Efficient Sampling Using Markov Chain Monte Carlo Methods

Ruoqi Shen

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Yin Tat Lee, Chair

Shayan Oveis Gharan

Simon S. Du

Program Authorized to Offer Degree:
Computer Science & Engineering

University of Washington

Abstract

Efficient Sampling Using Markov Chain Monte Carlo Methods

Ruoqi Shen

Chair of the Supervisory Committee:
Yin Tat Lee
Computer Science & Engineering

This thesis explores the challenge of efficient sampling from target distributions, a problem at the heart of statistics, machine learning, and theoretical computer science with applications in Bayesian estimation, volume computation, and bandit optimization. The focus is on designing optimal samplers using the Markov Chain Monte Carlo (MCMC) method, leveraging a gradient or value oracle for a given smooth function, aiming to match the output distribution closely to the target distribution without direct access to the density function or its normalization constant. The research addresses the inefficiencies of current algorithms, especially in high-dimensional, ill-conditioned, structured, or constrained distributions, and introduces novel sampling algorithms that optimize query complexity.

The thesis is structured into four main parts: The first part presents an improved discretization method for simulating stochastic differential equations like Langevin Diffusion, significantly enhancing sampling efficiency. The second part examines Metropolized Sampling Algorithms, offering new insights into their query complexity and establishing upper and lower bounds for widely used algorithms like Metropolized Hamiltonian Monte Carlo and Metropolis-adjusted Langevin Dynamics. The third part introduces a proximal sampler, improving condition number dependence and presenting efficient algorithms for various structured log-concave families. Finally, the fourth part tackles the challenging task of sampling from constrained sets, overcoming the difficulties posed by maintaining the random walk within the constraints and slow convergence rates in ill-conditioned sets.

This work demonstrates theoretical advancements and practical efficiency in sampling algorithms. It contributes to understanding the fundamental aspects of sampling, the intricacies of discretization errors, and the impact of condition numbers on sampling complexity. This work improves over existing sampling methods, particularly in sampling from

high-dimensional, constrained, ill-conditioned, and structured distributions.

TABLE OF CONTENTS

Page

List of Figures iv

Chapter 1: Introduction 1

Part I: Efficient Discretization for Samplers 8

Chapter 2: Randomized Midpoint Method 9
2.1 Introduction 9
2.2 Preliminary 12
2.3 Randomized Midpoint Method 14
2.4 Numerical Experiments 19

Part II: Metropolized Sampling Algorithms 22

Chapter 3: Logsmooth Gradient Concentration and Runtimes Upper Bounds . 23
3.1 Introduction 23
3.2 Preliminaries 29
3.3 Gradient concentration 32
3.4 Mixing time bounds via blocking conductance 34

Chapter 4: Lower Bounds on Metropolized Sampling Methods for Well-Conditioned Distributions 39
4.1 Introduction 39
4.2 Preliminaries 46
4.3 Lower bound for MALA on Gaussians 50
4.4 Lower bound for MALA on well-conditioned distributions 57
4.5 Mixing time lower bound for MALA 64
4.6 Lower bounds for HMC 67
4.7 Conclusion 76

Part III: Proximal Sampling Method 78

Chapter 5: Structured Logconcave Sampling using Proximal Sampling Methods 79
5.1 Introduction 79
5.2 Preliminaries 95

5.3 Proximal reduction framework . . . 97
5.4 Tighter runtimes for structured densities . . . 102
5.5 Composite logconcave sampling with a restricted Gaussian oracle . . . 109
5.6 Logconcave finite sums . . . 113

Chapter 6: Algorithmic Aspects of the Log-Laplace Transform and a Non-Euclidean Proximal Sampler . . . 122
6.1 Introduction . . . 122
6.2 Preliminaries . . . 133
6.3 Properties of the LLT . . . 135
6.4 Proximal LLT sampler . . . 141
6.5 Applications . . . 149
6.6 Conclusion . . . 159

Part IV: Sampling in a Constrained Space . . . 161

Chapter 7: Sampling using Riemannian Hamiltonian Monte Carlo with Condition-number-independent Convergence Rate . . . 162
7.1 Introduction . . . 162
7.2 Constrained RHMC . . . 166
7.3 Experiments . . . 171

Bibliography . . . 177

Appendix A: Deferred contents from Chapter 2 . . . 202
A.1 Brownian Motion Simulation . . . 202
A.2 Properties of the ULD and the Brownian motion . . . 203
A.3 Discretization Error of Algorithm 1 . . . 207
A.4 Bounds on $\|\nabla f(x)\|$ and $\|v\|$ . . . 210
A.5 Proof of Theorem 3 . . . 216
A.6 Discretization Error of Algorithm 2 . . . 218

Appendix B: Deferred contents from Chapter 3 . . . 223
B.1 Equivalence of HMC and Metropolis-adjusted Langevin dynamics . . . 223
B.2 Improved concentration under Hessian log-Sobolev inequality . . . 224
B.3 Mixing time proofs . . . 225
B.4 Total variation bounds . . . 234
B.5 Deferred proofs . . . 237

Appendix C: Deferred contents from Chapter 4 . . . 238
C.1 Necessity of fixing a scale . . . 238
C.2 HMC lower bounds beyond $\kappa\sqrt{d}$ . . . 238

Appendix D: Deferred contents from Chapter 5 247
D.1 Discussion of inexactness tolerance 247
D.2 Deferred proofs from Section 5.5 248
D.3 Mixing time ingredients 258
D.4 Structural results 265

Appendix E: Deferred contents from Chapter 6 270
E.1 Information-theoretic lower bound 270
E.2 Lower bound on the range of $\psi_{1,1}$ 274
E.3 Deferred proofs from Section 6.4 275

Appendix F: Deferred contents from Chapter 7 278
F.1 Additional Experiment Details 278
F.2 Deferred details of CRHMC 281
F.3 Discretization 290
F.4 Condition Number Independence via Self-concordant Barrier 303
F.5 Missing Notations and Definitions 313

LIST OF FIGURES

Figure Number Page

2.1 Error of random walks with different choice of step size. 20

4.1 Second derivative of our hard function f_{1d}, $\kappa = 10$, $h = 0.01$. Starting from inside the hard region, on average over $g \sim \mathcal{N}(0, \mathrm{id})$, a move by $\sqrt{2h}g$ decreases the second derivative. 45

7.1 Mixing rate of CRHMC and the competitors. Mixing rate of CRHMC was sub-linear in dimension and the nnz of a preprocessed matrix A in a model, whereas the others needed quadratically many steps to converge to uniform distribution. In particular for our dataset, CRHMC mixed up to 6 orders of magnitude earlier than the others. Note that mixing rate of CHAR was very close to quadratic growth when using the full-dimensional scale (the first column in Table F.1). 173

7.2 Sampling time of CRHMC and the competitors. The sampling time per effective sample of CRHMC was sub-quadratic in dimension and the nnz of a preprocessed matrix A in a model, while the others indicates at least a cubic dependency on dimension. In particular for our dataset, CRHMC was able to obtain a statistically independent sample up to 4 orders of magnitude faster than the others. This benefit of speed-up was actually straightforward from the figure, since CHRR could not obtain enough samples from instances with more than 5000 variables until it ran out of time. 173

7.3 Mixing rate and sampling time on structured polytopes including hybercubes, simplices, and Birkhoff polytopes. CRHMC is scalable up to 0.5 million dimension on hypercubes and simplices and up to 0.1 million dimension on Birkhoff polytopes. We note that on the 0.5 million dimensional Birkhoff polytope the ESS is only 16, which is not reliable compared to the ESS on the other instances. 176

7.4 We plot the empirical cumulative distribution function of the radial distribution to the power of (1/dim) with 1000 ESS obtained by running CRHMC on *ATCC-49176* (952×1069, left) and *Aci-PHEA* (1319×1561, right), and in the plot x-axis is the scaling factor. We can observe the CDFs are very close to the CDFs of the uniform distribution over the polytopes defined by two instances. 176

F.1 Proof outline for the mixing rates of CRHMC 304

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Yin Tat Lee, who taught me the basics of research, introduced me to the problem of sampling, and supported my exploration of various research directions throughout my PhD journey. His expertise and dedication have been instrumental in shaping my research and academic growth.

I would like to thank other committee members, Shayan Oveis Gharan, Simon S. Du, and Marina Meila, for their valuable feedback and thoughtful suggestions since my qualifying and general exams. Their diverse perspectives have greatly enriched my work.

I am grateful to the theory group, the Paul G. Allen School of Computer Science & Engineering and the University of Washington for offering the PhD program and supporting their students in every aspect of the program.

Next, I am profoundly grateful to my other academic mentors. Sébastien Bubeck and Suriya Gunasekar mentored me during my internship and part-time research at Microsoft. They taught me a lot about research, especially how to ask the right questions and approach them, and offered me resources to explore different research questions. Kevin Tian and I started collaborating when he visited the University of Washington in 2019, and we have worked on many projects together since then, leading to Chapters 3-6 of this thesis. I learned a lot about research, writing papers, and giving talks from Kevin. Santosh Vempala introduced me to the history of sampling and offered me valuable research advice, which led to Chapter 7 of this thesis, a project that spanned more than four years.

I would like to thank all my collaborators and academic friends during my PhD: Yifang Chen, Sinho Chewi, Sally Dong, Ronen Eldan, Murat A. Erdogdu, Mars Liyao Gao, Sivakanth Gopi, Haotian Jiang, Yunbum Kook, Ananya Kumar, Mufan Li, Yuanzhi Li, Daogao Liu, Kuikui Liu, Swati Padmanabhan, Hao Peng, Victor Reis, Tianxiao Shen, Yuhao Wan, Zhihan Xiong, Han Zhang, Matthew Zhang, Yi Zhang, Mingyuan Zhong, Runlong Zhou and Xiangfeng Zhu.

Finally, I would like to express my deepest gratitude to my parents and grandparents

for their unconditional love, patience, and belief in me, and to my partner, Qiwen, for being a source of support and joy. Without their support, none of this would have been possible.

Chapter 1

INTRODUCTION

We study the problem of efficiently sampling from a target distribution. This problem is central in statistics, machine learning, and theoretical computer science, with applications such as Bayesian estimation [ADFDJ03], volume computation [Vem10] and bandit optimization [RVR$^+$18]. Since the seminal result [DFK91a], which first presented a polynomial-time algorithm based on the Markov Chain Monte Carlo (MCMC) method (for an equivalent problem), decades of research has been focused on using MCMC methods to sample. Mathematically, we are given a target distribution with a density function proportional to $\exp(-f(x))$, where $f : \mathbb{R}^d \to \mathbb{R}$ is a smooth function. We have access to a gradient oracle and/or a value oracle to the function f. Our goal is to design samplers that can output a sample x such that the distribution of x, π_x is close to the target distribution. Common metrics used to measure the distance between the output distributions and the target distribution include Wasserstein distance, total variation distance, KL divergence, etc. Notice that we don't assume access to the density function itself, nor do we need to know the normalization constant of the distribution. For this problem, we are interested in how we can design samplers with optimal query complexity.

The general approach of MCMC-based sampling algorithms often includes two steps. The first step involves choosing a Markov process with a stationary distribution equal to or close to the target distribution. The second step is discretizing the process and simulating it until the distribution of the points generated is sufficiently close to the target distribution. One commonly used Markov process is the Langevin diffusion (LD) [RT96a, GM91, DMM19], which evolves according to the SDE,

$$\mathrm{d}x(t) \quad = \quad -\nabla f(x(t))\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \tag{1.1}$$

where B_t is the Brownian motion. (1.1) converges to the stationary distribution $\pi^* \sim \exp(-f(x))$. However, to simulate this continuous process, we need to discretize it, which can introduce discretization errors that affect the accuracy. The key to designing efficient samplers lies in both choosing the appropriate Markov process and accurate implementation of the process.

Sampling has a close relationship with optimization, one of the most studied problems in recent years. The problem of optimizing a function f can be solved by a reduction to sampling from a distribution with density proportional to $\exp(-\beta f)$ for a large constant β. Indeed, many sampling algorithms are inspired by optimization algorithms. However, sampling is more powerful and strictly harder than optimization. Intuitively, sampling from a distribution needs to not only identify the region with the highest density but also explore the landscape of other regions. This exploration ability makes sampling an especially powerful tool, but at the same time introduces substantial challenges, both theoretically and practically.

One challenge in designing samplers is that provable correctness and mixing time guarantees are central to sampling algorithms. Unlike in optimization problems where the value of the target function can be evaluated easily, in sampling problems, evaluating whether output samples follow a target distribution is hard. As a result, rigorous mathematical analysis is necessary to ensure the correctness and mixing of sampling methods.

In large-scale and high-dimensional scenarios, many sampling algorithms suffer from slow mixing and high computation costs. Only one trajectory is sufficient to obtain the optimizer in optimization problems, but sampling problem usually needs a large number of samples in practical applications. Therefore, efficiency is especially important in designing sampling methods.

To tackle these challenges, we focus on the "hard" cases, where the distributions are ill-conditioned and high-dimensional. We are interested in the theoretical guarantees of the samplers as well as their practical efficiency. Our theoretical study aims not only to ensure the correctness of the algorithms but also to gain a fundamental understanding of the efficiency of the samplers.

Part I: Efficient Discretization for Samplers

MCMC sampling algorithms involve running a Markov chain to convergence, making the query complexity contingent on the rate of convergence. A significant factor in the sampler's inefficiency is the discretization error. For instance, to simulate the Langevin Diffusion (1.1) for a small time interval h starting from x_0, a straightforward method is to query the gradient $\nabla f(x_0)$ and estimate x_h using the Euler method,

$$\tilde{x}_h = x_0 - h\nabla f(x_0) + \zeta_t,$$

where ζ_t is drawn from $\mathcal{N}(0,t)$. Therefore, to achieve the desired accuracy, it is necessary to select a sufficiently small step size h. However, a smaller h necessitates a higher number of steps for the Markov chain to converge, leading to slower convergence rates. As a result, an improved discretization method can substantially improve the efficiency of the sampling algorithm.

In Chapter 2, we propose a new framework to discretize stochastic differential equations. We apply this framework to discretize and simulate underdamped Langevin diffusion (ULD), which can be viewed as a version of the Langevin diffusion with momentum. The framework can be used to solve not only the log-concave sampling problem, but any problem that involves simulating (stochastic) differential equations. Our algorithm achieve $\epsilon \cdot D$ error (in 2-Wasserstein distance) in $\tilde{O}\left(\kappa^{7/6}/\epsilon^{1/3} + \kappa/\epsilon^{2/3}\right)$ steps, where $D \stackrel{\text{def}}{=} \sqrt{\frac{d}{m}}$ is the effective diameter of the problem and $\kappa \stackrel{\text{def}}{=} \frac{L}{m}$ is the condition number. Our algorithm performs significantly faster than the previously best known algorithm for solving this problem, which requires $\tilde{O}\left(\kappa^{1.5}/\epsilon\right)$ steps [CV19, DRD18]. Moreover, our algorithm can be easily parallelized to require only $O(\kappa \log \frac{1}{\epsilon})$ parallel steps.

Part II: Metropolized Sampling Algorithms

In the preceding part, we improve the efficiency of the samplers by adopting a better discretization method. To further reduce the error resulting from discretization, one can apply a Metropolis-Hastings(MH) filter in each iteration to adjust the stationary distribution of the Markov. This approach enables the creation of samplers that exhibit high accuracy with a logarithmic dependence on the inverse of ϵ, where ϵ represents the distance to the stationary distribution, compared to a mere polynomial dependence in the absence of the MH filter. Two of the most common sampling algorithms with the MH filter are Metropolized Hamiltonian Monte Carlo (HMC) and Metropolis-adjusted Langevin Dynamics (MALA). However, a comprehensive understanding of their potential and fundamental limitations remains lacking. In Chapter 3 and Chapter 4, we study the upper bound and the lower bound on the query complexity of Metropolized HMC and MALA. In particular, we show a matching upper and lower bound for one-step Metropolized HMC and MALA, addressing a longstanding open question regarding the query complexity of these widely utilized sampling algorithms.

In Chapter 3, we show that the gradient norm $\|\nabla f(x)\|$ for $x \sim \exp(-f(x))$, where f is strongly convex and smooth, concentrates tightly around its mean. This removes a

barrier in the prior state-of-the-art analysis for the well-studied Metropolized HMC and MALA for sampling from a strongly logconcave distribution [DCWY19]. We correspondingly demonstrate that Metropolized HMC mixes in $\tilde{O}(\kappa d)$ iterations[1], improving upon the $\tilde{O}(\kappa^{1.5}\sqrt{d} + \kappa d)$ runtime of [DCWY19, CDWY20] by a factor $(\kappa/d)^{1/2}$ when the condition number κ is large. Our mixing time analysis introduces several techniques which to our knowledge have not appeared in the literature and may be of independent interest, including restrictions to a nonconvex set with good conductance behavior, and a new reduction technique for boosting a constant-accuracy total variation guarantee under weak warmness assumptions. This is the first high-accuracy mixing time result for logconcave distributions using only first-order function information which achieves linear dependence on κ.

In Chapter 4, we give lower bounds on the performance of two of the most popular sampling methods in practice, MALA and multi-step HMC with a leapfrog integrator, when applied to well-conditioned distributions. Our main result is a nearly-tight lower bound of $\widetilde{\Omega}(\kappa d)$ on the mixing time of MALA from an exponentially warm start, matching a line of algorithmic results [DCWY19, CDWY20, LST20] up to logarithmic factors and answering an open question of [CLA+21]. We also show that a polynomial dependence on dimension is necessary for the relaxation time of HMC under any number of leapfrog steps, and bound the gains achievable by changing the step count. Our HMC analysis draws upon a novel connection between leapfrog integration and Chebyshev polynomials, which may be of independent interest.

Part III: Proximal Sampler

The complexity of sampling algorithms can be influenced by the condition number of the target distributions, rendering the sampling process from ill-conditioned distributions inefficient. This inefficiency is particularly pronounced in the case of structured densities, where the necessity for more sophisticated sampling algorithms, which have a marked reliance on the condition number, becomes apparent. Structured distributions possess unique features such as separability, enabling specialized samplers to outperform general-purpose ones in terms of efficiency. These distributions are of significant practical importance, and their optimization counterparts have been extensively explored in the literature. In Chapter 5, we design samplers for structured distributions and introduce a proximal reduction

[1]We use $\tilde{O}$ to hide logarithmic factors in problem parameters.

framework aimed at enhancing the condition number dependence of these samplers. Chapter 6 expands upon this by adapting the proximal framework for the non-Euclidean case. The development of efficient sampling algorithms catering to non-Euclidean geometries has been a challenging endeavor, as discretization techniques that succeed in the Euclidean setting do not readily carry over to more general settings. The proximal frameworks presented are applicable not only to samplers for structured distributions but also to general samplers.

In Chapter 5, we give algorithms for sampling several structured logconcave families to high accuracy.[2] We further develop a reduction framework, inspired by *proximal point methods* in convex optimization, which bootstraps samplers for regularized densities to generically improve dependences on problem conditioning κ from polynomial to linear. A key ingredient in our framework is the notion of a "restricted Gaussian oracle" (RGO) for $g : \mathbb{R}^d \to \mathbb{R}$, which is a sampler for distributions whose negative log-likelihood sums a quadratic (in a multiple of the identity) and g. By combining our reduction framework with our new samplers, we obtain the following bounds for sampling structured distributions to total variation distance ϵ.

- For composite densities $\exp(-f(x) - g(x))$, where f has condition number κ and convex (but possibly non-smooth) g admits an RGO, we obtain a mixing time of $O(\kappa d \log^3 \frac{\kappa d}{\epsilon})$, matching the state-of-the-art non-composite bound [LST20], shown in Chapter 3. No composite samplers with better mixing than general-purpose logconcave samplers were previously known.

- For logconcave finite sums $\exp(-F(x))$, where $F(x) = \frac{1}{n}\sum_{i\in[n]} f_i(x)$ has condition number κ, we give a sampler querying $\widetilde{O}(n + \kappa \max(d, \sqrt{nd}))$ gradient oracles[3] to $\{f_i\}_{i\in[n]}$. No high-accuracy samplers with nontrivial gradient query complexity were previously known.

- For densities with condition number κ, we give an algorithm obtaining mixing time $O(\kappa d \log^2 \frac{\kappa d}{\epsilon})$, improving [LST20] by a logarithmic factor with a significantly simpler

[2] We say a sampler is "high-accuracy" if its mixing time has polylogarithmic dependence on the target accuracy ϵ.

[3] For convenience of exposition, the $\widetilde{O}$ notation hides logarithmic factors in the dimension d, problem conditioning κ, desired accuracy ϵ, and summand count n (when applicable). A first-order (gradient) oracle for $f : \mathbb{R}^d \to \mathbb{R}$ returns $(f(x), \nabla f(x))$ on input x, and a zeroth-order (value) oracle only returns $f(x)$.

analysis. We also show a zeroth-order algorithm attains the same query complexity.

In Chapter 6, we develop a non-Euclidean analog of the recent proximal sampler of [LST21b], which naturally induces regularization by an object known as the log-Laplace transform (LLT) of a density. We prove new mathematical properties (with an algorithmic flavor) of the LLT, such as strong convexity-smoothness duality and an isoperimetric inequality, which are used to prove a mixing time on our proximal sampler matching [LST21b] under a warm start. We find our investigation of the LLT to be a promising proof-of-concept of its utility as a tool for designing samplers and outline directions for future exploration.

Part IV: Sampling from a Constrained Set

A common yet challenging type of distribution to sample from is the constrained distribution. This is defined by a set of equality or inequality constraints, such as polytopes. More precisely, we consider the problem of sampling from the distribution

$$e^{-f(x)} \text{ subject to } Ax = b, x \in K \tag{1.2}$$

for some convex set K. To estimate characteristics like the volume of complex, high-dimensional objects, sampling from the object is often the sole method available to researchers. However, despite their importance, current algorithms for sampling from constrained distributions, especially those that are high-dimensional or ill-conditioned, are highly inefficient. This makes the process extremely costly or even unfeasible.

In Part IV, we focus on designing samplers for constrained distributions. There are two main challenges in using MCMC method to sample from constrained distributions. Firstly, maintaining the random walk strictly within the boundaries of the constrained set can be difficult. Secondly, if the constrained set is ill-conditioned, the Markov chain might converge very slowly in certain directions. We overcome these two challenges and demonstrate for the first time that ill-conditioned, non-smooth, constrained distributions in very high dimensions, upwards of 100,000, can be sampled efficiently *in practice*. Our algorithm incorporates constraints into the Riemannian version of Hamiltonian Monte Carlo and maintains sparsity. This allows us to use the local geometry of the constrained set, and achieve a mixing rate independent of condition numbers.

On benchmark data sets from systems biology and linear programming, our algorithm outperforms existing packages by orders of magnitude. In particular, we achieve a 1,000-

fold speed-up for sampling from the largest published human metabolic network (RECON3D). Our package has been incorporated into the COBRA toolbox [HAP+19].

Part I

EFFICIENT DISCRETIZATION FOR SAMPLERS

Chapter 2

RANDOMIZED MIDPOINT METHOD

This chapter is based on [SL19], with Yin Tat Lee.

2.1 Introduction

In this chapter, we study the problem of sampling from a high-dimensional log-concave distribution. We call a distribution *log-concave* if its density is proportional to $e^{-f(x)}$ with a convex function f. The standard assumption is that f is m-strongly convex with an L-Lipschitz gradient (see Section 2.2.4). In this chapter, we present an algorithm with no dependence on d and a much smaller dependence on κ and ϵ than shown in previous research in terms of Wasserstein distance convergence. Moreover, our algorithm is the first algorithm with better than $1/\epsilon$ dependence that is not Metropolis-adjusted and does not make any extra assumption, such as high-order smoothness [MS17, MV18, CFM$^+$18, MMW$^+$19].

To explain our main result, we note that this problem has an effective diameter $D \stackrel{\text{def}}{=} \sqrt{\frac{d}{m}}$ because the distance between the minimizer x^* of f and a random point $y \sim e^{-f}$ satisfies $\mathbb{E}_{y\sim e^{-f}}\|x^* - y\|^2 \leq \frac{d}{m}$[DM16]. Therefore, a natural problem definition[1] is to find a random x that makes the Wasserstein distance small:

$$W_2(x, y) \leq \epsilon \cdot D. \tag{2.1}$$

This choice of distance is also common in previous papers [DM16, DM17, CCBJ17, MS17, LSV18, MV18, CFM$^+$18].

For $\epsilon = 1$, we can simply output the minimizer x^* of f as the "random" point. We first consider the question how quickly we can find a random point satisfying $\epsilon = \frac{1}{2}$. For convex optimization under the same assumption, it takes $\sqrt{\kappa}$ iterations via acceleration methods

[1] Previous papers addressing this problem defined ϵ as $W_2(x, e^{-f}) \leq \epsilon$. This definition is not scale invariant, i.e., the number of steps changes when we scale f. In comparison, our definition yields results that are invariant under: (1) the scaling of f, namely, replacing $f(x)$ by $\alpha f(x)$ for $\alpha > 0$, and (2) the tensor power of f, namely, replacing $f(x)$ by $g(x) \stackrel{\text{def}}{=} \sum_i f(x_i)$. Our new definition of ϵ also clarifies definitions in previous research. Under the prior definition of ϵ, the algorithms [DM16, CCBJ17, CV19] take $\tilde{O}(\kappa^2(\sqrt{\frac{d}{m}}/\epsilon)^2)$, $\tilde{O}(\kappa^2\sqrt{\frac{d}{m}}/\epsilon)$, and $\tilde{O}(\kappa^{1.5}\sqrt{\frac{d}{m}}/\epsilon)$ steps, respectively. Our new definition shows that these different dependences on d and m all relate to their dependence on ϵ.

or d iterations via cutting plane methods, and these results are tight. For sampling, the current fastest algorithms take either $\tilde{O}(\kappa^{1.5})$ steps [CV19, DRD18] without Metropolis-Hasting filter or $\tilde{O}(d^4)$ steps [LV06a] when the distribution is not well-conditioned. Although there is no rigorous lower bound for this problem, it is believed that $\min(\kappa, d^2)$ is the natural barrier.[2] We present an algorithm that takes only $\tilde{O}(\kappa^{7/6})$ steps, much closer to the natural barrier of κ for the high-dimensional regime.

For general $0 < \epsilon < 1$, our algorithm takes $\tilde{O}(\kappa^{7/6}/\epsilon^{1/3} + \kappa/\epsilon^{2/3})$ steps, which is almost linear in κ and sub-linear in ϵ . It has significantly better dependence on both κ and ϵ than previous algorithms (See the detailed comparison in Table 2.1.) Moreover, if we query gradient ∇f at multiple points in parallel in each step, we can improve the number to $O(\kappa \log \frac{1}{\epsilon})$ steps.

2.1.1 Contributions

We propose a new framework to discretize stochastic differential equations (SDEs), which is a crucial step of log-sampling algorithms. Since our techniques can also be applied to ordinary differential equations (ODEs), we focus on the following ODE here:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = F(x(t)).$$

There are two main frameworks to discretize a differential equation. One is the Taylor expansion, which approximates $x(t)$ by $x(0) + x'(0)t + x''(0)\frac{t^2}{2} + \cdots$. We use the second framework, called the *collocation method*. This method uses the fact that the differential equation is equivalent to the integral equation $x = \mathcal{T}(x)$, where $\mathcal{T}$ maps continuous functions to continuous functions:

$$\mathcal{T}(x)(t) = x(0) + \int_0^t F(x(s))\,\mathrm{d}s \text{ for all } t \geq 0.$$

Since x is a fixed point of $\mathcal{T}$, we can approximate x by computing $\mathcal{T}(\mathcal{T}(\cdots(\mathcal{T}(x_0))\cdots))$ for some approximate initial function x_0. Algorithmically, two key questions are how to: (1) show when and how quickly $\mathcal{T}$ iterations converge, and (2) compute the integration. The convergence rate of $\mathcal{T}$ was shown by the Picard–Lindelöf Theorem in the 1890s [Lin94,

[2]The corresponding optimization problem takes at least $\min(\sqrt{\kappa}, d)$ steps [NY83]. If we represent each point the optimization algorithm visited by a vertex and each step the algorithm takes by an edge, then the existing lower bound in fact shows that this graph has a diameter of at least $\min(\sqrt{\kappa}, d)$. Since a random walk on a graph of diameter D takes D^2 to mix, a random walk on the graph takes at least $\min(\sqrt{\kappa}, d)^2$ steps.

Pic98] and was key to achieving $O(\kappa^{1.75})$ and $O\left(\kappa^{1.5}\right)$ in the previous papers [LSV18, CV19]. To approximate the integration, one standard approach is to approximate

$$\int_0^t F(x(s))\,\mathrm{d}s \sim \sum_i w_i F(x(s_i))$$

for some carefully chosen w_i and s_i. The key drawback of this approach is its introduction of a deterministic error, which accumulates linearly to the number of steps. Since we expect to take at least κ-many iterations, the approximation error must be κ times smaller than the target accuracy.

In this paper, we improve upon the collocation method for sampling by developing a new algorithm, called the *randomized midpoint method*, that yields three distinct benefits:

1. We generalize fixed point iteration to stochastic differential equations and hence avoid the cost of reducing SDEs to ODEs, as was done in [LSV18].

2. We greatly reduce the error accumulation by simply approximating $\int_0^t F(x(s))ds$ by $t \cdot F(x(s))$ where s is randomly chosen from 0 to t uniformly.

3. We show that two iterations of $\mathcal{T}$ suffice to achieve the best theoretical guarantee.

Although we discuss only strongly convex functions with a Lipschitz gradient, we believe our framework can be applied to other classes of functions as well. By designing suitable unbiased estimators of integrals, researchers can easily use our approach to obtain faster algorithms for solving SDEs that are unrelated to sampling problems.

2.1.2 Organization

Section 2.2 provides background information on solving the log-concave sampling problem, while Section 2.2.3 introduces our notations and assumptions about the function f. We introduce our algorithm in Section 2.3, where we present the main result of our paper. We show our proofs in appendices: Appendix A.1–how we simulate the Brownian motion; Appendix A.2–important properties of ULD and the Brownian motion; Appendix A.3– bounds for the discretization error of our algorithm; Appendix A.4–a bound on the average value of $\|\nabla f(x_n)\|$ and $\|v_n\|$ in our algorithm, which is useful for bounding the discretization error; Appendix A.5–proofs for our main result; Appendix A.6–additional proofs on how to parallelize our algorithm.

2.2 Preliminary

Many different algorithms have been proposed to solve the log-concave sampling problem. The general approach uses a MCMC-based algorithm that often includes two steps. The first step involves the choice of a Markov process with a stationary distribution equal or close to the target distribution. The second step is discretizing the process and simulating it until the distribution of the points generated is sufficiently close to the target distribution.

2.2.1 Choosing the Markov Process

One commonly used Markov process is the Langevin diffusion (LD), which evolves according to the SDE

$$\mathrm{d}x(t) \quad = \quad -\nabla f(x(t))\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \tag{2.2}$$

where B_t is the standard Brownian motion. Under the assumption that f is L-smooth and m-strongly convex (see Section 2.2.4) with $\kappa = \frac{L}{m}$ as the condition number, [DM16, Dal17b, CB17] show that algorithms based on LD can achieve less than ϵ error in $\tilde{O}\left(\frac{\kappa^2}{\epsilon^2}\right)$ steps. Other related works include LD with stochastic gradient [DK19, ZLC17, RRT17, CFM$^+$18] and LD in the non-convex setting [RRT17, CCAY$^+$18].

One important breakthrough introduced the Hamiltonian Monte Carlo (HMC), originally proposed in [Kra40]. In this process, SDE (2.2) is approximated by a piece-wise curve, where each piece is governed by an ODE called the Hamiltonian dynamics. The Hamiltonian dynamics maintains a velocity v in addition to a position x and conserves the value of the Hamiltonian $H(x,v) = f(x) + \frac{1}{2}\left\|v\right\|^2$. HMC has been widely studied in [Nea11, MCF15, MS17, MV18, LSV18, CV19, LV18]. The works [CV19, DRD18] show that algorithms based on HMC can achieve less than ϵ error in $\tilde{O}\left(\frac{\kappa^{1.5}}{\epsilon}\right)$ steps.

The underdamped Langevin diffusion (ULD) can be viewed as a version of HMC that replaces multiple ODEs with one SDE; it has been studied in [CCBJ17, EGZ17, DRD18]. ULD follows the SDE:

$$\mathrm{d}v(t) = -2v(t)\,\mathrm{d}t - u\nabla f(x(t))\,\mathrm{d}t + 2\sqrt{u}\,\mathrm{d}B_t, \qquad \mathrm{d}x(t) = v(t)\,\mathrm{d}t, \tag{2.3}$$

where $u = \frac{1}{L}$. [CCBJ17] shows that even a basic discretization of ULD has a fast convergence rate that can achieve less than ϵ error in $\tilde{O}\left(\frac{\kappa^2}{\epsilon}\right)$ steps. Recently, it was shown that ULD can be viewed as an accelerated gradient descent for sampling [MCC$^+$21]. This suggests that ULD might be one of the right dynamics for sampling in the same way as

the accelerated gradient descent method is appropriate for convex optimization. For this reason, we focus on how to discretize ULD. We note that our framework can be applied to both LD and HMC to improve on previous results for these dynamics as well.

2.2.2 Discretizing the Process

To simulate the random process mentioned, previous works usually apply the Euler method [CCBJ17, DM16] or the Leapfrog method [MS17, MV18] to discretize the SDEs or the ODEs. In Section 2.3.2, we introduce a 2-step fixed point iteration method to solve general differential equations. We apply this method to ULD and significantly reduce the discretization error compared to existing methods. In particular, ULD can achieve less than ϵ error in $\tilde{O}\left(\frac{\kappa^{7/6}}{\epsilon^{1/3}} + \frac{\kappa}{\epsilon^{2/3}}\right)$ steps. Table 2.1 summarizes the number of steps needed by previous algorithms versus our algorithm. Moreover, with slightly more effort, our algorithm can be parallelized so that it needs only $O\left(\kappa \log \frac{1}{\epsilon}\right)$ parallel steps.

On top of the discretization method, one can use a Metropolis-Hastings accept-reject step to ensure that the post-discretization random process results in a stationary distribution equal to the target distribution Since this chpater focuses on achieving a dimension independent result, we do not discuss how to combine our process with a Metropolis-Hastings step in this chapter. We will have a more depth discussion on Metropolis-adjusted algorithms in Chapter 3 and Chapter 4.

Finally, we note that all results–including ours–can be improved if we assume that f has bounded higher-order derivatives. To ensure a fair comparison in Table 2.1, we only include results that only assume f is strongly convex and has a Lipschitz gradient.

2.2.3 Notations and Definitions

For any function f, we use $\tilde{O}(f)$ to denote the class $O(f) \cdot \log^{O(1)}(f)$. For vector $v \in \mathbb{R}^d$, we use $\|v\|$ to denote the Euclidean norm of v.

2.2.4 Assumptions on f

We assume that the function f is a twice continuously differentiable function from $\mathbb{R}^d$ to $\mathbb{R}$ that has an L-Lipschitz continuous gradient and is m-strongly convex. That is, there exist positive constants L and m such that for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \text{ and } f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2} \|x - y\|^2.$$

It is easy to show that these inequalities are equivalent to $mI_d \preceq \nabla^2 f(x) \preceq LI_d$, where I_d is the identity matrix of dimension d. Let $\kappa = \frac{L}{m}$ be the condition number. We assume that we have access to an oracle that, given a point $x \in \mathbb{R}^d$, can return the gradient of f at point x, $\nabla f(x)$.

2.2.5 Wasserstein Distance

The pth Wasserstein distance between two probability measures μ and ν is defined as

$$W_p(\mu,\nu) \quad = \quad \left(\inf_{(X,Y)\in\mathcal{C}(\mu,\nu)} \mathbb{E}\left[\|X-Y\|^p\right]\right)^{1/p},$$

where $\mathcal{C}(\mu,\nu)$ is the set of all couplings of μ and ν. For any $0 < \epsilon < 1$, we study the number of steps needed so that the W_2 distance between the distribution of the point our algorithms generate and the target distribution is smaller than $\epsilon \cdot D$.

2.3 Randomized Midpoint Method

2.3.1 Underdamped Langevin Diffusion (ULD)

ULD is a random process that evolves according to (2.3). We study (2.3) with $u = \frac{1}{L}$. Under mild conditions, it can be shown that the stationary distribution of (2.3) is proportional to $\exp\left(-f(x) + L\|v\|^2/2\right)$. Then, the marginal distribution of x is proportional to $\exp(-f(x))$. It can also be shown that the solution to (2.3) has a contraction property [CCBJ17, EGZ17], shown in the following lemma.

Lemma 1 (Theorem 5 of [CCBJ17])**.** *Let (x_0, v_0) and (y_0, w_0) be two arbitrary points in $\mathbb{R}^d \times \mathbb{R}^d$. Let (x_t, v_t) and (y_t, w_t) be the exact solutions of the underdamped Langevin diffusion after time t. If (x_t, v_t) and (y_t, w_t) are coupled through a shared Brownian motion, then,*

$$\mathbb{E}\left[\|x_t - y_t\|^2 + \|(x_t + v_t) - (y_t + w_t)\|^2\right] \le e^{-\frac{t}{\kappa}}\mathbb{E}\left[\|x_0 - y_0\|^2 + \|(x_0 + v_0) - (y_0 + w_0)\|^2\right].$$

This contraction bound can be very useful for showing the convergence of the continuous process (2.3). In our algorithm, we discretize the continuous process to implement it; therefore we need to use this contraction bound together with a discretization error bound to show the guarantee of our algorithm. In Section 2.3.2, we show how we discretize (2.3).

Algorithm 1 Randomized Midpoint Method for ULD

Procedure RandomMidpoint(x_0, v_0, N, h)

For $n = 0, ..., N-1$

Randomly sample α uniformly from $[0,1]$.

Generate Gaussian random variable $\left(W_1^{(n)}, W_2^{(n)}, W_3^{(n)}\right) \in \mathbb{R}^{3d}$ as in Appendix A.1

$x_{n+\frac{1}{2}} = x_n + \frac{1}{2}\left(1 - e^{-2\alpha h}\right) v_n - \frac{1}{2} u \left(\alpha h - \frac{1}{2}(1 - e^{-2\alpha h})\right) \nabla f(x_n) + \sqrt{u} W_1^{(n)}$.

$x_{n+1} = x_n + \frac{1}{2}\left(1 - e^{-2h}\right) v_n - \frac{1}{2} uh \left(1 - e^{-2(h-\alpha h)}\right) \nabla f(x_{n+\frac{1}{2}}) + \sqrt{u} W_2^{(n)}$.

$v_{n+1} = v_n e^{-2h} - uhe^{-2(h-\alpha h)} \nabla f(x_{n+\frac{1}{2}}) + 2\sqrt{u} W_3^{(n)}$.

end for

end procedure

2.3.2 Randomized Midpoint Method

Our step size for each iteration is h. In iteration n of our algorithm, to simulate (2.3), we need to approximate the solution to SDE (2.3) at time h, $(x_n^*(h), v_n^*(h))$, with initial value, (x_n, v_n). The simplest way to do so is to use the Euler method:

$$v_n(h) = (1 - 2h)v_n - uh\nabla f(x_n) + 2\sqrt{uh}\zeta, \quad x_n(h) = x_n + hv_n,$$

where $\zeta \in \mathbb{R}^d$ is drawn from the standard normal distribution. This discretization was considered in [DM17, Dal17b] due to its simplicity.

As discussed in Section 2.1.1, we improve the accuracy by studying the integral formulation of (2.3):

$$x_n^*(t) = x_n + \frac{1 - e^{-2t}}{2} v_n - \frac{u}{2} \int_0^t \left(1 - e^{-2(t-s)}\right) \nabla f(x_n^*(s))\, \mathrm{d}s + \sqrt{u} \int_0^t \left(1 - e^{-2(t-s)}\right) \mathrm{d}B_s,$$

$$v_n^*(t) = v_n e^{-2t} - u \left(\int_0^t e^{-2(t-s)} \nabla f(x_n^*(s))\, \mathrm{d}s\right) + 2\sqrt{u} \int_0^t e^{-2(t-s)}\, \mathrm{d}B_s. \tag{2.4}$$

[CCBJ17] considered the same integral formulation and used $\nabla f(x_n)$ to approximate $\nabla f(x_n^*(t))$ for $t \in [0, h]$ to get the following algorithm:

$$\hat{x}_n(h) = x_n + \frac{1 - e^{-2h}}{2} v_n - \frac{u}{2} \int_0^h \left(1 - e^{-2(h-s)}\right) \nabla f(x_n)\, \mathrm{d}s + \sqrt{u} \int_0^h \left(1 - e^{-2(h-s)}\right) \mathrm{d}B_s,$$

$$\hat{v}_n(h) = v_n e^{-2h} - u \left(\int_0^h e^{-2(h-s)} \nabla f(x_n)\, \mathrm{d}s\right) + 2\sqrt{u} \int_0^h e^{-2(h-s)}\, \mathrm{d}B_s.$$

However, this approximation method can still generate a relatively large error. We propose a new method, the randomized midpoint method, to solve (2.4), which yields a more accurate approximation and significantly reduces the total runtime of the algorithm.

We first need to identify an accurate estimator of the integral $\int_0^h \left(1 - e^{-2(h-s)}\right) \nabla f(x_n^*(s))\,\mathrm{d}s$. To do so, we sample a random number α uniformly from $[0,1]$ so that αh gives a random point from $[0,h]$. Then, $h\left(1 - e^{-2(h-\alpha h)}\right) \nabla f(x_n^*(\alpha h))$ is an accurate estimator of the integral
$\int_0^h \left(1 - e^{-2(h-s)}\right) \nabla f(x_n^*(s))\,\mathrm{d}s$. We can further show that this estimator is unbiased.

For brevity, we use $x_{n+\frac{1}{2}}$ to denote our approximation of $x_n^*(\alpha h)$. To approximate $x_n^*(\alpha h)$, we use equation (2.4) again:

$$x_{n+\frac{1}{2}} = x_n + \frac{1 - e^{-2\alpha h}}{2} v_n - \frac{u}{2} \int_0^{\alpha h} \left(1 - e^{-2(\alpha h - s)}\right) \nabla f(x_n)\mathrm{d}s + \sqrt{u} \int_0^{\alpha h} \left(1 - e^{-2(\alpha h - s)}\right) \mathrm{d}B_s.$$

Then, $(x_n^*(h), v_n^*(h))$ can be approximated as

$$x_{n+1} = x_n + \frac{1 - e^{-2h}}{2} v_n - \frac{u}{2} h \left(1 - e^{-2(h-\alpha h)}\right) \nabla f(x_{n+\frac{1}{2}}) + \sqrt{u} \int_0^h \left(1 - e^{-2(h-s)}\right) \mathrm{d}B_s,$$

$$v_{n+1} = v_n e^{-2h} - uhe^{-2(h-\alpha h)} \nabla f(x_{n+\frac{1}{2}}) + 2\sqrt{u} \int_0^h e^{-2(h-s)}\,\mathrm{d}B_s.$$

Note that we can view (2.4) as the fixed point of the operator $\mathcal{T}$, $x_n^* = \mathcal{T}(x_n^*)$, where for all t,

$$\mathcal{T}(x)(t) = x_n + \frac{1 - e^{-2t}}{2} v_n - \frac{u}{2} \int_0^t \left(1 - e^{-2(t-s)}\right) \nabla f(x(s))\,\mathrm{d}s + \sqrt{u} \int_0^t \left(1 - e^{-2(t-s)}\right) \mathrm{d}B_s. \tag{2.5}$$

Then, our randomized algorithm is essentially approximating $\mathcal{T}(\mathcal{T}(x_n))$. Under the assumption f is twice differentiable, we show that two iterations suffice to achieve the best theoretical guarantee, but we suspect more iterations might be useful if f has higher order derivatives. As emphasized in Section 2.1.1, the way we obtain our algorithm forms a general framework that can be applied to other SDEs.

In Lemma 5, we show that the stochastic terms $W_1 = \int_0^{\alpha h} \left(1 - e^{-2(\alpha h - s)}\right) \mathrm{d}B_s$,
$W_2 = \int_0^h \left(1 - e^{-2(h-s)}\right) \mathrm{d}B_s$, and $W_3 = \int_0^h e^{-2(h-s)}\,\mathrm{d}B_s$ conditional on the choice of α follow a multi-dimensional Gaussian distribution and therefore can be easily sampled. The steps mentioned above are summarized in Algorithm 1. Using this randomized midpoint method, we can solve (2.4) much more accurately than previous works. We show that the discretization error satisfies:

Lemma 2. *For each iteration n of Algorithm 1, let $\mathbb{E}_\alpha$ be the expectation taken over the random choice of α in iteration n. Let $\mathbb{E}$ be the expectation taken over other randomness in iteration n. Let $(x_n^*(t), v_n^*(t))_{t\in[0,h]}$ be the solution of the exact underdamped Langevin*

diffusion starting from (x_n, v_n) *coupled through a shared Brownian motion with* $x_{n+\frac{1}{2}}$, v_n *and* x_{n+1}. *Assume that* $h \leq \frac{1}{20}$ *and* $u = \frac{1}{L}$. *Then,* x_{n+1} *and* v_{n+1} *of Algorithm 1 satisfy*

$$\begin{aligned}
\mathbb{E}\left\|\mathbb{E}_\alpha x_{n+1} - x_n^*(h)\right\|^2 &\leq O\left(h^{10}\left\|v_n\right\|^2 + u^2h^{12}\left\|\nabla f(x_n)\right\|^2 + udh^{11}\right), \\
\mathbb{E}\left\|x_{n+1} - x_n^*(h)\right\|^2 &\leq O\left(h^{6}\left\|v_n\right\|^2 + u^2h^{4}\left\|\nabla f(x_n)\right\|^2 + udh^{7}\right), \\
\mathbb{E}\left\|\mathbb{E}_\alpha v_{n+1} - v_n^*(h)\right\|^2 &\leq O\left(h^{8}\left\|v_n\right\|^2 + u^2h^{10}\left\|\nabla f(x_n)\right\|^2 + udh^{9}\right), \\
\mathbb{E}\left\|v_{n+1} - v_n^*(h)\right\|^2 &\leq O\left(h^{4}\left\|v_n\right\|^2 + u^2h^{4}\left\|\nabla f(x_n)\right\|^2 + udh^{5}\right).
\end{aligned}$$

In Appendix A.4, we show that the average value of $\|v_n\|^2$ is of order $\tilde{O}\left(\frac{d}{L}\right)$; that of $\|\nabla f(x_n)\|^2$ is of order $\tilde{O}\left(Ld\right)$. Then, Lemma 2 shows that the bias of the discretization is of order $\tilde{O}\left(h^4\sqrt{\frac{d}{L}}\right)$ and the standard deviation is of order $\tilde{O}\left(h^2\sqrt{\frac{d}{L}}\right)$, which implies the error is larger when h is larger. However, by Lemma 1, in order for the algorithm to converge in a small number of steps, we need to avoid choosing an h that is too small. Therefore, it is important to choose the largest possible h that can still make the algorithm converge. By Lemma 1, it is sufficient to run our algorithm for $\tilde{O}\left(\frac{\kappa}{h}\right)$ iterations. Then, the bias will cumulate to $\tilde{O}\left(h^4\sqrt{\frac{d}{L}}\cdot\frac{\kappa}{h}\right) = \tilde{O}\left(h^3\sqrt{\frac{d\kappa}{m}}\right)$, and the standard deviation will cumulate to $\tilde{O}\left(h^2\sqrt{\frac{d}{L}}\cdot\sqrt{\frac{\kappa}{h}}\right) = \tilde{O}\left(h^{1.5}\sqrt{\frac{d}{m}}\right)$. Thus, in order to make the W_2 distance less than $\tilde{O}\left(\epsilon\sqrt{\frac{d}{m}}\right)$, we show in Theorem 3 that it is enough to choose h to be $\tilde{\Theta}\left(\min\left(\frac{\epsilon^{1/3}}{\kappa^{1/6}}, \epsilon^{2/3}\right)\right)$. This choice of h yields our main result, which is stated in Theorem 3. (See Appendix A.5 for the full proof.)

Theorem 3 (Main Result)**.** *Let* f *be a function such that* $0 \prec m \cdot I_d \preceq \nabla^2 f(x) \preceq L \cdot I_d$ *for all* $x \in \mathbb{R}^d$. *Let* Y *be a random point drawn from the density proportional to* e^{-f}. *Let the starting point* x_0 *be the point that minimizes* $f(x)$ *and* $v_0 = 0$. *For any* $0 < \epsilon < 1$, *if we set the step size of Algorithm* 1 *as* $h = C\min\left(\frac{\epsilon^{1/3}}{\kappa^{1/6}}\log^{-1/6}\left(\frac{1}{\epsilon}\right), \epsilon^{2/3}\log^{-1/3}\left(\frac{1}{\epsilon}\right)\right)$, *for some small constant* C *and run the algorithm for* $N = \frac{2\kappa}{h}\log\left(\frac{20}{\epsilon^2}\right) \leq \tilde{O}\left(\frac{\kappa^{7/6}}{\epsilon^{1/3}} + \frac{\kappa}{\epsilon^{2/3}}\right)$ *iterations, then Algorithm* 1 *after* N *iterations can generate a random point* X *such that* $W_2(X, Y) \leq \epsilon\sqrt{\frac{d}{m}}$. *Furthermore, each iteration of Algorithm* 1 *involves computing* ∇f *exactly twice.*

2.3.3 A More General Algorithm

Now we show how our algorithm can be parallelized. The algorithm studied in this section can be viewed as a more general version of Algorithm 1. Instead of choosing one random

Algorithm 2 Randomized Midpoint Method for ULD (Parallel)

Procedure RandomMidpoint_P(x_0, v_0, N, h, R)

For $n = 0, ..., N-1$

Randomly sample α_1, ..., α_R uniformly from $\left[0, \frac{1}{R}\right]$, $\left[\frac{1}{R}, \frac{2}{R}\right]$, ..., $\left[\frac{R-1}{R}, 1\right]$.

Generate Gaussian r.v. $\left(W_{1,1}^{(n)}, ..., W_{1,R}^{(n)}, W_2^{(n)}, W_3^{(n)}\right) \in \mathbb{R}^{(R+2)d}$ similar to Appendix A.1

$x_n^{(0,i)} = x_n$ for $i = 1, ..., R$.

For $k = 1, ..., K-1, i = 1, ..., R$

$x_n^{(k,i)} = x_n + \frac{1}{2}\left(1 - e^{-2\alpha_i h}\right) v_n$

$-\frac{1}{2}u \sum_{j=1}^{i} \left[\int_{(j-1)\delta}^{\min(j\delta, \alpha_i h)} \left(1 - e^{-2(\alpha_i h - s)}\right) \mathrm{d}s \cdot \nabla f(x_n^{(k-1,j)})\right] + \sqrt{u} W_{1,i}^{(n)}$

end for

$x_{n+1} = x_n + \frac{1}{2}\left(1 - e^{-2h}\right) v_n - \frac{1}{2}u \sum_{i=1}^{R} \delta \left(1 - e^{-2(h-\alpha_i h)}\right) \nabla f(x_n^{(K-1,i)}) + \sqrt{u} W_2^{(n)}$,

$v_{n+1} = v_n e^{-2h} - u \sum_{i=1}^{R} \delta e^{-2(h-\alpha_i h)} \nabla f(x_n^{(K-1,i)}) + 2\sqrt{u} W_3^{(n)}$.

end for

end procedure

point from $[0, h]$, we divide the time interval $[0, h]$ into R pieces, each of length $\delta = \frac{h}{R}$, and choose one random point from each piece. That is, we randomly choose α_1, α_2, ..., α_R uniformly from $\left[0, \frac{1}{R}\right]$, $\left[\frac{1}{R}, \frac{2}{R}\right]$, ..., $\left[\frac{R-1}{R}, 1\right]$. As in Algorithm 1, to approximate $(x_n^*(h), v_n^*(h))$, we use

$$\tilde{x} = x_n + \frac{1 - e^{-2h}}{2} v_n - \frac{u}{2} \sum_{i=1}^{R} \delta \left(1 - e^{-2(h-\alpha_i h)}\right) \nabla f(x_n^*(\alpha_i h)) + \sqrt{u} \int_0^h \left(1 - e^{-2(h-s)}\right) \mathrm{d}B_s,$$

$$\tilde{v} = v_n e^{-2h} - u \sum_{i=1}^{R} \delta e^{-2(h-\alpha_i h)} \nabla f(x^*(\alpha_i h)) + 2\sqrt{u} \int_0^h e^{-2(h-s)} \, \mathrm{d}B_s,$$

which gives an unbiased estimator of $(x_n^*(h), v_n^*(h))$. The next step is to approximate $x_n^*(\alpha_i h)$ for $i = 1, .., R$. We know that the solution x_n^* is the fixed point of the operator $\mathcal{T}$ defined in (2.5). To solve the fixed point of $\mathcal{T}$, we can use the fixed point iteration method, which applies the operator $\mathcal{T}$ multiple times on some initial point. By the Banach fixed point theorem, the resulting points can converge to the fixed point of $\mathcal{T}$. Instead of applying $\mathcal{T}$, which involves computing an integral, we apply the operator $\tilde{\mathcal{T}}$, which approximates $\mathcal{T}$, on $X = \left(x^{(1)}, ..., x^{(R)}\right)$,

$$\tilde{\mathcal{T}}(X)_i = x_n + \frac{1}{2}\left(1 - e^{-2\alpha_i h}\right) v_n - \frac{1}{2}u \sum_{j=1}^{i} \left[\int_{(j-1)\delta}^{\min(j\delta, \alpha_i h)} \left(1 - e^{-2(\alpha_i h - s)}\right) \mathrm{d}s \cdot \nabla f(x^{(j)})\right]$$

$$+\sqrt{u}\int_0^{\alpha_i h}\left(1-e^{-2(\alpha_i h-s)}\right)\mathrm{d}B_s.$$

We set the initial points to $x_n^{(0,j)}=x_n$ for $j=1,...,R$. Then, we apply $\tilde{\mathcal{T}}$ for K times and get $(x^{(K,1)},...,x^{(K,R)})=\tilde{\mathcal{T}}^{\circ K}(x^{(0,1)},...,x^{(0,R)})$. The preceding steps are summarized in Algorithm 2. It is easy to see Algorithm 1 is a special case of Algorithm 2 with $R=1$ and $K=2$.

This algorithm can be parallelized since we can compute $\tilde{\mathcal{T}}(x^{(k,1)},...,x^{(k,R)})_j$ for each j parallelly. It can be shown that it is sufficient to choose K to depend logarithmically on κ and ϵ. Similar to Algorithm 1, we can show that Algorithm 2 has the guarantee that the bias of the discretization is of order $\tilde{O}\left(\frac{h^4}{R}\sqrt{\frac{d}{L}}\right)$ and the standard deviation is of order $\tilde{O}\left(\frac{h^2}{R}\sqrt{\frac{d}{L}}\right)$ (Appendix A.6). Then, summing from $\tilde{O}\left(\frac{\kappa}{h}\right)$ iterations, the total bias would be $\tilde{O}\left(\frac{h^4}{R}\sqrt{\frac{d}{L}}\cdot\frac{\kappa}{h}\right)=\tilde{O}\left(\frac{h^3}{R}\sqrt{\frac{d\kappa}{m}}\right)$, and the total standard deviation would be $\tilde{O}\left(\frac{h^2}{R}\sqrt{\frac{d}{L}}\cdot\sqrt{\frac{\kappa}{h}}\right)=\tilde{O}\left(\frac{h^{1.5}}{R}\sqrt{\frac{d}{m}}\right)$. By choosing $R=\tilde{\Theta}\left(\frac{\sqrt{\kappa}}{\epsilon}\right)$, it is enough to choose h to be a constant to achieve less than $\epsilon\sqrt{\frac{d}{m}}$ error, which shows that the algorithm needs only $O\left(\frac{\kappa}{h}\log\frac{1}{\epsilon}\right)=O(\kappa\log\frac{1}{\epsilon})$ parallel steps. Appendix A.6 gives a partial proof of the guarantee of Algorithm 2. The other part of the proof is similar to that in Algorithm 1, so we omit it here.

Theorem 4. *Let f be a function such that $0\prec m\cdot I_d\preceq\nabla^2 f(x)\preceq L\cdot I_d$ for all $x\in\mathbb{R}^d$. Let Y be a random point drawn from the density proportional to e^{-f}. Algorithm 2 can generate a random point X such that $W_2(X,Y)\leq\epsilon\sqrt{\frac{d}{m}}$ in $O(\kappa\log\frac{1}{\epsilon})$ parallel steps. Furthermore, each iteration of Algorithm 2 involves computing $\tilde{\Theta}\left(\frac{\sqrt{\kappa}}{\epsilon}\right)$ of ∇fs.*

2.4 Numerical Experiments

In this section, we compare the algorithm from our paper, randomized midpoint method, with the one from [CCBJ17]. We test the algorithms on the liver-disorders dataset and the breast-cancer dataset from UCL machine learning [DG17]. In both datasets, we observe a set of independent samples $\{x_i,y_i\}_{i=1}^m$, where y_i is the label, x_i is the feature and m is the number of samples. We sample from the target distribution $p^*(\theta)\propto\exp\left(-f(\theta)\right)$, where

$$f(\theta)=\frac{\lambda}{2}\left\|\theta\right\|^2+\frac{1}{m}\sum_{i=1}^m\log\left(\exp\left(-y_i x_i^T\theta\right)+1\right),$$

for regularization parameters λ. We set λ to be 10^{-2} in our experiments. Figure 2.1 shows the error of randomized midpoint method and the algorithm from [CCBJ17] with different

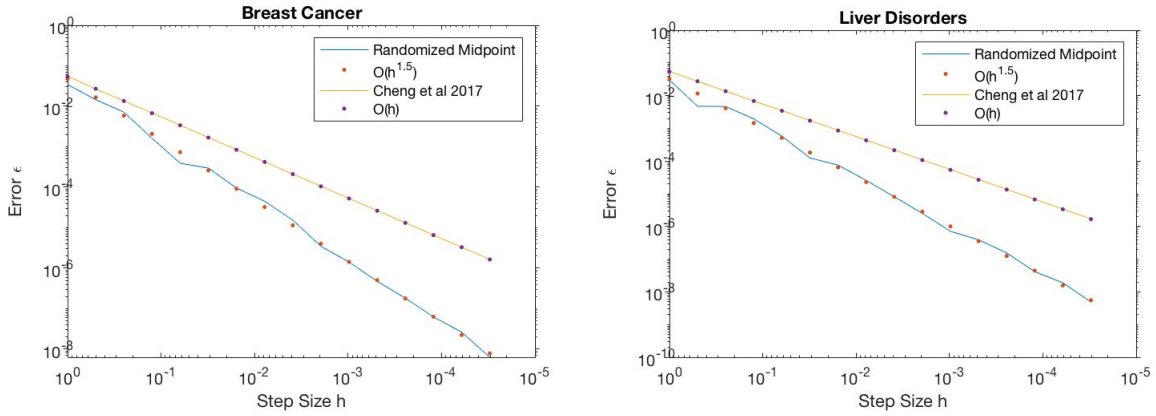


Figure 2.1: Error of random walks with different choice of step size.

step size h. The error is measured by the ℓ_2 distance to the true solution of (2.3) at time $N = 5000$, a time much greater than the mixing time of (2.3) for both datasets. Our results show that the ϵ dependence analysis of our algorithm and that of [CCBJ17] are both tight. However, we note that the logistic function is infinitely differentiable, so there are methods of higher orders for this objective such as the standard midpoint method and Runge–Kutta methods.

Algorithm	# Step	
	Warm Start	**Cold Start**
Langevin Diffusion[DM16, Dal17b]	$\tilde{O}\left(\kappa^2/\epsilon^2\right)$	
Underdamped Langevin Diffusion [CCBJ17]	$\tilde{O}\left(\kappa^2/\epsilon\right)$	
Underdamped Langevin Diffusion2 [DRD18]	$\tilde{O}\left(\kappa^{1.5}/\epsilon + \kappa^2\right)$	
High-Order Langevin Diffusion[MMW$^+$19]	$\tilde{O}\left(\kappa^{19/4}/\epsilon^{1/2} + \kappa^{13/3}/\epsilon^{2/3}\right)$	
Hamiltonian Monte Carlo with Euler Method [MS17]	$\tilde{O}\left(\kappa^{6.5}/\epsilon\right)$	
Hamiltonian Monte Carlo with Collocation Method [LSV18]	$\tilde{O}\left(\kappa^{1.75}/\epsilon\right)$	
Hamiltonian Monte Carlo with Collocation Method 2 [CV19]	$\tilde{O}\left(\kappa^{1.5}/\epsilon\right)$	
Underdamped Langevin Diffusion with Randomized Midpoint Method (This Work)	$\tilde{O}\left(\kappa^{7/6}/\epsilon^{1/3} + \kappa/\epsilon^{2/3}\right)$	

Table 2.1: Summary of iteration complexity. Each step involves $O(1)$-gradient computation. We exclude metropolis-adjusted algorithms in this table. See Chapter 3 for metropolis-adjusted algorithms.

Part II

METROPOLIZED SAMPLING ALGORITHMS

Chapter 3

LOGSMOOTH GRADIENT CONCENTRATION AND RUNTIMES UPPER BOUNDS

This chapter is based on [LST20], with Yin Tat Lee and Kevin Tian.

3.1 Introduction

The problem we address in this chapter is determining the mixing time of the well-studied Metropolized Hamiltonian Monte Carlo (HMC) algorithm[1], when sampling from a target distribution whose log-density is smooth and strongly concave. Indeed, as it is the default sampler implementation in a variety of popular packages [Aba16, CGH$^+$17], understanding Metropolized HMC is of high practical importance. Moreover, the specific setting we study, where the target distribution has a density proportional to $\exp(-f)$ for a function f with quadratic upper and lower bounds, is commonplace in applications arising from multivariate Gaussians, logistic regression models, and structured mixture models [DCWY19]. This setting is also of great theoretical interest because of its connection to a well-understood setting in convex optimization [Nes03], where matching upper and lower bounds have long-been known. Similar guarantees are much less well-understood in sampling settings, and exploring the connection is an active research area (e.g. [MCJ$^+$18, Tal19] and references therein). Throughout the introduction, we will refer to this setting as the "condition number regime" for logconcave sampling, as without a finite condition number, black-box sampling guarantees exist, but typically have a large dimension dependence in the mixing time [Vem05].

Many algorithms have been proposed for sampling from logconcave distributions, mainly falling into two categories: zeroth-order methods and first-order methods. Zeroth-order methods only use the information on the density of the distribution by querying the value of f to inform the algorithm trajectory. First-order methods have access to the gradient information of f in addition to the value of f at a query point. This class of methods

[1]Metropolized HMC also refers to a family of algorithms which takes multiple *leapfrog* steps, see Algorithm 3. In this work, we study the variant which takes one leapfrog step, to analyze convergence behavior under minimal assumptions on the log-density (i.e. in the absence of higher-derivative bounds past smoothness).

usually involves simulating a continuous Markov process whose stationary distribution is exactly the target distribution. To simulate a random process in discrete time, one approach is to choose a small-enough step size so that the behavior of the discrete Markov process is not too different from that of the original Markov process over a small time interval. This discretization strategy is typical of sampling algorithms with a polynomial dependence on ϵ^{-1}, where ϵ is the target total variation distance to the stationary distribution [Dal17b, CCBJ17, DM$^+$19, MMW$^+$19, MS17, LSV18, CV19, SL19]. However, for precise values of ϵ, bounding the error incurred by the discretization is typically not enough, leading to prohibitively large runtimes.

On top of the discretization, one further can apply a Metropolis-Hastings filter to adjust the stationary distribution of the Markov process, so that the target distribution is attained in the long run. Studying the non-asymptotic behavior of Metropolized variants of the Langevin dynamics and HMC has been considered in a large number of works [RT96a, RT96b, PST$^+$12, BRH13, XSL$^+$14, DCWY19, CDWY20]. Indeed, the standard discretizations of these methods are identical, which was observed in prior work (see Appendix B.1); we will refer to them both as Metropolized HMC. The works which inspired this study in particular were due to [DCWY19, CDWY20], which showed that the mixing time of Metropolized HMC was bounded by roughly $\max(\kappa^{1.5}\sqrt{d}, \kappa d)$, with logarithmic dependence on the target accuracy ϵ, where κ is the *condition number*[2] of the negative log-density f. In the poly(ϵ^{-1}) runtime regime, the recent work [DMM19] obtains a total variation mixing time bound which scales as $\tilde{O}(\kappa d^2/\epsilon^4)$, which is to our knowledge the only bound known with linear dependence on κ; on the other hand, [SL19] gives an algorithm that depends on $\kappa^{7/6}$ for Wasserstein-2 distance, but with better dependence on all other parameters (see Table 3.1).

By a plausible assumption on the existence of a gap between the complexity of sampling and optimization in the logconcave setting, it is reasonable to believe that a linear dependence on κ is necessary. More specifically, it is well-known that gradient-based optimization algorithms require at least $\min(d, \sqrt{\kappa})$ queries to an oracle providing first-order information [Bub15]; for the worst-case instance, a quadratic in the graph Laplacian of a length-d path, there is a corresponding quadratic gap with sampling a uniform point via a random walk, which mixes in roughly d^2 iterations. We believe understanding the tight

[2]The condition number of a function is the ratio of its smoothness and strong convexity parameters, and is the standard parameter in measuring the complexity of algorithms in sampling and optimization in this regime.

dependence of the mixing time of popular sampling algorithms on natural parameters such as the condition number is fundamental to the development of the field of sampling, just as characterizing the tight complexity of convex optimization algorithms has resulted in rapid recent progress in the area, by giving researchers goalposts in algorithm design. To that end, this work addresses the following question.

Question 1. *What is the mixing time of the Metropolized HMC algorithm?*

We give a comparison of (selected recent) prior work in Table 3.1; for a more complete discussion, we refer the reader to the excellent discussion in [DCWY19, CDWY20]. We note that for the last two rows, the dependence on ϵ is logarithmic, and the notion of mixing is in total variation distance, a much stronger notion than the Wasserstein metric used in all other runtimes listed. We omit logarithmic factors for simplicity. We remark that several works obtain different rates under stronger assumptions on the log-density f, such as higher-order smoothness (e.g. a Lipschitz Hessian) or moment bounds; as this work studies the basic condition number setting with no additional assumptions, we omit comparison to runtimes of this type.

Algorithm	**Mixing Time**	**Metric**
Langevin Diffusion [Dal17b]	κ^2/ϵ^2	$W_2{}^3$
High-Order Langevin Diffusion [MMW+19]	$\kappa^{19/4}/\epsilon^{1/2} + \kappa^{13/3}/\epsilon^{2/3}$	
HMC 1 (Collocation Method) [LSV18]	$\kappa^{1.5}/\epsilon$	
HMC 2 (Collocation Method) [CV19]	$\kappa^{1.5}/\epsilon$	
ULD 1 (Euler Method) [CCBJ17]	$\kappa^{1.5}/\epsilon$	
ULD 2 (Euler Method) [DRD18]	$\kappa^{1.5}/\epsilon + \kappa^2$	
ULD 3 (Random Midpoint Method) [SL19]	$\kappa^{7/6}/\epsilon^{1/3} + \kappa/\epsilon^{2/3}$	
Unadjusted Langevin Dynamics [DCWY19]	$\kappa d^2/\epsilon^4$	TV
Metropolized HMC & MALA [CDWY20]	$\kappa^{1.5}\sqrt{d} + \kappa d$	
Metropolized HMC & MALA (This work)	κd	

Table 3.1: Mixing times for algorithms in the condition number regime of logconcave sampling.

3.1.1 Contribution

Towards improving our understanding of Question 1, we show that there is an algorithm which runs Metropolized HMC (defined in Algorithm 3) for $O(\kappa d \log^3(\kappa d/\epsilon))$ iterations[4], for sampling from a density $\exp(-f(x))$ defined on $\mathbb{R}^d$, where f has a condition number of κ, and produces a point from a distribution with total variation at most ϵ away from the target density, for any $\epsilon > 0$. This is the first mixing-time guarantee for any algorithm in the high-accuracy regime accessing first-order function information from the log-density f attaining linear dependence on the condition number κ, without additional smoothness assumptions (i.e. higher-order derivative bounds). Our mixing time bound improves upon a recent bound attaining linear dependence on κ due to [DMM19], of $\tilde{O}(\kappa d^2/\epsilon^4)$, in all parameters. Moreover, our dependence on the dimension d matches the prior state-of-the-art [DCWY19, CDWY20], and our algorithm does not require a warm start, as it explicitly bounds warmness dependence from a known starting distribution.

The starting point of our analysis is the mixing time analysis framework for the HMC algorithm in [DCWY19, CDWY20]. However, we introduce several technical modifications to overcome barriers in their work to obtain our improved mixing time bound, which we now discuss. We hope these tools may be of broader interest to both the community studying first-order sampling methods in the smooth, strongly logconcave regime, and sampling researchers in general.

Gradient concentration

How large is the norm of the gradient of a "typical" point drawn from the density $\exp(-f)$? It has been observed in a variety of recent works studying sampling algorithms [LSV18, SL19, VW19] that the *average* gradient norm of a point drawn from the target density is bounded by $\sqrt{Ld}$, where L is the smoothness parameter of the function f and d is the ambient dimension; this observation has been used in obtaining state-of-the-art sampling algorithms in the $\text{poly}(\epsilon^{-1})$ runtime regime. However, for runtimes obtaining a $\text{polylog}(\epsilon^{-1})$ runtime, this guarantee is not good enough, as it must hold for all points in a set of substantially larger measure than guaranteed by e.g. Markov's inequality. The weaker high-probability guarantee that the gradient norm is bounded by $\sqrt{Ld} \cdot \sqrt{\kappa}$ follows directly from sub-Gaussian concentration on the point x, and a Lipschitz guarantee on

[4]The precise statement of our algorithmic guarantee can be found as Theorem 11.

the gradient norm. Indeed, this weaker bound is the bottleneck term in the analysis of [DCWY19, CDWY20], and prevents a faster algorithm when $\kappa > d$. Can we improve upon the average-case guarantee more generally when the log density f is smooth?

For quadratic f, it is easy to see that the average gradient norm bound can be converted into a high-probability guarantee. We show that a similar concentration guarantee holds for *all* logsmooth, strongly logconcave densities, which is the starting point of our improved mixing time bound. Our concentration proof follows straightforwardly from a Hessian-weighted variant of the well-known Poincaré inequality, combined with a reduction due to Herbst, as explored in [Led99].

Mixing time analysis

The study of Markov chains producing iterates $\{x_k\}$, where the transition $x_k \to x_{k+1}$ is described by an algorithm whose steady-state is a stationary distribution π^*, and x_0 is drawn from an initial distribution π_0, primarily focuses on characterizing the rate at which the distribution of iterates of the chain approaches π^*. To obtain a mixing time bound, i.e. a bound on the number of iterations needed for our algorithm to obtain a distribution within total variation distance ϵ of the stationary π^*, we follow the general framework of bounding the *conductance* of the random walk defined by Metropolized HMC, initiated in a variety of works on Markov chain mixing times (e.g. [SJ89, LS93]). In particular, [LS93] showed how to use the generalized notion of s-conductance to account for a small-probability "bad" region with poor random walk behavior. In our work, the "good" region Ω will be the set of points whose gradient has small norm. However, our mixing time analysis requires several modifications from prior work to overcome subtle technical issues.

Average conductance. As in prior work [CDWY20], because of the exponential warmness $\kappa^{d/2}$ of the starting distribution used, we require extensions in the theory of *average conductance* [LK99] to obtain a milder dependence on the warmness, i.e. doubly logarithmic rather than singly logarithmic, to prevent an additional dimension dependence in the mixing time. The paper [CDWY20] obtained this improved dependence on the warmness by generalizing the analysis of [GMT06] to continuous-time walks and restrictions to high-probability regions. This analysis becomes problematic in our setting, as our region Ω may be nonconvex, and the restriction of a strongly logconcave function to a noncon-

vex set is possibly not even logconcave. This causes difficulties when bounding standard conductance notions which may depend on sets of small measure, because these sets may behave poorly under restriction by Ω (e.g. in the proof of Lemma 63).

Blocking conductance. To mitigate the difficulty of poor small-set conductance due to the nonconvexity of Ω, we use the *blocking conductance* analysis of [KLM06], which averages conductance bounds of sets with measure *equal to* some specified values in a range lower-bounded by roughly the inverse-warmness. In our case, this is potentially problematic, as the set where our concentration result guarantees that the norm of the gradient is not much larger than its mean has measure roughly $1-\exp(-\sqrt{d})$, which is too small to bound the behavior of sets of size $\kappa^{-d/2}$ required by the quality of the warm start. However, we show that, perhaps surprisingly, the analysis of the blocking conductance is not bottlenecked by the worse quality of the gradient concentration required. In particular, the $\kappa^{1.5}\sqrt{d}$ runtime of [DCWY19, CDWY20] resulted from the statement, with probability at most $\exp(-d)$, the gradient norm is bounded by $\sqrt{L\kappa d}$. We are able to sharpen this by Corollary 2 to $\sqrt{L}d$, trading off a κ for a d, which is sufficient for our tighter runtime.

Boosting to high accuracy. Finally, the blocking conductance analysis of [KLM06] makes an algorithmic modification. In particular, letting $d\pi_k$ be the density after running k steps of the Markov chain from π_0, the analysis of [KLM06] is able to guarantee that the *average* density $d\rho_k \stackrel{\text{def}}{=} \frac{1}{k}\sum_{0\leq i<k} d\pi_i$ converges to $d\pi^*$ at a rate roughly $1/k$, with a factor depending on the average conductance[5]. In our case, we can show that in roughly $O(\kappa d)$ iterations of Algorithm 3, the distance $\|\rho_k - \pi^*\|_{\text{TV}}$ is bounded by a constant. However, as the analysis requires averaging with a potentially poor starting distribution, it is not straightforward to obtain a rate of convergence with dependence $\log \epsilon^{-1}$ for potentially small values of ϵ, rather than the ϵ^{-1} dependence typical of $1/k$ rates. Moreover, it is unclear in our setting how to apply standard arguments [AD86, LW95] which convert mixing time guarantees for obtaining a constant total variation distance to guarantees for total variation distance ϵ with a logarithmic overhead on ϵ, because the definition of mixing time used is a worst-case notion over all starting points. We propose an alternative reduction based on mixing-time guarantees over arbitrary starting distributions of a specified warmness, which we use to boost our constant-accuracy mixing-time guarantee (see Ap-

[5]We note averaging has been observed to improve sampling accuracy in a different setting [DMM19]; we leave as an interesting open direction whether this averaging is necessary for our method.

pendix B.3.4 for a more formal treatment). While it is simple and inspired by classical coupling-based reduction arguments, to the best of our knowledge this reduction is new in the literature, and may be of independent interest.

3.2 Preliminaries

3.2.1 Notation

We denote the set $1 \le i \le d$ by $[d]$. For $S \subseteq \mathbb{R}^d$, S^c is its complement $\mathbb{R}^d \setminus S$. $\|\cdot\|$ is the ℓ_2 norm ($\|x\|^2 = \sum_{i\in[d]} x_i^2$ for $x \in \mathbb{R}^d$). Differentiable $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth if

$$f(x)+\langle\nabla f(x), y-x\rangle+\frac{\mu}{2}\|y-x\|^2 \le f(y) \le f(x)+\langle\nabla f(x), y-x\rangle+\frac{L}{2}\|y-x\|^2,\ \forall x, y \in \mathbb{R}^d.$$

It is well-known that smoothness is equivalent to having a Lipschitz gradient, i.e. $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$, and when f is twice-differentiable, smoothness and strong convexity imply

$$\mu I_d \preceq \nabla^2 f(x) \preceq L I_d$$

everywhere, where I_d is the identity and $\preceq$ is the Loewner order. In this paper, function $f : \mathbb{R}^d \to \mathbb{R}$ will always be differentiable, L-smooth, and μ-strongly convex, with minimizer x^*. We let $\kappa \stackrel{\text{def}}{=} L/\mu \ge 1$ be the *condition number* of f. We define the Hamiltonian $\mathcal{H}$ of $(x, v) \in \mathbb{R}^d \times \mathbb{R}^d$ by

$$\mathcal{H}(x, v) = f(x) + \frac{1}{2}\|v\|^2.$$

$\mathcal{N}(\mu, \Sigma)$ is the Gaussian density centered at a point $\mu \in \mathbb{R}^d$ with covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. For $A \subseteq \mathbb{R}^d$ and a distribution π, we write

$$\pi(A) \stackrel{\text{def}}{=} \int_{x\in A} d\pi(x).$$

We fix the definition of the distribution density $d\pi^*(x)$, where $d\pi^*(x)/dx \propto \exp(-f(x))$ has

$$d\pi^*(x) = \frac{\exp(-f(x))dx}{\int_{\mathbb{R}^d}\exp(-f(y))dy},\ \int_{\mathbb{R}^d} d\pi^*(x) = 1.$$

The marginal in the first argument of the density on $\mathbb{R}^d\times\mathbb{R}^d$ proportional to $\exp(-\mathcal{H}(x, v))$ is $d\pi^*$; we overload $d\pi^*(x, v)$ to mean this density. For distributions ρ, π on $\mathbb{R}^d$, the total variation is

$$\|\rho - \pi\|_{\text{TV}} \stackrel{\text{def}}{=} \sup_{A\subseteq\mathbb{R}^d} |\rho(A) - \pi(A)| = \frac{1}{2}\int_{\mathbb{R}^d}\left|\frac{d\rho}{d\pi}(x) - 1\right| d\pi(x).$$

We say that a distribution π is β-warm with respect to another distribution ρ if

$$\sup_{x \in \mathbb{R}^d} \frac{d\pi}{d\rho}(x) \le \beta.$$

We define the expectation and variance with respect to a distribution in the usual way:

$$\mathbb{E}_\pi[g] \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} g(x) d\pi(x), \ \mathrm{Var}_\pi[g] \stackrel{\text{def}}{=} \mathbb{E}_\pi[g^2] - (\mathbb{E}_\pi[g])^2 .$$

Finally, to simplify some calculations, we assume that d is bounded below by a small constant. In the absence of this bound, general-purpose mixing times for logconcave functions with no dependence on κ attain our stated guarantees.

