bigl[\|X-Z\|^2 \bigm| W\bigr] + 2\,\mathbb{E}\bigl[|\langle X-Z, Z\rangle| \bigm| W\bigr] \\
&\le (1+\lambda)\,\mathbb{E}\bigl[\|X-Z\|^2 \bigm| W\bigr] + \frac{1}{\lambda}\,\mathbb{E}\bigl[\|Z\|^2 \bigm| W\bigr] \\
&\le (1+\lambda)\,c^3 d^2 + \frac{\operatorname{tr}\Sigma}{\lambda} \le (1+\lambda)\,\frac{c^3}{\bar{c}}\operatorname{tr}\Sigma + \frac{\operatorname{tr}\Sigma}{\lambda}\,.
\end{aligned}$$

If we take $\lambda = 18/\delta$, $m \ge (36/\delta)^2$, and if c is sufficiently small (depending only on δ), we obtain

$$\mathbb{E}\bigl[|\widehat{\mathrm{tr}} - \operatorname{tr}\Sigma| \bigm| W\bigr] \le \frac{\delta\operatorname{tr}\Sigma}{6}\,.$$

By Markov's inequality,

$$\begin{aligned}
\mathbb{P}\bigl\{\bigl|\widehat{\mathrm{tr}} - \operatorname{tr}\Sigma\bigr| \ge \frac{1}{2}\operatorname{tr}\Sigma\bigr\} &\le \mathbb{P}(\mathcal{E}^{\mathsf{c}}) + \mathbb{P}(\mathcal{E}'^{\mathsf{c}}) + \mathbb{E}\bigl[\mathbb{P}\bigl\{\bigl|\widehat{\mathrm{tr}} - \operatorname{tr}\Sigma\bigr| \ge \frac{1}{2}\operatorname{tr}\Sigma \bigm| W\bigr\}\,\mathbb{1}_{\mathcal{E}\cap\mathcal{E}'}\bigr] \\
&\le \frac{\delta}{3} + \frac{\delta}{3} + \frac{\delta}{3} \le \delta\,.
\end{aligned}$$

We conclude as before. □

Lower bound for inverse trace estimation

In this section, we prove Theorem 39. The idea is that due to the heavy tails of $\lambda_{\min}(W^{-1})$ implied by Proposition 20, with some small probability δ, $\operatorname{tr}(W^{-1})$ will be very large. An algorithm for inverse trace estimation which succeeds with probability at least $1-\delta$ must be able to detect this event, and we show that this requires making $\Omega(d)$ queries.

The key technical tools are the following propositions, due to [Bra+20].

Proposition 18 ([Bra+20, Lemma 3.4]). *Let $W \sim \mathsf{Wishart}(d)$. Then, for any sequence of $n < d$ (possibly adaptive) queries $v_1, \dots, v_n$ and responses $w_1 = Wv_1, \dots, w_n = Wv_n$, there exists an orthogonal matrix $V \in \mathbb{R}^{d\times d}$ and matrices $Y_1 \in \mathbb{R}^{n\times n}, Y_2 \in \mathbb{R}^{(d-n)\times n}$ that only depend on $v_1, \dots, v_n, w_1, \dots, w_n$, such that $VWV^\intercal$ has the block form*

$$VWV^\intercal = \begin{bmatrix} Y_1Y_1^\intercal & Y_1Y_2^\intercal \\ Y_2Y_1^\intercal & Y_2Y_2^\intercal + \widetilde{W} \end{bmatrix}.$$

Here, conditionally on $v_1, \dots, v_n, w_1, \dots, w_n$, the matrix $\widetilde{W}$ has the $\mathsf{Wishart}(d-n)$ distribution.

Proposition 19 ([Bra+20, Lemma 3.5]). *For any matrices $Y_1 \in \mathbb{R}^{n\times n}$, $Y_2 \in \mathbb{R}^{(d-n)\times n}$, and any symmetric matrix $\widetilde{W} \in \mathbb{R}^{(d-n)\times(d-n)}$, it holds that*

$$\lambda_{\min}\Bigl(\begin{bmatrix} Y_1Y_1^\intercal & Y_1Y_2^\intercal \\ Y_2Y_1^\intercal & Y_2Y_2^\intercal + \widetilde{W} \end{bmatrix}\Bigr) \le \lambda_{\min}(\widetilde{W}).$$

We are now ready to prove Theorem 39. Note that this result is very similar to that of [Bra+20], except that we work with the inverse trace rather than the minimum eigenvalue.

Proof of Theorem 39. Let $\delta > 0$ be chosen later. We first argue that $\widehat{\operatorname{tr}}$ must not be too large. Applying Proposition 21, we conclude that there is a universal constant $C' > 0$ such that $\operatorname{tr}(W^{-1}) \le C'd^2$ with probability at least $1/2$. Hence,

$$\begin{aligned} \mathbb{P}\bigl\{\widehat{\operatorname{tr}} \le CC'd^2\bigr\} &\ge \mathbb{P}\bigl\{\operatorname{tr}(W^{-1}) \le C'd^2 \text{ and } \widehat{\operatorname{tr}} \le C\operatorname{tr}(W^{-1})\bigr\} \\ &\ge \mathbb{P}\{\operatorname{tr}(W^{-1}) \le C'd^2\} - \mathbb{P}\{\widehat{\operatorname{tr}} > C\operatorname{tr}(W^{-1})\} \ge \frac{1}{2} - \delta. \end{aligned}$$

Next, suppose for the sake of contradiction that $n \le d/2$. Let $\mathscr{F}_n$ denote the σ-algebra generated by the information available to the algorithm up to iteration n, that is, the queries $v_1, \dots, v_n$, the responses $w_1, \dots, w_n$, and any external randomness used by the algorithm (which is independent of W). Applying Propositions 18 and 19,

$$\mathbb{P}\bigl\{\widehat{\operatorname{tr}} < C^{-1}\operatorname{tr}(W^{-1})\bigr\} \ge \mathbb{P}\bigl\{\widehat{\operatorname{tr}} \le CC'd^2 \text{ and } \lambda_{\min}(W^{-1}) > C^2C'd^2\bigr\}$$

$$
\begin{aligned}
&\geq \mathbb{P}\bigl\{\widehat{\mathrm{tr}} \leq CC'd^2 \text{ and } \lambda_{\min}(\widetilde{W}^{-1}) > C^2C'\,d^2\bigr\} \\
&= \mathbb{E}\bigl[\mathbb{1}\{\widehat{\mathrm{tr}} \leq CC'd^2\}\,\mathbb{P}\{\lambda_{\min}(\widetilde{W}^{-1}) \geq C^2C'd^2 \mid \mathscr{F}_n\}\bigr]\,.
\end{aligned}
$$

According to Proposition 18, conditionally on $\mathscr{F}_n$, $\widetilde{W}$ has the $\mathsf{Wishart}(d-n)$ distribution. By applying Proposition 20,

$$
\mathbb{P}\{\lambda_{\min}(\widetilde{W}^{-1}) \geq C^2C'd^2 \mid \mathscr{F}_n\} \geq \mathbb{P}\{\lambda_{\min}(\widetilde{W}^{-1}) \geq 4C^2C'\,(d-n)^2 \mid \mathscr{F}_n\} \gtrsim \frac{1}{C\sqrt{C'}}\,.
$$

Therefore,

$$
\mathbb{P}\bigl\{\widehat{\mathrm{tr}} < C^{-1}\,\mathrm{tr}(W^{-1})\bigr\} \gtrsim \mathbb{P}\bigl\{\widehat{\mathrm{tr}} \leq CC'd^2\bigr\}\,\frac{1}{C\sqrt{C'}} \geq \frac{1/2-\delta}{C\sqrt{C'}}\,,
$$

which is larger than δ provided that δ is chosen sufficiently small (depending only on C). This contradicts the success probability of the algorithm, and hence we deduce that $n \geq d/2$. □

Useful facts about Wishart matrices

We collect together useful facts about Wishart matrices which are used in the proofs.

Proposition 20 (extreme singular values of a Gaussian matrix)**.** *Let $W \sim \mathsf{Wishart}(d)$. For any $x \in [0,1]$,*

$$
\mathbb{P}\bigl\{\lambda_{\min}(W) \leq \frac{x}{d^2}\bigr\} \asymp \sqrt{x}\,.
$$

Also, there is a universal constant $C > 0$ such that

$$
\mathbb{P}\{\lambda_{\max}(W) \geq C\,(1+t)\} \leq 2\exp(-dt)\,.
$$

Proof. See, e.g., [Ede89, Theorem 5.1] and [Ver18, Theorem 4.4.5]. □

Proposition 21 (bound on the inverse trace)**.** *Let $W \sim \mathsf{Wishart}(d)$. Then, for any $\delta > 0$, with probability at least $1-\delta$, it holds that $\mathrm{tr}(W^{-1}) \leq C_\delta d^2$ where C_δ is a constant depending only on δ.*

Proof. According to [Sza91, Theorem 1.2], there is a universal constant $C > 0$ such that for each $j = 1,\dots,d$ and $\alpha \geq 0$,

$$
\mathbb{P}\Bigl\{\frac{1}{\lambda_j(W)} \geq \frac{d^2}{\alpha^2 j^2}\Bigr\} \leq (C\alpha)^{j^2}\,.
$$

Let $\alpha < 1/C$ and let $E_\alpha \coloneqq \{1/\lambda_j(W) \ge d^2/(\alpha^2 j^2)$ for some $j = 1, \dots, d\}$. By the union bound,

$$\mathbb{P}(E_\alpha) \le \sum_{j=1}^{d} (C\alpha)^{j^2} \lesssim \frac{1}{\sqrt{\log(1/(C\alpha))}}\,.$$

On the event E_α^{c},

$$\operatorname{tr}(W^{-1}) \le \sum_{j=1}^{d} \frac{d^2}{\alpha^2 j^2} = \frac{\pi^2 d^2}{6\alpha^2}\,,$$

which is the claimed result upon taking α sufficiently small. □

Remark 11. The proof only shows that $\mathbb{P}\{\operatorname{tr}(W^{-1}) \ge \eta d^2\} \lesssim 1/\sqrt{\log \eta}$ for $\eta \gg 1$, which is not enough to conclude that $\mathbb{E}\operatorname{tr}(W^{-1})$ is finite. In fact, it holds that $\mathbb{E}\operatorname{tr}(W^{-1}) = \infty$, which can already be seen from Proposition 20.

11.4 A lower bound via reduction to block Krylov

In this section, we prove Theorem 37. Our proof procedes in two parts: we first show a lower bound against the block Krylov method, and then a reduction showing that an arbitrary adaptive algorithm can be simulated via a block Krylov method.

Preliminaries

We first record some important facts that we will use later on. The following is a standard approximation-theoretic result:

Proposition 22 ([SV14, Proposition 2.4, rephrased])**.** *Let T_K be the degree-K Chebyshev polynomial, and let $1 = \beta_1 > \cdots > \beta_{K+1} = -1$ be the set of real values β such that $T_K(\beta) \in \{-1, 1\}$. Then, for any real degree-K polynomial p such that $|p(\beta_i)| \le 1$ for all β_i, we have $|p(x)| \le |T_K(x)| \le (|x| + \sqrt{x^2 - 1})^K$ for all $|x| > 1$.*

Let $c_0 > 0$ be a constant to be chosen later. The above proposition immediately implies:

Corollary 14 (approximation error)**.** *Suppose that $K \le c_0\sqrt{\kappa}\log d$. Then, there exist $\kappa = \lambda_1 > \cdots > \lambda_{K+2} = 1$ (that only depend on K and κ) such that for any real degree-K polynomial P, $\max_{1 \le i \le K+2} |\frac{1}{\lambda_i} - P(\lambda_i)| \ge d^{-2c_0 - O(1/\sqrt{\kappa})}/\kappa$.*

Proof. Set $\beta_1, \dots, \beta_{K+2}$ to be the solutions of $T_{K+1} \in \{-1, 1\}$, and for each $1 \le i \le K+2$, set $\lambda_i \coloneqq \frac{(\kappa-1)}{2}(\beta_i + 1) + 1$; by construction, $\kappa = \lambda_1 > \cdots > \lambda_{K+2} = 1$. Given any polynomial Q of degree at most $K+1$, note that if $|Q(\lambda_i)| \le 1$ for all i, then the polynomial p given by $p(x) \coloneqq Q(\frac{\kappa-1}{2}(x+1)+1)$ satisfies $|p(\beta_i)| \le 1$ for all i. By Proposition 22, for $x_0 \coloneqq -(1 + \frac{2}{\kappa-1})$,

$$|Q(0)| = |p(x_0)| \le \bigl(|x_0| + \sqrt{x_0^2 - 1}\bigr)^{K+1} \le \Bigl(1 + \frac{2}{\sqrt{\kappa}} + O\bigl(\frac{1}{\kappa}\bigr)\Bigr)^{K+1}$$

$$< \exp\Bigl(\bigl(\frac{2}{\sqrt{\kappa}} + O\bigl(\frac{1}{\kappa}\bigr)\bigr)\,\bigl(c_0\sqrt{\kappa}\log d + 1\bigr)\Bigr) = d^{2c_0 + O(1/\sqrt{\kappa})}\,.$$

Next, for a degree-K polynomial P, consider $Q(x) \coloneqq d^{2c_0+O(1/\sqrt{\kappa})}\,(1 - xP(x))$. Note that Q has degree $K+1$ and $|Q(0)| = d^{2c_0+O(1/\sqrt{\kappa})}$, which implies that $|Q(\lambda_i)| > 1$ for some i, which in turn shows that $|\frac{1}{\lambda_i} - P(\lambda_i)| \ge d^{-2c_0-O(1/\sqrt{\kappa})}/\kappa$. □

We also introduce random matrix ensembles that are used in the proof, together with basic facts and properties.

Interestingly, as in the previous section, Wishart matrices are also useful for understanding block Krylov algorithms, but for a completely different reason. This time, we will study inner products between random vectors, which is also captured by a Wishart matrix. We denote by $\mathsf{Wishart}(K, N)$ the law of the random matrix $XX^\intercal \in \mathbb{R}^{K\times K}$, where the entries of $X \in \mathbb{R}^{K\times N}$ are i.i.d. *standard* Gaussians. Note that this is a different convention from the previous section, in which each entry of X was i.i.d. $\mathcal{N}(0, \frac{1}{d})$.

We also define the Gaussian orthogonal ensemble (GOE) of size K, denoted $\mathsf{GOE}(K)$. This is the law of a random symmetric matrix $G \in \mathbb{R}^{K\times K}$ where each diagonal entry $G_{i,i}$ is distributed as $\mathcal{N}(0,1)$, and each off-diagonal entry $G_{i,j} = G_{j,i}$ is distributed as $\mathcal{N}(0, \frac{1}{2})$. Also, the entries $\{G_{i,j} : 1 \le i \le K,\ j \le i\}$ are independent.

A long line of work (see, e.g., [JL15; Bub+16; BG18; RR19; BBH21; Mik22]) shows that when $N \gg K^3$, the Wishart ensemble is well-approximated by a scaled and shifted GOE, a fact which we shall invoke in the sequel.