3.2.2 Algorithm

We state the Metropolized HMC algorithm to be analyzed throughout the remainder of this paper. We remark that it may be thought of as a symplectic discretization of the continuous-time Hamiltonian dynamics for $\mathcal{H}(x, v) = f(x) + \frac{1}{2}\|v\|^2$,

$$\frac{dx}{dt} = \frac{\partial \mathcal{H}(x,v)}{\partial v} = v, \ \frac{dv}{dt} = -\frac{\partial \mathcal{H}(x,v)}{\partial x} = -\nabla f(x).$$

The HMC process can be thought of as a dual velocity v accumulating the gradient of the primal point x, with the primal point being guided by the velocity, similar to the classical mirror descent algorithm. The algorithm resamples v each timestep to attain the correct stationary distribution.

From a point $x \in \mathbb{R}^d$, we define $\mathcal{P}_x$ to be the distribution of $\tilde{x}_k$ after one step of Algorithm 3 starting from $x_k = x$. Similarly, $\mathcal{T}_x$ is the distribution of x_{k+1} starting at $x_k = x$, i.e. after the accept-reject step. Algorithm 3 uses the subprocedure `Leapfrog`, which enjoys the following property.

Lemma 1. *If* `Leapfrog`$(\eta, x, -v) = (\tilde{x}, \tilde{v})$, *then* `Leapfrog`$(\eta, \tilde{x}, -\tilde{v}) = (x, v)$.

Proof. Recall that `Leapfrog`$(x, -v) = (\tilde{x}, \tilde{v})$ implies

$$\tilde{v} = -v - \frac{\eta}{2}\nabla f(x) - \frac{\eta}{2}\nabla f(\tilde{x}), \ \tilde{x} = x - \eta v - \frac{\eta^2}{2}\nabla f(x).$$

Reversing these definitions yields the claim. □

Corollary 1. *$d\pi^*$ is a stationary distribution for the Markov chain defined by Algorithm 3.*

Algorithm 3 Metropolized Hamiltonian Monte Carlo: $\texttt{HMC}(\eta, x_0, f)$

Input: Initial point $x_0 \in \mathbb{R}^d$, step size η.

Output: Sequence $\{x_k\}$, $k \geq 0$.

for $k \geq 0$ **do**
  Draw $v_k \sim \mathcal{N}(0, I_d)$.
  $(\tilde{x}_k, \tilde{v}_k) \leftarrow \texttt{Leapfrog}(\eta, x_k, v_k)$.
  Draw u uniformly in $[0, 1]$.
  if $u \leq \min\{1, \exp(\mathcal{H}(x, v) - \mathcal{H}(\tilde{x}, \tilde{v}))\}$ **then**
    $x_{k+1} \leftarrow \tilde{x}_k$.
  else
    $x_{k+1} \leftarrow x_k$.
  end if
end for

Algorithm 4 Leapfrog: $\texttt{Leapfrog}(\eta, x, v)$

Input: Points $x, v \in \mathbb{R}^d$, step size η.

Output: Points $\tilde{x}, \tilde{v} \in \mathbb{R}^d$.

$v' \leftarrow v - \frac{\eta}{2}\nabla f(x)$.
$\tilde{x} \leftarrow x + \eta v'$.
$\tilde{v} \leftarrow v' - \frac{\eta}{2}\nabla f(\tilde{x})$.

Proof. We show that for $z = (x, v)$, $d\pi^*(z)/dz \propto \mathcal{H}(z)$ is the stationary distribution on (x_k, v_k); correctness then follows from π^* having the correct marginal. Stationarity follows if and only if

$$d\pi^*(x, v)\mathcal{T}_{x,v}(\tilde{x}, \tilde{v}) = d\pi^*(x, v)\mathcal{T}_{\tilde{x},\tilde{v}}(x, v)$$

for all pairs (x, v), $(\tilde{x}, \tilde{v})$, where we overload the definition of $\mathcal{T}$ to be the transition distribution from a point (x, v). By the standard proof of correctness for the Metropolis-Hastings correction, i.e. choosing an acceptance probability proportional to

$$\min\left\{1, \frac{d\pi^*(\tilde{x}, \tilde{v})\mathcal{P}_{\tilde{x},\tilde{v}}(x, v)}{d\pi^*(x, v)\mathcal{P}_{x,v}(\tilde{x}, \tilde{v})}\right\},$$

it suffices to show that $\mathcal{P}_{\tilde{x},\tilde{v}}(x, v) = \mathcal{P}_{x,v}(\tilde{x}, \tilde{v})$. Note that $\mathcal{P}_{x,v}$ is a deterministic proposal, and uniquely maps to a point $(\tilde{x}, \tilde{v})$. Moreover, by symmetry of $\mathcal{H}$ in the second argument, iteration k of Algorithm 3 is equivalent to drawing v_k, negating it, and then running

`Leapfrog`. Correctness for this equivalent algorithm follows by Lemma 1. □

3.3 Gradient concentration

In this section, we give a bound on how well the norm of the gradient $\|\nabla f(x)\|$ concentrates when f is smooth and $x \sim d\pi^*(x)/dx \propto \exp(-f(x))$. First, we recall the following "Hessian-weighted" variant of the Poincaré inequality, which first appeared in [BL76].

Theorem 2 (Hessian Poincaré)**.** *For probability density $d\pi^*(x)/dx \propto \exp(-f(x))$, and continuously differentiable function $g : \mathbb{R}^d \to \mathbb{R}$ with bounded variance with respect to π^*,*

$$\mathrm{Var}_{\pi^*}[g] \leq \int_{\mathbb{R}^d} \left\langle \left(\nabla^2 f(x)\right)^{-1} \nabla g(x), \nabla g(x) \right\rangle d\pi^*(x).$$

An immediate corollary of Theorem 2 is that the Poincaré constant of a μ-strongly logconcave distribution is at most μ^{-1}. While it does not appear to have been previously stated in the literature, our concentration bound can be viewed as a simple application of an argument of Herbst which reduces concentration to an isoperimetric inequality such as Theorem 2; an exposition of this technique can be found in [Led99]. We now state the concentration result.

Theorem 3 (Gradient norm concentration)**.** *If twice-differentiable $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and convex, then for $d\pi^*(x)/dx \propto \exp(-f(x))$, and all $c > 0$,*

$$\Pr_{\pi^*}\left[\|\nabla f(x)\| \geq \mathbb{E}_{\pi^*}\left[\|\nabla f\|\right] + c\sqrt{L}\log d\right] \leq 3d^{-c}.$$

Proof. Let $G(x) \stackrel{\text{def}}{=} \|\nabla f(x)\|$, and let $g(x) \stackrel{\text{def}}{=} \exp(\frac{1}{2}\lambda G(x))$. Clearly g is continuously differentiable. Moreover, suppose first for simplicity that f is strongly convex; then the existence of the variance of g follows from the well-known fact that f has sub-Gaussian tails (e.g. [DCWY19], Lemma 1) and Lipschitzness of its gradient, from which the sublevel sets of the gradient norm grow more slowly than the decay of $\|x - x^*\|_2$. The final conclusion has no dependence on the strong concavity of f, and we can extend this to arbitrary convex functions by regularizing by a small amount of quadratic regularizer (which only affects smoothness) and taking a limit as the regularizer amount vanishes. We now apply Theorem 2, which implies (noting that the gradient of $\|\nabla f\|$ is $\nabla^2 f \frac{\nabla f}{\|\nabla f\|}$)

$$\begin{aligned}\mathbb{E}_{\pi^*}\left[\exp(\lambda G)\right] - \mathbb{E}_{\pi^*}\left[\exp\left(\frac{\lambda G}{2}\right)\right]^2 &\leq \frac{\lambda^2}{4}\mathbb{E}_{\pi^*}\left[\left\langle (\nabla^2 f)\frac{\nabla f}{\|\nabla f\|}, \frac{\nabla f}{\|\nabla f\|}\right\rangle \exp(\lambda G)\right] \\ &\leq \frac{L\lambda^2}{4}\mathbb{E}_{\pi^*}\left[\exp(\lambda G)\right].\end{aligned}$$

In the last inequality we used smoothness. Letting $H(\lambda) \stackrel{\text{def}}{=} \mathbb{E}_{\pi^*}[\exp(\lambda G)]$, for $\lambda < \frac{2}{\sqrt{L}}$,

$$H(\lambda) \leq \frac{1}{1 - \frac{L\lambda^2}{4}} H\left(\frac{\lambda}{2}\right)^2.$$

Using this recursively, we have

$$H(\lambda) \leq \prod_{k=0}^{\infty} \left(\frac{1}{1 - \frac{L\lambda^2}{4^{k+1}}}\right)^{2^k} \lim_{\ell \to \infty} H\left(\frac{\lambda}{\ell}\right)^{\ell}.$$

There are two things to estimate on the right hand side. First, for sufficiently large ℓ,

$$\mathbb{E}_{\pi^*}\left[\exp\left(\frac{\lambda G}{\ell}\right)\right]^{\ell} \approx \left(1 + \mathbb{E}_{\pi^*}\left[\frac{\lambda G}{\ell}\right]\right)^{\ell} \approx \exp\left(\lambda \mathbb{E}_{\pi^*}[G]\right).$$

Second, letting $C = \frac{L\lambda^2}{4} < 1$, [BL97] showed that

$$\prod_{k=0}^{\infty} \left(\frac{1}{1 - \frac{C}{4^k}}\right)^{2^k} \leq \frac{1 + \sqrt{C}}{1 - \sqrt{C}}.$$

For completeness, we show this in Appendix D.2. Altogether, we have that for all $\lambda < \frac{2}{\sqrt{L}}$,

$$\mathbb{E}_{\pi^*}[\exp(\lambda G)] \leq \frac{1 + \frac{1}{2}\sqrt{L}\lambda}{1 - \frac{1}{2}\sqrt{L}\lambda} \exp\left(\lambda \mathbb{E}_{\pi^*}[G]\right).$$

By Markov's inequality on the exponential, we thus conclude that

$$\Pr_{\pi^*}[G \geq \mathbb{E}_{\pi^*}[G] + r] \leq \exp(-\lambda r) \frac{1 + \frac{1}{2}\sqrt{L}\lambda}{1 - \frac{1}{2}\sqrt{L}\lambda}.$$

Finally, letting $\lambda = \frac{1}{\sqrt{L}}$ and $r = c\sqrt{L}\log d$,

$$\Pr_{\pi^*}\left[\|\nabla f\| \geq \sqrt{Ld} + c\sqrt{L}\log d\right] \leq 3d^{-c}.$$

□

As an immediate corollary, we obtain the following.

Corollary 2. *If twice-differentiable $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and strongly convex, then $\forall c > 0$,*

$$\Pr_{\pi^*}\left[\|\nabla f\| \geq \sqrt{Ld} + c\sqrt{L}\log d\right] \leq 3d^{-c}.$$

Proof. It suffices to show that

$$\mathbb{E}_{\pi^*}\left[\|\nabla f\|\right] \leq \sqrt{Ld}. \tag{3.1}$$

This was observed in [Dal17a, VW19]; we adapt a proof here. Observe that because

$$\nabla \cdot (\nabla f(x)\pi^*(x)) = \Delta f(x)\pi^*(x) - \langle \nabla f(x), \nabla f(x) \rangle \pi(x),$$

where $\nabla\cdot$ is divergence and Δ is the Laplacian operator, integrating both sides and noting that the boundary term vanishes,

$$\mathbb{E}_{\pi^*}\left[\|\nabla f\|^2\right] = \mathbb{E}_{\pi^*}\left[\Delta f\right] \leq Ld.$$

In the last equality, we used smoothness of f. (3.1) then follows from concavity of the square root. □

We remark that for densities $d\pi^*$ where a log-Sobolev variant of the inequality in Theorem 2 holds, we can sharpen the bound in Corollary 2 to $O(d^{-c^2})$; we provide details in Appendix B.2. This sharpening is desirable for reasons related to the warmness of starting distributions for sampling from π^*, as will become clear in Section 3.4. However, the "Hessian log-Sobolev" inequality is strictly stronger than Theorem 2, and does not hold for general strongly logconcave distributions [BL00]. Correspondingly, the concentration arguments derivable from Poincaré inequalities appear to be weaker [Led99]: we find exploring the tightness of Corollary 2 to be an interesting open question.

3.4 Mixing time bounds via blocking conductance

We first give a well-known bound of the warmness of an initial distribution; this starting distribution also was used in prior work in this setting [Dal17a, DCWY19]

Lemma 2 (Initial warmness)**.** *For $d\pi^* \propto \exp(-f(x))dx$ where f is L-smooth and μ-strongly convex with minimizer x^*[6], $\pi_0 = \mathcal{N}(x^*, L^{-1}I_d)$ is a $\kappa^{d/2}$-warm distribution with respect to π^*.*

Proof. By smoothness and strong convexity, and the density of a Gaussian distribution,

$$d\pi_0(x) = \frac{\exp\left(-\frac{L}{2}\|x-x^*\|^2\right)dx}{(2\pi L^{-1})^{d/2}},\ d\pi^*(x) = \frac{\exp(-f(x))dx}{\int_{\mathbb{R}^d}\exp(-f(y))dy} \geq \frac{\exp\left(-\frac{L}{2}\|x-x^*\|^2\right)dx}{(2\pi\mu^{-1})^{d/2}}.$$

In the last inequality we normalized by $\exp(-f(x^*))$. Combining these bounds yields the result. □

[6]We remark that the minimizer x^* can be efficiently found using e.g. an accelerated gradient method, to a degree of accuracy which does not bottleneck the runtime of Metropolized HMC by more than mild logarithmic factors. We defer a discussion of performance under inexact knowledge of the parameters x^*, L to [DCWY19], and assume their exact knowledge for simplicity in this work.

Let $d\pi_k$ be the density of x_k after running k steps of Algorithm 3, where x_0 is drawn from $\pi_0 = \mathcal{N}(x^*, L^{-1}I_d)$. Moreover, let $d\rho_k \stackrel{\text{def}}{=} \frac{1}{k}\sum_{0\le i<k} d\pi_i$ be the average density over the first k iterations. In Section 3.4.1, we will show how to use the blocking conductance framework of [KLM06] to obtain a bound on the number of iterations k required to obtain a constant-accuracy distribution. We then show in Section 3.4.2 that we can boost this guarantee to obtain total variation ϵ for arbitrary $\epsilon > 0$ with logarithmic overhead, resulting in our main mixing time claim, Theorem 11.

3.4.1 Constant-accuracy mixing

We state the results required to prove a mixing-time bound for constant levels of total variation from the stationary measure π^*. All proofs are deferred to Appendix B.3. The first result is a restatement of the main result of [KLM06], modified for our purposes; recall that ρ_k is an average over the distributions π_i for $0 \le i < k$. Finally, we define $Q(S) \stackrel{\text{def}}{=} \int_S \mathcal{T}_x(S^c)d\pi^*(x)$ to be the probability one step of the walk starting at random point in a set S leaves the set.

Theorem 4 (Blocking conductance mixing bound)**.** *Suppose the starting distribution π_0 is β-warm with respect to π^*. Moreover, suppose for some c, and for all $c \le t \le \frac{1}{2}$, we have a bound*

$$\frac{\pi^*(S)}{Q(S)^2} \le \phi(t), \text{ for all } S \subseteq \mathbb{R}^d \text{ with } \pi^*(S) = t, \tag{3.2}$$

for a decreasing function ϕ on the range $[c, \frac{1}{4}]$, and $\phi(x) \le M$ for $x \in [\frac{1}{4}, \frac{1}{2}]$. Then,

$$\left\|\rho_k - \pi^*\right\|_{\text{TV}} \le \beta c + \frac{32}{k}\left(\int_c^{1/4} \phi(x)dx + \frac{M}{4}\right).$$

At a high level, the mixing time requires us to choose a threshold c which is inversely-proportional to the warmness, and bound the average value of a function $\phi(t)$ in the range $[c, \frac{1}{2}]$, where $\phi(t)$ serves as an indicator of how "bottlenecked" sets of measure exactly equal to t are.

Next, by using a logarithmic isoperimetric inequality from [CDWY20], we show in the following lemma that we can bound $\frac{\pi^*(S)}{Q(S)^2}$ when $\pi^*(S)$ is in some range.

Lemma 3. *Suppose for $\Omega \subset R^d$ with $\pi^*(\Omega) = 1 - s$, and all $x, y \in \Omega$ with $\|x - y\| \le \eta$,*

$$\left\|\mathcal{T}_x - \mathcal{T}_y\right\|_{\text{TV}} \le 1 - \alpha. \tag{3.3}$$

Then, if π^* is μ-strongly logconcave, $\eta\sqrt{\mu} < 1$, and $s \le \frac{\eta\sqrt{\mu}t}{16}$, for all $t \le \frac{1}{2}$,

$$\frac{\pi^*(S)}{Q(S)^2} \le \frac{2^{16}}{\alpha^2\eta^2\mu t\log(1/t)},\ \forall S \text{ with } \pi^*(S) = t.$$

For a more formal statement and proof, see Lemma 63. Note that in particular the lower range of t required by Theorem 4 is at least inversely proportional to the warmness, which causes the gradient norm bound obtained by the high-probability set Ω to lose roughly a $\sqrt{d}$ factor. To this end, for a fixed positive $\epsilon \le 1$, denote

$$\Omega \stackrel{\text{def}}{=} \left\{x \in \mathbb{R}^d \mid \|\nabla f(x)\|_2 \le 5\sqrt{L}d\log\frac{\kappa}{\epsilon}\right\}. \tag{3.4}$$

In Appendix B.4, we show the following.

Lemma 4. *For $\eta^2 \le \frac{1}{20Ld\log\frac{\kappa}{\epsilon}}$ and all x, $y \in \Omega$ with $\|x - y\| \le \eta$,*

$$\|\mathcal{T}_x - \mathcal{T}_y\|_{\text{TV}} \le \frac{7}{8}.$$

By combining these pieces, we are able to obtain an algorithm which mixes to constant total variation distance from the stationary distribution π^* in $\tilde{O}(\kappa d)$ iterations.

Proposition 1. *Let $\epsilon \in [0, 1]$, $\beta = \kappa^{d/2}$. From any β/ϵ-warm initial distribution π_0, running Algorithm 3 for j iterations, where j is uniform between 0 and $k - 1$ for $k > C\kappa d\log\kappa\log\log\beta$ for universal constant C, returns from distribution ρ_k with $\|\rho_k - \pi^*\|_{\text{TV}} < (2e)^{-1}$.*

Proof. Note that ρ_k as defined in the theorem statement is precisely the ρ_k of Theorem 4. Moreover, for the set Ω in (5.16), the probability $x \sim \pi^*$ is not in Ω is bounded via Corollary 2 by

$$s < 3d^{-5d\log_d(\kappa/\epsilon)} < (\kappa/\epsilon)^{-4d}. \tag{3.5}$$

For $\eta = \sqrt{\frac{1}{20Ld\log\frac{\kappa}{\epsilon}}}$ and $c \stackrel{\text{def}}{=} \epsilon/(4\beta e)$, $s \le \frac{\eta\sqrt{\mu}t}{16}$ is satisfied for all t in the range $[c, \frac{1}{2}]$. Thus, we can apply Lemma 3 and conclude that (3.2) holds for the function

$$\phi(t) = \left(20\cdot 2^{22}\kappa d\log\frac{\kappa}{\epsilon}\right)\frac{1}{t\log(1/t)}.$$

Next, note that $\phi(t)$ is decreasing in the range $[c, 1/e]$, and attains its maximum at $t = \frac{1}{2}$ within the range $t \in [\frac{1}{4}, \frac{1}{2}]$, where $2/\log 2 < 3$. Thus, the conditions of Theorem 4 apply, such that

$$\|\rho_k - \pi^*\|_{\text{TV}} \le \frac{1}{4e} + \frac{20\cdot 2^{27}\kappa d\log\frac{\kappa}{\epsilon}}{k}\left(\int_{\epsilon/(4\beta e)}^{1/4}\frac{1}{x\log(1/x)}dx + \frac{3}{4}\right)$$

$$\leq \frac{1}{4e} + \frac{20 \cdot 2^{27} \kappa d \log \frac{\kappa}{\epsilon}}{k} \left(\log \log \left(\frac{4\beta e}{\epsilon} \right) - \log \log 4 + \frac{3}{4} \right).$$

Thus, by choosing k to be a sufficiently large multiple of $\kappa d \log \frac{\kappa}{\epsilon} \log \log \frac{\beta}{\epsilon}$, the guarantee follows. □

3.4.2 High-accuracy mixing

We now state a general framework for turning guarantees such as Proposition 1 into a ϵ-accuracy mixing bound guarantee, with logarithmic overhead in the quantity ϵ. We defer a more specific statement and proof to Appendix B.3.4.

Lemma 5. *Suppose there is a Markov chain with transitions given by $\widetilde{\mathcal{T}}$, and some non-negative integer T_{mix}, such that for every π which is a β/ϵ-warm distribution with respect to π^*,*

$$\|\widetilde{\mathcal{T}}^{T_{\mathrm{mix}}} \pi - \pi^*\|_{\mathrm{TV}} \leq \frac{1}{2e}. \tag{3.6}$$

Then, if π_0 is a β-warm start, and $k \geq T_{\mathrm{mix}} \log(\epsilon^{-1})$,

$$\|\widetilde{\mathcal{T}}^{k} \pi_0 - \pi^*\|_{\mathrm{TV}} \leq \epsilon.$$

At a high level, the proof technique is to couple points according to the total variation bound between $\mathcal{T}^i \pi_0$ and π^* every T_{mix} iterations, while the total variation distance is at least ϵ. This in turn allows us to bound the warmness of the "conditional distribution" of uncoupled points by β/ϵ using the fact that the total variation bound measures the size of the set of uncoupled points, and use the guarantee (3.6) iteratively. We can now state our main claim.

Theorem 5 (Mixing of Hamiltonian Monte Carlo)**.** *There is an algorithm initialized from a point drawn from $\mathcal{N}(x^*, L^{-1} I_d)$, which iterates Algorithm 3 for*

$$O\left(\kappa d \log \left(\frac{\kappa}{\epsilon} \right) \log \left(d \log \frac{\kappa}{\epsilon} \right) \log \left(\frac{1}{\epsilon} \right) \right)$$

iterations, and produces a point from a distribution ρ such that $\|\rho - \pi^*\|_{\mathrm{TV}} \leq \epsilon$.

Proof. Define a Markov chain with transitions $\widetilde{\mathcal{T}}$, whose one-step distribution from an initial point is to run the algorithm of Proposition 1. Note that each step of the Markov chain with transitions $\widetilde{\mathcal{T}}$ requires $O\left(\kappa d \log\left(\frac{\kappa}{\epsilon}\right) \log\left(d \log \frac{\kappa}{\epsilon}\right)\right)$ iterations of Algorithm 3, and the averaging step is easily implementable by sampling a random stopping time at uniform. Moreover, the Markov chain with transitions $\widetilde{\mathcal{T}}$ satisfies (3.6) with $T_{\mathrm{mix}} = 1$, by

the guarantees of Corollary 2. Thus, by running $\log(\epsilon^{-1})$ iterations of this Markov chain, we obtain the required guarantee. $\square$

Chapter 4

LOWER BOUNDS ON METROPOLIZED SAMPLING METHODS FOR WELL-CONDITIONED DISTRIBUTIONS

This chapter is based on [LST21a], with Yin Tat Lee and Kevin Tian.

4.1 Introduction

In this chapter, we address the lower bound on the complexity of Metropolized sampling methods. Demonstrating *lower bounds* on the complexity of sampling tasks (in the well-conditioned regime or otherwise) has proven to be a remarkably challenging problem. To our knowledge, there are very few unconditional lower bounds for sampling tasks (i.e. the complexity of sampling from a family of distributions under some query model). This is in stark contrast to the theory of optimization, where there are matching upper and lower bounds for a variety of fundamental tasks and query models, such as optimization of a convex function under first-order oracle access [Nes03]. This gap in the development of the algorithmic theory of sampling is the primary motivation for our work, wherein we aim to answer the following more restricted question.

What is the complexity of the popular sampling methods, MALA and HMC, for sampling well-conditioned distributions?

The problem we study is still less general than *unconditional query lower bounds* for sampling, in that our lower bounds are *algorithm-specific*; we characterize the performance of particular algorithms for sampling a distribution family. However, we believe asking this question, and developing an understanding of it, is an important first step towards a theory of complexity for sampling. On the one hand, lower bounds for specific algorithms highlight weaknesses in their performance, pinpointing their shortcomings in attaining faster rates. This is useful from an algorithm design perspective, as it clarifies what the key technical barriers are to overcome. On the other hand, the hard instances which arise in designing lower bounds may have important structural properties which pave the way to stronger and more general (i.e. *algorithm-agnostic*) lower bounds.

For these reasons, in this work we focus on characterizing the complexity of the MALA

and HMC algorithms , which are often the samplers of choice in practice, by lower bounding their performance when they are used to sample from densities proportional to $\exp(-f(x))$, where $f : \mathbb{R}^d \to \mathbb{R}$ has a finite condition number. In particular, f is said to have a condition number of $\kappa < \infty$ if it is L-smooth and μ-strongly convex (has second derivatives in all directions in the range $[\mu, L]$), where $\kappa = \frac{L}{\mu}$. We will also overload this terminology and say the density itself has condition number κ. We call such a density (with finite κ) "well-conditioned." Finally, we explicitly assume throughout that $\kappa = O(d^4)$, as otherwise in light of our lower bounds the general-purpose logconcave samplers of [LV07, JLLV20, Che21a] are preferable.

4.1.1 Our results

Our primary contribution is a nearly-tight characterization of the performance of MALA for sampling from two high-dimensional distribution families without a warm start assumption: well-conditioned Gaussians, and the more general family of well-conditioned densities. In Sections 4.3 and 4.4, we prove the following two lower bounds on MALA's complexity, which is a one-parameter algorithm (for a given target distribution) depending only on step size. We also note that we fix a scale $[1, \kappa]$ on the eigenvalues of the function Hessian up front, because otherwise the non-scale-invariance of the step size can be exploited to give much more trivial lower bounds (cf. Appendix C.1).

Theorem 6. *For every step size, there is a target Gaussian on $\mathbb{R}^d$ whose negative log-density always has Hessian eigenvalues in $[1, \kappa]$, such that the relaxation time of MALA is $\Omega(\frac{\kappa\sqrt{d}}{\sqrt{\log d}})$.*

Theorem 7. *For every step size, there is a target density on $\mathbb{R}^d$ whose negative log-density always has Hessian eigenvalues in $[1, \kappa]$, such that the relaxation time of MALA is $\Omega(\frac{\kappa d}{\log d})$.*

To give more context on Theorems 6 and 7, MALA is an example of a Metropolis-adjusted Markov chain, which in every step performs updates which preserve the stationary distribution. Indeed, it can be derived by applying a Metropolis filter on the standard forward Euler discretization of the Langevin dynamics, a stochastic differential equation with stationary density $\propto \exp(-f(x))$:

$$dx_t = -\nabla f(x_t)dt + \sqrt{2}dB_t,$$

where B_t is Brownian motion. Such *Metropolis-adjusted* methods typically provide total variation distance guarantees, and attain logarithmic dependence on the target accuracy.[1] The mixing of such chains is governed by their relaxation time, also known as the inverse *spectral gap* (the difference between 1 and the second-largest eigenvalue of the Markov chain transition operator).

However, in the continuous-space setting, it is not always clear how to relate the relaxation time to the *mixing time*, which we define as the number of iterations it takes to reach total variation distance $\frac{1}{e}$ from the stationary distribution from a given warm start (we choose $\frac{1}{e}$ for consistency with the literature, but indeed any constant bounded away from 1 will do). There is an extensive line of research on when it is possible to relate these two quantities (see e.g. [BGL14]), but typically these arguments are tailored to properties of the specific Markov chain, causing relaxation time lower bounds to not be entirely satisfactory in some cases. We thus complement Theorems 6 and 7 with a *mixing time* lower bound from an exponentially warm start, as follows.

Theorem 8. *For every step size, there is a target density on $\mathbb{R}^d$ whose negative log-density always has Hessian eigenvalues in $[1, \kappa]$, such that MALA initialized at an $\exp(d)$-warm start requires $\Omega(\frac{\kappa d}{\log^2 d})$ iterations to reach e^{-1} total variation distance to the stationary distribution.*

We remark that Theorem 8 is the first *mixing time* lower bound for discretizations of the Langevin dynamics we are aware of, as other related lower bounds have primarily been on relaxation times [CV19, LST20, CLA$^+$20]. Up to now, it is unknown how to obtain a starting distribution for a general distribution with condition number κ with warmness better than κ^d (which is obtained by the starting distribution $\mathcal{N}(x^*, \frac{1}{L}\,\mathrm{id})$ where L is the smoothness parameter and x^* is the mode).[2] A line of work [DCWY19, CDWY20, LST20] analyzed the performance of MALA under this warm start, culminating in a mixing time of $\widetilde{O}(\kappa d)$ as shown in Chapter 3, where $\widetilde{O}$ hides logarithmic factors in κ, d, and the target accuracy. On the other hand, a recent work [CLA$^+$20] demonstrated that MALA obtains

[1] We note this is in contrast with a different family of *unadjusted* discretizations, which are analyzed by coupling them with the stochastic differential equation they simulate (see e.g. [Dal17b, CCBJ17] for examples), at the expense of a polynomial dependence on the target accuracy; we focus on Metropolis-adjusted discretizations in this work.

[2] The warmness of a distribution is the worst-case ratio between the measures it and the stationary assign to a set.

a mixing time scaling as $\widetilde{O}(\text{poly}(\kappa)\sqrt{d})$, when initialized at a *polynomially* warm start,[3] and further showed that such a mixing time is tight (in its dependence on d). They posed as an open question whether it was possible to obtain $\widetilde{O}(\text{poly}(\kappa)d^{1-\Omega(1)})$ mixing from an explicit starting distribution.

We address this question by proving Theorem 8, showing that the $\widetilde{O}(\kappa d)$ rate of [LST20] for MALA applied to a κ-conditioned density is tight up to logarithmic factors from an explicit "bad" warm start. Concretely, to prove Theorems 6-8, in each case we exhibit an $\exp(-d)$-sized set according to the stationary measure where either the chain cannot move in $\text{poly}(d)$ steps with high probability, or must choose a very small step size. Beyond exhibiting a mixing bound, this demonstrates the subexponential warmness assumption in [CLA$^+$20] is truly necessary for their improved bound. To our knowledge, this is the first *nearly-tight* characterization of a specific sampling algorithm's performance in all parameters, and improves lower bounds of [CLA$^+$20, LST20]. It also implies that to go beyond $\widetilde{O}(\kappa d)$ mixing requires a subexponential warm start.

The lower bound statement of Theorem 8 is warmness-sensitive, and is of the following (somewhat non-standard) form: for $\beta = \exp(d)$, we provide a lower bound on the quantity

$$\inf_{\text{algorithm parameters}} \sup_{\substack{\text{starts of warmness} \le \beta \\ \text{densities in target family}}} \text{mixing time of algorithm.}$$

In other words, we are allowed to choose both the hard density and starting distribution adaptively based on the algorithm parameters (in the case of MALA, our choices respond to the step size). We note that this type of lower bound is compatible with standard conductance-based upper bound analyses, which typically only depend on the starting distribution through the warmness parameter.

In Section 4.6, we further study the multi-step generalization of MALA, known in the literature as Hamiltonian Monte Carlo with a leapfrog integrator (which we refer to in this paper as HMC). In addition to a step size η, HMC is parameterized by a number of steps per iteration K; in particular, HMC makes K gradient queries in every step to perform a K-step discretization of the Langevin dynamics, before applying a Metropolis filter. It was recently shown in [CDWY20] that under higher derivative bounds, balancing η and K more carefully depending on problem parameters could break the apparent κd barrier of MALA, even from an exponentially warm start.

[3] As discussed, it is currently unknown how to obtain such a warm start generically.

It is natural to ask if there is a stopping point for improving HMC. We demonstrate that HMC cannot obtain a better relaxation time than $\widetilde{O}(\kappa\sqrt{d}K^{-1})$ for any K, even when the target is a Gaussian. Since every HMC step requires K gradients, this suggests $\widetilde{\Omega}(\kappa\sqrt{d})$ queries are necessary.

Theorem 9. *For every step size and count, there is a target Gaussian on $\mathbb{R}^d$ whose negative log-density always has Hessian eigenvalues in $[1, \kappa]$, such that the relaxation time of HMC is $\Omega(\frac{\kappa\sqrt{d}}{K\sqrt{\log d}})$.*

In Appendix C.2, we also give some lower bounds on how much increasing K can help the performance of HMC in the in-between range $\kappa\sqrt{d}$ to κd. In particular, we demonstrate that if $K \leq d^c$ for some constant $c \approx 0.1$, then the K-step HMC Markov chain can only improve the relaxation time of Theorem 9 by roughly a factor K^2, showing that to truly go beyond a κd relaxation time by more than a $d^{o(1)}$ factor, the step size must scale polynomially with the dimension (Proposition 16). We further demonstrate how to extend the mixing time lower bound of Theorem 8 in a similar manner, demonstrating formally for small K that (up to logarithmic factors) the gradient query complexity of HMC cannot be improved beyond κd by more than roughly a K factor (Proposition 17).

Our mixing lower bound technique in Theorem 8 does not directly extend to give a complementary mixing lower bound for Theorem 9 for all K, but we defer this to interesting future work.

4.1.2 Technical overview

In this section, we give an overview of the techniques we use to show our lower bounds. Throughout for the sake of fixing a scale, we assume the negative log-density has Hessian between id and κ id.

MALA. Our starting point is the observation made in [CLA+20] that for a MALA step size h, the spectral gap of the MALA Markov chain scales no better than $O(h + h^2)$, witnessed by a simple one-dimensional Gaussian. Thus, our strategy for proving Theorems 6 and 7 is to show a dichotomy on the choice of step size: either h is so large such that we can construct an $\exp(d)$-warm start where the chain is extremely unlikely to move (e.g. the step almost always is filtered), or it is small enough to imply a poor spectral gap. In the Gaussian case, we achieve this by explicitly characterizing the rejection probability

and demonstrating that choosing the "small ball" warm start where $\|x\|_2^2$ is smaller than its expectation by a constant ratio suffices to upper bound h.

Given the result of Theorem 6, we see that if MALA is to move at all with decent probability from an exponentially warm start, we must take $h \ll 1$, so the spectral gap in this regime is simply $O(h)$. We now move onto the more general well-conditioned setting. As a thought experiment, we note that the upper bound analyses of [DCWY19, CDWY20, LST20] for MALA have a dimension dependence which is bottlenecked by the *noise term* only. In particular, the MALA iterates apply a filter to the move $x' \leftarrow x - h\nabla f(x) + \sqrt{2h}g$, where $g \sim \mathcal{N}(0, \mathrm{id})$ is a standard Gaussian vector. However, even for the more basic "Metropolized random walk" where the proposal is simply $x' \leftarrow x + \sqrt{2h}g$, the dimension dependence of upper bound analyses scales linearly in d. Thus, it is natural to study the effect of the noise, and construct a hard distribution based around it.

We first formalize this intuition, and demonstrate that for step sizes not ruled out by Theorem 6, all terms in the rejection probability calculation other than those due to the effect of the noise g are low-order. Moreover, because the effect of the noise is coordinatewise separable (since $\mathcal{N}(0, \mathrm{id})$ is a product distribution), to demonstrate a $\widetilde{O}(\frac{1}{\kappa d})$ upper bound on h it suffices to show a hard one-dimensional distribution where the log-rejection probability has expectation $-\Omega(h\kappa)$, and apply sub-Gaussian concentration to show a product distribution has expectation $-\Omega(h\kappa d)$.

At this point, we reduce to the following self-contained problem: let $x \in \mathbb{R}$, let $\pi^* \propto \exp(-f_{1\mathrm{d}})$ be one-dimensional with second derivative $\leq \kappa$, and let $x_g = x + \sqrt{2h}g$ for $g \sim \mathcal{N}(0, 1)$. We wish to construct $f_{1\mathrm{d}}$ such that for x in a constant probability region over $\exp(-f_{1\mathrm{d}})$ (the "bad set"),

$$\mathbb{E}_{g\sim\mathcal{N}(0,1)}\left[-f_{1\mathrm{d}}(x_g) + f_{1\mathrm{d}}(x) - \frac{1}{2}\left\langle x - x_g, f'_{1\mathrm{d}}(x) + f'_{1\mathrm{d}}(x_g)\right\rangle\right] = -\Omega(h\kappa), \tag{4.1}$$

where the contents of the expectation in (4.1) are the log-rejection probability along one coordinate by a straightforward calculation. By forming a product distribution using $f_{1\mathrm{d}}$ as a building block, and combining with the remaining low-order terms due to the drift $\nabla f(x)$, we attain an $\exp(-d)$-sized region where the rejection probability is $\exp(-\Omega(h\kappa d))$, completing Theorem 7.

It remains to construct such a hard $f_{1\mathrm{d}}$. The calculation

$$-f_{1\mathrm{d}}(x_g) + f_{1\mathrm{d}}(x) - \frac{1}{2}\left\langle x - x_g, f'_{1\mathrm{d}}(x) + f'_{1\mathrm{d}}(x_g)\right\rangle = -2h\int_0^1 \left(\frac{1}{2} - s\right) g^2 f''_{1\mathrm{d}}(x + s(x_g - s))ds$$

suggests the following approach: because the above integral places more mass closer to the starting point, we wish to make sure our bad set has large second derivative, but most moves g result in a much smaller second derivative. Our construction patterns this intuition: we choose[4]

$$f_{1\mathrm{d}}(x) = \frac{\kappa}{3}x^2 - \frac{\kappa h}{3}\cos\frac{x}{\sqrt{h}} \implies f''_{1\mathrm{d}}(x) = \frac{2\kappa}{3} + \frac{\kappa}{3}\cos\frac{x}{\sqrt{h}},$$

such that our bad set is when $\cos\frac{x}{\sqrt{h}}$ is relatively large (which occurs with probability $\to \frac{1}{2}$ for small h in one dimension). The period of our construction scales with $\sqrt{h}$, so that most moves $\sqrt{2h}g$ of size $O(\sqrt{h})$ will "skip a period" and hence hit a region with small second derivative, satisfying (4.1).

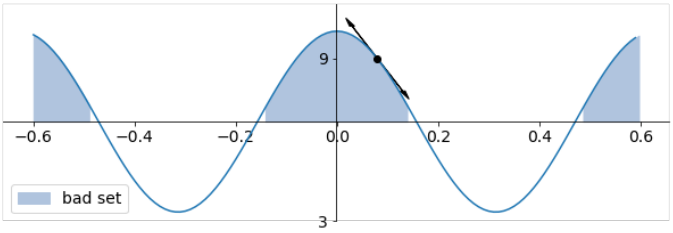


Figure 4.1: Second derivative of our hard function $f_{1\mathrm{d}}$, $\kappa = 10$, $h = 0.01$. Starting from inside the hard region, on average over $g \sim \mathcal{N}(0, \mathrm{id})$, a move by $\sqrt{2h}g$ decreases the second derivative.

HMC. We further demonstrate that similar hard Gaussians as the one we use for MALA also place an upper bound on the step size of HMC for any number of steps K. Our starting point is a novel characterization of HMC iterates on Gaussians: namely, when the negative log-density is quadratic, we show that the HMC iterates implement a linear combination between the starting position and velocity, where the coefficients are given by *Chebyshev polynomials*. For step size η of size $\Omega(\frac{1}{K\sqrt{\kappa}})$ for specific constants, we show the HMC chain begins to cycle because of the locations of the Chebyshev polynomials' zeroes, and cannot move. Moreover, for sufficiently small step size η outside of this range, it is straightforward by examining the coefficients of Chebyshev polynomials to show that they are the same (up to constant factors) as in the MALA case, at which point our previous lower bound holds. It takes some care to modify our hard Gaussian construction to rule out all constant ranges in the $\eta \approx \frac{1}{K\sqrt{\kappa}}$ region, but by doing so we obtain Theorem 9.

[4]We note [CLA+20] also used a (different, but similar) cosine-based construction for their lower bound.

We remark that the observation that HMC iterates are implicitly implementing a Chebyshev polynomial approximation appears to be unknown in the literature, and is a novel contribution of our work. We believe understanding this connection is a worthwhile endeavor, as a similar connection between polynomial approximation and first-order convex optimization has led to various interesting interpretations of Nesterov's accelerated gradient descent method [Har13, Bac19].

4.1.3 Prior work

The bounds most closely relevant to those in this paper are given by [LST20], who showed that the step size of MALA must scale inversely in κ for the chain to have a constant chance of moving, and [CLA$^+$20], who showed that the step size must scale as $d^{-\frac{1}{2}}$. Theorem 7 matches or improves both bounds simultaneously, proving that up to logarithmic factors the relaxation time of MALA scales *linearly* in both κ and d, while giving an explicit hard distribution and $\exp(-d)$-sized bad set. Moreover, both [LST20, CLA$^+$20] gave strictly spectral lower bounds, which are complemented by our Theorem 8, a mixing time lower bound.

We briefly mention several additional lower bound results in the sampling and sampling-adjacent literature, which are related to this work. Recently, [CLW20] exhibited an information-theoretic lower bound on unadjusted discretizations simulating the *underdamped Langevin dynamics*, whose dimension dependence matches the upper bound of [SL19] shown in Chapter 2 (while leaving the precise dependence on κ open). Finally, [GLL20] and [CBL20] give information-theoretic lower bounds for estimating normalizing constants of well-conditioned distributions and the number of stochastic gradient queries required by first-order sampling methods under noisy gradient access respectively.

4.2 Preliminaries

In Section 5.2.1, we give an overview of notation and technical definitions used throughout the paper. We state standard helper concentration bounds we frequently use in Section 4.2.2. We then recall the definitions of the sampling methods which we study in this paper in Sections 4.2.3 and 4.2.4.

4.2.1 Notation

General notation. For $d \in \mathbf{N}$ we let $[d] \stackrel{\text{def}}{=} \{i \in \mathbf{N} \mid 1 \le i \le d\}$. We let $\|\cdot\|_2$ denote the Euclidean norm on $\mathbb{R}^d$ for any d; for any positive semidefinite matrix $\mathbf{A}$, we let $\|\cdot\|_{\mathbf{A}}$ be its induced seminorm $\|x\|_{\mathbf{A}} = \sqrt{x^\top \mathbf{A} x}$. We use $\|\cdot\|_p$ to denote the ℓ_p norm for $p \ge 1$, and $\|\cdot\|_\infty$ is the maximum absolute value of entries. We let $\mathcal{N}(\mu, \mathbf{\Sigma})$ denote the multivariate Gaussian with mean $\mu \in \mathbb{R}^d$ and covariance $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$. We let $\mathrm{id} \in \mathbb{R}^{d \times d}$ denote the identity matrix when dimensions are clear from context, and $\preceq$ is the Loewner order on the positive semidefinite cone. We let $\{W_t\}_{t \ge 0} \subset \mathbb{R}^d$ denote the standard Brownian motion when dimensions are clear from context.

Functions. We say twice-differentiable $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and μ-strongly convex for $0 \le \mu \le L$ if $\mu \,\mathrm{id} \preceq \nabla^2 f(x) \preceq L \,\mathrm{id}$ for all $x \in \mathbb{R}^d$. It is well-known that for any $x, y \in \mathbb{R}^d$, this implies f has a Lipschitz gradient (i.e. $\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x - y\|_2$), and satisfies the quadratic bounds

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2 \le f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2.$$

We define the condition number of such a function f by $\kappa \stackrel{\text{def}}{=} \frac{L}{\mu}$. We will assume that κ is at least a constant for convenience of stating bounds; a lower bound of 10 suffices for all our results.

Distributions. For distribution π on $\mathbb{R}^d$, we say π is logconcave if $\frac{d\pi}{dx}(x) = \exp(-f(x))$ for convex f; we say π is μ-strongly logconcave if f is μ-strongly convex. For $A \subseteq \mathbb{R}^d$ we let A^c denote its complement and $\pi(A) \stackrel{\text{def}}{=} \int_{x \in A} d\pi(x)$ denote its measure under π. We say distribution ρ is β-warm with respect to π if $\frac{d\rho}{d\pi}(x) \le \beta$ everywhere; we define their total variation $\|\pi - \rho\|_{\mathrm{TV}} \stackrel{\text{def}}{=} \sup_{A \subseteq \mathbb{R}^d} \pi(A) - \rho(A)$. Finally, we denote the expectation and variance of $g : \mathbb{R}^d \to \mathbb{R}$ under π by

$$\mathbb{E}_\pi[g] = \int g(x) d\pi(x), \ \mathrm{Var}_\pi[g] = \mathbb{E}_\pi\left[g^2\right] - \left(\mathbb{E}_\pi[g]\right)^2.$$

Sampling. Consider a Markov chain defined on $\mathbb{R}^d$ with transition kernel $\{\mathcal{T}_x\}_{x \in \mathbb{R}^d}$, so that $\int \mathcal{T}_x(y) dy = 1$ for all x. Further, denote the stationary distribution of the Markov chain by π^*. Define the Dirichlet form of functions $g, h : \mathbb{R}^d \to \mathbb{R}$ with respect to the Markov chain by

$$\mathcal{E}(g, h) \stackrel{\text{def}}{=} \int g(x) h(x) d\pi^*(x) - \iint g(y) h(x) \mathcal{T}_x(y) d\pi^*(x) dy.$$

A standard calculation demonstrates that

$$\mathcal{E}(g,g) = \frac{1}{2}\iint (g(x)-g(y))^2 \mathcal{T}_x(y) d\pi^*(x) dy.$$

The mixing of the chain is governed by its spectral gap, a classical quantity we now define:

$$\lambda\left(\{\mathcal{T}_x\}_{x\in\mathbb{R}^d}\right) \stackrel{\text{def}}{=} \inf_g \left\{\frac{\mathcal{E}(g,g)}{\mathrm{Var}_{\pi^*}[g]}\right\}. \tag{4.2}$$

The relaxation time is the inverse spectral gap. We also recall a result of Cheeger [Che69], showing the spectral gap is $O(\Phi)$, where Φ is the conductance of the chain:

$$\Phi\left(\{\mathcal{T}_x\}_{x\in\mathbb{R}^d}\right) \stackrel{\text{def}}{=} \inf_{A\subset\mathbb{R}^d | \pi^*(A)\leq\frac{1}{2}} \frac{\int_{x\in A} \mathcal{T}_x(A^c) d\pi^*(x)}{\pi^*(A)} \tag{4.3}$$

Finally, we recall the definition of a Metropolis filter. A Markov chain with transitions $\{\mathcal{T}_x\}_{x\in\mathbb{R}^d}$ and stationary distribution π^* is said to be reversible if for all $x, y \in \mathbb{R}^d$,

$$d\pi^*(x)\mathcal{T}_x(y) = d\pi^*(y)\mathcal{T}_y(x).$$

The Metropolis filter is a way of taking an arbitrary set of proposal distributions $\{\mathcal{P}_x\}_{x\in\mathbb{R}^d}$ and defining a reversible Markov chain with stationary distribution π^*. In particular, the Markov chain induced by the Metropolis filter has transition distributions $\{\mathcal{T}_x\}_{x\in\mathbb{R}^d}$ defined by

$$\mathcal{T}_x(y) \stackrel{\text{def}}{=} \mathcal{P}_x(y)\min\left(1, \frac{d\pi^*(y)\mathcal{P}_y(x)}{d\pi^*(x)\mathcal{P}_x(y)}\right) \text{ for all } y \neq x. \tag{4.4}$$

Whenever the proposal is rejected by the modified distributions above, the chain does not move.

4.2.2 Concentration

Here we state several frequently used (standard) concentration facts.

Fact 1 (Mill's inequality)**.** *For one-dimensional Gaussian random variable $Z \sim \mathcal{N}(0,\sigma^2)$,*

$$\Pr[Z > t] \leq \sqrt{\frac{2}{\pi}}\frac{\sigma}{t}\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Fact 2 (χ^2 tail bounds, Lemma 1 [LM00])**.** *Let $\{Z_i\}_{i\in[n]} \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$ and $a \in \mathbb{R}^n_{\geq 0}$. Then*

$$\Pr\left[\sum_{i\in[n]} a_i Z_i^2 - \|a\|_2^2 \geq 2\|a\|_2\sqrt{t} + 2\|a\|_\infty t\right] \leq \exp(-t),$$

$$\Pr\left[\sum_{i\in[n]} a_i Z_i^2 - \|a\|_2^2 \leq -2\|a\|_2\sqrt{t}\right] \leq \exp(-t).$$

Fact 3 (Bernstein's inequality)**.** *Let $\{Z_i\}_{i\in[n]}$ be independent mean-zero random variables with sub-exponential parameter λ. Then*

$$\Pr\left[\left|\sum_{i\in[n]} Z_i\right| > t\right] \leq \exp\left(-\frac{1}{2}\min\left(\frac{t^2}{n\lambda^2}, \frac{t}{\lambda}\right)\right).$$

4.2.3 Metropolis-adjusted Langevin algorithm

In this section, we formally define the Metropolis-adjusted Langevin algorithm (MALA) which we study in Sections 4.3 and 4.4. Throughout this discussion, fix a distribution π on $\mathbb{R}^d$, with density $\frac{d\pi}{dx}(x) = \exp(-f(x))$, and suppose that f is twice-differentiable for simplicity.

The MALA Markov chain is given by a discretization of the (continuous-time) Langevin dynamics

$$dx_t = -\nabla f(x_t)dt + \sqrt{2}dW_t,$$

which is well-known to have stationary density $\exp(-f(x))$. MALA is defined by performing a simple Euler discretization of the Langevin dynamics up to time $h > 0$, and then applying a Metropolis filter. In particular, define the proposal distribution at a point x by

$$\mathcal{P}_x \stackrel{\text{def}}{=} \mathcal{N}\left(x - h\nabla f(x), 2h\,\mathrm{id}\right).$$

We obtain the MALA transition distribution by applying the definition (4.4), which yields

$$\mathcal{T}_x(y) \propto \exp\left(-\frac{\|y - (x - h\nabla f(x))\|_2^2}{4h}\right)\min\left(1, \frac{\exp\left(-f(y) - \frac{\|x-(y-h\nabla f(y))\|_2^2}{4h}\right)}{\exp\left(-f(x) - \frac{\|y-(x-h\nabla f(x))\|_2^2}{4h}\right)}\right). \tag{4.5}$$

The normalization constant above is that of the multivariate Gaussian with covariance $2h\,\mathrm{id}$.

4.2.4 Hamiltonian Monte Carlo

In this section, we formally define the (Metropolized) Hamiltonian Monte Carlo (HMC) method which we study in Section 4.6. We assume the same setting as Section 4.2.3.

The Metropolized HMC algorithm is governed by two parameters, a step size $\eta > 0$ and a step count $K \in \mathbf{N}$, and can be viewed as a multi-step generalization of MALA. In particular, when $K = 1$ it is straightforward to show that HMC is a reparameterization of MALA, see e.g. Appendix A of [LST20]. More generally, from an iterate x, HMC performs the following updates.

1. $x_0 \leftarrow x$, $v_0 \sim \mathcal{N}(0, \mathrm{id})$

2. For $0 \le k < K$:

   (a) $v_{k+\frac{1}{2}} \leftarrow v_k - \frac{\eta}{2}\nabla f(x_k)$

   (b) $x_{k+1} \leftarrow x_k + \eta v_{k+\frac{1}{2}}$

   (c) $v_{k+1} \leftarrow v_k - \frac{\eta}{2}\nabla f(x_{k+1})$

3. Return x_K

Each loop of step 2 is known in the literature as a "leapfrog" step, and is a discretization of Hamilton's equations for the Hamiltonian function $\mathcal{H}(x, v) \stackrel{\text{def}}{=} f(x) + \frac{1}{2}\|v\|_2^2$; for additional background, we refer the reader to [CDWY20]. This discretization is well-known to have reversible transition probabilities (i.e. the transition density is the same if the endpoints are swapped) because it satisfies a property known as *symplecticity*. Moreover, the Markov chain has stationary density on the expanded space $(x, v) \in \mathbb{R}^d \times \mathbb{R}^d$ proportional to $\exp(-\mathcal{H}(x, v))$. Correspondingly, the Metropolized HMC Markov chain performs the above algorithm from a point x, and accepts with probability

$$\min\left\{1, \frac{\exp\left(-\mathcal{H}(x_K, v_K)\right)}{\exp\left(-\mathcal{H}(x_0, v_0)\right)}\right\}. \tag{4.6}$$

4.3 Lower bound for MALA on Gaussians

In this section, we derive a upper bound on the spectral gap of MALA when the target distribution is restricted to being a multivariate Gaussian (i.e. its negative log-density is a quadratic in some well-conditioned matrix $\mathbf{A}$). Throughout this section we will let $f(x) = \frac{1}{2}x^\top \mathbf{A} x$ for some $\mathrm{id} \preceq \mathbf{A} \preceq \kappa\,\mathrm{id}$. We remark here that without loss of generality, we have assumed that the minimizer of f is the all-zeros vector and the strong convexity parameter is $\mu = 1$. These follow from invariance of condition number under linear translations and scalings of the variable.

Next, we define a specific hard quadratic function we will consider in this section, $f_{\mathrm{hq}} : \mathbb{R}^d \to \mathbb{R}$. Specifically, f_{hq} will be a quadratic in a diagonal matrix $\mathbf{A}$ which has $\mathbf{A}_{11} = 1$ and $\mathbf{A}_{ii} = \kappa$ for $2 \le i \le d$. We can rewrite this as

$$f_{\mathrm{hq}}(x) \stackrel{\text{def}}{=} \sum_{i \in [d]} f_i(x_i), \text{ where } f_i(c) = \begin{cases} \frac{1}{2}c^2 & i = 1 \\ \frac{\kappa}{2}c^2 & 2 \le i \le d \end{cases}. \tag{4.7}$$

Notice that f_{hq} is coordinate-wise separable, and behaves identically on coordinates $2 \leq i \leq d$ (and differently on coordinate 1). To this end for a vector $v \in \mathbb{R}^d$, we will denote its first coordinate by $v_1 \in \mathbb{R}$, and its remaining coordinates by $v_{-1} \in \mathbb{R}^{d-1}$. This will help us analyze the behavior of these components separately, and simplify notation.

We next show that for coordinate-separable functions with well-behaved first coordinate, such as our f_{hq}, the spectral gap (defined in (4.2)) of the MALA Markov chain is governed by the step size h. The following is an extension of an analogous proof in [CLA$^+$20].

Lemma 6. *Consider the MALA Markov chain* (4.5)*, with stationary distribution π^* with negative log-density f. Suppose f is coordinate-wise separable (i.e. $f(x) = \sum_{i\in[d]} f_i(x_i)$). If $f(x) = f(-x)$ for all $x \in \mathbb{R}^d$, f_1 is $O(1)$-smooth, and $\mathbb{E}_{x_1\sim\exp(-f_1)}[x_1^2] = \Theta(1)$, the spectral gap* (4.2) *is $O(h + h^2)$.*

Proof. Recalling the definition (4.2), we choose $g(x) = x_1$; note that by symmetry of f around the origin, we have $\mathbb{E}_{\pi^*}[g] = 0$, and thus by our assumption,

$$\mathrm{Var}_{\pi^*}[g] = \mathbb{E}_{x\sim\pi^*}[x_1^2] = \Theta(1).$$

Here we used that π^* is a product distribution. Thus it suffices to upper bound $\mathcal{E}(g, g)$:

$$\begin{aligned}
\mathcal{E}(g,g) &= \frac{1}{2}\iint (x_1 - y_1)^2 \mathcal{T}_x(y) d\pi^*(x) dy \\
&\leq \frac{1}{2}\iint (x_1 - y_1)^2 \mathcal{P}_x(y) d\pi^*(x) dy \\
&= \frac{1}{2}\mathbb{E}_{x\sim\pi^*,\xi\sim\mathcal{N}(0,1)}\left[\left(hf_1'(x_1) - \sqrt{2h}\xi\right)^2\right] \\
&\leq \mathbb{E}_{x\sim\pi^*}\left[h^2\left(f_1'(x_1)\right)^2\right] + 2\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[h\xi^2\right] \\
&\leq O(h^2)\mathbb{E}_{x\sim\pi^*}\left[x_1^2\right] + 2h = O\left(h + h^2\right).
\end{aligned}$$

In the second line, we used that whenever the Markov chain rejects the distribution both terms are zero; in the third, we used the definition of the MALA proposals; in the fourth, we used $(a+b)^2 \leq 2a^2 + 2b^2$ for $a, b \in \mathbb{R}$. Finally, the last line used that symmetry implies that the minimizer of f is the origin, so applying Lipschitzness and $f_1'(0) = 0$ yields the desired bound. □

This immediately implies a spectral gap bound on our hard function f_{hq}.

Corollary 3. *The spectral gap of the MALA Markov chain for sampling from the density proportional to* $\exp(-f_{\mathrm{hq}})$*, where f_{hq} is defined in* (4.7)*, is $O(h + h^2)$.*

It remains to give a lower bound on the step size h, which we accomplish by upper bounding the acceptance probability of MALA. We will give a step size analysis for a fairly general characterization of Markov chains, where the proposal distribution from a point x is

$$y = \begin{pmatrix} y_1 \\ y_{-1} \end{pmatrix}, \text{ where } y_1 = (1-\alpha_1)x_1 + \beta_1 g_1 \\ \text{and } y_{-1} = (1-\alpha_{-1})x_{-1} + \beta_{-1} g_{-1}, \text{ for } g \sim \mathcal{N}(0, \mathrm{id}). \tag{4.8}$$

To be concrete, recall that the proposal distribution for MALA (4.5) is given by $y = x - h\mathbf{A}x + \sqrt{2h}g$. For the $\mathbf{A}$ used in defining f_{hq}, this is of the form (4.8) with the specific parameters

$$\alpha_1 = h,\ \alpha_{-1} = h\kappa,\ \beta_1 = \beta_{-1} = \sqrt{2h}.$$

However, this more general characterization will save significant amounts of recalculation when analyzing updates of the HMC Markov chain in Section 4.6. Recalling the formula (4.5), we first give a closed form for the acceptance probability.

Lemma 7. *For $f(x) = \frac{1}{2}x^\top \mathbf{A}x$, we have*

$$f(x) - f(y) + \frac{1}{4h}\left(\|y - (x - h\nabla f(x))\|_2^2 - \|x - (y - h\nabla f(y))\|_2^2\right) = \frac{h}{4}\|x\|_{\mathbf{A}^2}^2 - \frac{h}{4}\|y\|_{\mathbf{A}^2}^2.$$

Supposing y is of the form in (4.8) *and $\mathbf{A}$ is as in* (4.7)*, we have*

$$\frac{h}{4}\|x\|_{\mathbf{A}^2}^2 - \frac{h}{4}\|y\|_{\mathbf{A}^2}^2 = \frac{h}{4}\left(\left(2\alpha_1 - \alpha_1^2\right)x_1^2 - \beta_1^2 g_1^2 - 2(1-\alpha_1)\beta_1 x_1 g_1\right) \\ + \frac{h\kappa^2}{4}\left(\left(2\alpha_{-1} - \alpha_{-1}^2\right)\|x_{-1}\|_2^2 - \beta_{-1}^2\|g_{-1}\|_2^2 - 2(1-\alpha_{-1})\beta_{-1}\langle x_{-1}, g_{-1}\rangle\right).$$

Proof. This is a direct computation which we perform here for completeness: the given quantity is

$$\frac{1}{2}\|x\|_{\mathbf{A}}^2 - \frac{1}{2}\|y\|_{\mathbf{A}}^2 + \frac{1}{4h}\left(\|y - x + h\mathbf{A}x\|_2^2 - \|x - y + h\mathbf{A}y\|_2^2\right) \\ = \frac{1}{2}\|x\|_{\mathbf{A}}^2 - \frac{1}{2}\|y\|_{\mathbf{A}}^2 + \frac{1}{2}\langle y - x, \mathbf{A}x\rangle + \frac{h}{4}\|x\|_{\mathbf{A}^2}^2 - \frac{1}{2}\langle x - y, \mathbf{A}y\rangle - \frac{h}{4}\|y\|_{\mathbf{A}^2}^2 = \frac{h}{4}\|x\|_{\mathbf{A}^2}^2 - \frac{h}{4}\|y\|_{\mathbf{A}^2}^2.$$

The second equality follows from expanding the definition of y:

$$\|x\|_{\mathbf{A}^2}^2 - \|y\|_{\mathbf{A}^2}^2 = x_1^2 - \left((1-\alpha_1)x_1 + \beta_1 g_1\right)^2 + \kappa^2\left(\|x_{-1}\|_2^2 - \|(1-\alpha_{-1})x_{-1} + \beta_{-1}g_{-1}\|_2^2\right) \\ = \left(2\alpha_1 - \alpha_1^2\right)x_1^2 - \beta_1^2 g_1^2 - 2(1-\alpha_1)\beta_1 x_1 g_1 \\ + \kappa^2\left(\left(2\alpha_{-1} - \alpha_{-1}^2\right)\|x_{-1}\|_2^2 - \beta_{-1}^2\|g_{-1}\|_2^2 - 2(1-\alpha_{-1})\beta_{-1}\langle x_{-1}, g_{-1}\rangle\right).$$

□

Corollary 4. *For any fixed $x \in \mathbb{R}^d$, and supposing y is of the form in* (4.8) *and* $\mathbf{A}$ *is as in* (4.7)*,*

$$\mathbb{E}_{g\sim\mathcal{N}(0,\mathrm{id})}\left[f(x) - f(y) + \frac{1}{4h}\left(\|y - (x - h\nabla f(x))\|_2^2 - \|x - (y - h\nabla f(y))\|_2^2\right)\right]$$
$$= \frac{h}{4}\left(\left(2\alpha_1 - \alpha_1^2\right)x_1^2 - \beta_1^2\right) + \frac{h\kappa^2}{4}\left(\left(2\alpha_{-1} - \alpha_{-1}^2\right)\|x_{-1}\|_2^2 - \beta_{-1}^2(d-1)\right).$$

Proof. This follows from Lemma 7, independence of g and x, and linearity of trace and expectation applied on squared coordinates of g, where we recognize $\mathbb{E}_{g\sim\mathcal{N}(0,\mathrm{id})}[gg^\top] = \mathrm{id}$. □

Next, for a fixed x, consider the random variables R_i^x:

$$R_i^x = \begin{cases} \frac{h}{4}\left(\left(2\alpha_1 - \alpha_1^2\right)x_1^2 - \beta_1^2 g_1^2 - 2(1-\alpha_1)\beta_1 x_1 g_1\right) & i = 1 \\ \frac{h\kappa^2}{4}\left(\left(2\alpha_{-1} - \alpha_{-1}^2\right)x_i^2 - \beta_{-1}^2 g_i^2 - 2(1-\alpha_{-1})\beta_{-1} x_i g_i\right) & 2 \le i \le d \end{cases}$$

where $g \sim \mathcal{N}(0, \mathrm{id})$ is a standard Gaussian random vector. Notice that for a given realization of g, we have by Lemma 7 that

$$\sum_{i\in[d]} R_i^x = \frac{h}{4}\|x\|_{\mathbf{A}^2} - \frac{h}{4}\|y\|_{\mathbf{A}^2}. \tag{4.9}$$

We computed the expectation of $\sum_{i\in[d]} R_i^x$ in Corollary 4. We next give a high-probability bound on the deviation of $\sum_{i\in[d]} R_i^x$ from its expectation.

Lemma 8. *With probability at least $1-\delta$ over the randomness of $g \sim \mathcal{N}(0, \mathrm{id})$,*

$$\sum_{i\in[d]} R_i^x - \mathbb{E}_{g\sim\mathcal{N}(0,\mathrm{id})}\left[\sum_{i\in[d]} R_i^x\right] \le 2h|\alpha_1 - 1||\beta_1||x_1|\sqrt{\log\left(\frac{4}{\delta}\right)} + h\beta_1^2\sqrt{\log\left(\frac{4}{\delta}\right)}$$
$$+ 2h\kappa^2|\alpha_{-1} - 1||\beta_{-1}|\|x_{-1}\|_2\sqrt{\log\left(\frac{4}{\delta}\right)} + h\kappa^2\beta_{-1}^2\sqrt{d\log\left(\frac{4}{\delta}\right)}.$$

Proof. In defining $\{R_i^x\}_{i\in[d]}$, the terms involving $\{x_i^2\}_{i\in[d]}$ are deterministic. Thus, we need to upper bound the deviations of the remaining terms, namely

$$S_1^{(1)} \stackrel{\mathrm{def}}{=} \frac{h}{2}(\alpha_1 - 1)\beta_1 x_1 g_1,\ S_1^{(2)} \stackrel{\mathrm{def}}{=} \frac{h\beta_1^2}{4}\left(1 - g_1^2\right),$$
$$S_{-1}^{(1)} \stackrel{\mathrm{def}}{=} \frac{h\kappa^2}{2}(\alpha_{-1} - 1)\beta_{-1}\sum_{2\le i\le d} x_i g_i,\ S_{-1}^{(2)} \stackrel{\mathrm{def}}{=} \frac{h\kappa^2\beta_{-1}^2}{4}\sum_{2\le i\le d}\left(1 - g_i^2\right).$$

To motivate these definitions, $S_1^{(1)} + S_1^{(2)} + S_{-1}^{(1)} + S_{-1}^{(2)}$ is the left hand side of the display in the lemma statement. We begin with $S_{-1}^{(1)}$. Notice that this is a one-dimensional Gaussian random variable distributed as

$$\mathcal{N}\left(0, \sigma_1^2\right) \text{ where } \sigma_1 \stackrel{\mathrm{def}}{=} \frac{h\kappa^2}{2}|\alpha_{-1} - 1||\beta_{-1}|\|x_{-1}\|_2.$$

Thus, applying Mill's inequality yields

$$\Pr\left[S_{-1}^{(1)} > t\right] \le \sqrt{\frac{2}{\pi}}\frac{\sigma_1}{t}\exp\left(-\frac{t^2}{2\sigma_1^2}\right) \le \frac{\delta}{4}, \text{ for } t = 4\sigma_1\sqrt{\log\left(\frac{4}{\delta}\right)}.$$

Next, to bound the term $S_{-1}^{(2)}$, define

$$\sigma_2 \stackrel{\text{def}}{=} \frac{h\kappa^2\beta_{-1}^2}{4}\sqrt{d-1}.$$

Standard χ^2 concentration results (Fact 2) then yield

$$\Pr\left[S_{-1}^{(2)} > t\right] \le \exp\left(-\frac{t^2}{4\sigma_2^2}\right) \le \frac{\delta}{4}, \text{ for } t = 2\sigma_2\sqrt{\log\left(\frac{4}{\delta}\right)}.$$

Similar bounds follow for $S_1^{(1)}$ and $S_1^{(2)}$, whose computations we omit for brevity. Taking a union bound over these four terms yields the desired claim. □

Finally, we have a complete characterization of a bad set $\Omega \subset \mathbb{R}^d$ where, with high probability over the proposal distribution, the acceptance probability is extremely small.

Proposition 2. *Let $x \in \mathbb{R}^d$ satisfy $\|x_{-1}\|_2 \le \sqrt{\frac{2d}{3\kappa}}$ and $|x_1| \le 5\sqrt{\log d}$, and suppose y is of the form in* (4.8) *and* $\mathbf{A}$ *is as in* (4.7)*. Also suppose that*

$$|\alpha_{-1}| \le \frac{3}{5}\beta_{-1}^2\kappa,\ \beta_{-1} = \omega\left(\sqrt{\frac{\log d}{\kappa d}}\right),\ |\alpha_1| = O(|\alpha_{-1}|),\ \beta_1 = O(\beta_{-1}).$$

Then with probability at least $1 - d^{-5}$ over the randomness of $g \sim \mathcal{N}(0, \mathrm{id})$, we have

$$\frac{h}{4}\|x\|_{\mathbf{A}^2} - \frac{h}{4}\|y\|_{\mathbf{A}^2} = -\Omega\left(h\kappa^2\beta_{-1}^2 d\right).$$

Proof. We first handle terms involving x_{-1} and g_{-1}. Combining (4.9), Corollary 4, and Lemma 8, we have with probability at least $1 - \frac{1}{2}d^{-5}$ over the randomness of $g \sim \mathcal{N}(0, \mathrm{id})$ that $\|x_{-1}\|_{\mathbf{A}_{-1}^2}^2 - \|y_{-1}\|_{\mathbf{A}_{-1}^2}^2$ (where $\mathbf{A}_{-1}$ is the Hessian of f_{hq} on the last $d-1$ coordinates) is upper bounded by

$$\begin{aligned}\frac{h\kappa^2}{4}\left(\left(2\alpha_{-1} - \alpha_{-1}^2\right)\|x_{-1}\|_2^2 - \beta_{-1}^2(d-1)\right) + 5h\kappa^2|\alpha_{-1}-1|\beta_{-1}\|x_{-1}\|_2\sqrt{\log d} + 3h\kappa^2\beta_{-1}^2\sqrt{d\log d}\\ \le -\frac{h\kappa^2}{4.5}\beta_{-1}^2 d + \frac{h\kappa^2}{4}(2\alpha_{-1} - \alpha_{-1}^2)\|x_{-1}\|_2^2 + 5h\kappa^2|\alpha_{-1}-1|\beta_{-1}\|x_{-1}\|_2\sqrt{\log d}.\end{aligned} \tag{4.10}$$

Here we dropped the last term in the first line by adjusting a constant since it is dominated for sufficiently large d. It remains to show that all the terms in the second line other than $-\frac{h\kappa^2}{4.5}\beta_{-1}^2 d$ are bounded by $O(h\kappa^2\beta_{-1}^2 d)$. We will perform casework on the size of α_{-1}.

Case 1: $|\alpha_{-1}| > 3$. In this case, we have for sufficiently large d, by Young's inequality

$$\begin{aligned} 5h\kappa^2|\alpha_{-1}-1|\beta_{-1}\|x_{-1}\|_2\sqrt{\log d} &\le \frac{1}{40}h\kappa^2|\alpha_{-1}|\beta_{-1}\|x_{-1}\|_2\sqrt{d} \\ &\le \frac{1}{80}h\kappa^2\beta_{-1}^2 d + \frac{1}{80}h\kappa^2\alpha_{-1}^2\|x_{-1}\|_2^2. \end{aligned}$$

Plugging this bound into (4.10), we have the desired

$$\|x_{-1}\|_{\mathbf{A}_{-1}^2}^2 - \|y_{-1}\|_{\mathbf{A}_{-1}^2}^2 \le -\frac{h\kappa^2}{5}\beta_{-1}^2 d + \frac{h\kappa^2}{4}\left(2\alpha_{-1} - 0.9\alpha_{-1}^2\right)\|x_{-1}\|_2^2 \le -\frac{h\kappa^2}{5}\beta_{-1}^2 d.$$

In the last inequality we used $2\alpha_{-1} - 0.9\alpha_{-1}^2 \le 0$ for $|\alpha_{-1}| > 3$.

Case 2: $|\alpha_{-1}| \le 3$. In this case, we first observe by our assumed bounds on $\|x_{-1}\|_2$ and β_{-1},

$$5h\kappa^2|\alpha_{-1}-1|\beta_{-1}\|x_{-1}\|_2\sqrt{\log d} \le 20h\kappa^{1.5}\beta_{-1}\sqrt{d\log d} = o\left(h\kappa^2\beta_{-1}^2 d\right).$$

Thus, substituting into (4.10) and dropping the (nonpositive) term corresponding to α_{-1}^2,

$$\begin{aligned} \|x_{-1}\|_{\mathbf{A}_{-1}^2}^2 - \|y_{-1}\|_{\mathbf{A}_{-1}^2}^2 &\le -\frac{h\kappa^2}{4.8}\beta_{-1}^2 d + \frac{h\kappa^2}{2}\alpha_{-1}\|x_{-1}\|_2^2 \\ &\le -\frac{h\kappa^2}{4.8}\beta_{-1}^2 d + \frac{h\kappa\alpha_{-1}d}{3} = -\Omega\left(h\kappa^2\beta_{-1}^2 d\right). \end{aligned}$$

In the second inequality, we used the assumed bound on $\|x_{-1}\|_2^2$, and in the last we used the bound $|\alpha_{-1}| \le \frac{3}{5}\beta_{-1}^2\kappa$ to reach the conclusion.

To complete the proof we need to show terms involving x_1 and g_1 are small. In particular, combining (4.9), Corollary 4, and Lemma 8 and dropping nonnegative terms, it suffices to argue

$$\frac{h}{2}\alpha_1 x_1^2 + 5h|\alpha_1 - 1|\beta_1|x_1|\sqrt{\log d} + 3h\beta_1^2\sqrt{\log d} = o\left(h\kappa^2\beta_{-1}^2 d\right).$$

This bound clearly holds for the last term $h\beta_1^2\sqrt{\log d}$ using $\beta_1 = O(\beta_{-1})$. For the first term, it suffices to use our assumed bounds on $|\alpha_1|$ and x_1. Finally, the middle term $5h|\alpha_1 - 1|\beta_1|x_1|\sqrt{\log d}$ is low-order compared to the term $5h\kappa^2|\alpha_{-1}-1|\beta_{-1}\|x_{-1}\|_2\sqrt{\log d}$ which we argued about earlier, and hence does not affect any of our earlier bounds by more than a constant. The left-hand side of the above display is an upper bound of the first coordinate's contribution with probability at least $1 - \frac{1}{2}d^{-5}$, so a union bound shows the proof succeeds with probability $\ge 1 - d^{-5}$. □

Finally, we are ready to give the main lower bound of this section.

Theorem 6. *For every step size, there is a target Gaussian on $\mathbb{R}^d$ whose negative log-density always has Hessian eigenvalues in $[1, \kappa]$, such that the relaxation time of MALA is $\Omega(\frac{\kappa\sqrt{d}}{\sqrt{\log d}})$.*

Proof. Let π^* be the Gaussian with log-density $-f_{\text{hq}}$ (4.7) throughout this proof. If $h = O\left(\frac{\sqrt{\log d}}{\kappa\sqrt{d}}\right)$, then Corollary 3 immediately implies the result, so for the remainder of the proof suppose

$$h = \omega\left(\frac{\sqrt{\log d}}{\kappa\sqrt{d}}\right). \tag{4.11}$$

We first recall that MALA Markov chains are an instance of (4.8) with

$$\alpha_1 = h,\ \alpha_{-1} = h\kappa,\ \beta_1 = \beta_{-1} = \sqrt{2h}.$$

It is easy to see that these parameters satisfy the assumptions in Proposition 2, for the given range of h. We define a "bad starting set" as follows:

$$\Omega \stackrel{\text{def}}{=} \left\{ x \;\middle|\; \|x_{-1}\|_2^2 \le \frac{2d}{3\kappa},\ x_1^2 \le 25 \log d \right\}. \tag{4.12}$$

For any $x \in \Omega$, and h satisfying (4.11), Proposition 2 is applicable, and by our definition of Ω, any $x \in \Omega$ has proposals which will be accepted with probability

$$\exp\left(-\Omega\left(h\kappa^2\beta_{-1}^2 d\right)\right) = \exp\left(-\Omega(h^2\kappa^2 d)\right) \le \frac{1}{d^{10}}.$$

The conductance of the Markov chain (4.3) is then at most $\frac{2}{d^5}$ by the witness set Ω and the failure probability of Proposition 2, which concludes the proof by Cheeger's inequality [Che69], where we use the assumption that $\kappa = O(d^4)$. □

Finally, as it clarifies the required warmness to escape the bad set in the proof of Theorem 6 (and is used in our mixing time bounds in Section 4.5), we lower bound the measure of Ω according to π^*. Applying Lemma 9 shows with probability at least $\exp(-\frac{1}{12}d)$, $\|x_{-1}\|_2^2 \le \frac{d}{2\kappa}$, and Fact 1 shows that $x_1^2 \le 25 \log d$ with probability at least $\frac{1}{2}$; combining shows that the measure is at least $\exp(-d)$. We required one helper technical fact, a small-ball probability bound for Gaussians.

Lemma 9. *Let $v \sim \mathcal{N}(0, \text{id})$ be a random Gaussian vector in n dimensions. For large enough n,*

$$\Pr\left[\|v\|_2^2 \le \frac{n}{2}\right] \ge \exp\left(-\frac{n}{12}\right).$$

Proof. Observe that $\|v\|_2^2$ follows a χ^2 distribution with n degrees of freedom. Thus this probability is governed by the χ^2 cumulative density function, and is

$$\frac{1}{\Gamma(k)}\gamma\left(k, ck\right)$$

where we define $k \stackrel{\text{def}}{=} \frac{n}{2}$ and $c \stackrel{\text{def}}{=} \frac{1}{2}$; here Γ is the standard gamma function, and γ is the lower incomplete gamma function. Next, we have the bound from [ODL+20]

$$\frac{1}{\Gamma(k)}\gamma\left(k, ck\right) \geq \left(1 - \exp\left(-\ell ck\right)\right)^k, \ \ell \stackrel{\text{def}}{=} \left(\Gamma(k+1)\right)^{-\frac{1}{k-1}}.$$

A direct calculation yields $\ell \geq \frac{2.5}{k} \implies 1 - \exp(-\ell ck) \geq \exp(-\frac{1}{6})$ for large enough k. Recalling we defined $k = \frac{n}{2}$ yields the conclusion. □

4.4 Lower bound for MALA on well-conditioned distributions

In this section, we derive a lower bound on the relaxation time of MALA for sampling from a distribution with density proportional to $\exp(-f(x))$, where $f : \mathbb{R}^d \to \mathbb{R}$ is a (non-quadratic) target function with condition number κ. In particular, by exploiting the structure of non-cancellations which do not occur for quadratics, we will attain a stronger lower bound.

Our first step is to derive an upper bound on the acceptance probability for a general target function f according to the MALA updates (4.5), analogously to Lemma 7 in the Gaussian case.