Lemma 38 (equivalence of Wishart and GOE)**.** *Let $W \sim \mathsf{Wishart}(K, N)$ be drawn from the Wishart distribution, and let W_0 be drawn from the distribution of symmetric matrices where the diagonal and above-diagonal entries are mutually independent, each diagonal entry is drawn as $\mathcal{N}(N, 2N)$, and each above-diagonal entry is drawn as $\mathcal{N}(0, N)$. (Equivalently, we can write $W_0 = NI + \sqrt{2N}\,G$, where $G \sim \mathsf{GOE}(K)$.) Then,*

$$\|\mathrm{law}(W) - \mathrm{law}(W_0)\|_{\mathrm{TV}} \le O\Bigl(\frac{K^{3/2}}{N^{1/2}}\Bigr)\,.$$

Finally, we also require the following basic linear algebraic fact:

Proposition 23 (rotating the right singular vectors)**.** *Let $V, V' \in \mathbb{R}^{K\times N}$ be such that $VV^\intercal = (V')(V')^\intercal$. Then, there exists an orthogonal matrix $U \in \mathbb{R}^{N\times N}$ such that $VU = V'$.*

Lower bound against block Krylov algorithms

We start with the following proposition, which will be useful in establishing the existence of matrices with different inverse traces but which generate similar power method iterates.

Proposition 24 (polynomial approximation and duality)**.** *Suppose that $K \le c_0\sqrt{\kappa}\log d$. Then, there exist $\kappa = \lambda_1 > \lambda_2 > \cdots > \lambda_{K+2} = 1$ and non-negative real numbers $x_1, \dots, x_{K+2}; x_1', \dots, x_{K+2}'$, such that:*

1. *For all $0 \le j \le K$, $\sum_{i=1}^{K+2} x_i\lambda_i^j = \sum_{i=1}^{K+2} x_i'\lambda_i^j$.*
2. *$\sum_{i=1}^{K+2} x_i = \sum_{i=1}^{K+2} x_i' = d$.*
3. *$\sum_{i=1}^{K+2} x_i/\lambda_i - \sum_{i=1}^{K+2} x_i'/\lambda_i \ge 2d^{1-2c_0-O(1/\sqrt{\kappa})}/\kappa$.*

Proof. If we fix the values of the λ_i to be the choices in Corollary 14, this becomes a linear program in the variables $\{x_i\}_{i=1}^{K+2}, \{x_i'\}_{i=1}^{K+2}$. By writing $\mathbf{x} = (x_1, \dots, x_{K+2}, x_1', \dots, x_{K+2}')$, our goal is to maximize $\mathbf{c}^\intercal\mathbf{x}$ over $\mathbf{x} \ge 0$ subject to $A\mathbf{x} = \mathbf{b}$. In our case, we set

$$\mathbf{c} \coloneqq \begin{bmatrix} \lambda_1^{-1} \\ \vdots \\ \lambda_{K+2}^{-1} \\ -\lambda_1^{-1} \\ \vdots \\ -\lambda_{K+2}^{-1} \end{bmatrix}, \qquad A \coloneqq \begin{bmatrix} 1 & \cdots & 1 & 1 & \cdots & 1 \\ 1 & \cdots & 1 & -1 & \cdots & -1 \\ \lambda_1 & \cdots & \lambda_{K+2} & -\lambda_1 & \cdots & -\lambda_{K+2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \lambda_1^K & \cdots & \lambda_{K+2}^K & -\lambda_1^K & \cdots & -\lambda_{K+2}^K \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 2d \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

We can consider the dual linear program, and by strong duality this maximization is equivalent to minimizing $\mathbf{b}^\intercal\mathbf{y}$ over $\mathbf{y}$ such that $A^\intercal\mathbf{y} \ge \mathbf{c}$. By writing $\mathbf{y} = (z, y_0, y_1, \dots, y_K)$, this means we wish to minimize $2dz$ subject to $z + (y_0 + y_1\lambda_i + \cdots + y_K\lambda_i^K) \ge \frac{1}{\lambda_i}$ and $z - (y_0 + y_1\lambda_i + \cdots + y_K\lambda_i^K) \ge -\frac{1}{\lambda_i}$ for all $1 \le i \le K+2$. Equivalently, we wish to minimize $2dz$ subject to the existence of a polynomial P of degree at most K (with coefficients $y_0, \dots, y_K$) such that $z \ge |\frac{1}{\lambda_i} - P(\lambda_i)|$ for all $i \le K+2$.

The minimum for the dual linear program (and thus the maximum for the primal linear program), is $2d\inf_{P\in\mathcal{P}_K}\max_{1\le i\le K+2}|\frac{1}{\lambda_i} - P(\lambda_i)|$, where $\mathcal{P}_K$ is the set of polynomials of degree at most K with real coefficients. By Corollary 14, this quantity is at least $2d^{1-2c_0-O(1/\sqrt{\kappa})}/\kappa$. □

We note that a slightly strengthened version of Proposition 24 holds. Let $0 < c_1 < 1$.

Corollary 15 (existence of good solutions)**.** *Proposition 24 holds, where we also ensure that each x_i and x'_i is at least $\frac{d}{2\,(K+2)}$ and $\frac{|x_i - x'_i|}{x_i} \leq \frac{2c_1}{1-c_1}$, though the right-hand side of the third condition becomes $c_1 d^{1-2c_0-O(1/\sqrt{\kappa})}/\kappa$.*

Proof. First, replace every x_i with $\frac{1}{2}\,(x_i + \frac{d}{K+2})$ and x'_i with $\frac{1}{2}\,(x'_i + \frac{d}{K+2})$. Then, we have that the replaced x_i, x'_i are at least $\frac{d}{2\,(K+2)}$, and the remaining statements in Proposition 24 hold, except the third which has the right-hand side replaced with $d^{1-2c_0-O(1/\sqrt{\kappa})}/\kappa$.

Next, replace every x_i with $\tilde{x}_i := \frac{1+c_1}{2}\,x_i + \frac{1-c_1}{2}\,x'_i$, and every x'_i with $\tilde{x}'_i := \frac{1+c_1}{2}\,x'_i + \frac{1-c_1}{2}\,x_i$. We still have that every $\tilde{x}_i, \tilde{x}'_i$ is at least $\frac{d}{2\,(K+2)}$, the first two conditions still hold, and the right-hand side of third condition is now $c_1 d^{1-2c_0-O(1/\sqrt{\kappa})}/\kappa$. Finally, note that $|\tilde{x}_i - \tilde{x}'_i| \leq c_1\,|x_i - x'_i|$, whereas $\tilde{x}_i \geq \frac{1-c_1}{2}\,(x_i + x'_i)$. This implies that $\frac{|\tilde{x}_i - \tilde{x}'_i|}{\tilde{x}_i} \leq \frac{2c_1}{1-c_1}$. □

We now have the necessary tools to prove our lower bound against block Krylov algorithms. Before doing so, we establish that there exist diagonal matrices D, D' which have substantially different inverse traces, but block Krylov algorithms cannot distinguish between them. To prove our actual lower bound, we show the same claim holds even if D, D' are randomly rotated, and the inverse trace difference is enough for a single sample to distinguish between them.

Lemma 39 (construction of diagonal matrices)**.** *Suppose that $K \leq c_0\sqrt{\kappa}\log d$ and $K \leq O(d)$. Then, there exist diagonal matrices $D, D' \in \mathbb{R}^{d\times d}$ with all diagonal entries between 1 and κ with the following properties.*

1. $|\operatorname{Tr}(D^{-1}) - \operatorname{Tr}(D'^{-1})| \geq c_1 d^{1-2c_0-O(1/\sqrt{\kappa})}/\kappa - 2\,(K+2)$.
2. *Consider sampling K d-dimensional random vectors $v^{(1)}, \ldots, v^{(K)} \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$. Then, the distributions of $\{\langle v^{(k)}, D^j\, v^{(\ell)}\rangle\}_{j\leq K+2;\, k,\ell\leq K}$ and $\{\langle v^{(k)}, D'^j\, v^{(\ell)}\rangle\}_{j\leq K+2; k,\ell\leq K}$ differ in total variation distance by at most $O(c_1 K^3 + K^3/d^{1/2})$.*

Proof. Choose $\{x_i\}_{i=1}^{K+2}$, $\{x'_i\}_{i=1}^{K+2}$, and $\{\lambda_i\}_{i=1}^{K=2}$ satisfying Corollary 15. Define integers $\{N_i\}_{i=1}^{K+2}$ such that each N_i is either $\lfloor x_i \rfloor$ or $\lceil x_i \rceil$ and $\sum_{i=1}^{K+2} N_i = d$; define $\{N'_i\}_{i=1}^{K+2}$ similarly in terms of $\{x'_i\}_{i=1}^{K+2}$. We let D, D' be diagonal matrices such that for all i, D has N_i diagonal entries equal λ_i, and D' has N'_i diagonal entries equal to λ_i. Now, let $v^{(1)}, \ldots, v^{(K)} \in \mathbb{R}^d$ be K random vectors drawn i.i.d. from $\mathcal{N}(0, I_d)$ (and define $v^{(1)\prime}, \ldots, v^{(K)\prime}$ similarly). For $1 \leq i \leq K+2$, we define $v^{(k,i)}$ to be the projection of $v^{(k)}$ onto the dimensions corresponding to the diagonal entry λ_i for D. Note that $\{v^{(k,i)}\}_{i\leq K+2,\, k\leq K}$ are independent, and $v^{(k,i)} \sim \mathcal{N}(0, I_{N_i})$. Likewise, define $\{v^{(k,i)\prime}\}_{i \le K+2,\, k \le K}$ accordingly in terms of D'.

Note that $\operatorname{Tr}(D^{-1}) - \operatorname{Tr}(D'^{-1}) = \sum_{i=1}^{K+2} N_i/\lambda_i - \sum_{i=1}^{K+2} N_i'/\lambda_i$. Since $|N_i - x_i|, |N_i' - x_i'| \le 1$, and since each $\lambda_i \ge 1$, it implies

$$\operatorname{tr}(D^{-1}) - \operatorname{tr}(D'^{-1}) \ge \frac{c_1\, d^{1-2c_0-O(1/\sqrt{\kappa})}}{\kappa} - 2\,(K+2)\,.$$

Next, we let $W^{(i)}$ represent the $K \times K$ matrix with entries $W^{(i)}_{k,\ell} = \langle v^{(k,i)}, v^{(\ell,i)}\rangle$ and define $W^{(i)\prime}$ similarly. Note that the matrices $W^{(i)}, W^{(i)\prime}$ over all i are independent. In addition, $W^{(i)}$ has the $\mathsf{Wishart}(K, N_i)$ distribution, and $W^{(i)\prime}$ has the $\mathsf{Wishart}(K, N_i')$ distribution. In addition, for any $k, \ell \le K$ and $j \le T$, we have that $\langle v^{(k)}, D^j\, v^{(\ell)}\rangle = \sum_{i=1}^{K+2} \lambda_i^j W^{(i)}_{k,\ell}$.

Now, we attempt to design a coupling between the matrices $\{W^{(i)}\}_{i=1}^{K+2}$ and $\{W^{(i)\prime}\}_{i=1}^{K+2}$ such that $W^{(i)} - W^{(i)\prime} = (x_i - x_i')\, I_K$ for all $i \le K+2$, with high probability. Note that this implies our claim, due to Corollary 15. To design this coupling, first note that by Lemma 38, if we draw $Z^{(i)} \sim N_i\, I_K + \sqrt{2N_i}\, \mathsf{GOE}(K)$, then $\|\operatorname{law}(W^{(i)}) - \operatorname{law}(Z^{(i)})\|_{\mathrm{TV}} \le O(K^{3/2}/N_i^{1/2})$, and a similar statement holds if we define $Z^{(i)\prime}$ and compare its law to that of $W^{(i)\prime}$.

Note that the entries of $Z^{(i)}$ and $Z^{(i)\prime}$ are independent (apart from the requirement of symmetry), so we will attempt a coupling between the entries $Z^{(i)}_{k,\ell}$ and $Z^{(i)\prime}_{k,\ell}$. For $k < \ell$, since $Z^{(i)}_{k,\ell} \sim \mathcal{N}(0, N_i)$ and $Z^{(i)\prime}_{k,\ell} \sim \mathcal{N}(0, N_i')$, the total variation distance between their distributions is bounded up to a constant, using Corollary 15, by

$$\Bigl|\frac{N_i'}{N_i} - 1\Bigr| \le \Bigl|\frac{x_i'}{x_i} - 1\Bigr| + \Bigl|\frac{N_i' - x_i'}{N_i}\Bigr| + \Bigl|\frac{x_i'\,(N_i - x_i)}{N_i x_i}\Bigr| \le O\bigl(c_1 + \frac{K}{d}\bigr)$$

under our assumptions. Therefore, we can couple $Z^{(i)}_{k,\ell}$ and $Z^{(i)\prime}_{k,\ell}$ such that they fail to coincide with this probability. For $k = \ell$, we have $Z^{(i)}_{k,k} \sim \mathcal{N}(N_i, 2N_i)$ and $Z^{(i)\prime}_{k,k} + x_i - x_i' \sim \mathcal{N}(N_i' + x_i - x_i', 2N_i')$. The total variation distance between their distributions is bounded by a constant times

$$\Bigl|\frac{N_i'}{N_i} - 1\Bigr| + \frac{|N_i' - x_i' + x_i - N_i|}{\sqrt{N_i}} \le O\bigl(c_1 + \frac{K^{1/2}}{d^{1/2}}\bigr)\,.$$

Therefore, we can couple the two random variables together so that $Z^{(i)}_{k,k} = Z^{(i)\prime}_{k,k} + x_i - x_i'$ fails with the above probability.

By a union bound, the coupling $Z^{(i)} = Z^{(i)\prime} + (x_i - x_i')\, I_K$ for all i fails with probability at most

$$O\Big(K^3\,\big(c_1+\frac{K}{d}\big)+K^2\,\big(c_1+\frac{K^{1/2}}{d^{1/2}}\big)\Big)=O\big(c_1K^3+\frac{K^{5/2}}{d^{1/2}}\big)\,.$$

Combining this with comparison between the Wishart and GOE ensembles and another union bound, we obtain the result. □

Finally, we are able to prove our main lower bound against block Krylov algorithms.

Lemma 40 (lower bound against block Krylov algorithms)**.** *Let κ, K, D, D' be as in Lemma 39. Then, let U be a uniformly random orthogonal matrix in $\mathbb{R}^{d\times d}$, and let $\Lambda = U^\mathsf{T} D U$ and $\Lambda' = U^\mathsf{T} D' U$. Let $v^{(1)},\dots,v^{(K)} \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$. Then, for any $\delta > 0$, provided $K \le O_\delta(\sqrt{\kappa}\log d)$ and $\kappa \le d^{1/5-\delta}$, the distributions of $\{\Lambda^j v^{(k)}\}_{j\le (K+2)/2,\, k\le K}$ and $\{\Lambda'^j v^{(k)}\}_{j\le (K+2)/2;\, k\le K}$ differ in total variation distance by at most $o(1)$. On the other hand, drawing a sample either from $\mathcal{N}(0,\Lambda^{-1})$ or $\mathcal{N}(0,\Lambda'^{-1})$ can, with probability $1-o(1)$, distinguish between the two cases.*

Proof. From Lemma 39, there is a coupling such that the tuples $\{\langle v^{(k)}, D^j\, v^{(\ell)}\rangle\}_{j\le K+2,\, k,\ell\le K}$ and $\{\langle v^{(k)\prime}, D'^j\, v^{(\ell)\prime}\rangle\}_{j\le K+2,\, k,\ell\le K}$ are equal with high probability. In particular, it holds that $\langle D^i\, v^{(k)}, D^j\, v^{(\ell)}\rangle = \langle D'^i\, v^{(k)\prime}, D'^j\, v^{(\ell)\prime}\rangle$ for all $i,j \le (K+2)/2$ and $k \le K$ with high probability. By Proposition 23, there is a unitary matrix U_0 such that $D'^j\, v^{(k)\prime} = U_0 D^j\, v^{(k)}$ for all $j \le (K+2)/2$ and all $k\le K$ with high probability. Note then that the tuples $\{U^\mathsf{T} D^j U\, U^\mathsf{T} v^{(k)}\}_{j\le (K+2)/2,\, k\le K}$ and $\{U^\mathsf{T} U_0^\mathsf{T} D'^j U_0 U\, U^\mathsf{T} U_0^\mathsf{T} v^{(k)\prime}\}_{j\le (K+2)/2,\, k\le K}$ are equal with high probability, and this is a coupling which witnesses the fact that the distributions of $\{\Lambda^j v^{(k)}\}_{j\le (K+2)/2,\, k\le K}$ and $\{\Lambda'^j v^{(k)}\}_{j\le (K+2)/2,\, k\le K}$ are at most $O(c_1K^3 + K^3/d^{1/2})$ apart in total variation distance.

Finally, we note that from a single sample it is easy to distinguish between $\mathcal{N}(0,\Lambda^{-1})$ and $\mathcal{N}(0,\Lambda'^{-1})$. This is because if $X \sim \mathcal{N}(0,\Lambda^{-1})$, then $\mathbb{E}[\|X\|^2] = \operatorname{Tr}(\Lambda^{-1}) = \operatorname{Tr}(D^{-1}) = \sum_{i=1}^{K+2} N_i/\lambda_i$, but one checks that $\operatorname{var}(\|X\|^2) = O(\sum_{i=1}^{K+2} N_i/\lambda_i^2) \le O(d)$. Likewise, if $X' \sim \mathcal{N}(0,\Lambda'^{-1})$, then we have $\mathbb{E}[\|X'\|^2] = \sum_{i=1}^{K+2} N_i'/\lambda_i$ but $\operatorname{var}(\|X'\|^2) = O(d)$. So, the difference in their expectations at least $c_1 d^{1-2c_0-O(1/\sqrt{\kappa})}/\kappa - 2\,(K+2)$, whereas the standard deviations are bounded by $O(d^{1/2})$.

To finish the proof, we must choose the values of c_0 and c_1. We require the following conditions:

1. $c_1 K^3 = o(1)$.
2. $K^3/d^{1/2} = o(1)$.
3. $d^{1/2} = o(c_1 d^{1-2c_0-O(1/\sqrt{\kappa})}/\kappa - 2\,(K+2))$.

For the second condition, we can assume $\kappa \le d^{1/3}/\log^4(d)$. To satisfy the first condition, we can set $c_1 = 1/(\kappa^{3/2}\log^4(d))$. Finally, if κ is sufficiently large and if c_0 is chosen depending on δ, then the third condition requires $\sqrt{\kappa}\log d + d^{1/2} = o(d^{1-\delta}/\kappa^{5/2})$, and it suffices for $\kappa \le d^{1/5-\delta}$. □

Remark 12. We did not attempt to optimize the exponent in the condition $\kappa \le d^{1/5-\delta}$. Indeed, by using the chain rule for the KL divergence rather than a union bound in the proof of Lemma 39, we believe that the total variation bound can be improved to $O(c_1 K^{3/2} + K^{5/2}/d^{1/2})$, and a back-of-the-envelope calculation suggests that this could improve the condition to $\kappa \le d^{2/7-\delta}$. Nevertheless, this falls short of capturing the full regime $\sqrt{\kappa}\log d \le O(d)$, and we leave this as an open question.

Reduction to block Krylov algorithms

In this section, we show that in order to prove a lower bound for sampling from Gaussians against any query algorithm, it suffices to prove a lower bound against block Krylov algorithms.

Setup

Let $\Lambda = U^\intercal D U$, where D is a (possibly random) diagonal matrix, U is a Haar-random orthogonal matrix, and U and D are independent. We consider the following model, which is a strengthening of the matrix-vector product model:

Definition 18 (extended oracle model)**.** *Given $K \in \mathbb{N}$, for all $k \in [K]$, the algorithm chooses a new query point v_k, and receives the information $\{\Lambda^i v_j\}_{(i,j)\in H_k}$, where $H_k := \{(i,j) : i + j \le k+1, i \ge 0, 1 \le j \le k\}$ is a set of ordered pairs of nonnegative integers. We use the following notation $\{\Lambda^i v_j\}_S$ for any set S to denote $\{\Lambda^i v_j\}_{(i,j)\in S}$.*

This is clearly a stronger oracle model than before, so a lower bound against algorithms in the extended oracle model implies a lower bound against algorithms in the original matrix-vector model.

Definition 19 (adaptive deterministic algorithm)**.** *An adaptive deterministic algorithm $\mathcal{A}$ that makes K extended oracle queries (see Definition 18) is given by a deterministic collection of functions $v_1, v_2(\cdot), \ldots, v_K(\cdot)$, where v_1 is constant and each $v_k(\cdot)$ is a function of $\frac{k\,(k+1)}{2} - 1$ inputs. This corresponds to a sequence of queries where the k-th query $v_k(\{\Lambda^i v_j\}_{H_{k-1}})$ is chosen adaptively based on the information available to the algorithm at the start of iteration k. (Note that v_1 has no inputs.) When the choice of the inputs is clear from context, we may simply write $v_k = v_k(\{\Lambda^i v_j\}_{H_{k-1}})$.*

In the extended oracle model, the next lemma shows that we can assume that each v_k is a unit vector orthogonal to its inputs.

Lemma 41 (extended oracle and orthogonal queries)**.** *For $k \in [2, K]$, let v_k be as stated in Definition 19 and let $\{\Lambda^i v_j\}_{H_{k-1}}$ be as stated in Definition 18. Then, v_k is orthogonal to the subspace spanned by the vectors in $\{\Lambda^i v_j\}_{H_{k-1}}$.*

Proof. Assume for sake of contradiction that this were not the case. Then, we can decompose $v_k = \sum_{(i,j)\in H_{k-1}} c_{i,j}\Lambda^i v_j + c^\perp v_k^\perp$ where $v_k^\perp$ is a unit vector orthogonal to $\{\Lambda^i v_j\}_{H_{k-1}}$ and each $c_{i,j}$ and $c^\perp$ is a scalar. At the end of iteration k, the new information obtained by the algorithm is $\{\Lambda^i v_j\}_{i+j=k+1, j\le k}$. For all $(i,j) \neq (1,k)$, the new information does not depend on v_k. Also, $\Lambda v_k = \sum_{(i,j)\in H_{k-1}} c_{i,j}\Lambda^{i+1} v_j + c^\perp \Lambda v_k^\perp$, where each $\Lambda^{i+1}v_j$ is information obtained by the algorithm at the end of iteration $k+1$ regardless (due to our extended query model). Since $(i+1,j) \in H_k$ if $(i,j) \in H_{k-1}$, and since $(1,k) \in H_k$, this expression shows that the algorithm would receive the same amount of information (or more, if $c^\perp = 0$) if it queries $v_k^\perp$ instead of v_k. Applying this reasoning inductively proves the claim. □

We compare to a *block Krylov algorithm*, which makes i.i.d. standard Gaussian queries $z_1, \dots, z_K$ and then receives $\{\Lambda^i z_j\}$ for all $i, j \le K$. Recall that a block Krylov algorithm does not make *adaptive* queries, so it is easier to prove lower bounds against block Krylov algorithms. Our goal is to now show that block Krylov algorithms can simulate an adaptive deterministic algorithm.

Conditioning lemma

We start by proving a general conditioning lemma which will be invoked repeatedly in the reduction to block Krylov algorithms. This lemma roughly shows that if the adaptive algorithm knows $\{\Lambda^i v_j\}_{H_k}$, the posterior distribution of Λ given $\{\Lambda^i v_j\}_{H_k}$ is indeed rotationally symmetric on the orthogonal complement $\{\Lambda^i v_j\}_{H_k}$.

We will use the notation $\overset{\mathsf{d}}{=}$ to denote that two random variables are equal in probability distribution (possibly conditioned on other information).

Lemma 42 (conditioning lemma, preliminary version)**.** *Let U be a Haar-random orthogonal matrix, and $\Lambda = U^\intercal D U$, where D is a (possibly random) positive diagonal matrix. Suppose that $\mathcal{A}$ is an adaptive deterministic algorithm that generates extended oracle queries $v_1, \dots, v_K$, and after the k-th query knows $\Lambda^i v_j$ for all $(i,j) \in H_k$. For any integer $m \ge 1$, let k be the integer such that $\frac{k(k+1)}{2} \le m < \frac{(k+1)(k+2)}{2}$, i.e., m is at least the k-th triangular number but less than the $(k+1)$-th triangular number. Consider the order of vectors $v_1, \Lambda v_1, v_2, \Lambda^2 v_1, \Lambda v_2, v_3, \Lambda^3 v_1, \dots$ (this enumerates $\Lambda^i v_j$ in order of $i + j$, breaking ties with smaller values of j first). Let W_m be the set of first*

m of these vectors and X_k be the set $\{v_1, \dots, v_k\}$. Let V be a Haar-random orthogonal matrix fixing W_m and acting on the orthogonal complement $W_m^\perp$. Then, $(X_k, U) \overset{\mathsf{d}}{=} (X_k, UV)$.

Before proving this lemma, we note that since the algorithm is deterministic and D is fixed, W_m and X_k are deterministic functions of Λ, and thus of U. Hence, we can write $v_k(U'), W_m(U'), X_k(U')$ to be the v_k, W_m, X_k that would have been generated if we started with $\Lambda' = (U')^\intercal D U'$. (If no argument is given, v_k, W_m, X_k are assumed to mean $v_k(U), W_m(U), X_k(U)$, respectively.) We note the following proposition.

Proposition 25 (fixing the first m queries and responses)**.** *Suppose that V is any orthogonal matrix fixing $W_m(U)$. Then, $W_m(U) = W_m(UV)$.*

Proof. We prove $W_{m'}(U) = W_{m'}(UV)$ for all $m' \le m$. The base case of $k = 1$ is trivial, since v_1 is fixed. We now prove the induction step for m'.

If $m' \le m$ is a triangular number, $m' = \frac{k(k+1)}{2}$, then the m'-th vector in W_m is v_k. But note that $v_k(U)$ is a deterministic function of $W_{m'-1}(U)$, and $v_k(UV)$ is the same deterministic function of $W_{m'-1}(UV)$. Hence, if the induction hypothesis holds for $m' - 1$, it also holds for m.

If $m' \le m$ is not a triangular number, then the m'-th number in $W_m(U)$ is $\Lambda^i v_j$ for some $i \ge 1$. Likewise, the m'-th number in $W_m(UV)$ is $V^\intercal \Lambda^i V v_j(UV)$. Since $i \ge 1$, we know that $v_j(U) = v_j(UV)$, by the induction hypothesis on $\frac{j(j+1)}{2} < m'$. But, we know that V fixes W_m, which means it fixes v_j and $\Lambda^i v_j$. Thus, $V^\intercal \Lambda^i V v_j(UV) = V^\intercal \Lambda^i V v_j = \Lambda^i v_j$. □

We are now ready to prove Lemma 42.

Proof of Lemma 42. We prove this by induction on m. For the base case $m = 1$, U is a random matrix and V is a random matrix that fixes v_1. Note that v_1 is chosen independently of Λ (and thus of U), so U and V are independent. Even for any fixed V, the distribution UV is a uniformly random orthogonal matrix, so overall $U \overset{\mathsf{d}}{=} UV$. Also, v_1 is deterministic, so $(v_1, U) \overset{\mathsf{d}}{=} (v_1, UV)$.

For the induction step, we split the proof into 2 cases. The proofs in both cases will be very similar, but with minor differences.

Case 1: m is a triangular number. This means that the m-th vector added is v_k, where $m = \frac{k(k+1)}{2}$. Let V_1 be a random orthogonal matrix fixing W_{m-1} and V_2 be a random orthogonal matrix fixing W_m. Our goal is then to show $(X_k, U) \overset{\mathsf{d}}{=} (X_k, UV_2)$.

To make this rigorous, we note an order of generating the random variables. First, we generate U randomly: W_m and X_k are deterministic in terms of U. Next, we define V_1 to be a random rotation fixing W_{m-1}. Finally, we define V_2

to be a random rotation fixing W_m, where V_1, V_2 are conditionally independent on U.

First, we prove that $(X_k, U) \overset{\mathsf{d}}{=} (X_k, UV_1)$. Note that $U \overset{\mathsf{d}}{=} UV_1$ by our inductive hypothesis. In addition, since V_1 fixes $W_{m-1}(U)$, $W_{m-1}(U) = W_{m-1}(UV_1)$ by Proposition 25. Since $m = \frac{k(k+1)}{2}$ is a triangular number, $X_k(\cdot)$ is a deterministic function of $W_{m-1}(\cdot)$, which means $X_k(U) = X_k(UV_1)$. Hence, $(X_k, U) \overset{\mathsf{d}}{=} (X_k(UV_1), UV_1) = (X_k, UV_1)$.

Next, we prove that $(X_k, UV_2) \overset{\mathsf{d}}{=} (X_k, UV_1V_2)$. It suffices to prove that

$$(X_k, U, V_2) \overset{\mathsf{d}}{=} (X_k, UV_1, V_2)\,.$$

To do so, we first show that $V_2 = f(U, R)$, where f is a deterministic function and R represents a random orthogonal matrix over $d - \dim(W_m)$ dimensions that is independent of U. (Recall that W_m is a deterministic function of U.) To define $f(U, R)$, we consider some deterministic map that sends each W_m to a set of $d - \dim(W_m)$ basis vectors in $W_m^{\perp}$. We then define $V_2 = f(U, R)$ to act on $W_m^{\perp}$ using R and the correspondence of basis vectors. Since W_m and X_k are deterministic in terms of U, this means $f(U, R)$ is well-defined. We will now show that

$$V_2 = f(U, R) = f(UV_1, R) \quad \text{ and } \quad X_k = X_k(UV_1)\,.$$

Since $U \overset{\mathsf{d}}{=} UV_1$ by our inductive hypothesis,

$$(X_k, U, V_2) \overset{\mathsf{d}}{=} (X_k(UV_1), UV_1, f(UV_1, R)) = (X_k, UV_1, V_2)\,.$$

By Proposition 25, $W_{m-1}(U) = W_{m-1}(UV_1)$, and since $X_k(\cdot)$ is deterministic given $W_{m-1}(\cdot)$ for $m = \frac{k(k+1)}{2}$, $X_k(U) = X_k(UV_1)$. This implies $W_m(U) = W_m(UV_1)$, which means $f(UV_1, R) = f(U, R)$, since $f(\cdot, R)$ only depends on $W_m(\cdot)$ and R. This completes the proof.