Lemma 10. *For any function $f : \mathbb{R}^d \to \mathbb{R}$, we have*

$$\begin{aligned}
&f(x) - f(y) + \frac{1}{4h}\left(\|y - (x - h\nabla f(x))\|_2^2 - \|x - (y - h\nabla f(y))\|_2^2\right) \\
&= -f(y) + f(x) - \frac{1}{2}\left\langle x - y, \nabla f(x) + \nabla f(y)\right\rangle + \frac{h}{4}\|\nabla f(x)\|_2^2 - \frac{h}{4}\|\nabla f(y)\|_2^2.
\end{aligned}$$

Proof. This is a direct computation which we perform here for completeness:

$$\begin{aligned}
&f(x) - f(y) + \frac{1}{4h}\left(\|y - (x - h\nabla f(x))\|_2^2 - \|x - (y - h\nabla f(y))\|_2^2\right) \\
&= f(x) - f(y) + \frac{1}{2h}\left\langle y - x, h\nabla f(x)\right\rangle - \frac{1}{2h}\left\langle x - y, h\nabla f(y)\right\rangle + \frac{h}{4}\|\nabla f(x)\|_2^2 - \frac{h}{4}\|\nabla f(y)\|_2^2 \\
&= -f(y) + f(x) - \frac{1}{2}\left\langle x - y, \nabla f(x) + \nabla f(y)\right\rangle + \frac{h}{4}\|\nabla f(x)\|_2^2 - \frac{h}{4}\|\nabla f(y)\|_2^2.
\end{aligned}$$

□

Next, recall the proposal distribution of the MALA updates (4.5) sets $y = x - h\nabla f(x) + \sqrt{2h}g$ where $g \sim \mathcal{N}(0, \mathrm{id})$. We further split this update into a random step and a deterministic step, by defining

$$x_g \stackrel{\text{def}}{=} x + \sqrt{2h}g, \text{ where } g \sim \mathcal{N}(0, \mathrm{id}) \text{ and } y = x_g - h\nabla f(x). \tag{4.13}$$

This will allow us to reason about the effects of the stochastic and drift terms separately. We crucially will use the following decomposition of the equation in Lemma 10:

$$\begin{aligned} -f(y) + f(x) - \frac{1}{2}\langle x - y, \nabla f(x) + \nabla f(y)\rangle + \frac{h}{4}\|\nabla f(x)\|_2^2 - \frac{h}{4}\|\nabla f(y)\|_2^2 \\ = -f(x_g) + f(x) - \frac{1}{2}\langle x - x_g, \nabla f(x) + \nabla f(x_g)\rangle \\ + f(x_g) - f(y) - \frac{1}{2}\langle x - x_g, \nabla f(y) - \nabla f(x_g)\rangle \\ - \frac{1}{2}\langle x_g - y, \nabla f(x) + \nabla f(y)\rangle + \frac{h}{4}\|\nabla f(x)\|_2^2 - \frac{h}{4}\|\nabla f(y)\|_2^2. \end{aligned} \tag{4.14}$$

We will use the following observation, which gives an alternate characterization of the second line of (4.14), as well as a bound on the third and fourth lines for smooth functions.

Lemma 11. *For twice-differentiable $f : \mathbb{R}^d \to \mathbb{R}$, letting $x_s \stackrel{\text{def}}{=} x + s(x_g - x)$ for $s \in [0, 1]$, we have*

$$-f(x_g) + f(x) - \frac{1}{2}\langle x - x_g, \nabla f(x) + \nabla f(x_g)\rangle = -2h\int_0^1 \left(\frac{1}{2} - s\right) g^\top \nabla^2 f(x_s) g ds.$$

Moreover, assuming f is κ-smooth,

$$\begin{aligned} f(x_g) - f(y) - \frac{1}{2}\langle x - x_g, \nabla f(y) - \nabla f(x_g)\rangle \\ - \frac{1}{2}\langle x_g - y, \nabla f(x) + \nabla f(y)\rangle + \frac{h}{4}\|\nabla f(x)\|_2^2 + \frac{h}{4}\|\nabla f(y)\|_2^2 \\ \le 2\left(h^2\kappa + h^3\kappa^2\right)\|\nabla f(x)\|_2^2 + 3\left(h^{1.5}\kappa + h^{2.5}\kappa^2\right)\|g\|_2\|\nabla f(x)\|_2 + h^2\kappa^2\|g\|_2^2. \end{aligned}$$

Proof. By integrating twice and using the definition $x_g = x + \sqrt{2h}g$,

$$\begin{aligned} f(x_g) &= f(x) + \int_0^1 \langle \nabla f(x_s), x_g - x\rangle\, ds \\ &= f(x) + \langle \nabla f(x), x_g - x\rangle + \int_0^1 \left\langle \int_0^s \nabla^2 f(x_t)\,(x_g - x)\, dt, x_g - x \right\rangle ds \\ &= f(x) + \langle \nabla f(x), x_g - x\rangle + 2h\int_0^1 (1 - s)\, g^\top \nabla^2 f(x_s) g ds. \end{aligned} \tag{4.15}$$

Similarly,

$$f(x) = f(x_g) + \langle \nabla f(x_g), x - x_g\rangle + 2h\int_0^1 s g^\top \nabla^2 f(x_s) g ds. \tag{4.16}$$

The first conclusion follows from combining (4.15) and (4.16). Next, assuming f is κ-smooth,

$$\begin{aligned}
&f(x_g) - f(y) - \frac{1}{2}\langle x - x_g, \nabla f(y) - \nabla f(x_g)\rangle - \frac{1}{2}\langle x_g - y, \nabla f(x) + \nabla f(y)\rangle \\
&= f(x_g) - f(y) + \frac{\sqrt{2h}}{2}\langle g, \nabla f(y) - \nabla f(x_g)\rangle - \langle x_g - y, \nabla f(y)\rangle - \frac{h}{2}\langle \nabla f(x), \nabla f(x) - \nabla f(y)\rangle \\
&\le f(x_g) - f(y) - \langle x_g - y, \nabla f(y)\rangle + \frac{\sqrt{2h}}{2}\|g\|_2\|\nabla f(x_g) - \nabla f(y)\|_2 + \frac{h}{2}\|\nabla f(x)\|_2\|\nabla f(x) - \nabla f(y)\|_2 \\
&\le \frac{\kappa}{2}\|x_g - y\|_2^2 + \frac{\sqrt{2h}\kappa}{2}\|g\|_2\|x_g - y\|_2 + \frac{h\kappa}{2}\|\nabla f(x)\|_2\|x - y\|_2 \\
&\le \frac{h^2\kappa}{2}\|\nabla f(x)\|_2^2 + \frac{\sqrt{2}}{2}h^{1.5}\kappa\|g\|_2\|\nabla f(x)\|_2 + \frac{h\kappa}{2}\|\nabla f(x)\|_2\left(\sqrt{2h}\|g\|_2 + h\|\nabla f(x)\|_2\right) \\
&= h^2\kappa\|\nabla f(x)\|_2^2 + \sqrt{2}h^{1.5}\kappa\|g\|_2\|\nabla f(x)\|_2.
\end{aligned} \tag{4.17}$$

The second line used the definitions of x_g and y in (4.13), and the third used Cauchy-Schwarz. The fourth used smoothness (which implies gradient Lipschitzness), and the fifth again used (4.13) and the triangle inequality. Next, we bound the remaining terms $\frac{h}{4}\|\nabla f(x)\|_2^2 - \frac{h}{4}\|\nabla f(y)\|_2^2$:

$$\begin{aligned}
&\frac{h}{4}\|\nabla f(x)\|_2^2 - \frac{h}{4}\|\nabla f(y)\|_2^2 = \frac{h}{4}\langle \nabla f(x) + \nabla f(y), \nabla f(x) - \nabla f(y)\rangle \\
&\le \frac{h\kappa}{4}\left(2\|\nabla f(x)\|_2 + \kappa\|x - y\|_2\right)\|x - y\|_2 \\
&\le \frac{h\kappa}{4}\left(2\|\nabla f(x)\|_2 + h\kappa\|\nabla f(x)\|_2 + \sqrt{2h}\kappa\|g\|_2\right)\left(h\|\nabla f(x)\|_2 + \sqrt{2h}\|g\|_2\right) \\
&\le \frac{1}{2}\left(h^2\kappa + h^3\kappa^2\right)\|\nabla f(x)\|_2^2 + \frac{\sqrt{2}}{2}\left(h^{1.5}\kappa + h^{2.5}\kappa^2\right)\|g\|_2\|\nabla f(x)\|_2 + h^2\kappa^2\|g\|_2^2.
\end{aligned} \tag{4.18}$$

Combining (4.17) and (4.18) yields the conclusion. □

We will ultimately use the second bound in Lemma 11 to argue that the third and fourth lines in (4.14) are low-order, so it remains to concentrate on the remaining term,

$$-f(x_g) + f(x) - \frac{1}{2}\langle x - x_g, \nabla f(x) + \nabla f(x_g)\rangle = -2h\int_0^1 \left(\frac{1}{2} - s\right) g^\top \nabla^2 f(x_s) g ds. \tag{4.19}$$

Our goal is to demonstrate this term is $-\Omega(h\kappa d)$ over an inverse-exponentially sized region, for a particular hard distribution. As it is coordinate-wise separable, our proof strategy will be to construct a hard one-dimensional function, and replicate it to obtain a linear dependence on d.

We now define the specific hard function $f_{\text{hard}} : \mathbb{R}^d \to \mathbb{R}$ we work with for the remainder of the section; it is straightforward to see f_{hard} is κ-smooth and 1-strongly convex.

$$f_{\text{hard}}(x) \stackrel{\text{def}}{=} \sum_{i\in[d]} f_i(x_i), \text{ where } f_i(c) = \begin{cases} \frac{1}{2}c^2 & i = 1 \\ \frac{\kappa}{3}c^2 - \frac{\kappa h}{3}\cos\frac{c}{\sqrt{h}} & 2 \le i \le d \end{cases}. \tag{4.20}$$

We will now show that sampling from the distribution with density proportional to $\exp(-f_{\text{hard}})$ is hard. First, notice that the function f_{hard} has condition number κ and is coordinate-wise separable. It immediately follows from Lemma 6 that the spectral gap (defined in (4.2)) of the MALA Markov chain is governed by the step size h as follows.

Corollary 5. *The spectral gap of the MALA Markov chain for sampling from the density proportional to* $\exp(-f_{\text{hard}})$*, where* f_{hard} *is defined in* (4.20)*, is* $O(h + h^2)$*.*

For the remainder of the section, we focus on upper bounding (4.19) over a large region according to the density proportional to $\exp(-f_{\text{hard}})$. Recall $\{f_i\}_{i\in[d]}$ are the summands of f_{hard}. For a fixed x, consider the random variables S_i^x:

$$S_i^x = -f_i([x_g]_i) + f_i(x_i) - \frac{1}{2}(x_i - [x_g]_i)(f_i'(x_i) + f_i'([x_g]_i)).$$

It is easy to check that for a given realization of g, we have

$$\sum_{i\in[d]} S_i^x = -f(x_g) + f(x) - \frac{1}{2}\left\langle x - x_g, \nabla f(x) + \nabla f(x_g)\right\rangle,$$

where the right-hand side of the above display is the left-hand side of (4.19). We bound the expectation of $\sum_{i\in[d]} S_i^x$, and its deviation from its expectation, in Lemma 12 and Lemma 13 respectively.

Lemma 12. *For any fixed* $x \in \left\{x \;\middle|\; -\frac{1}{2}\pi\sqrt{h} + 2\pi k_i\sqrt{h} \le x_i \le \frac{1}{2}\pi\sqrt{h} + 2\pi k_i\sqrt{h}, k_i \in \mathbb{N}, \forall 2 \le i \le d\right\}$ *and* $h \le 1$*, the random variables* S_i^x*,* $1 \le i \le d$ *satisfy*

$$\mathbb{E}_{g\sim\mathcal{N}(0,\text{id})}\left[S_i^x\right] \le \begin{cases} 0 & i = 1 \\ -0.08h\kappa\cos\frac{x_i}{\sqrt{h}} & 2 \le i \le d \end{cases}.$$

Proof. We remark that the condition on x simply enforces coordinatewise in $2 \le i \le d$, $\cos\frac{x_i}{\sqrt{h}} > 0$. Consider some coordinate $2 \le i \le d$: since $[x_g]_i = x_i + \sqrt{2h}g_i$,

$$\begin{aligned} S_i^x &= -f_i\left(x_i + \sqrt{2h}g_i\right) + f_i(x_i) + \frac{\sqrt{2h}}{2}g_i\left(f_i'(x_i) + f_i'\left(x_i + \sqrt{2h}g_i\right)\right) \\ &= -\frac{\kappa}{3}\left(x_i + \sqrt{2h}g_i\right)^2 + \frac{\kappa h}{3}\cos\left(\frac{x_i}{\sqrt{h}} + \sqrt{2}g_i\right) + \frac{\kappa}{3}x_i^2 - \frac{\kappa h}{3}\cos\left(\frac{x_i}{\sqrt{h}}\right) \\ &\quad + \frac{\sqrt{2h}}{2}g_i\left(\frac{4\kappa}{3}x_i + \frac{2\sqrt{2h}\kappa}{3}g_i + \frac{\kappa\sqrt{h}}{3}\sin\left(\frac{x_i}{\sqrt{h}} + \sqrt{2}g_i\right) + \frac{\kappa\sqrt{h}}{3}\sin\left(\frac{x_i}{\sqrt{h}}\right)\right) \\ &= \frac{\kappa h}{3}\left(\cos\left(\frac{x_i}{\sqrt{h}} + \sqrt{2}g_i\right) - \cos\left(\frac{x_i}{\sqrt{h}}\right)\right) + \frac{\sqrt{2}h\kappa}{6}g_i\left(\sin\left(\frac{x_i}{\sqrt{h}} + \sqrt{2}g_i\right) + \sin\left(\frac{x_i}{\sqrt{h}}\right)\right) \end{aligned}$$

Here, we used that the quadratic terms in the second and third lines cancel (this also follows from examining the proof of Lemma 7):

$$-\frac{\kappa}{3}\left(x_i+\sqrt{2h}g_i\right)^2+\frac{\kappa}{3}x_i^2+\frac{\sqrt{2h}}{2}g_i\left(\frac{4\kappa}{3}x_i+\frac{2\sqrt{2h}\kappa}{3}g_i\right)=0.$$

By a direct computation, taking an expectation over $g_i\sim\mathcal{N}(0,1)$ yields

$$\mathbb{E}_{g_i\sim\mathcal{N}(0,1)}\left[\cos\left(\frac{x_i}{\sqrt{h}}+\sqrt{2}g_i\right)\right]=\frac{\cos\left(\frac{x_i}{\sqrt{h}}\right)}{\exp\left(1\right)},$$
$$\mathbb{E}_{g_i\sim\mathcal{N}(0,1)}\left[\sin\left(\frac{x_i}{\sqrt{h}}+\sqrt{2}g_i\right)g_i\right]=\frac{\sqrt{2}\cos\left(\frac{x_i}{\sqrt{h}}\right)}{\exp\left(1\right)}.$$

Putting these pieces together,

$$\mathbb{E}_{g_i\sim\mathcal{N}(0,1)}\left[S_i^x\right]=\frac{\kappa h}{3}\left(\frac{2}{\exp(1)}-1\right)\cos\left(\frac{x_i}{\sqrt{h}}\right)\ \ \leq -0.08\kappa h\cos\left(\frac{x_i}{\sqrt{h}}\right).$$

Here, we used $\cos\frac{x_i}{\sqrt{h}}>0$. For $i=1$, Lemma 7 shows $\mathbb{E}_{g_1\sim\mathcal{N}(0,1)}\left[S_1^x\right]=0$. □

Lemma 13. *With probability at least $1-\frac{1}{d^5}$ over the randomness of $g\sim\mathcal{N}(0,\mathrm{id})$,*

$$\sum_{i\in[d]}S_i^x-\mathbb{E}_{g\sim\mathcal{N}(0,\mathrm{id})}\left[\sum_{i\in[d]}S_i^x\right]\leq 10h\kappa\sqrt{d\log d}.$$

Proof. By Lemma 11, for coordinate $1\leq i\leq d$,

$$S_i^x=-2h\int_0^1\left(\frac{1}{2}-s\right)f_i''([x_s]_i)ds g_i^2,\ \text{where}\ \left|2h\int_0^1\left(\frac{1}{2}-s\right)f_i''([x_s]_i)ds\right|\leq\frac{h\kappa}{2}.$$

We attained the latter bound by smoothness. Now, each random variable $S_i^x-\mathbb{E}[S_i^x]$ is sub-exponential with parameter $\frac{h\kappa}{2}$ (for coordinates where the coefficient is negative, note the negation of a sub-exponential random variable is still sub-exponential). Hence, by Fact 3,

$$\Pr\left[\sum_{i\in[d]}S_i^x-\mathbb{E}_{g\sim\mathcal{N}(0,\mathrm{id})}\left[\sum_{i\in[d]}S_i^x\right]\geq 10h\kappa\sqrt{d\log d}\right]\leq\frac{1}{d^5}.$$

□

Now, we build a bad set Ω_{hard} with lower bounded measure that starting from a point $x\in\Omega_{\mathrm{hard}}$, with high probability, $-\mathbb{E}_{g\sim\mathcal{N}(0,\mathrm{id})}\left[\sum_{i\in[d]}S_i^x\right]$ is negative:

$$\Omega_{\mathrm{hard}}=\Bigg\{x\ \Bigg|\ |x_1|\leq 2,\forall 2\leq i\leq d,\exists k_i\in\mathbb{Z},|k_i|\leq\left\lfloor\frac{5}{\pi\sqrt{h\kappa}}\right\rfloor,\ \text{such that}\\ -\frac{9}{20}\pi\sqrt{h}+2\pi k_i\sqrt{h}\leq x_i\leq\frac{9}{20}\pi\sqrt{h}+2\pi k_i\sqrt{h}\Bigg\}. \tag{4.21}$$

In other words, Ω_{hard} is the set of points where $\cos x_i$ is large for $2 \le i \le d$, and coordinates are bounded. We first lower bound the measure of Ω_{hard}, and show $\|\nabla f(x)\|_2$ is small within Ω_{hard}. Our measure lower bound will not be used in this section, but will become relevant in Section 4.5.

Lemma 14. *Let $h \le \frac{1}{10000\pi^2\kappa}$. Let π^* have log-density $-f_{\text{hard}}$ (4.20). Then, $\pi^*(\Omega_{\text{hard}}) \ge \exp(-d)$. Moreover, for all $x \in \Omega_{\text{hard}}$, $\|\nabla f(x)\|_2 \le 10\sqrt{\kappa d}$.*

Proof. We first consider a superset of Ω_{hard}. We define the set, for $K \stackrel{\text{def}}{=} \left\lfloor \frac{5}{\pi\sqrt{h\kappa}} \right\rfloor$,

$$\Omega' = \left\{ x \;\middle|\; |x_1| \le 2, \forall 2 \le i \le d, -\frac{9}{20}\pi\sqrt{h} - 2\pi K\sqrt{h} \le x_i \le \frac{9}{20}\pi\sqrt{h} + 2\pi K\sqrt{h} \right\}.$$

It is easy to verify that $\Omega' \supseteq \Omega_{\text{hard}}$. We first show $\pi^*(\Omega')$ is lower bounded by 1.1^{-d}. Since f_{hard} is separable, the coordinates are independent, so it suffices to show each one-dimensional measure is lower bounded by $\frac{1}{1.1}$. This is a standard computation of Gaussian measure for the first coordinate, which we omit. For $2 \le i \le d$, since the marginal distribution is $\frac{\kappa}{3}$-strongly logconcave, it is sub-Gaussian with parameter $\frac{3}{\kappa}$ (see Lemma 1, [DCWY19]). It follows from a standard sub-Gaussian tail bound that the measure of the set $|x_i| \le \frac{9}{\sqrt{\kappa}}$ is at least $\frac{1}{1.1}$. For our choice of K, by assumption on h, $2\pi\sqrt{h}K \ge \frac{10}{\sqrt{\kappa}} - 2\pi\sqrt{h} \ge \frac{9}{\sqrt{\kappa}}$. Combining across coordinates gives $\pi^*(\Omega') \ge 1.1^{-d}$.

Next, we lower bound $\frac{\pi^*(\Omega_{\text{hard}})}{\pi^*(\Omega')}$. We divide the support of the set Ω_{hard} and Ω' into small disjoint regions and bound $\frac{\pi^*(\Omega_{\text{hard}})}{\pi^*(\Omega')}$ for each small region and each coordinate separately. For $2 \le i \le d$, $k \in \left[-\left\lfloor \frac{5}{\pi\sqrt{h\kappa}} \right\rfloor - 1, \left\lfloor \frac{5}{\pi\sqrt{h\kappa}} \right\rfloor \right]$, $k \in \mathbb{Z}$, let

$$\Omega'^{(i,k)} = \left(2\pi k\sqrt{h}, 2\pi(k+1)\sqrt{h} \right],$$

and

$$\Omega_{\text{hard}}^{(i,k)} = \left(2\pi k\sqrt{h}, 2\pi k\sqrt{h} + \frac{9}{20}\pi\sqrt{h} \right] \cup \left[2\pi(k+1)\sqrt{h} - \frac{9}{20}\pi\sqrt{h}, 2\pi(k+1)\sqrt{h} \right].$$

Then, letting π_i^* be the marginal of π^* on coordinate i, we have

$$\begin{aligned}
\frac{\pi_i^*\left(\Omega_{\text{hard}}^{(i,k)}\right)}{\pi_i^*\left(\Omega'^{(i,k)}\right)} &= \frac{\int_{2\pi k\sqrt{h}}^{2\pi k\sqrt{h}+\frac{9}{20}\pi\sqrt{h}} \exp\left(-\frac{\kappa}{3}x_i^2 + \frac{\kappa h}{3}\cos\frac{x_i}{\sqrt{h}}\right) dx_i + \int_{2\pi(k+1)\sqrt{h}-\frac{9}{20}\pi\sqrt{h}}^{2\pi(k+1)\sqrt{h}} \exp\left(-\frac{\kappa}{3}x_i^2 + \frac{\kappa h}{3}\cos\frac{x_i}{\sqrt{h}}\right) dx_i}{\int_{2\pi k\sqrt{h}}^{2\pi(k+1)\sqrt{h}} \exp\left(-\frac{\kappa}{3}x_i^2 + \frac{\kappa h}{3}\cos\frac{x_i}{\sqrt{h}}\right) dx_i} \\
&\ge \frac{\int_{2\pi k\sqrt{h}}^{2\pi k\sqrt{h}+\frac{9}{20}\pi\sqrt{h}} \exp\left(-\frac{\kappa}{3}x_i^2\right) dx_i + \int_{2\pi(k+1)\sqrt{h}-\frac{9}{20}\pi\sqrt{h}}^{2\pi(k+1)\sqrt{h}} \exp\left(-\frac{\kappa}{3}x_i^2\right) dx_i}{\int_{2\pi k\sqrt{h}}^{2\pi(k+1)\sqrt{h}} \exp\left(-\frac{\kappa}{3}x_i^2\right) dx_i \exp\left(\frac{\kappa h}{3}\right)} \\
&\ge \exp\left(-\frac{\kappa h}{3}\right) \cdot \frac{\frac{9}{10}\pi\sqrt{h}}{2\pi\sqrt{h}} \cdot \frac{\exp\left(-\frac{\kappa}{3}\left(2\pi(k+1)\sqrt{h}\right)^2\right)}{\exp\left(-\frac{\kappa}{3}\left(2\pi k\sqrt{h}\right)^2\right)} \ge 0.42.
\end{aligned}$$

The second step used $\cos\frac{x_i}{\sqrt{h}} \geq 0$ for $x_i \in \Omega_{\text{hard}}^{(i,k)}$. The fourth step used the assumption $\kappa h \leq \frac{1}{10000\pi^2}$.

Finally, letting $\Omega'^{(i)}$ and $\Omega_{\text{hard}}^{(i)}$ be the projections of Ω' and Ω_{hard} on the i^{th} coordinate. For any $x_i \in \Omega_i'$ with $x \notin \Omega'^{(i,k)}$, and for all integers $k \in [-K-1, K]$, $x_i \in \Omega_{\text{hard}}^{(i)}$, so $\frac{\pi^*\left(\Omega_{\text{hard}}^{(i)}\right)}{\pi^*\left(\Omega'^{(i)}\right)} \geq 0.42$. Since the coordinates are independent under π^*, $\frac{\pi^*(\Omega_{\text{hard}})}{\pi^*(\Omega')} \geq 0.42^d$. Combining our lower bounds,

$$\pi^*\left(\Omega_{\text{hard}}\right) = \pi^*(\Omega')\frac{\pi^*(\Omega_{\text{hard}})}{\pi^*(\Omega')} \geq \left(\frac{1.1}{0.42}\right)^{-d} \geq \exp\left(-d\right).$$

Finally, we bound $\left\|\nabla f_{\text{hard}}(x)\right\|_2$ for $x \in \Omega'$, from the definition of f_{hard} (4.20),

$$\left\|\nabla f_{\text{hard}}(x)\right\|_2 = \sqrt{f_1'(x)^2 + \sum_{i=2}^{d} f_i'(x)^2} = \sqrt{x_1^2 + \sum_{i=2}^{d}\left(\frac{2\kappa}{3}x_i + \frac{\kappa\sqrt{h}}{3}\sin\frac{x_i}{\sqrt{h}}\right)^2}.$$

Then directly plugging in the definition of Ω_{hard} and using $|\sin c| \leq |c|$ for all c,

$$\left\|\nabla f_{\text{hard}}(x)\right\|_2 \leq \sqrt{1.5^2 + (d-1)\left(9\sqrt{\kappa}\right)^2} \leq 10\sqrt{\kappa d}.$$

□

Finally, we combine the bounds we derived to show the acceptance probability is small within Ω_{hard}.

Lemma 15. *Let $h = o\left(\frac{1}{\kappa \log d}\right)$. For any $x \in \Omega_{\text{hard}}$, let $y = x - h\nabla f_{\text{hard}}(x) + \sqrt{2h}g$ for $g \sim \mathcal{N}(0, \text{id})$. With probability at least $1 - \frac{2}{d^5}$, we have*

$$f_{\text{hard}}(x) - f_{\text{hard}}(y) + \frac{1}{4h}\left(\left\|y - (x - h\nabla f_{\text{hard}}(x))\right\|_2^2 - \left\|x - (y - h\nabla f_{\text{hard}}(y))\right\|_2^2\right) = -\Omega\left(h\kappa d\right).$$

Proof. By combining Lemma 10 and the decomposition (4.14), the conclusion is equivalent to showing that the following quantity is $-\Omega(h\kappa d)$:

$$\begin{aligned}
&-f_{\text{hard}}(x_g) + f_{\text{hard}}(x) - \frac{1}{2}\left\langle x - x_g, \nabla f_{\text{hard}}(x) + \nabla f_{\text{hard}}(x_g)\right\rangle \\
&+f_{\text{hard}}(x_g) - f_{\text{hard}}(y) - \frac{1}{2}\left\langle x - x_g, \nabla f_{\text{hard}}(y) - \nabla f_{\text{hard}}(x_g)\right\rangle \\
&-\frac{1}{2}\left\langle x_g - y, \nabla f_{\text{hard}}(x) + \nabla f_{\text{hard}}(y)\right\rangle + \frac{h}{4}\left\|\nabla f_{\text{hard}}(x)\right\|_2^2 - \frac{h}{4}\left\|\nabla f_{\text{hard}}(y)\right\|_2^2.
\end{aligned}$$

For $x \in \Omega_{\text{hard}}$, every x_i for $2 \leq i \leq d$ has $\cos\frac{x_i}{\sqrt{h}}$ bounded away from 0 by a constant and hence combining Lemmas 12 and 13 implies the first line is $-\Omega(h\kappa d)$ with probability at least $\frac{1}{d^5}$. Regarding the second and third lines, Lemma 11 shows that it suffices to bound (over the set Ω_{hard})

$$(h^2\kappa + h^3\kappa^2)\left\|\nabla f_{\text{hard}}(x)\right\|_2^2 + \left(h^{1.5}\kappa + h^{2.5}\kappa^2\right)\left\|g\right\|_2\left\|\nabla f_{\text{hard}}(x)\right\|_2 + h^2\kappa^2\left\|g\right\|_2^2 = o(h\kappa d).$$

Fact 2 implies $\|g\|_2 \leq \sqrt{2d}$ with probability at least $1 - \frac{1}{d^5}$. Combining this bound, the bound on $\|\nabla f_{\text{hard}}(x)\|_2$ from Lemma 14, and the upper bound on h yields the conclusion. □

We conclude by giving the main result of this section.

Proposition 3. *For $h = o(\frac{1}{\kappa \log d})$, there is a target density on $\mathbb{R}^d$ whose negative log-density always has Hessian eigenvalues in $[1, \kappa]$, such that the relaxation time of MALA is $\Omega(\frac{\kappa d}{\log d})$.*

Proof. The proof is identical to that of Theorem 6, where we use Lemma 15 in place of Proposition 2. □

Theorem 7. *For every step size, there is a target density on $\mathbb{R}^d$ whose negative log-density always has Hessian eigenvalues in $[1, \kappa]$, such that the relaxation time of MALA is $\Omega(\frac{\kappa d}{\log d})$.*

Proof. This is immediate from combining Theorem 6 (with the hard function f_{hq} in the range $h = \Omega(\frac{1}{\kappa \log d})$) and Proposition 3 (with the hard function f_{hard} in the range $h = o(\frac{1}{\kappa \log d})$). □

4.5 Mixing time lower bound for MALA

In this section, we derive a mixing time lower bound for MALA. Concretely, we show that for any step size h, there is a hard distribution $\pi^* \propto \exp(-f)$ such that $\nabla^2 f$ always has eigenvalues in $[1, \kappa]$, yet there is a $\exp(d)$-warm start π_0 such that the chain cannot mix in $o(\frac{\kappa d}{\log^2 d})$ iterations, starting from π_0. We begin by giving such a result for $h = O(\frac{\log d}{\kappa d})$ in Section 4.5.1, and combine it with our developments in Sections 4.3 and 4.4 to prove our main mixing time result.

4.5.1 Mixing time lower bound for small h

Throughout this section, let $h = O\left(\frac{\log d}{\kappa d}\right)$, and let $\pi^* = \mathcal{N}(0, \mathrm{id})$ be the standard d-dimensional multivariate Gaussian. We will let π_0 be the marginal distribution of π^* on the set

$$\Omega \stackrel{\text{def}}{=} \left\{ x \mid \|x\|_2^2 \leq \frac{1}{2} d \right\}.$$

Recall from Lemma 9 that π_0 is a $\exp(d)$-warm start. Our main proof strategy will be to show that for such a small value of h, after $T = O(\frac{\kappa d}{\log^2 d})$ iterations, with constant probability both of the following events happen: no rejections occur throughout the Markov

chain, and $\|x_t\|_2^2 \le \frac{9}{10}d$ holds for all $t \in [T]$. Combining these two facts will demonstrate our total variation lower bound.

Lemma 16. *Let $\{x_t\}_{0\le t<T}$ be the iterates of the MALA Markov chain with step size $h = O\left(\frac{\log d}{\kappa d}\right)$, for $T = o(\frac{\kappa d}{\log^2 d})$ and $x_0 \sim \pi_0$. With probability at least $\frac{99}{100}$, both of the following events occur:*

1. *Throughout the Markov chain, $\|x_t\|_2 \le 0.9\sqrt{d}$.*

2. *Throughout the Markov chain, the Metropolis filter never rejected.*

Proof. We inductively bound the failure probability of the above events in every iteration by $\frac{0.01}{T}$, which will yield the claim via a union bound. Take some iteration $t+1$, and note that by triangle inequality, and assuming all prior iterations did not reject,

$$\|x_{t+1}\|_2 \le \|x_0\|_2 + h\sum_{s=0}^{t}\|x_s\|_2 + \sqrt{2h}\|\sum_{s=0}^{t} g_s\|_2 \le \|x_0\|_2 + 0.9hT\sqrt{d} + \sqrt{2h}\|G_t\|_2 \le 0.8\sqrt{d} + \sqrt{2h}\|G_t\|_2.$$

Here, we applied the inductive hypothesis on all $\|x_s\|_2$, the initial bound $\|x_0\|_2 \le \sqrt{\frac{1}{2}d}$, and that $hT = o(1)$ by assumption. We also defined $G_t = \sum_{s=0}^{t} g_s$, where g_s is the random Gaussian used by MALA in iteration s; note that by independence, $G_t \sim \mathcal{N}(0, t+1)$. By Fact 2, with probability at least $\frac{1}{200T}$, $\|G_t\|_2 \le 2\sqrt{Td}$, and hence $0.8\sqrt{d} + \sqrt{2h}\|G_t\|_2 \le 0.9\sqrt{d}$, as desired.

Next, we prove that with probability $\ge 1 - \frac{1}{200T}$, step t does not reject. This concludes the proof by union bounding over both events in iteration t, and then union bounding over all iterations. By the calculation in Lemma 7, the accept probability is

$$\min\left(1, \exp\left(\frac{h}{4}\left(\left(2h - h^2\right)\|x_t\|_2^2 - 2h\|g\|_2^2 - 2\sqrt{2h}\,(1-h)\,\langle x_t, g\rangle\right)\right)\right).$$

We lower bound the argument of the exponential as follows. With probability at least $1 - d^{-5} \ge 1 - \frac{1}{400T}$, Facts 1 and 2 imply both of the events $\|g\|_2^2 \le 2d$ and $\langle x_t, g\rangle \le 10\sqrt{\log d}\|x\|_2$ occur. Conditional on these bounds, we compute (using $2h \ge h^2$ and the assumption $\|x_t\|_2 \le 0.9\sqrt{d}$)

$$\left(2h - h^2\right)\|x_t\|_2^2 - 2h\|g\|_2^2 - 2\sqrt{2h}\,(1-h)\,\langle x_t, g\rangle \ge -4hd - 40\sqrt{hd\log d} \ge -44\log d.$$

Hence, the acceptance probability is at least

$$\exp\left(-11h\log d\right) \ge 1 - \frac{1}{400T},$$

by our choice of T with $Th\log d = o(1)$, concluding the proof. □

Proposition 4. *The MALA Markov chain with step size $h = O\left(\frac{\log d}{\kappa d}\right)$ requires $\Omega(\frac{\kappa d}{\log^2 d})$ iterations to reach total variation distance $\frac{1}{e}$ to π^*, starting from π_0.*

Proof. Let $\tilde{\pi}$ be the distribution of the MALA Markov chain after $T = o(\frac{\kappa d}{\log^2 d})$ steps without applying a Metropolis filter in any step, and let $\hat{\pi}$ be the distribution after applying the actual MALA chain (including rejections). To show $\|\hat{\pi} - \pi^*\|_{\mathrm{TV}} \geq \frac{1}{e}$, it suffices to show the bounds

$$\|\tilde{\pi} - \pi^*\|_{\mathrm{TV}} \geq \frac{2}{5}, \ \|\tilde{\pi} - \hat{\pi}\|_{\mathrm{TV}} \leq 0.01,$$

and then we apply the triangle inequality. By the coupling characterization of total variation, the second bound follows immediately from the second claim in Lemma 16, wherein we couple the two distributions whenever a rejection does not occur. To show the first bound, the measure of

$$\Omega_{\mathrm{large}} \stackrel{\mathrm{def}}{=} \left\{x \mid \|x\|_2^2 \geq 0.81d\right\}$$

according to π^* is at least 0.99 by Fact 2, and according to $\tilde{\pi}$ it can be at most 0.01 by the first conclusion of Lemma 16. This yields the bound via the definition of total variation. □

4.5.2 Proof of Theorem 8

Finally, we put together the techniques of Sections 4.3, 4.4, and 4.5.1 to prove Theorem 8.

Theorem 8. *For every step size, there is a target density on $\mathbb{R}^d$ whose negative log-density always has Hessian eigenvalues in $[1, \kappa]$, such that MALA initialized at an $\exp(d)$-warm start requires $\Omega(\frac{\kappa d}{\log^2 d})$ iterations to reach e^{-1} total variation distance to the stationary distribution.*

Proof. We consider three ranges of h. First, if $h = \Omega\left(\frac{1}{\kappa \log d}\right)$, we use the hard function f_{hq} and the hard set in (4.12), which has measure at least $\exp(-d)$ according to the stationary distribution by Lemma 9. Then, applying Proposition 2 demonstrates that the chance the Markov chain can move over d^5 iterations is $o(\frac{1}{d})$, and hence it does not reach total variation $\frac{1}{e}$ in this time. Next, if $h = o\left(\frac{1}{\kappa \log d}\right) \cap \omega\left(\frac{\log d}{\kappa d}\right)$, we use the hard function f_{hard} and the hard set in (4.21), which has measure at least $\exp(-d)$ by Lemma 14. Applying Lemma 15 again implies the chain does not mix in d^5 iterations. Finally, if $h = O\left(\frac{\log d}{\kappa d}\right)$, applying Proposition 4 yields the claim. □

4.6 Lower bounds for HMC

In this section, we derive a lower bound on the spectral gap of HMC. We first analyze some general structural properties of HMC in Section 4.6.1, as a prelude to later sections. We then provide a lower bound for HMC on quadratics in Section 4.6.2, with any number of leapfrog steps K.

4.6.1 Structure of HMC: a detour to Chebyshev polynomials

We begin our development with a bound on the acceptance probability for general HMC Markov chains. Recall from (4.6) that this probability is (for $\mathcal{H}(x,v) \stackrel{\text{def}}{=} f(x) + \frac{1}{2}\|v\|_2^2$)

$$\min\left\{1, \frac{\exp\left(-\mathcal{H}(x_K, v_K)\right)}{\exp\left(-\mathcal{H}(x_0, v_0)\right)}\right\}. \tag{4.6}$$

We first state a helper calculation straightforwardly derived from the exposition in Section 4.2.4.

Fact 4. *One step of the HMC Markov chain starting from x_0 generates iterates $\{(v_{k-\frac{1}{2}}, x_k, v_k)\}_{0\le k\le K}$ defined recursively by the closed-form equations:*

$$\begin{aligned}
v_{k-\frac{1}{2}} &= v_0 - \frac{\eta}{2}\nabla f(x_0) - \eta \sum_{j\in[k-1]} \nabla f(x_j),\\
v_k &= v_0 - \frac{\eta}{2}\nabla f(x_0) - \eta \sum_{j\in[k-1]} \nabla f(x_j) - \frac{\eta}{2}\nabla f(x_k),\\
x_k &= x_0 + \eta k v_0 - \frac{\eta^2 k}{2}\nabla f(x_0) - \eta^2 \sum_{j\in[k-1]} (k-j)\nabla f(x_j).
\end{aligned}$$

When expanding the acceptance probability (4.6) using the equations in Fact 4, many terms conveniently cancel, which we capture in Lemma 17. This phenomenon underlies the improved performance of HMC on densities with highly-Lipschitz Hessians [CDWY20].

Lemma 17. *For the iterates given by Fact 4,*

$$\begin{aligned}
\mathcal{H}\left(x_0, v_0\right) - \mathcal{H}\left(x_K, v_K\right) = \sum_{0\le k\le K-1} \left(f(x_k) - f(x_{k+1}) + \frac{1}{2}\left\langle \nabla f(x_k) + \nabla f(x_{k+1}), x_{k+1} - x_k\right\rangle\right)\\
+ \frac{\eta^2}{8}\|\nabla f(x_0)\|_2^2 - \frac{\eta^2}{8}\|\nabla f(x_K)\|_2^2.
\end{aligned}$$

Proof. Recall $\mathcal{H}(x_0,v_0) - \mathcal{H}(x_K,v_K) = f(x_0) - f(x_K) + \frac{1}{2}\|v_0\|_2^2 - \frac{1}{2}\|v_K\|_2^2$. We begin by expanding

$$\frac{1}{2}\|v_0\|_2^2 - \frac{1}{2}\|v_K\|_2^2 = \frac{1}{2}\|v_0\|_2^2 - \frac{1}{2}\|v_0 - \frac{\eta}{2}\nabla f(x_0) - \eta\sum_{j\in[K-1]}\nabla f(x_j) - \frac{\eta}{2}\nabla f(x_K)\|_2^2$$

$$
\begin{aligned}
&= \eta \left\langle v_0, \frac{1}{2}\nabla f(x_0) + \sum_{j\in[K-1]} \nabla f(x_j) + \frac{1}{2}\nabla f(x_K) \right\rangle \\
&- \frac{\eta^2}{2}\|\frac{1}{2}\nabla f(x_0) + \sum_{j\in[K-1]} \nabla f(x_j) + \frac{1}{2}\nabla f(x_K)\|_2^2 \\
&= \frac{\eta}{2} \sum_{0\le k\le K-1} \langle v_0, \nabla f(x_k) + \nabla f(x_{k+1})\rangle \\
&- \frac{\eta^2}{2} \sum_{0\le k\le K-1} \left\langle \nabla f(x_k) + \nabla f(x_{k+1}), \frac{1}{2}\nabla f(x_0) + \sum_{j\in[k]} \nabla f(x_j) \right\rangle \\
&+ \frac{\eta^2}{8} \langle \nabla f(x_0) - \nabla f(x_K), \nabla f(x_0) + \nabla f(x_K)\rangle \,.
\end{aligned}
$$

Here the first equality used Fact 4. Moreover, for each $0 \le k \le K-1$, by Fact 4

$$
\begin{aligned}
\frac{1}{2}\langle \nabla f(x_k) + \nabla f(x_{k+1}), x_{k+1} - x_k\rangle &= \frac{\eta}{2}\langle \nabla f(x_k) + \nabla f(x_{k+1}), v_0\rangle \\
&- \frac{\eta^2}{2}\left\langle \nabla f(x_k) + \nabla f(x_{k+1}), \frac{1}{2}\nabla f(x_0) + \sum_{j\in[k]} \nabla f(x_j) \right\rangle .
\end{aligned}
$$

Combining yields the result. □

We state a simple corollary of Lemma 17 in the case of quadratics.

Corollary 6. *For $f(x) = \frac{1}{2}x^\top \mathbf{A} x$, the iterates given by Fact 4 satisfy*

$$
\mathcal{H}(x_0, v_0) - \mathcal{H}(x_K, v_K) = \frac{\eta^2}{8}\|\nabla f(x_0)\|_2^2 - \frac{\eta^2}{8}\|\nabla f(x_K)\|_2^2.
$$

Proof. It suffices to observe that for any two points $x, y \in \mathbb{R}^d$,

$$
f(x) - f(y) + \frac{1}{2}\langle \nabla f(x) + \nabla f(y), y - x\rangle = \frac{1}{2}x^\top \mathbf{A} x - \frac{1}{2}y^\top \mathbf{A} y + \frac{1}{2}\langle \mathbf{A}(x+y), y - x\rangle = 0.
$$

□

Finally, it will be convenient to have a more explicit form of iterates in the case of quadratics, which follows directly from examining the recursion in Fact 4.

Lemma 18. *For $f(x) = \frac{1}{2}x^\top \mathbf{A} x$, the iterates $\{x_k\}_{0\le k\le K}$ given by Fact 4 satisfy*

$$
\begin{aligned}
x_k &= \left(\sum_{0\le j\le k} D_{j,k}(\eta^2 \mathbf{A})^j \right) x_0 + \left(\eta \sum_{0\le j\le k-1} E_{j,k}(\eta^2 \mathbf{A})^j \right) v_0, \\
\text{where } D_{j,k} &\stackrel{\text{def}}{=} (-1)^j \cdot \frac{k}{k+j} \cdot \binom{k+j}{2j},\ E_{j,k} \stackrel{\text{def}}{=} (-1)^j \cdot \binom{k+j}{2j+1}.
\end{aligned} \tag{4.22}
$$

Proof. This formula can be verified to match the recursions of Fact 4 by checking the base cases $D_{0,k} = 1$, $D_{1,k} = -\frac{k^2}{2}$, $E_{0,k} = k$, and (where $D_{j,k} \stackrel{\text{def}}{=} 0$ for $j > k$ and $E_{j,k} \stackrel{\text{def}}{=} 0$ for $j \geq k$)

$$D_{j,k} = -\sum_{i\in[k-1]} (k-i)D_{j-1,i},\ E_{j,k} = -\sum_{i\in[k-1]} (k-i)E_{j-1,i}.$$

In particular, by using the third displayed line of Fact 4, the coefficient of $(\eta^2\mathbf{A})^j x_0$ in x_k for $j \geq 2$ is the negated sum of the coefficients of $(\eta^2\mathbf{A})^{j-1}$ in all $(k-i)x_i$. Similarly, the coefficient of $\eta(\eta^2\mathbf{A})^j v_0$ in x_k for $j \geq 1$ is the negated sum of the coefficients of $\eta(\eta^2\mathbf{A})^{j-1}$ in all $(k-i)x_i$. The displayed coefficient identities follow from the binomial coefficient identities

$$\frac{k}{k+j}\binom{k+j}{2j} = \sum_{j-1\leq i\leq k-1} \frac{(k-i)i}{i+j-1}\binom{i+j-1}{2j-2},\ \binom{k+j}{2j+1} = \sum_{j\leq i\leq k-1} (k-i)\binom{i+j-1}{2j-1}.$$

□

Lemma 18 motivates the definition of the polynomials

$$p_k(z) \stackrel{\text{def}}{=} \sum_{0\leq j\leq k} D_{j,k}z^j,\ q_k(z) \stackrel{\text{def}}{=} \sum_{0\leq j\leq k-1} E_{j,k}z^j. \tag{4.23}$$

In this way, at least in the case when $\mathbf{A} = \mathbf{diag}(\lambda)$ for a vector of eigenvalues $\lambda \in \mathbb{R}^d$, we can concisely express the coordinates of iterates in (4.22) by

$$[x_k]_i = p_k(\eta^2\lambda_i)[x_0]_i + \eta q_k(\eta^2\lambda_i)[v_0]_i. \tag{4.24}$$

Interestingly, the polynomial p_k turns out to have a close relationship with the k^{th} *Chebyshev polynomial* (of the first kind), which we denote by T_k. Similarly, the polynomial q_k is closely related to the $(k-1)^{\text{th}}$ *Chebyshev polynomial* of the second kind, denoted U_{k-1}. The relationship between the Chebyshev polynomials and the phenomenon of *acceleration* for optimizing quadratics via first-order methods has been known for some time (see e.g. [Har13, Bac19] for discussions), and we find it interesting to further explore this relationship. Concretely, the following identities hold.

Lemma 19. *Following definitions* (4.22), (4.23),

$$p_k(z) = T_k\left(1 - \frac{z}{2}\right),\ q_k(z) = U_{k-1}\left(1 - \frac{z}{2}\right).$$

Proof. It is easy to check $p_0(z) = 1$ and $p_1(z) = 1 - \frac{z}{2}$, so the former conclusion would follow from

$$p_{k+1}(z) = (2-z)p_k(z) - p_{k-1}(z) \iff D_{j,k+1} = 2D_{j,k} - D_{j-1,k} - D_{j,k-1},$$

following well-known recursions defining the Chebyshev polynomials of the first kind. This identity can be verified by direct expansion. Moreover, for the latter conclusion, recalling the definition of Morgan-Voyce polynomials of the first kind $B_k(z)$, we can directly match $q_k(z) = B_{k-1}(-z)$. The conclusion follows from Section 4 of [AJ94], which shows $B_{k-1}(-z) = U_{k-1}(1 - \frac{z}{2})$ as desired (note that in the work [AJ94], the indexing of Chebyshev polynomials is off by one from ours). □

Now for $z = \eta^2\lambda_i$, we have from (4.24) and Lemma 19 that $[x_k]_i = \pm[x_0]_i$ precisely when

$$p_k(z) = T_k\left(1 - \frac{z}{2}\right) = \pm 1,\ q_k(z) = U_{k-1}\left(1 - \frac{z}{2}\right) = 0.$$

Hence, this occurs whenever $1 - \frac{z}{2}$ is both an *extremal point* of T_k in the range $[-1, 1]$ and a root of U_{k-1}. Both of these occur exactly at the points $\cos(\frac{j}{k}\pi)$, for $0 \le j \le k$.

Proposition 5. *For $\kappa \ge \pi^2$ and $K \ge 2$, no K-step HMC Markov chain with step size $1 \ge \eta^2 \ge \frac{\pi^2}{\kappa K^2}$ can mix in finite time for all densities on $\mathbb{R}^d$ whose negative log-density's Hessian has eigenvalues between 1 and κ for all points $x \in \mathbb{R}^d$, initialized at a constant-warm start.*

Proof. Fix a value of $1 \ge \eta \ge \sqrt{\frac{\pi^2}{\kappa K^2}}$. We claim there exists a $1 \le j \le K - 1$ such that for

$$\lambda \stackrel{\text{def}}{=} \frac{2\left(1 - \cos\left(\frac{j\pi}{K}\right)\right)}{\eta^2},\ 1 \le \lambda \le \kappa.$$

Since λ is a monotone function of η, it suffices to check the endpoints of the interval $[\frac{\pi^2}{\kappa K^2}, 1]$. For $\eta^2 = 1$, we choose $j = K - 1$, which using $\frac{2x^2}{\pi^2} \le 1 - \cos(x) \le \frac{x^2}{2}$ for all $-\pi \le x \le \pi$, yields

$$1 \le \frac{4(K-1)^2}{K^2} \le \lambda \le \frac{(K-1)^2\pi^2}{K^2} \le \pi^2 \le \kappa.$$

Similarly, for $\eta^2 = \frac{\pi^2}{\kappa K^2}$, we choose $j = 1$, which yields

$$1 \le \frac{4}{\eta^2 K^2} \le \lambda \le \frac{\pi^2}{\eta^2 K^2} \le \kappa.$$

Now, consider the quadratic $f(x) = \frac{1}{2}x^\top \mathbf{A}x$ where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is a diagonal matrix, $\mathbf{A}_{11} = 1$, $\mathbf{A}_{ii} = \kappa$ for all $3 \le i \le d$, and $\mathbf{A}_{22} = \lambda \stackrel{\text{def}}{=} \frac{2\left(1-\cos\left(\frac{j\pi}{K}\right)\right)}{\eta^2}$ for the choice of j which makes $1 \le \lambda \le \kappa$. For any symmetric starting set capturing a constant amount of measure along the second coordinate, by Lemma 18 and the following exposition, $x_K = \pm x_0$ along the second coordinate regardless of the random choice of velocity and thus the chain cannot leave the starting set. □

4.6.2 HMC lower bound for all K

We now give our HMC lower bound, via improving Proposition 5 by a dimension dependence. We begin in Section 4.6.2, where we give a stronger upper bound on η in the range $\eta^2 \leq \frac{1}{\kappa K^2}$. Noting that there is a constant-sized gap between this range and the bound in Proposition 5, we rule out this gap in Section 4.6.2. Finally, we handle the case of extremely large $\eta^2 \geq 1$ in Section 4.6.2. We put these pieces together to prove Theorem 9 in Section 4.6.2.

Upper bounding $\eta = O(K^{-1}\kappa^{-\frac{1}{2}})$ under a constant gap

For this section, we let $\mathbf{A}$ be the $d \times d$ diagonal matrix which induces the hard quadratic function f_{hq}, defined in (4.7) and reproduced here for convenience:

$$f_{\mathrm{hq}}(x) \stackrel{\mathrm{def}}{=} \sum_{i \in [d]} f_i(x_i), \text{ where } f_i(c) = \begin{cases} \frac{1}{2}c^2 & i = 1 \\ \frac{\kappa}{2}c^2 & 2 \leq i \leq d \end{cases}.$$

We also let $h \stackrel{\mathrm{def}}{=} \frac{\eta^2}{2}$, $x \stackrel{\mathrm{def}}{=} x_0$, $g \stackrel{\mathrm{def}}{=} v_0$, and $y \stackrel{\mathrm{def}}{=} x_K$ throughout for analogy to Section 4.3, so that we can apply Proposition 2. Next, note that by the closed-form expression given by Lemma 18, we can write the iterates of the HMC chain in the form (4.8), reproduced here:

$$y = \begin{pmatrix} y_1 \\ y_{-1} \end{pmatrix}, \text{ where } y_1 = (1 - \alpha_1)x_1 + \beta_1 g_1$$
$$\text{and } y_{-1} = (1 - \alpha_{-1})x_{-1} + \beta_{-1} g_{-1}, \text{ for } g \sim \mathcal{N}(0, \mathrm{id}).$$

Concretely, we have by Lemma 18 that

$$\begin{aligned} \alpha_1 &= -\sum_{1 \leq j \leq K} (-1)^j (2h)^j \left(\frac{K}{K+j}\right)\binom{K+j}{2j}, \\ \alpha_{-1} &= -\sum_{1 \leq j \leq K} (-1)^j (2h\kappa)^j \left(\frac{K}{K+j}\right)\binom{K+j}{2j}, \\ \beta_1 &= \sqrt{2h} \sum_{0 \leq j \leq K-1} (-1)^j (2h)^j \binom{K+j}{2j+1}, \\ \beta_{-1} &= \sqrt{2h} \sum_{0 \leq j \leq K-1} (-1)^j (2h\kappa)^j \binom{K+j}{2j+1}. \end{aligned} \tag{4.25}$$

By a straightforward computation, the parameters in (4.25) satisfy the conditions of Proposition 2.

Lemma 20. *Supposing* $\eta^2 \le \frac{1}{\kappa K^2}$, α_1, α_{-1}, β_1, β_{-1} *defined in* (4.25) *satisfy*

$$|\alpha_{-1}| \le \frac{3}{5}\beta_{-1}^2\kappa,\ |\alpha_1| = O(|\alpha_{-1}|),\ \beta_1 = O(\beta_{-1}).$$

Proof. The proof follows since under $\eta^2 \le \frac{1}{10\kappa K^2}$, all of the parameters in (4.25) are dominated by their first summand. We will argue this for α_{-1} and β_{-1}; the corresponding conclusions for α_1 and β_1 follow analogously since $\kappa \ge 1$. Define the summands of α_{-1} and β_{-1} by

$$c_j \stackrel{\text{def}}{=} (-1)^{j+1}(2h\kappa)^j \left(\frac{K}{K+j}\right)\binom{K+j}{2j},\ 1 \le j \le K,$$
$$d_j \stackrel{\text{def}}{=} \sqrt{2h}(-1)^j(2h\kappa)^j \binom{K+j}{2j+1},\ 0 \le j \le K-1.$$

Then, we compute that for all $1 \le j \le K-1$, assuming $2h\kappa K^2 \le 1$,

$$0 \ge \frac{c_{j+1}}{c_j} = (-2h\kappa)\frac{(K+j)(K-j)}{(2j+2)(2j+1)} \ge -\frac{2h\kappa K^2}{12} \ge -0.1. \tag{4.26}$$

Similarly, for all $0 \le j \le K-2$,

$$0 \ge \frac{d_{j+1}}{d_j} = (-2h\kappa)\,\frac{(K+j+1)(K-j-1)}{(2j+3)(2j+2)} \ge -\frac{2h\kappa K^2}{6} \ge -0.2. \tag{4.27}$$

By repeating these calculations for α_1 and β_1, we see that all parameters are given by rapidly decaying geometric sequences, and thus the conclusion follows by examination from

$$\alpha_1 \in \left[0.8hK^2, hK^2\right],\ \alpha_{-1} \in \left[0.8h\kappa K^2, h\kappa K^2\right],$$
$$\beta_1 \in \left[0.8\sqrt{2h}K, \sqrt{2h}K\right],\ \beta_{-1} \in \left[0.8\sqrt{2h}K, \sqrt{2h}K\right].$$

□

We obtain the following corollary by combining Lemma 20, Corollary 6, and Proposition 2.

Corollary 7. *Let* $x \in \mathbb{R}^d$ *satisfy* $\|x_{-1}\|_2 \le \sqrt{\frac{2d}{3\kappa}}$ *and* $|x_1| \le 5\sqrt{\log d}$, *let* (x_K, v_K) *be the result of the* K*-step HMC Markov chain with step size* $\eta = \sqrt{2h}$ *with* $\eta^2 \le \frac{1}{\kappa K^2}$ *from* $x_0 = x$, *and let* $\mathbf{A}$ *be as in* (4.7). *Then with probability at least* $1 - d^{-5}$ *over the randomness of* $v_0 \sim \mathcal{N}(0, \mathrm{id})$, *we have*

$$\mathcal{H}(x_0, v_0) - \mathcal{H}(x_K, v_K) = -\Omega\left(h^2\kappa^2K^2d\right).$$

Proof. It suffices to use the bounds on $\beta_{-1} = \Theta(\sqrt{h}K)$ shown in the proof of Lemma 20 and the conclusions of Corollary 6 and Proposition 2. □

Removing the constant gap

We show how to improve the bound in Corollary 7 to only require $\eta^2 \le \frac{\pi^2}{\kappa K^2}$, which removes the constant gap between the requirement of Corollary 7 and the bound in Proposition 5. First, let $\mathbf{A}_c$ be the $D \times d$ diagonal matrix which induces the following hard quadratic function f_{hqc}:

$$f_{\text{hqc}}(x) \stackrel{\text{def}}{=} \sum_{i \in [d]} f_i(x_i), \text{ where } f_i(c) = \begin{cases} \frac{1}{2}c^2 & i = 1 \\ \frac{\kappa}{2\pi^2}c^2 & 2 \le i \le d-1 \\ \frac{\kappa}{2}c^2 & i = d \end{cases}. \tag{4.28}$$

In other words, along the first $d-1$ coordinates, f_{hqc} is the same as a $d-1$-dimensional variant of f_{hq} with condition number $\frac{\kappa}{\pi^2}$. We define a coordinate partition of x and g into x_1, x_{-1d}, x_d, and g_1, g_{-1d}, g_d, and we define α_1, α_{-1d}, α_d, β_1, β_{-1d}, β_d in analogy with (4.8).

We first note that because of separability of f_{hqc}, and since the assumption of Corollary 7 holds on the first $d-1$ coordinates for $\eta^2 \le \frac{\pi^2}{\kappa K^2}$, we can immediately obtain a bound on the change in the Hamiltonian along these coordinates.

Corollary 8. *Let $x \in \mathbb{R}^d$ satisfy $\|x_{-1}\|_2 \le \sqrt{\frac{2\pi^2 d}{3\kappa}}$ and $|x_1| \le 5\sqrt{\log d}$, let (x_K, v_K) be the result of the K-step HMC Markov chain with step size $\eta = \sqrt{2h}$ where $\eta^2 \le \frac{\pi^2}{\kappa K^2}$ from $x_0 = x$, and let $\mathbf{A}_c$ be as in (4.28). Then with probability at least $1 - 2d^{-5}$ over the randomness of $v_0 \sim \mathcal{N}(0, \mathrm{id})$, we have*

$$\mathcal{H}\left([x_0]_{[d-1]}, [v_0]_{[d-1]}\right) - \mathcal{H}\left([x_K]_{[d-1]}, [v_K]_{[d-1]}\right) = -\Omega\left(h^2\kappa^2 K^2 d\right).$$

We now move to bounding the contribution of the last coordinate.

Lemma 21. *Let (y, v_K) be the result of the K-step HMC Markov chain with step size $\eta = \sqrt{2h}$ where $\eta^2 \le \frac{\pi^2}{\kappa K^2}$, and write $y_d = (1-\alpha_d)x_d + \beta_d g_d$, for*

$$\alpha_d = -\sum_{1 \le j \le K} (-1)^j (2h\kappa)^j \left(\frac{K}{K+j}\right)\binom{K+j}{2j}, \; \beta_d = \sqrt{2h} \sum_{0 \le j \le K-1} (-1)^j (2h\kappa)^j \binom{K+j}{2j+1}.$$

Then, we have $|\alpha_d| = O(h\kappa K^2)$, $|\beta_d| = O(\sqrt{h}K)$.

Proof. After the index j is a sufficiently large constant, the geometric argument sequence of Lemma 20 applies (since the denominators of the ratios (4.26) and (4.27) grow with the index j); before then, each coefficient is within a constant factor of the first in absolute

value. Thus, the coefficients can be at most a constant factor larger than the first in absolute value. □

Lemma 22. *Let* $|[x_0]_d| \le \frac{\log d}{\sqrt{\kappa}}$, $|[v_0]_d| \le \log d$, *and let* (x_K, v_K) *be the result of the* K*-step HMC Markov chain with step size* $\eta = \sqrt{2h}$ *where* $\eta^2 \le \frac{\pi^2}{\kappa K^2}$. *Then with probability at least* $1 - d^{-5}$ *over the randomness of* $v_0 \sim \mathcal{N}(0, \mathrm{id})$, *we have*

$$\mathcal{H}\left([x_0]_d, [v_0]_d\right) - \mathcal{H}\left([x_K]_d, [v_K]_d\right) = o\left(h^2\kappa^2K^2d\right).$$

Proof. We can assume $|[v_0]_d| = |g_d| \le \log d$, which passes the high probability bound. By Corollary 6 and Lemma 7, we wish to bound

$$\frac{h\kappa^2}{4}\left(\left(2\alpha_d - \alpha_d^2\right)x_d^2 - \beta_d^2 g_d^2 - 2(1-\alpha_d)\beta_d x_d g_d\right) = o\left(h^2\kappa^2K^2d\right).$$

Dropping all clearly negative terms, and since $|\alpha_d| = O(1)$ by Lemma 21, it is enough to show

$$\left|h\kappa^2\alpha_d x_d^2\right| = o\left(h^2\kappa^2K^2d\right), \ \left|h\kappa^2\beta_d x_d g_d\right| = o\left(h^2\kappa^2K^2d\right).$$

The first bound is immediate from assumptions. The second follows from assumptions as well since $\sqrt{h\kappa K^2}$ is at most a constant, so $\left|h\kappa^2\beta_d x_d g_d\right| = O(h^{1.5}\kappa^{1.5}K\log^2 d) = O(h^2\kappa^2K^2\log^2 d)$. □

By combining Lemma 22 and Corollary 8, we obtain the following strengthening of Corollary 7.

Corollary 9. *Let* $x \in \mathbb{R}^d$ *satisfy* $\|x_{-1d}\|_2 \le \sqrt{\frac{2d}{3\kappa}}$, $|x_1| \le 5\sqrt{\log d}$, *and* $|x_d| \le \frac{\log d}{\sqrt{\kappa}}$, *let* (x_K, v_K) *be the result of the* K*-step HMC Markov chain with step size* $\eta = \sqrt{2h}$ *with* $\eta^2 \le \frac{\pi^2}{\kappa K^2}$ *from* $x_0 = x$, *and let* $\mathbf{A}_c$ *be as in* (4.28). *Then with probability at least* $1 - d^{-5}$ *over the randomness of* $v_0 \sim \mathcal{N}(0, \mathrm{id})$, *we have*

$$\mathcal{H}(x_0, v_0) - \mathcal{H}(x_K, v_K) = -\Omega\left(h^2\kappa^2K^2d\right).$$

Ruling out $\eta \ge 1$

Finally, we give a short argument ruling out the case $\eta \ge 1$ not covered by Proposition 5. In this section, let $\pi^* = \mathcal{N}(0, \kappa^{-1}\,\mathrm{id})$, with negative log-density $f(x) = \frac{\kappa}{2}\|x\|_2^2$. For $\eta \ge 1$ and $\kappa \ge 10$, (4.24) and straightforward lower bounds on Chebyshev polynomials outside the range $[-1, 1]$ demonstrate the proposal distribution is of the form (from starting point $x_0 \in \mathbb{R}^d$)

$$x_K \leftarrow \alpha x_0 + \beta v_0, \ v_0 \sim \mathcal{N}(0,1), \ |\alpha| \ge 10, \ |\beta| \ge 1. \tag{4.29}$$

Lemma 23. *Letting (x_K, v_K) be the result of K-step HMC from any x_0, and $f(x) = \frac{\kappa}{2}\|x\|_2^2$, for $\eta \geq 1$, with probability at least $1 - d^{-5}$ over the randomness of $v_0 \sim \mathcal{N}(0, \mathrm{id})$, we have*

$$\mathcal{H}(x_0, v_0) - \mathcal{H}(x_K, v_K) = -\Omega(d).$$

Proof. Following notation (4.29) and applying Corollary 6, it suffices to show

$$\|x_0\|_2^2 - \|\alpha x_0 + \beta v_0\|_2^2 = -\Omega(d).$$

Expanding, it suffices to upper bound

$$\left(1 - \alpha^2\right)\|x_0\|_2^2 - 2\alpha\beta\,\langle x_0, v_0\rangle - \beta^2\|v_0\|_2^2.$$

With probability at least $1 - d^{-5}$, Fact 2 shows $\|v_0\|_2^2 \geq \frac{1}{2}d$ and $\langle x_0, v_0\rangle \geq -4\sqrt{\log d}\|x_0\|_2$. Hence,

$$\begin{aligned}\left(1 - \alpha^2\right)\|x_0\|_2^2 - 2\alpha\beta\,\langle x_0, v_0\rangle - \beta^2\|v_0\|_2^2 &\leq -0.99\alpha^2\|x_0\|_2^2 + 8\alpha\beta\sqrt{\log d}\|x_0\|_2 - \frac{\beta^2}{2}d \\ &\leq 20\beta^2\log d - \frac{\beta^2}{2}d = -\Omega(d).\end{aligned}$$

Here, we used that $\alpha^2 \geq 100$ and took d larger than a sufficiently large constant. □

Proof of Theorem 9

A consequence of Corollary 7 is that if the step size $h = \omega(\frac{\sqrt{\log d}}{\kappa K\sqrt{d}})$, initializing the chain from any x_0 in the set Ω defined in (4.12) leads to a polynomially bad mixing time. We further relate the step size to the spectral gap of the HMC Markov chain in the following.

Lemma 24. *The spectral gap of the K-step HMC Markov chain for sampling from the density proportional to $\exp(-f_{\mathrm{hq}})$, where f_{hq} is defined in (4.7), is $O(hK^2 + h^2K^4)$.*

Proof. We follow the proof of Lemma 6; again let $g(x) = x_1$, and π^* be the stationary distribution. For our function f, it is clear again that $\mathrm{Var}_{\pi^*}[g] = \Theta(1)$. Thus it suffices to upper bound $\mathcal{E}(g, g)$: letting $\mathcal{P}_x(y)$ be the proposal distribution of K-step HMC, and α_1, β_1 be as in (4.25),

$$\begin{aligned}\mathcal{E}(g, g) &\leq \frac{1}{2}\iint (x_1 - y_1)^2\mathcal{P}_x(y)d\pi^*(x)dy \\ &\leq \mathbb{E}_{x\sim\pi^*}\left[\alpha_1^2 x_1^2\right] + \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\beta_1^2\xi^2\right] \\ &= \alpha_1^2 + \beta_1^2 = O\left(hK^2 + h^2K^4\right).\end{aligned}$$

□

Finally, by combining Lemma 24 and Corollary 9, we arrive at the main result of this section.

Theorem 9. *For every step size and count, there is a target Gaussian on $\mathbb{R}^d$ whose negative log-density always has Hessian eigenvalues in $[1, \kappa]$, such that the relaxation time of HMC is $\Omega(\frac{\kappa\sqrt{d}}{K\sqrt{\log d}})$.*

Proof. For $1 \geq \eta^2 \geq \frac{\pi^2}{\kappa K^2}$ it suffices to apply Proposition 5. For $\eta^2 \geq 1$, we apply Lemma 23. Otherwise, in the relevant range of $h = 2\eta^2$, the dominant term in Lemma 24 is $O(hK^2)$. Applying Corollary 9 with the hard quadratic function f_{hqc}, the remainder of the proof follows analogously to that of Theorem 6. □

We remark that as in Theorem 6, it is straightforward to see that the measure of the bad region $\|x_{-1d}\|_2 \leq \sqrt{\frac{2d}{3\kappa}}$, $|x_1| \leq 5\sqrt{\log d}$, and $|x_d| \leq \frac{\log d}{\sqrt{\kappa}}$ used in the proof is at least $\exp(-d)$.

4.7 Conclusion

We presented relaxation time lower bounds for the MALA and HMC Markov chains at every step size and scale, as well as a mixing time bound for MALA from an exponentially warm start. We highlight in this section a number of unexplored directions left open by our work, beyond direct strengthenings of our results, which we find interesting and defer to a future exploration.

Variable or random step sizes. The lower bounds of this paper were for MALA and HMC Markov chains with a *fixed step size*. For variable step sizes which take e.g. values in a bounded multiplicative range, we believe our arguments can be modified to also give relaxation time lower bounds for the resulting Markov chains. However, the arguments of Section 4.6 (our HMC lower bound) are particularly brittle to large multiplicative ranges of candidate step sizes, because they rely on the locations of Chebyshev polynomial zeroes, which only occur in a bounded range. From an algorithm design perspective, this suggests that adaptively or randomly choosing step size ranges may be effective in improving the performance of HMC. Such a result would also give theoretical justification to the No-U-Turn sampler of [HG14a], a common HMC alternative in practice. We state as an explicit open problem: can one obtain improved upper bounds, such as a $\sqrt{\kappa}$ dependence or a dimension-independent rate, for example by using variations of these strategies (variable step sizes)?

Necessity of κ lower bound. All of our witness sets throughout the paper are $\exp(-d)$ sized. It was observed in [DCWY19] that it is possible to construct a starting distribution with warmness arbitrarily close to $\sqrt{\kappa}^d$; the marginal restriction of our witness set obtains this bound for all $\kappa \geq e^2$. However, recently [LST21b] proposed a *proximal point reduction* for sampling, which we will discuss in Chapter 5, showing (for mixing bounds scaling at least linearly in κ) it suffices to sample a small number of regularized distributions, with conditioning arbitrarily close to 1. Adjusting constants, we can modify our Gaussian lower bounds (Theorems 6 and 9) to have witness sets with measure c^d for c arbitrarily close to 1. However, our witness set for the family of hard non-Gaussian distributions encounters a natural barrier at measure 2^d, as it is sign-restricted by the cosine function. We find it interesting to see if a stronger construction rules out existing warm starts for all $\kappa \geq 1$, or if an upper bound can take advantage of the [LST21b] reduction to obtain improved dependences on dimension.

Part III

PROXIMAL SAMPLING METHOD

Chapter 5

STRUCTURED LOGCONCAVE SAMPLING USING PROXIMAL SAMPLING METHODS

This chapter is based on [LST21b], with Yin Tat Lee and Kevin Tian.

5.1 Introduction

Developing efficient algorithms for sampling from *structured* logconcave densities is a topic that has received significant recent interest due to its widespread practical applications. There are many types of structure which densities commonplace in applications may possess that are exploitable for improved runtimes. Examples of such structure include derivative bounds ("well-conditioned densities") and various types of separability (e.g. "composite densities" corresponding to possibly non-smooth regularization or restrictions to a set, and "logconcave finite sums" corresponding to averages over multiple data points).[1] Building an algorithmic theory for sampling these latter two families, which are not well-understood in the literature, is a primary motivation of this work.

There are strong parallels between the types of structured logconcave families garnering recent attention and the classes of convex functions known to admit efficient first-order optimization algorithms. Notably, gradient descent and its accelerated counterpart [Nes83] are well-known to quickly optimize a well-conditioned function, and have become ubiquitous in both practice and theory. Similarly, various methods have been developed for efficiently optimizing non-smooth but structured composite objectives [BT09] and well-conditioned finite sums [All17].

Logconcave sampling and convex optimization are intimately related primitives (cf. e.g. [BV04, AH16]), so it is perhaps unsurprising that there are analogies between the types of structure algorithm designers may exploit. Nonetheless, our understanding of the complexity landscape for sampling is quite a bit weaker in comparison to counterparts in the field of optimization; few lower bounds are known for the complexity of sampling tasks, and obtaining stronger upper bounds is an extremely active research area (contrary

[1] We make this terminology more precise in Section 5.2.1, which contains various definitions used in this paper.

to optimization, where matching bounds exist in many cases). Moreover (and perhaps relatedly), the toolkit for designing logconcave samplers is comparatively lacking; for many important primitives in optimization, it is unclear if there are analogs in sampling, possibly impeding improved bounds. Our work broadly falls under the themes of (1) understanding which types of structured logconcave distributions admit efficient samplers, and (2) leveraging connections between optimization and sampling for algorithm design. We address these problems on two fronts, which constitute the primary technical contributions of this paper.

1. We give a general reduction framework for bootstrapping samplers with mixing times with polynomial dependence on a conditioning measure κ to mixing times with linear dependence on κ. The framework is heavily motivated by a perspective on *proximal point methods* in structured convex optimization as instances of optimizing composite objectives, and leverages this connection via a surprisingly simple analysis (cf. Theorem 10).

2. We develop novel "base samplers" for composite logconcave distributions and logconcave finite sums (cf. Theorems 11, 12). The former is the first composite sampler with stronger guarantees than those known in the general logconcave setting. The latter constitutes the first high-accuracy finite sum sampler whose gradient query complexity improves upon the naïve strategy of querying full gradients of the negative log-density in each iteration.

Using our novel base samplers within our reduction framework, we obtain state-of-the-art samplers for all of the aforementioned structured families, i.e. well-conditioned, composite, and finite sum, as Corollaries 10, 11, and 12. We emphasize that even without our reduction technique, the guarantees of our base samplers for composite and finite sum-structured densities are the first of their kind. However, by boosting their mixing via our reduction, we obtain guarantees for these structured distribution families which are essentially the best one can hope for without a significant improvement in the most commonly studied well-conditioned regime (cf. discussion in Section 5.1.1).

We now formally state our results in Section 5.1.1, and situate them in the literature in Section 5.1.2. Section 5.1.3 is a technical overview of our approaches for developing our base samplers for composite and finite sum-structured densities (Sections 5.1.3 and 5.1.3),

as well as our proximal reduction framework (Section 5.1.3). Finally, Section 5.1.5 gives a roadmap for the rest of the paper.

5.1.1 Our results

Before stating our results, we first require the notion of a restricted Gaussian oracle, whose definition is a key ingredient in giving our reduction framework as well as our later composite samplers.

Definition 1 (Restricted Gaussian oracle)**.** $\mathcal{O}(\lambda, v)$ *is a* restricted Gaussian oracle *(RGO) for convex* $g : \mathbb{R}^d \to \mathbb{R}$ *if it returns*

$$\mathcal{O}(\lambda, v) \leftarrow \text{sample from the distribution with density} \propto \exp\left(-\frac{1}{2\lambda}\|x - v\|_2^2 - g(x)\right).$$

In other words, an RGO asks to sample from a multivariate Gaussian (with covariance a multiple of the identity), "restricted" by some convex function g. Intuitively, if we can reduce a sampling problem for the density $\propto \exp(-g)$ to calling an RGO a small number of times with a small value of λ, each RGO subproblem could be much easier to solve than the original problem. This can happen for a variety of reasons, e.g. if the regularized density is extremely well-conditioned, or because it inherits concentration properties of a Gaussian. This idea of reducing a sampling problem to multiple subproblems, each implementing an RGO, underlies the framework of Theorem 10. Because the idea of regularization by a large Gaussian component repeatedly appears in this paper, we make the following more specific definition for convenience, which lower bounds the size of the Gaussian.

Definition 2 (η-RGO)**.** *We say* $\mathcal{O}(\lambda, v)$ *is an* η*-restricted Gaussian oracle (*η*-RGO) if it satisfies Definition 1 with the restriction that parameter* λ *is required to be always at most* η *in calls to* $\mathcal{O}$*.*

Variants of our notion of an RGO have implicitly appeared previously [CV18, MFWB19], and efficient RGO implementation was a key subroutine in the fastest sampling algorithm for general logconcave distributions [CV18]. It also extends a similar oracle used in composite optimization, which we will discuss shortly. However, the explicit use of RGOs in a framework such as Theorem 10 is a novel technical innovation of our work, and we believe this abstraction will find further uses.

Proximal reduction framework. In Section 5.3, we prove correctness of our proximal reduction framework, whose guarantees are stated in the following Theorem 10.

Theorem 10. *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-f_{\text{oracle}}(x))$ such that f_{oracle} is μ-strongly convex, and let $\epsilon \in (0,1)$. Let $\eta \le \frac{1}{\mu}$, $T = \Theta(\frac{1}{\eta\mu}\log\frac{d}{\eta\mu\epsilon})$ for some $\beta \ge 1$, and $\mathcal{O}$ be a η-RGO for f_{oracle}. Algorithm 11, initialized at the minimizer of f_{oracle}, runs in T iterations, each querying $\mathcal{O}$ a constant number of times, and obtains ϵ total variation distance to π.*

In other words, if we can implement an η-RGO for a μ-strongly convex function f_{oracle} in time $\mathcal{T}_{\text{RGO}}$, we can sample from $\exp(-f_{\text{oracle}})$ in time $\widetilde{O}(\frac{1}{\eta\mu}\cdot\mathcal{T}_{\text{RGO}})$. To highlight the power of this reduction framework, suppose there was an existing sampler $\mathcal{A}$ for densities $\propto \exp(-f)$ with mixing time $\widetilde{O}(\kappa^{10}\sqrt{d})$, where $f:\mathbb{R}^d\to\mathbb{R}$ is L-smooth, μ-strongly convex, and has condition number $\kappa = \frac{L}{\mu}$ (cf. Section 5.2.1 for definitions).[2] Choosing $\eta = \frac{1}{L}$ and $f_{\text{oracle}} \leftarrow f$ in Theorem 10 yields a sampler whose mixing time is $\widetilde{O}(\kappa\cdot\mathcal{T}_{\text{RGO}})$, where $\mathcal{T}_{\text{RGO}}$ is the cost of sampling from a density proportional to

$$\exp\left(-\frac{L}{2}\|x-v\|_2^2 - f(x)\right),$$

for some $v\in\mathbb{R}^d$. However, observe that this distribution has a negative log-density with constant condition number $\frac{L+L}{L+\mu} \le 2$! By using $\mathcal{A}$ as our RGO, we have $\mathcal{T}_{\text{RGO}} = \widetilde{O}(\sqrt{d})$, and the overall mixing time is $\widetilde{O}(\kappa\sqrt{d})$. Leveraging Theorem 10 in applications, we obtain the following new results, improving mixing of various "base samplers" which we bootstrap as RGOs for regularized densities.

Well-conditioned densities. In [LST20] (Chapter 3), it was shown that a variant of Metropolized Hamiltonian Monte Carlo obtains a mixing time of $\widetilde{O}(\kappa d\log^3\frac{\kappa d}{\epsilon})$ for sampling a density on $\mathbb{R}^d$ with condition number κ. The analysis of [LST20] was somewhat delicate, and required reasoning about conditioning on a nonconvex set with desirable concentration properties. In Section 5.4.1, we prove Corollary 10, improving [LST20] by roughly a logarithmic factor with a significantly simpler analysis.

Corollary 10. *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-f(x))$ such that f is L-smooth and μ-strongly convex, and let $\epsilon\in(0,1)$, $\kappa = \frac{L}{\mu}$. Assume access to $x^* = \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$. Algorithm 11 with $\eta = \frac{1}{8Ld\log(\kappa)}$ using Algorithm 6 as a restricted Gaussian oracle for f uses $O(\kappa d\log\kappa\log\frac{\kappa d}{\epsilon})$ gradient queries in expectation, and obtains ϵ total variation distance to π.*

[2] No sampler with mixing time scaling as $\text{poly}(\kappa)\sqrt{d}$ is currently known.

We include Corollary 10 as a warmup for our more complicated results, as a way to showcase the use of our reduction framework in a slightly different way than the one outlined earlier. In particular, in proving Corollary 10, we will choose a significantly smaller value of η, at which point a simple rejection sampling scheme implements each RGO with expected constant gradient queries.

We give another algorithm matching Corollary 10 with a deterministic query complexity bound as Corollary 14. The algorithm of Corollary 14 is interesting in that it is entirely a *zeroth-order* algorithm, and does not require access to a gradient oracle. To our knowledge, in the well-conditioned optimization setting, no zeroth-order query complexities better than roughly $\sqrt{\kappa}d$ are known, e.g. simulating accelerated gradient descent with a value oracle; thus, our sampling algorithm has a query bound off by only $\tilde{O}(\sqrt{\kappa})$ from the best-known optimization algorithm. We are hopeful this result may help in the search for query lower bounds for structured logconcave sampling.