Next, we show that $(X_k, UV_1V_2) \overset{\mathsf{d}}{=} (X_k, UV_1)$. Since we chose the order with U being defined first, we are allowed to condition on U. Since X_k is deterministic in terms of U, it suffices to show that $V_1V_2 \mid U \overset{\mathsf{d}}{=} V_1 \mid U$. Since W_{m-1}, W_m are also deterministic given U, note that V_1 is a uniformly random orthogonal matrix fixing W_{m-1}, and V_2 is a random orthogonal matrix fixing $W_m \supset W_{m-1}$. Since V_1 and V_2 are conditionally independent given U, this means $V_1V_2 \mid U$ is a uniformly random orthogonal matrix fixing W_{m-1}, so $V_1V_2 \mid U \overset{\mathsf{d}}{=} V_1 \mid U$.

In summary, we have that

$$(X_k, U) \overset{\mathsf{d}}{=} (X_k, UV_1)$$

$$\begin{aligned} &\overset{\mathsf{d}}{=} (X_k, UV_1V_2) \\ &\overset{\mathsf{d}}{=} (X_k, UV_2)\,. \end{aligned}$$

Case 2: m is not a triangular number. Again, let V_1 be a random orthogonal matrix fixing W_{m-1} and V_2 be a random orthogonal matrix fixing W_m. Our goal is again to show that $(X_k, U) \overset{\mathsf{d}}{=} (X_k, UV_2)$.

First, we again have $(X_k, UV_1) \overset{\mathsf{d}}{=} (X_k, U)$ by our inductive hypothesis.

Next, we show that $(X_k, UV_2) \overset{\mathsf{d}}{=} (X_k, UV_2V_1)$. It suffices to prove that

$$(X_k, U, V_2) \overset{\mathsf{d}}{=} (X_k, UV_1, V_1^{\mathsf{T}}V_2V_1)\,,$$

since $(UV_1)(V_1^{\mathsf{T}}V_2V_1) = UV_2V_1$. We recall the random variable R and use the same function $V_2 = f(U,R)$. Since we have already shown that $U \overset{\mathsf{d}}{=} UV_1$, this implies that $(X_k, U, V_2) \overset{\mathsf{d}}{=} (X_k(UV_1), UV_1, f(UV_1, R))$. Since m is not triangular, $X_k(\cdot)$ is contained in $W_{m-1}(\cdot)$, so by Proposition 25, $X_k(U) = X_k(UV_1)$. So, we have

$$(X_k, U, V_2) \overset{\mathsf{d}}{=} (X_k(UV_1), UV_1, f(UV_1, R)) = (X_k, UV_1, f(UV_1, R))\,.$$

Now, if we fix U and V_1, $W_{m-1}(UV_1) = W_{m-1}(U)$ by Proposition 25. However, since the m-th (i,j) pair has $i \geq 1$ when m is not triangular, the final vector in $W_m(UV_1)$ will be $V_1^{\mathsf{T}}\Lambda^i V_1 v_j = V_1^{\mathsf{T}}(\Lambda^i v_j)$. For fixed U, V_1, $f(U,R)$ is a random rotation fixing W_{m-1} and $\Lambda^i v_j$, but $f(UV_1, R)$ is a random rotation fixing W_{m-1} and $V_1^{\mathsf{T}}(\Lambda^i v_j)$. Since V_1^{T} fixes W_{m-1} by how we defined V_1, this means that for fixed U, V_1, $f(U,R)$ is a random rotation fixing W_m but $f(UV_1, R)$ is a random rotation fixing $V_1^{\mathsf{T}}W_m$. Therefore, conditioned on U, V_1, $f(UV_1, R)$ has the same distribution as $V_1^{\mathsf{T}}f(U,R)V_1$. Since X_k is deterministic in terms of U, this means

$$(X_k, UV_1, f(UV_1, R)) \mid U, V_1 \overset{\mathsf{d}}{=} (X_k, UV_1, V_1^{\mathsf{T}}f(U,R)V_1) \mid U, V_1\,.$$

We can remove the conditioning to establish that $(X_k, UV_1, f(UV_1, R)) \overset{\mathsf{d}}{=} (X_k, UV_1, V_1^{\mathsf{T}}f(U,R)V_1) = (X_k, UV_1, V_1^{\mathsf{T}}V_2V_1)$, which completes the proof.

Next, we show that $(X_k, UV_2V_1) \overset{\mathsf{d}}{=} (X_k, UV_1)$. The proof is essentially the same as in the case when m is triangular. We again condition on U, and we have that $V_2V_1 \mid U \overset{\mathsf{d}}{=} V_1 \mid U$ have the same distribution as uniform orthogonal matrices fixing $W_{m-1}(U)$. Since X_k is a deterministic function of U, this means $(X_k, UV_2V_1) \mid U \overset{\mathsf{d}}{=} (X_k, UV_1) \mid U$, and removing the conditioning finishes the proof.

In summary,

$$\begin{aligned}(X_k, U) &\overset{\mathsf{d}}{=} (X_k, UV_1) \\ &\overset{\mathsf{d}}{=} (X_k, UV_2V_1) \\ &\overset{\mathsf{d}}{=} (X_k, UV_2)\,. \end{aligned}$$ □

We now prove our main conditioning lemma, which will be a modification of Lemma 42.

Lemma 43 (conditioning lemma)**.** *Let all notation be as in Lemma 42, and let V_0 be a fixed orthogonal matrix fixing W_m. Importantly, V_0 is a deterministic function only depending on W_m (and not directly on U). Then, $(X_k, U) \overset{\mathsf{d}}{=} (X_k, UV_0)$.*

Proof. First, note that since V_0 is a deterministic function of W_m, it is also a deterministic function of U. We can write $V_0(\cdot)$ as this function, and $V_0 = V_0(U)$.

Now, Lemma 42 proves that $(X_k, U) \overset{\mathsf{d}}{=} (X_k, UV)$. Note that conditioned on U, V is a random matrix fixing W_m and V_0 is a fixed matrix fixing W_m, which means that $VV_0 \mid U \overset{\mathsf{d}}{=} V \mid U$. Hence, $(X_k, UV) \overset{\mathsf{d}}{=} (X_k, UVV_0)$. But from Proposition 25, $X_k(UV) = X_k(U)$ and $W_m(UV) = W_m(U)$, which means that $V_0(\cdot)$, which only depends on $W_m(\cdot)$, satisfies $V_0(UV) = V_0(U)$. Hence, because $U \overset{\mathsf{d}}{=} UV$, we have $(X_k, UVV_0) = (X_k(UV), UV \cdot V_0(UV)) \overset{\mathsf{d}}{=} (X_k(U), U \cdot V_0(U)) = (X_k, UV_0)$.

In summary, we have that $(X_k, U) \overset{\mathsf{d}}{=} (X_k, UV) \overset{\mathsf{d}}{=} (X_k, UVV_0) \overset{\mathsf{d}}{=} (X_k, UV_0)$, which completes the proof. □

From query algorithms to block Krylov algorithms

In this section, we carry out the high-level outline from Section 11.2. We aim to prove the following result, which implies that any adaptive deterministic algorithm in the extended oracle model can be simulated by rotating the output of a block Krylov algorithm.

Lemma 44 (reduction to block Krylov)**.** *Suppose $\Lambda = U^\mathsf{T} D U$, where U is a Haar-random orthogonal matrix and D is a diagonal matrix drawn from some (possibly unknown) distribution. Let $v_1, v_2(\cdot), \dots, v_K(\cdot)$ be an adaptive deterministic algorithm that makes K queries, where $K^2 < d$. Let $v_1^{\mathsf{alg}}, v_2^{\mathsf{alg}}, \dots, v_K^{\mathsf{alg}}$ be recursively defined as follows: $v_1^{\mathsf{alg}} = v_1$, and $v_k^{\mathsf{alg}} = v_k(\{\Lambda^i v_j^{\mathsf{alg}}\}_{H_{k-1}})$ for $k \geq 2$. Let $z_1, \dots, z_K$ be i.i.d. standard Gaussian vectors. Then, from the collection $\{\Lambda^i z_j\}_{H_K}$ (without knowledge of D or Λ), we can construct a set of unit vectors $\tilde{v}_1, \tilde{v}_2, \dots, \tilde{v}_K$, and a set of rotation matrices $U_1^{\mathsf{sim}}, U_2^{\mathsf{sim}}, \dots, U_K^{\mathsf{sim}}$,* *where $\tilde{v}_k$ and U_k^{sim} only depend on $\{\Lambda^i z_j\}_{H_{k-1}}$ and z_k, and such that*

$$\{(U_{1:K}^{\mathsf{sim}})^{\mathsf{T}} \Lambda^i \tilde{v}_j\}_{H_K} \overset{\mathsf{d}}{=} \{\Lambda^i v_j^{\mathsf{alg}}\}_{H_K},$$

where $U_{1:K}^{\mathsf{sim}} \coloneqq U_1^{\mathsf{sim}} \cdots U_K^{\mathsf{sim}}$, and the equivalence in distribution is over the randomness of Λ and $\{z_i\}_{i \le K}$. Moreover, $\{\Lambda^i \tilde{v}_j\}_{H_K}$ is deterministically determined by $\{\Lambda^i z_j\}_{H_K}$.

Lemma 44 says that the knowledge of $\Lambda^i z_j$ alone is sufficient to reconstruct the distribution of any adaptive algorithm's queries and responses. The proof of the lemma requires introducing a hefty amount of notation, but we emphasize that it follows along the lines of Section 11.2.

First, we describe how to construct $\tilde{v}_k$. Let $\tilde{v}_1 = \frac{z_1}{\|z_1\|}$, and for $k \ge 2$, let $\tilde{v}_k$ be the unit vector parallel to the component of z_k that is orthogonal to the span of $\{\Lambda^i z_j\}_{H_{k-1}}$. (With probability 1, this is well-defined.)

Because each $\tilde{v}_k$ is a linear combination of $\{\Lambda^i z_j\}_{H_{k-1}}$ and z_k, we can construct the set $\{\Lambda^i \tilde{v}_j\}_{H_K}$ from the set $\{\Lambda^i z_j\}_{H_K}$.

We now construct the rotation matrices U_k^{sim}. First, we define matrix-valued functions $U_k(\cdot)$, for $k = 1, \dots, K$, as follows.

Definition 20 (rotations fixing previous queries and responses)**.** *For $1 \le k \le K$, the function $U_k(\cdot)$ takes arguments $\{x_{i,j}\}_{H_{k-1}}$, y_k, z_k, where the vectors y_k and z_k have unit norm and are both orthogonal to the collection $\{x_{i,j}\}_{H_{k-1}}$.*

To define $U_1(\cdot)$: since H_0 is empty, the first function U_1 only takes arguments y_1, z_1, and is such that $U_1(y_1, z_1)$ is a deterministic orthogonal matrix that satisfies $U_1(y_1, z_1)^{\mathsf{T}} y_1 = z_1$. Note that $U_1(\cdot)$ exists because y_1 and z_1 both have unit norm; for example, we can complete y_1 and z_1 to orthonormal bases $(y_1, y_2, \dots, y_d)$, $(z_1, z_2, \dots, z_d)$ and take $U_1(y_1, z_1) = \sum_{i=1}^d y_i z_i^{\mathsf{T}}$.

To define $U_k(\cdot)$: $U_k(\{x_{i,j}\}_{H_{k-1}}, y_k, z_k)$ is a deterministic orthogonal matrix that satisfies

$$\begin{aligned} U_k^{\mathsf{T}} x_{i,j} &= x_{i,j}\,, \qquad \textit{for all } (i,j) \in H_{k-1}\,, \\ U_k^{\mathsf{T}} y_k &= z_k\,. \end{aligned} \tag{11.2}$$

Such a choice of U_k is always possible, because $k^2 < d$, and because y_k and z_k are orthogonal to $x_{i,j}$; for example, we can start with the identity matrix on the subspace spanned by $\{x_{i,j}\}_{H_{k-1}}$ and add to it a sum of outer products formed by completing y_k and z_k to two orthonormal bases of the orthogonal complement.

Next, we describe how to construct U_k^{sim}. We will define U_k^{sim} along with an auxiliary sequence $\{v_k^{\mathsf{sim}}\}_{k=1,2,\dots,K-1}$.

Definition 21 (simulated sequences)**.** *We let* $v_1^{\mathsf{sim}} = v_1$*, and* $U_1^{\mathsf{sim}} = U_1(\tilde{v}_1, v_1^{\mathsf{sim}})$*. For* $k \geq 2$*,* v_k^{sim} *and* U_k^{sim} *are defined recursively as follows:*

$$\begin{aligned} v_k^{\mathsf{sim}} &= v_k\big(\{(U_{1:(k-1)}^{\mathsf{sim}})^\intercal \Lambda^i \tilde{v}_j\}_{H_{k-1}}\big) \\ U_k^{\mathsf{sim}} &= U_k\big(\{(U_{1:(k-1)}^{\mathsf{sim}})^\intercal \Lambda^i \tilde{v}_j\}_{H_{k-1}},\ (U_{1:(k-1)}^{\mathsf{sim}})^\intercal \tilde{v}_k,\ v_k^{\mathsf{sim}}\big)\,. \end{aligned} \tag{11.3}$$

Intuitively, one can think of v_k^{sim} as the kth vector the simulator thinks the algorithm is querying, and U_k^{sim} as a rotation that corresponds v_k^{sim} to the random unit vector known by block Krylov.

Proposition 26 (existence of rotations)**.** *Each* U_k^{sim} *is well-defined.*

Proof. To show that this choice of U_k^{sim} is possible, we need to check that $(U_{1:(k-1)}^{\mathsf{sim}})^\intercal \tilde{v}_k$, v_k^{sim} both have unit norm and are orthogonal to the subspace S_k spanned by $(U_{1:(k-1)}^{\mathsf{sim}})^\intercal \Lambda^i \tilde{v}_j$ for $(i,j) \in H_{k-1}$. They both have unit norm because $\tilde{v}_k$ and v_k^{sim} are constructed to have unit norm, and inductively we can assume $U_{1:(k-1)}^{\mathsf{sim}}$ is orthogonal. Note that v_k^{sim} is orthogonal to S_k by our assumption on the function $v_k(\cdot)$, and $(U_{1:(k-1)}^{\mathsf{sim}})^\intercal \tilde{v}_k$ is also orthogonal to S_k because

$$\langle (U_{1:(k-1)}^{\mathsf{sim}})^\intercal \Lambda^i \tilde{v}_j, (U_{1:(k-1)}^{\mathsf{sim}})^\intercal \tilde{v}_k \rangle = \langle \Lambda^i \tilde{v}_j, \tilde{v}_k \rangle = 0\,,$$

where the second line follows from the definition of $\tilde{v}_k$. □

We summarize some additional properties of v_k^{sim} and U_k^{sim} in the following lemma.

Lemma 45 (properties of the simulated sequences)**.** *The variables* U_k^{sim} *and* v_k^{sim} *for* $k = 1, \ldots, K$ *defined above satisfy the following properties:*

(P1) v_k^{sim} *depends only on* $\{\Lambda^i \tilde{v}_j\}_{H_{k-1}}$*, and* U_k^{sim} *depends only on* $\{\Lambda^i \tilde{v}_j\}_{i+j \leq k}$.

(P2) For any $k \geq j$*, we have*

$$\tilde{v}_j = U_{1:k}^{\mathsf{sim}} v_j^{\mathsf{sim}}\,.$$

(P3) For $k \geq 2$*,* v_k^{sim} *satisfies*

$$v_k^{\mathsf{sim}} = v_k\big(\{(U_{1:(k-1)}^{\mathsf{sim}})^\intercal \Lambda^i U_{1:(k-1)}^{\mathsf{sim}} v_j^{\mathsf{sim}}\}_{H_{k-1}}\big)\,.$$

(P4) For $k \geq 2$*,* U_k^{sim} *satisfies*

$$U_k^{\mathsf{sim}} = U_k\big(\{(U_{1:(k-1)}^{\mathsf{sim}})^\intercal \Lambda^i U_{1:(k-1)}^{\mathsf{sim}} v_j^{\mathsf{sim}}\}_{H_{k-1}},\ (U_{1:(k-1)}^{\mathsf{sim}})^\intercal \tilde{v}_k,\ v_k^{\mathsf{sim}}\big)\,.$$

Proof. (P1) is immediate from the definitions, since $\{(i,j) : i+j \le k\} = H_{k-1} \cup \{(0,k)\}$.

To show (P2), note that the second property of the function U_k from (11.2) implies that

$$v_j^{\mathsf{sim}} = (U_j^{\mathsf{sim}})^\intercal (U_{1:(j-1)}^{\mathsf{sim}})^\intercal \tilde{v}_j = (U_{1:j}^{\mathsf{sim}})^\intercal \tilde{v}_j \,. \tag{11.4}$$

This proves (P2) for $k=j$. To prove (P2) for $k>j$, we use induction on k. If (P2) holds for $k-1 \ge j$, then

$$(U_{1:k}^{\mathsf{sim}})^\intercal \tilde{v}_j = (U_k^{\mathsf{sim}})^\intercal (U_{1:(k-1)}^{\mathsf{sim}})^\intercal \tilde{v}_j = (U_{1:(k-1)}^{\mathsf{sim}})^\intercal \tilde{v}_j = v_j^{\mathsf{sim}}. \tag{11.5}$$

Above, the middle equality holds by the first property of (11.2), since U_k^{sim} fixes $(U_{1:(k-1)}^{\mathsf{sim}})^\intercal \tilde{v}_j$ because $j \le k-1$. The final equality holds by our inductive hypothesis. So, (P2) holds for k.

Finally, (P3) and (P4) then follow from (P2), since $k-1 \ge j$ if $j \in H_{k-1}$. □

We highlight the importance of (P2) for $k=K$, which roughly states that $(U_{1:K}^{\mathsf{sim}})^\intercal$ actually sends each block Krylov-generated vector $\tilde{v}_j$ to the simulated vector v_j^{sim}.

Before proving Lemma 44, we must make one more basic definition.

Definition 22 (queries and data)**.** *For $k \ge 2$, given the matrix Λ and a set $\{v_j\}_{1\le j\le k-1}$, define $\mathfrak{C}_k$ as the function that satisfies $\mathfrak{C}_k(\Lambda, \{v_j\}_{1\le j\le k-1}) = \{\Lambda^i v_j\}_{H_{k-1}}$. In addition, define $\mathfrak{D}_k = v_k \circ \mathfrak{C}_k$.*

We are now ready to prove Lemma 44. Although the proof is notationally burdensome, the message is that we can show the equality of distributions inductively by repeatedly invoking the conditioning lemma (Lemma 43), which is designed precisely for the present situation.

Proof of Lemma 44. For $1 \le k \le K$, let $\Lambda_k := (U_{1:k}^{\mathsf{sim}})^\intercal \Lambda U_{1:k}^{\mathsf{sim}}$. Since we can write $(U_{1:k}^{\mathsf{sim}})^\intercal \Lambda^i \tilde{v}_j = (U_{1:k}^{\mathsf{sim}})^\intercal \Lambda^i (U_{1:k}^{\mathsf{sim}}) v_j^{\mathsf{sim}} = \Lambda_k^i v_j^{\mathsf{sim}}$ for any $k \ge j$ by (P2) of Lemma 45, it suffices to inductively prove that for all $1 \le k \le K$,

$$(\Lambda_k, \{v_j^{\mathsf{sim}}\}_{1\le j\le k}) \overset{\mathsf{d}}{=} (\Lambda, \{v_j^{\mathsf{alg}}\}_{1\le j\le k})\,. \tag{11.6}$$

For the base case of $k=1$, it suffices to show that $(\Lambda_1, v_1^{\mathsf{sim}}) \overset{\mathsf{d}}{=} (\Lambda, v_1^{\mathsf{alg}})$. Note, however, that $v_1^{\mathsf{sim}} = v_1^{\mathsf{alg}} = v_1$, and $\Lambda_1 = (U_1^{\mathsf{sim}})^\intercal \Lambda (U_1^{\mathsf{sim}}) = U_1(\tilde{v}_1, v_1)^\intercal \Lambda U_1(\tilde{v}_1, v_1)$. Since v_1 is a deterministic vector, $\tilde{v}_1$ is independent of Λ, and the distribution of Λ is rotationally invariant, the claim follows.

For the inductive step, assume we know $(\Lambda_k, \{v_j^{\mathsf{sim}}\}_{1\le j\le k}) \overset{\mathsf{d}}{=} (\Lambda, \{v_j^{\mathsf{alg}}\}_{1\le j\le k})$. Then, note that $v_{k+1}^{\mathsf{alg}} = v_{k+1}(\{\Lambda^i v_j^{\mathsf{alg}}\}_{H_k})$ and $v_{k+1}^{\mathsf{sim}} = v_{k+1}(\{\Lambda_k^i v_j^{\mathsf{sim}}\}_{H_k})$. Thus, we have $v_{k+1}^{\mathsf{alg}} = \mathfrak{D}_{k+1}(\Lambda, \{v_j^{\mathsf{alg}}\}_{1\le j\le k})$ and $v_{k+1}^{\mathsf{sim}} = \mathfrak{D}_{k+1}(\Lambda_k, \{v_j^{\mathsf{sim}}\}_{1\le j\le k})$. In addition, because U^{sim}_{k+1} fixes $\Lambda^i_k v^{\mathsf{sim}}_j$ for all $(i,j) \in H_k$ by (P4), we also have that $\Lambda^i_{k+1} v^{\mathsf{sim}}_j = \Lambda^i_k v^{\mathsf{sim}}_j$ for all $(i,j) \in H_k$, which means $v^{\mathsf{sim}}_{k+1} = \mathfrak{D}_{k+1}(\Lambda_{k+1}, \{v^{\mathsf{sim}}_j\}_{1\le j\le k})$. Therefore, it suffices to show

$$(\Lambda_{k+1}, \{v^{\mathsf{sim}}_j\}_{1\le j\le k}) \overset{\mathsf{d}}{=} (\Lambda, \{v^{\mathsf{alg}}_j\}_{1\le j\le k})\,, \tag{11.7}$$

as this implies $(\Lambda_{k+1}, \{v^{\mathsf{sim}}_j\}_{1\le j\le k+1}) \overset{\mathsf{d}}{=} (\Lambda, \{v^{\mathsf{alg}}_j\}_{1\le j\le k+1})$, which completes the inductive step.

Next, we show that U^{sim}_{k+1} sends $\tilde{v}_{k+1}$ to a random unit vector orthogonal to the simulated queries so far. Note that $\Lambda_{k+1} = (U^{\mathsf{sim}}_{k+1})^\mathsf{T} \Lambda_k (U^{\mathsf{sim}}_{k+1})$, where, by (P4),

$$U^{\mathsf{sim}}_{k+1} = U_{k+1}(\{\Lambda^i_k v^{\mathsf{sim}}_j\}_{H_k}, (U^{\mathsf{sim}}_{1:k})^\mathsf{T} \tilde{v}_{k+1}, v^{\mathsf{sim}}_{k+1})\,. \tag{11.8}$$

Note that $\tilde{v}_{k+1}$ is a random unit vector orthogonal to $\{\Lambda^i z_j\}_{H_k}$, or equivalently, it is a random unit vector orthogonal to $\{\Lambda^i \tilde{v}_j\}_{H_k}$. Since $(U^{\mathsf{sim}}_{1:k})^\mathsf{T} \Lambda^i \tilde{v}_j = (U^{\mathsf{sim}}_{1:k})^\mathsf{T} \Lambda^i (U^{\mathsf{sim}}_{1:k}) v^{\mathsf{sim}}_j = \Lambda^i_k v^{\mathsf{sim}}_j$ for all $(i,j) \in H_k$ (by (P2)), this means that $(U^{\mathsf{sim}}_{1:k})^\mathsf{T} \tilde{v}_{k+1}$ is orthogonal to $\{\Lambda^i_k v^{\mathsf{sim}}_j\}_{H_k}$. The random direction of $\tilde{v}_{k+1}$ has no dependence on $\{\Lambda^i \tilde{v}_j\}_{H_k}$ apart from being orthogonal to them, which means by (P1), $(U^{\mathsf{sim}}_{1:k})^\mathsf{T} \tilde{v}_{k+1}$ is a *uniformly random* unit vector orthogonal to $\{\Lambda^i_k v^{\mathsf{sim}}_j\}_{H_k}$.

Recalling that $v^{\mathsf{sim}}_{k+1} = \mathfrak{D}_{k+1}(\Lambda_k, \{v^{\mathsf{sim}}_j\}_{1\le j\le k})$, this means that we can rewrite (11.8) as

$$U^{\mathsf{sim}}_{k+1} = U_{k+1}\left(\{\Lambda^i_k v^{\mathsf{sim}}_j\}_{H_k}, \hat{v}^{\mathsf{sim}}, \mathfrak{D}_{k+1}(\Lambda_k, \{v^{\mathsf{sim}}_j\}_{1\le j\le k})\right)\,, \tag{11.9}$$

where $\hat{v}^{\mathsf{sim}}$ is a random unit vector orthogonal to $\{\Lambda^i_k v^{\mathsf{sim}}_j\}_{H_k}$. As a result, if we define

$$U^{\mathsf{alg}}_{k+1} \coloneqq U_{k+1}\left(\{\Lambda^i v^{\mathsf{alg}}_j\}_{H_k}, \hat{v}^{\mathsf{alg}}, \mathfrak{D}_{k+1}(\Lambda, \{v^{\mathsf{alg}}_j\}_{1\le j\le k})\right)\,, \tag{11.10}$$

where $\hat{v}^{\mathsf{alg}}$ is a random unit vector orthogonal to $\{\Lambda^i v^{\mathsf{alg}}_j\}_{H_k}$, then

$$\begin{aligned}(\Lambda_{k+1}, \{v^{\mathsf{sim}}_j\}_{1\le j\le k}) &= \left((U^{\mathsf{sim}}_{k+1})^\mathsf{T} \Lambda_k (U^{\mathsf{sim}}_{k+1}), \{v^{\mathsf{sim}}_j\}_{1\le j\le k}\right) \\ &\overset{\mathsf{d}}{=} \left((U^{\mathsf{alg}}_{k+1})^\mathsf{T} \Lambda (U^{\mathsf{alg}}_{k+1}), \{v^{\mathsf{alg}}_j\}_{1\le j\le k}\right)\,.\end{aligned}$$

Above, the first equality follows by definition, and the second follows from our inductive hypothesis that $(\Lambda_k, \{v^{\mathsf{sim}}_j\}_{1\le j\le k}) \overset{\mathsf{d}}{=} (\Lambda, \{v^{\mathsf{alg}}_j\}_{1\le j\le k})$, along with (11.9) and (11.10).

We are now in a position to apply the conditioning lemma (Lemma 43). Note that U^{alg}_{k+1} only depends on $\{\Lambda^i v^{\mathsf{alg}}_j\}_{H_k}$ (as well as some randomness in $\hat{v}^{\mathsf{alg}}$, but the randomness is independent of everything else given $\{\Lambda^i v^{\mathsf{alg}}_j\}_{H_k}$, so we can safely condition on it). Hence, we can apply the conditioning lemma with U^{alg}_{k+1}, to obtain that

$$(\Lambda_{k+1}, \{v^{\mathsf{sim}}_j\}_{1\le j\le k}) \overset{\mathsf{d}}{=} \big((U^{\mathsf{alg}}_{k+1})^\mathsf{T} \Lambda (U^{\mathsf{alg}}_{k+1}), \{v^{\mathsf{alg}}_j\}_{1\le j\le k}\big) \overset{\mathsf{d}}{=} \big(\Lambda, \{v^{\mathsf{alg}}_j\}_{1\le j\le k}\big)\,,$$

which establishes (11.7) and thereby concludes the proof. □

With the block Krylov reduction in hand, we can now establish our second lower bound for sampling from Gaussians.

Theorem 40 (second lower bound for sampling from Gaussians)**.** *There is a universal constant $\varepsilon_0 > 0$ such that the query complexity of sampling from Gaussian distributions $\mathcal{N}(0,\Sigma)$ in $\mathbb{R}^d$, where the condition number κ of Σ satisfies $\kappa \le d^{1/5-\delta}$, with accuracy ε_0 in total variation distance is at least $\Omega_\delta(\sqrt{\kappa}\log d)$.*

Proof. Let U be a random orthogonal matrix, and let $\Lambda = U^\mathsf{T} D U$, $\Lambda' = U^\mathsf{T} D' U$ be as in Lemma 40. We first show that if $\kappa \le d^{1/5-\delta}$ and c is a sufficiently small constant, no adaptive algorithm that makes less than $c_\delta\sqrt{\kappa}\log d$ queries to the extended oracle can distinguish between Λ and Λ', with $\Omega(1)$ probability.

First we assume that the algorithm is deterministic, so its behavior is characterized by functions $v_1, v_2(\cdot), \dots, v_K(\cdot)$, as in Lemma 44. The algorithm then proceeds to make queries $v^{\mathsf{alg}}_1, v^{\mathsf{alg}}_2, \dots, v^{\mathsf{alg}}_K$, where $v^{\mathsf{alg}}_k = v_k(\{\Lambda^i v^{\mathsf{alg}}_j\}_{H_{k-1}})$. Lemma 44 shows that the output of the algorithm $\{\Lambda^i v^{\mathsf{alg}}_j\}_{H_K}$ can be entirely simulated by a block Krylov algorithm, which receives $\{\Lambda^i z_k\}_{H_K}$, where $z_1, \dots, z_K$ are i.i.d. standard Gaussians. Lemma 40 says that a block Krylov algorithm that makes $K = c_\delta\sqrt{\kappa}\log d$ queries, where c_δ is a small constant depending on δ and $\kappa \le d^{1/5-\delta}$, cannot distinguish between Λ and Λ' with $\Omega(1)$ advantage, which then implies the same for any deterministic algorithm.