Composite densities with a restricted Gaussian oracle. In Section 5.5, we develop a sampler for densities on $\mathbb{R}^d$ proportional to $\exp(-f(x) - g(x))$, where f has condition number κ and g admits a restricted Gaussian oracle $\mathcal{O}$. We state its guarantees here.

Theorem 11. *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp\left(-f(x) - g(x)\right)$ such that f is L-smooth and μ-strongly convex, and let $\epsilon \in (0, 1)$. Let $\eta \leq \frac{1}{32L\kappa d \log(\kappa/\epsilon)}$ (where $\kappa = \frac{L}{\mu}$), $T = \Theta(\frac{1}{\eta\mu}\log(\frac{\kappa d}{\epsilon}))$, and let $\mathcal{O}$ be a η-RGO for g. Further, assume access to the minimizer $x^* = \operatorname{argmin}_{x\in\mathbb{R}^d}\{f(x) + g(x)\}$. There is an algorithm which runs in T iterations in expectation, each querying a gradient oracle of f and $\mathcal{O}$ a constant number of times, and obtains ϵ total variation distance to π.*

The assumption that the composite component g admits an RGO can be thought of as a measure of "simplicity" of g. This mirrors the widespread use of a proximal oracle as a measure of simplicity in the context of composite optimization [BT09], which we now define.

Definition 3 (Proximal oracle)**.** *$\mathcal{O}(\lambda, v)$ is a* proximal oracle *for convex $g : \mathbb{R}^d \to \mathbb{R}$ if it returns*

$$\mathcal{O}(\lambda, v) \leftarrow \operatorname{argmin}_{x\in\mathbb{R}^d}\left\{\frac{1}{2\lambda}\|x - v\|_2^2 + g(x)\right\}.$$

Many regularizers g in defining composite optimization objectives, which are often used to enforce a quality such as sparsity or "simplicity" in a solution, admit efficient proximal

oracles. In particular, if the proximal oracle subproblem admits a closed form solution (or otherwise is computable in $O(d)$ time), the regularized objective can be optimized at essentially no asymptotic loss. It is readily apparent that our RGO (Definition 1) is the extension of Definition 3 to the sampling setting. In [MFWB19], a variety of regularizations arising in practical applications including coordinate-separable g (such as restrictions to a coordinate-wise box, e.g. for a Bayesian inference task where we have side information on the ranges of parameters) and ℓ_1 or group Lasso regularized densities were shown to admit RGOs. Our composite sampling results achieve a similar "no loss" phenomenon for such regularizations, with respect to existing well-conditioned samplers.

By choosing the largest possible value of η in Theorem 11, we obtain an iteration bound of $\widetilde{O}(\kappa^2 d)$. In Section 5.4.2, we prove Corollary 11, which improves Theorem 11 by roughly a κ factor.

Corollary 11. *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-f(x) - g(x))$ such that f is L-smooth and μ-strongly convex, and let $\epsilon \in (0,1)$, $\kappa = \frac{L}{\mu}$. Assume access to $x^* = \operatorname{argmin}_{x\in\mathbb{R}^d}\{f(x) + g(x)\}$ and let $\mathcal{O}$ be a restricted Gaussian oracle for g. There is an algorithm (Algorithm 11 using Theorem 11 as a restricted Gaussian oracle) which runs in $O(\kappa d \log^3 \frac{\kappa d}{\epsilon})$ iterations in expectation, each querying a gradient of f and $\mathcal{O}$ a constant number of times, and obtains ϵ total variation distance to π.*

To sketch the proof, choosing $\eta = \frac{1}{L}$ in Theorem 10 yields an algorithm running in $\widetilde{O}(\frac{1}{\eta\mu}) = \widetilde{O}(\kappa)$ iterations. In each iteration, the RGO subproblem asks to sample from the distribution whose negative log-density is $f(x) + g(x) + \frac{L}{2}\|x - v\|_2^2$ for some $v \in \mathbb{R}^d$, so we can call Theorem 11, where the "well-conditioned" portion $f(x) + \frac{L}{2}\|x - v\|_2^2$ has constant condition number. Thus, Theorem 11 runs in $\widetilde{O}(d)$ iterations to solve the subproblem, yielding the result. In fact, Corollary 11 nearly matches Corollary 10 in the case $g = 0$ uniformly. Surprisingly, this recovers the runtime of [LST20] without appealing to strong gradient concentration bounds (e.g. [LST20], Theorem 3.2).

Logconcave finite sums. In Section 5.6, we initiate the study of mixing times for sampling logconcave finite sums with polylogarithmic dependence on accuracy. We give the following result.

Theorem 12. *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-F(x))$, where $F(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ is μ-strongly convex, f_i is L-smooth and convex $\forall i \in [n]$, $\kappa = \frac{L}{\mu}$, and $\epsilon \in$*

$(0,1)$. *Assume access to* $x^* = \operatorname{argmin}_{x\in\mathbb{R}^d} F(x)$. *Algorithm 10 uses* $O\left(\kappa^2 d \log^4 \frac{n\kappa d}{\epsilon}\right)$ *value queries to summands* $\{f_i\}_{i\in[n]}$, *and obtains* ϵ *total variation distance to* π.

For a zeroth-order algorithm, Theorem 12 serves as a surprisingly strong baseline as it nearly matches the previously best-known bound for zeroth-order well-conditioned sampling when $n = 1$; however, when e.g. $\kappa \approx d$, the complexity bound is at least cubic. By using Theorem 12 within the framework of Theorem 10, we obtain the following improved result.

Corollary 12 (Improved first-order logconcave finite sum sampling)**.** *In the setting of Theorem 12, Algorithm 11 using Algorithm 10 and SVRG [JZ13] as a restricted Gaussian oracle for* F *uses*

$$O\left(n \log\left(\frac{n\kappa d}{\epsilon}\right) + \kappa\sqrt{nd}\log^{3.5}\left(\frac{n\kappa d}{\epsilon}\right) + \kappa d \log^5\left(\frac{n\kappa d}{\epsilon}\right)\right) = \widetilde{O}\left(n + \kappa \max\left(d, \sqrt{nd}\right)\right)$$

queries to first-order oracles for summands $\{f_i\}_{i\in[n]}$, *and obtains* ϵ *total variation distance to* π.

Corollary 12 has several surprising properties. First, its bound when $n = 1$ gives yet another way of (up to polylogarithmic factors) recovering the runtime of [LST20] without gradient concentration. Second, up to a $\widetilde{O}(\max(1, \sqrt{\frac{n}{d}}))$ factor, it is essentially the best runtime one could hope for without an improvement when $n = 1$. This is in the sense that $\widetilde{O}(\kappa d)$ is the best runtime for $n = 1$, and to our knowledge every efficient well-conditioned sampler requires minimizer access, i.e. $\widetilde{O}(n)$ gradient queries [WS16]. Interestingly, when $n = 1$, Algorithm 10 can be significantly simplified, and becomes the standard Metropolized random walk [DCWY19]; this yields Corollary 14, an algorithm attaining the iteration complexity of Corollary 10 while only querying a value oracle for f.

5.1.2 Previous work

Composite densities. Recent works have studied sampling from densities of the form (5.1), or its specializations (e.g. restrictions to a convex set). Several [Per16, BDMP17, Ber18] are based on Moreau envelope or proximal regularization strategies, and demonstrate efficiency under more stringent assumptions on the structure of the composite function g, but under minimal assumptions obtain fairly large provable mixing times $\Omega(d^5)$. Proximal regularization algorithms have also been considered for non-composite sampling

[Wib19]. Another discretization strategy based on projections was studied by [BEL18], but obtained mixing time $\Omega(d^7)$. Finally, improved algorithms for special constrained sampling problems have been proposed, such as simplex restrictions [HKRC18].

Of particular relevance and inspiration to this work is [MFWB19]. By generalizing and adapting Metropolized HMC algorithms of [DCWY19, CDWY20], adopting a Moreau envelope strategy, and using (a stronger version of) the RGO access model, [MFWB19] obtained a runtime which in the best case scales as $\tilde{O}\left(\kappa^2 d\right)$, similar to the guarantee of our base sampler in Theorem 11. However, this result required a variety of additional assumptions, such as access to the normalization factor of restricted Gaussians, Lipschitzness of g, warmness of the start, and various problem parameter tradeoffs. The general problem of sampling from (5.1) under minimal assumptions more efficiently than general-purpose logconcave algorithms is to the best of our knowledge unresolved (even under restricted Gaussian oracle access), a novel contribution of our mixing time bound. Our results also suggest that the RGO is a natural notion of tractability for the composite sampling problem.

Logconcave finite sums. Since [WT11] proposed the stochastic gradient Langevin dynamics, which at each step stochastically estimates the full gradient of the function, there has been a long line of work giving bounds for this method and other similar algorithms [DK19, GGZ18, SKR19, BCM$^+$18, NF19]. These convergence rates depend heavily on the variance of the stochastic estimates. Inspired by variance-reduced methods in convex optimization, samplers based on low-variance estimators have also been proposed [DRW$^+$16, DSM$^+$16, BFR$^+$19, BFFN19, NDH$^+$17, CWZ$^+$17, ZXG18, CFM$^+$18]. Although our reduction-based approach is not designed specifically for solving problems of finite sum structure, our speedup can be viewed as due to a lower variance estimator implicitly defined through the oracle subproblems of Theorem 10 via repeated re-centering.

In Table 5.1, we list prior runtimes [ZXG18, CFM$^+$18] for sampling logconcave finite sums; note these results additionally require bounded higher derivatives (with the exception of the κ^4 dependence), obtain guarantees only in Wasserstein distance, and have polynomial dependences on ϵ^{-1}. On the other hand, our reduction-based approach obtains total variation bounds with linear dependence on κ and polylogarithmic dependence on ϵ^{-1}. Our bound also simultaneously matches the state-of-the-art bound when $n = 1$, a fea-

Method	Gradient oracle complexity	
	$W_2 \le \epsilon$, $\mu = 1$	$W_2 \le \epsilon\sqrt{d\mu^{-1}}$
SAGA-LD [CFM+18]	$n + \frac{\kappa^{1.5}\sqrt{d}+\kappa d+Md}{\epsilon} + \frac{\kappa d^{4/3}}{\epsilon^{2/3}}$	$n + \frac{\kappa^{1.5}+\kappa\sqrt{d}+M\sqrt{d}}{\epsilon} + \frac{\kappa d^{2/3}}{\epsilon^{2/3}}$
SVRG-LD [CFM+18]	$n + \frac{\kappa^{1.5}\sqrt{d}+\kappa d+Md}{\epsilon} + \frac{\kappa d^{4/3}}{\epsilon^{2/3}}$	$n + \frac{\kappa^3}{\epsilon^2} + \frac{\kappa^{1.5}+M\sqrt{d}}{\epsilon}$
CV-ULD [CFM+18]	$n + \frac{\kappa^4 d^{1.5}}{\epsilon^3}$	$n + \frac{\kappa^4}{\epsilon^3}$
SVRG-LD [ZXG18]	$n + \frac{\kappa^{1.5}\sqrt{d}+Md}{\epsilon} + \frac{\kappa\sqrt{nd}}{\epsilon}$	$n + \frac{\kappa^{1.5}+M\sqrt{d}}{\epsilon} + \frac{\kappa\sqrt{n}}{\epsilon}$
State-of-the-art, $n = 1$ [SL19]	$\frac{\kappa^{7/6}d^{1/6}}{\epsilon^{1/3}} + \frac{\kappa d^{1/3}}{\epsilon^{2/3}}$	$\frac{\kappa^{7/6}}{\epsilon^{1/3}} + \frac{\kappa}{\epsilon^{2/3}}$

Method	Gradient oracle complexity (TV $\le \epsilon$)
Corollary 12	$n + \kappa d + \kappa\sqrt{nd}$
State-of-the-art, $n = 1$ [LST20]	κd

Table 5.1: Complexity of sampling from $e^{-F(x)}$ where $F(x) = \frac{1}{n}\sum_{i\in[n]} f_i(x)$ on $\mathbb{R}^d$ is μ-strongly convex, each f_i is convex and L-smooth, and $\kappa = \frac{L}{\mu}$. For relevant lines, M is the Lipschitz constant of the Hessian $\nabla^2 F$, which our algorithm has no dependence on. Complexity is measured in terms of the number of calls to f_i or ∇f_i for summands $\{f_i\}_{i\in[n]}$. We hide polylog$(\frac{n\kappa d}{\epsilon})$ factors for simplicity.

ture not shared by prior stochastic algorithms. To our knowledge, no previous nontrivial[3] bounds were known in the high-accuracy regime before our work.

5.1.3 Technical overview

Composite logconcave sampling

We study the problem of approximately sampling from a distribution π on $\mathbb{R}^d$, with density

$$\frac{d\pi(x)}{dx} \propto \exp\left(-f(x) - g(x)\right). \tag{5.1}$$

Here, $f : \mathbb{R}^d \to \mathbb{R}$ is assumed to be "well-behaved" (i.e. has finite condition number), and $g : \mathbb{R}^d \to \mathbb{R}$ is a convex, but possibly non-smooth function. This problem generalizes the special case of sampling from $\exp(-f(x))$ for well-conditioned f, simply by letting g vanish. Even the specialization of (5.1) where g indicates a convex set (i.e. is 0 inside the set, and ∞ outside) is not well-understood; existing mixing time bounds for this

[3] As mentioned previously, one can always compute the full ∇F in every iteration in a well-conditioned sampler.

restricted setting are large polynomials in d [BDMP17, BEL18], and are typically weaker than guarantees in the general logconcave setting [LV06a, LV06b]. This is in contrast to the convex optimization setting, where first-order methods readily generalize to solve problem families such as $\min_{x\in\mathcal{X}} f(x)$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex set, as well as its generalization

$$\min_{x\in\mathbb{R}^d} f(x) + g(x), \text{ where } g : \mathbb{R}^d \to \mathbb{R} \text{ is convex and admits a proximal oracle.} \tag{5.2}$$

We defined proximal oracles in Definition 3; in short, they are prodecures which minimize the sum of a quadratic and g. Definition 3 is desirable as many natural non-smooth composite objectives arising in learning settings, such as the Lasso [Tib96] and elastic net [ZH05], admit efficient proximal oracles. It is clear that the definition of a proximal oracle implies it can also handle arbitrary sums of linear functions and quadratics, as the resulting function can be rewritten as the sum of a constant and a single quadratic. The seminal work [BT09] extends fast gradient methods to solve (5.2) via proximal oracles, and has prompted many follow-up studies.

Motivated by the success of the proximal oracle framework in convex optimization, we study sampling from the family (5.1) through the lens of RGOs, a natural extension of Definition 3. The main result of Section 5.5 is a "base" algorithm efficiently sampling from (5.1), assuming access to an RGO for g. We now survey the main components of this algorithm.

Reduction to shared minimizers. We first observe that without loss of generality, f and g share a minimizer: by shifting f and g by linear terms, i.e. $\tilde{f}(x) \stackrel{\text{def}}{=} f(x) - \langle \nabla f(x^*), x\rangle$, $\tilde{g}(x) \stackrel{\text{def}}{=} g(x) + \langle \nabla f(x^*), x\rangle$, where x^* minimizes $f + g$, first-order optimality implies both $\tilde{f}$ and $\tilde{g}$ are minimized by x^*. Moreover, implementation of a first-order oracle for $\tilde{f}$ and an RGO for $\tilde{g}$ are immediate without additional assumptions. This modification becomes crucial for our later developments, and we hope this simple observation, reminiscent of "variance reduction" techniques in stochastic optimization [JZ13], is broadly applicable to improving algorithms for the sampling problem induced by (5.1).

Beyond Moreau envelopes: expanding the space. A typical approach in convex optimization in handling non-smooth objectives g is to instead optimize its *Moreau envelope*, defined by

$$g^\eta(y) \stackrel{\text{def}}{=} \min_{x\in\mathbb{R}^d} \left\{ g(x) + \frac{1}{2\eta}\|x - y\|_2^2 \right\}. \tag{5.3}$$

Intuitively, the envelope g^η trades off function value with proximity to y; a standard exercise shows that g^η is smooth (has a Lipschitz gradient), with smoothness depending

on η, and moreover that computing gradients of g^η reduces to calling a proximal oracle (Definition 3). It is natural to extend this idea to the composite sampling setting, e.g. via sampling from the density

$$\exp\left(-f(x) - g^\eta(x)\right).$$

However, a variety of complications prevent such strategies from obtaining rates comparable to their non-composite, well-conditioned counterparts, including difficulty in bounding closeness of the resulting distribution, as well as biased drifts of the sampling process due to error in gradients.

Our approach departs from this smoothing strategy in a crucial way, inspired by Hamiltonian Monte Carlo (HMC) methods [Kra40, Nea11]. HMC can be seen as a discretization of the ubiquitous Langevin dynamics, on an expanded space. In particular, discretizations of Langevin dynamics simulate the stochastic differential equation $\frac{dx_t}{dt} = -\nabla f(x_t) + \sqrt{2}\frac{dW_t}{dt}$, where W_t is Brownian motion. HMC methods instead simulate dynamics on an extended space $\mathbb{R}^d \times \mathbb{R}^d$, via an auxiliary "velocity" variable which accumulates gradient information. This is sometimes interpreted as a discretization of the underdamped Langevin dynamics [CCBJ17]. HMC often has desirable stability properties, and expanding the dimension via an auxiliary variable has been used in algorithms obtaining the fastest rates in the well-conditioned logconcave sampling regime [SL19, LST20]. Inspired by this phenomenon, we consider the density on $\mathbb{R}^d \times \mathbb{R}^d$

$$\frac{d\hat{\pi}}{dz}(z) \stackrel{\text{def}}{=} \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|x - y\|_2^2\right) \text{ where } z = (x, y). \tag{5.4}$$

Due to technical reasons, the family of distributions we use in our final algorithms are of slightly different form than (5.4), but this simplification is useful to build intuition. Note in particular that the form of (5.4) is directly inspired by (5.3), where rather than maximizing over x, we directly expand the space. The idea is that for small enough η and a set on x of large measure, smoothness of f will guarantee that the marginal of (5.4) on x will concentrate y near x, a fact we make rigorous. To sample from (5.1), we then show that a rejection filter applied to a sample x from the marginal of (5.4) will terminate in constant steps. Consequently, it suffices to develop a fast sampler for (5.4).

Alternating sampling with an oracle. The form of the distribution (5.4) suggests a natural strategy for sampling from it: starting from a current state (x_k, y_k), we iterate

1. Sample $y_{k+1} \sim \exp\left(-f(y) - \frac{1}{2\eta}\|x_k - y\|_2^2\right)$.

2. Sample $x_{k+1} \sim \exp\left(-g(x) - \frac{1}{2\eta}\|x - y_{k+1}\|_2^2\right)$, via a restricted Gaussian oracle.

When f and g share a minimizer, taking a first-order approximation in the first step, i.e. sampling $y_{k+1} \sim \exp(-f(x_k) - \langle \nabla f(x_k), y - x_k\rangle - \frac{1}{2\eta}\|y - x_k\|_2^2)$, can be shown to generalize the `Leapfrog` step of HMC updates. However, for η very small (as in our setting), we observe the first step itself reduces to the case of sampling from a distribution with constant condition number, performable in $\tilde{O}(d)$ gradient calls by e.g. Metropolized HMC [DCWY19, CDWY20, LST20]. Moreover, it is not hard to see that this "alternating marginal" sampling strategy preserves the stationary distribution exactly, so no filtering is necessary. Directly bounding the conductance of this random walk, for small enough η, leads to an algorithm running in $\tilde{O}\left(\kappa^2 d^2\right)$ iterations, each calling an RGO once, and a gradient oracle for f roughly $\tilde{O}\left(d\right)$ times. This latter guarantee is by an appeal to known bounds [CDWY20, LST20] on the mixing time in high dimensions of Metropolized HMC for a well-conditioned distribution, a property satisfied by the y-marginal of (5.4) for small η.

Stability of Gaussians under bounded perturbations. To obtain our tightest runtime result, we use that η is chosen to be much smaller than L^{-1} to show structural results about distributions of the form (5.4), yielding tighter concentration for bounded perturbations of a Gaussian (i.e. the Gaussian has covariance $\frac{1}{\eta}\,\mathrm{id}$, and is restricted by L-smooth f for $\eta \ll L^{-1}$). To illustrate, let

$$\frac{d\mathcal{P}_x(y)}{dy} \propto \exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right)$$

and let its mean and mode be $\bar{y}_x$, y_x^*. It is standard that $\|\bar{y}_x - y_x^*\|_2 \le \sqrt{d\eta}$, by η^{-1}-strong logconcavity of $\mathcal{P}_x$. Informally, we show that for $\eta \ll L^{-1}$ and x not too far from the minimizer of f, we can improve this to $\|\bar{y}_x - y_x^*\|_2 = O(\sqrt{\eta})$; see Proposition 22 for a precise statement.

Using our structural results, we sharpen conductance bounds, improve the warmness of a starting distribution, and develop a simple rejection sampling scheme for sampling the y variable in expected constant gradient queries. Our proofs are continuous in flavor and based on gradually perturbing the Gaussian and solving a differential inequality; we believe they may of independent interest. These improvements lead to an algorithm running in $\tilde{O}\left(\kappa^2 d\right)$ iterations; ultimately, we use our reduction framework, stated in Theorem 10, to improve this dependence to $\tilde{O}\left(\kappa d\right)$.

Logconcave finite sums

We initiate the algorithmic study of the following task in the high-accuracy regime: sample $x \sim \pi$ within total variation distance ϵ, where $\frac{d\pi}{dx}(x) \propto \exp(-F(x))$ and

$$F(x) = \frac{1}{n} \sum_{i \in [n]} f_i(x), \tag{5.5}$$

all $f_i : \mathbb{R}^d \to \mathbb{R}$ are convex and L-smooth, and F is μ-strongly convex. We call such a distribution π a (well-conditioned) *logconcave finite sum*.

In applications (where summands correspond to points in a dataset, e.g. in Bayesian linear and logistic regression tasks [DCWY19]), querying ∇F may be prohibitively expensive, so a natural goal is to obtain bounds on the number of required queries to summands ∇f_i for $i \in [n]$. This motivation also underlies the development of stochastic gradient methods in optimization, a foundational tool in modern statistics and data processing. Naïvely, one can complete the task by using existing samplers for well-conditioned distributions and querying the full gradient ∇F in each iteration, resulting in a summand gradient query complexity of $\widetilde{O}(n\kappa d)$ [LST20]. Many recent works, inspired from recent developments in the complexity of optimizing a well-conditioned finite sum, have developed subsampled gradient methods for the sampling problem. However, to our knowledge, all such guarantees depend polynomially on the accuracy ϵ and are measured in the 2-Wasserstein distance; in the high-accuracy, total variation case no nontrivial query complexity is currently known.

We show in Section 5.6 that given access to the minimizer of F, a simple zeroth-order algorithm which queries $\widetilde{O}(\kappa^2 d)$ values of summands $\{f_i\}_{i \in [n]}$ succeeds (i.e. it never requires a full value or gradient query of F). The algorithm is essentially the Metropolized random walk proposed in [DCWY19] for the $n = 1$ case with a cheaper subsampled filter step. Notably, because the random walk is conducted with respect to F, we cannot efficiently query the function value at any point; nonetheless, by randomly sampling to compute a nearly-unbiased estimator of the rejection probability, we do not incur too much error. This random walk was shown in [CDWY20] to mix in $\widetilde{O}(\kappa^2 d)$ iterations; we implement each step to sufficient accuracy using $\widetilde{O}(1)$ function evaluations.

It is natural to ask if first-order information can be used to improve this query complexity, perhaps through "variance reduction" techniques (e.g. [JZ13]) developed for stochastic optimization. The idea behind variance reduction is to recenter gradient estimates in phases, occasionally computing full gradients to improve the estimate quality. One fundamental difficulty which arises from using variance reduction in high-accuracy sampling is

that the resulting algorithms are not *stateless*. By this, we mean that the variance-reduced estimates depend on the history of the algorithm, and thus it is difficult to ascertain correctness of the stationary distribution. We take a different approach to achieve a linear query dependence on the conditioning κ, described in the following section.

Proximal point reduction framework

To motivate Theorem 10, we first recast existing "proximal point" reduction-based optimization methods through the lens of composite optimization, and subsequently show that similar ideas underlying our composite sampler in Section 5.1.3 yield an analagous "proximal point reduction framework" for sampling. We hope these insights prove fruitful for further development of proximal approaches to sampling.

Proximal point methods as composite optimization. Proximal point methods are a well-studied primitive in optimization, developed by [Roc76]; cf. [PB14] for a modern survey. The principal idea is that to minimize convex $F : \mathbb{R}^d \to \mathbb{R}$, it suffices to solve a sequence of subproblems

$$x_{k+1} \leftarrow \text{argmin}_{x \in \mathbb{R}^d} \left\{ F(x) + \frac{1}{2\lambda} \|x - x_k\|_2^2 \right\}. \tag{5.6}$$

Intuitively, by tuning the parameter $\lambda \geq 0$, we trade off how regularized the subproblems (5.6) are with how rapidly the overall method converges. Smaller values of λ result in larger regularization amounts which are amenable to algorithms for minimizing well-conditioned objectives.

For optimizing functions of the form (5.5) via stochastic gradient estimates to ϵ error, [JZ13, DBL14, SRB17] developed variance-reduced methods obtaining a query complexity of $\widetilde{O}(n + \kappa)$. To match a known lower bound of $\widetilde{O}(n + \sqrt{n\kappa})$ due to [WS16], two works [LMH15, FGKS15] appropriately applied instances of accelerated proximal point methods [Gul92] with careful analyses of how accurately subproblems (5.6) needed to be solved. These algorithms black-box called the $\widetilde{O}(n + \kappa)$ runtime as an oracle to solve the subproblems (5.6) for an appropriate choice of λ, obtaining an accelerated rate.[4] To shed some light on this acceleration procedure, we adopt an alternative view on proximal point methods.[5] Consider the following known composite optimization result.

[4] We note that an improved runtime without extraneous logarithmic factors was later obtained by [All17].

[5] This perspective can also be found in the lecture notes [Lee18].

Proposition 6 (Informal statement of [BT09])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-strongly convex, and $g : \mathbb{R}^d \to \mathbb{R}$ admit a proximal oracle $\mathcal{O}(\lambda, v)$ (cf. Definition 3). There is an algorithm taking $\widetilde{O}(\sqrt{\kappa})$ iterations for $\kappa = \frac{L}{\mu}$ to find an ϵ-approximate minimizer to $f + g$, each querying ∇f and $\mathcal{O}$ a constant number of times. Further, $\lambda = \frac{1}{L}$ in all calls to $\mathcal{O}$.*

Ignoring subtleties of the error tolerance of $\mathcal{O}$, we show how to use an instance of Proposition 6 to recover the $\widetilde{O}(n + \sqrt{n\kappa})$ query complexity for optimizing (5.5). Let $f(x) = \frac{\mu}{2}\|x\|_2^2$, and $g = F - f$. For any $\Lambda \geq \mu$, f is both μ-strongly convex and Λ-smooth. Moreover, note that all calls to the proximal oracle $\mathcal{O}$ for g require solving subproblems of the form

$$\text{argmin}_{x \in \mathbb{R}^d} \left\{ F(x) - \frac{\mu}{2}\|x\|_2^2 + \frac{\Lambda}{2}\|x - v\|_2^2 \right\}. \tag{5.7}$$

The upshot of choosing a smoothness bound $\Lambda \geq \mu$ is that the regularization amount in (5.7) increases, improving the conditioning of the subproblem, which is Λ-strongly convex and $L+\Lambda$-smooth. The algorithm of e.g. [JZ13] solves each subproblem (5.7) in $\widetilde{O}(n+\frac{L+\Lambda}{\Lambda})$ gradient queries, leading to an overall query complexity (for Proposition 6) of

$$\widetilde{O}\left(\sqrt{\frac{\Lambda}{\mu}} \cdot \left(n + \frac{L}{\Lambda}\right)\right).$$

Optimizing over $\Lambda \geq \mu$, i.e. taking $\Lambda = \max(\mu, \frac{L}{n})$, yields the desired bound of $\widetilde{O}(n+\sqrt{n\kappa})$.

Applications to sampling. In Sections 5.5 and 5.6, we develop samplers for structured families with quadratic dependence on the conditioning κ. The proximal point approach suggests a strategy for accelerating these runtimes. Namely, if there is a framework which repeatedly calls a sampler for a regularized density (analogous to calls to (5.6)), one could trade off the regularization with the rate of the outer loop. Fortunately, in the spirit of interpreting proximal point methods as composite optimization, the composite sampler of Section 5.5 itself meets these reduction framework criteria.

We briefly recall properties of our composite sampler here. Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-f_{\text{wc}}(x) - f_{\text{oracle}}(x))$,[6] where f_{wc} is well-conditioned (has finite condition number κ) and f_{oracle} admits an RGO, which solves subproblems of the form

$$\mathcal{O}(\eta, v) \sim \text{the density proportional to } \exp\left(-\frac{1}{2\eta}\|x - v\|_2^2 - f_{\text{oracle}}(x)\right). \tag{5.8}$$

[6] To disambiguate, we sometimes also use the notation $f_{\text{wc}} + f_{\text{oracle}}$ rather than $f+g$ in defining instances of our reduction framework or composite samplers, when convenient in the context.

The algorithm of Section 5.5 only calls $\mathcal{O}$ with a fixed η; note the strong parallel between the RGO subproblem and the proximal oracle of Proposition 6. For a given value of $\eta \geq 0$, our composite sampler runs in $\widetilde{O}(\frac{1}{\eta\mu})$ iterations, each requiring a call to $\mathcal{O}$. Smaller η improve the conditioning of the negative log-density of subproblem (5.8), but increase the overall iteration count, yielding a range of trade-offs. The algorithm of Section 5.5 has an upper bound requirement on η (cf. Theorem 11); in Section 5.3, we observe that this may be lifted when $f_{\text{wc}} = 0$ uniformly, allowing for a full range of choices. Moreover, the analysis of the composite sampler becomes much simpler when $f_{\text{wc}} = 0$, as in Theorem 10. We give the framework as Algorithm 11, as well as a full (fairly short) convergence analysis. By trading off the regularization amount with the cost of implementing (5.8) via bootstrapping base samplers, we obtain a host of improved runtimes.

Beyond our specific applications, the framework we provide has strong implications as a generic reduction from mixing times scaling polynomially in κ to improved methods scaling linearly in κ. This is akin to the observation in [LMH15] that accelerated proximal point methods generically improve poly(κ) dependences to $\sqrt{\kappa}$ dependences for optimization. We are optimistic this insight will find further implications in the logconcave sampling literature.

5.1.4 Erratum, and a word of warning for $o(d)$ mixing

The initial version of this paper, presented at COLT 2021, had an incorrect proof of Theorem 10. This was due to our reliance on the average conductance ("spectral profile") technique of [LK99] for bounding mixing. Roughly speaking, the mistake was caused by a misunderstanding that for stationary measures satisfying μ-log isoperimetry (for example, μ-strongly logconcave densities) and with transition distributions of Δ-close points having constant overlap, [LK99] provides mixing time bounds of the form (where β is a warmness parameter of the starting distribution)

$$\int_{\frac{1}{\beta}}^{\frac{1}{2}} \frac{1}{s\Phi(s)^2} ds \lesssim \frac{1}{\mu\Delta^2} \int_{\frac{1}{\beta}}^{\frac{1}{2}} \frac{1}{s\log(s)} ds \approx \frac{1}{\mu\Delta^2} \log\log\beta. \tag{5.9}$$

Here, $\Phi(s)$ is the s-conductance of the Markov chain, which can typically be lower bounded by $\Omega(\sqrt{\mu\log(s)}\Delta)$ under a stationary density exhibiting log-isoperimetry. However, the trivial bound $\Phi(s) \leq 1$ demonstrates that there is an additive $\log(\beta)$ term in (5.9). This is a bottleneck towards mixing times scaling as $o(d)$ for distributions where only an $\exp(\Omega(d))$-warm start is feasible; in particular, the conductance actually scales as

$\min(1, \Omega(\sqrt{\mu \log(s)}\Delta))$, causing this additive term. In settings where $\mu\Delta^2 \geq d^{-1}$ (such as our reductions, where this term often scales as a condition number of the problem), this additive term $\log(\beta) = \Omega(d)$ may dominate. This observation (and the fix) came out of conversations with Sinho Chewi; we are immensely greatful for his help.

For the particular structure of the algorithm in Theorem 10, we are able to give an alternative analysis going through W_2 convergence bounds, preserving the correctness of the theorem. However, this bottleneck is a general phenomenon which may cause future attempts to use Metropolized algorithms from exponentially warm starts to be stuck at $\Omega(d)$ iterations, which merits further investigation. We write this section as a word of warning to future researchers aiming at sublinear dimension dependences in Metropolized algorithms, and as a suggested open research direction.

5.1.5 Roadmap

We give notations and preliminaries in Section 6.2. In Section 5.3 we give our framework for bootstrapping fast regularized samplers, and prove its correctness (Theorem 10). Assuming the "base samplers" of Theorems 11 and 12, in Section 5.4 we apply our reduction to obtain all of our strongest guarantees, namely Corollaries 10, 11, and 12. We then prove Theorems 11 and 12 in Sections 5.5 and 5.6.

5.2 Preliminaries

5.2.1 Notation

General notation. For $d \in \mathbf{N}$, $[d]$ refers to the set of naturals $1 \leq i \leq d$; $\|\cdot\|_2$ is the Euclidean norm on $\mathbb{R}^d$ when d is clear from context. $\mathcal{N}(\mu, \mathbf{\Sigma})$ is the multivariate Gaussian of specified mean and variance, id is the identity of appropriate dimension when clear from context, and $\preceq$ is the Loewner order on symmetric matrices.

Functions. We say twice-differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and μ-strongly convex if $\mu \,\mathrm{id} \preceq \nabla^2 f(x) \preceq L \,\mathrm{id}$ for all $x \in \mathbb{R}^d$; it is well-known that L-smoothness implies that f has an L-Lipschitz gradient, and that for any $x, y \in \mathbb{R}^d$,

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2 \leq f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2.$$

If f is L-smooth and μ-strongly convex, we say it has a condition number $\kappa \stackrel{\text{def}}{=} \frac{L}{\mu}$. We call a zeroth-order oracle, or "value oracle", an oracle which returns $f(x)$ on any input

point $x \in \mathbb{R}^d$; similarly, a first-order oracle, or "gradient oracle", returns both the value and gradient $(f(x), \nabla f(x))$.

Distributions. We call distribution π on $\mathbb{R}^d$ logconcave if $\frac{d\pi}{dx}(x) = \exp(-f(x))$ for convex f; π is μ-strongly logconcave if f is μ-strongly convex. For $A \subseteq \mathbb{R}^d$, A^c is its complement, and we let $\pi(A) \stackrel{\text{def}}{=} \int_{x \in A} d\pi(x)$. We say distribution ρ is β-warm with respect to π if $\frac{d\pi}{d\rho}(x) \le \beta$ everywhere, and define the total variation $\|\pi - \rho\|_{\text{TV}} \stackrel{\text{def}}{=} \sup_{A \subseteq \mathbb{R}^d} \pi(A) - \rho(A)$. We will frequently use the fact that $\|\pi - \rho\|_{\text{TV}}$ is also the probability that $x \sim \pi$ and $x' \sim \rho$ are unequal under the best coupling of (π, ρ); this allows us to "locally share randomness" when comparing two random walk procedures. We define the expectation $\mathbb{E}_\pi$ and variance Var_π with respect to distribution π in the standard way,

$$\mathbb{E}_\pi[h(x)] \stackrel{\text{def}}{=} \int h(x) d\pi(x), \ \text{Var}_\pi[h(x)] \stackrel{\text{def}}{=} \mathbb{E}_\pi\left[(h(x))^2\right] - (\mathbb{E}_\pi[h(x)])^2 .$$

Structured distributions. This work considers two types of distributions with additional structure, which we call *composite logconcave densities* and *logconcave finite sums*. A composite logconcave density has the form $\exp(-f(x) - g(x))$, where both f and g are convex. In this context throughout, f will either be uniformly 0 or have a finite condition number (be "well-conditioned"), and g will represent a "simple" but possibly non-smooth function, as measured by admitting an RGO (cf. Definition 1). We will sometimes refer to the components as f and g as f_{wc} and f_{oracle} respectively, to disambiguate when the functions f and g are already defined in context. In our reduction-based approaches, we have additional structure on the parameter λ which an RGO is called with (cf. Definition 2). Specifically, in our instances typically $\lambda^{-1} \gg L$ (or some other "niceness" parameter associated with the negative log-density); this can be seen as heavily regularizing the negative log-density, and often makes the implementation simpler.

Finally, a logconcave finite sum has density of the form $\exp(-F(x))$ where $F(x) = \frac{1}{n}\sum_{i \in [n]} f_i(x)$. When treating such densities, we make the assumption that each constituent summand f_i is L-smooth and convex, and the overall function F is μ-strongly convex. We measure complexity of algorithms for logconcave finite sums by gradient queries to summands, i.e. $\nabla f_i(x)$ for some $i \in [n]$.

Optimization. Throughout this work, we are somewhat liberal with assuming access to minimizers to various functions (namely, the negative log-densities of target distributions). We give a more thorough discussion of this assumption in Appendix D.1, but

note here that for all function families we consider (well-conditioned, composite, and finite sum), efficient first-order methods exist for obtaining high accuracy minimizers, and this optimization query complexity is never the leading-order term in any of our algorithms assuming polynomially bounded initial error.

5.2.2 Technical facts

We will repeatedly use the following results.

Fact 5 (Gaussian integral)**.** *For any $\lambda \geq 0$ and $v \in \mathbb{R}^d$,*

$$\int \exp\left(-\frac{1}{2\lambda}\|x - v\|_2^2\right) dx = (2\pi\lambda)^{\frac{d}{2}}.$$

Fact 6 ([DCWY19], Lemma 1)**.** *Let π be a μ-strongly logconcave distribution, and let x^* minimize its negative log-density. Then, for $x \sim \pi$ and any $\delta \in [0, 1]$, with probability at least $1 - \delta$,*

$$\|x - x^*\|_2 \leq \sqrt{\frac{d}{\mu}}\left(2 + 2\max\left(\sqrt[4]{\frac{\log(1/\delta)}{d}}, \sqrt{\frac{\log(1/\delta)}{d}}\right)\right).$$

Fact 7 ([Har04], Theorem 1.1)**.** *Let π be a μ-strongly logconcave density. Let $d\gamma_\mu(x)$ be the Gaussian density with covariance matrix $\mu^{-1}\,\mathrm{id}$. For any convex function h,*

$$\mathbb{E}_\pi[h(x - \mathbb{E}_\pi[x])] \leq \mathbb{E}_{\gamma_\mu}[h(x - \mathbb{E}_{\gamma_\mu}[x])].$$

Fact 8 ([DM16], Theorem 1)**.** *Let π be a μ-strongly logconcave distribution, and let x^* minimize its negative log-density. Then, $\mathbb{E}_\pi[\|x - x^*\|_2^2] \leq \frac{d}{\mu}$.*

5.3 Proximal reduction framework

The reduction framework of Theorem 10 can be thought of as a specialization of a more general composite sampler which we develop in Section 5.5, whose guarantees are reproduced here.

Theorem 11. *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-f(x) - g(x))$ such that f is L-smooth and μ-strongly convex, and let $\epsilon \in (0, 1)$. Let $\eta \leq \frac{1}{32L\kappa d\log(\kappa/\epsilon)}$ (where $\kappa = \frac{L}{\mu}$), $T = \Theta(\frac{1}{\eta\mu}\log(\frac{\kappa d}{\epsilon}))$, and let $\mathcal{O}$ be a η-RGO for g. Further, assume access to the minimizer $x^* = \mathrm{argmin}_{x\in\mathbb{R}^d}\{f(x) + g(x)\}$. There is an algorithm which runs in T iterations in expectation, each querying a gradient oracle of f and $\mathcal{O}$ a constant number of times, and obtains ϵ total variation distance to π.*

Our main observation, elaborated on more formally for specific applications in Section 5.4, is that a variety of structured logconcave densities have negative log-densities f_{oracle}, where we can implement an efficient restricted Gaussian oracle for f_{oracle} via calling an existing sampling method. Crucially, in these instantiations we use the fact that the distributions which $\mathcal{O}$ is required to sampled from are heavily regularized (restricted by a quadratic with large leading coefficient) to obtain fast samplers. We further note that the upper bound requirement on η in Theorem 11 can be lifted when the "well-conditioned" component is uniformly 0. Instead of setting $f = 0$ and $g = f_{\text{oracle}}$ in Theorem 11, and refining the analysis for this special case to tolerate arbitrary η, we provide a self-contained proof here. This particular structure (the composite setting where f_{wc} is uniformly zero and f_{oracle} is strongly convex) admits significant simplifications from the more general case, so using slightly different proof techniques, we are able to obtain stronger convergence guarantees for this particular problem allowing for mixing in fewer than d iterations from a feasible start (see Section 5.1.4).

Theorem 10. *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-f_{\text{oracle}}(x))$ such that f_{oracle} is μ-strongly convex, and let $\epsilon \in (0,1)$. Let $\eta \leq \frac{1}{\mu}$, $T = \Theta(\frac{1}{\eta\mu}\log\frac{d}{\eta\mu\epsilon})$ for some $\beta \geq 1$, and $\mathcal{O}$ be a η-RGO for f_{oracle}. Algorithm 11, initialized at the minimizer of f_{oracle}, runs in T iterations, each querying $\mathcal{O}$ a constant number of times, and obtains ϵ total variation distance to π.*

For simplicity of notation, we replace f_{oracle} in the statement of Theorem 10 with g throughout just this section. Let π be a density on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-g(x))$ where g is μ-strongly convex (but possibly non-smooth), and let $\mathcal{O}$ be a restricted Gaussian oracle for g. Consider the joint distribution $\hat{\pi}$ supported on an expanded space $z = (x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ with density, for some $\eta > 0$,

$$\frac{d\hat{\pi}}{dz}(z) \propto \exp\left(-g(x) - \frac{1}{2\eta}\|x - y\|_2^2\right).$$

Note that the x-marginal of $\hat{\pi}$ is precisely π, so it suffices to sample from the x-marginal. We consider a simple alternating Markov chain for sampling from $\hat{\pi}$, described in the following Algorithm 11.

By observing that the distributions π_x and π_y in the above method are precisely the marginal distributions of $\hat{\pi}$ with one variable fixed, it is straightforward to see that $\hat{\pi}$ is a stationary distribution of the process. We make this formal in the following lemma.

Algorithm 5 `AlternateSample`(g, η, T)

Input: μ-strongly convex $g : \mathbb{R}^d \to \mathbb{R}$, $\eta > 0$, $T \in \mathbf{N}$, $x_0 = \min_x g(x)$.

1: **for** $k \in [T]$ **do**

2: Sample $y_k \sim \pi_{x_{k-1}}$, where for all $x \in \mathbb{R}^d$, $\frac{d\pi_x}{dy}(y) \propto \exp\left(-\frac{1}{2\eta}\|x - y\|_2^2\right)$.

3: Sample $x_k \sim \pi_{y_k}$, where for all $y \in \mathbb{R}^d$, $\frac{d\pi_y}{dx}(x) \propto \exp\left(-g(x) - \frac{1}{2\eta}\|x - y\|_2^2\right)$.

4: **end for**

5: **return** x_T

Lemma 25 (Alternating marginal sampling)**.** *Let $\hat{\pi}$ be a density on two blocks (x, y). Sample $(x, y) \sim \hat{\pi}$, and then sample $\tilde{x} \sim \hat{\pi}(\cdot, y)$, $\tilde{y} \sim \hat{\pi}(\tilde{x}, \cdot)$. Then, the distribution of $(\tilde{x}, \tilde{y})$ is $\hat{\pi}$. Moreover, the alternating marginal sampling Markov chain on either marginal is reversible.*

Proof. The density of the resulting distribution at $(\tilde{x}, y)$ is proportional to the product of the (marginal) density at y and the conditional distribution of $\tilde{x} \mid y$, which by definition is $\hat{\pi}$. Therefore, $(\tilde{x}, y)$ is distributed as $\hat{\pi}$, and the argument for $\tilde{y}$ follows symmetrically. To see reversibility on the x marginal, it suffices to note that the probability we move from x to x' is proportional to

$$\int_y \hat{\pi}(x, y)\hat{\pi}(x', y)dy,$$

which is a symmetric function of x and x'. A similar argument holds for the y marginals. □

We also state a simple observation about alternating schemes such as Algorithm 11, which will be useful later. Let $\mathcal{P}_x$ be the density of y_k after one step of the above procedure starting from $x_{k-1} = x$, and let $\mathcal{T}_x$ be the resulting density of x_k.

Observation 1. *For any two points x, $x' \in \mathbb{R}^d$, $\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\mathrm{TV}} \le \|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\mathrm{TV}}$.*

Proof. This follows by the coupling characterization of total variation (see e.g. Chapter 5 of [LPW09]). Per the optimal coupling of $y \sim \mathcal{P}_x$ and $y' \sim \mathcal{P}_{x'}$, whenever the total variation sets $y = y'$ in Line 2 of `AlternateSample`, we can couple the resulting distributions in Line 3 as well. □

In order to prove Theorem 10, we first show that the random walk in Algorithm 11 converges rapidly in the 2-Wasserstein distance (denoted W_2 in this section).

Lemma 26. *Let π_0 be the starting distribution of x in Algorithm 11. Let π_k be the distribution of x_k and π be the x-marginal of $\hat{\pi}$. For all $k \geq 0$,*

$$W_2^2(\pi_{k+1}, \pi) \leq \frac{1}{(1+\eta\mu)^2} W_2^2(\pi_k, \pi).$$

Hence, for any $\eta \leq \frac{1}{\mu}$, in $T' = O\left(\frac{1}{\eta\mu} \log \frac{d}{\mu\Delta}\right)$ iterations, the random walk mixes to

$$W_2(\pi_{T'}, \pi) \leq \Delta.$$

Proof. Let Γ_{x_k} be the optimal coupling between $x_k \sim \pi_k$ and $\hat{x} \sim \pi$ according to the W_2 distance. Coupling the Gaussian random variable generating $y_{k+1} \sim \pi_{x_k}$ and $\hat{y} \sim \pi_{\hat{x}}$ gives a coupling $\Gamma_{y_{k+1}}$ between y_{k+1} and $\hat{y}$ such that

$$\mathbb{E}_{\Gamma_{y_{k+1}}}\left[\|y_{k+1} - \hat{y}\|_2^2\right] = \mathbb{E}_{\Gamma_{x_k}}\left[\|x_k - \hat{x}\|_2^2\right]. \tag{5.10}$$

Then, let π_y be the distribution of x_{k+1} in a run of Line 3 of Algorithm 11 starting from $y_{k+1} = y$, and $\pi_{\hat{y}}$ be the distribution of $\hat{x}$ in Line 3 starting from $\hat{y}$, respectively. Since $\pi_{\hat{y}}$ is $\mu + \frac{1}{\eta}$ strongly log-concave, $\pi_{\hat{y}}$ satisfies a log-Sobolev inequality with constant $\mu + \frac{1}{\eta}$ (Theorem 2 of [OV00]). Hence,

$$\begin{aligned} W_2^2(\pi_y, \pi_{\hat{y}}) &\leq \frac{2}{\mu + \frac{1}{\eta}} d_{\mathrm{KL}}(\pi_y \| \pi_{\hat{y}}) \\ &\leq \frac{1}{\left(\mu + \frac{1}{\eta}\right)^2} \mathbb{E}_{\pi_y}\left[\|\nabla \log \frac{\pi_y}{\pi_{\hat{y}}}\|_2^2\right] \\ &\leq \frac{1}{(1+\eta\mu)^2} \|y - \hat{y}\|_2^2. \end{aligned}$$

The first step used the Talagrand transportation inequality (Theorem 1 of [OV00]). The second step used the log-Sobolev inequality. The third step used

$$\begin{aligned} \nabla \log \frac{\pi_y(x)}{\pi_{\hat{y}}(x)} &= \nabla \log \frac{\exp(-g(x) - \frac{1}{2\eta}\|x - y\|_2^2) \int_{x'} \exp(-g(x) - \frac{1}{2\eta}\|x - \hat{y}\|_2^2) dx'}{\exp(-g(x) - \frac{1}{2\eta}\|x - \hat{y}\|_2^2) \int_{x'} \exp(-g(x) - \frac{1}{2\eta}\|x - y\|_2^2) dx'} \\ &= \frac{1}{2\eta} \nabla \left(\|x - \hat{y}\|_2^2 - \|x - y\|_2^2\right) = \frac{1}{\eta}(y - \hat{y}). \end{aligned} \tag{5.11}$$

Taking expectation over $\Gamma_{y_{k+1}}$ and using (5.10) shows that

$$W_2^2(\pi_{k+1}, \pi) \leq \frac{1}{(1+\eta\mu)^2} W_2^2(\pi_k, \pi).$$

Algorithm 11 starts from the distribution $\pi_0 = \delta_{x^*}$, where $x^* = \min_x g(x)$. Since π is μ-strongly logconcave, we have (see e.g. Proposition 1 of [DM$^+$19])

$$W_2^2(\pi_0, \pi) = \mathbb{E}_{\hat{\pi}}\left[\|x^* - x\|^2\right] \leq \frac{d}{\mu}.$$

Then, for $\eta < \frac{1}{\mu}$, $\frac{1}{1+\eta\mu} \leq 1 - \frac{\eta\mu}{2}$, so after $T' = O(\frac{1}{\eta\mu} \log \frac{d}{\mu\Delta})$ iterations, $W_2(\pi_{T'}, \pi) \leq \Delta$. □

Next, we bound the KL divergence between the output of Algorithm 11 and the target distribution π. We need the following standard lemma regarding KL divergences of marginal distributions.

Lemma 27. *Let P_z and Q_z be distributions supported on $\mathcal{X}$ indexed by z, a random variable distributed as π_z. Let $\widetilde{P}$ be the joint distribution of (x, z) for $x \sim P_z$ and $z \sim \pi_z$, and $\widetilde{Q}$ be the joint distribution of (x, z) as $x \sim Q_z$ and $z \sim \pi_z$. Let P and Q be the marginal distribution of $\widetilde{P}$ and $\widetilde{Q}$ on x, averaged over z. Then,*

$$d_{\mathrm{KL}}(P\|Q) \le \mathbb{E}_{z\sim\pi_z}\left[d_{\mathrm{KL}}(P_z\|Q_z)\right].$$

Proof. By the definition of d_{KL},

$$\begin{aligned} d_{\mathrm{KL}}(\widetilde{P}\|\widetilde{Q}) &= \mathbb{E}_{(x,z)\sim\widetilde{P}}\left[\log\frac{\widetilde{P}(x,z)}{\widetilde{Q}(x,z)}\right] \\ &= \mathbb{E}_{z\sim\pi_z}\left[\mathbb{E}_{x\sim P_z}\left[\log\frac{\widetilde{P}(x,z)}{\widetilde{Q}(x,z)}\right]\right] \\ &= \mathbb{E}_{z\sim\pi_z}\left[\mathbb{E}_{x\sim P_z}\left[\log\frac{P_z(x)}{Q_z(x)}\right]\right] \\ &= \mathbb{E}_{z\sim\pi_z}\left[d_{\mathrm{KL}}(P_z\|Q_z)\right]. \end{aligned}$$

Finally, by the data processing inequality,

$$d_{\mathrm{KL}}(P\|Q) \le d_{\mathrm{KL}}(\widetilde{P}\|\widetilde{Q}) = \mathbb{E}_{z\sim\pi_z}\left[d_{\mathrm{KL}}(P_z\|Q_z)\right].$$

□

The following lemma shows that a 2-Wasserstein distance bound on the distribution at iteration k implies a KL divergence bound on iteration $k+1$.

Lemma 28. *Let π_k be the distribution of x_k for some k such that $W_2(\pi_k, \pi) \le \Delta$ and π be the x-marginal of $\hat{\pi}$. Then,*

$$d_{\mathrm{KL}}(\pi_{k+1}\|\pi) \le \frac{\Delta^2}{2\eta}.$$

Proof. As in Lemma 26, let Γ_{x_k} be the optimal coupling between $x_k \sim \pi_k$ and $\hat{x} \sim \pi$, which yields a coupling $\Gamma_{y_{k+1}}$ between y_{k+1} and $\hat{y}$ such that

$$\mathbb{E}_{\Gamma_{y_{k+1}}}\left[\|y_{k+1} - \hat{y}\|_2^2\right] = \mathbb{E}_{\Gamma_{x_k}}\left[\|x_k - \hat{x}\|_2^2\right] \le \Delta^2. \tag{5.12}$$

Then,

$$d_{\mathrm{KL}}(\pi_{k+1}\|\pi) \le \mathbb{E}_{(y_{k+1},\hat{y})\sim\Gamma_{y_k}}\left[d_{\mathrm{KL}}(\pi_{y_{k+1}}\|\pi_{\hat{y}})\right]$$
$$\le \frac{1}{2\eta^2\left(\mu+\frac{1}{\eta}\right)}\mathbb{E}_{(y_{k+1},\hat{y})\sim\Gamma_{y_{k+1}}}\left[\left\|y_{k+1}-\hat{y}\right\|_2^2\right] \le \frac{\Delta^2}{2\eta}.$$

The first inequality followed from Lemma 28 by taking $P = \pi_{k+1}$, $Q = \pi$ and $z = (y_{k+1}, y)$. The second inequality used the log-Sobolev inequality and (5.11). The last inequality used (5.12). □

Finally, putting the pieces together, Theorem 10 follows from Lemma 26 and Lemma 28.

Proof of Theorem 10. By Lemma 26 and Lemma 28, there is $T = O\left(\frac{1}{\eta\mu}\log\frac{d}{\eta\mu\epsilon}\right)$ so that $d_{\mathrm{KL}}(\pi_T\|\pi) \le 2\epsilon^2$. By Pinsker's inequality,

$$\|\pi_T - \pi\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}d_{\mathrm{KL}}(\pi_T\|\pi)} = \epsilon.$$

□

We note that Theorem 10 is robust to a small amount of error tolerance in the sampler $\mathcal{O}$. Specifically, if $\mathcal{O}$ has tolerance $\frac{\epsilon}{2T}$, then by calling Theorem 10 with desired accuracy $\frac{\epsilon}{2}$ and adjusting constants appropriately, the cumulative error incurred by all calls to $\mathcal{O}$ is within the total requisite bound (formally, this can be shown via the coupling characterization of total variation). We defer a more formal elaboration on this inexactness argument to Appendix D.1 and the proof of Proposition 10.

5.4 Tighter runtimes for structured densities

In this section, we use applications of Theorem 10 to obtain simple analyses of novel state-of-the-art high-accuracy runtimes for the well-conditioned densities studied in [DCWY19, CDWY20, LST20], as well as the composite and finite sum densities studied in this work. We will assume the conclusions of Theorems 11 and 12 respectively in deriving the results of Sections 5.4.2 and 5.4.3.

5.4.1 Well-conditioned logconcave sampling: proof of Corollary 10

In this section, let π be a distribution on $\mathbb{R}^d$ with density proportional to $\exp(-f(x))$, where f is L-smooth and μ-strongly convex (and $\kappa = \frac{L}{\mu}$) and has pre-computed minimizer

x^*. We will instantiate Theorem 10 with $f_{\text{oracle}}(x) = f(x)$, and choose $\eta = \frac{1}{8Ld\log(\kappa)}$. We now require an η-RGO $\mathcal{O}$ for $f_{\text{oracle}} = f$ to use in Theorem 10.

Our implementation of $\mathcal{O}$ is a rejection sampling scheme. We use the following helpful guarantee.

Lemma 29 (Rejection sampling)**.** *Let π, $\hat{\pi}$ be distributions on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto p(x)$, $\frac{d\hat{\pi}}{dx}(x) \propto \hat{p}(x)$. Suppose for some $C \geq 1$ and all $x \in \mathbb{R}^d$, $\frac{p(x)}{\hat{p}(x)} \leq C$. The following is termed "rejection sampling": repeat independent runs of the following procedure until a point is outputted.*

1. *Draw $x \sim \hat{\pi}$.*
2. *With probability $\frac{p(x)}{C\hat{p}(x)}$, output x.*

Rejection sampling terminates in $\frac{C\int \hat{p}(x)dx}{\int p(x)dx}$ runs in expectation, and the output distribution is π.

Proof. The second claim follows from Bayes' rule which implies the conditional density of the output point is proportional to $\hat{p}(x) \cdot \frac{p(x)}{C\hat{p}(x)} \propto p(x)$, so the distribution is π. To see the first claim, the probability any sample outputs is

$$\int_x \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x) = \frac{1}{C}\int_x \frac{\int_x p(x)dx}{\int_x \hat{p}(x)dx} d\pi(x) = \frac{\int_x p(x)dx}{C\int_x \hat{p}(x)dx}.$$

The conclusion follows by independence and linearity of expectation. □

We further state a concentration bound shown first in [LST20] regarding the norm of the gradient of a point drawn from a logsmooth distribution.

Proposition 7 (Logsmooth gradient concentration, Corollary 3.3, [LST20])**.** *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-f(x))$ where f is convex and L-smooth. With probability at least $1 - \kappa^{-d}$,*

$$\|\nabla f(x)\|_2 \leq 3\sqrt{L}d\log\kappa \text{ for } x \sim \pi. \tag{5.13}$$

By the requirements of Theorem 10, the restricted Gaussian oracle $\mathcal{O}$ only must be able to draw samples from densities of the form, for some $y \in \mathbb{R}^d$,

$$\exp\left(-f_{\text{oracle}}(x) - \frac{1}{2\eta}\|x-y\|_2^2\right) = \exp\left(-f(x) - 4Ld\log\kappa\|x-y\|_2^2\right). \tag{5.14}$$

We will use the following Algorithm 6 to implement $\mathcal{O}$.

Algorithm 6 `XSample`(f, y, η)

Input: L-smooth, μ-strongly convex $f : \mathbb{R}^d \to \mathbb{R}$, $y \in \mathbb{R}^d$, $\eta > 0$

1: **if** $\|\nabla f(y)\|_2 \le 3\sqrt{L} d \log \kappa$ **then**
2: **while true do**
3: Draw $x \sim \mathcal{N}(y - \nabla f(y), \eta \,\mathrm{id})$
4: $\tau \sim \mathrm{Unif}[0, 1]$
5: **if** $\tau \le \exp(f(y) + \langle \nabla f(y), x - y\rangle - f(x))$ **then**
6: **return** x
7: **end if**
8: **end while**
9: **end if**
10: Use [CDWY20] to sample x from (5.14) to total variation distance $\frac{\epsilon}{\Theta(\kappa d^2 \log^3(\frac{\kappa d}{\epsilon}))}$ using $O(d \log \frac{\kappa d}{\epsilon})$ queries to ∇f (Theorem 1, [CDWY20], where (5.14) has constant condition number)
11: **return** x

Lemma 30. *Let $\eta = \frac{1}{8Ld\log(\kappa)}$, and suppose y satisfies the bound in* (5.13)*, i.e. $\|\nabla f(y)\|_2 \le 3\sqrt{L} d \log \kappa$. Then, Line 3 of Algorithm 6 runs an expected 2 times, and Algorithm 6 samples exactly from* (5.14)*, whenever the condition of Line 1 is met.*

Proof. Note that when the assumption of Line 1 is met, Algorithm 6 is an instantiation of rejection sampling (Lemma 29) with

$$p(x) = \exp\left(-f(x) - \frac{1}{2\eta}\|x - y\|_2^2\right),$$
$$\hat{p}(x) = \exp\left(-f(y) - \langle \nabla f(y), x - y\rangle - \frac{1}{2\eta}\|x - y\|_2^2\right).$$

By convexity, we may take $C = 1$. Next, by applying Fact 5 twice and L-smoothness of f_{oracle},

$$\begin{aligned}
\int_x p(x)dx &\ge \int_x \exp\left(-f(y) - \langle \nabla f(y), x - y\rangle - \frac{1 + \eta L}{2\eta}\|x - y\|_2^2\right) dx \\
&= \exp\left(-f(y) + \frac{\eta}{2(1 + \eta L)}\|\nabla f(y)\|_2^2\right) \int_x \exp\left(-\frac{1 + \eta L}{2\eta}\|x - y + \frac{\eta}{1 + \eta L}\nabla f(y)\|_2^2\right) dx \\
&= \exp\left(-f(y) + \frac{\eta}{2(1 + \eta L)}\|\nabla f(y)\|_2^2\right) \left(\frac{2\pi\eta}{1 + \eta L}\right)^{\frac{d}{2}}, \\
\int_x \hat{p}(x)dx &= \exp\left(-f(y) + \frac{\eta}{2}\|\nabla f(y)\|_2^2\right) (2\pi\eta)^{\frac{d}{2}},
\end{aligned}$$

which implies the desired bound (recalling Lemma 29 and our assumed bound on $\|\nabla f(y)\|_2$)

$$\frac{\int \hat{p}(x)dx}{\int p(x)dx} \leq \exp\left(\left(\frac{\eta}{2} - \frac{\eta}{2(1+\eta L)}\right)\|\nabla f(y)\|_2^2\right)(1+\eta L)^{\frac{d}{2}}$$
$$\leq 1.5\exp\left(\frac{\eta^2 L}{2(1+\eta L)}\|\nabla f(y)\|_2^2\right) \leq 2.$$

□

We are now equipped to prove our main result concerning well-conditioned densities.

Corollary 10. *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-f(x))$ such that f is L-smooth and μ-strongly convex, and let $\epsilon \in (0,1)$, $\kappa = \frac{L}{\mu}$. Assume access to $x^* = \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$. Algorithm 11 with $\eta = \frac{1}{8Ld\log(\kappa)}$ using Algorithm 6 as a restricted Gaussian oracle for f uses $O(\kappa d \log\kappa \log\frac{\kappa d}{\epsilon})$ gradient queries in expectation, and obtains ϵ total variation distance to π.*

Proof. By applying Theorem 10 with the chosen η, and noting that the cumulative error due to all calls to Line 10 cannot amount to more than $\frac{\epsilon}{2}$ total variation error throughout the algorithm, it suffices to show that Algorithm 6 uses $O(1)$ gradient queries each iteration in expectation. This happens whenever the condition in Line 1 is met via Lemma 30, so we must show Line 10 is executed with probability $O((d\log\frac{\kappa d}{\epsilon})^{-1})$.

To show this, note that combining Proposition 7 with the warmness of the start x_0 in Algorithm 6, this event occurs with probability at most $\kappa^{-\frac{d}{2}}$ in the first iteration.[7] Since warmness is monotonically decreasing[8] throughout using an exact oracle in Algorithm 11, and the total error accumulated due to Line 10 throughout the algorithm is $O((d\log\frac{\kappa d}{\epsilon})^{-1})$, we have the desired conclusion. □

We show a bound nearly-matching Corollary 10 using only value access to f, and with a deterministic iteration complexity (rather than an expected one), as Corollary 14 in Section 5.4.3.

---

[7]Formally, Line 2 of Algorithm 11 has $y_1 \sim \mathcal{N}(x_0, \eta\,\mathrm{id})$, but by smoothness $\|\nabla f(y_1)\|_2 \leq \|\nabla f(x_0)\|_2 + L\|x-y\|_2$ and $L\|x-y\|_2 \leq \widetilde{O}(L\sqrt{\eta})$ with high probability, adding a negligible constant to the bound of Proposition 7.

[8]This is a standard fact in the literature, and can be seen as follows: each transition step in the chain is a convex combination of warm point masses, preserving warmness.

5.4.2 Composite logconcave sampling: proof of Corollary 11

In this section, let π be a distribution on $\mathbb{R}^d$ with density proportional to $\exp(-f(x)-g(x))$, where f is L-smooth and μ-strongly convex (and $\kappa = \frac{L}{\mu}$), and g is convex and admits a restricted Gaussian oracle $\mathcal{O}$. Without loss of generality, we assume that f and g share a minimizer x^* which we have pre-computed; if this is not the case, we can redefine $f(x) \leftarrow f(x) - \langle \nabla f(x^*), x\rangle$ and $g(x) \leftarrow g(x) + \langle \nabla f(x^*), x\rangle$; see Section 5.5.1 for this reduction.

We will instantiate Theorem 10 with $f_{\text{oracle}} = f + g$, which is a μ-strongly convex function. Our main result of this section follows directly from Theorem 10 and using Theorem 11 as the required oracle $\mathcal{O}$, stated more precisely in the following.

Corollary 11. *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-f(x) - g(x))$ such that f is L-smooth and μ-strongly convex, and let $\epsilon \in (0,1)$, $\kappa = \frac{L}{\mu}$. Assume access to $x^* = \operatorname{argmin}_{x\in\mathbb{R}^d}\{f(x) + g(x)\}$ and let $\mathcal{O}$ be a restricted Gaussian oracle for g. There is an algorithm (Algorithm 11 using Theorem 11 as a restricted Gaussian oracle) which runs in $O(\kappa d \log^3 \frac{\kappa d}{\epsilon})$ iterations in expectation, each querying a gradient of f and $\mathcal{O}$ a constant number of times, and obtains ϵ total variation distance to π.*

Proof. As discussed at the beginning of this section, assume without loss that f and g both are minimized by x^*. We apply the algorithm of Theorem 10 with $\eta = \frac{1}{L}$ to the μ-strongly convex function $f + g$, which requires one call to $\mathcal{O}$ to implement. Thus, the iteration count parameter in Theorem 10 is $T = O(\kappa \log \frac{\kappa d}{\epsilon})$.

Recall that we chose $\eta = \frac{1}{L}$. To bound the total complexity of this algorithm, it suffices to give an η-RGO $\mathcal{O}^+$ for sampling from distributions with densities of the form, for some $y \in \mathbb{R}^d$,

$$\exp\left(-f(x) - g(x) - \frac{1}{2\eta}\|x-y\|_2^2\right) = \exp\left(-f(x) - g(x) - \frac{L}{2}\|x-y\|_2^2\right)$$

to total variation distance $\frac{\epsilon}{\Theta(T)}$ (see discussion at the end of Section 5.3). To this end, we apply Theorem 11 with the well-conditioned component $f(x) + \frac{L}{2}\|x-y\|_2^2$, the composite component $g(x)$, and the largest possible choice of η. Note that we indeed have access to a restricted Gaussian oracle for g (namely, $\mathcal{O}$), and this choice of well-conditioned component is $2L$-smooth and L-strongly convex, so its condition number is a constant. Thus, Theorem 11 requires $O(d \log^2 \frac{\kappa d}{\epsilon})$ calls to $\mathcal{O}$ and gradients of f to implement the desired $\mathcal{O}^+$ on any query y (where we note $\frac{\epsilon}{\Theta(T)} = \frac{1}{\text{poly}(\kappa, d, \epsilon^{-1})}$). Combining these complexity bounds yields the desired conclusion. □

5.4.3 Sampling logconcave finite sums: proof of Corollary 12

In this section, let π be a distribution on $\mathbb{R}^d$ with density proportional to $\exp(-F(x))$, where $F(x) = \frac{1}{n}\sum_{i\in[n]} f_i(x)$ is μ-strongly convex, and for all $i \in [n]$, f_i is L-smooth (and $\kappa = \frac{L}{\mu}$). We will instantiate Theorem 10 with $f_{\text{oracle}}(x) = F(x)$, and Theorem 12 as an η-RGO for some choice of η.

More precisely, Theorem 12 shows that given access to the minimizer x^*, only zeroth-order access to the summands of F is necessary to obtain the iteration bound. In order to obtain the minimizer to high accuracy however, variance reduced stochastic gradient methods (e.g. [JZ13]) require $\Omega(n + \kappa)$ gradient queries, which amounts to $\Omega((n + \kappa)d)$ function evaluations. We state a convenient corollary of Theorem 12 which removes the requirement of accessing x^*, via an optimization pre-processing step using the method of [JZ13] (see further discussion in Appendix D.1). This is useful to us in proving Theorem 12 because in the sampling tasks required by the RGO, the minimizer changes (and thus must be recomputed every time).

Corollary 13 (First-order logconcave finite sum sampling)**.** *In the setting of Theorem 12, using [JZ13] to precompute the minimizer x^* and running Algorithm 10 uses $O(n \log \frac{\kappa d}{\epsilon} + \kappa^2 d \log^4 \frac{n\kappa d}{\epsilon})$ first-order oracle queries to summands $\{f_i\}_{i\in[n]}$ and obtains ϵ total variation distance to π.*

We now apply the reduction framework developed in Section 6.2 to our Algorithm 10 to obtain an improved query complexity for sampling from logconcave finite sums.

Corollary 12 (Improved first-order logconcave finite sum sampling)**.** *In the setting of Theorem 12, Algorithm 11 using Algorithm 10 and SVRG [JZ13] as a restricted Gaussian oracle for F uses*

$$O\left(n \log\left(\frac{n\kappa d}{\epsilon}\right) + \kappa\sqrt{nd}\log^{3.5}\left(\frac{n\kappa d}{\epsilon}\right) + \kappa d \log^5\left(\frac{n\kappa d}{\epsilon}\right)\right) = \widetilde{O}\left(n + \kappa \max\left(d, \sqrt{nd}\right)\right)$$

queries to first-order oracles for summands $\{f_i\}_{i\in[n]}$, and obtains ϵ total variation distance to π.

Proof. We apply Theorem 10 with μ-strongly convex $f_{\text{oracle}} = F(x)$, using Algorithm 10 as the required η-RGO $\mathcal{O}$ for sampling from distributions with densities of the form

$$\exp\left(-F(x) - \frac{1}{\eta}\|x - y\|_2^2\right)$$

for some $y \in \mathbb{R}^d$, to total variation $\frac{\epsilon}{\Theta(T)}$ (see Section 5.3) for T the iteration bound of Algorithm 11. We apply Theorem 12 to the function $\widetilde{F}(x) = F(x) + \frac{1}{\eta}\|x - y\|_2^2$; we can express this in finite sum form by adding $\frac{1}{\eta}\|x - y\|_2^2$ to every constituent function, and the effect on gradient oracles is $\frac{1}{\eta}(x - y)$. Note $\widetilde{F}$ has condition number $O(1 + \eta L)$. For a given η, the overall complexity is

$$\frac{\log \frac{\kappa d}{\epsilon}}{\eta\mu}\left(n \log\left(\frac{n\kappa d}{\epsilon}\right) + d\log^4\left(\frac{n\kappa d}{\epsilon}\right) + (\eta L)^2 d \log^4\left(\frac{n\kappa d}{\epsilon}\right)\right)$$

Here, the inner loop complexity uses Corollary 13 to also find the minimizer (for warm starts), and the outer loop complexity is by Theorem 10. The result follows by optimizing over η, namely picking $\eta = \max(\frac{1}{L}, \sqrt{\frac{n}{L^2 d \log^3(n\kappa d/\epsilon)}})$, and that Algorithm 11 always must have at least one iteration. □

Note the only place that Corollary 12 used gradient evaluations was in determining minimizers of subproblems, via the first step of Corollary 13. Consider now the $n = 1$ case. By running e.g. accelerated gradient descent for smooth and strongly convex functions, it is well-known [Nes83] that we can obtain a minimizer in $\widetilde{O}(\sqrt{\kappa})$ iterations, each querying a gradient oracle, where κ is the condition number. By smoothness, we can approximate every coordinate of the gradient to arbitrary precision using 2 function evaluations, so this is a $\widetilde{O}(\sqrt{\kappa}d)$ value oracle complexity.

Finally, for every optimization subproblem in Corollary 12 where $\eta = (L \cdot \text{polylog}\frac{\kappa d}{\epsilon})^{-1}$, the condition number is a constant, which amounts to a $\widetilde{O}(d)$ value oracle complexity for computing a minimizer. This is never the dominant term compared to Theorem 12, yielding the following conclusion.

Corollary 14. *In the setting of Corollary 10, Algorithm 11 using Algorithm 10 as a restricted Gaussian oracle uses $O(\kappa d \log^2 \frac{\kappa d}{\epsilon})$ value queries and obtains ϵ total variation distance to π.*

We note that the polylogarithmic factor is significantly improved when compared to Corollary 12 by removing the random sampling steps in Algorithm 10. A precise complexity bound of the resulting Metropolized random walk, a zeroth-order algorithm mixing in $O(\kappa^2 d \log \frac{\kappa d}{\epsilon})$ for a logconcave distribution with condition number κ, is given as Theorem 2 of [CDWY20].

Finally, in the case $n \geq 1$, we also exhibit an improved query complexity in terms of an entirely zeroth-order sampling algorithm which interpolates with Corollary 14 (up to

logarithmic factors). By trading off the $\widetilde{O}(nd+\kappa d)$ zeroth-order complexity of minimizing a finite sum function [JZ13], and the $\widetilde{O}(\kappa^2 d)$ zeroth-order complexity of sampling, we can run Theorem 10 for the optimal choice of $\eta = \widetilde{O}(\frac{\sqrt{n}}{L})$. The overall zeroth-order complexity can be seen to be $\widetilde{O}(nd + \sqrt{n}\kappa d)$.

5.5 Composite logconcave sampling with a restricted Gaussian oracle

In this section, we provide our "base sampler" for composite logconcave densities as Algorithm 7, and give its guarantees by proving Theorem 11. Throughout, fix distribution π with density

$$\frac{d\pi}{dx}(x) \propto \exp\left(-f(x) - g(x)\right), \text{where } f : \mathbb{R}^d \to \mathbb{R} \text{ is } L\text{-smooth, } \mu\text{-strongly convex,} \\ \text{and } g : \mathbb{R}^d \to \mathbb{R} \text{ admits a restricted Gaussian oracle } \mathcal{O}. \tag{5.15}$$

We will define $\kappa \stackrel{\text{def}}{=} \frac{L}{\mu}$, and assume that we have precomputed $x^* \stackrel{\text{def}}{=} \text{argmin}_{x\in\mathbb{R}^d} \{f(x) + g(x)\}$. Our algorithm proceeds in stages following the outline in Section 5.1.3.

1. `Composite-Sample` is reduced to `Composite-Sample-Shared-Min`, which takes as input a distribution with negative log-density $f+g$, where f and g share a minimizer; this reduction is given in Section 5.5.1, and the remainder of the section handles the shared-minimizer case.

2. The algorithm `Composite-Sample-Shared-Min` is a rejection sampling scheme built on top of sampling from a joint distribution $\hat{\pi}$ on $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ whose x-marginal approximates π. We give this reduction in Section 5.5.2.

3. The bulk of our analysis is for `Sample-Joint-Dist`, an alternating marginal sampling algorithm for sampling from $\hat{\pi}$. To implement marginal sampling, it alternates calls to $\mathcal{O}$ and a rejection sampling algorithm `YSample`. We prove its correctness in Section 5.5.3.

We put these pieces together in Section 5.5.4 to prove Theorem 11. We remark that for simplicity, we will give the algorithms corresponding to the largest value of step size η in the theorem statement; it is straightforward to modify the bounds to tolerate smaller values of η, which will cause the mixing time to become correspondingly larger (in particular, the value of K in Algorithm 9).

Algorithm 7 `Composite-Sample`(π, x^*, ϵ)

Input: Distribution π of form (5.15), x^* minimizing negative log-density of π, $\epsilon \in [0, 1]$.

Output: Sample x from a distribution π' with $\|\pi' - \pi\|_{\mathrm{TV}} \le \epsilon$.

1: $\tilde{f}(x) \leftarrow f(x) - \langle \nabla f(x^*), x \rangle$, $\tilde{g}(x) \leftarrow g(x) + \langle \nabla f(x^*), x \rangle$

2: **return** `Composite-Sample-Shared-Min`$(\pi, \tilde{f}, \tilde{g}, x^*, \epsilon)$

Algorithm 8 `Composite-Sample-Shared-Min`$(\pi, f, g, x^*, \epsilon)$

Input: Distribution π of form (5.15), where f and g are both minimized by x^*, $\epsilon \in [0, 1]$.

Output: Sample x from a distribution π' with $\|\pi' - \pi\|_{\mathrm{TV}} \le \epsilon$.

1: **while true do**

2: Define the set

$$\Omega \stackrel{\text{def}}{=} \left\{ x \mid \|x - x^*\|_2 \le 4\sqrt{\frac{d \log(288\kappa/\epsilon)}{\mu}} \right\} \tag{5.16}$$

3: $x \leftarrow$ `Sample-Joint-Dist`$(f, g, x^*, \mathcal{O}, \frac{\epsilon}{18})$

4: **if** $x \in \Omega$ **then**

5: $\tau \sim \mathrm{Unif}[0, 1]$

6: $y \leftarrow$ `YSample`(f, x, η)

7: $\alpha \leftarrow \exp\left(f(y) - \langle \nabla f(x), y - x \rangle - \frac{L}{2}\|y - x\|_2^2 + g(x) + \frac{\eta L^2}{2}\|x - x^*\|_2^2\right)$

8: $\hat{\theta} \leftarrow \exp\left(-f(x) - g(x) + \frac{\eta}{2(1+\eta L)}\|\nabla f(x)\|_2^2\right)(1 + \eta L)^{\frac{d}{2}}\alpha$

9: **if** $\tau \le \frac{\hat{\theta}}{4}$ **then**

10: **return** x

11: **end if**

12: **end if**

13: **end while**

5.5.1 Reduction from `Composite-Sample` to `Composite-Sample-Shared-Min`

Correctness of `Composite-Sample` is via the following properties.

Proposition 8. *Let $\tilde{f}$ and $\tilde{g}$ be defined as in `Composite-Sample`.*

1. *The density $\propto \exp(-f(x) - g(x))$ is the same as the density $\propto \exp(-\tilde{f}(x) - \tilde{g}(x))$.*

2. *Assuming first-order (function and gradient evaluation) access to f, and restricted Gaussian oracle access to g, we can implement the same accesses to $\tilde{f}$, $\tilde{g}$ with con-*

Algorithm 9 `Sample-Joint-Dist`$(f, g, x^*, \eta, \mathcal{O}, \delta)$

Input: f, g of form (5.15) both minimized by x^*, $\delta \in [0,1]$, $\eta > 0$, $\mathcal{O}$ restricted Gaussian oracle for g.

Output: Sample x from a distribution $\hat{\pi}'$ with $\|\hat{\pi}' - \hat{\pi}\|_{\mathrm{TV}} \le \delta$, where we overload $\hat{\pi}$ to mean the marginal of (5.17) on the x variable.