If the algorithm is randomized, then it uses a random seed ξ that is independent of Λ and Λ'. So conditional on the random seed, the algorithm will not be able to distinguish Λ and Λ' with $\Omega(1)$ advantage, so the overall probability that the randomized algorithm successfully distinguishes Λ and Λ' also cannot be $\Omega(1)$.

Finally, we note that a sample from $\mathcal{N}(0,\Lambda^{-1})$ versus $\mathcal{N}(0,\Lambda'^{-1})$ can distinguish between the two cases. This means that even if we were able to draw a sample that was $\frac{1}{3}$-far in total variation distance, we could output the correct answer with probability at least $\frac{2}{3}$. This implies that any sampling algorithm must require at least $\Omega_\delta(\sqrt{\kappa}\log d)$ queries to the extended oracle, and hence at least same number of queries to the standard oracle. □

11.5 Upper bound for sampling from Gaussians

Finally, we show a simple proof that, using only $O(\min(\sqrt{\kappa}\log d, d))$ gradient queries, one can generate an approximate sample from a Gaussian $\mathcal{N}(0,\Sigma)$ in d dimensions. Note that the density evaluated at x, up to an additive constant, equals $-\frac{1}{2}\,x^\mathsf{T}\Lambda x$ for $\Lambda = \Sigma^{-1}$, which means that a gradient query at x amounts to receiving the matrix-vector product Λx.

First, we require a well-known proposition from approximation theory.

Proposition 27 ([SV14, Theorem 3.3])**.** *For any positive integer s and $0 < \delta < 1$, there exists a polynomial $p_{s,\delta}$ of degree $\lceil \sqrt{2s\ln(2/\delta)}\rceil$ such that $|p_{s,\delta}(x) - x^s| \le \delta$ for all $x \in [-1,1]$.*

As a corollary, we have the following result.

Proposition 28 (polynomial approximation of inverse square root)**.** *For any $\kappa \ge 2$ and $\delta < \frac{1}{2}$, there exists a polynomial $q_{\kappa,\delta}$ of degree $O(\sqrt{\kappa}\log\frac{\kappa}{\delta})$ such that $|q_{\kappa,\delta}(x) - x^{-1/2}| \le \delta/\sqrt{\kappa}$ for all $1 \le x \le \kappa$.*

Proof. First, consider the function $(1+x)^{-1/2}$. For $|x| \le 1 - \frac{1}{\kappa} < 1$, we can use the Taylor series to write

$$(1+x)^{-1/2} = 1 + \sum_{t=1}^{\infty} \frac{(\frac{1}{2}-1)\,(\frac{1}{2}-2)\,(\frac{1}{2}-3)\cdots(\frac{1}{2}-t)}{t!}\,x^t = 1 + \sum_{i=1}^{\infty} c_t x^t\,,$$

where $|c_t| \le 1$ for all $t \ge 1$.

Note that for $|x| \le 1 - \frac{1}{\kappa}$, $\bigl|\sum_{t>T} c_t x^t\bigr| \le \sum_{t>T} |x|^t \le \frac{|x|^T}{1-|x|}$. For $T = O(\kappa \log\frac{\kappa}{\delta})$, we can bound this by $\frac{(1-1/\kappa)^T}{1/\kappa} \le \frac{\delta}{2}$. Therefore, for all such x,

$$\Bigl|(1+x)^{-1/2} - \sum_{t=0}^{T} c_t x^t\Bigr| \le \frac{\delta}{2}\,,$$

where we have set $c_0 \coloneqq 1$.

Next, using Proposition 27, we can replace each x^t with $p_{t,\delta}(x)$ where $p_{t,\delta}$ is a polynomial of degree $O(\sqrt{t\log(t/\delta)})$ such that $|p_{t,\delta}(x) - x^t| \le \delta/(4t^2)$ for all $|x| \le 1$. (We also let $p_{0,\delta}$ simply be the constant function 1.) Therefore,

$$\Bigl|(1+x)^{-1/2} - \sum_{t=0}^{T} c_t p_{t,\delta}(x)\Bigr| \le \frac{\delta}{2} + \sum_{t=1}^{T} |c_t|\,\frac{\delta}{4t^2} \le \delta\,.$$

In addition, the polynomial $\hat{p} \coloneqq \sum_{t=0}^{T} c_t p_{t,\delta}$ has degree at most $O(\sqrt{T\log(T/\delta)}) = O(\sqrt{\kappa}\log\frac{\kappa}{\delta})$.

To finish, $|\hat{p}(x-1) - x^{-1/2}| \le \frac{\delta}{\kappa}$ for all $\frac{1}{\kappa} \le x \le 1$, which means that

$$\Bigl|\hat{p}\bigl(\frac{x}{\kappa} - 1\bigr)\,\frac{1}{\sqrt{\kappa}} - x^{-1/2}\Bigr| \le \frac{\delta}{\sqrt{\kappa}} \qquad \text{for all } 1 \le x \le \kappa\,.$$

So, there exists a polynomial $q_{\kappa,\delta}$ with $q_{\kappa,\delta}(x) = \hat{p}(\frac{x}{\kappa} - 1)\,\frac{1}{\sqrt{\kappa}}$, such that $q_{\kappa,\delta}$ has degree $O(\sqrt{\kappa}\log\frac{\kappa}{\delta})$ and $|q_{\kappa,\delta}(x) - x^{-1/2}| \le \delta/\sqrt{\kappa}$ for all $1 \le x \le \kappa$. □

We are now ready to prove our query complexity upper bound.

Theorem 41 (optimal algorithm for sampling from Gaussians)**.** *Let $\Lambda = \Sigma^{-1}$ be an unknown positive definite matrix with all eigenvalues between 1 and κ. Then, using $O(\min(\sqrt{\kappa}\log\frac{d}{\varepsilon}, d))$ adaptive matrix-vector queries to Λ, we can produce a sample from a distribution $\hat{\pi}$ such that $\mathsf{KL}(\hat{\pi} \,\|\, \mathcal{N}(0,\Sigma)) \le \varepsilon^2$.*

Proof. Choose $X \sim \mathcal{N}(0, I_d)$, define $R = O(\sqrt{\kappa}\log\frac{\kappa}{\delta})$ be the degree of $q_{\kappa,\delta}$, and for simplicity write $q(x) \coloneqq q_{\kappa,\delta}(x) \coloneqq \sum_{i=0}^R a_i x^i$. The algorithm works as follows. Using the power method, we compute $X, \Lambda X, \Lambda^2 X, \dots, \Lambda^R X$. We output $Y = \sum_{i=0}^R a_i\,\Lambda^i X$. Note that $Y \sim \mathcal{N}(0, \hat{\Sigma})$, where we set $\hat{\Sigma} \coloneqq (\sum_{i=0}^R a_i \Lambda^i)^2$. If $\lambda_1, \dots, \lambda_d$ denote the eigenvalues of Λ, then the eigenvalues of $\hat{\Sigma}$ are $q(\lambda_1), \dots, q(\lambda_d)$. The KL divergence is given by

$$\begin{aligned}\mathsf{KL}\bigl(\mathcal{N}(0,\hat{\Sigma}) \bigm\| \mathcal{N}(0,\Sigma)\bigr) &\lesssim \sum_{k=1}^d |q(\lambda_k)^2\,\lambda_k - 1|^2 \\ &\lesssim \sum_{k=1}^d |q(\lambda_k)\,\lambda_k^{1/2} - 1|^2 \\ &\lesssim \sum_{k=1}^d \lambda_k\,|q(\lambda_k) - \lambda_k^{-1/2}|^2 \\ &\lesssim d\kappa\,\frac{\delta^2}{\kappa}\,.\end{aligned}$$

If we set $\delta \asymp \log(\varepsilon/d)$, then we obtain a KL divergence of at most ε^2.

Finally, we can also learn Λ by querying Λe_i for each unit basis vector $e_1, \dots, e_d$. So, we can thus learn Σ, and then generate a perfect random sample from $\mathcal{N}(0,\Sigma)$. Hence, the query complexity of generating a sample from $\mathcal{N}(0,\Sigma)$ is at most $O(\min(\sqrt{\kappa}\log\frac{\kappa d}{\varepsilon}, d)) = O(\min(\sqrt{\kappa}\log\frac{d}{\varepsilon}, d))$. □

Remark 13. If π is an α-strongly log-concave distribution, then from Pinsker's inequality and Talagrand's transport inequality,

$$\max\{\|\mu - \pi\|_{\mathrm{TV}}^2,\ \alpha\, W_2^2(\mu,\pi)\} \lesssim \mathsf{KL}(\mu \,\|\, \pi)\,.$$

Hence, this algorithmic result for Gaussians complements the two lower bounds in Corollary 13.

11.6 Conclusion

We proved lower bounds for Gaussian sampling in high dimensions that combine to give a $\widetilde{\Omega}(\min(\sqrt{\kappa}\log d, d))$ rate, which is nearly tight. There are two interesting consequences of this result.

First, our lower bound shows that log-concave sampling in high dimensions indeed will have dimension dependence. This shows a difference between sampling and convex optimization in high dimensions, where the latter has dimension independent rates.

Second, we see a similarity between sampling and optimization, in that the optimal algorithms for low dimensional and high dimensional settings are qualitatively different. In contrast to the low dimensional setting, where the sampling lower bounds were achieved by rejection sampling, the nearly tight upper bound algorithm given in Theorem 41 produces a sample by taking a linear combination of a standard Gaussian vector, as well as past query points and past gradient queries. This mechanism is more similar to the gradient based sampling algorithms, such as the Langevin algorithm or Hamiltonian Monte Carlo. It will be interesting to see whether such gradient-based algorithms are optimal for high dimensional log-concave sampling in general, and what the optimal dimension dependence will be. We hope that this question can be resolved in future works.

Bibliography

[AB15] D. Alonso-Gutiérrez and J. Bastero. *Approaching the Kannan-Lovász-Simonovits and variance conjectures*. Vol. 2131. Lecture Notes in Mathematics. Springer, Cham, 2015, pp. x+148.

[AC21] K. Ahn and S. Chewi. "Efficient constrained sampling via the mirror-Langevin algorithm". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 28405–28418.

[AC23] J. M. Altschuler and S. Chewi. "Faster high-accuracy log-concave sampling via algorithmic warm starts". In: *arXiv preprint arXiv:2302.10249* (2023).

[AGS08] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[AWR17] J. Altschuler, J. Weed, and P. Rigollet. "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. 2017, pp. 1961–1971.

[Bal+22] K. Balasubramanian et al. "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 P.-L. Loh and M. Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, Feb. 2022, pp. 2896–2923.

[BBH21] M. Brennan, G. Bresler, and B. Huang. "De Finetti-style results for Wishart matrices: combinatorial structure and phase transitions". In: *arXiv e-prints*, arXiv:2103.14011 (2021).

[BC12] S. Bubeck and N. 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.

[BCG08] D. Bakry, P. Cattiaux, and A. Guillin. “Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré”. In: *Journal of Functional Analysis* 254.3 (2008), pp. 727–759.

[BCW22] A. Bakshi, K. L. Clarkson, and D. P. Woodruff. “Low-rank approximation with $1/\epsilon^{1/3}$ matrix-vector products”. In: *54th Annual ACM SIGACT Symposium on Theory of Computing.* ACM, 2022, pp. 1130–1143.

[BD01] L. J. Billera and P. Diaconis. “A geometric interpretation of the Metropolis-Hastings algorithm”. In: *Statistical Science* (2001), pp. 335–339.

[BE15] S. Bubeck and R. Eldan. “The entropic barrier: a simple and optimal universal self-concordant barrier”. In: *Conference on Learning Theory.* 2015, pp. 279–279.

[BEL18] S. Bubeck, R. Eldan, and J. Lehec. “Sampling from a log-concave distribution with projected Langevin Monte Carlo”. In: *Discrete & Computational Geometry* 59.4 (2018), pp. 757–783.

[Ber18] E. Bernton. “Langevin Monte Carlo and JKO splitting”. In: *Proceedings of the 31st Conference On Learning Theory.* Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1777–1798.

[BFG96] Z. Bai, G. Fahey, and G. Golub. “Some large-scale matrix computation problems”. In: *Journal of Computational and Applied Mathematics* 7 (1-2 1996), pp. 71–89.

[BG18] S. Bubeck and S. Ganguly. “Entropic CLT and phase transition in high-dimensional Wishart matrices”. In: *Int. Math. Res. Not. IMRN* 2 (2018), pp. 588–606.

[BGG12] F. Bolley, I. Gentil, and A. Guillin. “Convergence to equilibrium in Wasserstein distance for Fokker-Planck equations”. In: *Journal of Functional Analysis* 263.8 (2012), pp. 2430–2457.

[BGK05] A. Bovier, V. Gayrard, and M. 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] D. Bakry, I. Gentil, and M. 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.

[BH97] S. G. Bobkov and C. Houdré. “Some connections between isoperimetric and Sobolev-type inequalities”. In: *Mem. Amer. Math. Soc.* 129.616 (1997), pp. viii+111.

[BKM22] V. Braverman, A. Krishnan, and C. Musco. “Sublinear time spectral density estimation”. In: *54th Annual ACM SIGACT Symposium on Theory of Computing.* ACM, 2022, pp. 1144–1157.

[BL00] S. G. Bobkov and M. Ledoux. “From Brunn-Minkowski to Brascamp-Lieb and to logarithmic Sobolev inequalities”. In: *Geom. Funct. Anal.* 10.5 (2000), pp. 1028–1052.

[BL06] J. M. Borwein and A. 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] H. J. Brascamp and E. 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] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities.* A nonasymptotic theory of independence, With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013, pp. x+481.

[BM20] S. Bubeck and D. Mikulincer. “How to trap a gradient flow”. In: *Proceedings of Thirty Third Conference on Learning Theory.* Ed. by J. Abernethy and S. Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, Sept. 2020, pp. 940–960.

[Bob03] S. 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] S. G. Bobkov. “Isoperimetric and analytic inequalities for log-concave probability measures”. In: *The Annals of Probability* 27.4 (1999), pp. 1903–1921.

[Bov+02] A. Bovier et al. “Metastability and low lying spectra in reversible Markov chains”. In: *Comm. Math. Phys.* 228.2 (2002), pp. 219–255.

[Bov+04] A. Bovier et al. “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.

[BP17] K. Bringmann and K. Panagiotou. "Efficient sampling methods for discrete distributions". In: *Algorithmica* 79.2 (2017), pp. 484–508.

[Bra+20] M. Braverman et al. "The gradient complexity of linear regression". In: *Conference on Learning Theory, (COLT)*. Vol. 125. Proceedings of Machine Learning Research. PMLR, 2020, pp. 627–647.

[Bro+19] N. Brosse et al. "The tamed unadjusted Langevin algorithm". In: *Stochastic Processes and their Applications* 129.10 (2019), pp. 3638–3663.

[Bub+16] S. Bubeck et al. "Testing for high-dimensional geometry in random graphs". In: *Random Structures Algorithms* 49.3 (2016), pp. 503–532.

[Bub15] S. Bubeck. "Convex optimization: algorithms and complexity". In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357.

[Bur73] D. L. Burkholder. "Distribution function inequalities for martingales". In: *Ann. Probability* 1 (1973), pp. 19–42.