1: $\eta \leftarrow \frac{1}{32L\kappa d \log(16\kappa/\delta)}$

2: Let $\hat{\pi}$ be the density with

$$\frac{d\hat{\pi}}{dx}(z) \propto \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y - x\|_2^2 - \frac{\eta L^2}{2}\|x - x^*\|_2^2\right) \tag{5.17}$$

3: Call $\mathcal{O}$ to sample $x_0 \sim \pi_{\mathrm{start}}$, for

$$\frac{d\pi_{\mathrm{start}}(x)}{dx} \propto \exp\left(-\frac{L + \eta L^2}{2}\|x - x^*\|_2^2 - g(x)\right) \tag{5.18}$$

4: $K \leftarrow \frac{2^{26}\cdot 100}{\eta\mu} \log\left(\frac{d\log(16\kappa)}{4\delta}\right)$ (see Remark 1)

5: **for** $k \in [K]$ **do**

6: Call `YSample` $\left(f, x_{k-1}, \eta, \frac{\delta}{2Kd\log(\frac{d\kappa}{\delta})}\right)$ to sample $y_k \sim \pi_{x_{k-1}}$ (Algorithm 14), for

$$\frac{d\pi_x}{dy}(y) \propto \exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right) \tag{5.19}$$

7: Call $\mathcal{O}$ to sample $x_k \sim \pi_{y_k}$, for

$$\frac{d\pi_y}{dx}(x) \propto \exp\left(-g(x) - \frac{1}{2\eta}\|y - x\|_2^2 - \frac{\eta L^2}{2}\|x - x^*\|_2^2\right) \tag{5.20}$$

8: **end for**

9: **return** x_K

stant overhead.

3. $\tilde{f}$ and $\tilde{g}$ are both minimized by x^*.

Proof. For f and g with properties as in (5.15), with x^* minimizing $f + g$, define the functions

$$\tilde{f}(x) \stackrel{\mathrm{def}}{=} f(x) - \langle \nabla f(x^*), x\rangle\,,\ \tilde{g}(x) \stackrel{\mathrm{def}}{=} g(x) + \langle \nabla f(x^*), x\rangle\,,$$

and observe that $\tilde{f} + \tilde{g} = f + g$ everywhere. This proves the first claim. Further, implementation of a first-order oracle for $\tilde{f}$ and a restricted Gaussian oracle for $\tilde{g}$ are immediate

assuming a first-order oracle for f and a restricted Gaussian oracle for g, showing the second claim; any quadratic shifted by a linear term is the sum of a quadratic and a constant. We now show $\tilde{f}$ and $\tilde{g}$ have the same minimizer. By strong convexity, $\tilde{f}$ has a unique minimizer; first-order optimality shows that

$$\nabla \tilde{f}(x^*) = \nabla f(x^*) - \nabla f(x^*) = 0,$$

so this unique minimizer is x^*. Moreover, optimality of x^* for $f + g$ implies that for all $x \in \mathbb{R}^d$,

$$\langle \partial g(x^*) + \nabla f(x^*), x^* - x \rangle \leq 0.$$

Here, ∂g is a subgradient. This shows first-order optimality of x^* for $\tilde{g}$ also, so x^* minimizes $\tilde{g}$. □

5.5.2 Reduction from `Composite-Sample-Shared-Min` to `Sample-Joint-Dist`

`Composite-Sample-Shared-Min` is a rejection sampling scheme, which accepts samples from subroutine `Sample-Joint-Dist` in the high-probability region Ω defined in (5.16). We give a general analysis for approximate rejection sampling in Appendix D.2.1, and Appendix D.2.1 bounds relationships between distributions π and $\hat{\pi}$, defined in (5.15) and (5.17) respectively (i.e. relative densities and normalization constant ratios). Combining these pieces proves the following main claim.

Proposition 9. *Let $\eta = \frac{1}{32L\kappa d \log(288\kappa/\epsilon)}$, and assume `Sample-Joint-Dist`$(f, g, x^*, \mathcal{O}, \delta)$ samples within δ total variation of the x-marginal on* (5.17)*. `Composite-Sample-Shared-Min` outputs a sample within total variation ϵ of* (5.15) *in an expected $O(1)$ calls to `Sample-Joint-Dist`.*

5.5.3 Implementing `Sample-Joint-Dist`

`Sample-Joint-Dist` alternates between sampling marginals in the joint distribution $\hat{\pi}$, as seen by definitions (5.19), (5.20). We showed that marginal sampling attains the correct stationary distribution as Lemma 25. We bound the conductance of the induced walk on iterates $\{x_k\}$ by combining an isoperimetry bound with a total variation guarantee between transitions of nearby points in Appendix D.2.2. Finally, we give a simple rejection sampling scheme `YSample` as Algorithm 14 for implementing the step (5.19). Since the y-marginal of $\hat{\pi}$ is a bounded perturbation of a Gaussian (intuitively, f is L-smooth and $\eta^{-1} \gg L$), we show in a high probability region that rejecting from the sum of a first-order approximation to f and the Gaussian succeeds in 2 iterations.

Remark 1. *For simplicity of presentation, we were conservative in bounding constants throughout; in practice, we found that the constant in Line 4 is orders of magnitude too large (a constant < 10 sufficed), which can be found as Section 4 of [SL19]. Several constants were inherited from prior analyses, which we do not rederive to save on redundancy.*

We now give a complete guarantee on the complexity of `Sample-Joint-Dist`.

Proposition 10. *`Sample-Joint-Dist` outputs a point with distribution within δ total variation distance from the x-marginal of $\hat{\pi}$. The expected number of gradient queries per iteration is constant.*

5.5.4 Putting it all together: proof of Theorem 11

We show Theorem 11 follows from the guarantees of Propositions 8, 9, and 10. Formally, Theorem 11 is stated for an arbitrary value of η which is upper bounded by the value in Line 1 of Algorithm 9; however, it is straightforward to see that all our proofs go through for any smaller value. By observing the value of K in `Sample-Joint-Dist`, we see that the number of total iterations in each call to `Sample-Joint-Dist` $O\left(\frac{1}{\eta\mu}\log(\frac{\kappa d}{\epsilon})\right) = O\left(\kappa^2 d \log^2\left(\frac{\kappa d}{\delta}\right)\right)$. Proposition 10 also shows that every iteration, we require an expected constant number of gradient queries and calls to $\mathcal{O}$, the restricted Gaussian oracle for g, and that the resulting distribution has δ total variation from the desired marginal of $\hat{\pi}$. Next, Proposition 9 implies that the number of calls to `Sample-Joint-Dist` in a run of `Composite-Sample-Shared-Min` is bounded by a constant, the choice of δ is $\Theta(\epsilon)$, and the resulting point has total variation ϵ from the original distribution π. Finally, Proposition 8 shows sampling from a general distribution of the form (5.1) is reducible to one call of `Composite-Sample-Shared-Min`, and the requisite oracles are implementable.

5.6 Logconcave finite sums

In this section, we provide our "base sampler" for logconcave finite sums as Algorithm 10, and give its guarantees by proving Theorem 12. Throughout, fix distribution π with density

$$\frac{d\pi}{dx}(x) \propto \exp(-F(x)), \text{ where } F(x) = \frac{1}{n}\sum_{i\in[n]} f_i(x) \text{ is } \mu\text{-strongly convex,}$$
$$\text{and for all } i \in [n],\ f_i \text{ is } L\text{-smooth.}$$

We will define $\kappa \stackrel{\text{def}}{=} \frac{L}{\mu}$, and assume that we have precomputed $x^* \stackrel{\text{def}}{=} \operatorname{argmin}_{x \in \mathbb{R}^d}\{F(x)\}$. We will also assume explicitly that $\nabla f_i(x^*) = 0$ for all $i \in [n]$ throughout this section (i.e. all f_i are minimized at the same point); this is without loss of generality, by a similar argument as in Proposition 8.

Algorithm 10 `FiniteSum-MRW`(F, h, x_0, p, K)

Input: $F(x) = \frac{1}{n}\sum_{i\in[n]} f_i(x)$, step size $h > 0$, initial x_0, $p \in [0,1]$, iteration count $K \in \mathbf{N}$

for $0 \le k < K$ **do**
  Draw $\xi_k \sim \mathcal{N}(0, \mathrm{id})$
  $y_{k+1} \leftarrow x_k + \sqrt{2h}\xi_k$
  Draw $S_k \subseteq [n]$ by including each $i \in S_k$ independently with probability p
  For each $i \in [n]$,

$$\gamma_k^{(i)} \leftarrow \begin{cases} \frac{1}{p}\left(\sqrt{\exp\left(-\frac{1}{n}f_i(y_{k+1}) + \frac{1}{n}f_i(x_k)\right)} - 1\right) + 1 & i \in S_k \\ 1 & i \notin S_k \end{cases}$$

  $\gamma_k \leftarrow \prod_{i=1}^n \gamma_k^{(i)}$, $\tau \sim \mathrm{Unif}[0,1]$
  if $\tau \le \frac{3}{4}\gamma_k$ and $|S_k| \le 2pn$ **then**
    $x_{k+1} \leftarrow y_{k+1}$
  else
    $x_{k+1} \leftarrow x_k$
  end if
end for
return x_K.

Algorithm 10 is the zeroth-order Metropolized random walk of [DCWY19] with an efficient, but biased, filter step; the goal of our analysis is to show this bias does not incur significant error.

5.6.1 Approximate Metropolis-Hastings

We first recall the following well-known fact underlying Metropolis-Hastings (MH) filters.

Proposition 11. *Consider a random walk on $\mathbb{R}^d$ with proposal distributions $\{\mathcal{P}_x\}_{x\in\mathbb{R}^d}$ and acceptance probabilities $\{\alpha(x,x')\}_{x,x'\in\mathbb{R}^d}$ conducted as follows: at a current point x,*

1. *Draw a point* $x' \sim \mathcal{P}_x$.

2. *Move the random walk to* x' *with probability* $\alpha(x, x')$*, else stay at* x.

Suppose $\mathcal{P}_x(x') = \mathcal{P}_{x'}(x)$ *for all pairs* $x, x' \in \mathbb{R}^d$*, and further* $\frac{d\pi}{dx}(x)\alpha(x, x') = \frac{d\pi}{dx}(x')\alpha(x', x)$*. Then,* π *is a stationary distribution for the random walk.*

Proof. This follows because the walk satisfies detailed balance (reversibility) with respect to π. □

We propose an algorithm that applies a variant of the Metropolis-Hastings filter to a Gaussian random walk. Specifically, we define the following algorithm, which we call `Inefficient-MRW`.

Definition 4 (`Inefficient-MRW`)**.** *Consider the following random walk for some step size* $h > 0$*: for each iteration* k *at a current point* $x_k \in \mathbb{R}^d$*,*

1. *Set* $y_{k+1} \leftarrow x_k + \sqrt{2h}\xi$*, where* $\xi \sim \mathcal{N}(0, \mathrm{id})$.

2. $x_{k+1} \leftarrow y_{k+1}$ *with probability* $\alpha(x_k, y_{k+1})$ *(otherwise,* $x_{k+1} \leftarrow x_k$*), where*

$$\alpha(x, y) = \begin{cases} 1 & \sqrt{\frac{\exp(-F(y))}{\exp(-F(x))}} > \frac{4}{3}, \\ \frac{3}{4}\sqrt{\frac{\exp(-F(y))}{\exp(-F(x))}} & \frac{3}{4} \leq \sqrt{\frac{\exp(-F(y))}{\exp(-F(x))}} \leq \frac{4}{3}, \\ \frac{\exp(-F(y))}{\exp(-F(x))} & \sqrt{\frac{\exp(-F(y))}{\exp(-F(x))}} < \frac{3}{4}. \end{cases} \tag{5.21}$$

Lemma 31. *Distribution* π *with* $\frac{d\pi}{dx}(x) \propto \exp(-F(x))$ *is stationary for* `Inefficient-MRW`*.*

Proof. Without loss of generality, assume that π has been normalized so that $\frac{d\pi}{dx}(x) = \exp(-F(x))$. We apply Proposition 11, dropping subscripts in the following. It is clear that $\mathcal{P}_x(y) = \mathcal{P}_y(x)$ for any x, y, so it suffices to check the second condition. When $\frac{3}{4} \leq \sqrt{\frac{\exp(-F(y))}{\exp(-F(x))}} \leq \frac{4}{3}$, this follows from

$$\frac{d\pi}{dx}(x)\alpha(x, x') = \frac{3}{4}\sqrt{\exp(-F(x) - F(y))} = \frac{d\pi}{dx}(x')\alpha(x', x).$$

The other case is similar (as it is a standard Metropolis-Hastings filter). □

In Algorithm 10, we implement an approximate version of the modified MH filter in Definition 4, where we always assume the pair x, y are in the second case of (5.21). In Lemma 32, we show that if a certain boundedness condition holds, then Algorithm 10

approximates `Inefficient-MRW` well. We then show that the output distributions of `Inefficient-MRW` and our Algorithm 10 have small total variation distance in Lemma 33.

Lemma 32. *Suppose that in an iteration $0 \le k < K$ of Algorithm 10, the following three conditions hold for some parameters R_x, C_ξ, $C_x \in \mathbb{R}_{\ge 0}$:*

1. $\|x_k - x^*\|_2 \le R_x$.
2. $\|\xi_k\|_2 \le C_\xi \sqrt{d}$.
3. *For all $i \in [n]$, $|\nabla f_i(x_k)^\top \xi_k| \le C_x \|\nabla f_i(x_k)\|_2$.*

Then, for any

$$h \le \frac{1}{98 C_x^2 L^2 R_x^2 + 7 L C_\xi^2 d}, \tag{5.22}$$

$\frac{3}{4} \le \sqrt{\frac{\exp(-F(y_{k+1}))}{\exp(-F(x_k))}} \le \frac{4}{3}$. *Moreover, we have $\mathbb{E}[\gamma_k] = \sqrt{\frac{\exp(-F(y_{k+1}))}{\exp(-F(x_k))}}$, and when $|S_k| \le 2pn$, $\gamma_k \le \frac{4}{3}$.*

Proof. We first show $\mathbb{E}[\gamma_k] = \sqrt{\frac{\exp(-F(y_{k+1}))}{\exp(-F(x_k))}}$. Since each $i \in S_k$ is generated independently,

$$\begin{aligned}
\mathbb{E}[\gamma_k] &= \prod_{i \in [n]} \mathbb{E}\left[\gamma_k^{(i)}\right] \\
&= \prod_{i \in [n]} \left[(1-p) + p\left(\frac{1}{p}\left(\sqrt{\exp\left(-\frac{1}{n} f_i(y_{k+1}) + \frac{1}{n} f_i(x_k)\right)} - 1\right) + 1\right)\right] \\
&= \prod_{i \in [n]} \sqrt{\exp\left(-\frac{1}{n} f_i(y_{k+1}) + \frac{1}{n} f_i(x_k)\right)} = \sqrt{\frac{\exp(-F(y_{k+1}))}{\exp(-F(x_k))}}.
\end{aligned}$$

Next, for any $i \in [n]$, we lower and upper bound $-f_i(y_{k+1}) + f_i(x_k)$. First,

$$\begin{aligned}
-f_i(y_{k+1}) + f_i(x_k) &\le \nabla f_i(x_k)^\top (x_k - y_{k+1}) \\
&\le \sqrt{2h} C_x \|\nabla f_i(x_k)\|_2 \le \sqrt{2h} C_x L R_x.
\end{aligned}$$

The first inequality followed from convexity of f_i, the second from $y_{k+1} - x_k = \sqrt{2h}\xi_k$ and our assumed bound, and the third from smoothness and $\nabla f(x^*) = 0$. To show a lower bound,

$$\begin{aligned}
f_i(y_{k+1}) - f_i(x_k) &\le \nabla f_i(x_k)^\top (y_{k+1} - x_k) + \frac{L}{2} \|y_{k+1} - x_k\|_2^2 \\
&\le \sqrt{2h} C_x L R_x + h L C_\xi^2 d.
\end{aligned}$$

The first inequality was smoothness. Repeating this argument for each $i \in [n]$ and averaging,

$$-\sqrt{2h}C_x L R_x - hLC_\xi^2 d \le -F(y_{k+1}) + F(x_k) \le \sqrt{2h}C_x L R_x. \tag{5.23}$$

Then, when $h \le \frac{1}{98C_x^2L^2R_x^2+7LC_\xi^2 d}$,

$$\frac{3}{4} \le \sqrt{\frac{\exp(-F(y_{k+1}))}{\exp(-F(x_k))}} \le \frac{4}{3}, \text{ and for all } i \in [n],\ -f_i(y_{k+1}) + f_i(x_k) \le \frac{1}{4}.$$

Thus, we can bound each $\gamma_k^{(i)}$:

$$\gamma_k^{(i)} \le \frac{1}{p}\left(\exp\left(\frac{1}{8n}\right) - 1\right) + 1 \le 1 + \frac{1}{7pn}.$$

Finally, when $|S_k| \le 2pn$, $\gamma_k \le (1 + \frac{1}{7pn})^{2pn} \le \frac{4}{3}$ as desired. □

Lemma 33. *Draw $x_0 \sim \mathcal{N}(x^*, \frac{1}{L}\,\mathrm{id})$. Let $\hat{\pi}_K$ be the output distribution of the algorithm of Definition 4 for K steps starting from x_0, and let π_K be the output distribution of Algorithm 10 starting from x_0. For any $\delta \in [0,1]$, let $p = \frac{5\log\frac{12K}{\delta}}{n}$ in Algorithm 10. There exist*

$$C_\xi = O\left(1 + \sqrt{\frac{\log\frac{K}{\delta}}{d}}\right),\ C_x = O\left(\sqrt{\log\frac{nK}{\delta}}\right),\ \text{ and } R_x = O\left(\sqrt{\frac{d\log\frac{\kappa K}{\delta}}{\mu}}\right),$$

so that when $h \le \frac{1}{98C_x^2L^2R_x^2+7LC_\xi^2 d}$, we have $\|\pi_K - \hat{\pi}_K\|_{\mathrm{TV}} \le \delta$.

Proof. By the coupling definition of total variation, it suffices to upper bound the probability that the algorithms' trajectories, sharing all randomness in proposing points y_{k+1}, differ. This can happen for two reasons: either we used an incorrect filtering step (i.e. the pair (x_k, y_{k+1}) did not lie in the second case of (5.21)), or we incorrectly rejected in Line 7 of Algorithm 10 because $|S_k| \ge 2pn$. We bound the error due to either happening over any iteration by δ, yielding the conclusion.

Incorrect filtering. Consider some iteration k. Lemma 32 shows that as long as its three conditions hold in iteration k, we are in the second case of (5.21), so it suffices to show all conditions hold. By Fact 6 and as ξ_k is independent of all $\{\nabla f_i(x_k)\}_{i\in[n]}$, with probability at least $1 - \frac{\delta}{2K}$, both of the conditions $\|\xi_k\|_2 \le C_\xi\sqrt{d}$ and[9] $|\nabla f_i(x_k)^\top \xi_k| \le C_x\|\nabla f_i(x_k)\|_2$ for all $i \in [n]$ hold for some

$$C_\xi = O\left(1 + \sqrt{\frac{\log\frac{K}{\delta}}{d}}\right),\ C_x = O\left(\sqrt{\log\frac{nK}{\delta}}\right).$$

[9]We recall that the distribution of $v^\top\xi$ for $\xi \sim \mathcal{N}(0, \mathrm{id})$ is the one-dimensional $\mathcal{N}(0, \|v\|_2^2)$.

Next, $x_0 \sim \mathcal{N}(x^*, \frac{1}{L}\,\mathrm{id})$ is drawn from a $\kappa^{\frac{d}{2}}$ warm start for π. By Fact 6, we have $\|x_0 - x^*\|_2 \le R_x$ for x_0 drawn from π with probability at least $1 - \frac{\delta}{4K} \cdot \kappa^{-\frac{d}{2}}$, for some

$$R_x = O\left(\sqrt{\frac{d \log \frac{\kappa K}{\delta}}{\mu}}\right).$$

Since warmness of the exact algorithm of Definition 4 is monotonic, as long as the trajectories have not differed up to iteration k, $\|x_k - x^*\|_2 \le R_x$ also holds with probability $\ge 1 - \frac{\delta}{4K}$. Inductively, the total variation error caused by incorrect filtering over K steps is at most $\frac{3\delta}{4}$.

Error due to large $|S_k|$. Supposing all the conditions of Lemma 32 are satisfied in iteration k, we show that with high probability, `Inefficient-MRW` and Algorithm 10 make the same accept or reject decision. By Lemma 32, `Inefficient-MRW` (5.21) accepts with probability $\alpha_k' = \frac{3}{4}\sqrt{\frac{\exp(-F(y_{k+1}))}{\exp(-F(x_k))}}$. On the other hand, Algorithm 10 accepts with probability

$$\alpha_k = \frac{3}{4}\mathbb{E}\left[\gamma_k \mid |S_k| \le 2pn\right] \cdot \Pr[|S_k| \le 2pn].$$

The total variation between the output distributions is $|\alpha_k - \alpha_k'|$. Further, since by Lemma 32,

$$\begin{aligned}
\alpha_k' &= \frac{3}{4}\mathbb{E}\left[\gamma_k\right] \\
&= \frac{3}{4}\left(\mathbb{E}\left[\gamma_k \mid |S_k| \le 2pn\right] \cdot \Pr[|S_k| \le 2pn] + \mathbb{E}\left[\gamma_k \mid |S_k| > 2pn\right] \cdot \Pr[|S_k| > 2pn]\right) \\
&= \alpha_k + \frac{3}{4}\mathbb{E}\left[\gamma_k \mid |S_k| > 2pn\right] \cdot \Pr[|S_k| > 2pn],
\end{aligned}$$

it suffices to upper bound this latter quantity. First, by Lemma 34, when $p = \frac{5 \log \frac{12K}{\delta}}{n}$, we have $\Pr[|S_k| > 2pn] \le \frac{\delta}{12K}$. Finally, since each $i \in S_k$ is generated independently,

$$\begin{aligned}
\mathbb{E}\left[\gamma_k \mid |S_k| > 2pn\right] &\le \max_{S' : |S'| = 2pn} \mathbb{E}\left[\prod_{i \in [n]} \gamma_k^{(i)} \mid S' \subseteq S_k\right] \\
&\le 2\mathbb{E}\left[\prod_{i \in [n] \setminus S'} \gamma_k^{(i)}\right] = 2\sqrt{\prod_{[n] \setminus S'} \exp\left(-\frac{1}{n} f_i(y_{k+1}) + \frac{1}{n} f_i(x_k)\right)} \le 4.
\end{aligned}$$

Here, we used Lemma 32 applied to the set S', and the upper bound (5.23) we derived earlier. Combining these calculations shows that the total variation distance incurred in any iteration k due to $|S_k|$ being too large is at most $\frac{\delta}{4K}$, so the overall contribution over K steps is at most $\frac{\delta}{4}$. $\square$

We used the following helper lemma in our analysis.

Lemma 34. *Let $S \subseteq [n]$ be formed by independently including each $i \in [n]$ with probability p. Then,*

$$\Pr\left[|S| > 2pn\right] \leq \exp\left(-\frac{3pn}{14}\right).$$

Proof. For $i \in [n]$, let $\mathbf{1}_{i\in S}$ be the indicator random variable of the event $i \in S$, so $\mathbb{E}\left[\mathbf{1}_{i\in S}\right] = p$ and

$$\mathrm{Var}\left[\mathbf{1}_{i\in S} - p\right] = p(1-p)^2 + (1-p)p^2 \leq 2p.$$

By Bernstein's inequality,

$$\Pr\left[\sum_{i\in[n]} \mathbf{1}_{i\in S} \geq np + r\right] \leq \exp\left(-\frac{\frac{1}{2}r^2}{2np + \frac{1}{3}r}\right).$$

In particular, when $r = pn$, we have the desired conclusion. □

5.6.2 Conductance analysis

We next bound the mixing time of `Inefficient-MRW`, using the following result from prior work. We remark that (see Section 5.1.4) in our application, the $\log \beta$ term is non-dominant.

Proposition 12 (Lemma 1, Lemma 2, [CDWY20])**.** *Let a random walk with a μ-strongly logconcave stationary distribution π on $x \in \mathbb{R}^d$ have transition distributions $\{\mathcal{T}_x\}_{x\in\mathbb{R}^d}$. For some $\epsilon \in [0,1]$, let convex set $\Omega \subseteq \mathbb{R}^d$ have $\pi(\Omega) \geq 1 - \frac{\epsilon^2}{2\beta^2}$. Let π_{start} be a β-warm start for π, and let the algorithm be initialized at $x_0 \sim \pi_{\mathrm{start}}$. Suppose for any $x, x' \in \Omega$ with $\|x - x'\|_2 \leq \Delta$,*

$$\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\mathrm{TV}} \leq \frac{7}{8}. \tag{5.24}$$

Then, the random walk mixes to total variation distance within ϵ of π in $O(\log\beta + \frac{1}{\Delta^2\mu}\log\frac{\log\beta}{\epsilon})$ iterations.

Consider an iteration of `Inefficient-MRW` from $x_k = x$. Let $\mathcal{P}_x$ be the density of y_{k+1}, and let $\mathcal{T}_x$ be the density of x_{k+1} after filtering. Define a convex set $\Omega \subseteq \mathbb{R}^d$ parameterized by $R_\Omega \in \mathbb{R}_{\geq 0}$:

$$\Omega = \{x \in \mathbb{R}^d : \|x - x^*\|_2 \leq R_\Omega\}.$$

We show that for two close points $x, x' \subseteq \Omega$, the total variation between $\mathcal{T}_x$ and $\mathcal{T}_{x'}$ is small.

Lemma 35. *For some* $h = O(\frac{1}{L^2R_\Omega^2+Ld})$ *and* $x, x' \subseteq \Omega$ *with* $\|x - x'\|_2 \le \frac{1}{8}\sqrt{h}$, $\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\mathrm{TV}} \le \frac{7}{8}$.

Proof. By the triangle inequality of total variation distance,

$$\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\mathrm{TV}} \le \|\mathcal{T}_x - \mathcal{P}_x\|_{\mathrm{TV}} + \|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\mathrm{TV}} + \|\mathcal{T}_{x'} - \mathcal{P}_{x'}\|_{\mathrm{TV}}.$$

First, by Pinsker's inequality and the KL divergence between Gaussian distributions,

$$\|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\mathrm{TV}} \le \sqrt{2\mathrm{KL}(\mathcal{P}_x||\mathcal{P}_{x'})} = \frac{\|x - x'\|_2}{\sqrt{2h}}.$$

When $\|x - x'\|_2 \le \frac{1}{8}\sqrt{h}$, $\|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\mathrm{TV}} \le \frac{1}{8}$. Next, we bound $\|\mathcal{T}_x - \mathcal{P}_x\|_{\mathrm{TV}}$: by a standard calculation (e.g. Lemma D.1 of [LST20]), we have

$$\|\mathcal{T}_x - \mathcal{P}_x\|_{\mathrm{TV}} = 1 - \frac{3}{4}\mathbb{E}_{\xi_{k+1}}\left[\sqrt{\frac{\exp\left(-F(y_{k+1})\right)}{\exp\left(-F(x_k)\right)}}\right].$$

We show that $\|\mathcal{T}_x - \mathcal{P}_x\|_{\mathrm{TV}} \le \frac{3}{8}$. It suffices to show that $\mathbb{E}_{\xi_{k+1}}\left[\sqrt{\exp\left(-F(y_{k+1}) + F(x_k)\right)}\right] \ge \frac{5}{6}$.

Since $\frac{15}{16}\sqrt{\exp\left(-\frac{1}{16}\right)} \ge \frac{5}{6}$, it suffices to show that with probability at least $\frac{15}{16}$ over the randomness of ξ_{k+1}, $-F(y_{k+1}) + F(x_k) \ge -\frac{1}{16}$. As $\xi_{k+1} \sim \mathcal{N}(0, \mathbb{I}_d)$, by applying Fact 6 twice,

$$\begin{aligned} \Pr\left[\|\xi_{k+1}\|_2^2 > 36d\right] &\le \exp(-4) \le \frac{1}{32}, \\ \Pr\left[\left|\nabla F(x_k)^\top \xi_{k+1}\right|^2 \ge 36\|\nabla F(x_k)\|_2^2\right] &\le \frac{1}{32}. \end{aligned} \tag{5.25}$$

We upper bound the term $F(y_{k+1}) - F(x_k)$ by smoothness and Cauchy-Schwarz:

$$\begin{aligned} F(y_{k+1}) - F(x_k) &\le \nabla F(x_k)^\top (y_{k+1} - x_k) + \frac{L}{2}\|y_{k+1} - x_k\|_2^2 \\ &\le \sqrt{2h}\left|\nabla F(x_k)^\top \xi_{k+1}\right| + hL\|\xi_{k+1}\|_2^2. \end{aligned}$$

Then, since $\|\nabla F(x_k)\| \le LR_\Omega$ when $x \in \Omega$, it is enough to choose $h = O(\frac{1}{L^2R_\Omega^2+Ld})$ so that

$$-F(y_{k+1}) + F(x_k) \ge -\frac{1}{16},$$

as long as the events of (5.25) hold, which occurs with probability at least $\frac{15}{16}$. Similarly, we can show that $\|\mathcal{T}_{x'} - \mathcal{P}_{x'}\|_{\mathrm{TV}} \le \frac{3}{8}$. Combining the three bounds, we have the desired conclusion.

□

Theorem 12. *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-F(x))$, where $F(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ is μ-strongly convex, f_i is L-smooth and convex $\forall i \in [n]$, $\kappa = \frac{L}{\mu}$, and $\epsilon \in (0,1)$. Assume access to $x^* = \operatorname{argmin}_{x\in\mathbb{R}^d} F(x)$. Algorithm 10 uses $O\left(\kappa^2 d \log^4 \frac{n\kappa d}{\epsilon}\right)$ value queries to summands $\{f_i\}_{i\in[n]}$, and obtains ϵ total variation distance to π.*

Proof. First, $\mathcal{N}(x^*, \frac{1}{L}\,\mathrm{id})$ yields a $\beta = \kappa^{\frac{d}{2}}$-warm start for π (see e.g. [DCWY19]). For this value of β, by Fact 6 it suffices to choose

$$R_\Omega = \Theta\left(\sqrt{\frac{d\log\frac{\kappa}{\epsilon}}{\mu}}\right)$$

for $\pi(\Omega) \geq 1 - \frac{\epsilon^2}{2\beta^2}$. Letting $\delta = \frac{\epsilon}{2}$, we will choose the step size h and iteration count K so that

$$\frac{1}{h} = \Theta\left(L\kappa d \log^2 \frac{n\kappa d}{\epsilon}\right),\ K = \Theta\left(\kappa^2 d \log^3 \frac{n\kappa d}{\epsilon}\right)$$

have constants compatible with Lemma 33. Note that this choice of h is also sufficiently small to apply Lemma 35 for our choice of R_Ω. By applying Proposition 12 to the algorithm of Definition 4, and using the bound from Lemma 35, in K iterations `Inefficient-MRW` will mix to total variation distance δ to π. Furthermore, applying Lemma 33, we conclude that Algorithm 10 has total variation distance at most $2\delta = \epsilon$ from π.

It remains to bound the oracle complexity of Algorithm 10. Note in every iteration, we never compute more than $4pn$ values of $\{f_i\}_{i\in[n]}$, since we always reject if $|S_k| \geq 2pn$, and we only compute values for indices in S_k. For the value of p in Lemma 33, this amounts to $O(\log \frac{n\kappa d}{\epsilon})$ value queries. □

Chapter 6

ALGORITHMIC ASPECTS OF THE LOG-LAPLACE TRANSFORM AND A NON-EUCLIDEAN PROXIMAL SAMPLER

This chapter is based on [GLL+23a], with Sivakanth Gopi, Yin Tat Lee, Daogao Liu, and Kevin Tian.

6.1 Introduction

The theory of continuous optimization under regularity assumptions stated for non-Euclidean geometries has played an important role in algorithm design. These geometries naturally arise when the optimization problem is over a structured constraint set, such as an ℓ_p ball or a polytope. In diverse applications such as learning from experts [AHK12], sparse recovery [CRT06], multi-armed bandits [BC12], matrix completion [ANW10], fair resource allocation [DFO20], and robust PCA [JLT20], first-order mirror descent techniques for ℓ_p or Schatten-p geometries have been a remarkable success story. Beyond these applications, the theory of self-concordant barriers (and the Riemannian geometries induced by their Hessians) has been greatly influential to the theory of convex programming and interior point methods [NT02, Nem04].[1]

Non-Euclidean samplers. A natural direction for building the theory of logconcave sampling (the analog of convex optimization) is thus to develop samplers which can handle non-Euclidean regularity assumptions and constraint sets. Unfortunately, progress in this direction has relatively lagged behind optimization counterparts, as discretization tools which work well in the Euclidean case do not readily generalize. Briefly (with an extended discussion deferred to Section 6.1.3), most prior attempts at giving non-Euclidean samplers have focused on analyzing variants of the *mirrored Langevin dynamics*, building upon the ubiquitous mirror descent algorithm in optimization [NY83]. The key idea of mirror descent is to choose a regularizer $\phi : \mathcal{X} \to \mathbb{R}$ over a constraint set $\mathcal{X}$, such that ϕ is strongly convex in an appropriate (possibly non-Euclidean) norm $\|\cdot\|_{\mathcal{X}}$. The regularizer ϕ is then

[1] Self-concordance requires that the second derivative of a function is stable to perturbations which are measured in the induced norm. For notation and definitions used throughout the paper, see Section 6.2.

used to define iterative methods for optimizing functions f with regularity in $\|\cdot\|_{\mathcal{X}}$.

The sampling analog of this non-Euclidean generalization is to extend the *Langevin dynamics*, a stochastic process inherently catered to the ℓ_2 geometry, to use Brownian motion reweighted by the Hessian of a regularizer ϕ. This process, which we call the mirrored Langevin dynamics (MLD), was introduced recently by [ZPFP20] (see also [HKRC18] for an earlier incarnation). Several follow-up works attempted to bound convergence rates for discretizations of the MLD process, e.g. [AC21, Jia21, LTVW22]. Unfortunately, many of these analyses have imposed rather strong conditions on ϕ beyond strong convexity, e.g. a "modified self-concordance" assumption used in [ZPFP20, Jia21, LTVW22] which (to our knowledge) is not known to be satisfied by standard regularizers. Even more problematically, these analyses (as well as an empirical evaluation by [Jia21]) suggest that without strong relative regularity assumptions between the target density and ϕ, naïve discretizations of MLD inherently do not converge to the target even in the limit. A notable exception is the work of [AC21], which circumvented both issues (the modified self-concordance assumption and a biased limit) using a different MLD discretization; however, it is not always clear that this discretization is feasible for standard choices of ϕ and $\mathcal{X}$.

An alternative to directly discretizing MLD is to use a filter to control bias, akin to the MALA or Metropolized HMC algorithms which are well-studied in the Euclidean case [Bes94, RT96a, BRH13, DCWY19, CDWY20, LST20]. However, here too generalizing existing analyses runs into obstacles: for example, typical analyses of MALA and Metropolized HMC rely on bounding the conductance of random walks via isoperimetric inequalities on the target distribution. Prior isoperimetry bounds appear to be tailored to the ℓ_2 geometry and properties of Gaussians (the basic strongly logconcave distribution in Euclidean settings). Potentially due to this difficulty, to our knowledge no general-purpose extension of MALA or its variants to non-Euclidean norms exists in the literature.[2]

Proximal samplers. In this paper, we overcome these difficulties by following a third strategy for the design of efficient samplers: a proximal approach [LST21b], as discussed in Chapter 5. To sample from a density π on $\mathbb{R}^d$ proportional to $\exp(-f)$, the algorithm of [LST21b] first extends the space to $\mathbb{R}^d \times \mathbb{R}^d$, and defines a joint density $\hat{\pi}$ such that, for

[2]We mention that in certain geometries induced by structured manifolds (discussed in part in Section 6.1.3), generalizations of MALA or Metropolized HMC have been previously proposed, e.g. [GC11, Bar20]. These works are motivated by related, but different, settings to the ones considered in this work (we mainly study norm regularity, akin to first-order convex optimization), and their focus is not on establishing non-asymptotic mixing time bounds.

some parameter $\eta > 0$,

$$\mathrm{d}\hat{\pi}(z) \propto \exp\left(-f(x) - \frac{1}{2\eta}\|x-y\|_2^2\right)\mathrm{d}z \text{ where } z = (x,y) \in \mathbb{R}^d \times \mathbb{R}^d. \tag{6.1}$$

It is straightforward to see that for any η, the x-marginal of $\hat{\pi}$ is the original distribution π, and further [LST21b] shows that alternating sampling from the conditional distributions of $\hat{\pi}$, i.e. $\hat{\pi}(x \mid y)$ or $\hat{\pi}(y \mid x)$, mixes rapidly. We give an extended discussion on recent activity on designing and harnessing proximal samplers building upon [LST21b] in Section 6.1.3, but mention that instantiations of the framework have resulted in state-of-the-art runtimes for many structured density families [CCSW22, LC22, GLL22]. Motivated by the success of proximal methods in the Euclidean setting, one goal of our work is to extend this technique to non-Euclidean geometries.

Our approach. Our main insight is that a generalization of the strategy in [LST21b] induces a well-studied object in probability theory called the *log-Laplace transform* (LLT). Letting $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a convex function in the dual space $y \in \mathbb{R}^d$, our generalization of (6.1) defines the joint density

$$\begin{aligned} \mathrm{d}\hat{\pi}(z) &\propto \exp\left(-f(x) + (\langle x, y\rangle - \varphi(y) - \psi(x))\right)\mathrm{d}z, \\ \text{where } \psi(x) &\stackrel{\text{def}}{=} \log\left(\int \exp\left(\langle x, y\rangle - \varphi(y)\right)\mathrm{d}y\right). \end{aligned} \tag{6.2}$$

The function ψ is called the LLT of φ, and it has an interpretation as a normalizing constant for induced densities $\mathcal{D}_x^\varphi$ on the dual space proportional to $\exp(\langle x, \cdot\rangle - \varphi)$. Indeed, $\mathcal{D}_x^\varphi$ is defined exactly so the x-marginal of $\hat{\pi}$ is $\pi \propto \exp(-f)$. When $\eta = 1$ and φ, ψ are quadratics, this is exactly (6.1); we discuss the case of general η in Section 6.1.2. Moreover, the LLT is a well-studied mathematical object: it arises in probability theory as a *cumulant-generating function*, i.e. derivatives of the LLT yield cumulants of the induced distributions $\mathcal{D}_x^\varphi$, just as derivatives of the MGF yield moments.

The LLT famously appeared in Cramér's theorem on large deviations [Cra38], and its cumulant-generating properties have yielded fundamental concentration results in convex geometry [Kla06, EK11, KM12]. More recently, algorithmically-motivated properties of the LLT have been studied in settings such as optimization [BE19], where it was used to define an optimal self-concordant barrier, as well as connections to localization schemes for sampling from discrete distributions [CE22].

We continue this investigation by demonstrating new mathematical properties of the LLT with an algorithmic flavor, and showcasing uses of the LLT as a tool for continuous

logconcave sampling. In particular, armed with a deeper understanding of the LLT, we overcome several of the aforementioned barriers to non-Euclidean sampler design and develop a generalized proximal sampler. We further give applications of our sampler to obtain new complexity results for non-Euclidean differentially private convex optimization, building upon a connection discovered by [GLL22, GLL$^+$23b]. We are optimistic that the LLT will find additional uses in sampler design (potentially beyond the proximal sampling framework, building upon the new properties we prove), and suggest a number of avenues of future exploration to the community in Section 6.6.

6.1.1 Our results

In this section, we overview our results, which separate cleanly into three categories.

Algorithmic aspects of the LLT. It is well-known that the derivatives of the LLT at a point $x \in \mathbb{R}^d$ are *cumulants* of the induced density on $y \in \mathbb{R}^d$:

$$\mathrm{d}\mathcal{D}_x^\varphi(y) \propto \exp\left(\langle x, y\rangle - \varphi(y)\right) \mathrm{d}y.$$

For example, $\nabla\psi(x) = \mathbb{E}_{y\sim\mathcal{D}_x^\varphi}[y]$, and $\nabla^2\psi(x)$ is the covariance of $\mathcal{D}_x^\varphi$. Further, it was shown in [BE19] that if ψ is the LLT of a convex function φ, then ψ is convex and self-concordant. Building upon these facts, in Section 6.3, we prove the following new properties of the LLT.

- *Strong convexity-smoothness duality.* Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. We prove that if $\varphi : \mathbb{R}^d \to \mathbb{R}$ is L-smooth in the dual norm $\|\cdot\|_*$, its LLT $\psi : \mathbb{R}^d \to \mathbb{R}$ is $\frac{1}{L}$-strongly convex in $\|\cdot\|$.[3] This fact parallels a similar, well-known form of strong convexity-smoothness duality for Fenchel conjugates [Sha07, KST09]. Our proof does not require φ to be convex. We further show that the converse holds as well: a $\frac{1}{L}$-strongly convex φ has a L-smooth LLT.

- *Isoperimetry in the Hessian norm.* We prove a one-dimensional isoperimetric inequality for densities of the form $\exp(-\phi)$, where $\phi : \mathbb{R} \to \mathbb{R}$ is self-concordant and convex. By appealing to (a strong variant of) the localization lemma of [LS93], this proves that measures which are strongly logconcave with respect to convex and self-concordant $\phi : \mathbb{R}^d \to \mathbb{R}$ satisfy a similar isoperimetric inequality in the Riemannian

[3]The constant factor 1 here is optimal, as demonstrated by quadratics.

geometry induced by $\nabla^2\phi$. Importantly, due to self-concordance of the LLT, this applies to strongly logconcave measures in an LLT.

- *Overlap of induced distributions $\mathcal{D}_x^\varphi$.* We provide a KL divergence bound on the distributions $\mathcal{D}_x^\varphi$ and $\mathcal{D}_{x'}^\varphi$ for x and x' which are close in the Riemannian distance induced by ψ. Combined with our isoperimetric inequality and a classical argument of [DFK91b], this proves a lower bound on the conductance of an alternating sampler for densities of the form (6.2).

These new properties of the LLT suggest that it may find uses in designing samplers under non-Euclidean geometries beyond those explored in Sections 6.4 and 6.5 of our paper. For example, the LLT of a smooth function is strongly convex and self-concordant, which are exactly the properties required by the mirror Langevin discretization scheme of [AC21]. In optimization, regularizers ϕ for mirror descent typically only require strong convexity (and not self-concordance). However, controlling the evolution of the geometry induced by $\nabla^2\phi$ is critical for discretizing MLD schemes, so imposing self-concordance (as opposed to more non-standard regularity such as the modified self-concordance of [ZPFP20, Jia21, LTVW22]) may be viewed as a minimal assumption. Problematically, standard strongly convex regularizers for mirror descent such as entropy or ℓ_p^2 *are not* self-concordant, so LLTs are a way of bridging this gap for sampling. Moreover, our new isoperimetric inequality and conductance bounds suggest that LLTs may find use in Metropolized sampling schemes, paving the way for non-Euclidean generalizations of MALA and its variants.

In some sense, our new duality result is a generic way of taking a strongly convex regularizer and transform it, via the *Fenchel transform* and the *log-Laplace transform*, to another regularizer which is strongly convex in the same norm, but also self-concordant. The first transform makes the function smooth in the dual [KST09], and the second effectively undoes this change. We will later discuss an application of this framework in improving the oracle complexity of the problem of private stochastic convex optimization in the ℓ_p geometry, using the LLT of the ℓ_q^2 regularizer.

Non-Euclidean proximal sampling. In Section 6.4, we build upon these aforementioned tools to analyze the mixing time of an alternating scheme for sampling densities π on convex, compact $\mathcal{X} \subset \mathbb{R}^d$ equipped with a norm $\|\cdot\|_{\mathcal{X}}$, where π is proportional to

$\exp\left(-F(x) - \eta\mu\psi(x)\right) \mathbb{1}_{\mathcal{X}}(x)$. Here, $F : \mathcal{X} \to \mathbb{R}$ is convex, $\eta, \mu > 0$ are tunable parameters, and ψ is the LLT of η-smooth $\varphi : \mathbb{R}^d \to \mathbb{R}$ in the dual norm $\|\cdot\|_{\mathcal{X}^*}$. We prove in Theorem 13 that alternately sampling from conditional distributions of the extended density on $z = (x, y) \in \mathcal{X} \times \mathbb{R}^d$ proportional to

$$\exp\left(-F(x) - \eta\mu\psi(x) + \left(\langle x, y\rangle - \varphi(y) - \psi(x)\right)\right) \mathbb{1}_{\mathcal{X}}(x) \tag{6.3}$$

has stationary distribution π, and converges in $\approx \frac{1}{\eta\mu}$ iterations for a warm start. More specifically, the convergence rate of our sampler depends polylogarithmically on both the warmness β of the point it is initialized with, and the inverse of the total variation error δ. The form of (6.3) is the same as (6.2), but we impose that f is $\eta\mu$-relatively strongly convex in ψ.

We first compare this result to the Euclidean proximal sampler of [LST21b], who proved a similar result for alternating sampling densities of the form (6.1). The main result of [LST21b] shows that if f is μ-strongly convex in the ℓ_2 norm, then alternating sampling from the marginals of (6.1) converges in $\approx \frac{1}{\eta\mu}$ iterations, also with polylogarithmic dependence on the target total variation error. Our result can be viewed as an extension of this result; instead of requiring μ-strong convexity in the ℓ_2 norm (which is equivalent to relative strong convexity with respect to the function $x \to \frac{1}{2}\|x\|_2^2$), we require μ-relative strong convexity in the function $\eta\psi$. In light of our duality result, $\eta\psi$ is 1-strongly convex in $\|\cdot\|_{\mathcal{X}}$, so it is the natural "unit" for measuring strong convexity.

We remark that the parameters η and μ play different roles: μ governs the strong logconcavity of the stationary distribution, and η controls the strong logconcavity of the x-conditional distribution of (6.3), which is tuned to govern the convergence rate of sampling from the conditional distribution. In particular, we further show that when F is G-Lipschitz in $\|\cdot\|_{\mathcal{X}}$, then as long as $\eta \lesssim G^{-2}$, the conditional sampling required by (6.3) can be performed in constant calls to a value oracle to F in expectation. This result holds even when F is a distribution over G-Lipschitz functions, and we only have sample access to this distribution. This extends a similar implementation of the marginal sampler required by [LST21b] for log-Lipschitz densities in the ℓ_2 norm, given by [GLL22]. The remaining complexity of the marginal sampling depends on the structure of the chosen φ and $\mathcal{X}$, but is independent of F; we give a discussion of this aspect of our sampler in Sections 6.5.3 and 6.6.

One shortcoming of Theorem 14's rate is that it depends polylogarithmically on the

warmness parameter. In contrast, the rate of [LST21b] depends *doubly logarithmically* on the warmness, which is important because in many sampling applications, standard starting distributions have warmness bounds growing exponentially in problem parameters such as the dimension d. We refer the reader to a discussion in Section 1.1 of [LST21a] on warmness assumptions under ℓ_2 geometry, which have created a $\approx \sqrt{d}$-sized gap on mixing time bounds for MALA, with and without a polynomially-bounded warm start [CLA$^+$21, LST20]. We believe it is an interesting future direction to close this gap in warmness assumptions for our sampler in Section 6.4, analogously to the result of [LST21b]. Notably, there has been an ongoing exploration of new proof techniques for the convergence of proximal samplers by the community [CCSW22, CE22], and we are optimistic similar advancements can be made in non-Euclidean settings, discussed further in Section 6.1.3.

Zeroth-order private convex optimization. As the main application of our techniques, in Section 6.5 we design LLTs based on the smoothness of the function $\varphi_q(x) = \frac{p-1}{2}\|x\|_q^2$ in the norm ℓ_q, where $\frac{1}{p} + \frac{1}{q} = 1$ and $p \in [1, 2]$, $q \geq 2$. We show that the additive range of $\psi_{\eta,p}$,[4] the LLT of $\eta\varphi_q$ for $\eta \lesssim \frac{1}{d}$,[5] is bounded by $O(\frac{1}{(p-1)\eta})$ over the unit ℓ_p ball. This makes $\eta\psi_{\eta,p}$ competitive with the canonical choice of regularizer in ℓ_p norms for optimization, namely $r_p(x) \stackrel{\text{def}}{=} \frac{1}{2(p-1)}\|x\|_p^2$, which has the same additive range and strong convexity parameters as $\eta\psi_{\eta,p}$ (up to constants). We further build efficient value oracles and samplers for induced densities for $\psi_{\eta,p}$ in Section 6.5.3.

A critical difference between $\eta\psi$ and r_p, however, is that regularizing by a multiple of $\eta\psi$ admits efficient samplers via the machinery in Section 6.4; to our knowledge no similar technique is known for r_p. This difference is particularly important in the setting of *differentially private convex optimization*: see Problem 2 for a formal statement of the problem we study. Recently, [GLL$^+$23b] showed that to privately minimize either population or empirical risk for a distribution over convex functions which are Lipschitz in a (possibly non-Euclidean) norm $\|\cdot\|_{\mathcal{X}}$, it suffices to sample from a regularized density $\propto \exp(-k(F_{\text{erm}} + \mu r))$. Here, $F_{\text{erm}} = \frac{1}{n}\sum_{i\in[n]} f_i$ is the empirical risk over n samples $\{f_i\}_{i\in[n]}$, k, μ are tunable parameters, and r is a 1-strongly convex regularizer in $\|\cdot\|_{\mathcal{X}}$.

Our new sampling results show a demonstrable algorithmic advantage of using $\eta\psi_{\eta,p}$ as a regularizer for ℓ_p geometries, as opposed to r_p. In Theorem 14, we give algorithms for

[4] We use slightly different notation than in Section 6.5 for convenience of exposition here.

[5] This restriction is discussed further in Section 6.1.2, but does not bottleneck our privacy applications.

private convex optimization matching the state-of-the-art excess risk bounds for private convex optimization recently attained by [GLL$^+$23b] (who used r_p as their regularizer). Under a warm start, our new algorithms further improve the *value (zeroth-order) oracle* complexities of private convex optimization under ℓ_p regularity in dimension d compared to [GLL$^+$23b] by poly(d) factors, i.e. the number of queries to $\{f_i\}_{i\in[n]}$ used. We also show these new value oracle complexities extend straightforwardly to improve private convex optimization over matrix spaces satisfying Schatten-p norm regularity.

We note that our results match (up to logarithmic factors) the value oracle complexities in the ℓ_2 setting obtained by [GLL22], for all ℓ_p norms where $p \in [1, 2]$. In Appendix E.1, we extend lower bounds for stochastic optimization from [DJWW15, GLL22] to the ℓ_p setting to show the value oracle complexities of Theorems 13 and 14 are near-optimal, assuming a polynomially warm start.

6.1.2 Our techniques

Analogously to Section 6.1.1, in this section we split our discussion of our techniques into three parts.

Algorithmic aspects of the LLT. We first discuss our strong convexity-smoothness duality result. From a convex geometry perspective, smoothness of φ (with LLT ψ) ensures that the induced distributions $\propto \exp(\langle x, \cdot\rangle - \varphi)$ are heavy-tailed (because their log-densities cannot grow quickly), which means their variances are "large." We also know that $\nabla^2\psi$ is the covariance matrix of the induced distribution which means that $\nabla^2\psi$ should be lower-bounded. We formalize this using a version of the Cramér-Rao bound from [CP22]. An older arXiv version of this paper contains a more elementary proof of this result inspired by differential privacy, achieving a worse constant of $\approx \frac{1}{12}$; the (optimal) improvement was suggested by Sam Power. Our converse proof is similar, and follows by applying the Brascamp-Lieb inequality [BL76].

To prove our isoperimetric inequality, we draw inspiration from a similar bound shown in Lemma 35 of [LV18], but for a family of convex functions ϕ satisfying a strange condition that ϕ'' was convex (which fortunately includes the log barrier function). Noticing that $-\log$ is self-concordant, we extend the [LV18] result to hold for all self-concordant functions. Further we show by a direct calculation that the KL divergence between the induced distributions of two nearby points x and x' is essentially the LLT ψ at one of the

points, up to a linear term. This lets us use stability of the Hessian of self-concordance functions to demonstrate stability of nearby induced distributions, a key ingredient in proving conductance bounds by the machinery of [DFK91b].

Non-Euclidean proximal sampling. Given the results of Section 6.3, establishing our main proximal sampling result Theorem 13 is fairly routine. Our algorithm consists of an "outer loop" and an "inner loop" for sampling from the x marginal of (6.3) which is stated and analyzed in Section 6.4.1. Our outer loop analysis is directly based on the mixing time-to-conductance reduction of [LS93] and the technique of [DFK91b] to lower bound conductance, using facts from Section 6.3. Our inner loop handling functions F in (6.3) which are Lipschitz (or distributions over Lipschitz functions) is a small modification of a similar result in [GLL22]. The only property we need of the LLT is strong convexity: this implies a rejection sampler terminates quickly via the concentration of Lipschitz functions under strongly logconcave distributions (in any norm) [Led99, BL00].

We do note there is a design decision to be made on how to define "scaling up the LLT by $\frac{1}{\eta}$," unlike in the case of (6.1) where using the induced density $\mathcal{N}(x, \eta^{-1} \mathrm{id}_d)$ is natural. Given r, a 1-strongly convex function in $\|\cdot\|_{\mathcal{X}}$, and letting r^* be its (smooth) Fenchel conjugate, two natural ways of defining a scaled up induced distribution at x are to choose densities

$$\propto \exp\left(\langle x, y\rangle - \eta r^*(y) - \psi(x)\right), \tag{6.4}$$

or

$$\propto \exp\left(\frac{1}{\eta}\left(\langle x, y\rangle - r^*(y) - \psi(x)\right)\right). \tag{6.5}$$

The choice (6.4) clearly results in ψ which is $\Omega(\eta^{-1})$-strongly convex, rendering it suitable for our proximal sampling applications. It is not difficult to see that the second results in $\eta^{-1}\psi$ which is also $\Omega(\eta^{-1})$-strongly convex. More interestingly, plugging in $r = r^* = \frac{1}{2}\|\cdot\|_2^2$ makes (6.1) agree with (6.5) rather than (6.4). Unfortunately, the ψ which results from (6.5) is not self-concordant, as its Hessian scales with η^{-1} and its third derivative with η^{-2}. Our choice to use (6.4) has further implications, elaborated on next, but a deeper understanding of this discrepancy seems interesting.

Zeroth-order private convex optimization. As outlined in Section 6.1.1, the frameworks of [GLL22, GLL$^+$23b] show that to use our proximal sampler for ℓ_p norm private convex optimization, it suffices to design an LLT which has small additive range. Perhaps

surprisingly, we exploit the *non-scale invariance* of LLT for this task: the LLT of $\eta\varphi$ does not behave like η^{-1} times the LLT of φ.[6] To see why this is helpful, consider the case when $\varphi = \frac{1}{2}\|\cdot\|_\infty^2$: then,

$$\psi(x) = \log\left(\int \exp\left(\langle x, y\rangle - \frac{1}{2}\|y\|_\infty^2\right) \mathrm{d}y\right).$$

Although one would hope $\psi(x)$ has additive range comparable to $\frac{1}{2}\|x\|_1^2$, the Fenchel conjugate of $\frac{1}{2}\|x\|_\infty^2$, it is not hard to show that $\psi(e_1) - \psi(0) = \Omega(\sqrt{d})$; we give a proof in Appendix E.2. Intuitively, the ℓ_∞ radius of a typical point $\sim \exp(-\frac{1}{2}\|\cdot\|_\infty^2)$ is about $\sqrt{d}$, and a constant fraction of points on the surface of this ℓ_∞ ball have inner product with e_1 of $\Omega(\sqrt{d})$. This shows the additive range of ψ on the ℓ_1 ball is larger than $\frac{1}{2}\|\cdot\|_1^2$ by dimension-dependent factors.

We show that the non-scale invariance of (6.4) is actually helpful in controlling additive ranges. Specifically, letting ψ_η denote the LLT of $\eta\|x\|_q^2$, we show the additive range of $\eta\psi_\eta$ (a $\approx$ 1-strongly convex function) is $\approx \max(\eta, 1, \sqrt{d\eta})$. For sufficiently small η, this implies $\eta\psi_\eta$ is actually a much smaller regularizer than ψ; graciously, our differential privacy applications require $\eta \lesssim \frac{1}{d^2}$. We find it potentially useful to explore how generic this non-scale invariance of the LLT is.

6.1.3 Prior work

Non-Euclidean sampling. A recurring issue that arises in bounding the convergence rate of non-Euclidean samplers is that naïve discretizations can result in significant error. As a result, most prior works either require strong assumptions or oracles for accurate discretization or adopt more sophisticated discretization methods that are difficult to analyze. For example, earlier in the introduction this was discussed for discretizations of MLD [ZPFP20, Jia21, AC21, LTVW22]. Part of the intrinsic difficulty of bounding discretized MLD lies in third-order error terms emerging from non-Euclidean geometries, which are hard to control under standard assumptions.

Under structured settings different than, but related to, those in this paper, an interesting alternative sampling strategy is discretizing Riemannian Langevin or Hamiltonian dynamics. For example, [GV22] studied the Riemmanian Langevin dynamics assuming access to an oracle to sample from Brownian motion on a manifold, whose complexity

[6] On the other hand, the Fenchel conjugate of $\eta\varphi$ is η^{-1} times the Fenchel conjugate of φ.

heavily depends on the manifold. Further, the convergence rate of Riemannian Hamiltonian Monte Carlo (RHMC) in polytopes was studied in [LV18], and a discretized version was analyzed in [KLSV22b]; the results apply to a limited family of distributions, and the convergence rate is fairly large. For RHMC to converge to the correct target distribution, sophisticated discretization methods such as Implicit Midpoint Method are necessary. Though efficient in practice, these methods are challenging to analyze theoretically.

Proximal sampling. A long line of works has studied the use of proximal methods in sampling (inspired by optimization). Several considered proximal Langevin algorithms [Per16, BDMP17, Ber18, Wib19], which combine proximal methods and discretizations of Langevin dynamics. Further, [MFWB19] proposed a sampler based on a proximal sampling oracle. However, these algorithms required either stringent assumptions or a large mixing time. Recently, [LST21b] proposed a new proximal sampler overcoming many of the assumptions and efficiency issues in prior methods. Several works have focused on generalizing [LST21b] and applying it in different settings: [CCSW22] proved convergence results using weaker assumptions than strong logconcavity. The framework has been used to obtain state-of-the-art samplers for various structured families, including smooth, composite, and finite-sum densities [LST21b] as well as non-smooth densities [GLL22, LC22].

Log-Laplace transform. The LLT is a powerful tool that emerges frequently in probability theory and convex geometry. Notably, [BE19, Che21c] showed that the Legendre-Fenchel dual of LLT of the uniform measure on a convex body in $\mathbb{R}^n$ is an n-self-concordant barrier, giving the first universal barrier for convex bodies with optimal self-concordance parameter. In [CE22], the LLT serves as one of the key ingredients of entropy conservation in localization schemes for sampling. In addition, the LLT shows up in the solution to the entropic optimal transport problem, where a KL divergence is added to regularize the optimal transport objective [CP22].

Private convex optimization. Differentially private convex optimization is one of the most extensively studied problems in the privacy literature and captures an increasing number of critical applications in various domains, including machine learning, statistics, and data analysis. There is a rich body of works on this topic [CM08, CMS11, KST12, BST14, WYX17, BFTGT19, FKT20], which have mainly focused on the Euclidean geometry, e.g. assuming the ℓ_2 diameter of the domain and ℓ_2 norms of gradients are bounded.

Motivated by applications not captured by these assumptions, there has been growing interest in studying differentially private convex optimization in non-Euclidean geometries, as seen in [TTZ15, AFKT21, BGN21, HLL$^+$22, GLL$^+$23b]. Of particular relevance, [GLL$^+$23b] develops an exponential mechanism based method attaining state-of-the-art excess risk bounds for ℓ_p and Schatten-p norms, which are matched by our algorithms in Section 6.5.

6.2 Preliminaries

General notation. In Section 6.1 only, $\widetilde{O}$, $\approx$, and $\lesssim$ hide logarithmic factors in problem parameters for expositional convenience. For $n \in \mathbf{N}$, $[n]$ refers to the naturals $1 \le i \le n$. We use $\mathcal{X}$ to denote a compact convex subset of $\mathbb{R}^d$. For all $p \ge 1$ including $p = \infty$, we let $\|\cdot\|_p$ applied to a vector argument denote the ℓ_p norm. We denote matrices in boldface and when $\|\cdot\|_p$ is applied to a matrix argument it denotes the corresponding Schatten-p norm (ℓ_p norm of the singular values).

For any $\mathcal{X} \subset \mathbb{R}^d$ we let its indicator function (i.e. the function which is 1 on $\mathcal{X}$ and 0 otherwise) be denoted $\mathbb{1}_{\mathcal{X}}$. We will be concerned with optimizing functions $f : \mathcal{X} \to \mathbb{R}$, and $\|\cdot\|_{\mathcal{X}}$ refers to a norm on $\mathcal{X}$. We let $\mathcal{X}^*$ be the dual space to $\mathcal{X}$, and equip it with the dual norm $\|y\|_{\mathcal{X}^*} \stackrel{\text{def}}{=} \sup_{\|x\|_{\mathcal{X}}=1} x^\top y$. We let $\mathcal{N}(\mu, \mathbf{\Sigma})$ be the Gaussian density of given mean and covariance. For a positive definite matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, we denote the induced norm by $\|v\|_{\mathbf{M}} \stackrel{\text{def}}{=} \sqrt{v^\top \mathbf{M} v}$. When making asymptotic statements we will typically assume the dimension d is at least a sufficently large constant, else we can pad and affect statements by at most constant factors.

Optimization. In the following, fix $f : \mathcal{X} \to \mathbb{R}$. We say f is G-Lipschitz in $\|\cdot\|_{\mathcal{X}}$ if for all $x, x' \in \mathcal{X}$, $|f(x) - f(x')| \le G\|x - x'\|_{\mathcal{X}}$. If f is differentiable, we say it is L-smooth in $\|\cdot\|_{\mathcal{X}}$ if for all $x, x' \in \mathcal{X}$, $\|\nabla f(x) - \nabla f(x')\|_{\mathcal{X}^*} \le L\|x - x'\|_{\mathcal{X}}$. Taylor expanding then shows $f(x') \le f(x) + \langle \nabla f(x), x' - x\rangle + \frac{L}{2}\|x - x'\|_{\mathcal{X}}^2$. We say f is m-relatively strongly convex in ϕ if $f - m\phi$ is convex. For k-times differentiable f, $\nabla^k f(x)[v_1, v_2, \ldots, v_k]$ denotes the corresponding k^{th} order directional derivative at f. We say twice-differentiable f is m-strongly convex in $\|\cdot\|_{\mathcal{X}}$ if for all $x \in \mathcal{X}$, $v \in \mathbb{R}^d$, $\nabla^2 f(x)[v, v] \ge m\|v\|_{\mathcal{X}}^2$. We say convex $\phi : \mathbb{R}^d \to \mathbb{R}$ is self-concordant if

$$\left|\nabla^3 \phi(x)[h, h, h]\right| \le 2\left(\nabla^2 \phi(x)[h, h]\right)^{\frac{3}{2}}, \text{ for all } x, h \in \mathbb{R}^d.$$

A key fact we use about self-concordant functions is that their Hessians are stable under small distances, where the distance is measured in the Hessian norm: see Lemma 37 for a formal statement.

Probability. For a density π supported on $\mathcal{X}$, we let $\pi(S) \stackrel{\text{def}}{=} \Pr_{x\sim\pi}[x \in S]$. For two densities μ, π, we define their total variation distance by $\|\mu - \pi\|_{\text{TV}} \stackrel{\text{def}}{=} \frac{1}{2}\int |\mu(x) - \pi(x)|\mathrm{d}x$ and (when the Radon-Nikodym derivative exists) their KL divergence by $d_{\text{KL}}(\mu\|\pi) \stackrel{\text{def}}{=} \int \mu(x) \log \frac{\mu(x)}{\pi(x)}\mathrm{d}x$. For $1 < \alpha < \infty$, we also define the α-Rényi divergence between densities μ, π by

$$D_\alpha(\mu\|\pi) \stackrel{\text{def}}{=} \frac{1}{\alpha - 1} \log\left(\int \left(\frac{\mu(x)}{\pi(x)}\right)^\alpha \pi(x)\mathrm{d}x\right).$$

We say density π is logconcave (respectively, m-strongly logconcave in $\|\cdot\|_{\mathcal{X}}$) if $-\log\pi$ is convex (respectively, m-strongly convex in $\|\cdot\|_{\mathcal{X}}$). We similarly say π is m-relatively strongly logconcave in ϕ if $-\log\pi$ is m-relatively strongly convex in ϕ. If $\log\pi$ is affine, we say π is logaffine. We say a density π_0 is β-warm with respect to a density π if for all x in the support of π, $\frac{\mathrm{d}\pi_0(x)}{\mathrm{d}\pi(x)} \le \beta$.

Log-Laplace transform. We define the log-Laplace transform (LLT) of $\varphi : \mathbb{R}^d \to \mathbb{R}$ by

$$\psi(x) \stackrel{\text{def}}{=} \log\left(\int \exp\left(\langle x, y\rangle - \varphi(y)\right)\mathrm{d}y\right).$$

When φ, ψ are clear from context, we define the density

$$\mathcal{D}_x^\varphi(y) = \exp\left(\langle x, y\rangle - \varphi(y) - \psi(x)\right). \tag{6.6}$$

Note that the normalization constant is exactly given by $\psi(x)$ and hence $\mathcal{D}_x^\varphi$ is indeed a valid density. We use $\propto$ to indicate proportionality, e.g. if μ is a density and we write $\mu \propto \exp(-f)$, we mean $\mu(x) = \frac{\exp(-f)}{Z}$ where $Z \stackrel{\text{def}}{=} \int \exp(-f(x))\mathrm{d}x$ and the integration is over the support of μ.

Riemannian geometry. In Sections 6.3 and 6.4 we will use geometry induced by the Hessian of a self-concordant, convex function $\phi : \mathbb{R}^d \to \mathbb{R}$. We summarize the important points here, and defer a more extended treatment to [NT02]. When ϕ is clear from context, we denote the norm $\|h\|_x \stackrel{\text{def}}{=} \|h\|_{\nabla^2\phi(x)}$. Throughout this discussion let $M \subseteq \mathbb{R}^d$ be a Riemannian manifold equipped with the local metric $\|\cdot\|_x$. The induced Riemannian distance of a curve $c : [0, 1] \to M$ is defined as

$$L_\phi(c) \stackrel{\text{def}}{=} \int_0^1 \|\frac{\mathrm{d}}{\mathrm{d}t}c(t)\|_{c(t)}\mathrm{d}t,$$

where $\frac{\mathrm{d}}{\mathrm{d}t}c(t)$ is the velocity element of the curve in the tangent space at $c(t)$. For $x, y \in M$, we then define $d_\phi(x, y)$ to be the infimum of the length $L_\phi(c)$ over all curves c such that $c(0) = x$ and $c(1) = y$. We will use the following two important properties of the Riemannian geometry over $M = \mathbb{R}^d$ induced by self-concordant, convex functions.

Lemma 36 ([NT02], Lemma 3.1)**.** *Suppose $\phi : \mathbb{R}^d \to \mathbb{R}$ is convex and self-concordant. For $x, y \in \mathbb{R}^d$, if $d_\phi(x, y) \leq \delta - \delta^2 < 1$ for some $\delta \in (0, 1)$, then $\|y - x\|_x \leq \delta$.*

Lemma 37 ([Nem04], Section 2.2.1)**.** *Suppose $\phi : \mathbb{R}^d \to \mathbb{R}$ is convex and self-concordant. For any $h, x \in \mathbb{R}^d$ such that $\|h\|_x < 1$, $(1 - \|h\|_x)^2 \nabla^2 \phi(x) \preceq \nabla^2 \phi(x + h) \preceq (1 - \|h\|_x)^{-2} \nabla^2 \phi(x)$.*

6.3 Properties of the LLT

In this section, we collect a variety of facts about the log-Laplace transform which we will use to develop our sampling scheme in Section 6.4. We begin by proving basic facts about the LLT in Section 6.3.1. We then use them to derive isoperimetric properties of induced distributions in Section D.3.3 and total variation bounds in Section 6.3.3. Throughout this section we will fix a convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$, and let $\psi : \mathbb{R}^d \to \mathbb{R}$ be its LLT. We will also follow the notation (6.6).

6.3.1 Basic properties and duality

The log-Laplace transform ψ at x is the cumulant-generating function of the distribution $\mathcal{D}_x^\varphi$, which means that ψ is infinitely-differentiable and that $\nabla^k \psi$ is the k^{th} cumulant tensor of $\mathcal{D}_x^\varphi$. We will only use the first three derivatives of ψ which we compute below for completeness.

Lemma 38 (LLT derivatives)**.** *For any $x, h \in \mathbb{R}^d$, we have*

$$\begin{aligned} \nabla \psi(x) &= \mu(\mathcal{D}_x^\varphi) \stackrel{\text{def}}{=} \mathbb{E}_{y \sim \mathcal{D}_x^\varphi}[y], \\ \nabla^2 \psi(x) &= \mathrm{Cov}(\mathcal{D}_x^\varphi) \stackrel{\text{def}}{=} \mathbb{E}_{y \sim \mathcal{D}_x^\varphi}\left[(y - \mu(\mathcal{D}_x^\varphi))(y - \mu(\mathcal{D}_x^\varphi))^\top\right], \\ \nabla^3 \psi(x)[h, h, h] &= \mathbb{E}_{y \sim \mathcal{D}_x^\varphi}\left[\langle y - \mu(\mathcal{D}_x^\varphi), h\rangle^3\right]. \end{aligned}$$

Proof. For any $x \in \mathbb{R}^d$, a straightforward calculation shows that

$$\nabla \psi(x) = \nabla \left(\log \int \exp\left(\langle x, y\rangle - \varphi(y)\right) \mathrm{d}y \right) = \frac{\int \exp\left(\langle x, y\rangle - \varphi(y)\right) y \mathrm{d}y}{\int \exp\left(\langle x, y\rangle - \varphi(y)\right) \mathrm{d}y} = \mu(\mathcal{D}_x^\varphi).$$

Further,

$$\nabla^2\psi(x) = \nabla\left(\frac{\int \exp\left(\langle x,y\rangle - \varphi(y)\right) y \mathrm{d}y}{\int \exp\left(\langle x,y\rangle - \varphi(y)\right) \mathrm{d}y}\right)$$
$$= \frac{\int \exp\left(\langle x,y\rangle - \varphi(y)\right) yy^\top \mathrm{d}y}{\int \exp\left(\langle x,y\rangle - \varphi(y)\right) \mathrm{d}y} - \frac{\left(\int \exp\left(\langle x,y\rangle - \varphi(y)\right) y \mathrm{d}y\right)\left(\int \exp\left(\langle x,y\rangle - \varphi(y)\right) y \mathrm{d}y\right)^\top}{\left(\int \exp\left(\langle x,y\rangle - \varphi(y)\right) \mathrm{d}y\right)^2}.$$

Finally,

$$\nabla^3\psi(x)[h,h,h] = h^\top \nabla\left(\frac{\int \exp\left(\langle x,y\rangle - \varphi(y)\right)\left(y^\top h\right)^2 \mathrm{d}y}{\int \exp\left(\langle x,y\rangle - \varphi(y)\right) \mathrm{d}y} - \frac{\left(\int \exp\left(\langle x,y\rangle - \varphi(y)\right) y^\top h \mathrm{d}y\right)^2}{\left(\int \exp\left(\langle x,y\rangle - \varphi(y)\right) \mathrm{d}y\right)^2}\right)$$
$$= \frac{\int \exp\left(\langle x,y\rangle - \varphi(y)\right)\left(y^\top h\right)^3 \mathrm{d}y}{\int \exp\left(\langle x,y\rangle - \varphi(y)\right) \mathrm{d}y} + 2\left(\frac{\int \exp\left(\langle x,y\rangle - \varphi(y)\right) y^\top h \mathrm{d}y}{\int \exp\left(\langle x,y\rangle - \varphi(y)\right) \mathrm{d}y}\right)^3$$
$$- \frac{3\int \exp\left(\langle x,y\rangle - \varphi(y)\right)\left(y^\top h\right)^2 \mathrm{d}y \int \exp\left(\langle x,y\rangle - \varphi(y)\right) y^\top h \mathrm{d}y}{\left(\int \exp\left(\langle x,y\rangle - \varphi(y)\right) \mathrm{d}y\right)^2}.$$

□

By using a fact on one-dimensional logconcave distributions in [BE19], this implies the following.

Lemma 39 (Self-concordance)**.** *If ψ is the LLT of a convex function, it is self-concordant.*

Proof. By the definition of self-concordance and Lemma 38, it suffices to show for any $h \in \mathbb{R}^d$,

$$\mathbb{E}_{y\sim\mathcal{D}_x^\varphi}\left[\langle y - \mu(\mathcal{D}_x^\varphi), h\rangle\right]^3 \le 2\left(\mathbb{E}_{y\sim\mathcal{D}_x^\varphi}\left[\langle y - \mu(\mathcal{D}_x^\varphi), h\rangle^2\right]\right)^{\frac{3}{2}}. \tag{6.7}$$

We then note that the random variable $\langle y - \mu(\mathcal{D}_x^\varphi), h\rangle$ for $y \sim \mathcal{D}_x^\varphi$ follows a logconcave distribution because affine transformations preserve logconcavity. Finally Lemma 2 of [BE19] implies (6.7) holds. □

Next, we prove that a form of strong convexity-smoothness duality (and its converse) holds with respect to φ and ψ, analogous to the type of duality satisfied by Fenchel conjugates [KST09].

Lemma 40 (Strong convexity-smoothness duality)**.** *If $\varphi : \mathbb{R}^d \to \mathbb{R}$ is L-smooth with respect to $\|\cdot\|_*$, then $\psi : \mathbb{R}^d \to \mathbb{R}$ is $\frac{1}{L}$-strongly convex with respect to $\|\cdot\|$.*

Proof. By definition of strong convexity it suffices to prove for any $x, v \in \mathbb{R}^d$, $v^\top \nabla^2\psi(x) v \ge \frac{1}{L}\|v\|^2$. Without loss of generality, by scale invariance we can assume $\|v\| = 1$. Let $Y = \langle y, v\rangle$, where $y \sim \mathcal{D}_x^\varphi$. By Lemma 38, $\nabla^2\psi(x) = \mathrm{Cov}(\mathcal{D}_x^\varphi)$, so it suffices to prove that

$$\mathrm{Var}(Y) = \mathbb{E}_{y\sim\mathcal{D}_x^\varphi}\left[\langle y - \mu(\mathcal{D}_x^\varphi), v\rangle^2\right] \ge \frac{1}{L}.$$

Letting $\mathbf{M} \stackrel{\text{def}}{=} \mathbb{E}_{y \sim \mathcal{D}_x^\varphi} \nabla^2 \varphi(y)$, we first observe

$$\frac{L}{2} v^\top \mathbf{M}^{-1} v = \max_{u \in \mathbb{R}^d} \langle u, v \rangle - \frac{1}{2L} u^\top \mathbf{M} u \geq \max_{u \in \mathbb{R}^d} \langle u, v \rangle - \frac{1}{2} \|u\|_*^2 = \frac{1}{2} \|v\|^2.$$

In the only inequality, we used that $u^\top \mathbf{M} u = \mathbb{E}_{y \sim \mathcal{D}_x^\varphi} u^\top \nabla^2 \varphi(y) u \leq L \|u\|_*^2$ by smoothness of φ, and the last equality follows by optimizing over $\|u\|_*$. This shows $v^\top \mathbf{M}^{-1} v \geq \frac{1}{L}$. The Cramér-Rao inequality (see Lemma 2, [CP22]) then implies

$$\mathrm{Var}(Y) \geq v^\top \mathbf{M}^{-1} v \geq \frac{1}{L},$$

since the Hessian of $-\log \mathcal{D}_x^\varphi$ at any $x \in \mathbb{R}^d$ is $\nabla^2 \varphi$. □

Lemma 41 (Smoothness-strong convexity duality)**.** *If $\varphi : \mathbb{R}^d \to \mathbb{R}$ is $\frac{1}{L}$-strongly convex with respect to $\|\cdot\|_*$, then $\psi : \mathbb{R}^d \to \mathbb{R}$ is L-smooth with respect to $\|\cdot\|$.*

Proof. Let $v, x \in \mathbb{R}^d$ and assume $\|v\| = 1$. As in Lemma 40, defining $Y = \langle y, v \rangle$ for $y \sim \mathcal{D}_x^\varphi$, we have $v^\top \nabla^2 \psi(x) v = \mathrm{Var}(Y)$, and want to show $\mathrm{Var}(Y) \leq L$. First note that for any $y \in \mathbb{R}^d$,

$$\frac{1}{2L} v^\top \left(\nabla^2 \varphi(y)\right)^{-1} v = \max_{u \in \mathbb{R}^d} \langle u, v \rangle - \frac{L}{2} u^\top \nabla^2 \varphi(y) u \leq \max_{u \in \mathbb{R}^d} \langle u, v \rangle - \frac{1}{2} \|u\|_*^2 = \frac{1}{2} \|v\|^2.$$

The first inequality used strong convexity of φ and the last equality follows by optimizing over $\|u\|_*$. This shows $v^\top (\nabla^2 \varphi(y))^{-1} v \leq L$ for all y. The Brascamp-Lieb inequality [BL76] then implies

$$\mathrm{Var}(Y) \leq \mathbb{E}_{y \sim \mathcal{D}_x^\varphi} \left[v^\top \left(\nabla^2 \varphi(y)\right)^{-1} v \right] \leq L,$$

since the Hessian of $-\log \mathcal{D}_x^\varphi$ at any $x \in \mathbb{R}^d$ is $\nabla^2 \varphi$. □

6.3.2 Isoperimetry

In this section we prove an isoperimetric inequality for densities which are relatively strongly logconcave with respect to an appropriate LLT. The only LLT property we use is Lemma 39, i.e. self-concordance, via the following generic fact which generalizes Lemma 35 of [LV18].