[Caf00] L. A. Caffarelli. "Monotonicity properties of optimal transportation and the FKG and related inequalities". In: *Comm. Math. Phys.* 214.3 (2000), pp. 547–563.

[Car+20] Y. Carmon et al. "Lower bounds for finding stationary points I". In: *Math. Program.* 184.1-2, Ser. A (2020), pp. 71–120.

[Car+21] Y. Carmon et al. "Lower bounds for finding stationary points II: first-order methods". In: *Math. Program.* 185.1-2, Ser. A (2021), pp. 315–355.

[Car11] E. A. Carlen. "Functional inequalities and dynamics". In: *Nonlinear PDE's and Applications*. Springer, 2011, pp. 17–85.

[CB18] X. Cheng and P. Bartlett. "Convergence of Langevin MCMC in KL-divergence". In: *Algorithmic Learning Theory 2018*. Vol. 83. Proc. Mach. Learn. Res. (PMLR). Proceedings of Machine Learning Research PMLR, 2018, p. 26.

[CBL22] N. S. Chatterji, P. L. Bartlett, and P. M. Long. "Oracle lower bounds for stochastic gradient sampling algorithms". In: *Bernoulli* 28.2 (2022), pp. 1074–1092.

[CBS22] S. Chewi, S. Bubeck, and A. Salim. "On the complexity of finding stationary points of smooth functions in one dimension". In: *arXiv e-prints*, arXiv:2209.07513 (2022).

[CG09] P. Cattiaux and A. Guillin. "Trends to equilibrium in total variation distance". In: *Ann. Inst. Henri Poincaré Probab. Stat.* 45.1 (2009), pp. 117–145.

[CGT00] A. R. Conn, N. I. Gould, and P. L. Toint. *Trust region methods.* Vol. 1. SIAM, 2000.

[Che+18a] X. Cheng et al. "Sharp convergence rates for Langevin dynamics in the nonconvex setting". In: *arXiv e-prints*, arXiv:1805.01648 (2018).

[Che+18b] X. Cheng et al. "Underdamped Langevin MCMC: A non-asymptotic analysis". In: *Proceedings of the 31st Conference On Learning Theory.* Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 300–323.

[Che+18c] X. Cheng et al. "Underdamped Langevin MCMC: A non-asymptotic analysis". In: *Proceedings of the 31st Conference On Learning Theory.* Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 300–323.

[Che+19] X. Cheng et al. "Quantitative W_1 convergence of Langevin-like stochastic processes with non-convex potential and state-dependent noise". In: *arXiv e-prints*, arXiv:1907.03215 (July 2019).

[Che+20a] Y. Chen et al. "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] S. Chewi et al. "Exponential ergodicity of mirror-Langevin diffusions". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 19573–19585.

[Che+20c] S. Chewi et al. "Gradient descent algorithms for Bures-Wasserstein barycenters". In: *Proceedings of Thirty Third Conference on Learning Theory.* Ed. by J. Abernethy and S. Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, Sept. 2020, pp. 1276–1304.

[Che+20d] S. Chewi et al. "SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 2098–2109.

[Che+21] S. Chewi et al. "Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm". In: *Conference on Learning Theory.* PMLR. 2021, pp. 1260–1300.

[Che+22a] Y. Chen et al. “Improved analysis for a proximal algorithm for sampling”. In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by P.-L. Loh and M. Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, Feb. 2022, pp. 2984–3014.

[Che+22b] S. Chewi et al. “Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev”. In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by P.-L. Loh and M. Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, Feb. 2022, pp. 1–2.

[Che+22c] S. Chewi et al. “Rejection sampling from shape-constrained distributions in sublinear time”. In: *International conference on artificial intelligence and statistics*. PMLR. 2022, pp. 2249–2265.

[Che+22d] S. Chewi et al. “The query complexity of sampling from strongly log-concave distributions in one dimension”. In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by P.-L. Loh and M. Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, Feb. 2022, pp. 2041–2059.

[Che+23a] S. Chewi et al. “Fisher information lower bounds for sampling”. In: *Proceedings of the 34th International Conference on Algorithmic Learning Theory*. Ed. by S. Agrawal and F. Orabona. Vol. 201. Proceedings of Machine Learning Research. PMLR, Feb. 2023, pp. 375–410.

[Che+23b] S. Chewi et al. “Query lower bounds for log-concave sampling”. In: *arXiv preprint arXiv:2304.02599* (2023).

[CL89] M.-F. Chen and S.-F. Li. “Coupling methods for multidimensional diffusion processes”. In: *The Annals of Probability* (1989), pp. 151–177.

[CLL19] Y. Cao, J. Lu, and Y. Lu. “Exponential decay of Rényi divergence under Fokker-Planck equations”. In: *J. Stat. Phys.* 176.5 (2019), pp. 1172–1184.

[CLW21] Y. Cao, J. Lu, and L. Wang. “Complexity of randomized algorithms for underdamped Langevin dynamics”. In: *Commun. Math. Sci.* 19.7 (2021), pp. 1827–1853.

[Coh+18] D. Cohen-Steiner et al. “Approximating the spectrum of a graph”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* ACM, 2018, pp. 1263–1271.

[CT02] J. A. Carrillo and G. Toscani. “Long-time asymptotics for strong solutions of the thin film equation”. In: *Comm. Math. Phys.* 225.3 (2002), pp. 551–571.

[CT06] T. M. Cover and J. A. Thomas. *Elements of information theory.* Second. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006, pp. xxiv+748.

[Cut13] M. Cuturi. “Sinkhorn distances: lightspeed computation of optimal transport”. In: *Advances in Neural Information Processing Systems 26.* Ed. by C. J. C. Burges et al. Curran Associates, Inc., 2013, pp. 2292–2300.

[Dal17a] A. 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 S. Kale and O. Shamir. Vol. 65. Proceedings of Machine Learning Research. Amsterdam, Netherlands: PMLR, 2017, pp. 678–689.

[Dal17b] A. S. Dalalyan. “Theoretical guarantees for approximate sampling from smooth and log-concave densities”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.3 (2017), pp. 651–676.

[Dal17c] A. 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] B. Davis. “On the L^p norms of stochastic integrals and other martingales”. In: *Duke Math. J.* 43.4 (1976), pp. 697–704.

[Det+18] G. Detommaso et al. “A Stein variational Newton method”. In: *Advances in Neural Information Processing Systems.* 2018, pp. 9169–9179.

[Dev87] L. Devroye. “A simple generator for discrete log-concave distributions”. In: *Computing* 39.1 (1987), pp. 87–91.

[Din15] Y. Ding. “A note on quadratic transportation and divergence inequality”. In: *Statist. Probab. Lett.* 100 (2015), pp. 115–123.

[DK19] A. S. Dalalyan and A. Karagulyan. “User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient”. In: *Stoch. Proc. Appl.* 129.12 (2019), pp. 5278–5311.

[DKR19] A. S. Dalalyan, A. Karagulyan, and L. Riou-Durand. “Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets”. In: *arXiv e-prints*, arXiv:1906.08530 (June 2019).

[DM+19] A. Durmus, E. Moulines, et al. “High-dimensional Bayesian inference via the unadjusted Langevin algorithm”. In: *Bernoulli* 25.4A (2019), pp. 2854–2882.

[DM15] A. Durmus and É. Moulines. “Quantitative bounds of convergence for geometrically ergodic Markov chain in the Wasserstein distance with application to the Metropolis adjusted Langevin algorithm”. In: *Statistics and Computing* 25.1 (Jan. 2015), pp. 5–19.

[DM17] A. Durmus and É. Moulines. “Nonasymptotic convergence analysis for the unadjusted Langevin algorithm”. In: *Ann. Appl. Probab.* 27.3 (2017), pp. 1551–1587.

[DM21] P. Dharangutte and C. Musco. “Dynamic trace estimation”. In: *Advances in Neural Information Processing Systems 34*. 2021, pp. 30088–30099.

[DMM19] A. Durmus, S. Majewski, and B. Miasojedow. “Analysis of Langevin Monte Carlo via convex optimization”. In: *J. Mach. Learn. Res.* 20 (2019), Paper No. 73, 46.

[DNS19] A. Duncan, N. Nuesken, and L. Szpruch. “On the geometry of Stein variational gradient descent”. In: *arXiv e-prints*, arXiv:1912.00894 (Dec. 2019).

[DR20] A. S. Dalalyan and L. Riou-Durand. “On sampling from a log-concave density using kinetic Langevin diffusions”. In: *Bernoulli* 26.3 (2020), pp. 1956–1988.

[DRK19] A. S. Dalalyan, L. Riou-Durand, and A. Karagulyan. “Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets”. In: *arXiv e-prints*, arxiv:1906.08530 (June 2019).

[DT12] A. S. Dalalyan and A. B. Tsybakov. “Sparse regression learning by aggregation and Langevin Monte-Carlo”. In: *Journal of Computer and System Sciences* 78.5 (2012), pp. 1423–1443.

[Dvi09] Z. Dvir. “On the size of Kakeya sets in finite fields”. In: *Journal of the American Mathematical Society* 22.4 (2009), pp. 1093–1097.

[Dwi+19] R. Dwivedi et al. “Log-concave sampling: Metropolis-Hastings algorithms are fast”. In: *Journal of Machine Learning Research* 20.183 (2019), pp. 1–42.

[Ebe16] A. Eberle. “Reflection couplings and contraction rates for diffusions”. In: *Probability Theory and Related Fields* 166.3-4 (2016), pp. 851–886.

[Eck87] R. Eckhardt. “Stan Ulam, John von Neumann, and the Monte Carlo method”. In: *Los Alamos Science* 15 (1987), pp. 131–136.

[Ede89] A. Edelman. “Eigenvalues and condition numbers of random matrices”. PhD thesis. Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, 1989.

[EH21] M. A. Erdogdu and R. 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 M. Belkin and S. Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, 15–19 Aug 2021, pp. 1776–1822.

[Eva10] L. C. Evans. *Partial differential equations*. Second. Vol. 19. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2010, pp. xxii+749.

[FGP20] M. Fathi, N. Gozlan, and M. 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] A. Frieze, R. Kannan, and N. Polson. “Sampling from log-concave distributions”. In: *The Annals of Applied Probability* 4.3 (1994), pp. 812–837.

[Fri34] K. Friedrichs. “Spektraltheorie Halbbeschränkter Operatoren Und Anwendung Auf Die Spektralzerlegung von Differentialoperatoren”. In: *Mathematische Annalen* 109.1 (1934), pp. 465–487.

[Gen08] I. Gentil. “From the Prékopa-Leindler inequality to modified logarithmic Sobolev inequality”. In: *Ann. Fac. Sci. Toulouse Math. (6)* 17.2 (2008), pp. 291–308.

[GJ14] P. Groeneboom and G. Jongbloed. *Nonparametric estimation under shape constraints*. Vol. 38. Cambridge University Press, 2014.

[GLL20] R. Ge, H. Lee, and J. Lu. “Estimating normalizing constants for log-concave distributions: Algorithms and lower bounds”. In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. 2020, pp. 579–586.

[GLL22] S. Gopi, Y. T. Lee, and D. Liu. “Private convex optimization via exponential mechanism”. In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by P.-L. Loh and M. Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, Feb. 2022, pp. 1948–1989.

[Gol10] O. Goldreich. *Property testing—current research and surveys*. Vol. 6390. Lecture Notes in Computer Science. Springer, 2010.

[Gol17] O. Goldreich. *Introduction to property testing*. Cambridge University Press, 2017.

[Gor41] R. D. Gordon. "Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument". In: *The Annals of Mathematical Statistics* 12.3 (1941), pp. 364–366.

[Gra05] P. Graham. 2005. URL: http://www.paulgraham.com/college.html.

[GT20] A. Ganesh and K. Talwar. "Faster differentially private samplers via Rényi divergence analysis of discretized Langevin MCMC". In: *arXiv e-prints*, arXiv:2010.14658 (Oct. 2020).

[Gui+09] A. Guillin et al. "Transportation-information inequalities for Markov processes". In: *Probab. Theory Related Fields* 144.3-4 (2009), pp. 669–695.

[Han16] R. van Handel. *Probability in high dimension*. 2016.

[Has70] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1 (1970), pp. 97–109.

[HBE22] Y. He, K. Balasubramanian, and M. A. Erdogdu. "Heavy-tailed sampling via transformed unadjusted Langevin algorithm". In: *arXiv e-prints*, arXiv:2201.08349 (2022).

[HS86] R. Holley and D. W. Stroock. "Logarithmic Sobolev inequalities and stochastic Ising models". In: (1986).

[Hsi+18] Y.-P. Hsieh et al. "Mirrored Langevin dynamics". In: *Advances in Neural Information Processing Systems*. 2018, pp. 2878–2887.

[HSV14] M. Hairer, A. M. Stuart, and S. J. Vollmer. "Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions". In: *The Annals of Applied Probability* 24.6 (2014), pp. 2455–2490.

[Hut90] M. F. Hutchinson. "A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines". In: *Communications in Statistics-Simulation and Computation* 19 (2 1990), pp. 433–450.

[JKO98] R. Jordan, D. Kinderlehrer, and F. Otto. "The variational formulation of the Fokker-Planck equation". In: *SIAM Journal on Mathematical Analysis* 29.1 (1998), pp. 1–17.

[JL15] T. Jiang and D. Li. "Approximation of rectangular beta-Laguerre ensembles and large deviations". In: *J. Theoret. Probab.* 28.3 (2015), pp. 804–847.

[Juk11] S. Jukna. *Extremal combinatorics: with applications in computer science*. Vol. 571. Springer, 2011.

[KLS95] R. Kannan, L. Lovász, and M. Simonovits. "Isoperimetric problems for convex bodies and a localization lemma". In: *Discrete Comput. Geom.* 13.3-4 (1995), pp. 541–559.

[KNS16] H. Karimi, J. Nutini, and M. Schmidt. "Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2016, pp. 795–811.

[KS91] I. Karatzas and S. E. Shreve. *Brownian motion and stochastic calculus*. Second. Vol. 113. Graduate Texts in Mathematics. Springer-Verlag, New York, 1991, pp. xxiv+470.

[KS98] I. Karatzas and S. E. Shreve. "Brownian motion". In: *Brownian Motion and Stochastic Calculus*. Springer, 1998, pp. 47–127.

[Le 16] J.-F. Le Gall. *Brownian motion, martingales, and stochastic calculus*. French. Vol. 274. Graduate Texts in Mathematics. Springer, [Cham], 2016, pp. xiii+273.

[Led18] M. Ledoux. *Remarks on some transportation cost inequalities*. 2018.

[Li+19] X. Li et al. "Stochastic Runge–Kutta accelerates Langevin Monte Carlo and beyond". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.

[Li+20] L. Li et al. "A stochastic version of Stein variational gradient descent for efficient sampling". In: *Commun. Appl. Math. Comput. Sci.* 15.1 (2020), pp. 37–63.

[Liu+19] C. Liu et al. "Understanding and accelerating particle-based variational inference". In: *International Conference on Machine Learning*. 2019, pp. 4082–4092.

[Liu17] Q. Liu. "Stein variational gradient descent as gradient flow". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 3115–3123.

[Liu20] Y. Liu. "The Poincaré inequality and quadratic transportation-variance inequalities". In: *Electron. J. Probab.* 25 (2020), Paper No. 1, 16.