Lemma 42. *Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is convex and self-concordant. For any $x \in \mathbb{R}$,*

$$\frac{\exp(-\phi(x))}{\sqrt{\phi''(x)}} \geq \frac{1}{12} \min \left\{ \int_{-\infty}^x \exp(-\phi(t)) dt, \int_x^\infty \exp(-\phi(t)) dt \right\}.$$

Proof. Assume $\phi'(x) \geq 0$ (the other case will follow analogously by bounding the integral on $(-\infty, x]$). Define $r \stackrel{\text{def}}{=} x + \frac{1}{4\sqrt{\phi''(x)}}$. By self-concordance (Lemma 37), for all $t \in [x, r]$,

$$\frac{1}{2}\phi''(x) \leq \phi''(t) \leq 2\phi''(x).$$

Hence, we have for all $t \in [x, r]$, since $\phi'(x) \geq 0$,

$$\phi(t) = \phi(x) + \phi'(x)(t-x) + \int_x^t (t-s)\phi''(s)\mathrm{d}s \geq \phi(x) + \frac{1}{4}(t-x)^2\phi''(x). \tag{6.8}$$

We use (6.8) to bound the integral on $[x, r]$:

$$\begin{aligned}\int_x^r \exp(-\phi(t))\mathrm{d}t &\leq \exp(-\phi(x)) \int_x^r \exp\left(-\frac{1}{4}(t-x)^2\phi''(x)\right)\mathrm{d}t \\ &\leq \exp(-\phi(x)) \int_{-\infty}^{\infty} \exp\left(-\frac{1}{4}(t-x)^2\phi''(x)\right)\mathrm{d}t = 2\sqrt{\pi}\cdot\frac{\exp(-\phi(x))}{\sqrt{\phi''(x)}}.\end{aligned} \tag{6.9}$$

Next, to bound the integral on $[r, \infty)$, we first observe

$$\phi'(r) \geq \phi'(x) + \int_x^r \phi''(r)\mathrm{d}t \geq \frac{1}{2}\int_x^r \phi''(x)\mathrm{d}t \geq \frac{1}{8}\sqrt{\phi''(x)}.$$

Hence, by convexity from r,

$$\begin{aligned}\int_r^\infty \exp(-\phi(t))\mathrm{d}t &\leq \int_r^\infty \exp\left(-\phi(r) - \phi'(r)(t-r)\right)\mathrm{d}t \\ &\leq \exp(-\phi(x)) \int_r^\infty \exp\left(-\frac{1}{8}\sqrt{\phi''(x)}(t-r)\right)\mathrm{d}t = 8\cdot\frac{\exp(-\phi(x))}{\sqrt{\phi''(x)}}.\end{aligned} \tag{6.10}$$

We used $\phi(r) \geq \phi(x)$ by convexity and $\phi'(x) \geq 0$. Combining (6.9) and (6.10) yields the claim. □

Next, we reduce the problem of proving isoperimetry for relatively strongly logconcave densities to the same problem in one dimension (captured via Lemma 42), via the localization lemma.

Lemma 43 (Modification of the localization lemma, [KLS95], Theorem 2.7)**.** *Let f_1, f_2, f_3, f_4 be four nonnegative functions on $\mathbb{R}^d$ such that f_1 and f_2 are upper semicontinuous and f_3 and f_4 are lower semicontinuous, let $c_1, c_2 > 0$, and let $\phi : \mathbb{R}^d \to \mathbb{R}$ be convex. Then, the following are equivalent:*

- *For every density $\pi : \mathbb{R}^d \to \mathbb{R}$ which is 1-relatively strongly logconcave in ϕ,*

$$\left(\int f_1(x)\pi(x)\mathrm{d}x\right)^{c_1}\left(\int f_2(x)\pi(x)\mathrm{d}x\right)^{c_2} \leq \left(\int f_3(x)\pi(x)\mathrm{d}x\right)^{c_1}\left(\int f_4(x)\pi(x)\mathrm{d}x\right)^{c_2}.$$

- *For every $a, b \in \mathbb{R}^d$ and $\gamma \in \mathbb{R}$,*

$$\left(\int_0^1 f_1((1-t)a+tb)e^{\gamma t-\phi((1-t)a+tb)}\mathrm{d}t\right)^{c_1}\left(\int_0^1 f_2((1-t)a+tb)e^{\gamma t-\phi((1-t)a+tb)}\mathrm{d}t\right)^{c_2}$$
$$\leq \left(\int_0^1 f_3((1-t)a+tb)e^{\gamma t-\phi((1-t)a+tb)}\mathrm{d}t\right)^{c_1}\left(\int_0^1 f_4((1-t)a+tb)e^{\gamma t-\phi((1-t)a+tb)}\mathrm{d}t\right)^{c_2}.$$

Proof. The proof follows identically to the case where $\phi = 0$, which was proven in [LS93, KLS95] via a bisection argument (see Lemma 2.5, [LS93]). The only fact the bisection argument relies on is that restricting logconcave densities to subsets of $\mathbb{R}^d$ preserves logconcavity, which remains true for densities which are relatively strongly logconcave with respect to a given convex function. For a more formal treatment of this generalized bisection argument, see Lemma 1 of [GLL+23b]. Finally the change on the continuity assumptions on the $\{f_i\}_{i\in[4]}$ follows by Remark 2.3 of [KLS95]. □

Finally, we combine these tools to prove the main result of this section.

Lemma 44 (Self-concordant isoperimetry)**.** *Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be convex and self-concordant, and let $f : \mathbb{R}^d \to \mathbb{R}$ be m-relatively strongly convex in ϕ. Given any partition S_1, S_2, S_3 of $\mathbb{R}^d$,*

$$\frac{\int_{S_3}\exp\left(-f(x)\right)\mathrm{d}x}{\min\left\{\int_{S_1}\exp(-f(x))\mathrm{d}x, \int_{S_2}\exp(-f(x))\mathrm{d}x\right\}} = \Omega\left(\sqrt{m}d_\phi(S_1,S_2)\right),$$

where $d_\phi(S_1, S_2) = \min_{x\in S_1, y\in S_2} d_\phi(x, y)$.

Proof. We assume $m = 1$ by rescaling $\phi \leftarrow m\phi$ which results in $d_\phi(S_1, S_2) \leftarrow \sqrt{m}d_\phi(S_1, S_2)$. We first show that without loss of generality, we can assume

$$\max_{i\in\{1,2\}} \frac{\int_{S_i}\exp(-f(x))\mathrm{d}x}{\int\exp(-f(x))\mathrm{d}x} = \Omega(1). \tag{6.11}$$

To see this, let $S_1^\star, S_2^\star$ and $S_3^\star$ be the partition that achieves the minimum of

$$\beta(S_1,S_2,S_3) = \frac{\int_{S_3}\exp\left(-f(x)\right)\mathrm{d}x}{d_\phi(S_1,S_2)\min\left\{\int_{S_1}\exp(-f(x))\mathrm{d}x, \int_{S_2}\exp(-f(x))\mathrm{d}x\right\}}.$$

Let $\delta = d_\phi(S_1^\star, S_2^\star)$. For any $z \in S_3^\star$, let $x \in S_1^\star$ minimize $d_\phi(x, z)$ and let $y \in S_2^\star$ minimize $d_\phi(y, z)$. By the triangle inequality we have

$$d_\phi(x,z) + d_\phi(y,z) \geq \delta$$

and hence $\max(d_\phi(x, z), d_\phi(y, z)) \geq \frac{\delta}{2}$. Consequently we can partition $S_3^\star$ into S_3' and S_3'' such that $d_\phi(S_1^\star, S_3') \geq \frac{\delta}{2}$ and $d_\phi(S_2^\star, S_3'') \geq \frac{\delta}{2}$ by placing each z into an appropriate set.

Moreover, we can assume without loss of generality that

$$\frac{\int_{S_3'} \exp\left(-f(x)\right) \mathrm{d}x}{\frac{\delta}{2} \min\left\{\int_{S_1^\star} \exp(-f(x))\mathrm{d}x, \int_{S_2^\star} \exp(-f(x))\mathrm{d}x\right\}} \leq \beta.$$

as otherwise the above is true for S_3''. Thus, $\beta(S_1^\star \cup S_3'', S_2^\star, S_3') \leq \beta(S_1^\star, S_2^\star, S_3^\star)$, proving (6.11) (else we may halve the measure of S_3). Given (6.11), it suffices to show that there is a constant C with

$$\begin{aligned} C d_\phi(S_1, S_2) \int \exp(-f(x)) \mathbb{1}_{S_1}(x)\mathrm{d}x \int \exp(-f(x)) \mathbb{1}_{S_2}(x)\mathrm{d}x \\ \leq \int \exp(-f(x))\mathrm{d}x \int \exp(-f(x)) \mathbb{1}_{S_3}(x)\mathrm{d}x. \end{aligned}$$

Using the localization lemma (Lemma 43), letting $f_i = \mathbb{1}_{S_i}$ for $i \in [3]$ and $f_4 = (C d_\phi(S_1, S_2))^{-1}$,[7] it suffices to prove for every $a, b \in \mathbb{R}^d$ and $\gamma \in \mathbb{R}$,

$$\begin{aligned} & C d_\phi(S_1, S_2) \int_0^1 \exp\left(\gamma t - \phi((1-t)a + tb)\right) \mathbb{1}_{S_1}((1-t)a + tb)\mathrm{d}t \\ & \cdot \int_0^1 \exp\left(\gamma t - \phi((1-t)a + tb)\right) \mathbb{1}_{S_2}((1-t)a + tb)\mathrm{d}t \\ \leq & \int_0^1 \exp\left(\gamma t - \phi((1-t)a + tb)\right) \mathrm{d}t \int_0^1 \exp\left(\gamma t - \phi((1-t)a + tb)\right) \mathbb{1}_{S_3}((1-t)a + tb)\mathrm{d}t. \end{aligned}$$

Redefine $\phi(t) \leftarrow \phi((1-t)a + tb) - \gamma t$ for $t \in \mathbb{R}$, which is a one-dimensional self-concordant function, and redefine $S_i \leftarrow \{t \mid (1-t)a + tb \in S_i\}$ for $i \in [3]$, such that each S_i is a union of intervals. It is straightforward to check that the distance $d_\phi(S_1, S_2)$ only increases under this transformation, because it can only take fewer paths, and each path has the same length (the change in $\sqrt{\phi''}$ is negated by the change in distance traveled by the path).

So, it suffices to consider the special one-dimensional case with $\gamma = 0$, where $d_\phi(x, y) = \int_x^y \sqrt{\phi''(t)}\mathrm{d}t$. We next note that it suffices to consider the case when S_3 is a single interval, i.e. for any $a \leq a' \leq b' \leq b$, we have $S_1 = [a, a']$, $S_2 = [b', b]$, $S_3 = [a', b']$, and wish to show for some constant C

$$\frac{\int_{a'}^{b'} \exp(-\phi(t))\mathrm{d}t}{\int_{a'}^{b'} \sqrt{\phi''(t)}dt} \geq C \frac{\int_a^{a'} \exp(-\phi(t))\mathrm{d}t \int_{b'}^{b} \exp(-\phi(t))\mathrm{d}t}{\int_a^b \exp(-\phi(t))\mathrm{d}t}. \tag{6.12}$$

When S_3 has multiple intervals, by Theorem 2.6 in [LS93], we show (6.12) for each interval in S_3 and its adjacent segments in S_1 and S_2, and sum over all such inequalities. By Lemma 42, when ϕ is convex and self-concordant, we have for any $x \in [a, b]$,

$$\frac{\exp(-\phi(x))}{\sqrt{\phi''(x)}} \geq \frac{1}{12} \min\left(\int_a^x \exp(-\phi(t))\mathrm{d}t, \int_x^b \exp(-\phi(t))\mathrm{d}t\right)$$

[7] Without loss of generality we can assume S_1 and S_2 are closed (implying S_3 is open) by taking their closures. This implies f_1, f_2 are upper semicontinuous and f_3, f_4 are lower semicontinuous.

which combined with $\frac{\int_{a'}^{b'} \exp(-\phi(t))\mathrm{d}t}{\int_{a'}^{b'} \sqrt{\phi''(t)}\mathrm{d}t} \geq \min_{x\in[a',b']} \frac{\exp(-\phi(x))}{\sqrt{\phi''(x)}}$ shows (6.12). □

6.3.3 Total variation bounds

In this section, we provide a bound on the total variation distance of induced distributions $\mathcal{D}_x^\varphi$ and $\mathcal{D}_{x'}^\varphi$, when x and x' are close in the Riemannian distance given by ψ.

Lemma 45 (TV distance between $\mathcal{D}_x^\varphi$ and $\mathcal{D}_{x'}^\varphi$)**.** *For any $x, x' \in \mathbb{R}^d$ such that $d_\psi(x, x') \leq \frac{1}{4}$,*

$$\left\|\mathcal{D}_x^\varphi - \mathcal{D}_{x'}^\varphi\right\|_{\mathrm{TV}} \leq \frac{1}{2}.$$

Proof. Let $h = x' - x$ and note that the KL divergence between $\mathcal{D}_x^\varphi$ and $\mathcal{D}_{x'}^\varphi$ may be rewritten as

$$\begin{aligned} D_{\mathrm{KL}}\left(\mathcal{D}_x^\varphi \| \mathcal{D}_{x'}^\varphi\right) &= \mathbb{E}_{y\sim\mathcal{D}_x^\varphi}\left[\log \frac{\mathrm{d}\mathcal{D}_x^\varphi}{\mathrm{d}\mathcal{D}_{x'}^\varphi}(y)\right] = \mathbb{E}_{y\sim\mathcal{D}_x^\varphi}\left[\psi(x') - \psi(x) - \langle h, y\rangle\right] \\ &= \psi(x') - \psi(x) - \langle h, \nabla\psi(x)\rangle\,. \end{aligned}$$

In the last equation, we used Lemma 38. We recognize that the KL divergence is the Bregman divergence (first-order Taylor approximation) in ψ, and hence letting $x_t = x + th$ for $t \in [0, 1]$ such that $x_0 = x$ and $x_1 = x'$, we continue bounding

$$\begin{aligned} D_{\mathrm{KL}}\left(\mathcal{D}_x^\varphi \| \mathcal{D}_{x'}^\varphi\right) &= \int_0^1 (1-t)\nabla^2\psi(x_t)[h,h]\mathrm{d}t \\ &\leq \int_0^1 4(1-t)\nabla^2\psi(x)[h,h]\mathrm{d}t \leq \frac{1}{2}. \end{aligned}$$

The first inequality used that when $d_\psi(x, x') \leq \frac{1}{4}$, Lemma 36 shows $\left\|x_t - x\right\|_x \leq \left\|x' - x\right\|_x \leq \frac{1}{2}$, so Lemma 37 gives $\nabla^2\psi(x_t) \preceq 4\nabla^2\psi(x)$; the second used $\left\|h\right\|_x \leq \frac{1}{2}$. Finally by Pinsker's inequality,

$$\left\|\mathcal{D}_x^\varphi - \mathcal{D}_{x'}^\varphi\right\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathcal{D}_x^\varphi \| \mathcal{D}_{x'}^\varphi)} \leq \frac{1}{2}.$$

□

6.4 Proximal LLT sampler

In this section, we study a sampling problem in the following setting, assumed throughout.

Problem 1. *For $D, G, \eta > 0$, let $\mathcal{X} \subset \mathbb{R}^d$ be compact and convex, with diameter in a norm $\|\cdot\|_{\mathcal{X}}$ at most D. Let $F : \mathcal{X} \to \mathbb{R}$ have the stochastic form $F(x) \overset{\mathrm{def}}{=} \mathbb{E}_{i\sim\mathcal{I}}\left[f_i(x)\right]$, for a distribution $\mathcal{I}$ over (a possibly infinite) family of indices i, such that each $f_i : \mathcal{X} \to \mathbb{R}$ is*

convex and G-Lipschitz in $\|\cdot\|_{\mathcal{X}}$. *Finally, let* $\varphi : \mathbb{R}^d \to \mathbb{R}$ *be convex and* η*-smooth in the dual norm* $\|\cdot\|_{\mathcal{X}^*}$. *Given* $\mu > 0$, *and letting* $\psi : \mathbb{R}^d \to \mathbb{R}$ *be the LLT of* φ, *the goal is to sample from the density* π *satisfying*

$$\mathrm{d}\pi(x) \propto \exp\left(-F(x) - \eta\mu\psi(x)\right) \mathbb{1}_{\mathcal{X}}(x)\mathrm{d}x. \tag{6.13}$$

Note that by Lemma 40, $\eta\mu\psi$ is μ-strongly convex in $\|\cdot\|_{\mathcal{X}}$. Letting $z = (x, y)$ denote a variable on $\mathcal{X} \times \mathbb{R}^d$, it is convenient for us to define the extended density on the joint space of z:

$$\mathrm{d}\hat{\pi}(z) \propto \exp\left(-F(x) - \eta\mu\psi(x) + (\langle x, y\rangle - \psi(x) - \varphi(y))\right) \mathbb{1}_{\mathcal{X}}(x)\mathrm{d}z. \tag{6.14}$$

Our sampling framework for (6.13) generalizes an approach pioneered by [LST21b], and is stated in the following Algorithm 11. The algorithm simply alternately samples from each marginal of (6.14). Before stating it, we define the following notation for conditional densities throughout the section:

$$\begin{aligned} &\mathrm{d}\pi_x(y) = \exp\left(\langle x, y\rangle - \psi(x) - \varphi(y)\right)\mathrm{d}y \text{ for all } x \in \mathcal{X}, \\ &\mathrm{d}\pi_y(x) \propto \exp\left(-F(x) - (1+\eta\mu)\psi(x) + \langle x, y\rangle\right) \mathbb{1}_{\mathcal{X}}(x)\mathrm{d}x \text{ for all } y \in \mathbb{R}^d. \end{aligned} \tag{6.15}$$

In particular, we observe that $\mathrm{d}\pi_x(y) = \mathrm{d}\hat{\pi}(\cdot \mid x)$ and $\mathrm{d}\pi_y(x) = \mathrm{d}\hat{\pi}(\cdot \mid y)$.

Algorithm 11 `AlternateSample`$(\mathcal{X}, F, \varphi, T, \mu, x_0)$

Input: $\mathcal{X}, F, \varphi$ in the setting of Problem 1, $T \in \mathbf{N}$, $\mu > 0$, $x_0 \in \mathcal{X}$.

1: **for** $k \in [T]$ **do**
2: Sample $y_k \sim \pi_{x_{k-1}}$.
3: Sample $x_k \sim \pi_{y_k}$.
4: **end for**
5: **return** x_T

Correctness of Algorithm 11 for sampling from (6.14) builds upon the following basic facts.

Lemma 46. *The total* x*-marginal of* $\hat{\pi}$ *in* (6.14) *is* π *in* (6.13). *Furthermore, the stationary distribution of Algorithm 11 is* $\hat{\pi}$, *and the induced Markov chains in Algorithm 11 restricted to either* $\{x_k\}_{0 \le k \le T}$ *(a Markov chain on* $\mathcal{X}$*) or* $\{y_k\}_{k \in [T]}$ *(a Markov chain on* $\mathbb{R}^d$*) are both reversible.*

Proof. The first conclusion is a direct calculation, and the remainder is Lemma 1 in [LST21b]. □

In Section 6.4.1 we develop a subroutine based on rejection sampling for implementing Line 3 of Algorithm 11, extending [GLL22]. We then give our complete analysis of Algorithm 11 in Section 6.4.2.

6.4.1 Sampling from the x-conditional distribution

Throughout this section, we assume the setting in Problem 1, and fix some $y \in \mathbb{R}^d$. We provide a sampler for the marginal density π_y (following notation (6.15)), and denote the component of the density independent of F by γ_y, i.e.

$$\mathrm{d}\gamma_y(x) \propto \exp\left(-\eta\mu\psi(x) - (\psi(x) - \langle x, y\rangle)\right) \mathbb{1}_{\mathcal{X}}(x)\mathrm{d}x. \tag{6.16}$$

By Lemma 40, γ_y (and hence π_y) is $\frac{1}{\eta}$-strongly logconcave in $\|\cdot\|_{\mathcal{X}}$. Our rejection sampler leverages this fact and the stochastic nature of F to build a rejection sampling scheme similarly to [GLL22]. For completeness, we state our Algorithm 12 below, and provide the details of its analysis here.

In order to analyze Algorithm 12, we first state a general result about concentration of Lipschitz functions with respect to a strongly logconcave measure, in general norms. The following is a direct adaptation of standard results on log-Sobolev inequalities contained in [Led99, BL00].

Lemma 47 ([Led99], Section 2.3 and [BL00], Proposition 3.1)**.** *Let $X \sim \pi$ for density $\pi : \mathcal{X} \to \mathbb{R}$ which is μ-strongly logconcave in $\|\cdot\|_{\mathcal{X}}$, and let $\ell : \mathcal{X} \to \mathbb{R}$ be G-Lipschitz in $\|\cdot\|_{\mathcal{X}}$. For all $t \geq 0$,*

$$\Pr_{x\sim\pi}\left[\ell(x) \geq \mathbb{E}_\pi[\ell] + t\right] \leq \exp\left(-\frac{\mu t^2}{2G^2}\right).$$

In the remainder of the section, let $\tilde{\pi}_y$ be the distribution of the output of Algorithm 12 and recall the target stationary distribution is π_y. When ρ is clear from context, we define $\bar{\rho} \stackrel{\text{def}}{=} \mathrm{med}(0, \rho, 2)$ to be the truncation of ρ to $[0, 2]$. We also denote the index set drawn on Line 7 by

$$\mathcal{J} \stackrel{\text{def}}{=} \{j_{i,b}\}_{b\in[a],i\in[b]},$$

when a is clear from context. We first provide the following characterization of $\|\pi_y - \tilde{\pi}_y\|_{\mathrm{TV}}$.

Algorithm 12 `InnerLoop`$(y, \delta, \mathcal{X}, F, \varphi, \mu)$

Input: $\delta \in (0, \frac{1}{2})$, $y \in \mathbb{R}^d$, $\mathcal{X}, F, \varphi$ in the setting of Problem 1 for $\frac{1}{\eta} \geq 10^4 G^2 \log \frac{1}{\delta}$

Output: Sample within total variation distance δ of

$$\mathrm{d}\pi_y(x) \propto \exp\left(-F(x) - \eta\mu\psi(x) - (\psi(x) - \langle x, y\rangle)\right) \mathbb{1}_{x\in\mathcal{X}}\mathrm{d}x.$$

$u \leftarrow 1, \rho \leftarrow 1$
while $u > \frac{1}{2}\rho$ **do**
  Sample $x_1, x_2 \sim \gamma_y$ defined in (6.16) independently
  $\rho \leftarrow 1$, $u \sim_{\text{unif.}} [0, 1]$
  Draw $a \in \mathbf{N}$ such that for all $b \in \mathbf{N}$, $\Pr[a \geq b] = \frac{1}{b!}$
  for $b \in [a]$ **do**
    Draw $j_{i,b} \sim \mathcal{I}$ for $i \in [b]$
    $\rho \leftarrow \rho + \prod_{i\in[b]}(f_{j_{i,b}}(x_2) - f_{j_{i,b}}(x_1))$
  end for
end while
return x_1

Lemma 48. *Define r_x to be the random variable $\mathbb{E}[\rho \mid x_1 = x]$ (where the expectation is over x_2, a, and the random indices $\mathcal{J}$, and similarly let $\bar{r}_x \stackrel{\text{def}}{=} \mathbb{E}[\bar{\rho} \mid x_1 = x]$. Then,*

$$\left\|\pi_y - \tilde{\pi}_y\right\|_{\mathrm{TV}} \leq \mathbb{E}_{x\sim\gamma_y}\left|r_x - \bar{r}_x\right|.$$

Proof. First, by definition of π_y, we have

$$\pi_y(x) = \frac{\exp(-F(x))\gamma_y(x)}{\int \exp(-F(w))\gamma_y(w)\mathrm{d}w} = \gamma_y(x) \cdot \frac{\exp(-F(x))}{\mathbb{E}_{w\sim\gamma_y}\exp(-F(w))}. \tag{6.17}$$

Moreover, by definition of the algorithm,

$$\tilde{\pi}_y(x) = \frac{\gamma_y(x)\Pr[u \leq \frac{1}{2}\rho \mid x_1 = x]}{\Pr[u \leq \frac{1}{2}\rho]} = \frac{\gamma_y(x)\mathbb{E}[\bar{\rho} \mid x_1 = x]}{\mathbb{E}[\bar{\rho}]} \tag{6.18}$$

where all probabilities and expectations are x_2, a, and $\mathcal{J}$. Furthermore, note that for fixed $b \in [a]$,

$$\mathbb{E}_{\mathcal{J}}\left[\prod_{i\in[b]}(f_{j_{i,b}}(x_2) - f_{j_{i,b}}(x_1))\right] = \left(\mathbb{E}_{j\sim\mathcal{I}}\left[f_j(x_2) - f_j(x_1)\right]\right)^b = (F(x_2) - F(x_1))^b.$$

Hence, taking expectations over a, we have for any fixed x_1, x_2,

$$\begin{aligned}\mathbb{E}\left[\rho \mid x_1, x_2\right] &= \sum_{b\geq 0} \Pr[a \geq b](F(x_2) - F(x_1))^b \\ &= \sum_{b\geq 0} \frac{1}{b!}(F(x_2) - F(x_1))^b = \exp\left(F(x_2) - F(x_1)\right).\end{aligned} \tag{6.19}$$

Next, by combining (6.17) and (6.18), we have

$$\begin{aligned}\left\|\pi - \tilde{\pi}\right\|_{\mathrm{TV}} &= \frac{1}{2}\int \left|\frac{\exp(-F(x))}{\mathbb{E}_{w\sim\gamma_y}\exp(-F(w))} - \frac{\mathbb{E}[\bar{\rho} \mid x_1 = x]}{\mathbb{E}[\bar{\rho}]}\right| \gamma_y(x)\mathrm{d}x \\ &= \frac{1}{2}\mathbb{E}_{x\sim\gamma_y}\left[\left|\frac{\exp(-F(x))}{\mathbb{E}_{w\sim\gamma_y}\exp(-F(w))} - \frac{\mathbb{E}[\bar{\rho} \mid x_1 = x]}{\mathbb{E}[\bar{\rho}]}\right|\right].\end{aligned}$$

By taking expectations over x_2 in (6.19), and recalling the definitions of $r_x, \bar{r}_x$, we obtain $r_x = \mathbb{E}[\rho \mid x_1 = x] = \exp(-F(x))\mathbb{E}_{x_2\sim\gamma_y}\exp(F(x_2))$. We thus have

$$\left\|\pi - \tilde{\pi}\right\|_{\mathrm{TV}} = \frac{1}{2}\mathbb{E}_{x\sim\gamma_y}\left[\left|\frac{r_x}{\mathbb{E}_{w\sim\gamma_y} r_w} - \frac{\bar{r}_x}{\mathbb{E}_{w\sim\gamma_y}\bar{r}_w}\right|\right].$$

Next, we lower bound $\mathbb{E}_{w\sim\gamma_y} r_w$ as follows. By taking expectations over (6.19) and using independence of x_1 and x_2, we have that for the random variable $Z = \exp(-F(x))$ where $x \sim \gamma_y$, we have

$$\mathbb{E}_{w\sim\gamma_y} r_w = (\mathbb{E}Z)\cdot\left(\mathbb{E}Z^{-1}\right) \geq 1, \tag{6.20}$$

where we used Jensen's inequality which implies the last inequality for any nonnegative random variable Z. Finally, combining the above two displays, we derive the desired bound as follows:

$$\begin{aligned}\frac{1}{2}\mathbb{E}_{x\sim\gamma_y}\left[\left|\frac{r_x}{\mathbb{E}_{w\sim\gamma_y} r_w} - \frac{\bar{r}_x}{\mathbb{E}_{w\sim\gamma_y}\bar{r}_w}\right|\right] &\leq \frac{1}{2}\mathbb{E}_{x\sim\gamma_y}\left[\left|\frac{r_x}{\mathbb{E}_{w\sim\gamma_y} r_w} - \frac{\bar{r}_x}{\mathbb{E}_{w\sim\gamma_y} r_w}\right|\right] \\ &\quad + \frac{1}{2}\mathbb{E}_{x\sim\gamma_y}\left[\left|\frac{\bar{r}_x}{\mathbb{E}_{w\sim\gamma_y} r_w} - \frac{\bar{r}_x}{\mathbb{E}_{w\sim\gamma_y}\bar{r}_w}\right|\right] \\ &\leq \frac{1}{2}\mathbb{E}_{x\sim\gamma_y}\left[|r_x - \bar{r}_x|\right] + \frac{\mathbb{E}_{x\sim\gamma_y}[|\bar{r}_x|]}{2}\cdot\left|\frac{1}{\mathbb{E}_{w\sim\gamma_y}\bar{r}_w} - \frac{1}{\mathbb{E}_{w\sim\gamma_y} r_w}\right| \\ &= \frac{1}{2}\mathbb{E}_{x\sim\gamma_y}\left[|r_x - \bar{r}_x|\right] + \frac{1}{2}\left|1 - \frac{\mathbb{E}_{x\sim\gamma_y}\bar{r}_x}{\mathbb{E}_{x\sim\gamma_y} r_x}\right| \\ &\leq \frac{1}{2}\mathbb{E}_{x\sim\gamma_y}\left[|r_x - \bar{r}_x|\right] + \frac{1}{2|\mathbb{E}_{x\sim\gamma_y} r_x|}\cdot\mathbb{E}_{x\sim\gamma_y}\left[|r_x - \bar{r}_x|\right] \\ &\leq \mathbb{E}_{x\sim\gamma_y}\left[|r_x - \bar{r}_x|\right].\end{aligned}$$

In the second and last inequalities, we use the bound (6.20). The third line follows since $\bar{r}_x$ is always nonnegative by definition, and the third inequality used convexity of $|\cdot|$. $\square$

Lemma 48 shows it remains to bound $\mathbb{E}_{x\sim\gamma_y}|r_x - \bar{r}_x|$. Fixing x_1 and x_2, we know ρ and $\bar{\rho}$ as random variables of a and $\mathcal{J}$ are equal, except for the effect of truncating ρ to $[0, 2]$. Hence,

$$\mathbb{E}_{x\sim\gamma_y}|r_x - \bar{r}_x| \le \mathbb{E}[|\rho|\mathbb{1}_{\rho\notin[0,2]}]. \tag{6.21}$$

In the remainder of the section, define

$$H \stackrel{\text{def}}{=} \left\lceil 10 \log \frac{1}{\delta} \right\rceil. \tag{6.22}$$

We then let

$$\begin{aligned} \lambda &\stackrel{\text{def}}{=} \sum_{b>H} \mathbb{1}_{a\ge b} \prod_{i\in[b]} (f_{j_{i,b}}(x_2) - f_{j_{i,b}}(x_1)), \\ \sigma &\stackrel{\text{def}}{=} \sum_{b=0}^{H} \mathbb{1}_{a\ge b} \prod_{i\in[b]} (f_{j_{i,b}}(x_2) - f_{j_{i,b}}(x_1)), \end{aligned} \tag{6.23}$$

be random variables depending on the choices of $x_1, x_2, a, \mathcal{J}$, where λ captures the effect of the "large" b, and σ captures the effect of the "small" b (where the $b = 0$ term is 1 by convention). Since $\rho = \sigma + \lambda$, in light of (6.21) it suffices to bound $\mathbb{E}[|\sigma|\mathbb{1}_{\rho\notin[0,2]}] + \mathbb{E}[|\lambda|\mathbb{1}_{\rho\notin[0,2]}]$, as

$$\mathbb{E}_{x\sim\gamma_y}|r_x - \bar{r}_x| \le \mathbb{E}[|\rho|\mathbb{1}_{\rho\notin[0,2]}] \le \mathbb{E}[|\sigma|\mathbb{1}_{\rho\notin[0,2]}] + \mathbb{E}[|\lambda|\mathbb{1}_{\rho\notin[0,2]}]. \tag{6.24}$$

We defer proofs of the following to Appendix E.3, using small modifications to [GLL22].

Lemma 49. *For λ defined in* (6.23)*,*

$$\mathbb{E}\left[|\lambda|\mathbb{1}_{\rho\notin[0,2]}\right] \le \frac{\delta}{4}.$$

Lemma 50. *For σ defined in* (6.23)*,*

$$\mathbb{E}\left[|\sigma|\mathbb{1}_{\rho\notin[0,2]}\right] \le \frac{\delta}{4}.$$

Putting together these pieces, we finally obtain the following guarantee on Algorithm 12.

Proposition 13. *The output of Algorithm 12 has total variation distance to π_y bounded by δ. In expectation, Algorithm 12 queries $O(1)$ random f_i and draws $O(1)$ samples from γ_y.*

Proof. The total variation distance bound comes from combining Lemma 48, (6.24), Lemma 49, and Lemma 50. Further, the end probability of each "while" loop is $\Pr[u \le \frac{1}{2}\rho] = \mathbb{E}[\bar{\rho}] =$

$\mathbb{E}_{x\sim\gamma}\bar{r}_x \geq \mathbb{E}_{x\sim\gamma_y} r_x - \mathbb{E}_{x\sim\gamma_y}|\bar{r}_x - r_x|$. We proved in (6.20) that $\mathbb{E}_{x\sim\gamma_y} r_x \geq 1$, and combining (6.24), Lemma 49 and Lemma 50, shows $\mathbb{E}_{x\sim\gamma_y}|\bar{r}_x - r_x| \leq \delta \leq \frac{1}{2}$. Hence the expected number of loops is ≤ 2, and each loop draws two samples from γ_y, and $O(1)$ many f_i in expectation since $\mathbb{E}a^2 = O(1)$. □

6.4.2 Analysis of Algorithm 11

We now prove a mixing time on Algorithm 11 using a standard conductance argument, by using tools developed in Section 6.3. We first define our notion of conductance.

Definition 5. *For a reversible Markov chain with stationary distribution π supported on $\mathcal{X}$ and transition distributions $\{\mathcal{T}_x\}_{x\in\mathcal{X}}$, we define the conductance of the Markov chain by*

$$\Phi := \inf_{S\subset\mathcal{X}} \frac{\int_S \mathcal{T}_x(\mathcal{X}\backslash S)\mathrm{d}\pi(x)}{\min\{\pi(S), \pi(\mathcal{X}\backslash S))\}}.$$

We further recall a standard way of lower bounding conductance via isoperimetry.

Lemma 51 ([LV18], Lemma 13)**.** *In the setting of Definition 5, let $d : \mathcal{X}\times\mathcal{X}$ be a metric on $\mathcal{X}$. Suppose for any $x, x' \in \mathcal{X}$ with $d(x, x') \leq \Delta$,*

$$\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\mathrm{TV}} \leq \frac{1}{2}.$$

Also, suppose that for any partition S_1, S_2, S_3 of $\mathbb{R}^d$, π satisfies the isoperimetric inequality

$$\pi(S_3) \geq C_{\mathrm{iso}} \left(\min_{x\in S_1, y\in S_2} d(x, y)\right) \min\left\{\pi(S_1), \pi(S_2)\right\}.$$

Then $\Phi = \Omega\left(\Delta C_{\mathrm{iso}}\right)$.

Finally, a classical result of [LS93] shows how to upper bound mixing time via conductance.

Lemma 52 ([LS93], Corollary 1.5)**.** *In the setting of Definition 5, let π_t be the distribution after t steps of the Markov chain. If the starting distribution π_0 is β-warm with respect to π*

$$\|\pi_t - \pi\|_{\mathrm{TV}} \leq \sqrt{\beta}\left(1 - \frac{\Phi^2}{2}\right)^t.$$

Leveraging Lemmas 51 and 52, we prove the following mixing time bound.

Proposition 14. *Assume the input x_0 to Algorithm 11 is drawn from a β-warm distribution with respect to π, $\eta\mu \leq 1$, and $T = \Omega(\frac{1}{\eta\mu}\log\frac{\beta}{\delta})$ for a sufficiently large constant. Then the output of Algorithm 11 has total variation distance to π bounded by δ.*

Proof. Following the optimal coupling characterization of total variation, whenever the optimal coupling of $y \sim \mathcal{D}_x^\varphi$ and $y' \sim \mathcal{D}_{x'}^\varphi$ sets $y = y'$ in Line 2 of Algorithm 11, we can couple the resulting distributions in Line 3 as well. This shows that $\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\mathrm{TV}} \leq \|\mathcal{D}_x^\varphi - \mathcal{D}_{x'}^\varphi\|_{\mathrm{TV}}$. By Lemma 39, since φ is convex, ψ is a self-concordant function. Then, combined with Lemma 45, for any $d_\psi(x, x') \leq \frac{1}{4}$,

$$\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\mathrm{TV}} \leq \|\mathcal{D}_x^\varphi - \mathcal{D}_{x'}^\varphi\|_{\mathrm{TV}} \leq \frac{1}{2}.$$

By Lemma 79, since $F + \eta\mu\psi$ is $\eta\mu$-relatively strongly convex in ψ, π satisfies the isoperimetric inequality such that for any partition S_1, S_2, S_3 of $\mathbb{R}^d$,

$$\pi(S_3) = \Omega(\sqrt{\eta\mu}) \left(\min_{x \in S_1, y \in S_2} d_\psi(x, y) \right) \min\left\{\pi(S_1), \pi(S_2)\right\}.$$

By Lemma 51, we can then lower bound the conductance by $\Phi = \Omega(\sqrt{\eta\mu})$. Choosing a sufficiently large constant in T, we conclude by Lemma 52 the desired $\|\pi_T - \pi\|_{\mathrm{TV}} \leq \sqrt{\beta}\exp(-\frac{T\Phi^2}{2}) \leq \delta$. □

By combining Proposition 13 with Proposition 14, we can now complete our analysis.

Theorem 13. *In the setting of Problem 1, let $\eta\mu \leq 1$ and assume x_0 has a β-warm distribution with respect to π defined in* (6.13)*. Further for sufficiently large constants suppose $\frac{1}{\eta} = \Omega(G^2 \log \frac{\log \beta}{\delta\eta\mu})$ and*

$$T = \Theta\left(\frac{1}{\eta\mu} \log \frac{\beta}{\delta}\right).$$

Algorithm 11 using Algorithm 12 with error parameter $\frac{\delta}{2T}$ to implement Line 3 returns a point with δ total variation distance to π, querying $O(T)$ random f_i in expectation.

Proof. Proposition 14 guarantees that if each call to Line 3 of Algorithm 11 is implemented exactly, we obtain $\frac{\delta}{2}$ total variation to π. Further, the total variation error accumulated over T calls to Algorithm 12 is less than $\frac{\delta}{2}$ by a union bound on Proposition 13. Combining these bounds results in the desired total variation guarantee, and the complexity bound follows from Proposition 13. □

We note that given sample access to $\exp(-\eta\mu\psi(x))\mathbb{1}_{x \in \mathcal{X}}$, a distribution which only depends on the choice of φ and $\mathcal{X}$ (and not the function F), we obtain $\beta \leq \exp(GD)$ in Theorem 13.

Lemma 53. *In the setting of Problem 1, the density ν satisfying*

$$\mathrm{d}\nu(x) \propto \exp(-\eta\mu\psi(x))\mathbb{1}_{\mathcal{X}}(x)\mathrm{d}x$$

is $\exp(GD)$*-warm for* π *defined in* (6.13)*.*

Proof. Note that for all $x, w \in \mathcal{X}$, $|F(x) - F(w)| \leq GD$. Further recall $\pi \propto \exp(-F)\nu$. We conclude by observing that for all $x \in \mathcal{X}$,

$$\frac{\exp\left(-F(x)\right)\nu(x)}{\int_{\mathcal{X}} \exp(-F(w))\nu(w)\mathrm{d}w} \cdot \frac{\int_{\mathcal{X}} \nu(w)\mathrm{d}w}{\nu(x)} = \frac{\int_{\mathcal{X}} \nu(w)\mathrm{d}w}{\int_{\mathcal{X}} \exp(F(x) - F(w))\nu(w)\mathrm{d}w} \leq \exp\left(GD\right).$$

□

6.5 Applications

In this section, we discuss applications of the sampling scheme we develop in Section 6.4. We begin by specializing our machinery to ℓ_p and Schatten-p norms in Section 6.5.1. We then give new algorithms with improved zeroth-order query complexity for private convex optimization in Section 6.5.2. Finally, in Section 6.5.3 we discuss computational issues regarding the specific LLT we introduce.

6.5.1 LLT for ℓ_p and Schatten-p norms

Throughout this section we fix some $p \in [1, 2]$, and define the dual value $q \geq 2$ such that $\frac{1}{q} + \frac{1}{p} = 1$. It is well-known that the ℓ_q norm and ℓ_p norm are dual, as are the corresponding Schatten norms. In light of Lemma 40, to obtain a sampler catering to the ℓ_p geometry for example, it suffices to take the LLT of a smooth function in ℓ_q. We provide the latter by recalling the following fact.

Fact 9. *Let $p \in [1, 2]$, $q \geq 2$ satisfy $\frac{1}{p} + \frac{1}{q}$. If $\|\cdot\|_q$ is a vector ℓ_q norm, $\frac{1}{2}\|\cdot\|_q^2$ is $\frac{1}{p-1}$-smooth in the ℓ_q norm, and if $\|\cdot\|_q$ is a matrix Schatten-q norm, $\frac{1}{2}\|\cdot\|_q^2$ is $\frac{1}{p-1}$-smooth in the Schatten-q norm.*

Proof. This follows (for example) from three well-known facts: 1) that $\frac{1}{2}\|\cdot\|_q^2$ and $\frac{1}{2}\|\cdot\|_p^2$ are conjugate functions in both the vector and matrix cases, 2) that the conjugate of a m-strongly convex function in a norm is $\frac{1}{m}$-smooth in the dual norm [KST09], and 3) that $\frac{1}{2}\|\cdot\|_p^2$ is $(p-1)$-strongly convex in $\|\cdot\|_p$ in both the vector and matrix cases [BCL94]. □

ℓ_p norms. Next, for any $a > 0$, when the context is clearly about vector spaces, we define

$$\psi_{p,a}(x) \stackrel{\text{def}}{=} \log\left(\int \exp\left(\langle x, y\rangle - a\|y\|_q^2\right) \mathrm{d}y\right). \tag{6.25}$$

Note that as the LLT of a $\frac{2a}{p-1}$-smooth function in ℓ_q, $\psi_{p,a}$ is $\Omega(\frac{p-1}{a})$-strongly convex in ℓ_p by Lemma 40. In applications we fix a value of $\eta > 0$, set $a = \Theta((p-1)\eta)$, and use $\eta\psi_{p,a}$ as our strongly convex regularizer in ℓ_p. We next provide a bound on the range of $\psi_{p,a}$.

Lemma 54. *Let $a > 0$ and let $d \in \mathbf{N}$ be at least a sufficiently large constant. The additive range of $\psi_{p,a}$ over $\{x \in \mathbb{R}^d \mid \|x\|_p \le 1\}$ is*

$$O\left(1 + \frac{1}{a} + \sqrt{\frac{d}{a}\log\left(a + \frac{d}{a}\right)}\right).$$

In particular, for $a \le \frac{1}{d\log d}$, the additive range is $O(\frac{1}{a})$.

Proof. Throughout the proof denote for simplicity $\psi \stackrel{\text{def}}{=} \psi_{p,a}$ and let

$$\mathcal{D}_x^{\varphi}(y) \propto \exp\left(\langle x, y\rangle - a\|y\|_q^2\right)$$

be the associated density. By the characterization of $\nabla\psi$ in Lemma 38 and the fact that the associated density $\mathcal{D}_x^{\varphi}$ is symmetric in y for $x = 0$, we have $\nabla\psi(0) = 0$ and hence it suffices to bound $\psi(x) - \psi(0)$ for $\|x\|_q \le 1$. We simplify this expression as

$$\begin{aligned}\psi(x) - \psi(0) &= \log\left(\int \exp\left(\langle x, y\rangle - a\|y\|_q^2\right) \mathrm{d}y\right) - \log\left(\int \exp\left(-a\|y\|_q^2\right) \mathrm{d}y\right) \\ &= \log\left(\int \exp\left(\langle x, y\rangle\right) \frac{\exp\left(-a\|y\|_q^2\right)}{\int \exp\left(-a\|y\|_q^2\right) \mathrm{d}y} \mathrm{d}y\right) = \log\left(\mathbb{E}_{y\sim\mathcal{D}_0^{\varphi}}\left[\exp\left(\langle x, y\rangle\right)\right]\right).\end{aligned} \tag{6.26}$$

Next, let π be the probability density on $\mathbb{R}_{\ge 0}$ such that

$$\mathrm{d}\pi(r) \propto r^{d-1}\exp\left(-ar^2\right)\mathrm{d}r.$$

We note $\mathrm{d}\pi(r)$ is the density of the scalar quantity $r = \|y\|_q$ for $y \sim \mathcal{D}_0^{\varphi}$. This can be seen by taking a derivative of the volume of the ℓ_p ball of radius r, which scales as r^d, so the surface area of the ball scales as r^{d-1}. By Hölder's inequality, $\langle x, y\rangle \le \|y\|_q$ for all y, since $\|x\|_p \le 1$. We then continue (6.26) and bound $\psi(x) - \psi(0) \le \log(\mathbb{E}_{r\sim\pi}\exp(r))$, and the conclusion follows from Lemma 55. □

Lemma 55. *For any $a > 0$ and $d \in \mathbf{N}$ at least a sufficiently large constant,*

$$\log\left(\frac{\int_0^\infty \exp\left((d-1)\log r + r - ar^2\right)\mathrm{d}r}{\int_0^\infty \exp\left((d-1)\log r - ar^2\right)\mathrm{d}r}\right) \le 8 + \frac{8}{a} + \sqrt{\frac{8d}{a}\log\left(a + \frac{d}{a}\right)}.$$

Proof. Throughout this proof let

$$Z \stackrel{\text{def}}{=} \int_0^\infty \exp\left((d-1)\log r - \alpha r^2\right) \mathrm{d}r = \frac{\Gamma(\frac{d}{2})}{2a^{\frac{d}{2}}},\ \tau \stackrel{\text{def}}{=} 7 + \frac{8}{a} + \sqrt{\frac{8d}{a}\log\left(a + \frac{d}{a}\right)}.$$

Next we split the numerator of the left-hand side into two integrals:

$$I_1 \stackrel{\text{def}}{=} \int_0^\tau \exp\left((d-1)\log r + r - ar^2\right) \mathrm{d}r,$$
$$I_2 \stackrel{\text{def}}{=} \int_\tau^\infty \exp\left((d-1)\log r + r - ar^2\right) \mathrm{d}r.$$

It is immediate that $I_1 \le \exp(\tau) Z$. Further, we recognize that for $r \ge \tau$,

$$\max\left(r, (d-1)\log r\right) \le \frac{ar^2}{4}.$$

The first piece in the maximum is clear from $\tau \ge \frac{4}{a}$. The second follows since $\frac{r^2}{\log r}$ is an increasing function for $r \ge 7$, and either $\frac{4d}{a} \le 10$ in which case we use $\frac{7^2}{\log 7} \ge 10$, or we let $C \stackrel{\text{def}}{=} \frac{4d}{a}$ and use

$$\frac{r^2}{\log r} \ge C \text{ for } r \ge \sqrt{2C\log\frac{C}{4}},\ C \ge 10.$$

Hence we may bound

$$I_2 \le \int_\tau^\infty \exp\left(-\frac{ar^2}{2}\right) = \sqrt{\frac{2\pi}{a}} \Pr_{t\sim\mathcal{N}(0,a^{-1})}[t \ge \tau] \le \frac{2}{a\tau}\exp\left(-\frac{a\tau^2}{2}\right).$$

Above, we used Mill's inequality

$$\Pr_{t\sim\mathcal{N}(0,\sigma^2)}[t \ge \tau] \le \sqrt{\frac{2}{\pi}}\frac{\sigma}{\tau}\exp\left(-\frac{\tau^2}{2\sigma^2}\right).$$

Further for our τ, our upper bound on I_1 is larger than our upper bound on I_2. To see this,

$$\tau\left(1 + \frac{a\tau}{2}\right) + \frac{d}{3}\log d \ge \frac{d}{2}\log a \implies \exp\left(\tau\left(1+\frac{a\tau}{2}\right)\right)\Gamma\left(\frac{d}{2}\right) \ge a^{\frac{d}{2}}$$
$$\implies \frac{\exp(\tau)\,\Gamma(\frac{d}{2})}{2a^{\frac{d}{2}}} \ge \frac{4}{a\tau}\exp\left(-\frac{a\tau^2}{2}\right).$$

The first inequality is because $a\tau^2 \ge d\log a$. The first implication then follows by exponentiating and using $\log\Gamma(\frac{d}{2}) \ge \frac{d}{3}\log d$ for sufficiently large d, and the second implication follows by rearranging and using $a\tau \ge 4$. Finally the conclusion follows from

$$\log\left(\frac{\int_0^\infty \exp\left((d-1)\log r + r - ar^2\right)\mathrm{d}r}{\int_0^\infty \exp\left((d-1)\log r - ar^2\right)\mathrm{d}r}\right) \le \log\left(\frac{2\exp(\tau)Z}{Z}\right) \le \tau + 1.$$

□

Schatten-p norms. When the context is clearly about matrix spaces, we analogously define

$$\psi_{p,a}(\mathbf{X}) \stackrel{\text{def}}{=} \log\left(\int \exp\left(\langle \mathbf{X}, \mathbf{Y}\rangle - a\|\mathbf{Y}\|_q^2\right) \mathrm{d}y\right).$$

The proof of Lemma 54 implies the following analogous range bound in this setting.

Corollary 15. *Let $a > 0$ and let $d_1, d_2 \in \mathbf{N}$ be at least sufficiently large constants. The additive range of $\psi_{p,a}$ over $\{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2} \mid \|\mathbf{X}\|_p \le 1\}$ is*

$$O\left(1 + \frac{1}{a} + \sqrt{\frac{d_1 d_2}{a}\log\left(a + \frac{d_1 d_2}{a}\right)}\right).$$

In particular, for $a \le \frac{1}{d_1 d_2 \log(d_1 d_2)}$, the additive range is $O(\frac{1}{a})$.

6.5.2 Zeroth-order private convex optimization

In this section, we consider a pair of closely-related problems in private convex optimization. Let $\mathcal{S}$ be a domain, and let $n \in \mathbf{N}$. We say that a mechanism (randomized algorithm) $\mathcal{M} : \mathcal{S}^n \to \Omega$ satisfies (ϵ, δ)-differential privacy (DP) if for any event $S \subseteq \Omega$ where Ω is the output space, and any two datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{S}^n$ which differ in exactly one element,

$$\Pr[\mathcal{M}(\mathcal{D}) \in S] \le \exp(\epsilon)\Pr[\mathcal{M}(\mathcal{D}') \in S] + \delta.$$

We next define the private optimization problems we study.

Problem 2 (DP-ERM and DP-SCO)**.** *Let $n \in \mathbf{N}$, $\epsilon, \delta \in (0, 1)$, $D, G \ge 0$, and let $\mathcal{X} \subset \mathbb{R}^d$ be compact and convex with diameter in a norm $\|\cdot\|_{\mathcal{X}}$ at most D. Let $\mathcal{P}$ be a distribution over a set $\mathcal{S}$ such that for any $s \in \mathcal{S}$, there is a $f(\cdot; s) : \mathcal{X} \to \mathbb{R}$ which is convex and G-Lipschitz in $\|\cdot\|_{\mathcal{X}}$. Let $\mathcal{D} \stackrel{\text{def}}{=} \{s_i\}_{i\in[n]}$ consist of n independent draws from $\mathcal{P}$, and let $f_i \stackrel{\text{def}}{=} f(\cdot; s_i)$ for all $i \in [n]$.*

In the differentially private empirical risk minimization (DP-ERM) *problem, we receive $\mathcal{D}$ and wish to design a mechanism $\mathcal{M}$ which satisfies (ϵ, δ)-DP and approximately minimizes*

$$F_{\text{erm}}(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i\in[n]} f_i(x).$$

In the differentially private stochastic convex optimization (DP-SCO) *problem, we receive $\mathcal{D}$ and wish to design a mechanism $\mathcal{M}$ which satisfies (ϵ, δ)-DP and approximately minimizes*

$$F_{\text{sco}}(x) \stackrel{\text{def}}{=} \mathbb{E}_{s\sim\mathcal{P}}\left[f(x; s)\right].$$

The following powerful general-purpose result was proven in [GLL+23b] reducing the DP-ERM and DP-SCO problems to logconcave sampling problems catered to the $\|\cdot\|_{\mathcal{X}}$ geometry. We slightly improve the parameter settings used by Theorem 4 of [GLL+23b] for DP-SCO by noting that a smaller value of k also suffices (due to the larger error bound), as observed by [GLL22].

Proposition 15 (Theorem 3, Theorem 4, [GLL+23b], Theorem 6.9, [GLL22])**.** *In the setting of Problem 2, let $k \geq 0$, and let $r : \mathcal{X} \to \mathbb{R}$ be 1-strongly convex with respect to $\|\cdot\|_{\mathcal{X}}$, with additive range at most Θ. Let ν be the density on $\mathcal{X}$ satisfying $\mathrm{d}\nu(x) \propto \exp(-k(F_{\mathrm{erm}}(x)+\mu r(x)))\mathbb{1}_{\mathcal{X}}(x)\mathrm{d}x$. Then the algorithm which returns a sample from ν for*

$$k = \frac{\sqrt{d}n\epsilon}{G\sqrt{2\Theta\log\frac{1}{2\delta}}},\ \mu = \frac{2G^2k\log\frac{1}{2\delta}}{n^2\epsilon^2},$$

satisfies (ϵ,δ)-DP, and guarantees

$$\mathbb{E}_{x\sim\nu}\left[F_{\mathrm{erm}}(x)\right] - \min_{x\in\mathcal{X}} F_{\mathrm{erm}}(x) \leq O\left(G\sqrt{\Theta}\cdot\frac{\sqrt{d\log\frac{1}{\delta}}}{n\epsilon}\right).$$

Further, the algorithm which returns a sample from ν for

$$k = \frac{1}{G\sqrt{\Theta}}\cdot\sqrt{\left(\frac{d\log\frac{1}{2\delta}}{\epsilon^2n^2}+\frac{1}{n}\right)}\cdot\min\left(\frac{\epsilon^2n^2}{\log\frac{1}{2\delta}}, nd\right),\ \mu = G^2k\cdot\max\left(\frac{\log\frac{1}{2\delta}}{n^2\epsilon^2}, \frac{1}{nd}\right)$$

satisfies (ϵ,δ)-DP, and guarantees

$$\mathbb{E}_{x\sim\nu}\left[F_{\mathrm{sco}}(x)\right] - \min_{x\in\mathcal{X}} F_{\mathrm{sco}}(x) \leq O\left(G\sqrt{\Theta}\cdot\left(\frac{\sqrt{d\log\frac{1}{\delta}}}{n\epsilon}+\frac{1}{\sqrt{n}}\right)\right).$$

Armed with Proposition 15 and the sampler in Theorem 13, we give our main results on Problem 2.

Assumption 1. *Fix $p \in [1,2]$ and $k, a, \eta, \mu > 0$. In the setting of Problem 2, assume there is an algorithm $\mathcal{A}$ which returns a point drawn from a β-warm start to the density ν satisfying*

$$\mathrm{d}\nu(x) \propto \exp\left(-k\left(F_{\mathrm{erm}}(x) + \eta\mu\psi_{p,a}(x)\right)\right)\mathbb{1}_{\mathcal{X}}(x)\mathrm{d}x.$$

Theorem 14. *Let $p \in [1,2]$, $\epsilon,\delta \in (0,1)$. In the setting of Problem 2 where $\|\cdot\|_{\mathcal{X}}$ is the ℓ_p norm on $\mathbb{R}^d$, there is an (ϵ,δ)-differentially private algorithm $\mathcal{M}_{\mathrm{erm}}$ which produces $x \in \mathcal{X}$ such that*

$$\mathbb{E}_{\mathcal{M}_{\mathrm{erm}}}\left[F_{\mathrm{erm}}(x)\right] - \min_{x\in\mathcal{X}} F_{\mathrm{erm}}(x) = O\left(\frac{GD}{\sqrt{p-1}}\cdot\frac{\sqrt{d\log\frac{1}{\delta}}}{n\epsilon}\right) \text{ for } p \in (1,2],$$

$$\mathbb{E}_{\mathcal{M}_{\text{erm}}}\left[F_{\text{erm}}(x)\right] - \min_{x\in\mathcal{X}} F_{\text{erm}}(x) = O\left(GD\sqrt{\log d}\cdot\frac{\sqrt{d\log\frac{1}{\delta}}}{n\epsilon}\right) \text{ for } p = 1.$$

Further, there is an (ϵ,δ)-differentially private algorithm $\mathcal{M}_{\text{sco}}$ which produces $x \in \mathcal{X}$ such that

$$\mathbb{E}_{\mathcal{M}_{\text{sco}}}\left[F_{\text{sco}}(x)\right] - \min_{x\in\mathcal{X}} F_{\text{sco}}(x) = O\left(\frac{GD}{\sqrt{p-1}}\cdot\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log\frac{1}{\delta}}}{n\epsilon}\right)\right) \text{ for } p \in (1,2],$$

$$\mathbb{E}_{\mathcal{M}_{\text{sco}}}\left[F_{\text{sco}}(x)\right] - \min_{x\in\mathcal{X}} F_{\text{sco}}(x) = O\left(GD\sqrt{\log d}\cdot\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log\frac{1}{\delta}}}{n\epsilon}\right)\right) \text{ for } p = 1.$$

Both $\mathcal{M}_{\text{erm}}$ and $\mathcal{M}_{\text{sco}}$ call $\mathcal{A}$ in Assumption 1, appropriately parameterized, once. $\mathcal{M}_{\text{erm}}$ uses

$$O\left(\left(1 + \frac{n^2\epsilon^2}{\log\frac{1}{\delta}}\right)\log\left(\frac{(1+n\epsilon)\log\beta}{\delta}\right)\log\frac{\beta}{\delta}\right).$$

additional value queries to some $f(\cdot; s_i)$, and $\mathcal{M}_{\text{sco}}$ uses

$$O\left(\min\left(nd,\ 1 + \frac{n^2\epsilon^2}{\log\frac{1}{\delta}}\right)\log\left(\frac{(1+n\epsilon)\log\beta}{\delta}\right)\log\frac{\beta}{\delta}\right)$$

additional value queries to some $f(\cdot; s_i)$.

Proof. First, we slightly simplify the setting of Problem 2. We may first assume that $D = 1$, i.e. $\mathcal{X}$ has diameter at most 1 in $\|\cdot\|_{\mathcal{X}}$. If the diameter is bounded by some $D \neq 1$, we can rescale the domain $\mathcal{X} \leftarrow \frac{1}{D}\mathcal{X}$, and remap to the modified functions $f(x; s) \leftarrow f(Dx; s)$ over this modified domain for all $s \in \mathcal{S}$. It is clear the Lipschitz constant rescales as $G \leftarrow GD$ as a result. Next, we assume $(n\epsilon)^2 \geq d\Theta\log\frac{1}{\delta}$ where $\Theta = \min(\frac{1}{p-1}, \log d)$. In the other case, in light of the diameter bound on $\mathcal{X}$ and the Lipschitz assumption, returning a random point in $\mathcal{X}$ attains the error bound claimed. Finally, assume $p \in (1,2]$, as otherwise we set $p \leftarrow 1 + \frac{1}{\log d}$, which only affects bounds by constant factors, since $\|\cdot\|_p$ is affected by $O(1)$ multplicatively everywhere under this change.

Under these simplifications, we choose the parameters k and μ according to Proposition 2 for each problem. Assume for now that Θ for the regularizer r we choose is bounded by a universal constant times $\frac{1}{p-1}$. Then the Lipschitz constant of kF_{erm} in either case of Proposition 2 is

$$kG = \Omega\left(\min\left(\frac{\sqrt{(p-1)d}n\epsilon}{\sqrt{\log\frac{1}{\delta}}}, d\sqrt{n}\right)\right) = \Omega(d),$$

as implied by our earlier simplification. We hence may choose $\mathcal{I}$ to be uniform over $[n]$, and

$$\eta = O\left(\frac{1}{k^2 G^2 \log \frac{(1+n\epsilon)\log\beta}{\delta}}\right)$$

for a sufficiently small constant to use Theorem 13. Under this setting we certainly have $\eta = O(\frac{1}{d^2})$, so letting $r \stackrel{\text{def}}{=} \eta\psi_{p,a}$ for $a \stackrel{\text{def}}{=} \frac{\eta(p-1)}{2}$ shows that r is η times the LLT of an η-smooth function in ℓ_q. By Lemma 40, r is indeed 1-strongly convex in ℓ_p, and Lemma 54 bounds its range by $\Theta = O(\frac{1}{p-1})$ satisfying our earlier assumption, where we use $a = O(\frac{1}{d^2})$. The runtime finally follows by applying our choices of k, μ in Proposition 15, with our choice of η, in Theorem 13, where we ensure that $\eta \cdot k\mu \le 1$ by choosing a smaller η if this is not the case (so Theorem 13 applies). Finally, to account for total variation error in our sampler, it suffices to adjust the failure probability δ by a constant and take a union bound over the privacy definition and the failure of Theorem 13. □

By combining the proof strategy of Theorem 14 with Corollary 15 instead of Lemma 54, we immediately obtain the following corollary in the case of Schatten norms.

Corollary 16. *Let $p \in [1,2]$, $\epsilon, \delta \in (0,1)$. In the setting of Problem 2 where $\|\cdot\|_{\mathcal{X}}$ is the Schatten-p norm on $\mathbb{R}^{d_1 \times d_2}$, there is an (ϵ,δ)-differentially private algorithm $\mathcal{M}_{\text{erm}}$ which produces $\mathbf{X} \in \mathcal{X}$ such that*

$$\mathbb{E}_{\mathcal{M}_{\text{erm}}}\left[F_{\text{erm}}(\mathbf{X})\right] - \min_{\mathbf{X}\in\mathcal{X}} F_{\text{erm}}(\mathbf{X}) = O\left(\frac{GD}{\sqrt{p-1}} \cdot \frac{\sqrt{d_1 d_2 \log\frac{1}{\delta}}}{n\epsilon}\right) \text{ for } p \in (1,2],$$

$$\mathbb{E}_{\mathcal{M}_{\text{erm}}}\left[F_{\text{erm}}(\mathbf{X})\right] - \min_{\mathbf{X}\in\mathcal{X}} F_{\text{erm}}(\mathbf{X}) = O\left(GD\sqrt{\log(d_1 d_2)} \cdot \frac{\sqrt{d_1 d_2 \log\frac{1}{\delta}}}{n\epsilon}\right) \text{ for } p = 1.$$

Further, there is an (ϵ,δ)-differentially private algorithm $\mathcal{M}_{\text{sco}}$ which produces $\mathbf{X} \in \mathcal{X}$ such that

$$\mathbb{E}_{\mathcal{M}_{\text{sco}}}\left[F_{\text{sco}}(\mathbf{X})\right] - \min_{\mathbf{X}\in\mathcal{X}} F_{\text{sco}}(\mathbf{X}) = O\left(\frac{GD}{\sqrt{p-1}} \cdot \left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d_1 d_2 \log\frac{1}{\delta}}}{n\epsilon}\right)\right) \text{ for } p \in (1,2],$$

$$\mathbb{E}_{\mathcal{M}_{\text{sco}}}\left[F_{\text{sco}}(\mathbf{X})\right] - \min_{\mathbf{X}\in\mathcal{X}} F_{\text{sco}}(\mathbf{X}) = O\left(GD\sqrt{\log(d_1 d_2)} \cdot \left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d_1 d_2 \log\frac{1}{\delta}}}{n\epsilon}\right)\right) \text{ for } p = 1.$$

Both $\mathcal{M}_{\text{erm}}$ and $\mathcal{M}_{\text{sco}}$ call $\mathcal{A}$ in Assumption 1, appropriately parameterized, once. $\mathcal{M}_{\text{erm}}$ uses

$$O\left(\left(1 + \frac{n^2\epsilon^2}{\log\frac{1}{\delta}}\right)\log\left(\frac{(1+n\epsilon)\log\beta}{\delta}\right)\log\frac{\beta}{\delta}\right).$$

additional value queries to some $f(\cdot; s_i)$, and $\mathcal{M}_{\mathrm{sco}}$ uses

$$O\left(\min\left(nd_1d_2,\ 1+\frac{n^2\epsilon^2}{\log\frac{1}{\delta}}\right)\log\left(\frac{(1+n\epsilon)\log\beta}{\delta}\right)\log\frac{\beta}{\delta}\right)$$

additional value queries to some $f(\cdot; s_i)$.

6.5.3 Oracle access for $\psi_{p,a}$

In Theorem 14 and Corollary 16, we only bounded the value oracle complexity of our sampling algorithms. The remainder of the steps in Algorithm 11 and its subroutine Algorithm 12 require samples from densities of the form $\mathrm{d}\pi_x$ (for some $x \in \mathcal{X}$) or $\mathrm{d}\gamma_y$ (for some $y \in \mathbb{R}^d$), defined in (6.15) and (6.16) respectively and reproduced here for convenience:

$$\begin{aligned}\mathrm{d}\pi_x(y) &= \exp\left(\langle x, y\rangle - \psi(x) - \varphi(y)\right)\mathrm{d}y,\\ \mathrm{d}\gamma_y(x) &\propto \exp\left(-\eta\mu\psi(x) - (\psi(x) - \langle x, y\rangle)\right)\mathbb{1}_{\mathcal{X}}(x)\mathrm{d}x.\end{aligned} \tag{6.27}$$

These densities are independent of the function F in Problem 1 and hence do not require additional value oracle queries in the setting of Problem 1. In general, the complexity of these steps depends on the complexity of the functions φ and ψ, and the set $\mathcal{X}$. We now discuss strategies for sampling from π_x and γ_y in specific settings described by Section 6.5.1, which we first briefly summarize.

1. We describe a method based on the inverse Laplace transform for sampling from π_x and evaluating $\psi_{p,a}$ with complexity linear in the dimension d in the vector setting.

2. Under efficient value oracle access to $\psi_{p,a}$ and membership oracle access to $\mathcal{X}$, general-purpose results [LV07, JLLV20, JLV22] imply polynomial-time samplers for γ_y.

3. We discuss generalizations of these methods to the matrix setting, and naïve sampling methods. We draw a loose connection to the HCIZ integral from harmonic analysis, and suggest how it may potentially help in the structured sampling task for LLTs in Schatten norms.

ℓ_p setting. We first discuss the case when $\mathcal{X} \subset \mathbb{R}^d$ is a set on vectors equipped with the ℓ_p norm for some $p \in [1, 2]$, and we let $q \geq 2$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. We follow the notation (6.25).

In order to sample from the density π_x, we use an *inverse Laplace transform* decomposition. For a parameter $c \in [0, 1)$, we define the density μ_c supported on $\mathbb{R}_{\geq 0}$, such that for all $t \geq 0$,

$$\exp(-t^c) = \int_0^\infty \exp(-\lambda t)\, \mu_c(\lambda) \mathrm{d}\lambda. \tag{6.28}$$

Intuitively, the density $\mu_c(\lambda)$ and the corresponding decomposition (inverse Laplace transform) (6.28) aims to express the more heavy-tailed function $\exp(-t^c)$ as a distribution over the lighter-tailed functions $\exp(-\lambda t)$. The inverse Laplace transform densities μ_c are well-studied in the probability theory literature, and correspond to *stable count distributions* parameterized by c. For example, it is well-known that $\mu_{\frac{1}{2}}$ is the Lévy distribution

$$\mathrm{d}\mu_{\frac{1}{2}}(\lambda) = \frac{1}{2\sqrt{\pi}\lambda^{\frac{3}{2}}} \exp\left(-\frac{1}{4\lambda}\right) \mathrm{d}\lambda.$$

We refer the reader to references e.g. [Mai07] on properties of the densities μ_c, and for now assume we can access and sample from these one-dimensional distributions in closed form for simplicity. Given this decomposition, we can then write

$$\begin{aligned} \exp(\psi_{p,a}(x)) &= \int \exp\left(\langle x, y\rangle - a\|y\|_q^2\right) \mathrm{d}y \\ &= \int_0^\infty \left(\int \exp\left(\langle x, y\rangle - \lambda a^{\frac{q}{2}} \|y\|_q^q\right) \mathrm{d}y\right) \mu_{\frac{2}{q}}(\lambda) \mathrm{d}\lambda \\ &= \int_0^\infty \prod_{i \in [d]} \left(\int_{-\infty}^\infty \exp\left(x_i y_i - \lambda a^{\frac{q}{2}} y_i^q\right) \mathrm{d}y_i\right) \mu_{\frac{2}{q}}(\lambda) \mathrm{d}\lambda. \end{aligned} \tag{6.29}$$

The decomposition (6.29) reduces the problem of sampling from π_x to d one-dimensional problems. To sample $\propto \exp(\langle x, y\rangle - a\|y\|_q^2)$, we can first sample λ from the density μ_c for $c = \frac{2}{q}$, and then sample each coordinate y_i proportionally to $\exp(x_i y_i - \lambda a^{\frac{q}{2}} y_i^q)$ conditioned on the sampled λ.

This decomposition also gives us an efficient value oracle for $\psi_{p,a}$, by evaluating (6.29) as a one-dimensional integral over λ, where the integrand may be evaluated as a product of d one-dimensional integrals. Under membership oracle access to $\mathcal{X}$, the problem of sampling from γ_y then falls under a generic logconcave sampling setup studied in a long line of work building upon [DFK91b]. The state-of-the-art general-purpose logconcave sampler, which combines the algorithms of [LV07, JLLV20] with the isoperimetric bound in [JLV22] (improving recent breakthroughs by [Che21a, KL22]), requires roughly d^3 value oracle calls to $\psi_{p,a}$ and membership oracle calls to $\mathcal{X}$.

In principle, for structured sets $\mathcal{X}$ (such as ℓ_p balls), the particular explicit structure of $\psi_{p,a}$ and $\mathcal{X}$ may be exploited to design more efficient samplers for the densities γ_y,

analogously to our custom linear-time sampler for π_x. However, it should be noted that the sampling problem for γ_y appears to be quite a bit more challenging than the problem for π_x. We leave the investigation of explicit sampler design for γ_y as an interesting open problem for future work.

Schatten-p setting. The situation is somewhat less straightforward in the matrix case. Here, the key computational problem in replicating the strategy suggested by (6.29) is evaluating the integral

$$\int \exp\left(\langle \mathbf{X}, \mathbf{Y}\rangle - C\|\mathbf{Y}\|_q^q\right) \mathrm{d}\mathbf{Y}, \tag{6.30}$$

where the integral is over $\mathbf{Y} \in \mathbb{R}^{d_1 \times d_2}$, and $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, $C > 0$ are fixed. The difficulty is $\langle \mathbf{X}, \mathbf{Y}\rangle$ decomposes coordinatewise, whereas $\|\mathbf{Y}\|_q^q$ decomposes spectrally.[8] At least superficially, this is similar to the challenge faced when evaluating the Harish-Chandra-Itzykson-Zuber (HCIZ) formula

$$\int \exp\left(\mathrm{Tr}\left(\mathbf{AUBU}^\dagger\right)\right) \mathrm{d}\mathbf{U}, \tag{6.31}$$

where the integral is over the Haar measure on (complex) unitary matrices $\mathbf{U}$, and $\mathbf{A}$, $\mathbf{B}$ are Hermitian. By dropping the $-C\|\mathbf{Y}\|_q^q$ term in (6.30) and only integrating over unitary conjugations of a fixed matrix $\mathbf{Y}$, we arrive at a generalization of (6.31). The difficulty in evaluating (6.31) is also a sort of tension between the eigenspaces of $\mathbf{A}$ and $\mathbf{B}$. Nonetheless, (6.31) has a (polynomial-time computable) exact formula, which was famously discovered independently by [HC57, IZ80]. Furthermore, [LMV21] recently obtained a polynomial-time sampler for the density induced by (6.31); while a sampler for (6.30) would follow from logconcavity and general-purpose results, it would be far from cheap, so ways of exploiting structure are fruitful to explore.

As a proof-of-concept, evaluating the integral (6.30) in (polynomial-time computable) closed form is a minimal requirement for implementing the $\mathbf{X}$-oracles in (6.27) used by our algorithm. Even this problem appears challenging, but (as summarized cleanly by [Tao13, McS21]) a plethora of techniques exist for proving the HCIZ formula, some based on tools from stochastic processes. We pose the efficient computability of the integral (6.30) as another explicit open question.

---

[8] Note that because $\|\cdot\|_q$ is unitarially invariant, we may assume $\mathbf{X}$ is diagonal.

6.6 Conclusion

We believe our work is a significant step towards developing the theory of LLTs and paving the way for their use in designing sampling algorithms. There are a number of important questions left open by our work, which we find interesting and potentially fruitful for the community to explore.

Stronger mixing time bounds. Perhaps the most immediate open question regarding our alternating sampling framework in Section 6.4 is to obtain a better understanding of its mixing time. As discussed in Section 6.1.1, Theorem 13's mixing time scales linearly in $\log \beta$, which as demonstrated by Lemma 53 (and related other settings, e.g. MALA [CLA$^+$21, LST20]) can result in additional polynomial overhead in problem parameters: for what φ, ψ is this avoidable? Notably, it is avoided for the Euclidean proximal sampler [LST21b] by working directly with KL divergence (as opposed to the larger χ^2 distance typically used by proofs using conductance bounds). Different proofs of this $\log \log \beta$ dependency for the Euclidean proximal sampler were then subsequently obtained by [CCSW22, CE22]. We also mention that $\log \log \beta$ dependences may sometimes follow via average conductance techniques (e.g. [LK99]), which may apply to our Markov chain.

Samplers for explicit distributions. Our results Theorem 13 and 14 mainly focused on bounding the query complexity to the function F, or samples f_i from the distribution defining it. The total computational complexity of a practical implementation of Algorithm 11 also includes the cost of sampling from the distributions (6.27), which are "data-independent" for this problem (only depending on explicit functions and sets instead of F). In Section 6.5.3, we give a linear-time sampler for π_x and a polynomial-time sampler for γ_y under the ℓ_p geometry, but it is interesting to obtain faster samplers for particular structured choices of $(\varphi, \mathcal{X})$ of importance in applications.

LLT beyond proximal sampling. More generally, we believe it is worthwhile to obtain a better understanding of specific choices of (φ, ψ), e.g. the examples in Section 6.5.1, from an algorithmic perspective. LLTs satisfy appealing properties such as self-concordance, strong convexity, and isoperimetry making them well-suited for frameworks beyond Algorithm 11, such as discretized MLD [AC21] and Metropolized sampling methods discussed in Section 6.1. Bounding the complexity of their use in these applications necessitates an

improved understanding of specific LLTs.

LLT as a dual object. Finally, a tantalizing open question in the theory of well-conditioned sampling (even in the ℓ_2 setting) is whether acceleration is achievable, i.e. mixing times scaling with the square root of the condition number (which is famously possible in optimization [Nes83]). The duality of Fenchel conjugates appears to play a key role in acceleration, as made explicit by [WA18, CST21], so a better understanding of duality may be helpful in the corresponding endeavor for sampling. The LLT is a natural candidate for a dual object in sampling, as it arises via joint densities on an extended space (6.2), and satisfies properties such as strong convexity-smoothness duality. Can we demystify this relationship, and use it to obtain faster samplers?

Part IV

SAMPLING IN A CONSTRAINED SPACE

Chapter 7

SAMPLING USING RIEMANNIAN HAMILTONIAN MONTE CARLO WITH CONDITION-NUMBER-INDEPENDENT CONVERGENCE RATE

This chapter is based on [KLSV22a], with Yunbum Kook, Yin Tat Lee, and Santosh S. Vempala.

7.1 Introduction

In this chapter, we study the problem of sampling from a constrained space. The need for efficient high-dimensional constrained sampling arises in many fields. A notable setting is *metabolic networks* in systems biology. A constraint-based model of a metabolic network consists of m metabolites and n reactions, and a set of equalities and inequalities that define a set of feasible steady state reaction rates (fluxes):

$$\Omega = \left\{ v \in \mathbb{R}^n \,|\, Sv = 0,\, l \le v \le u,\, c^T v = \alpha \right\},$$

where S is a stoichiometric matrix with coefficients for each metabolite and reaction. The linear equalities ensure that the fluxes into and out of every node are balanced. The inequalities arise from thermodynamical and environmental constraints. Sampling constraint-based models is a powerful tool for evaluating the metabolic capabilities of biochemical networks [LNP12, TSF$^+$13]. While the most common distribution used is uniform over the feasible region, researchers have also argued for sampling from the Gaussian density restricted to the feasible region; the latter has the advantage that the feasible set does not have to be bounded. A previous approach to sampling, using hit-and-run with rounding [HCT$^+$17], has been incorporated into the COBRA package [HAP$^+$19] for metabolic systems analysis (Bioinformatics).

A second example of mathematical interest is the problem of computing the volume of the Birkhoff polytope. For a given dimension n, the Birkhoff polytope is the set of all doubly stochastic $n \times n$ matrices (or the convex hull of all permutation matrices). This object plays a prominent role in algebraic geometry, probability, and other fields. Computing its volume has been pursued using algebraic representations; however exact

computations become intractable even for $n = 11$, requiring years of computation time. Hit-and-run has been used to show that sampling-based volume computation can go to higher dimension [CV16], with small error of estimation. However, with existing sampling implementations, going beyond $n = 20$ seems prohibitively expensive.

A third example is from machine learning, a field that is increasingly turning to *sampling* models of data according to their performance in some objective. One such commonly used criterion is the logistic regression function. The popularity of logistic regression has led to sampling being incorporated into widely used packages such as STAN [Sta20], PyMC3 [SWF16], and Pyro [BCJ+19]. However, those packages in general do not run on the constraint-based models we are interested in.

Problem Description. We consider the problem of sampling from distributions whose densities are of the form

$$e^{-f(x)} \text{ subject to } Ax = b, x \in K \tag{7.1}$$

where f is a convex function and K is a convex body. We assume that a self-concordant barrier ϕ for K is given. Note that any convex body has a self-concordant barrier [LY21] and there are explicit barriers for convex bodies that come up in practical applications [NN94], so this is a mild assumption. We introduce an efficient algorithm for the problem when K is a product of convex bodies K_i, each with small dimension. Many practical instances can be written in this form. As a special case, the algorithm can handle K in the form of $\{x \in \mathbb{R}^n : l_i \leq x_i \leq u_i \text{ for all } i \in [n]\}$ with $l_i \in \mathbb{R} \cup \{-\infty\}$ and $u_i \in \mathbb{R} \cup \{+\infty\}$, which is the common model structure in systems biology. Moreover, any generalized linear model $\exp(-\sum f_i(a_i^\top x - b_i))$, e.g., the logistic model, can be rewritten in the form

$$\exp(-\sum t_i) \text{ subject to } Ax = b + s, (s, t) \in K \tag{7.2}$$

where $K = \Pi K_i$ and each $K_i = \{(s_i, t_i) : f_i(s_i) \leq t_i\}$ is a two-dimensional convex body.

The Challenges of Practical Sampling. High dimensional sampling has been widely studied in both the theoretical computer science and the statistics communities. Many popular samplers are first-order methods, such as MALA [RT96a], basic HMC [Nea11, DKPR87] and NUTS [HG+14b], which update the Markov chain based on the gradient information of f. The runtime of such methods can depend on the condition number of

the function f [DCWY19, LST20, CDWY20, CCBJ17, SL19]. However, the condition number of real-world applications can be very large. For example, RECON1 [KLD$^+$16], a reconstruction of the human metabolic network, can have condition number as large as 10^6 due to the dramatically different orders of different chemicals' concentrations. Motivated by sampling from ill-conditioned distributions, another class of samplers use higher-order information such as Hessian of f to take into account the local structure of the problems [SBCR16, CLGL$^+$20]. However, such samplers cannot handle non-smooth distributions, such as hinge-loss, lasso, or uniform densities over polytopes.

For non-smooth distributions, the best polytime methods are based on discretizations of Brownian motion, e.g., the Ball walk [KLS97] (and its affine-invariant cousin, the Dikin walk [KN12]), which takes a random step in a ball of a fixed size around the current point. Hit-and-Run [LV06a] builds on these by avoiding an explicit step size and going to a random point along a random line through the current point. Both approaches hit the same bottleneck — in a polytope that contains a unit ball, the step size should be $O(1/\sqrt{n})$ to avoid stepping out of the body with large probability. This leads to quadratic bounds (in dimension) on the number of steps to "mix".

Due to the reduction mentioned in (7.2), non-smooth distributions can be translated to the form in (7.1) with constraint K. Both the first and higher-order sampler and the polytime non-smooth samplers have their limitations in handling distributions with non-smooth objective function or constraint K. Given the limitations of all previous samplers, a natural question we want to ask is the following.

Question. *Can we develop a practically efficient sampler that can handle the constrained problem in* (7.1) *and preserve sparsity*[1] *with mixing time independent of the condition number?*

In some applications, smoothness and condition number can be controlled with tailor-made models. Our goal here is to propose a general solver that can sample from any non-smooth distributions as given. For traditional samplers such as the Ball walk and Hit-and-Run, as mentioned earlier, the step size needs to be small so that the process does not step out. An approach that gets around this bottleneck is Hamiltonian Monte Carlo (HMC), where the next step is given by a point along a Hamiltonian-preserving curve according to a suitably chosen Hamiltonian. It has two advantages. First, the steps are no

[1]When A is sparse, preserving the sparsity of A can greatly enhance both the runtime and the space efficiency.

longer straight lines in Euclidean space, and we no longer have the concern of "stepping out". Second, the process is *symplectic* (so measure-preserving), and hence the filtering step is easy to compute. It was shown in [LV18] that significantly longer steps can be taken and the process with a convergence analysis in the setting of Hessian manifolds, leading to subquadratic convergence for uniformly sampling polytopes.

To make this practical, however, is a formidable challenge. There are two high-level difficulties. One is that many real-world instances are *highly skewed* (far from isotropic) and hence it is important to use the local geometry of the density function. This means efficiently computing or maintaining second-order information such as a Hessian of the logarithm of the density. This can be done in the Riemannian HMC (RHMC) framework [GC11, LV18], but the computation of the next step requires solving the Hamiltonian ODE to high accuracy, which in turn needs the computation of leverage scores, a procedure that takes at least matrix-multiplication time in the worst case. Another important difficulty is maintaining hard linear constraints. Existing high-dimensional packages do not allow for constraints (they must be somehow incorporated into the target density), and RHMC is usually considered with a full-dimensional feasible region such as a full-dimensional polytope. This can also be done in the presence of linear equalities by working in the affine subspace defined by the equalities, but this has the effect of *losing any sparsity* inherent in the problem and turning all coefficient matrices and objective coefficients into dense objects, thereby potentially incurring a quadratic blow-up.

Our Solution: Constrained Riemannian Hamiltonian Monte Carlo (CRHMC). We develop a constrained version of RHMC, maintaining both *sparsity* and *constraints*. Our refinement of RHMC ensures that the process satisfies the given constraints throughout, without incurring a significant overhead in time or sparsity. It works even if the resulting feasible region is poorly conditioned. Since many instances in practice are ill-conditioned and have degeneracies, we believe this is a crucial aspect. Our algorithm outperforms existing packages by orders of magnitude.

In Section 7.2, we give the main ingredients of the algorithm and discuss how we overcome the challenges that prevent us from sampling efficiently in practice. Following that, in Section 7.3, we present empirical results on several benchmark datasets, showing that CRHMC successfully samples much larger models than previously known to be possible, and is significantly faster in terms of rate of convergence ("number of steps") and total

sampling time. Our complete package is available on GitHub. We refer the reader to Appendix for theory, notations, and definitions.

7.2 Constrained RHMC

In this section, we propose a constrained Riemannian Hamiltonian Monte Carlo (CRHMC[2]) algorithm to sample from a distributions of the form

$$e^{-f(x)} \text{ subject to } c(x) = 0 \text{ and } x \in K \text{ for some convex body } K,$$

where the constraint function $c : \mathbb{R}^n \to \mathbb{R}^m$ satisfies the property that the Jacobian $Dc(x)$ has full rank for all x such that $c(x) = 0$. It is useful to keep in mind the case when $c(x) = 0$ is an affine subspace $Ax = b$, in which case $Dc(x) = A$, and the full-rank condition simply says that the rows of A are independent.

We refer readers to [And83, BSU12, Rei93] for preliminary versions of CRHMC called the constrained Hamiltonian Monte Carlo (CHMC). In particular, a framework in [BSU12] can be extended to CRHMC when $K = \mathbb{R}^n$, and in fact they mention CRHMC as a possible variant. However, their algorithm for CRHMC requires eigenvalue decomposition and is not efficient for large problems, which takes n^3 time and n^2 space per MCMC step in practice. In this section, we propose an algorithm that overcomes those limitations and satisfies the additional constraint K by using a local metric induced by the Hessian of self-concordant barriers, leading to $n^{1.5}$ time and n space in practice.

7.2.1 Basics of CRHMC

To introduce our algorithm, we first recall the RHMC algorithm (Algorithm 13). In RHMC, we extend the space x to the pair (x, v), where v denotes the *velocity*. Instead of sampling from $e^{-f(x)}$, RHMC samples from the distribution $e^{-H(x,v)}$, where $H(x, v)$ is the Hamiltonian, and then outputs x. To make sure the distribution is correct, we choose the Hamiltonian such that the marginal of $e^{-H(x,v)}$ along v is proportional to $e^{-f(x)}$. One common choice of $H(x, v)$ is

$$H(x, v) = f(x) + \frac{1}{2} v^\top M(x)^{-1} v + \frac{1}{2} \log \det M(x), \tag{7.3}$$

where $M(x)$ is a position-dependent positive definite matrix defined on $\mathbb{R}^n$.

[2]pronounced "crumch".

Algorithm 13 `Riemannian Hamiltonian Monte Carlo` (RHMC)

Input: Initial point $x^{(0)}$, step size h

1: **for** $k = 1, 2, \cdots$ **do**

Step 1: resample v

2: Sample $v^{(k-\frac{1}{2})} \sim \mathcal{N}(0, M(x^{(k-1)}))$ and set $x^{(k-\frac{1}{2})} \leftarrow x^{(k-1)}$.

Step 2: Hamiltonian dynamics

3: Solve the ODE

$$\frac{dx}{dt} = \frac{\partial H(x,v)}{\partial v}, \ \frac{dv}{dt} = -\frac{\partial H(x,v)}{\partial x} \tag{7.4}$$

with H defined in (7.3) and the initial point given by $(x^{(k-\frac{1}{2})}, v^{(k-\frac{1}{2})})$.

4: Set $x^{(k)} \leftarrow x(h)$ and $v^{(k)} \leftarrow v(h)$.

5: **end for**

Output: $x^{(k)}$

To extend RHMC to the constrained case, we need to make sure both Step 1 and Step 2 satisfy the constraints, so the Hamiltonian dynamics has to maintain $c(x) = 0$ throughout Step 2. Note that

$$\frac{d}{dt}c(x_t) = Dc(x_t) \cdot \frac{dx_t}{dt} = Dc(x_t) \cdot \frac{\partial H(x_t, v_t)}{\partial v_t}, \tag{7.5}$$

where $Dc(x)$ is the Jacobian of c at x. With H defined in (7.3), Condition (7.5) becomes $Dc(x)M(x)^{-1}v = 0$. However, for full rank $Dc(x)$, if $M(x)$ is invertible, then $\text{Range}(v) = \text{Range}(\mathcal{N}(0, M(x))) = \mathbb{R}^n$ immediately violates this condition due to $\dim(\text{Null}(Dc(x)M^{-1}(x))) = n - m$. To get around this issue, we use a non-invertible matrix $M(x)$ with its pseudo-inverse $M(x)^\dagger$ to satisfy $Dc(x)M(x)^\dagger v = 0$ for any $v \in \text{Range}(M(x))$. Since we want the step to be able to move in all directions satisfying $c(x) = 0$, we impose the following condition with $\text{Range}(M(x)) = \text{Range}(M(x)^\dagger)$ in mind:

$$\text{Range}(M(x)) = \text{Null}(Dc(x)) \text{ for all } x \in \mathbb{R}^n, \tag{7.6}$$

which can be achieved by $M(x)$ proposed soon.

Under the condition (7.6), we sample v from $\mathcal{N}(0, M(x))$ in Step 1, which is equivalent to sampling from $e^{-H(x,v)}$ subject to $v \in \text{Range}(M(x)) = \text{Null}(Dc(x))$. Also, the stationary distribution of CRHMC should be proportional to

$$e^{-H(x,v)} \text{ subject to } c(x) = 0 \text{ and } v \in \text{Null}(Dc(x)).$$

Here, to maintain $v \in \text{Null}(Dc(x))$ during Step 2 we add a Lagrangian term to H. Without the Lagrangian term, v_t would escape from $\text{Null}(Dc(x_t)) = \text{Range}(M(x_t))$ in Step 2 as seen in the proof of Lemma 15, which contradicts $\text{Range}(v_t) = \text{Range}(\mathcal{N}(0, M(x_t))) = \text{Range}(M(x_t))$. The constrained Hamiltonian we propose is (See its rigorous derivation in Lemma 15)

$$H(x,v) = \overline{H}(x,v) + \lambda(x,v)^\top c(x) \quad \text{with} \quad \overline{H}(x,v) = f(x) + \frac{1}{2} v^\top M(x)^\dagger v + \log \text{pdet}(M(x)) \tag{7.7}$$

where $\lambda(x,v) = (Dc(x)Dc(x)^\top)^{-1}\left(D^2 c(x)[v, \frac{dx}{dt}] - Dc(x)\frac{\partial \overline{H}(x,v)}{\partial x}\right)$. Here, pdet denotes pseudo-determinant and $\lambda(x,v)$ is picked so that $v \in \text{Null}(Dc(x))$. An algorithmic description of CRHMC is the same as Algorithm 13 with the constrained H in place of the unconstrained $\overline{H}$. We show the convergence of CRHMC to the correct distribution $\exp(-f(x))$ in Appendix F.2.3.

Choice of M via Self-concordant Barriers. The construction of the Hamiltonian (7.7) relies on having a family of positive semi-definite matrix $M(x)$ satisfying the condition (7.6) (i.e., $\text{Range}(M(x)) = \text{Null}(Dc(x))$). One natural choice is the orthogonal projection to $\text{Null}(Dc(x))$:

$$Q(x) = I - Dc(x)^\top (Dc(x)Dc(x)^\top)^{-1} Dc(x), \tag{7.8}$$

which is similar to the choice in [BSU12].

For the problem we care about, there are additional constraints on x other than $\{c(x) = 0\}$. In the standard HMC algorithm, we have $\frac{dx}{dt} \sim \mathcal{N}(0, M(x)^{-1})$. For example, for a simple constraint $K = [0,1]$, to ensure every direction is moving towards/away from $x = 0$ multiplicatively, a natural choice of M is $M(x) = \text{diag}(x^{-2})$. For general convex body K, we can use a *self-concordant barrier*, a function defined on K such that $\phi(x)$ is self-concordant and $\phi(x) \to +\infty$ as $x \to \partial K$. Using the barrier ϕ, we can define the local metric based on $g(x) = \nabla^2 \phi(x)$. Intuitively, as the sampler approaches ∂K, the local metric stretches accordingly so that the Hamiltonian dynamics never passes the barrier, respecting $x \in K$ throughout.

In summary, we need $M(x)$ to have its range match the null space of $Dc(x)$ and agree with $g(x)$ in its range. We can verify that $M(x) = Q(x)^\top g(x) Q(x)$, where $Q(x)$ is the symmetric matrix defined in (7.8), satisfies these two constraints.

7.2.2 Efficient Computation of $\partial H/\partial x$ and $\partial H/\partial v$

With $M(x) = Q(x)^\top g(x) Q(x)$, we have all the pieces of the algorithm. However, using this naive algorithm to compute $\partial H/\partial x$ and $\partial H/\partial v$, we face several challenges.

1. The algorithm involves computing the pseudo-inverse and its derivatives, which takes $O(n^3)$ except for very special matrices.

2. The Lagrangian term in the constrained Hamiltonian dynamics requires additional computation such as the second-order derivative of $c(x)$.

3. A naive approach to computing leverage scores in $\partial H/\partial x$ results in a very dense matrix.

Those challenges make the algorithm hard to implement and inefficient, especially when the dimension is high. In the following paragraphs, we give an overview of how we overcome each of the challenges above. We defer a more detailed discussion of our approaches and the proofs to Appendix F.2.2.

Avoiding Pseudo-inverse and Pseudo-determinant. We are able to show equivalent formulas for $M(x)^\dagger$ and $\log \operatorname{pdet} M(x)$ that can take advantage of sparse linear system solvers. In particular, we show that $M(x)^\dagger = g(x)^{-\frac{1}{2}} \cdot (I - P(x)) \cdot g(x)^{-\frac{1}{2}}$, where

$$P(x) = g(x)^{-\frac{1}{2}} \cdot Dc(x)^\top (Dc(x) \cdot g(x)^{-1} \cdot Dc(x)^\top)^{-1} Dc(x) \cdot g(x)^{-\frac{1}{2}}. \tag{7.9}$$

As mentioned earlier, a majority of convex bodies appearing in practice are of the form $K = \prod_i K_i$, where K_i are constant dimensional convex bodies. In this case, we will choose $g(x)$ to be a block diagonal matrix with each block of size $O(1)$. Hence, the bottleneck of applying $P(x)$ to a vector is simply solving a linear system of the form $(Dc \cdot g^{-1} \cdot Dc^\top)u = b$ for some b. The existing sparse linear system solvers can solve large classes of sparse linear system much faster than $O(n^3)$ time [Dem97]. For $\log \operatorname{pdet} M(x)$, we show

$$\log \operatorname{pdet}(M(x)) = \log \det g(x) + \log \det \left(Dc(x) \cdot g(x)^{-1} \cdot Dc(x)^\top \right) - \log \det \left(Dc(x) \cdot Dc(x)^\top \right). \tag{7.10}$$

This simplification allows us to take advantage of sparse Cholesky decomposition. We prove (7.9) and (7.10) in Lemma 16 and Lemma 17 in Appendix F.2.2. The formulas (7.9) and (7.10) avoid the expensive pseudo-inverse and pseudo-determinant computations, and significantly improve the practical performance of our algorithm.

Simplification for Subspace Constraints. For the case $c(x) = Ax - b$, the Hamiltonian is now

$$H(x,v) = f(x) + \frac{1}{2} v^\top g^{-\frac{1}{2}} (I - P) g^{-\frac{1}{2}} v + \frac{1}{2} \left(\log\det g + \log\det A g^{-1} A^\top - \log\det A A^\top \right) + \lambda^\top c,$$

where $P = g^{-\frac{1}{2}} A^\top (A g^{-1} A^\top)^{-1} A g^{-\frac{1}{2}}$. The key observation is that the algorithm only needs to know $x(h)$ in the HMC dynamics, and not $v(h)$. Thus, we can replace H by any other that produces the same $x(h)$. We show in Lemma 18 (Appendix F.2.2) that the dynamics corresponding to H above is equivalent to the dynamics that corresponds to a much simpler Hamiltonian:

$$H(x,v) = f(x) + \frac{1}{2} v^\top g^{-\frac{1}{2}} (I - P) g^{-\frac{1}{2}} v + \frac{1}{2} \left(\log\det g + \log\det A g^{-1} A^\top \right).$$

Furthermore, we have

$$\frac{dx}{dt} = g^{-\frac{1}{2}} (I - P) g^{-\frac{1}{2}} v, \quad \frac{dv}{dt} = -\nabla f(x) + \frac{1}{2} Dg \left[\frac{dx}{dt}, \frac{dx}{dt} \right] - \frac{1}{2} \mathrm{Tr}(g^{-\frac{1}{2}} (I - P) g^{-\frac{1}{2}} Dg).$$

Efficient Computation of Leverage Score. Even after simplifying the Hamiltonian as above, we still have a term for the leverage scores, $\mathrm{Tr}(g^{-\frac{1}{2}} (I - P) g^{-\frac{1}{2}} Dg)$ in $\frac{dv}{dt}$ so that we need to compute the diagonal entries of $P = g^{-\frac{1}{2}} A^\top (A g^{-1} A^\top)^{-1} A g^{-\frac{1}{2}}$ to compute $\frac{dv}{dt}$. Since $(A g^{-1} A^\top)^{-1}$ can be extremely dense even when A is very sparse, a naive approach such as direct computation of the inverse can lead to a dense-matrix multiplication. To avoid dense-matrix multiplication, our approach is based on the fact that certain entries of $(A g^{-1} A^\top)^{-1}$ can be computed as fast as computing sparse Cholesky decomposition of $A g^{-1} A^\top$ [Tak73, CD95], which can be $O(n)$ time faster than computing $(A g^{-1} A^\top)^{-1}$ in many settings. We first compute the Cholesky decomposition to obtain a sparse triangular matrix L such that $L L^\top = A g^{-1} A^\top$. Then, we show that only entries of $A g^{-1} A^\top$ in $\mathrm{sp}(L) \cup \mathrm{sp}(L^\top)$ matter in computing $\mathrm{diag}(A^\top (A g^{-1} A^\top)^{-1} A)$, where $\mathrm{sp}(L)$ is the sparsity pattern of L. We give the details of our approach in Appendix F.2.2.

7.2.3 Discretization

Explicit integrators such as leapfrog integrator, which are commonly used for Hamiltonian Monte Carlo, are no longer symplectic on general Riemannian manifolds (see Appendix F.3.1). Even though there have been some attempts [Pih15] to make explicit integrators work in the Riemannian setting, its variants do not work for ill-conditioned problems.

Our algorithm uses the *implicit midpoint method* (Algorithm 16) to discretize the Hamiltonian process into steps of step size h and run the process for T iterations. This integrator is reversible and symplectic (so measure-preserving) [HHIL06], which allows us to use a Metropolis filter to ensure the distribution is correct so that we no longer need to solve ODE to accuracy to maintain the correct stationary distribution. We write $H(x,v) = \overline{H}_1(x,v) + \overline{H}_2(x,v)$, where

$$\overline{H}_1(x,v) = f(x) + \frac{1}{2}\left(\log\det g(x) + \log\det Ag(x)^{-1}A^\top\right),$$
$$\overline{H}_2(x,v) = \frac{1}{2}v^\top g(x)^{-\frac{1}{2}}\left(I - P(x)\right)g(x)^{-\frac{1}{2}}v.$$

Starting from (x_0, v_0), in the first step of the integrator, we run the process on the Hamiltonian $\overline{H}_1$ with step size $\frac{h}{2}$ to get $(x_{1/3}, v_{1/3})$. In the second step of the integrator, we run the process on $\overline{H}_2$ with step size h by solving

$$x_{\frac{2}{3}} = x_{\frac{1}{3}} + h\frac{\partial\overline{H}_2}{\partial v}\left(\frac{x_{\frac{1}{3}} + x_{\frac{2}{3}}}{2}, \frac{v_{\frac{1}{3}} + v_{\frac{2}{3}}}{2}\right), \qquad v_{\frac{2}{3}} = v_{\frac{1}{3}} - h\frac{\partial\overline{H}_2}{\partial x}\left(\frac{x_{\frac{1}{3}} + x_{\frac{2}{3}}}{2}, \frac{v_{\frac{1}{3}} + v_{\frac{2}{3}}}{2}\right),$$

iteratively using the Newton's method. This step involves computing the Cholesky decomposition of $(Ag^{-1}A^\top)^{-1}$ using the Cholesky decomposition of $Ag^{-1}A^\top$. In the third step, we run the process on the Hamiltonian $\overline{H}_1$ with step size $\frac{h}{2}$ again to get (x_1, v_1).

We state the complete algorithm (Algorithm 15 and Algorithm 16) with details on the step size in Appendix F.3.1 and give the theoretical guarantees in Appendix F.3.2 (convergence of implicit midpoint method) and Appendix F.4 (independence of condition number).

7.3 Experiments

In this section, we demonstrate the efficiency of our sampler using experiments on real-world datasets and compare our sampler with existing samplers. We demonstrate that CRHMC is able to sample larger models than previously known to be possible, and is significantly faster in terms of rate of convergence and sampling time in Section 7.3.2, along with convergence test in Section 7.3.4. We examine its behavior on benchmark instances such as simplices and Birkhoff polytopes in Section 7.3.3.

7.3.1 Experimental Setting

Settings. We performed experiments on the Standard DS12 v2 model from MS Azure cloud, which has a 2.1GHz Intel Xeon Platinum 8171M CPU and 28GB memory. In

the experiments, we used our MATLAB and C++ implementation of CRHMC[3], which is available here and has been integrated into the COBRA toolbox.

We used twelve constraint-based metabolic models from molecular systems biology in the COBRA Toolbox v3.0 [HAP+19] and ten real-world LP examples randomly chosen from NETLIB LP test sets. A polytope from each model is defined by $\{x \in \mathbb{R}^n : Ax = b, l \leq x \leq u\}$ for $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$, and $l, u \in \mathbb{R}^n$, which is input to CRHMC for uniform sampling. We describe in Appendix F.1 how we preprocessed these dataset, along with full information about the datasets in Table F.1.

Comparison. We used as a baseline the Coordinate Hit-and-Run (CHAR) implemented in two different languages. The former is Coordinate Hit-and-Run with Rounding (CHRR) written in MATLAB [CV16, HCT+17] and the latter is the same algorithm (CDHR) with an R interface and a C++ library, VolEsti [CF20]. We refer readers to Appendix F.1 for the details of these algorithms and our comparison setup. We note that popular sampling packages such as STAN and Pyro were not included in the experiments as they do not support constrained-based models. Even after transforming our dataset to their formats, the transformed dataset were too ill-conditioned for those algorithms to run. CHMC in [BSU12] works only for manifolds implicitly defined by $\{c(x) = 0\}$ for continuously differentiable $c(x)$ with $Dc(x)$ full-rank everywhere, so we could not use it for comparison.

Measurements. To evaluate the quality of sampling methods, we measured two quantities, the *number of steps per effective sample* (i.e., mixing rate) and the *sampling time per effective sample,* T_s. The *effective sample size* (ESS)[4] can be thought of as the number of actual independent samples, taking into account correlation of samples from a target distribution. Thus the number of steps per effective sample is estimated by the total number of steps divided by the ESS, and the sampling time T_s is estimated as the total sampling time until termination divided by the ESS.

Each algorithm attempted to draw 1000 uniform samples, with limits on running time set to 1 day (3 days for the largest instance *ken_18*) and memory usage to 6GB. If an algorithm passes either the time or the memory limit, we stop the algorithm and measure the quantities of interest based on samples drawn until that moment. After getting uni-

[3]Our package can be run to sample from general logconcave densities and has a feature for parallelization.

[4]We use the minimum of the ESS of each coordinate.

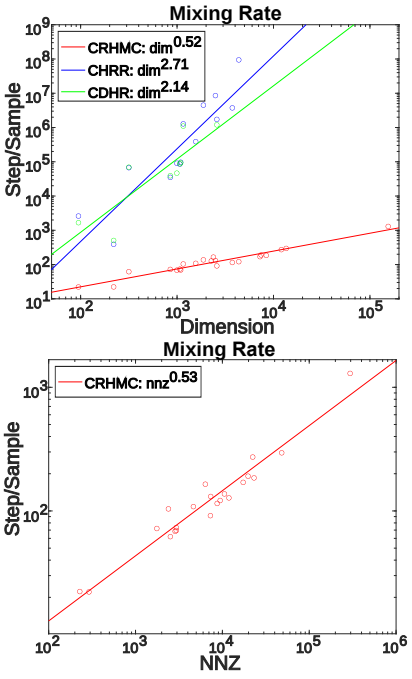


Figure 7.1: Mixing rate of CRHMC and the competitors. Mixing rate of CRHMC was sub-linear in dimension and the nnz of a preprocessed matrix A in a model, whereas the others needed quadratically many steps to converge to uniform distribution. In particular for our dataset, CRHMC mixed up to 6 orders of magnitude earlier than the others. Note that mixing rate of CHAR was very close to quadratic growth when using the full-dimensional scale (the first column in Table F.1).

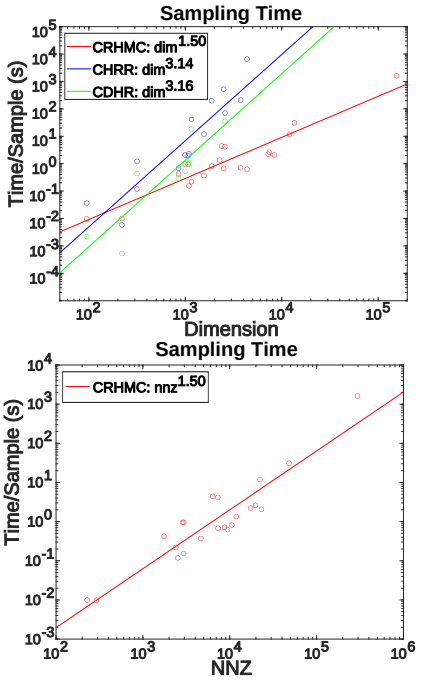


Figure 7.2: Sampling time of CRHMC and the competitors. The sampling time per effective sample of CRHMC was sub-quadratic in dimension and the nnz of a preprocessed matrix A in a model, while the others indicates at least a cubic dependency on dimension. In particular for our dataset, CRHMC was able to obtain a statistically independent sample up to 4 orders of magnitude faster than the others. This benefit of speed-up was actually straightforward from the figure, since CHRR could not obtain enough samples from instances with more than 5000 variables until it ran out of time.

form samples, we thinned the samples twice to ensure independence of samples; first we computed the ESS of the samples, only kept ESS many samples, and repeated this again. We estimated the above quantities only if the ESS is more than 10 and an algorithm does not run into any error while running[5].

7.3.2 Mixing Rate and Sampling Time

Sub-linear Mixing Rate. We examined how the number of steps per effective sample grows with the number of nonzeros (nnz) of matrix A (after preprocessing) and the number of variables (dimension in the plots). To this end, we counted the total number of steps taken until termination of algorithms and divided it by the effective sample size of drawn samples. Note that we thinned twice to ensure independence of samples used.

The mixing rate of CRHMC was sub-linear in both dimension and nnz, whereas previous implementations based on CHAR required at least n^2 steps per sample as seen in Figure 7.1. On the dataset, mixing rate attained was up to 6 orders of magnitude faster for CRHMC compared to CHAR, implying that CRHMC converged to uniform distribution substantially faster than the other competitors. This gap in mixing rate increased super-linearly in dimension, enabling CRHMC to run on large instances of dimension up to 100000.

Sub-quadratic Sampling Time. We next examined the sampling time T_s in terms of both the nnz of A and the dimension of the instance. We computed the runtime of algorithms until their termination divided by the effective sample size of drawn samples, where we ignored the time it takes for preprocessing. Note that the sampling time T_s is essentially multiplication of the mixing rate and the *per-step complexity* (i.e., how much time each step takes).

As shown in Figure 7.2 and Table 7.1, we found that the per-step complexity of CRHMC was small enough to make the sampling time sub-quadratic in both dimension and nnz, whereas CHAR had at least a cubic dependency on dimension, despite of a low per-step complexity. On our dataset, the sampling time of CRHMC was up to 4 orders of magnitude less than that of CHRR and CDHR. While CHRR can be used on dimension only up to a few thousands, increasing benefits of sampling time in higher dimension allows CRHMC

[5]When running CDHR from the VolEsti package on some instances, we got an error message "R session aborted and R encountered a fatal error".

Bio Model	Vars (n)	nnz	CRHMC	CHRR	CDHR
ecoli	95	291	0.0098	0.0365	0.0022
cardiac_mit	220	228	0.0100	0.0059	0.0005
Aci_D21	851	1758	0.4257	0.6884	0.2974
Aci_MR95	994	2859	0.9624	2.0668	0.5237
Abi_49176	1069	2951	0.9608	1.9395	0.9622
Aci_20731	1090	2946	0.1540	2.3014	1.1086
Aci_PHEA	1561	4640	0.3701	12.06	-
iAF1260	2382	6368	4.4355	3687.2	-
iJO1366	2583	7284	4.1608	70.5	35.556
Recon1	3742	8717	0.7184	208.5	-
Recon2	7440	19791	2.6116	10445*	-
Recon3	13543	48187	31.114	29211*	-

LP Model	Vars (n)	nnz	CRHMC	CHRR	CDHR
israel	316	2519	0.1186	1.2224	0.4426
gfrd_pnc	1160	2393	0.2199	40.988	18.468
25fv47	1876	10566	0.8159	199.9	-
pilot_ja	2267	11886	1.3490	5059*	-
sctap2	2500	7334	0.6752	520.2	-
ship08l	4363	9434	0.6258	6512	-
cre_a	7248	17368	2.2205	30455*	-
woodw	8418	23158	2.0689	30307*	-
80bau3b	12061	22341	11.881	47432*	-
ken_18	154699	295946	1616.3	-	-

Table 7.1: Sampling time per effective sample of CHRR and CRHMC. We note that CRHMC is 1000 times faster than CHRR on the latest metabolic network (Recon3). Sampling time with asterisk (*) indicates that the effective sample size is less than 10.

to run on dimension up to 0.1 million.

7.3.3 CRHMC on Structured Instances

To see the behavior of CRHMC on very large instances, we ran the algorithm on three families of structured polytopes – hypercube, simplex, and Birkhoff polytope – up to dimension half-million. We attempted to draw 500 uniform samples with a 1 day time limit (except for 2 days for half-million-dimensional Birkhoff polytope). The definitions of these polytopes are shown in Appendix F.1.1.

To the best of our knowledge, this is the first demonstration that it is possible to sample such a large model. As seen in Figure 7.3, CRHMC can scale smoothly up to half-million dimension on hypercubes and simplices and up to dimension 10^5 for Birkhoff polytopes (we could not obtain a reliable estimate of mixing rate and sampling time on the half-million dimensional Birkhoff polytope, as the ESS is only 16 after 2 days). However, we believe that one can find room for further improvement of CRHMC by tuning parameters or leveraging engineering techniques. We also expect that CRHMC enables us to estimate the volume of B_n for $n \geq 20$, going well beyond the previously best possible dimension.

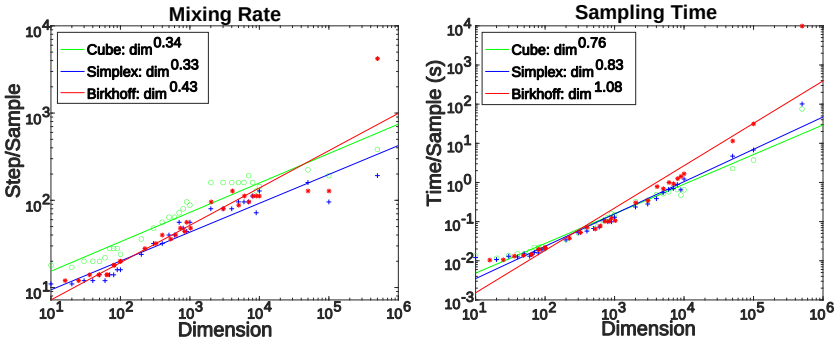


Figure 7.3: Mixing rate and sampling time on structured polytopes including hybercubes, simplices, and Birkhoff polytopes. CRHMC is scalable up to 0.5 million dimension on hypercubes and simplices and up to 0.1 million dimension on Birkhoff polytopes. We note that on the 0.5 million dimensional Birkhoff polytope the ESS is only 16, which is not reliable compared to the ESS on the other instances.

7.3.4 Uniformity Test

We used the following uniformity test to check whether samples from CRHMC form the uniform distribution over a polytope P: check that the fraction of the samples in the scaled set $x \cdot P$ is proportional to $x^{\dim}$. As seen in Figure 7.4, the empirical CDFs of the radial distribution to the power of (1/dim) are close to the CDFs of the uniform distribution over those polytopes.

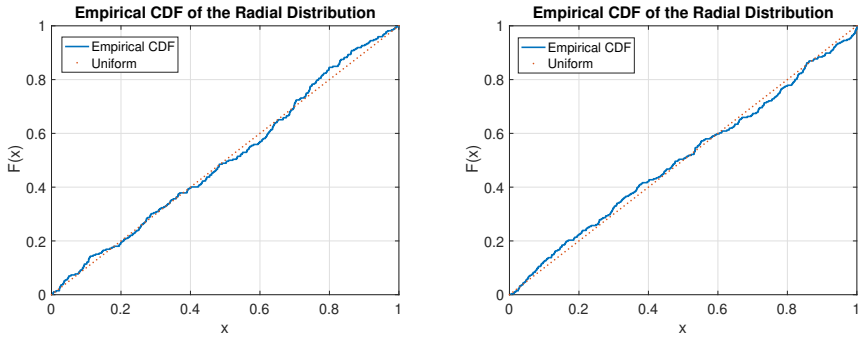


Figure 7.4: We plot the empirical cumulative distribution function of the radial distribution to the power of (1/dim) with 1000 ESS obtained by running CRHMC on *ATCC-49176* (952×1069, left) and *Aci-PHEA* (1319×1561, right), and in the plot x-axis is the scaling factor. We can observe the CDFs are very close to the CDFs of the uniform distribution over the polytopes defined by two instances.

BIBLIOGRAPHY

[Aba16] Martín Abadi. Tensorflow: learning functions at scale. In *Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming, ICFP 2016, Nara, Japan, September 18-22, 2016*, page 1, 2016. 3.1

[AC21] Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-langevin algorithm. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, pages 28405–28418, 2021. 6.1, 6.1.1, 6.1.3, 6.6

[ACCD12] Ery Arias-Castro, Emmanuel J Candes, and Mark A Davenport. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2012.

[AD86] David J. Aldous and Persi Diaconis. Shuffling cards and stopping times. *American Mathematical Monthly*, 93:333–348, 1986. 3.1.1

[ADFDJ03] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43, 2003. 1

[AFKT21] Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in ℓ_1 geometry. *arXiv preprint arXiv:2103.01516*, 2021. 6.1.3

[AH16] Jacob D. Abernethy and Elad Hazan. Faster convex optimization: Simulated annealing with an efficient universal barrier. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2520–2528, 2016. 5.1

[AHK12] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory Comput.*, 8(1):121–164, 2012. 6.1

[AJ94] Richard André-Jeannin. A generalization of morgan-voyce polynomials. *Fibonacci Quarterly*, 32(3), 1994. 4.6.1

[All17] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *J. Mach. Learn. Res.*, 18:221:1–221:51, 2017. 5.1, 4, D.1

[And83] Hans C Andersen. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *Journal of computational Physics*, 52(1):24–34, 1983. 7.2

[ANW10] Alekh Agarwal, Sahand N. Negahban, and Martin J. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010*, pages 37–45. Curran Associates, Inc., 2010. 6.1

[AWBR09] Alekh Agarwal, Martin J Wainwright, Peter Bartlett, and Pradeep Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

[Bac19] Francis Bach. Polynomial magic i: Chebyshev polynomials. `https://francisbach.com/chebyshev-polynomials/`, 2019. 4.1.2, 4.6.1

[Bar20] Alessandro Andrea Barp. *The bracket geometry of statistics*. PhD thesis, Imperial College London, 2020. 2

[BC12] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.*, 5(1):1–122, 2012. 6.1

[BCJ+19] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research (JMLR)*, 20:28:1–28:6, 2019. 7.1

[BCL94] Keith Ball, Eric A. Carlen, and Elliott H. Lieb. Sharp uniform convexity and smoothness estimates for trace norms. *Inventiones mathematicae*, 115(1):463–482, 1994. 6.5.1

[BCM+18] M Barkhagen, NH Chau, É Moulines, M Rásonyi, S Sabanis, and Y Zhang. On stochastic gradient langevin dynamics with dependent data streams in the logconcave case. *arXiv preprint arXiv:1812.02709*, 2018. 5.1.2

[BDMP17] Nicolas Brosse, Alain Durmus, Eric Moulines, and Marcelo Pereyra. Sampling from a log-concave distribution with compact support with proximal langevin monte carlo. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 319–342, 2017. 5.1.2, 5.1.3, 6.1.3

[BE19] Sébastien Bubeck and Ronen Eldan. The entropic barrier: Exponential families, log-concave geometry, and self-concordance. *Math. Oper. Res.*, 44(1):264–276, 2019. 6.1, 6.1.1, 6.1.3, 6.3.1, 6.3.1

[BEL18] Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *Discret. Comput. Geom.*, 59(4):757–783, 2018. 5.1.2, 5.1.3

[Ber18] Espen Bernton. Langevin monte carlo and JKO splitting. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, pages 1777–1798, 2018. 5.1.2, 6.1.3

[Bes94] Julian Besag. Comments on "representations of knowledge in complex systems" by u. grenander and mi miller. *Journal of the Royal Statistical Society, Series B*, 56:591–592, 1994. 6.1, B.1

[BFFN19] Jack Baker, Paul Fearnhead, Emily B Fox, and Christopher Nemeth. Control variates for stochastic gradient mcmc. *Statistics and Computing*, 29(3):599–615, 2019. 5.1.2

[BFR+19] Joris Bierkens, Paul Fearnhead, Gareth Roberts, et al. The zig-zag process and super-efficient sampling for bayesian analysis of big data. *The Annals of Statistics*, 47(3):1288–1320, 2019. 5.1.2

[BFTGT19] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019. 6.1.3

[BGL14] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014. 4.1.1

[BGN21] Raef Bassily, Cristóbal Guzmán, and Anupama Nandi. Non-euclidean differentially private stochastic convex optimization. *arXiv preprint arXiv:2103.01278*, 2021. 6.1.3

[BL76] Herm Jan Brascamp and Elliott H Lieb. On extensions of the brunn-minkowski and prékopa-leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4):366–389, 1976. 3.3, 6.1.2, 6.3.1

[BL97] Sergey G Bobkov and Michel Ledoux. Poincaré's inequalities and talagrand's concentration phenomenon for the exponential distribution. *Probability Theory and Related Fields*, 107(3):383–400, 1997. 3.3

[BL00] Sergey G Bobkov and Michel Ledoux. From brunn-minkowski to brascamp-lieb and to logarithmic sobolev inequalities. *GAFA, Geometric and Functional Analysis*, 10:1028–1052, 2000. 3.3, 6.1.2, 6.4.1, 47, B.2

[BRH13] Nawaf Bou-Rabee and Martin Hairer. Nonasymptotic mixing of the mala algorithm. *IMA Journal of Numerical Analysis*, 33(1):80–110, 2013. 3.1, 6.1

[BST14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pages 464–473. IEEE, 2014. 6.1.3

[BSU12] Marcus Brubaker, Mathieu Salzmann, and Raquel Urtasun. A family of MCMC methods on implicitly defined manifolds. In *Artificial intelligence and statistics (AISTATS)*, pages 161–172, 2012. 7.2, 7.2.1, 7.3.1, F.2.3, F.2.3

[BŠVV08] Ivona Bezáková, Daniel Štefankovič, Vijay V Vazirani, and Eric Vigoda. Accelerating simulated annealing for the permanent and combinatorial counting problems. *SIAM Journal on Computing (SICOMP)*, 37(5):1429–1454, 2008.

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2(1):183–202, 2009. 5.1, 5.1.1, 5.1.3, 6, D.1

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015. 3.1

[BV04] Dimitris Bertsimas and Santosh S. Vempala. Solving convex programs by random walks. *J. ACM*, 51(4):540–556, 2004. 5.1

[CB17] Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. *arXiv preprint arXiv:1705.09048*, 2017. 2.2.1

[CBL20] Niladri S. Chatterji, Peter L. Bartlett, and Philip M. Long. Oracle lower bounds for stochastic gradient sampling algorithms. *CoRR*, abs/2002.00291, 2020. 4.1.3

[CBMR19] Adam D Cobb, Atılım Güneş Baydin, Andrew Markham, and Stephen J Roberts. Introducing an explicit symplectic integration scheme for Riemannian manifold Hamiltonian Monte Carlo. *arXiv preprint arXiv:1910.06243*, 2019. F.3.1

[CCAY+18] Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018. 2.2.1

[CCBJ17] Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663*, 2017. 2.1, 1, 2.2.1, 2.2.1, 2.2.2, 2.3.1, 1, 2.3.2, 2.4, **??**, 3.1, **??**, 1, 5.1.3, 7.1

[CCSW22] Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2984–3014. PMLR, 2022. 6.1, 6.1.1, 6.1.3, 6.6

[CD95] Yogin E Campbell and Timothy A Davis. Computing the sparse inverse subset: an inverse multifrontal approach. *University of Florida, Technical Report TR-95-021*, 1995. 7.2.2, F.2.2

[CDWY20] Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Fast mixing of metropolized hamiltonian monte carlo: Benefits of multi-step gradients.

Journal of Machine Learning Research, 21(92):1–72, 2020. 1, 3.1, 3.1, **??**, 3.1.1, 3.1.1, 3.1.1, 3.1.1, 3.4.1, 4.1.1, 4.1.2, 4.2.4, 4.6.1, 5.1.2, 5.1.3, 5.1.3, 5.4, 10, 5.4.3, 12, 6.1, 7.1, B.1, B.3.3, 18, 19, D.2.2, 20, D.3.3, 84, 10

[CE22] Yuansi Chen and Ronen Eldan. Localization schemes: A framework for proving mixing bounds for markov chains (extended abstract). In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022*, pages 110–122. IEEE, 2022. 6.1, 6.1.1, 6.1.3, 6.6

[CF20] Apostolos Chalkis and Vissarion Fisikopoulos. volEsti: Volume approximation and sampling for convex polytopes in R. *arXiv preprint arXiv:2007.01578*, 2020. 7.3.1

[CFM+18] Niladri S Chatterji, Nicolas Flammarion, Yi-An Ma, Peter L Bartlett, and Michael I Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. *arXiv preprint arXiv:1802.05431*, 2018. 2.1, 2.1, 2.2.1, 5.1.2, **??**, **??**, **??**

[CGH+17] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017. 3.1

[Che69] Jeff Cheeger. A lower bound for the smallest eigenvalue of the laplacian. In *Proceedings of the Princeton conference in honor of Professor S. Bochner*, pages 195–199, 1969. 4.2.1, 4.3

[Che21a] Yuansi Chen. An almost constant lower bound of the isoperimetric coefficient in the kls conjecture. *CoRR*, abs/2011.13661, 2021. 4.1, 6.5.3

[Che21b] Sinho Chewi. The entropic barrier is n-self-concordant. *arXiv preprint arXiv:2112.10947*, 2021.

[Che21c] Sinho Chewi. The entropic barrier is n-self-concordant. *CoRR*, abs/2112.10947, 2021. 6.1.3

[Che23] Sinho Chewi. *Log-Concave Sampling*. 2023.

[CLA+20] Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the metropolis-adjusted langevin algorithm. *CoRR*, abs/2012.12810, 2020. 4.1.1, 4.1.2, 4, 4.1.3, 4.3

[CLA+21] Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the metropolis-adjusted langevin algorithm. In *Conference on Learning Theory*, pages 1260–1300. PMLR, 2021. 1, 6.1.1, 6.6

[CLGL+20] Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet, and Austin Stromme. Exponential ergodicity of mirror-Langevin diffusions. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:19573–19585, 2020. 7.1

[CLW20] Yu Cao, Jianfeng Lu, and Lihan Wang. Complexity of randomized algorithms for underdamped langevin dynamics. *CoRR*, abs/2003.09906, 2020. 4.1.3

[CM08] Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. *Advances in neural information processing systems*, 21, 2008. 6.1.3

[CMS11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011. 6.1.3

[CP22] Sinho Chewi and Aram-Alexandre Pooladian. An entropic generalization of Caffarelli's contraction theorem via covariance inequalities. *arXiv e-prints*, 2022. 6.1.2, 6.1.3, 6.3.1

[Cra38] Harald Cramér. Sur un nouveau théorème-limite de la théorie des probabilités. *Act. Sci. et Ind.*, 736, 1938. 6.1

[CRT06] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006. 6.1

[CST21] Michael B. Cohen, Aaron Sidford, and Kevin Tian. Relative lipschitzness in extragradient methods and a direct recipe for acceleration. In *12th Innovations in Theoretical Computer Science Conference, ITCS 2021*, volume 185 of *LIPIcs*, pages 62:1–62:18, 2021. 6.6

[CV16] Ben Cousins and Santosh Vempala. A practical volume algorithm. *Mathematical Programming Computation*, 8(2):133–160, 2016. 7.1, 7.3.1

[CV18] Ben Cousins and Santosh S. Vempala. Gaussian cooling and $o^*(n^3)$ algorithms for volume and gaussian volume. *SIAM J. Comput.*, 47(3):1237–1273, 2018. 5.1.1

[CV19] Zongchen Chen and Santosh S Vempala. Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions. *arXiv preprint arXiv:1905.02313*, 2019. 1, 1, 2.1, 2.1.1, 2.2.1, **??**, 3.1, **??**, 4.1.1

[CWZ+17] Changyou Chen, Wenlin Wang, Yizhe Zhang, Qinliang Su, and Lawrence Carin. A convergence analysis for a class of practical variance-reduction stochastic gradient mcmc. *arXiv preprint arXiv:1709.01180*, 2017. 5.1.2

[Dal17a] Arnak S. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 678–689, 2017. 3.3, 3.4, A.5

[Dal17b] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):651–676, 2017. 2.2.1, 2.3.2, **??**, 3.1, **??**, 1

[Dav06] Timothy A Davis. *Direct methods for sparse linear systems*. SIAM, 2006. F.2.2

[DBL14] Aaron Defazio, Francis R. Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1646–1654, 2014. 5.1.3

[DCWY19] Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183):1–42, 2019. 1, 3.1, 3.1, **??**, 3.1.1, 3.1.1, 3.1.1, 3.3, 3.4, 6, 4.1.1, 4.1.2, 4.4, 4.7, 5.1.1, 5.1.2, 5.1.3, 5.1.3, 6, 5.4, 5.6, 5.6.2, 6.1, 7.1, D.1

[Dem97] James W Demmel. *Applied numerical linear algebra*. SIAM, 1997. 7.2.2

[DFK91a] Martin Dyer, Alan Frieze, and Ravi Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM (JACM)*, 38(1):1–17, 1991. 1

[DFK91b] Martin E. Dyer, Alan M. Frieze, and Ravi Kannan. A random polynomial time algorithm for approximating the volume of convex bodies. *J. ACM*, 38(1):1–17, 1991. 6.1.1, 6.1.2, 6.1.2, 6.5.3

[DFO20] Jelena Diakonikolas, Maryam Fazel, and Lorenzo Orecchia. Fair packing and covering on a relative scale. *SIAM J. Optim.*, 30(4):3284–3314, 2020. 6.1

[DG17] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. 2.4

[DHK+20] Simon Shaolei Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2020.

[DJWW15] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015. 6.1.1, E.1, E.1, 85, E.1, 86, E.1

[DK19] Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and Their Applications*, 2019. 2.2.1, 5.1.2

[DKL18] Etienne De Klerk and Monique Laurent. Comparison of lasserre's measure-based bounds for polynomial optimization to bounds obtained by simulated annealing. *Mathematics of Operations Research*, 43(4):1317–1325, 2018. 87

[DKPR87] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222, 1987. 7.1

[DM16] Alain Durmus and Eric Moulines. High-dimensional bayesian inference via the unadjusted Langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016. 2.1, 2.1, 1, 2.2.1, 2.2.2, **??**, 8, A.5

[DM17] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017. 2.1, 2.3.2

[DM+19] Alain Durmus, Eric Moulines, et al. High-dimensional bayesian inference via the unadjusted langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019. 3.1, 5.3

[DMM19] Alain Durmus, Szymon Majewski, and Blazej Miasojedow. Analysis of langevin monte carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019. 1, 3.1, 3.1.1, 5

[Doo53] Joseph Leo Doob. *Stochastic Processes*, volume 101. New York, Wiley, 1953. 7

[DRD18] Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *arXiv preprint arXiv:1807.09382*, 2018. 1, 2.1, 2.2.1, **??**, **??**

[DRW+16] Kumar Avinava Dubey, Sashank J Reddi, Sinead A Williamson, Barnabas Poczos, Alexander J Smola, and Eric P Xing. Variance reduction in stochastic gradient langevin dynamics. In *Advances in neural information processing systems*, pages 1154–1162, 2016. 5.1.2

[DSM+16] Alain Durmus, Umut Simsekli, Eric Moulines, Roland Badeau, and Gaël Richard. Stochastic gradient richardson-romberg markov chain monte carlo. In *Advances in Neural Information Processing Systems*, pages 2047–2055, 2016. 5.1.2

[EGZ17] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *arXiv preprint arXiv:1703.01617*, 2017. 2.2.1, 2.3.1

[EK11] Ronen Eldan and Bo"az Klartag. Approximately gaussian marginals and the hyperplane conjecture. *Contemporary Math.*, 545:44–68, 2011. 6.1

[FG04] Matthieu Fradelizi and Olivier Guédon. The extreme points of subsets of s-concave probabilities and a geometric localization theorem. *Discrete & Computational Geometry*, 31(2):327–335, 2004.

[FGKS15] Roy Frostig, Rong Ge, Sham M. Kakade, and Aaron Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2540–2548, 2015. 5.1.3

[FKT20] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020. 6.1.3

[GC11] Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011. 2, 7.1

[GGZ18] Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Global convergence of stochastic gradient hamiltonian monte carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based acceleration. *arXiv preprint arXiv:1809.04618*, 2018. 5.1.2

[GLL20] Rong Ge, Holden Lee, and Jianfeng Lu. Estimating normalizing constants for log-concave distributions: algorithms and lower bounds. In *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 579–586, 2020. 4.1.3

[GLL22] Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. *arXiv preprint arXiv:2203.00263*, 2022. 6.1, 6.1, 6.1.1, 6.1.1, 6.1.2, 6.1.2, 6.1.3, 6.4, 6.4.1, 6.4.1, 6.5.2, 15, E.1, E.1, E.1, 86, E.1

[GLL+23a] Sivakanth Gopi, Yin Tat Lee, Daogao Liu, Ruoqi Shen, and Kevin Tian. Algorithmic aspects of the log-laplace transform and a non-euclidean proximal sampler. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2399–2439. PMLR, 2023. 6

[GLL+23b] Sivakanth Gopi, Yin Tat Lee, Daogao Liu, Ruoqi Shen, and Kevin Tian. Private convex optimization in general norms. In *Proceedings of the 2023*

ACM-SIAM Symposium on Discrete Algorithms, SODA 2023. SIAM, 2023. 6.1, 6.1.1, 6.1.2, 6.1.3, 6.3.2, 6.5.2, 15

[GM91] Saul B Gelfand and Sanjoy K Mitter. Recursive stochastic algorithms for global optimization in r^d. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991. 1

[GMT06] Sharad Goel, Ravi Montenegro, and Prasad Tetali. Mixing time bounds via the spectral profile. *Electronic Journal of Probability*, 11:1–26, 2006. 3.1.1

[Gul92] Osman Guler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992. 5.1.3

[GV22] Khashayar Gatmiry and Santosh S Vempala. Convergence of the riemannian langevin algorithm. *arXiv preprint arXiv:2204.10818*, 2022. 6.1.3

[GW08] Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008. F.2.2

[HAP+19] Laurent Heirendt, Sylvain Arreckx, Thomas Pfau, Sebastián N Mendoza, Anne Richelle, Almut Heinken, Hulda S Haraldsdóttir, Jacek Wachowiak, Sarah M Keating, Vanja Vlasov, et al. Creation and analysis of biochemical constraint-based models using the cobra toolbox v. 3.0. *Nature protocols*, 14(3):639–702, 2019. 1, 7.1, 7.3.1

[Har04] Gilles Hargé. A convex/log-concave correlation inequality for gaussian measure and an application to abstract wiener spaces. *Probability theory and related fields*, 130(3):415–440, 2004. 7

[Har13] Moritz Hardt. The zen of gradient descent. `http://blog.mrtz.org/2013/09/07/the-zen-of-gradient-descent.html`, 2013. 4.1.2, 4.6.1

[HC57] Harish-Chandra. Differential operators on a semisimple lie groups. *American Journal of Mathematics*, 79:87–120, 1957. 6.5.3

[HCT+17] Hulda S Haraldsdóttir, Ben Cousins, Ines Thiele, Ronan MT Fleming, and Santosh Vempala. Chrr: coordinate hit-and-run with rounding for uniform sampling of constraint-based models. *Bioinformatics*, 33(11):1741–1743, 2017. 7.1, 7.3.1, F.1

[HG14a] Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014. 4.7

[HG+14b] Matthew D Hoffman, Andrew Gelman, et al. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research (JMLR)*, 15(1):1593–1623, 2014. 7.1

[HHIL06] Ernst Hairer, Marlis Hochbruck, Arieh Iserles, and Christian Lubich. Geometric numerical integration. *Oberwolfach Reports*, 3(1):805–882, 2006. 7.2.3, F.3.1

[HKRC18] Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored langevin dynamics. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 2883–2892, 2018. 5.1.2, 6.1

[HLL+22] Yuxuan Han, Zhicong Liang, Zhipeng Liang, Yang Wang, Yuan Yao, and Jiheng Zhang. Private streaming sco in ℓ_p geometry with applications in high dimensional online decision making. In *International Conference on Machine Learning*, pages 8249–8279. PMLR, 2022. 6.1.3

[IZ80] C. Itzykson and J.-. Zuber. The planar approximation. ii. *Journal of Mathematical Physics*, 21:411–421, 1980. 6.5.3

[Jia21] Qijia Jiang. Mirror langevin monte carlo: the case under isoperimetry. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, pages 715–725, 2021. 6.1, 6.1.1, 6.1.3

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker-planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[JLLV20] He Jia, Aditi Laddha, Yin Tat Lee, and Santosh S. Vempala. Reducing isotropy and volume to KLS: an o($n^3\psi^2$) volume algorithm. *CoRR*, abs/2008.02146, 2020. 4.1, 2, 6.5.3

[JLT20] Arun Jambulapati, Jerry Li, and Kevin Tian. Robust sub-gaussian principal component analysis and width-independent schatten packing. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020. 6.1

[JLV22] Arun Jambulapati, Yin Tat Lee, and Santosh S. Vempala. A slightly improved bound for the KLS constant. *CoRR*, abs/2208.11644, 2022. 2, 6.5.3

[JSV04] Mark Jerrum, Alistair Sinclair, and Eric Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *Journal of the ACM (JACM)*, 51(4):671–697, 2004.

[JZ13] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 315–323, 2013. 12, 5.1.3, 5.1.3, 5.1.3, 5.1.3, 5.4.3, 13, 12, 5.4.3, D.1

[KJ16] Shiva Prasad Kasiviswanathan and Hongxia Jin. Efficient private empirical risk minimization for high-dimensional learning. In *International Conference on Machine Learning*, pages 488–497. PMLR, 2016.

[KL22] Bo"az Klartag and Joseph Lehec. Bourgain's slicing problem and kls isoperimetry up to polylog. *CoRR*, abs/2203.15551, 2022. 6.5.3

[Kla06] Bo"az Klartag. On convex perturbations with a bounded isotropic constant. *Geometric and Functional Analysis*, 16(6):1274–1290, 2006. 6.1

[KLD+16] Zachary A King, Justin Lu, Andreas Dräger, Philip Miller, Stephen Federowicz, Joshua A Lerman, Ali Ebrahim, Bernhard O Palsson, and Nathan E Lewis. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids research*, 44(D1):D515–D522, 2016. 7.1

[KLM06] Ravi Kannan, László Lovász, and Ravi Montenegro. Blocking conductance and mixing in random walks. *Combinatorics, Probability & Computing*, 15(4):541–570, 2006. 3.1.1, 3.1.1, 3.4, 3.4.1, B.3.1

[KLS95] Ravi Kannan, László Lovász, and Miklós Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13(3):541–559, 1995. 43, 6.3.2

[KLS97] Ravi Kannan, László Lovász, and Miklós Simonovits. Random walks and an $O^*(n^5)$ volume algorithm for convex bodies. *Random Structures & Algorithms*, 11(1):1–50, 1997. 7.1

[KLSV22a] Yunbum Kook, Yin-Tat Lee, Ruoqi Shen, and Santosh Vempala. Sampling with riemannian hamiltonian monte carlo in a constrained space. *Advances in Neural Information Processing Systems*, 35:31684–31696, 2022. 7

[KLSV22b] Yunbum Kook, Yin Tat Lee, Ruoqi Shen, and Santosh S Vempala. Condition-number-independent convergence rate of riemannian hamiltonian monte carlo with numerical integrators. *arXiv preprint arXiv:2210.07219*, 2022. 6.1.3

[KLSV22c] Yunbum Kook, Yin Tat Lee, Ruoqi Shen, and Santosh S. Vempala. Condition-number-independent Convergence Rate of Riemannian Hamiltonian Monte Carlo with Numerical Integrators. *arXiv preprint arXiv:2210.07219*, 2022. 25, **??**, F.4, F.4.1, F.4.1, F.4.2

[KM12] Bo"az Klartag and Emanuel Milman. Centroid bodies and the logarithmic laplace transform: a unified approach. *Journal of Functional Analysis*, 262(1):10–34, 2012. 6.1

[KN12] Ravindran Kannan and Hariharan Narayanan. Random walks on polytopes and an affine interior point method for linear programming. *Mathematics of Operations Research*, 37(1):1–20, 2012. 7.1

[Kra40] Hendrik Anthony Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940. 2.2.1, 5.1.3

[KST09] Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Applications of strong convexity–strong smoothness duality to learning with matrices. *arXiv e-prints*, abs/0910.0610, 2009. 6.1.1, 6.3.1, 6.5.1

[KST12] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on*

Learning Theory, pages 25–1. JMLR Workshop and Conference Proceedings, 2012. 6.1.3

[KV06] Adam Tauman Kalai and Santosh Vempala. Simulated annealing for convex optimization. *Mathematics of Operations Research*, 31(2):253–266, 2006.

[LC22] Jiaming Liang and Yongxin Chen. A proximal algorithm for sampling from non-smooth potentials. In *2022 Winter Simulation Conference (WSC)*, pages 3229–3240, 2022. 6.1, 6.1.3

[Led99] Michel Ledoux. *Concentration of measure and logarithmic Sobolev inequalities*. Seminaire de probabilities XXXIII, 1999. 3.1.1, 3.3, 3.3, 6.1.2, 6.4.1, 47, B.2

[Lee18] Yin Tat Lee. Lecture 8: Stochastic methods and applications. Class notes, UW CSE 599: Interplay between Convex Optimization and Geometry, 2018. 5

[LF09] Tom Lyche and Michael Floater. Lecture 2 inf-mat 4350 2009. `https://www.uio.no/studier/emner/matnat/ifi/nedlagte-emner/INF-MAT4350/h09/undervisningsmateriale/lecture2.pdf`, 2009.

[Lin94] Ernest Lindelof. Sur lapplication de la methode des approximations successives aux equations differentielles ordinaires du premier ordre. *Comptes rendus hebdomadaires des seances de lAcademie des sciences*, 116(3):454–457, 1894. 2.1.1

[LK99] László Lovász and Ravi Kannan. Faster mixing via average conductance. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing, May 1-4, 1999, Atlanta, Georgia, USA*, pages 282–287, 1999. 3.1.1, 5.1.4, 6.6

[LLV20] Aditi Laddha, Yin Tat Lee, and Santosh Vempala. Strong self-concordance and sampling. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1212–1222, 2020.

[LM00] Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. 2, B.4

[LMH15] Hongzhou Lin, Julien Mairal, and Zaïd Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3384–3392, 2015. 5.1.3, 5.1.3

[LMV21] Jonathan Leake, Colin S. McSwiggen, and Nisheeth K. Vishnoi. Sampling matrices from harish-chandra-itzykson-zuber densities with applications to quantum inference and differential privacy. In *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1384–1397. ACM, 2021. 6.5.3

[LNP12] Nathan E Lewis, Harish Nagarajan, and Bernhard O Palsson. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*, 10(4):291–305, 2012. 7.1

[LPW09] David Asher Levin, Yuval Peres, and Elizabeth Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009. 5.3, B.3.4

[LS93] László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random structures & algorithms*, 4(4):359–412, 1993. 3.1.1, 6.1.1, 6.1.2, 6.3.2, 6.3.2, 6.4.2, 52

[LS14] Yin Tat Lee and Aaron Sidford. Path finding methods for linear programming: Solving linear programs in $O(\sqrt{\text{rank}})$ iterations and faster algorithms for maximum flow. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 424–433. IEEE, 2014.

[LST20] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Logsmooth gradient concentration and tighter runtimes for metropolized hamiltonian monte carlo. In *Conference on learning theory*, pages 2565–2597. PMLR, 2020. 1, 1, 3, 4.1.1, 4.1.2, 4.1.3, 4.2.4, 5.1.1, 5.1.1, 5.1.1, **??**, 5.1.3, 5.1.3, 5.1.3, 5.4, 5.4.1, 7, 5.6.2, 6.1, 6.1.1, 6.6, 7.1

[LST21a] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Lower bounds on metropolized sampling methods for well-conditioned distributions. *Advances in Neural Information Processing Systems*, 34:18812–18824, 2021. 4, 6.1.1

[LST21b] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted gaussian oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021. 1, 4.7, 5, 6.1, 6.1, 6.1, 6.1.1, 6.1.3, 6.4, 6.4, 6.6

[LSV18] Yin Tat Lee, Zhao Song, and Santosh S Vempala. Algorithmic theory of ODEs and sampling from well-conditioned logconcave densities. *arXiv preprint arXiv:1812.06243*, 2018. 2.1, 2.1.1, 1, 2.2.1, **??**, 3.1, **??**, 3.1.1

[LTVW22] Ruilin Li, Molei Tao, Santosh S. Vempala, and Andre Wibisono. The mirror langevin algorithm converges with vanishing bias. In *International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pages 718–742. PMLR, 2022. 6.1, 6.1.1, 6.1.3

[LV06a] László Lovász and Santosh Vempala. Hit-and-Run from a corner. *SIAM Journal on Computing*, 35(4):985–1005, 2006. 2.1, 5.1.3, 7.1, F.1

[LV06b] László Lovász and Santosh S. Vempala. Fast algorithms for logconcave functions: Sampling, rounding, integration and optimization. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*, pages 57–68, 2006. 5.1.3

[LV07] László Lovász and Santosh S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Struct. Algorithms*, 30(3):307–358, 2007. 4.1, 2, 6.5.3

[LV17] Yin Tat Lee and Santosh S Vempala. Geodesic walks in polytopes. In *Proceedings of the 49th Annual ACM SIGACT Symposium on theory of Computing (STOC)*, pages 927–940, 2017.

[LV18] Yin Tat Lee and Santosh 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*, pages 1115–1121, 2018. 2.2.1, 6.1.2, 6.1.3, 6.3.2, 51, 7.1, F.2.3, F.4.1, F.4.1

[LV21] Aditi Laddha and Santosh Vempala. Convergence of Gibbs sampling: Coordinate Hit-and-Run mixes fast. *The 37th International Symposium on Computational Geometry (SoCG)*, 2021. F.1

[LW95] László Lovász and Peter Winkler. Efficient stopping rules for markov chains. In *Proceedings of the Twenty-Seventh Annual ACM Symposium on Theory of Computing, 29 May-1 June 1995, Las Vegas, Nevada, USA*, pages 76–82, 1995. 3.1.1

[LY21] Yin Tat Lee and Man-Chung Yue. Universal barrier is n-self-concordant. *Mathematics of Operations Research*, 2021. 7.1

[Mai07] Francesco Mainardi. Lévy stable distributions in the theory of probability. *Lecture Notes on Mathematical Physics*, 2007. 6.5.3

[MCC+21] Yi-An Ma, Niladri S Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L Bartlett, and Michael I Jordan. Is there an analog of nesterov acceleration for gradient-based mcmc? 2021. 2.2.1

[MCF15] Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pages 2917–2925, 2015. 2.2.1

[MCJ+18] Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I. Jordan. Sampling can be faster than optimization. *CoRR*, abs/1811.08413, 2018. 3.1

[McS21] Colin McSwiggen. The harish-chandra integral: An introduction with examples. *L'Enseignement Mathématique*, 67(3):229–299, 2021. 6.5.3

[MFWB19] Wenlong Mou, Nicolas Flammarion, Martin J. Wainwright, and Peter L. Bartlett. An efficient sampling algorithm for non-smooth composite potentials. *CoRR*, abs/1910.00551, 2019. 5.1.1, 5.1.1, 5.1.2, 6.1.3

[MMW+19] Wenlong Mou, Yi-An Ma, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. High-order langevin diffusion yields an accelerated mcmc algorithm. *arXiv preprint arXiv:1908.10859*, 2019. 2.1, **??**, 3.1, **??**

[MS17] Oren Mangoubi and Aaron Smith. Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. *arXiv preprint arXiv:1708.07114*, 2017. 2.1, 2.1, 2.2.1, 2.2.2, **??**, 3.1

[MV18] Oren Mangoubi and Nisheeth Vishnoi. Dimensionally tight bounds for second-order Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 6027–6037, 2018. 2.1, 2.1, 2.2.1, 2.2.2

[Nar16] Hariharan Narayanan. Randomized interior point methods for sampling and optimization. *The Annals of Applied Probability*, 26(1):597–641, 2016.

[NDH+17] Tigran Nagapetyan, Andrew B Duncan, Leonard Hasenclever, Sebastian J Vollmer, Lukasz Szpruch, and Konstantinos Zygalakis. The true cost of stochastic gradient langevin dynamics. *arXiv preprint arXiv:1706.02692*, 2017. 5.1.2

[Nea11] Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte cCarlo*, 2(11):2, 2011. 2.2.1, 5.1.3, 7.1

[Nem04] Arkadi Nemirovski. Interior point polynomial time methods in convex programming. *Lecture notes*, 42(16):3215–3224, 2004. 6.1, 37

[Nes83] Yurii Nesterov. A method for solving a convex programming problem with convergence rate $o(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983. 5.1, 5.4.3, 6.6, D.1

[Nes03] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course, volume I.* 2003. 3.1, 4.1

[NF19] Christopher Nemeth and Paul Fearnhead. Stochastic gradient markov chain monte carlo. *arXiv preprint arXiv:1907.06986*, 2019. 5.1.2

[NN94] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994. 7.1

[NS21] Hariharan Narayanan and Piyush Srivastava. On the mixing time of coordinate Hit-and-Run. *Combinatorics, Probability and Computing*, pages 1–13, 2021. F.1

[NT02] Yurii E Nesterov and Michael J Todd. On the riemannian geometry defined by self-concordant barriers and interior-point methods. *Foundations of Computational Mathematics*, 2(4):333–361, 2002. 6.1, 6.2, 36

[NY83] A. Nemirovski and D.B̃. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983. 2, 6.1

[ODL+20] F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. V. Saunders B. R. Mille and, H. S. Cohl, and eds. M. A. McClain. Nist digital library of mathematical functions, 2020. 4.3

[OV00] Felix Otto and Cédric Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000. 5.3

[PB14] Neal Parikh and Stephen P. Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, 2014. 5.1.3

[Per16] Marcelo Pereyra. Proximal markov chain monte carlo algorithms. *Stat. Comput.*, 26(4):745–760, 2016. 5.1.2, 6.1.3

[Pic98] Emile Picard. Sur les methodes dapproximations successives dans la theorie des equations differentielles. *American Journal of Mathematics*, pages 87–100, 1898. 2.1.1

[Pih15] Pauli Pihajoki. Explicit methods in extended phase space for inseparable hamiltonian problems. *Celestial Mechanics and Dynamical Astronomy*, 121(3):211–231, 2015. 7.2.3

[PST+12] Natesh S Pillai, Andrew M Stuart, Alexandre H Thiéry, et al. Optimal scaling and diffusion limits for the langevin algorithm in high dimensions. *The Annals of Applied Probability*, 22(6):2320–2356, 2012. 3.1

[Rei93] Sebastian Reich. *Symplectic integration of constrained Hamiltonian systems by Runge-Kutta methods*. University of British Columbia, Department of Computer Science, 1993. 7.2

[Roc76] R Tyrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976. 5.1.3

[RRT17] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017. 2.2.1

[RT96a] Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. 1, 3.1, 6.1, 7.1

[RT96b] Gareth O Roberts and Richard L Tweedie. Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms. *Biometrika*, 83(1):95–110, 1996. 3.1

[RVR+18] Daniel J Russo, Benjamin Van Roy, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018. 1

[SBCR16] Umut Simsekli, Roland Badeau, Taylan Cemgil, and Gaël Richard. Stochastic quasi-newton Langevin Monte Carlo. In *International Conference on Machine Learning (ICML)*, pages 642–651. PMLR, 2016. 7.1

[Sha07] Shai Shalev-Shwartz. Online learning: Theory, algorithms, and applications. PhD thesis, Hebrew University, 2007. 6.1.1

[SJ89] Alistair Sinclair and Mark Jerrum. Approximate counting, uniform generation and rapidly mixing markov chains. *Inf. Comput.*, 82(1):93–133, 1989. 3.1.1

[SKR19] Adil Salim, Dmitry Koralev, and Peter Richtárik. Stochastic proximal langevin algorithm: Potential splitting and nonasymptotic rates. In *Advances in Neural Information Processing Systems*, pages 6653–6664, 2019. 5.1.2

[SL19] Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems*, pages 2098–2109, 2019. 2, 3.1, **??**, 3.1.1, 4.1.3, **??**, 5.1.3, 1, 7.1

[SRB17] Mark Schmidt, Nicolas Le Roux, and Francis R. Bach. Minimizing finite sums with the stochastic average gradient. *Math. Program.*, 162(1-2):83–112, 2017. 5.1.3

[Sta20] Stan Development Team. RStan: the R interface to Stan, 2020. R package version 2.21.2. 7.1

[SWF16] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016. 7.1

[Tak73] Kazuhiro Takahashi. Formation of sparse bus impedance matrix and its application to short circuit study. In *Proceeding of PICA Conference, June, 1973*, 1973. 7.2.2, F.2.2

[Tal19] Kunal Talwar. Computational separations between sampling and optimization. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 14997–15007, 2019. 3.1

[Tao13] Terence Tao. The harish-chandra-itzykson-zuber integral formula. https://terrytao.wordpress.com/2013/02/08/the-harish-chandra-itzykson-zuber-integral-formula/, 2013. Accessed: 2023-02-05. 6.5.3

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1):267–288, 1996. 5.1.3

[TN10] Levent Tunçel and Arkadi Nemirovski. Self-concordant barriers for convex approximations of structured convex sets. *Foundations of Computational Mathematics*, 10(5):485–525, 2010.

[TSF+13] Ines Thiele, Neil Swainston, Ronan MT Fleming, Andreas Hoppe, Swagatika Sahoo, Maike K Aurich, Hulda Haraldsdottir, Monica L Mo, Ottar Rolfsson, Miranda D Stobbe, et al. A community-driven global reconstruction of human metabolism. *Nature biotechnology*, 31(5):419–425, 2013. 7.1

[TTZ15] Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Nearly-optimal private lasso. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 3025–3033, 2015. 6.1.3

[Vai96] Pravin M Vaidya. A new algorithm for minimizing convex functions over convex sets. *Mathematical programming*, 73(3):291–341, 1996.

[Vem05] Santosh Vempala. Geometric random walks: A survey. *MSRI Combinatorial and Computational Geometry*, 52:573–612, 2005. 3.1

[Vem10] Santosh S Vempala. Recent progress and open problems in algorithmic convex geometry. In *IARCS Annual Conference on Foundations of Software Tech-*

nology and Theoretical Computer Science (FSTTCS 2010). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2010. 1

[VW19] Santosh S. Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 8092–8104, 2019. 3.1.1, 3.3

[WA18] Jun-Kun Wang and Jacob D. Abernethy. Acceleration through optimistic no-regret dynamics. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, pages 3828–3838, 2018. 6.6

[Wib19] Andre Wibisono. Proximal langevin algorithm: Rapid convergence under isoperimetry. *CoRR*, abs/1911.01469, 2019. 5.1.2, 6.1.3

[WS16] Blake E. Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3639–3647, 2016. 5.1.1, 5.1.3

[WT11] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011. 5.1.2

[WYX17] Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017. 6.1.3

[XSL+14] Tatiana Xifara, Chris Sherlock, Samuel Livingstone, Simon Byrne, and Mark Girolami. Langevin diffusions and the metropolis-adjusted langevin algorithm. *Statistics & Probability Letters*, 91:14–19, 2014. 3.1

[ZH05] Hui Zou and Trevor Hastie. Regularization and variable section via the elastic net. *Journal of the Royal Statistical Society, Series B (Methodological)*, 67(2):301–320, 2005. 5.1.3

[ZLC17] Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. *arXiv preprint arXiv:1702.05575*, 2017. 2.2.1

[ZPFP20] Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. Wasserstein control of mirror langevin monte carlo. In *Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3814–3841. PMLR, 2020. 6.1, 6.1.1, 6.1.3

[ZXG18] Difan Zou, Pan Xu, and Quanquan Gu. Subsampled stochastic variance-reduced gradient langevin dynamics. In *International Conference on Uncertainty in Artificial Intelligence*, 2018. 5.1.2, **??**

Appendix A

DEFERRED CONTENTS FROM CHAPTER 2

A.1 Brownian Motion Simulation

In this section, we introduce how W_1, W_2 and W_3 can be sampled. Let $\{B_t\}_{t\in[0,h]}$ be the standard d-dimensional Brownian motion on $t \in [0,h]$. In Algorithm 1, $W_1 = \int_0^{\alpha h}\left(1-e^{-2(\alpha h-s)}\right)\mathrm{d}B_s$, $W_2 = \int_0^{h}\left(1-e^{-2(h-s)}\right)\mathrm{d}B_s$ and $W_3 = \int_0^{h} e^{-2(h-s)}\,\mathrm{d}B_s$. We define $G_1 = \int_0^{\alpha h} e^{2s}\,\mathrm{d}B_s$, $G_2 = \int_{\alpha h}^{h} e^{2s}\,\mathrm{d}B_s$, $H_1 = \int_0^{\alpha h}\mathrm{d}B_s$ and $H_2 = \int_{\alpha h}^{h}\mathrm{d}B_s$. Then, $W_1 = H_1 - e^{-2\alpha h}G_1$, $W_2 = (H_1+H_2) - e^{-2h}(G_1+G_2)$ and $W_3 = e^{-2h}(G_1+G_2)$. It is sufficient to sample H_1, H_2, G_1 and G_2. We can show that (G_1, H_1) is independent of (G_2, H_2), and (G_1, H_1) and (G_2, H_2) both follow a $2d$-dimensional Gaussian distribution, which can be easily sampled.

Lemma 5. *Define* $G_1 = \int_0^{\alpha h} e^{2s}\,\mathrm{d}B_s$, $G_2 = \int_{\alpha h}^{h} e^{2s}\,\mathrm{d}B_s$, $H_1 = \int_0^{\alpha h}\mathrm{d}B_s$ *and* $H_2 = \int_{\alpha h}^{h}\mathrm{d}B_s$. *Then,* (G_1, H_1) *is independent of* (G_2, H_2) *. Moreover,* (G_1, H_1) *and* (G_2, H_2) *both follow a* $2d$*-dimensional Gaussian distribution with mean zero. Conditional on the choice of* α*, their covariance is given by*

$$\begin{aligned}
\mathbb{E}\left[(G_1-\mathbb{E}G_1)(H_1-\mathbb{E}H_1)^T\right] &= \frac{1}{2}\left(e^{2\alpha h}-1\right)\cdot I_d,\\
\mathbb{E}\left[(G_1-\mathbb{E}G_1)(G_1-\mathbb{E}G_1)^T\right] &= \frac{1}{4}\left(e^{4\alpha h}-1\right)\cdot I_d,\\
\mathbb{E}\left[(H_1-\mathbb{E}H_1)(H_1-\mathbb{E}H_1)^T\right] &= \alpha h\cdot I_d,\\
\mathbb{E}\left[(G_2-\mathbb{E}G_2)(H_2-\mathbb{E}H_2)^T\right] &= \frac{1}{2}\left(e^{2h}-e^{2\alpha h}\right)\cdot I_d,\\
\mathbb{E}\left[(G_2-\mathbb{E}G_2)(G_2-\mathbb{E}G_2)^T\right] &= \frac{1}{4}\left(e^{4h}-e^{4\alpha h}\right)\cdot I_d,\\
\mathbb{E}\left[(H_2-\mathbb{E}H_2)(H_2-\mathbb{E}H_2)^T\right] &= (h-\alpha h)\cdot I_d.
\end{aligned}$$

Proof. By the definition of the standard Brownian motion, (G_1, H_1) is independent of (G_2, H_2) and (G_1, H_1) and (G_2, H_2) both have mean zero. Moreover,

$$\begin{aligned}
\mathbb{E}\left[(G_1-\mathbb{E}G_1)(H_1-\mathbb{E}H_1)^T\right] &= \mathbb{E}\left[\left(\int_0^{\alpha h} e^{2s}\,\mathrm{d}B_s\right)\left(\int_0^{\alpha h}\mathrm{d}B_s\right)^T\right] = \int_0^{\alpha h} e^{2s}\,\mathrm{d}s\cdot I_d\\
&= \frac{1}{2}\left(e^{2\alpha h}-1\right)\cdot I_d,
\end{aligned}$$

$$\mathbb{E}\left[(G_1 - \mathbb{E}G_1)(G_1 - \mathbb{E}G_1)^T\right] = \mathbb{E}\left[\left(\int_0^{\alpha h} e^{2s}\,\mathrm{d}B_s\right)\left(\int_0^{\alpha h} e^{2s}\,\mathrm{d}B_s\right)^T\right] = \int_0^{\alpha h} e^{4s}\,\mathrm{d}s \cdot I_d$$
$$= \frac{1}{4}\left(e^{4\alpha h} - 1\right) \cdot I_d,$$

and

$$\mathbb{E}\left[(H_1 - \mathbb{E}H_1)(H_1 - \mathbb{E}H_1)^T\right] \quad = \quad \alpha h \cdot I_d.$$

Similarly,

$$\mathbb{E}\left[(G_2 - \mathbb{E}G_2)(H_2 - \mathbb{E}H_2)^T\right] = \mathbb{E}\left[\left(\int_{\alpha h}^{h} e^{2s}\,\mathrm{d}B_s\right)\left(\int_{\alpha h}^{h} \mathrm{d}B_s\right)^T\right] = \int_{\alpha h}^{h} e^{2s}\,\mathrm{d}s \cdot I_d$$
$$= \frac{1}{2}\left(e^{2h} - e^{2\alpha h}\right) \cdot I_d,$$

$$\mathbb{E}\left[(G_1 - \mathbb{E}G_1)(G_1 - \mathbb{E}G_1)^T\right] = \mathbb{E}\left[\left(\int_{\alpha h}^{h} e^{2s}\,\mathrm{d}B_s\right)\left(\int_{\alpha h}^{h} e^{2s}\,\mathrm{d}B_s\right)^T\right] = \int_{\alpha h}^{h} e^{4s}\,\mathrm{d}s \cdot I_d$$
$$= \frac{1}{4}\left(e^{4h} - e^{4\alpha h}\right) \cdot I_d,$$

and

$$\mathbb{E}\left[(H_2 - \mathbb{E}H_2)(H_2 - \mathbb{E}H_2)^T\right] \quad = \quad (h - \alpha h) \cdot I_d.$$

□

A.2 Properties of the ULD and the Brownian motion

Here, we prove some properties of the ULD and the Brownian motion. These properties are used in Appendices A.3, A.4, A.5 and A.6 to prove the guarantee of our algorithm.