[LL08] C. Le Bris and P.-L. 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.

[LLN19] J. Lu, Y. Lu, and J. Nolen. “Scaling limit of the Stein variational gradient descent: The mean field regime”. In: *SIAM Journal on Mathematical Analysis* 51.2 (2019), pp. 648–671.

[LLT22] H. Lee, J. Lu, and Y. Tan. “Convergence for score-based generative modeling with polynomial complexity”. In: *Advances in Neural Information Processing Systems.* Ed. by A. H. Oh et al. 2022.

[LLT23] H. Lee, J. Lu, and Y. Tan. “Convergence of score-based generative modeling for general data distributions”. In: *Algorithmic Learning Theory.* Ed. by S. Agrawal and F. Orabona. PMLR. 2023.

[Loj63] S. Lojasiewicz. “Une propriété topologique des sous-ensembles analytiques réels”. In: *Les équations aux dérivées partielles* 117 (1963), pp. 87–89.

[LRG18] H. Lee, A. Risteski, and R. 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 et al. Vol. 31. Curran Associates, Inc., 2018.

[LS88] G. F. Lawler and A. 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] L. Lovász and M. Simonovits. “Random walks in a convex body and an improved volume algorithm”. In: *Random Structures & Algorithms* 4.4 (1993), pp. 359–412.

[LST21a] Y. T. Lee, R. Shen, and K. Tian. “Lower bounds on Metropolized sampling methods for well-conditioned distributions”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 18812–18824.

[LST21b] Y. T. Lee, R. Shen, and K. Tian. “Lower bounds on Metropolized sampling methods for well-conditioned distributions”. In: *Advances in Neural Information Processing Systems.* Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 18812–18824.

[LTV20] A. Laddha, Y. Tat Lee, and S. Vempala. “Strong self-concordance and sampling”. In: *STOC* (2020).

[LV06] L. Lovász and S. Vempala. “Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithm”. In: *J. Comput. System Sci.* 72.2 (2006), pp. 392–417.

[LV07] L. Lovász and S. Vempala. “The geometry of logconcave functions and sampling algorithms”. In: *Random Structures & Algorithms* 30.3 (2007), pp. 307–358.

[LV17] Y. T. Lee and S. S. Vempala. “Eldan’s stochastic localization and the KLS hyperplane conjecture: an improved lower bound for expansion”. In: *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*. IEEE Computer Soc., Los Alamitos, CA, 2017, pp. 998–1007.

[LV18a] Y. T. Lee and S. S. Vempala. “Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation”. In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. 2018, pp. 1115–1121.

[LV18b] Y. T. Lee and S. 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.

[LW16] Q. Liu and D. Wang. “Stein variational gradient descent: A general purpose Bayesian inference algorithm”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 2378–2386.

[Ma+19] Y.-A. Ma et al. “Is there an analog of Nesterov acceleration for MCMC?” In: *arXiv e-prints*, arXiv:1902.00996 (Feb. 2019).

[Ma+21] Y.-A. Ma et al. “Is there an analog of Nesterov acceleration for gradient-based MCMC?” In: *Bernoulli* 27.3 (2021), pp. 1942–1992.

[Mar+12] J. Martin et al. “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.

[Met+53] N. Metropolis et al. “Equation of state calculations by fast computing machines”. In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092.

[Mey+21] R. A. Meyer et al. “Hutch++: optimal stochastic trace estimation”. In: *4th Symposium on Simplicity in Algorithms*. SIAM, 2021, pp. 142–155.

[Mik22] D. Mikulincer. “A CLT in Stein’s distance for generalized Wishart matrices and higher-order tensors”. In: *Int. Math. Res. Not. IMRN* 10 (2022), pp. 7839–7872.

[MM15] C. Musco and C. Musco. “Randomized block Krylov methods for stronger and faster approximate singular value decomposition”. In: *Advances in Neural Information Processing Systems 28*. 2015, pp. 1396–1404.

[MMS09] D. Matthes, R. J. McCann, and G. Savaré. “A family of nonlinear fourth order equations of gradient flow type”. In: *Comm. Partial Differential Equations* 34.10-12 (2009), pp. 1352–1397.

[MMS20] M. B. Majka, A. Mijatović, and Ł. Szpruch. “Nonasymptotic bounds for sampling algorithms without log-concavity”. In: *Ann. Appl. Probab.* 30.4 (2020), pp. 1534–1581.

[Mou+19] W. Mou et al. “Improved bounds for discretization of Langevin diffusions: Near-optimal rates without convexity”. In: *arXiv e-prints*, arXiv:1907.11331 (July 2019).

[Mou+20] W. Mou et al. *High-order Langevin diffusion yields an accelerated MCMC algorithm.* 2020. arXiv: 1908.10859 [stat.ML].

[MT09] S. Meyn and R. L. Tweedie. *Markov chains and stochastic stability.* 2nd. USA: Cambridge University Press, 2009.

[MV18] O. Mangoubi and N. K. Vishnoi. “Dimensionally tight bounds for second-order Hamiltonian Monte Carlo”. In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., 2018, pp. 6027–6037.

[MV19] O. Mangoubi and N. K. Vishnoi. “Nonconvex sampling with the Metropolis-adjusted Langevin algorithm”. In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by A. Beygelzimer and D. Hsu. Vol. 99. Proceedings of Machine Learning Research. Phoenix, USA: PMLR, 25–28 Jun 2019, pp. 2259–2293.

[Nea12] R. Neal. “MCMC using Hamiltonian dynamics”. In: *Handbook of Markov Chain Monte Carlo* (June 2012).

[Nem94] A. Nemirovski. *Efficient methods in convex programming*. Lecture Notes (Tecehnion - Georgia Tech). 1994.

[Nes04] Y. Nesterov. *Introductory lectures on convex optimization.* Vol. 87. Applied Optimization. A basic course. Kluwer Academic Publishers, Boston, MA, 2004, pp. xviii+236.

[Nes13] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2013.

[Nes18] Y. Nesterov. *Lectures on convex optimization.* Vol. 137. Springer Optimization and Its Applications. Springer, Cham, 2018, pp. xxiii+589.

[Neu51] J. 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.

[NJ79] A. S. Nemirovskii and D. B. Judin. *Complexity of problems and efficiency of optimization methods.* 1979.

[NN94] Y. Nesterov and A. Nemirovskii. *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] Y. Nesterov and B. T. Polyak. "Cubic regularization of Newton method and its global performance". In: *Mathematical Programming* 108.1 (2006), pp. 177–205.

[NS22] A. Nishimura and M. A. Suchard. "Prior-preconditioned conjugate gradient method for accelerated Gibbs sampling in "large n, large p" Bayesian sparse regression". In: *Journal of the American Statistical Association* 0.0 (2022), pp. 1–14.

[NY83] A. S. Nemirovsky and D. B. a. Yudin. *Problem complexity and method efficiency in optimization.* A Wiley-Interscience Publication. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Inc., New York, 1983, pp. xv+388.

[OT11] S.-i. Ohta and A. Takatsu. "Displacement convexity of generalized relative entropies". In: *Adv. Math.* 228.3 (2011), pp. 1742–1787.

[OT13] S.-i. Ohta and A. Takatsu. "Displacement convexity of generalized relative entropies. II". In: *Comm. Anal. Geom.* 21.4 (2013), pp. 687–785.

[OV00] F. Otto and C. Villani. "Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality". In: *Journal of Functional Analysis* 173.2 (2000), pp. 361–400.

[Pav14] G. A. Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations.* Vol. 60. Springer, 2014.

[Per28] O. Perron. “Über einen Satz von Besicovitsch”. In: *Mathematische Zeitschrift* 28.1 (1928), pp. 383–386.

[PST12] N. S. Pillai, A. M. Stuart, and A. H. Thiéry. “Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions”. In: *Ann. Appl. Probab.* 22.6 (2012), pp. 2320–2356.

[RGG+97] G. O. Roberts, A. Gelman, W. R. Gilks, et al. “Weak convergence and optimal scaling of random walk Metropolis algorithms”. In: *The Annals of Applied Probability* 7.1 (1997), pp. 110–120.

[Roc97] R. T. Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Reprint of the 1970 original, Princeton Paperbacks. Princeton University Press, Princeton, NJ, 1997, pp. xviii+451.

[RR19] M. Z. Rácz and J. Richey. “A smooth transition from Wishart to GOE”. In: *J. Theoret. Probab.* 32.2 (2019), pp. 898–906.

[RR98] G. O. Roberts and J. S. Rosenthal. “Optimal scaling of discrete approximations to Langevin diffusions”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.1 (1998), pp. 255–268.

[RRT17] M. Raginsky, A. Rakhlin, and M. Telgarsky. “Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis”. In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by S. Kale and O. Shamir. Vol. 65. Proceedings of Machine Learning Research. Amsterdam, Netherlands: PMLR, July 2017, pp. 1674–1703.

[RS02] G. O. Roberts and O. Stramer. “Langevin diffusions and Metropolis-Hastings algorithms”. In: *Methodology and Computing in Applied Probability* 4.4 (2002), pp. 337–357.

[RS78] M. Reed and B. Simon. *Methods of modern mathematical physics. IV. Analysis of operators*. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1978, pp. xv+396.

[RV08] L. Rademacher and S. Vempala. “Dispersion of mass and the complexity of randomized geometric algorithms”. In: *Adv. Math.* 219.3 (2008), pp. 1037–1069.

[RWZ20] C. Rashtchian, D. P. Woodruff, and H. Zhu. “Vector-matrix-vector queries for solving linear algebra, statistics, and graph problems”. In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*. Vol. 176. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020, 26:1–26:20.

[San15] F. 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] F. Santambrogio. "{Euclidean, metric, and Wasserstein} gradient flows: an overview". In: *Bulletin of Mathematical Sciences* 7.1 (2017), pp. 87–154.

[SAR18] M. Simchowitz, A. E. Alaoui, and B. Recht. "Tight query complexity lower bounds for PCA via finite sample deformed Wigner law". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2018, pp. 1249–1259.

[Sim+16] U. Simsekli et al. "Stochastic quasi-Newton Langevin Monte Carlo". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16. New York, NY, USA: JMLR.org, 2016, pp. 642–651.

[SKL20] A. Salim, A. Korba, and G. Luise. "Wasserstein proximal gradient". In: *arXiv e-prints*, arxiv:2002.03035 (Feb. 2020).

[SL19] R. Shen and Y. T. Lee. "The randomized midpoint method for log-concave sampling". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.

[SS08] S. Saraf and M. Sudan. "An improved lower bound on the size of Kakeya sets over finite fields". In: *Analysis & PDE* 1.3 (2008), pp. 375–379.

[SS11] M. J. Silvapulle and P. K. Sen. *Constrained statistical inference: Order, inequality, and shape constraints*. Vol. 912. John Wiley & Sons, 2011.

[Str18] D. W. Stroock. *Elements of stochastic calculus and analysis*. Springer, 2018.

[Sun+19] X. Sun et al. "Querying a matrix through matrix-vector products". In: *46th International Colloquium on Automata, Languages, and Programming*. Vol. 132. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019, 94:1–94:16.

[SV06] D. W. Stroock and S. R. S. Varadhan. *Multidimensional diffusion processes*. Classics in Mathematics. Reprint of the 1997 edition. Springer-Verlag, Berlin, 2006, pp. xii+338.

[SV14] S. Sachdeva and N. K. Vishnoi. "Faster algorithms via approximation theory". In: *Found. Trends Theor. Comput. Sci.* 9.2 (2014), pp. 125–210.

[SW14] A. Saumard and J. A. Wellner. "Log-concavity and strong log-concavity: a review". In: *Stat. Surv.* 8 (2014), pp. 45–114.

[Sza91] S. J. Szarek. "Condition numbers of random matrices". In: *J. Complexity* 7.2 (1991), pp. 131–149.

[Tal19] K. Talwar. "Computational separations between sampling and optimization". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.

[Tsy09] A. 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.

[TY18] Y. Tat Lee and M.-C. Yue. "Universal barrier is n-self-concordant". In: *arXiv e-prints*, arxiv:1809.03011 (Sept. 2018).

[Ver18] R. 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] C. Villani. *Topics in optimal transportation.* Vol. 58. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2003, pp. xvi+370.

[Vil09] C. Villani. *Optimal transport.* Vol. 338. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Old and new. Springer-Verlag, Berlin, 2009, pp. xxii+973.

[VW19] S. Vempala and A. Wibisono. "Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices". In: *Advances in Neural Information Processing Systems 32.* Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8094–8106.

[Wai19] M. 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.

[Wan+19] D. Wang et al. "Stein variational gradient descent with matrix-valued kernels". In: *Advances in Neural Information Processing Systems 32.* Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 7836–7846.

[Wib18] A. Wibisono. "Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem". In: *Conference on Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2093–3027.

[Wib19] A. Wibisono. "Proximal Langevin algorithm: rapid convergence under isoperimetry". In: *arXiv preprint arXiv:1911.01469* (2019).

[WL20] Y. Wang and W. Li. "Information Newton's flow: second-order optimization method in probability space". In: *arXiv e-prints*, arxiv:2001.04341 (Jan. 2020).

[Woo14] D. P. Woodruff. "Sketching as a tool for numerical linear algebra". In: *Found. Trends Theor. Comput. Sci.* 10.1-2 (2014), pp. 1–157.

[WS17] B. Woodworth and N. Srebro. "Lower bound for randomized first order convex optimization". In: *arXiv e-prints*, arXiv:1709.03594 (2017).

[WSC22] K. Wu, S. Schmidler, and Y. Chen. "Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling". In: *The Journal of Machine Learning Research* 23.1 (2022), pp. 12348–12410.

[WWZ14] K. Wimmer, Y. Wu, and P. Zhang. "Optimal query complexity for estimating the trace of a matrix". In: *41st International Colloquium on Automata, Languages, and Programming*. Vol. 8572. Lecture Notes in Computer Science. Springer, 2014, pp. 1051–1062.

[Xu+18] P. Xu et al. "Global convergence of Langevin dynamics based algorithms for nonconvex optimization". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018.

[Yao77] A. C. C. Yao. "Probabilistic computations: toward a unified measure of complexity (extended abstract)". In: *18th Annual Symposium on Foundations of Computer Science (Providence, R.I., 1977)*. 1977, pp. 222–227.

[Zha+20a] J. Zhang et al. "Stochastic particle-optimization sampling and the non-asymptotic convergence theory". In: ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. Online: PMLR, 26–28 Aug 2020, pp. 1877–1887.

[Zha+20b] K. S. Zhang et al. “Wasserstein control of mirror Langevin Monte Carlo”. In: *arXiv e-prints*, arxiv:2002.04363 (Feb. 2020).

[ZWG13] T. Zhang, A. Wiesel, and M. S. Greco. “Multivariate generalized Gaussian distribution: convexity and graphical models”. In: *IEEE Transactions on Signal Processing* 61.16 (2013), pp. 4141–4148.

[ZXG21] D. Zou, P. Xu, and Q. 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 C. de Campos and M. H. Maathuis. Vol. 161. Proceedings of Machine Learning Research. PMLR, 27–30 Jul 2021, pp. 1152–1162.