A.2.1 Properties of the ULD

Lemma 6. *Let* $\{x(t)\}_{t\in[0,h]}$ *and* $\{v(t)\}_{t\in[0,h]}$ *be the solution to the underdamped Langevin diffusion* (2.3) *on* $t \in [0,h]$. *Assume that* $h \le \frac{1}{20}$ *and* $u = \frac{1}{L}$. *We have the following bounds.*

$$\begin{aligned}
\mathbb{E}\sup_{t\in[0,h]} \|v(t)\|^2 &\le O\left(\|v(0)\|^2 + u^2h^2\|\nabla f(x(0))\|^2 + udh\right), \\
\mathbb{E}\sup_{t\in[0,h]} \|\nabla f(x(t))\|^2 &\le O\left(\|\nabla f(x(0))\|^2 + L^2h^2\|v(0)\|^2 + Ldh^3\right), \\
\mathbb{E}\sup_{t\in[0,h]} \|x(0) - x(t)\|^2 &\le O\left(h^2\|v(0)\|^2 + u^2h^4\|\nabla f(x(0))\|^2 + udh^3\right),
\end{aligned}$$

and

$$\begin{aligned} -\mathbb{E}\inf_{t\in[0,h]}\|v(t)\|^2 &\leq -\frac{1}{3}\|v(0)\|^2 + O\left(u^2h^2\|\nabla f(x(0))\|^2 + udh\right), \\ -\mathbb{E}\inf_{t\in[0,h]}\|\nabla f(x(t))\|^2 &\leq -\frac{1}{3}\|\nabla f(x(0))\|^2 + O\left(h^2L^2\|v(0)\|^2 + Ldh^3\right). \end{aligned}$$

Proof. We first show the first three bounds. We can write $\mathbb{E}\sup_{t\in[0,h]}\|\nabla f(x(t))\|^2$ as

$$\begin{aligned} &\mathbb{E}\sup_{t\in[0,h]}\|\nabla f(x(t))\|^2 \\ \leq\ & 2\|\nabla f(x(0))\|^2 + 2\mathbb{E}\sup_{t\in[0,h]}\|\nabla f(x(0)) - \nabla f(x(t))\|^2 \\ \leq\ & 2\|\nabla f(x(0))\|^2 + 2L^2\mathbb{E}\sup_{t\in[0,h]}\|x(0) - x(t)\|^2, \end{aligned} \tag{A.1}$$

where the first step follows by Young's inequality and the second step follows by ∇f is L-Lipschitz. To bound $\mathbb{E}\sup_{t\in[0,h]}\|x(0) - x(t)\|^2$,

$$\begin{aligned} \mathbb{E}\sup_{t\in[0,h]}\|x(0) - x(t)\|^2 &= \mathbb{E}\sup_{t\in[0,h]}\left\|\int_0^t v(s)\,\mathrm{d}s\right\|^2 \\ &\leq \mathbb{E}\sup_{t\in[0,h]} t\int_0^t\|v(s)\|^2\,\mathrm{d}s \\ &\leq h^2\mathbb{E}\sup_{t\in[0,h]}\|v(t)\|^2, \end{aligned} \tag{A.2}$$

where the first step follows by the definition of x and the second follows by the Cauchy-Schwarz inequality. To bound $\mathbb{E}\sup_{t\in[0,h]}\|v(t)\|^2$,

$$\begin{aligned} \mathbb{E}\sup_{t\in[0,h]}\|v(t)\|^2 &= \mathbb{E}\sup_{t\in[0,h]}\left\|v(0)e^{-2t} - u\int_0^t e^{-2(t-s)}\nabla f(x(s))\,\mathrm{d}s + 2\sqrt{u}\int_0^t e^{-2(t-s)}\,\mathrm{d}B_s\right\|^2 \\ &\leq 3\|v(0)\|^2 + 3u^2h^2\mathbb{E}\sup_{t\in[0,h]}\|\nabla f(x(t))\|^2 + 12u\mathbb{E}\sup_{t\in[0,h]}\left\|\int_0^t e^{-2(t-s)}\,\mathrm{d}B_s\right\|^2 \\ &\leq 3\|v(0)\|^2 + 3u^2h^2\mathbb{E}\sup_{t\in[0,h]}\|\nabla f(x(t))\|^2 + 60udh, \end{aligned} \tag{A.3}$$

where the first step follows by the definition of ULD, the second step follows by the inequality $(a+b+c)^2 \leq 3a^2 + 3b^2 + 3c^2$ and the third step follows by Lemma 8. Then, combining $(A.1)$, (A.2) and (A.3), we have

$$\begin{aligned} \mathbb{E}\sup_{t\in[0,h]}\|\nabla f(x(t))\|^2 &\leq 2\|\nabla f(x(0))\|^2 + 2L^2\mathbb{E}\sup_{t\in[0,h]}\|x(0) - x(t)\|^2 \\ &\leq 2\|\nabla f(x(0))\|^2 + 2L^2h^2\mathbb{E}\sup_{t\in[0,h]}\|v(t)\|^2 \\ &\leq 2\|\nabla f(x(0))\|^2 + 6h^4\mathbb{E}\sup_{t\in[0,h]}\|\nabla f(x(t))\|^2 + 6L^2h^2\|v(0)\|^2 + 120Ldh^3. \end{aligned}$$

Since $6h^4 \leq \frac{1}{4}$,

$$\begin{aligned}\mathbb{E}\sup_{t\in[0,h]}\|\nabla f(x(t))\|^2 &\leq 3\|\nabla f(x(0))\|^2 + 8L^2h^2\|v(0)\|^2 + 160Ldh^3 \\ &\leq O\left(\|\nabla f(x(0))\|^2 + L^2h^2\|v(0)\|^2 + Ldh^3\right). \qquad (A.4)\end{aligned}$$

By $(A.3)$ and $(A.4)$,

$$\begin{aligned}\mathbb{E}\sup_{t\in[0,h]}\|v(t)\|^2 &\leq 3\|v(0)\|^2 + 3u^2h^2\mathbb{E}\sup_{t\in[0,h]}\|\nabla f(x(t))\|^2 + 60udh \\ &\leq 3\|v(0)\|^2 + 3u^2h^2\cdot O\left(\|\nabla f(x(0))\|^2 + L^2h^2\|v(0)\|^2 + Ldh^3\right) + 60udh \\ &\leq O\left(\|v(0)\|^2 + u^2h^2\|\nabla f(x(0))\|^2 + udh\right).\end{aligned}$$

where the last step follows by h is small.

By (A.2) and (A.4),

$$\begin{aligned}\mathbb{E}\sup_{t\in[0,h]}\|x(0)-x(t)\|^2 &\leq h^2\mathbb{E}\sup_{t\in[0,h]}\|v(t)\|^2 \\ &\leq O\left(h^2\|v(0)\|^2 + u^2h^4\|\nabla f(x(0))\|^2 + udh^3\right). \qquad (A.5)\end{aligned}$$

To prove the fourth claim,

$$\begin{aligned}&\inf_{t\in[0,h]}\|v(t)\|^2 \\ =\ &\inf_{t\in[0,h]}\left\|v(0)e^{-2t} - u\int_0^t e^{-2(t-s)}\nabla f(x(s))\,\mathrm{d}s + 2\sqrt{u}\int_0^t e^{-2(t-s)}\,\mathrm{d}B_s\right\|^2 \\ \geq\ &\inf_{t\in[0,h]}\left[e^{-4t}\|v(0)\|^2 - 2e^{-2t}v(0)^T\left(u\int_0^t e^{-2(t-s)}\nabla f(x(s))\,\mathrm{d}s\right)\right. \\ &\left.+2e^{-2t}v(0)^T\left(2\sqrt{u}\int_0^t e^{-2(t-s)}\,\mathrm{d}B_s\right)\right] \\ \geq\ &\inf_{t\in[0,h]}\left[e^{-4t}\|v(0)\|^2 - \frac{1}{2}e^{-4t}\|v(0)\|^2 - 4\left\|u\int_0^t e^{-2(t-s)}\nabla f(x(s))\,\mathrm{d}s\right\|^2\right. \\ &\left.-4\left\|2\sqrt{u}\int_0^t e^{-2(t-s)}\,\mathrm{d}B_s\right\|^2\right] \\ \geq\ &\inf_{t\in[0,h]}\left[\frac{1}{2}(1-4h)\|v(0)\|^2 - 4u^2h^2\sup_{s\in[0,t]}\|\nabla f(x(s))\|^2 - 16u\left\|\int_0^t e^{-2(t-s)}\,\mathrm{d}B_s\right\|^2\right] \\ \geq\ &\frac{1}{2}(1-4h)\|v(0)\|^2 - 4u^2h^2\sup_{t\in[0,h]}\|\nabla f(x(t))\|^2 - 16u\sup_{t\in[0,h]}\left\|\int_0^t e^{-2(t-s)}\,\mathrm{d}B_s\right\|^2,\end{aligned}$$

where the first step follows by the definition of v , the second step follows by the inequality $(a+b+c)^2 \geq a^2 + 2a(b+c)$, the third step follows by the inequality $2ab \leq a^2 + b^2$, the fourth step follows by $e^{-4t} \geq 1 - 4t$, and the last step follows by h is small.

Then, by $(A.4)$ and Lemma 8,

$$-\mathbb{E}\inf_{t\in[0,h]}\|v(t)\|^2 \;\leq\; -\frac{1}{3}\|v(0)\|^2 + O\left(u^2h^2\|\nabla f(x(0))\|^2 + udh\right).$$

To show the lower bound on $\mathbb{E}\inf_{t\in[0,h]}\|\nabla f(x(t))\|^2$, notice that

$$\begin{aligned}\mathbb{E}\inf_{t\in[0,h]}\|\nabla f(x(t))\|^2 &\geq \frac{1}{2}\|\nabla f(x(0))\|^2 - \mathbb{E}\sup_{t\in[0,h]}\|\nabla f(x(t)) - \nabla f(x(0))\|^2 \\ &\geq \frac{1}{2}\|\nabla f(x(0))\|^2 - L^2\mathbb{E}\sup_{t\in[0,h]}\|x(t) - x(0)\|^2.\end{aligned}$$

Then, by$(A.5)$ and $h \leq \frac{1}{20}$,

$$-\mathbb{E}\inf_{t\in[0,h]}\|\nabla f(x(t))\| \;\leq\; -\frac{1}{3}\|\nabla f(x(0))\|^2 + O\left(h^2L^2\|v(0)\|^2 + Ldh^3\right).$$

□

A.2.2 Properties of the Brownian Motion

Lemma 7 (Doob's maximal inequality [Doo53])**.** *Suppose* $\{X(t) : t \geq 0\}$ *is a continuous martingale. Then, for any* $t \geq 0$*,*

$$\mathbb{E}\left[\sup_{0\leq s\leq t}|X(s)|^2\right] \;\leq\; 4\mathbb{E}\left[|X(t)|^2\right].$$

Using the Doob's maximal inequality, we can show the following lemma.

Lemma 8. *For d-dimensional Brownian motion* B_t *on* $t \in [0, h]$*, assuming* $h \leq \frac{1}{10}$*,*

$$\mathbb{E}\left[\sup_{0\leq t\leq h}\|B(t)\|^2\right] \leq 4dh, \text{ and } \mathbb{E}\left[\sup_{0\leq t\leq h}\left\|\int_0^t e^{-2(t-s)}dB_s\right\|^2\right] \leq 5dh.$$

Proof. To show the first inequality,

$$\begin{aligned}\mathbb{E}\left[\sup_{0\leq t\leq h}\|B(t)\|^2\right] &\leq \sum_{i=1}^{d}\mathbb{E}\left[\sup_{0\leq t\leq h}|B_i(t)|^2\right] \\ &\leq 4d\mathbb{E}\left[|B_i(h)|^2\right] \\ &= 4dh,\end{aligned}$$

where the second step follows by Lemma 7. To show the second inequality,

$$\begin{aligned}\mathbb{E}\left[\sup_{0\leq t\leq h}\left\|\int_0^t e^{-2(t-s)}dB_s\right\|^2\right] &\leq \mathbb{E}\left[\sup_{0\leq t\leq h}e^{-4t}\left\|\int_0^t e^{2s}dB_s\right\|^2\right] \\ &\leq \mathbb{E}\left[\sup_{0\leq t\leq h}\left\|\int_0^t e^{2s}dB_s\right\|^2\right]\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{d} \mathbb{E}\left[\sup_{0\leq t\leq h}\left|\int_0^t e^{2s} dB_{s,i}\right|^2\right] \\
&\leq 4\sum_{i=1}^{d} \mathbb{E}\left[\left|\int_0^h e^{2s} dB_{s,i}\right|^2\right] \\
&= 4\sum_{i=1}^{d} \int_0^h e^{4s} ds \\
&\leq 5dh,
\end{aligned}
$$

where the second step follows by $e^{-4t} \leq 1$, the fourth step follows by Lemma 7 and the last inequality follows by $\int_0^h e^{4s} ds \leq \frac{5}{4}h$ for $h \leq \frac{1}{10}$. □

A.3 Discretization Error of Algorithm 1

In this section, we bound the discretization error of Algorithm 1 in each iteration. In order to prove Lemma 2, we first prove Lemma 9, stated next.

Lemma 9. *Let α be the random number chosen in iteration n. Let $x_{n+\frac{1}{2}}$ be the intermediate value computed in iteration n of Algorithm 1. Let $\{x_n^*(t)\}_{t\in[0,h]}$ be the ideal underdamped Langevin diffusion starting from $x_n^*(0) = x_n$ coupled through a shared Brownian motion with $x_{n+\frac{1}{2}}$. Assume that $h \leq \frac{1}{20}$. Then,*

$$
\mathbb{E}\left\|\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h))\right\|^2 \leq O\left(h^6 L^2 \left\|v_n\right\|^2 + h^8 \left\|\nabla f(x_n)\right\|^2 + Ldh^7\right).
$$

Proof. We have the bound

$$
\begin{aligned}
&\mathbb{E}\left\|\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h))\right\|^2 \\
&\leq L^2 \mathbb{E}\left\|x_{n+\frac{1}{2}} - x_n^*(\alpha h)\right\|^2 \\
&= L^2 \mathbb{E}\left\|\frac{1}{2}u\int_0^{\alpha h}\left(1 - e^{-2(\alpha h - s)}\right)\left(\nabla f(x_n^*(0)) - \nabla f(x_n^*(s))\right) \mathrm{d}s\right\|^2 \\
&\leq \frac{1}{4}\mathbb{E}\left[\int_0^{\alpha h}\left(1 - e^{-2(\alpha h - s)}\right)^2 \mathrm{d}s \cdot \alpha h \cdot \left(\sup_{t\in[0,h]} \left\|\nabla f(x_n^*(0)) - \nabla f(x_n^*(t))\right\|^2\right)\right] \\
&\leq h^4 \mathbb{E}\sup_{t\in[0,h]} \left\|\nabla f(x_n^*(0)) - \nabla f(x_n^*(t))\right\|^2 \\
&\leq L^2 h^4 \mathbb{E}\sup_{t\in[0,h]} \left\|x_n^*(0) - x_n^*(t)\right\|^2 \\
&\leq O\left(h^6 L^2 \left\|v_n\right\|^2 + h^8 \left\|\nabla f(x_n)\right\|^2 + Ldh^7\right),
\end{aligned}
$$

where the first and the fifth step follows by ∇f is L-Lipschitz, the third step follows by Cauchy-Schwarz inequality, the fourth step follows by $1 - e^{-2(\alpha h - t)} \leq 2h$ and the last step follows by Lemma 6. □

Now, we are ready to prove Lemma 2.

Proof. To show the first claim,

$$\begin{aligned}
& \left\|\mathbb{E}_\alpha x_{n+1} - x_n^*(h)\right\|^2 \\
= & \left\|\mathbb{E}_\alpha \frac{1}{2}uh\left(1-e^{-2(h-\alpha h)}\right)\nabla f(x_{n+\frac{1}{2}}) - \frac{1}{2}u\int_0^h \left(1-e^{-2(h-s)}\right)\nabla f(x_n^*(s))\,\mathrm{d}s\right\|^2 \\
\leq & \frac{1}{2}\mathbb{E}_\alpha\left\|uh\left(1-e^{-2(h-\alpha h)}\right)\nabla f(x_{n+\frac{1}{2}}) - uh\left(1-e^{-2(h-\alpha h)}\right)\nabla f(x_n^*(\alpha h))\right\|^2 \\
& + \frac{1}{2}\left\|\mathbb{E}_\alpha uh\left(1-e^{-2(h-\alpha h)}\right)\nabla f(x_n^*(\alpha h)) - u\int_0^h \left(1-e^{-2(h-s)}\right)\nabla f(x_n^*(s))\,\mathrm{d}s\right\|^2 \\
\leq & \frac{1}{2}u^2h^2\mathbb{E}_\alpha\left[\left(1-e^{-2(h-\alpha h)}\right)^2\left\|\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h))\right\|^2\right] + 0 \\
\leq & 2u^2h^4\mathbb{E}_\alpha\left\|\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h))\right\|^2,
\end{aligned}$$

where the first step follows by the definition of x_{n+1}, the second step follows by Young's inequality, the third step follows by

$$\mathbb{E}_\alpha h\left(1-e^{-2(h-\alpha h)}\right)\nabla f(x_n^*(\alpha h)) = \int_0^h \left(1-e^{-2(h-s)}\right)\nabla f(x_n^*(s))\,\mathrm{d}s,$$

and the fourth step follows by $1-e^{-2(h-\alpha h)} \leq 2h$. By Lemma 9,

$$\mathbb{E}\left\|\mathbb{E}_\alpha x_{n+1} - x_n^*(h)\right\|^2 \leq O\left(h^{10}\left\|v_n\right\|^2 + u^2h^{12}\left\|\nabla f(x_n)\right\|^2 + udh^{11}\right).$$

To show the second claim,

$$\begin{aligned}
& \mathbb{E}\left\|x_{n+1} - x_n^*(h)\right\|^2 \\
\leq & \frac{3}{4}\mathbb{E}\left\|uh\left(1-e^{-2(h-\alpha h)}\right)\nabla f(x_{n+\frac{1}{2}}) - uh\left(1-e^{-2(h-\alpha h)}\right)\nabla f(x_n^*(\alpha h))\right\|^2 \\
& + \frac{3}{4}\mathbb{E}\left\|uh\left(1-e^{-2(h-\alpha h)}\right)\nabla f(x_n^*(\alpha h)) - u\int_0^h \left(1-e^{-2(h-\alpha h)}\right)\nabla f(x_n^*(s))\,\mathrm{d}s\right\|^2 \\
& + \frac{3}{4}\mathbb{E}\left\|u\int_0^h \left(1-e^{-2(h-\alpha h)}\right)\nabla f(x_n^*(s))\,\mathrm{d}s - u\int_0^h \left(1-e^{-2(h-s)}\right)\nabla f(x_n^*(s))\,\mathrm{d}s\right\|^2,
\end{aligned}$$

which follows by definition and Young's inequality. To bound the second term,

$$\begin{aligned}
& \left\|uh\left(1-e^{-2(h-\alpha h)}\right)\nabla f(x_n^*(\alpha h)) - u\int_0^h \left(1-e^{-2(h-\alpha h)}\right)\nabla f(x_n^*(s))\,\mathrm{d}s\right\|^2 \\
= & \left\|u\int_0^h \left(1-e^{-2(h-\alpha h)}\right)\left(\nabla f(x_n^*(\alpha h)) - \nabla f(x_n^*(s))\right)\mathrm{d}s\right\|^2 \\
\leq & u^2\int_0^h \left(1-e^{-2(h-\alpha h)}\right)^2\mathrm{d}s \cdot \sup_{t\in[0,h]}\left\|\nabla f(x_n^*(\alpha h)) - \nabla f(x_n^*(t))\right\|^2 \cdot h \\
\leq & 4u^2h^4 \sup_{t\in[0,h]}\left\|\nabla f(x_n^*(\alpha h)) - \nabla f(x_n^*(t))\right\|^2
\end{aligned}$$

$$\leq \quad 16h^4 \sup_{t\in[0,h]} \|x_n^*(0) - x_n^*(t)\|^2 \tag{A.6}$$

where the second step follows by the Cauchy-Schwarz inequality. The third term satisfies

$$\begin{aligned}
&\left\| u\int_0^h \left(1 - e^{-2(h-\alpha h)}\right) \nabla f(x_n^*(s))\,\mathrm{d}s - u\int_0^h \left(1 - e^{-2(h-s)}\right) \nabla f(x_n^*(s))\,\mathrm{d}s \right\|^2 \\
= \quad & u^2 \left\| \int_0^h \left(e^{-2(h-s)} - e^{-2(h-\alpha h)}\right) \nabla f(x_n^*(s))\,\mathrm{d}s \right\|^2 \\
\leq \quad & 4u^2h^4 \sup_{t\in[0,h]} \|\nabla f(x_n^*(t))\|^2, \qquad (A.7)
\end{aligned}$$

where the second step follows by the Cauchy Schwarz inequality and $\left|e^{-2(h-s)} - e^{-2(h-\alpha h)}\right| \leq 2h$. Thus,

$$\begin{aligned}
& \mathbb{E}\,\|x_{n+1} - x_n^*(h)\|^2 \\
\leq \quad & 3u^2h^4\mathbb{E}\left\|\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h))\right\|^2 + 12h^4\mathbb{E}\sup_{t\in[0,h]} \|x_n^*(0) - x_n^*(t)\|^2 \\
& +3u^2h^4\mathbb{E}\sup_{t\in[0,h]} \|\nabla f(x_n^*(t))\|^2 \\
\leq \quad & 3h^4 \cdot O\left(h^6\|v_n\|^2 + h^8u^2\|\nabla f(x_n)\|^2 + udh^7\right) \\
& +12h^4 \cdot O\left(h^2\|v_n\|^2 + u^2h^4\|\nabla f(x_n)\|^2 + udh^3\right) \\
& +3u^2h^4 \cdot O\left(\|\nabla f(x_n)\|^2 + L^2h^2\|v_n\|^2 + Mdh^3\right) \\
\leq \quad & O\left(h^6\|v_n\|^2 + u^2h^4\|\nabla f(x_n)\|^2 + udh^7\right).
\end{aligned}$$

where the first step follows by (A.6) and (A.7), the second step follows by Lemma 6 and Lemma 9, and the last inequality follows by $h \leq 1$.

To show the third claim,

$$\begin{aligned}
\mathbb{E}\,\|\mathbb{E}_\alpha v_{n+1} - v_n^*(h)\|^2 \quad &= \quad \mathbb{E}\left\|\mathbb{E}_\alpha uhe^{-2(h-\alpha h)}\nabla f(x_{n+\frac{1}{2}}) - u\int_0^h e^{-2(h-s)}\nabla f(x_n^*(s))\,\mathrm{d}s\right\|^2 \\
&\leq \quad 2\mathbb{E}\left\|uhe^{-2(h-\alpha h)}\nabla f(x_{n+\frac{1}{2}}) - uhe^{-2(h-\alpha h)}\nabla f(x_n^*(\alpha h))\right\|^2 \\
&\quad +2\mathbb{E}\left\|\mathbb{E}_\alpha uhe^{-2(h-\alpha h)}\nabla f(x_n^*(\alpha h)) - u\int_0^h e^{-2(h-s)}\nabla f(x_n^*(s))\,\mathrm{d}s\right\|^2 \\
&\leq \quad 2u^2h^2\mathbb{E}\left\|\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h))\right\|^2 + 0 \\
&\leq \quad O\left(h^8\|v_n\|^2 + u^2h^{10}\|\nabla f(x_n)\|^2 + udh^9\right),
\end{aligned}$$

where the first step follows by Young's inequality, the second step follows by

$$\mathbb{E}_\alpha uhe^{-2(h-\alpha h)}\nabla f(x_n^*(\alpha h)) \quad = \quad u\int_0^h e^{-2(h-t)}\nabla f(x_n^*(t))\,\mathrm{d}t,$$

and $e^{-2(h-\alpha h)} \leq 1$, and the third step follows by Lemma 9.

To show the last claim,

$$\begin{aligned}
& \mathbb{E}\left\|v_{n+1}-v_n^*(h)\right\|^2 \\
= \;& \mathbb{E}\left\|uhe^{-2(h-\alpha h)}\nabla f(x_{n+\frac{1}{2}})-u\int_0^h e^{-2(h-s)}\nabla f(x^*(s))\,\mathrm{d}s\right\|^2 \\
\leq \;& 3\mathbb{E}\left\|uhe^{-2(h-\alpha h)}\nabla f(x_{n+\frac{1}{2}})-uhe^{-2(h-\alpha h)}\nabla f(x^*(\alpha h))\right\|^2 \\
& +3\mathbb{E}\left\|u\int_0^h e^{-2(h-\alpha h)}\nabla f(x_n^*(\alpha h))\,\mathrm{d}t-u\int_0^h e^{-2(h-\alpha h)}\nabla f(x_n^*(s))\,\mathrm{d}s\right\|^2 \\
& +3\mathbb{E}\left\|u\int_0^h e^{-2(h-\alpha h)}\nabla f(x_n^*(s))\,\mathrm{d}s-u\int_0^h e^{-2(h-s)}\nabla f(x_n^*(s))\,\mathrm{d}s\right\|^2 \\
\leq \;& 3u^2h^2\mathbb{E}\left\|\nabla f(x_{n+\frac{1}{2}})-\nabla f(x_n^*(\alpha h))\right\|^2+3h^2\mathbb{E}\sup_{t\in[0,h]}\left\|x_n^*(\alpha h)-x_n^*(t)\right\|^2 \\
& +12u^2h^4\mathbb{E}\sup_{t\in[0,h]}\left\|\nabla f(x_n^*(t))\right\|^2 \\
\leq \;& 3u^2h^2\cdot O\left(h^6L^2\left\|v_n\right\|^2+h^8\left\|\nabla f(x_n)\right\|^2+Ldh^7\right) \\
& +3h^2\cdot O\left(h^2\left\|v_n\right\|^2+u^2h^4\left\|\nabla f(x_n)\right\|^2+udh^3\right) \\
& +12u^2h^4\cdot O\left(\left\|\nabla f(x_n)\right\|^2+L^2h^2\left\|v_n\right\|^2+Ldh^3\right) \\
\leq \;& O\left(h^4\left\|v_n\right\|^2+u^2h^4\left\|\nabla f(x_n)\right\|^2+udh^5\right),
\end{aligned}$$

where the first step follows by the definition, the second step follows by Young's inequality, the third follows by $e^{-2(h-\alpha h)}-e^{-2(h-s)}\leq 2h$, the fourth step follows by Lemma 9 and Lemma 6 and the last inequality follows by $h\leq 1$. □

A.4 Bounds on $\|\nabla f(x)\|$ and $\|v\|$

In this section, we bound the sum of $\|\nabla f(x_n)\|^2$ and $\|v_n\|^2$ over all iterations n, $\sum_{n=0}^{N-1}\mathbb{E}\|\nabla f(x_n)\|^2$ and $\sum_{n=0}^{N-1}\mathbb{E}\|v_n\|^2$. In Appendix A.5, we use the results in this appendix together with Lemma 2 to prove the guarantee of our algorithm.

Lemma 10. *Assume $h\leq\frac{1}{20}$. For each iteration n, let x_n be the starting point of iteration n of Algorithm 1. Let $\{v_n(t),x_n(t)\}_{t\in[0,h]}$ be the solution of the exact underdamped Langevin diffusion starting from (v_n,x_n) . Let $\mathbb{E}_\alpha$ be the expectation over the random choice of α in iteration n. Then, the difference between the value of f on the starting point of iteration $n+1$, x_{n+1}, and that of $x_n(h)$ satisfies*

$$\mathbb{E}f(x_{n+1}(0))-f(x_n(h)) \;\leq\; O\left(uh^3\left\|\nabla f(x_n(0))\right\|^2+Lh^5\left\|v_n(0)\right\|^2+dh^6\right).$$

Proof. We first consider the expectation over the choice of α in iteration n,

$$\mathbb{E}_\alpha f(x_{n+1}(0))$$

$$
\begin{aligned}
&\leq f(x_n(h)) + \nabla f(x_n(h))^T \left(\mathbb{E}_\alpha x_{n+1}(0) - x_n(h)\right) + \frac{L}{2}\mathbb{E}_\alpha \left\|x_{n+1}(0) - x_n(h)\right\|^2 \\
&\leq f(x_n(h)) + \left\|\nabla f(x_n(h))\right\| \left\|\mathbb{E}_\alpha x_{n+1}(0) - x_n(h)\right\| + \frac{L}{2}\mathbb{E}_\alpha \left\|x_{n+1}(0) - x_n(h)\right\|^2 \\
&\leq f(x_n(h)) + uh^3 \left\|\nabla f(x_n(h))\right\|^2 + \frac{L}{h^3}\left\|\mathbb{E}_\alpha x_{n+1}(0) - x_n(h)\right\|^2 + \frac{L}{2}\mathbb{E}_\alpha \left\|x_{n+1}(0) - x_n(h)\right\|^2,
\end{aligned}
$$

where the first step follows by ∇f is L-Lipschitz, the second step follows by Cauchy-Schwarz inequality and the third step follows by Young's inequality. By Lemma 2 and Lemma 6,

$$
\begin{aligned}
\mathbb{E}f(x_{n+1}(0)) &\leq \mathbb{E}f(x_n(h)) + uh^3\mathbb{E}\left\|\nabla f(x_n(h))\right\|^2 + \frac{L}{h^3}\mathbb{E}\left\|\mathbb{E}_\alpha x_{n+1}(0) - x_n(h)\right\|^2 \\
&\quad + \frac{L}{2}\mathbb{E}\left\|x_{n+1}(0) - x_n(h)\right\|^2 \\
&\leq \mathbb{E}f(x_n(h)) + \mathbb{E}uh^3 \cdot O\left(\left\|\nabla f(x_n(0))\right\|^2 + L^2h^2\left\|v_n(0)\right\|^2 + Ldh^3\right) \\
&\quad + \mathbb{E}\frac{L}{h^3} \cdot O\left(h^{10}\left\|v_n(0)\right\|^2 + u^2h^{12}\left\|\nabla f(x_n(0))\right\|^2 + udh^{11}\right) \\
&\quad + \mathbb{E}\frac{L}{2} \cdot O\left(h^6\left\|v_n(0)\right\|^2 + h^4u^2\left\|\nabla f(x_n(0))\right\|^2 + udh^7\right) \\
&\leq \mathbb{E}f(x_n(h)) + O\left(uh^3\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + Lh^5\mathbb{E}\left\|v_n(0)\right\|^2 + dh^6\right).
\end{aligned}
$$

where the second step follows by Lemma 2 and Lemma 6, and the last step follows by $h \leq \frac{1}{20}$. □

Lemma 11. *Assume h is smaller than some given constant. For each iteration $n = 0, ..., N-1$, let (v_n, x_n) be the starting point of Algorithm 1 in iteration n. Then,*

$$
\sum_{n=0}^{N-1}\mathbb{E}\left\|v_n\right\|^2 \leq O\left(u^2h\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x_n)\right\|^2 + Nud\right).
$$

Proof. Let $\{v_n(t), x_n(t)\}_{t\in[0,h]}$ be the solution of the exact underdamped Langevin diffusion starting from (v_n, x_n) . By definition, for $t \in [0, h]$,

$$
\begin{aligned}
\frac{\mathrm{d}f(x_n(t))}{\mathrm{d}t} &= \nabla f(x_n(t))^T\frac{\mathrm{d}x_n(t)}{\mathrm{d}t} \\
&= \nabla f(x_n(t))^T v_n(t),
\end{aligned}
$$

so

$$
\begin{aligned}
f(x_n(h)) &= f(x_n(0)) + \int_0^h \mathrm{d}f(x_n(t)) \\
&= f(x_n(0)) + \int_0^h \nabla f(x_n(t))^T v_n(t)\,\mathrm{d}t. \qquad \text{(A.8)}
\end{aligned}
$$

Also, since

$$
\mathrm{d}v_n(t) = \left(-2v_n(t) - u\nabla f(x_n(t))\right)\mathrm{d}t + 2\sqrt{u}\,\mathrm{d}B_t,
$$

by Ito's lemma,

$$\begin{aligned}\mathrm{d}\frac{1}{2}\left\|v_n(t)\right\|^2 &= \left\langle v_n(t), 2\sqrt{u}\,\mathrm{d}B_t\right\rangle + \left(\left\langle v_n(t), -2v_n(t) - u\nabla f(x_n(t))\right\rangle + \frac{1}{2}\cdot 4u\mathrm{Tr}(I_d)\right)\mathrm{d}t \\ &= 2\sqrt{u}v_n(t)^T\,\mathrm{d}B_t + \left(-2\left\|v_n(t)\right\|^2 - uv_n(t)^T\nabla f(x_n(t)) + 2ud\right)\mathrm{d}t,\end{aligned}$$

and therefore

$$\mathbb{E}\frac{1}{2u}\left\|v_n(h)\right\|^2 = \mathbb{E}\frac{1}{2u}\left\|v_n(0)\right\|^2 + \mathbb{E}\int_0^h \left(4d - \frac{2}{u}\left\|v_n(t)\right\|^2 - v_n(t)^T\nabla f(x_n(t)) + 2d\right)\mathrm{d}t. \tag{A.9}$$

Now, we consider the term $\frac{1}{2u}\left\|v_n(h)\right\|^2 + f(x_n(h))$. By $(A.8)$and $(A.9)$,

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{2u}\left\|v_n(h)\right\|^2 + f(x_n(h))\right] \\ =\;& \mathbb{E}\left[\frac{1}{2u}\left\|v_n(0)\right\|^2 + f(x_n(0)) + \int_0^h\left(-\frac{2}{u}\left\|v_n(t)\right\|^2 + 6d\right)\mathrm{d}t\right] \\ \leq\;& \mathbb{E}\left[\frac{1}{2u}\left\|v_n(0)\right\|^2 + f(x_n(0)) - \frac{2}{u}h\inf_{t\in[0,h]}\left\|v_n(t)\right\|^2 + 6dh\right] \\ \leq\;& \mathbb{E}\left[\frac{1}{2u}\left\|v_n(0)\right\|^2 + f(x_n(0))\right] - \frac{2}{3}hL\mathbb{E}\left\|v_n(0)\right\|^2 + O\left(uh^3\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + dh\right),\end{aligned}$$

where the first step follows by $(A.8)$ and $(A.9)$ and the third step follows by Lemma 6.

Since

$$\begin{aligned}&\mathbb{E}\left[\left\|v_{n+1}(0)\right\|^2 - \left\|v_n(h)\right\|^2\right] \\ =\;& \mathbb{E}\left(v_{n+1}(0) - v_n(h)\right)^T\left(v_{n+1}(0) + v_n(h)\right) \\ \leq\;& \frac{1}{h^2}\mathbb{E}\left\|v_{n+1}(0) - v_n(h)\right\|^2 + \frac{1}{2}h^2\mathbb{E}\left\|v_{n+1}(0) + v_n(h)\right\|^2 \\ \leq\;& \frac{1}{h^2}\mathbb{E}\left\|v_{n+1}(0) - v_n(h)\right\|^2 + h^2\mathbb{E}\left\|v_{n+1}(0) - v_n(h)\right\|^2 + 4h^2\mathbb{E}\left\|v_n(h)\right\|^2 \\ \leq\;& \frac{2}{h^2}\mathbb{E}\left\|v_{n+1}(0) - v_n(h)\right\|^2 + 4h^2\mathbb{E}\left\|v_n(h)\right\|^2 \\ \leq\;& O\left(h^2\mathbb{E}\left\|v_n(0)\right\|^2 + u^2h^2\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + udh^3\right),\end{aligned}$$

where the first inequality follows by the inequality $2ab \leq a^2 + b^2$, the second inequality follows by Young's inequality and the last inequality follows by Lemma 2 and Lemma 6.

Since

$$\mathbb{E}f(x_{n+1}(0)) - f(x_n(h)) \;\leq\; O\left(uh^3\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + Lh^5\mathbb{E}\left\|v_n(0)\right\|^2 + dh^6\right),$$

which is shown in Lemma 10, we have

$$\mathbb{E}\left[\frac{1}{2u}\left\|v_{n+1}(0)\right\|^2 + f(x_{n+1}(0))\right]$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\frac{1}{2u}\left\|v_n(0)\right\|^2 + f(x_n(0))\right] - \frac{2}{3}hL\mathbb{E}\left\|v_n(0)\right\|^2 + O\left(uh^3\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + dh\right) \\
&\quad + O\left(h^2L\mathbb{E}\left\|v_n(0)\right\|^2 + uh^2\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + dh^3\right) \\
&\quad + O\left(uh^3\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + Lh^5\mathbb{E}\left\|v_n(0)\right\|^2 + dh^6\right) \\
&\leq \mathbb{E}\left[\frac{1}{2u}\left\|v_n(0)\right\|^2 + f(x_n(0))\right] - \frac{1}{3}hL\mathbb{E}\left\|v_n(0)\right\|^2 + O\left(uh^2\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + hd\right),
\end{aligned}
$$

where the last step follows by h is small. Summing n from 0 to $N-1$, we get

$$
\begin{aligned}
&\sum_{n=0}^{N-1}\mathbb{E}\left[\frac{1}{2u}\left\|v_{n+1}(0)\right\|^2 + f(x_{n+1}(0))\right] \\
\leq\ &\sum_{n=0}^{N-1}\mathbb{E}\left[\frac{1}{2u}\left\|v_n(0)\right\|^2 + f(x_n(0))\right] - \frac{1}{3}hL\sum_{n=0}^{N-1}\mathbb{E}\left\|v_n(0)\right\|^2 \\
&+ O\left(uh^2\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + Nhd\right).
\end{aligned}
$$

Since $\|v_0(0)\| = 0$ and $f(x_0(0)) \leq f(x_N(0))$,

$$
\frac{1}{3}hL\sum_{n=0}^{N-1}\mathbb{E}\left\|v_n(0)\right\|^2 \leq O\left(uh^2\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + Nhd\right),
$$

which implies

$$
\sum_{n=0}^{N-1}\mathbb{E}\left\|v_n(0)\right\|^2 \leq O\left(u^2h\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + Nud\right).
$$

□

Lemma 12. *Assume h is smaller than some given constant. For each iteration $n = 0, ..., N-1$, let (v_n, x_n) be the starting point of Algorithm 1 in iteration n. Then, the x_n in iteration $n = 0, ..., N-1$ satisfies*

$$
\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x_n)\right\|^2 \leq O\left(NLd + \frac{L}{h}\left|\mathbb{E}\nabla f(x_N)^T v_N\right|\right).
$$

Furthermore, the v_n in iteration $n = 0, ..., N-1$ satisfies

$$
\sum_{n=0}^{N-1}\mathbb{E}\left\|v_n\right\|^2 \leq O\left(Nud + u\left|\mathbb{E}\nabla f(x_N)^T v_N\right|\right).
$$

Proof. For each iteration $n = 0, ..., N-1$, let $\{v_n(t), x_n(t)\}_{t\in[0,h]}$ be the exact underdamped Langevin diffusion starting from (v_n, x_n) computed in Algorithm 1. By definition,

$$
\mathbb{E}\left[\mathrm{d}\nabla f(x_n(t))^T v_n(t)\right]
$$

$$
\begin{aligned}
&= \mathbb{E}\left[v_n(t)^T\nabla^2 f(x_n(t))v_n(t) + \nabla f(x_n(t))^T \,\mathrm{d}v_n(t)\right] \\
&= \mathbb{E}\left[v_n(t)^T\nabla^2 f(x_n(t))v_n(t) - 2\nabla f(x_n(t))^T v_n(t) - u\left\|\nabla f(x_n(t))\right\|^2\right].
\end{aligned}
$$

So we have

$$
\begin{aligned}
&\mathbb{E}\left[\nabla f(x_n(h))^T v_n(h)\right] \\
=\ &\mathbb{E}\left[\nabla f(x_n(0))^T v_n(0) + \int_0^h \mathrm{d}\nabla f(x_n(t))^T v_n(t)\right] \\
=\ &\mathbb{E}\Bigg[\nabla f(x_n(0))^T v_n(0) + \int_0^h v_n(t)^T\nabla^2 f(x_n(t))v_n(t) - 2\nabla f(x_n(t))^T v_n(t) \\
&-u\left\|\nabla f(x_n(t))\right\|^2 \mathrm{d}t\Bigg] \\
\le\ &\mathbb{E}\left[\nabla f(x_n(0))^T v_n(0) + 3L\int_0^h \left\|v_n(t)\right\|^2 \mathrm{d}t - \frac{1}{2}\int_0^h u\left\|\nabla f(x_n(t))\right\|^2 \mathrm{d}t\right] \\
\le\ &\mathbb{E}\left[\nabla f(x_n(0))^T v_n(0) + 3Lh\sup_{t\in[0,h]}\left\|v_n(t)\right\|^2 - \frac{1}{2}hu\inf_{t\in[0,h]}\left\|\nabla f(x_n(t))\right\|^2\right] \\
\le\ &\mathbb{E}\nabla f(x_n(0))^T v_n(0) - \frac{1}{6}hu\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + O\left(h^3L\mathbb{E}\left\|v_n(0)\right\|^2 + dh^4\right) \\
&+3Lh\cdot O\left(\mathbb{E}\left\|v_n(0)\right\|^2 + u^2h^2\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + udh\right) \\
\le\ &\mathbb{E}\nabla f(x_n(0))^T v_n(0) - \frac{1}{6}hu\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 \\
&+O\left(Lh\mathbb{E}\left\|v_n(0)\right\|^2 + uh^3\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + dh^2\right), \qquad \text{(A.10)}
\end{aligned}
$$

where the third step follows by Young's inequality, the fifth step follows by Lemma 6 and the last step follows by h is small. Also, we have

$$
\begin{aligned}
&\mathbb{E}\left[\nabla f(x_{n+1}(0))^T v_{n+1}(0) - \nabla f(x_n(h))^T v_n(h)\right] \\
=\ &\mathbb{E}\left(\nabla f(x_{n+1}(0)) - \nabla f(x_n(h)) + \nabla f(x_n(h))\right)^T\left(v_{n+1}(0) - v_n(h)\right) \\
&+\mathbb{E}\left(\nabla f(x_{n+1}(0) - \nabla f(x_n(h))\right)^T v_n(h) \\
\le\ &u\mathbb{E}\left\|\nabla f(x_{n+1}(0)) - \nabla f(x_n(h))\right\|^2 + L\mathbb{E}\left\|v_{n+1}(0) - v_n(h)\right\|^2 + uh^2\mathbb{E}\left\|\nabla f(x_n(h))\right\|^2 \\
&+\frac{L}{h^2}\mathbb{E}\left\|v_{n+1}(0) - v_n(h)\right\|^2 + \frac{u}{h}\mathbb{E}\left\|\nabla f(x_{n+1}(0)) - \nabla f(x_n(h))\right\|^2 + hL\mathbb{E}\left\|v_n(h)\right\|^2 \\
\le\ &\frac{2u}{h}\mathbb{E}\left\|\nabla f(x_{n+1}(0)) - \nabla f(x_n(h))\right\|^2 + \frac{2L}{h^2}\mathbb{E}\left\|v_{n+1}(0) - v_n(h)\right\|^2 + uh^2\mathbb{E}\left\|\nabla f(x_n(h))\right\|^2 \\
&+hL\mathbb{E}\left\|v_n(h)\right\|^2 \\
\le\ &\frac{2L}{h}\cdot O\left(h^6\mathbb{E}\left\|v_n(0)\right\|^2 + h^4u^2\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + udh^7\right) \\
&+\frac{2L}{h^2}\cdot O\left(h^4\mathbb{E}\left\|v_n(0)\right\|^2 + u^2h^4\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + udh^5\right) \\
&+uh^2\cdot O\left(\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 + L^2h^2\mathbb{E}\left\|v_n(0)\right\|^2 + Ldh^3\right)
\end{aligned}
$$

$$
\begin{aligned}
&+hL\cdot O\left(\mathbb{E}\left\|v_n(0)\right\|^2+u^2h^2\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2+udh\right)\\
\leq\;& O\left(hL\mathbb{E}\left\|v_n(0)\right\|^2+uh^2\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2+dh^2\right), \qquad (A.11)
\end{aligned}
$$

where the second step follows by Young's inequality and the fourth step follows by Lemma 2 and Lemma 6. Combining $(A.10)$ and $(A.11)$,

$$
\begin{aligned}
\mathbb{E}\nabla f(x_{n+1}(0))^Tv_{n+1}(0) &\leq \mathbb{E}\nabla f(x_n(0))^Tv_n(0)-\frac{1}{6}hu\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2\\
&\quad+O\left(Lh\mathbb{E}\left\|v_n(0)\right\|^2+uh^3\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2+dh^2\right)\\
&\quad+O\left(Lh\mathbb{E}\left\|v_n(0)\right\|^2+uh^2\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2+dh^2\right)\\
&\leq \mathbb{E}\nabla f(x_n(0))^Tv_n(0)-\frac{1}{6}hu\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2\\
&\quad+O\left(Lh\mathbb{E}\left\|v_n(0)\right\|^2+uh^2\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2+dh^2\right).
\end{aligned}
$$

Summing from $n=0$ to $N-1$,

$$
\begin{aligned}
\sum_{n=0}^{N-1}\mathbb{E}\nabla f(x_{n+1}(0))^Tv_{n+1}(0) &\leq \sum_{n=0}^{N-1}\mathbb{E}\nabla f(x_n(0))^Tv_n(0)-\frac{1}{6}hu\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2\\
&\quad+O\left(Lh\sum_{n=0}^{N-1}\mathbb{E}\left\|v_n(0)\right\|^2+uh^2\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2+Ndh^2\right)\\
&\leq \sum_{n=0}^{N-1}\mathbb{E}\nabla f(x_n(0))^Tv_n(0)-\frac{1}{6}hu\sum_{n=0}^{N-1}\left\|\nabla f(x_n(0))\right\|^2\\
&\quad+O\left(Lh\left(u^2h\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2+Nud\right)+Ndh^2\right)\\
&\leq \sum_{n=0}^{N-1}\mathbb{E}\nabla f(x_n(0))^Tv_n(0)-\frac{1}{8}hu\sum_{n=0}^{N-1}\left\|\nabla f(x_n(0))\right\|^2+O\left(Ndh\right),
\end{aligned}
$$

where the second step follows by Lemma 11 and the last step follows by h is small. Then, since $v_0=0$,

$$
\frac{1}{8}hu\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 \leq O\left(Ndh+\left|\mathbb{E}\nabla f(x_N(0))^Tv_N(0)\right|\right),
$$

which implies

$$
\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2 \leq O\left(NLd+\frac{L}{h}\left|\mathbb{E}\nabla f(x_N(0))^Tv_N(0)\right|\right).
$$

By Lemma 11,

$$
\begin{aligned}
\sum_{n=0}^{N-1}\mathbb{E}\left\|v_n(0)\right\|^2 &\leq O\left(u^2h\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x_n(0))\right\|^2+Nud\right)\\
&\leq O\left(Nud+u\left|\mathbb{E}\nabla f(x_N(0))^Tv_N(0)\right|\right).
\end{aligned}
$$

□

A.5 Proof of Theorem 3

Here, we combine Lemma 12 and Lemma 2 to prove our main result.

Proof. Let $x_{n+\frac{1}{2}}$, x_n and v_n be the iterates of Algorithm 1. Let (y_n, w_n) be the n-th step of the exact underdamped Langevin diffusion, starting from a random point $(y_0, w_0) \propto \exp\left(-\left(f(y) + \frac{L}{2}\|w\|^2\right)\right)$, coupled with (x_n, v_n) through the same Brownian motion. Let $\left(x^*_{n+1}, v^*_{n+1}\right)$ be the 1-step exact Langevin diffusion starting from (x_n, v_n) . For any iteration n, let $\mathbb{E}_\alpha$ be the expectation taken over the random choice of α in iteration n. Then,

$$
\begin{aligned}
& \mathbb{E}_\alpha\left[\|x_n - y_n\|^2 + \|(x_n + v_n) - (y_n + w_n)\|^2\right] \\
= \;& \mathbb{E}_\alpha\left[\|(x_n - x^*_n) - (y_n - x^*_n)\|^2 + \|(x_n + v_n - x^*_n - v^*_n) - (y_n + w_n - x^*_n - v^*_n)\|^2\right] \\
\leq \;& \|y_n - x^*_n\|^2 + \|y_n + w_n - x^*_n - v^*_n\|^2 + \mathbb{E}_\alpha\|x_n - x^*_n\|^2 + \mathbb{E}_\alpha\|x_n + v_n - x^*_n - v^*_n\|^2 \\
& -2\left(y_n - x^*_n\right)^T\left(\mathbb{E}_\alpha x_n - x^*_n\right) - 2\left(y_n + w_n - x^*_n - v^*_n\right)^T\left(\mathbb{E}_\alpha\left[x_n + v_n\right] - x^*_n - v^*_n\right) \\
\leq \;& \left(1 + \frac{h}{2\kappa}\right)\left(\|y_n - x^*_n\|^2 + \|y_n + w_n - x^*_n - v^*_n\|^2\right) \\
& + \frac{2\kappa}{h}\left(\|\mathbb{E}_\alpha x_n - x^*_n\|^2 + \|\mathbb{E}_\alpha\left[x_n + v_n\right] - x^*_n - v^*_n\|^2\right) + \mathbb{E}_\alpha\|x_n - x^*_n\|^2 \\
& + \mathbb{E}_\alpha\|x_n + v_n - x^*_n - v^*_n\|^2,
\end{aligned}
$$

where the second step follows by y_n, w_n, x^*_n and v^*_n are independent of the choice of α and the third follows by Young's inequality. Then,

$$
\begin{aligned}
& \mathbb{E}\left[\|x_N - y_N\|^2 + \|(x_N + v_N) - (y_N + w_N)\|^2\right] \\
\leq \;& \left(1 + \frac{h}{2\kappa}\right)e^{-\frac{h}{\kappa}}\mathbb{E}\left[\|y_{N-1} - x_{N-1}\|^2 + \|y_{N-1} + w_{N-1} - x_{N-1} - v_{N-1}\|^2\right] \\
& + \frac{2\kappa}{h}\left(\mathbb{E}\|\mathbb{E}_\alpha x_N - x^*_N\|^2 + \mathbb{E}\|\mathbb{E}_\alpha x_N + v_N - x^*_N - v^*_N\|^2\right) \\
& + \left(\mathbb{E}\|x_N - x^*_N\|^2 + \mathbb{E}\|x_N + v_N - x^*_N - v^*_N\|^2\right) \\
\leq \;& e^{-\frac{h}{2\kappa}}\mathbb{E}\left[\|y_{N-1} - x_{N-1}\|^2 + \|y_{N-1} + w_{N-1} - x_{N-1} - v_{N-1}\|^2\right] \\
& + \frac{2\kappa}{h}\left(2\mathbb{E}\|\mathbb{E}_\alpha v_N - v^*_N\|^2 + 3\mathbb{E}\|\mathbb{E}_\alpha x_N - x^*_N\|^2\right) + \left(2\mathbb{E}\|v_N - v^*_N\|^2 + 3\mathbb{E}\|x_N - x^*_N\|^2\right) \\
\leq \;& e^{-\frac{Nh}{2\kappa}}\mathbb{E}\left[\|y_0 - x_0\|^2 + \|y_0 + w_0 - x_0 - v_0\|^2\right] \\
& + \sum_{n=1}^{N}\frac{2\kappa}{h}\left(2\mathbb{E}\|\mathbb{E}_\alpha v_n - v^*_n\|^2 + 3\mathbb{E}\|\mathbb{E}_\alpha x_n - x^*_n\|^2\right) \\
& + \sum_{n=1}^{N}\left(2\mathbb{E}\|v_n - v^*_n\|^2 + 3\mathbb{E}\|x_n - x^*_n\|^2\right),
\end{aligned}
$$

where the first step follows by Lemma 1, the second step follows by $1+\frac{h}{2\kappa} \leq e^{\frac{h}{2\kappa}}$, and the last step follows by induction.

Since (y_N, w_N) follows the distribution $p^* \propto \exp\left(-\left(f(y)+\frac{L}{2}\|w\|^2\right)\right)$, $\mathbb{E}\|w_N\|^2 = \frac{d}{L}$. By Proposition 1 of [DM16], $\mathbb{E}\|y_0 - x_0\|^2 \leq \frac{d}{m}$. Then,

$$\begin{aligned}\mathbb{E}\left[\|y_0 - x_0\|^2 + \|y_0 + w_0 - x_0 - v_0\|^2\right] &\leq 3\mathbb{E}\|y_0 - x_0\|^2 + 2\mathbb{E}\|w_0 - v_0\|^2 \\ &\leq 5\frac{d}{m}.\end{aligned}$$

When $N = \frac{2\kappa}{h}\log\left(\frac{20}{\epsilon^2}\right)$,

$$e^{-\frac{Nh}{2\kappa}}\mathbb{E}\left[\|y_0 - x_0\|^2 + \|y_0 + w_0 - x_0 - v_0\|^2\right] \quad \leq \quad \frac{\epsilon^2 d}{4m}.$$

By Lemma 2,

$$\begin{aligned}&\sum_{n=1}^{N}\frac{2\kappa}{h}\left(2\mathbb{E}\|\mathbb{E}_\alpha v_n - v_n^*\|^2 + 3\mathbb{E}\|\mathbb{E}_\alpha x_n - x_n^*\|^2\right) \\ \leq\ & O\left(h^7\kappa\sum_{n=0}^{N-1}\mathbb{E}\|v_n\|^2 + \frac{u}{m}h^9\sum_{n=0}^{N-1}\mathbb{E}\|\nabla f(x_n)\|^2 + \frac{1}{m}Ndh^8\right),\end{aligned}$$

and

$$\begin{aligned}&\sum_{n=1}^{N}\left(2\mathbb{E}\|v_n - v_n^*\|^2 + 3\mathbb{E}\|x_n - x_n^*\|^2\right) \\ \leq\ & O\left(h^4\sum_{n=0}^{N-1}\mathbb{E}\|v_n\|^2 + u^2h^4\sum_{n=0}^{N-1}\mathbb{E}\|\nabla f(x_n)\|^2 + Nudh^5\right).\end{aligned}$$

By Lemma 2 of [Dal17a], $\mathbb{E}\|\nabla f(y_N)\|^2 \leq dL$. Then, by $\mathbb{E}\|\nabla f(y_N)\|^2 \leq dL$ and $\mathbb{E}\|w_N\|^2 = \frac{d}{L}$,

$$\begin{aligned}\left|\mathbb{E}\nabla f(x_N)^T v_N\right| &\leq \mathbb{E}\left[L\|v_N\|^2 + u\|\nabla f(x_N)\|^2\right] \\ &\leq 2\mathbb{E}\left[L\|w_N\|^2 + L\|v_N - w_N\|^2 + u\|\nabla f(y_N)\|^2 + L\|x_N - y_N\|^2\right] \\ &\leq 4d + 2L\mathbb{E}\left[\|v_N - w_N\|^2 + \|x_N - y_N\|^2\right] \\ &\leq 4d + 6L\mathbb{E}\left[\|x_N - y_N\|^2 + \|(x_N + v_N) - (y_N + w_N)\|^2\right],\end{aligned}$$

By Lemma 12 and our choice of N,

$$\sum_{n=0}^{N-1}\|\nabla f(x_n(0))\|^2 \leq O\left(\frac{\kappa dL}{h}\log\left(\frac{1}{\epsilon^2}\right) + \frac{L^2}{h}\mathbb{E}\left[\|x_N - y_N\|^2 + \|(x_N + v_N) - (y_N + w_N)\|^2\right]\right),$$

and

$$\sum_{n=0}^{N-1}\mathbb{E}\|v_n(0)\|^2 \leq O\left(\frac{d}{hm}\log\left(\frac{1}{\epsilon^2}\right) + \mathbb{E}\left[\|x_N - y_N\|^2 + \|(x_N + v_N) - (y_N + w_N)\|^2\right]\right).$$

Thus,

$$\sum_{n=1}^{N} \frac{2\kappa}{h} \left(2\mathbb{E} \left\| \mathbb{E}_\alpha v_n - v_n^* \right\|^2 + 3\mathbb{E} \left\| \mathbb{E}_\alpha x_n - x_n^* \right\|^2 \right) + \sum_{n=1}^{N} \left(2\mathbb{E} \left\| v_n - v_n^* \right\|^2 + 3\mathbb{E} \left\| x_n - x_n^* \right\|^2 \right)$$
$$\leq O\left(\left(\frac{\kappa d h^6}{m} + \frac{d h^3}{m} \right) \log\left(\frac{1}{\epsilon^2} \right) \right)$$
$$+ O\left(\kappa h^7 + h^3 \right) \mathbb{E} \left[\left\| x_N - y_N \right\|^2 + \left\| (x_N + v_N) - (y_N + w_N) \right\|^2 \right].$$

Then, we can choose a small constant C such that if we let

$$h = C \min\left(\frac{\epsilon^{1/3}}{\kappa^{1/6}} \log^{-1/6}\left(\frac{1}{\epsilon^2} \right), \epsilon^{2/3} \log^{-1/3}\left(\frac{1}{\epsilon^2} \right) \right),$$

then

$$\sum_{n=1}^{N} \frac{2\kappa}{h} \left(2\mathbb{E} \left\| \mathbb{E}_\alpha v_n - v_n^* \right\|^2 + 3\mathbb{E} \left\| \mathbb{E}_\alpha x_n - x_n^* \right\|^2 \right) + \sum_{n=1}^{N} \left(2\mathbb{E} \left\| v_n - v_n^* \right\|^2 + 3\mathbb{E} \left\| x_n - x_n^* \right\|^2 \right)$$
$$\leq \frac{\epsilon^2 d}{4m} + \frac{1}{2} \mathbb{E} \left[\left\| x_N - y_N \right\|^2 + \left\| (x_N + v_N) - (y_N + w_N) \right\|^2 \right].$$

Therefore,

$$\begin{aligned} & \mathbb{E} \left[\left\| x_N - y_N \right\|^2 + \left\| (x_N + v_N) - (y_N + w_N) \right\|^2 \right] \\ \leq \; & \frac{\epsilon^2 d}{4m} + \frac{\epsilon^2 d}{4m} + \frac{1}{2} \mathbb{E} \left[\left\| x_N - y_N \right\|^2 + \left\| (x_N + v_N) - (y_N + w_N) \right\|^2 \right] \\ = \; & \frac{\epsilon^2 d}{2m} + \frac{1}{2} \mathbb{E} \left[\left\| x_N - y_N \right\|^2 + \left\| (x_N + v_N) - (y_N + w_N) \right\|^2 \right], \end{aligned}$$

which implies

$$\mathbb{E} \left[\left\| x_N - y_N \right\|^2 \right] \leq \mathbb{E} \left[\left\| x_N - y_N \right\|^2 + \left\| (x_N + v_N) - (y_N + w_N) \right\|^2 \right] \leq \frac{\epsilon^2 d}{m}.$$

By our choice of h,

$$N \;\; \leq \;\; \tilde{O}\left(\frac{\kappa^{7/6}}{\epsilon^{1/3}} + \frac{\kappa}{\epsilon^{2/3}} \right).$$

□

A.6 Discretization Error of Algorithm 2

Here, we bound the discretization error in one step of Algorithm 2. Since the terms $\mathbb{E} \left\| \mathbb{E}_\alpha x_{n+1} - x_n^*(h) \right\|^2$ and $\mathbb{E} \left\| x_{n+1} - x_n^*(h) \right\|^2$ are dominated by the terms $\mathbb{E} \left\| \mathbb{E}_\alpha v_{n+1} - v_n^*(h) \right\|^2$ and $\mathbb{E} \left\| v_{n+1} - v_n^*(h) \right\|^2$, we bound only the later two terms.

Lemma 13. *Assume that* $R^4\delta^4 \leq \frac{1}{4}$. *Let* $x_n^{(k-1,i)}$ *for* $i = 1, ..., R$, $k = 1, ..., K$ *be the intermediate value computed in iteration* n *of Algorithm 2. Let* $\{x_n^*(t), v_n^*(t)\}_{t\in[0,h]}$ *be the ideal underdamped Langevin diffusion, starting from* $x_n^*(0) = x_n$ *and* $v_n^*(0) = v_n$, *coupled through a shared Brownian motion with* $\left\{x_n^{(k-1,i)}\right\}_{i=1,...,R,k=1,...,K}$. *Then, for any* $i = 1, ..., R$, *and* $k = 1, ..., K-1$,

$$\begin{aligned}\mathbb{E}\left\|x_n^{(k,i)} - x_n^*(\alpha_i h)\right\|^2 \quad &\leq \quad \left(2R^4\delta^4\right)^k \frac{1}{R}\sum_{j=1}^{R}\mathbb{E}\left\|x_n - x_n^*(\alpha_j h)\right\|^2 \\ &\quad +4R^3\delta^4\sum_{j=1}^{R}\mathbb{E}\sup_{s\in[(j-1)\delta, j\delta]}\left\|x_n^*(\alpha_j h) - x_n^*(s)\right\|^2.\end{aligned}$$

Proof. For any $i = 1, ..., R$, and $k = 1, ..., K-1$,

$$\begin{aligned}&\mathbb{E}\left\|x_n^{(k,i)} - x_n^*(\alpha_i h)\right\|^2 \\ \leq\ &\mathbb{E}\left\|\frac{1}{2}u\sum_{j=1}^{i}\left[\int_{(j-1)\delta}^{\min(j\delta,\alpha_i h)}\left(1 - e^{-2(\alpha_i h - s)}\right)\mathrm{d}s\cdot\nabla f(x_n^{(k-1,j)})\right]\right. \\ &\left.-\frac{1}{2}u\int_0^{\alpha_i h}\left(1 - e^{-2(\alpha_i h - s)}\right)\nabla f(x_n^*(s))\,\mathrm{d}s\right\|^2 \\ \leq\ &\frac{1}{2}\mathbb{E}\left\|u\sum_{j=1}^{i}\left[\int_{(j-1)\delta}^{\min(j\delta,\alpha_i h)}\left(1 - e^{-2(\alpha_i h - s)}\right)\mathrm{d}s\cdot\left(\nabla f(x_n^{(k-1,j)}) - \nabla f(x_n^*(\alpha_j h))\right)\right]\right\|^2 \\ &+\frac{1}{2}\mathbb{E}\left\|u\sum_{j=1}^{i}\left[\int_{(j-1)\delta}^{\min(j\delta,\alpha_i h)}\left(1 - e^{-2(\alpha_i h - s)}\right)\left(\nabla f(x_n^*(\alpha_j h)) - \nabla f(x_n^*(s))\right)\mathrm{d}s\right]\right\|^2,\end{aligned}$$

where the first step follows by the definition, and the second step follows by Young's inequality.

To compute the first term,

$$\begin{aligned}&\frac{1}{2}\mathbb{E}\left\|u\sum_{j=1}^{i}\left[\int_{(j-1)\delta}^{\min(j\delta,\alpha_i h)}\left(1 - e^{-2(\alpha_i h - s)}\right)\mathrm{d}s\cdot\left(\nabla f(x_n^{(k-1,j)}) - \nabla f(x_n^*(\alpha_j h))\right)\right]\right\|^2 \\ \leq\ &\frac{1}{2}u^2 R\sum_{j=1}^{i}\mathbb{E}\left\|\int_{(j-1)\delta}^{\min(j\delta,\alpha_i h)}\left(1 - e^{-2(\alpha_i h - s)}\right)\mathrm{d}s\cdot\left(\nabla f(x_n^{(k-1,j)}) - \nabla f(x_n^*(\alpha_j h))\right)\right\|^2 \\ \leq\ &2R^3\delta^4\sum_{j=1}^{R}\mathbb{E}\left\|x_n^{(k-1,j)} - x_n^*(\alpha_j h)\right\|^2, \qquad \text{(A.12)}\end{aligned}$$

where the first step follows by the inequality $\left(\sum_{i=1}^n a_i\right)^2 \leq n\sum_{i=1}^n a_i^2$, the second step follows by $1 - e^{-2(\alpha_i h - s)} \leq 2R\delta$ and ∇f is L-Lipschitz.

For the second term,

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}\left\|u\sum_{j=1}^{i}\left[\int_{(j-1)\delta}^{\min(j\delta,\alpha_i h)}\left(1-e^{-2(\alpha_i h-s)}\right)\left(\nabla f(x_n^*(\alpha_j h))-\nabla f(x_n^*(s))\right)\mathrm{d}s\right]\right\|^2\\
\leq\ & \frac{1}{2}u^2R\sum_{j=1}^{i}\mathbb{E}\left\|\int_{(j-1)\delta}^{\min(j\delta,\alpha_i h)}\left(1-e^{-2(\alpha_i h-s)}\right)\left(\nabla f(x_n^*(\alpha_j h))-\nabla f(x_n^*(s))\right)\mathrm{d}s\right\|^2\\
\leq\ & 2R^3\delta^4\sum_{j=1}^{R}\mathbb{E}\sup_{s\in[(j-1)\delta,j\delta]}\left\|x_n^*(\alpha_j h)-x_n^*(s)\right\|^2, \qquad \text{(A.13)}
\end{aligned}$$

where the first step follows by the inequality $(\sum_{i=1}^n a_i)^2 \leq n\sum_{i=1}^n a_i^2$ and the second step follows by $1-e^{-2(\alpha_i h-s)} \leq 2R\delta$ and ∇f is L-Lipschitz. Thus,

$$\begin{aligned}
&\mathbb{E}\left\|x_n^{(k,i)}-x_n^*(\alpha_i h)\right\|^2\\
\leq\ & 2R^3\delta^4\sum_{j=1}^{R}\mathbb{E}\left\|x_n^{(k-1,j)}-x_n^*(\alpha_j h)\right\|^2+2R^3\delta^4\sum_{j=1}^{R}\mathbb{E}\sup_{s\in[(j-1)\delta,j\delta]}\left\|x_n^*(\alpha_j h)-x_n^*(s)\right\|^2\\
\leq\ & \left(2R^4\delta^4\right)^k\frac{1}{R}\sum_{j=1}^{R}\mathbb{E}\left\|x_n-x_n^*(\alpha_j h)\right\|^2\\
&+\left(1+2R^4\delta^4+...+\left(2R^4\delta^4\right)^{k-1}\right)2R^3\delta^4\sum_{j=1}^{R}\mathbb{E}\sup_{s\in[(j-1)\delta,j\delta]}\left\|x_n^*(\alpha_j h)-x_n^*(s)\right\|^2\\
\leq\ & \left(2R^4\delta^4\right)^k\frac{1}{R}\sum_{j=1}^{R}\mathbb{E}\left\|x_n-x_n^*(\alpha_j h)\right\|^2+4R^3\delta^4\sum_{j=1}^{R}\mathbb{E}\sup_{s\in[(j-1)\delta,j\delta]}\left\|x_n^*(\alpha_j h)-x_n^*(s)\right\|^2,
\end{aligned}$$

where the first step follows by (A.12) and (A.13), the second step follows by induction, and the third step follows by $2R^4\delta^4 \leq \frac{1}{2}$. □

Lemma 14. *Let (v_n, x_n) be the iterates of iteration n. Let $x_n^{(k,i)}$ for $i = 1,...,R$, $k = 1,...,K-1$ be the intermediate value computed in iteration n of Algorithm 2. Let $\{x_n^*(t), v_n^*(t)\}_{t\in[0,h]}$ be the ideal underdamped Langevin diffusion, starting from $x_n^*(0) = x_n$ and $v_n^*(0) = v_n$, coupled through a shared Brownian motion with $\left\{x_n^{(k,i)}\right\}_{i=1,...,R,k=1,...,K-1}$. Assume that $h = R\delta \leq \frac{1}{10}$ and $K \geq \Omega\left(\log\frac{1}{\delta^4}\right)$. Let $\mathbb{E}_\alpha$ be the expectation taken over the choice of $\alpha_1,...,\alpha_R$ in iteration n. Let $\mathbb{E}$ be the expectation taken over other randomness in iteration n. Then,*

$$\begin{aligned}
\mathbb{E}\left\|\mathbb{E}_\alpha v_{n+1}-v_n^*(h)\right\|^2 &\leq O\left(R^6\delta^8\left\|v_n\right\|^2+u^2R^6\delta^{10}\left\|\nabla f(x_n)\right\|^2+R^6\delta^9 ud\right),\\
\mathbb{E}\left\|v_{n+1}-v_n^*(h)\right\|^2 &\leq O\left(R^2\delta^4\left\|v_n\right\|^2+u^2R^2\delta^4\left\|\nabla f(x_n)\right\|^2+R^2\delta^5 ud\right).
\end{aligned}$$

Proof. To show the first claim,

$$\mathbb{E}\left\|\mathbb{E}_\alpha v_{n+1}-v_n^*(h)\right\|^2$$

$$
\begin{aligned}
&\leq \mathbb{E}\left\|\mathbb{E}_\alpha u\sum_{i=1}^{R}\delta e^{-2(h-\alpha_i h)}\nabla f(x_n^{(K-1,i)}) - u\int_0^h e^{-2(h-s)}\nabla f(x_n^*(s))\,\mathrm{d}s\right\|^2\\
&\leq 2\mathbb{E}\left\|u\sum_{i=1}^{R}\delta e^{-2(h-\alpha_i h)}\nabla f(x_n^{(K-1,i)}) - u\sum_{i=1}^{R}\delta e^{-2(h-\alpha_i h)}\nabla f(x_n^*(\alpha_i h))\right\|^2\\
&\quad +2\mathbb{E}\left\|\mathbb{E}_\alpha u\sum_{i=1}^{R}\delta e^{-2(h-\alpha_i h)}\nabla f(x_n^*(\alpha_i h)) - u\int_0^h e^{-2(h-s)}\nabla f(x_n^*(s))\,\mathrm{d}s\right\|^2\\
&\leq 2\delta^2 R\sum_{i=1}^{R}\mathbb{E}\left\|x_n^{(K-1,i)} - x_n^*(\alpha_i h)\right\|^2 + 0\\
&\leq 2\delta^2 R\left(2R^4\delta^4\right)^{K-1}\sum_{i=1}^{R}\mathbb{E}\left\|x_n - x_n^*(\alpha_i h)\right\|^2\\
&\quad +8R^5\delta^6\sum_{i=1}^{R}\mathbb{E}\sup_{s\in[(i-1)\delta,i\delta]}\left\|x_n^*(\alpha_i h) - x_n^*(s)\right\|^2, \qquad \text{(A.14)}
\end{aligned}
$$

where the first step follows by the definition, the second step follows by Young's inequality, and the third step follows by

$$
\mathbb{E}_\alpha\delta e^{-2(h-\alpha_i h)}\nabla f(x_n^*(\alpha_i h)) \;=\; \int_{(i-1)\delta}^{i\delta} e^{-2(h-s)}\nabla f(x_n^*(s))\,\mathrm{d}s.
$$

To show the second claim,

$$
\begin{aligned}
&\mathbb{E}\left\|v_{n+1} - v_n^*(h)\right\|^2\\
&\leq \mathbb{E}\left\|u\sum_{i=1}^{R}\delta e^{-2(h-\alpha_i h)}\nabla f(x_n^{(K-1,i)}) - u\int_0^h e^{-2(h-s)}\nabla f(x_n^*(s))\,\mathrm{d}s\right\|^2\\
&\leq 3\mathbb{E}\left\|u\sum_{i=1}^{R}\delta e^{-2(h-\alpha_i h)}\nabla f(x_n^{(K-1,i)}) - u\sum_{i=1}^{R}\delta e^{-2(h-\alpha_i h)}\nabla f(x_n^*(\alpha_i h))\right\|^2\\
&\quad +3\mathbb{E}\left\|u\sum_{i=1}^{R}\int_{(i-1)\delta}^{i\delta} e^{-2(h-\alpha_i h)}\left(\nabla f(x_n^*(\alpha_i h)) - \nabla f(x_n^*(s))\right)\mathrm{d}s\right\|^2\\
&\quad +3\mathbb{E}\left\|u\sum_{i=1}^{R}\int_{(i-1)\delta}^{i\delta}\left(e^{-2(h-\alpha_i h)} - e^{-2(h-s)}\right)\nabla f(x_n^*(s))\,\mathrm{d}s\right\|^2.
\end{aligned}
$$

Like the proof of the third claim, the first term satisfies

$$
\begin{aligned}
&3\mathbb{E}\left\|u\sum_{i=1}^{R}\delta e^{-2(h-\alpha_i h)}\nabla f(x_n^{(K-1,i)}) - u\sum_{i=1}^{R}\delta e^{-2(h-\alpha_i h)}\nabla f(x_n^*(\alpha_i h))\right\|^2\\
&\leq 3\delta^2 R\left(2R^4\delta^4\right)^{K-1}\sum_{i=1}^{R}\mathbb{E}\left\|x_n - x_n^*(\alpha_i h)\right\|^2 + 12R^5\delta^6\sum_{i=1}^{R}\mathbb{E}\sup_{s\in[(i-1)\delta,i\delta]}\left\|x_n^*(\alpha_i h) - x_n^*(s)\right\|^2.
\end{aligned}
$$

The second term satisfies

$$
3\mathbb{E}\left\|u\sum_{i=1}^{R}\int_{(i-1)\delta}^{i\delta} e^{-2(h-\alpha_i h)}\left(\nabla f(x_n^*(\alpha_i h)) - \nabla f(x_n^*(s))\right)\mathrm{d}s\right\|^2
$$

$$\begin{aligned} &\leq 3u^2R\sum_{i=1}^{R}\mathbb{E}\left\|\int_{(i-1)\delta}^{i\delta}e^{-2(h-\alpha_ih)}\left(\nabla f(x_n^*(\alpha_ih))-\nabla f(x_n^*(s))\right)\mathrm{d}s\right\|^2 \\ &\leq 3\delta^2R\sum_{i=1}^{R}\mathbb{E}\sup_{s\in[(i-1)\delta,i\delta]}\left\|x_n^*(\alpha_ih)-x_n^*(s)\right\|^2, \end{aligned}$$

where the first step follows by $\left(\sum_{i=1}^n a_i\right)^2 \leq n\sum_{i=1}^n a_i^2$, and the second step follows by ∇f is L-Lipschitz.

The last term satisfies

$$3\mathbb{E}\left\|u\sum_{i=1}^{R}\int_{(i-1)\delta}^{i\delta}\left(e^{-2(h-\alpha_ih)}-e^{-2(h-s)}\right)\nabla f(x_n^*(s))\,\mathrm{d}s\right\|^2 \leq 12u^2R^2\delta^4\mathbb{E}\sup_{s\in[0,h]}\left\|\nabla f(x_n^*(s))\right\|^2,$$

which follows by $e^{-2(h-\alpha_ih)}-e^{-2(h-s)}\leq 2\delta$ for $s\in[(i-1)\delta,i\delta]$. Thus,

$$\begin{aligned} &\mathbb{E}\left\|v_{n+1}-v_n^*(h)\right\|^2 \\ \leq\ & 3\delta^2R\left(2R^4\delta^4\right)^{K-1}\sum_{i=1}^{R}\mathbb{E}\left\|x_n-x_n^*(\alpha_ih)\right\|^2 \\ &+12R^5\delta^6\sum_{i=1}^{R}\mathbb{E}\sup_{s\in[(i-1)\delta,i\delta]}\left\|x_n^*(\alpha_ih)-x_n^*(s)\right\|^2 \\ &+3\delta^2R\sum_{i=1}^{R}\mathbb{E}\sup_{s\in[(i-1)\delta,i\delta]}\left\|x_n^*(\alpha_ih)-x_n^*(s)\right\|^2+12u^2R^2\delta^4\mathbb{E}\sup_{s\in[0,h]}\left\|\nabla f(x_n^*(s))\right\|^2 \end{aligned} \quad \text{(A.15)}$$

By Lemma 6, for $i=1,...,R$,

$$\mathbb{E}\left\|x_n-x_n^*(\alpha_ih)\right\|^2 \leq O\left(R^2\delta^2\left\|v_n\right\|^2+u^2R^4\delta^4\left\|\nabla f(x_n)\right\|^2+udR^3\delta^3\right),$$

,and

$$\mathbb{E}\sup_{s\in[(i-1)\delta,i\delta]}\left\|x_n^*(\alpha_ih)-x_n^*(s)\right\|^2 \leq O\left(\delta^2\left\|v_n\right\|^2+u^2\delta^4\left\|\nabla f(x_n)\right\|^2+ud\delta^3\right).$$

Thus, when $K\geq\Omega\left(\log\frac{1}{\delta^4}\right)$, since $R\delta\leq\frac{1}{10}$, $\left(2R^4\delta^4\right)^{K-1}\leq O\left(\delta^4\right)$. By (A.14) and (A.15),

$$\mathbb{E}\left\|\mathbb{E}_\alpha v_{n+1}-v_n^*(h)\right\|^2 \leq O\left(R^6\delta^8\left\|v_n\right\|^2+u^2R^6\delta^{10}\left\|\nabla f(x_n)\right\|^2+R^6\delta^9ud\right),$$

and

$$\mathbb{E}\left\|v_{n+1}-v_n^*(h)\right\|^2 \leq O\left(R^2\delta^4\left\|v_n\right\|^2+u^2R^2\delta^4\mathbb{E}\left\|\nabla f(x_n)\right\|^2+R^2\delta^5ud\right).$$

□

Appendix B

DEFERRED CONTENTS FROM CHAPTER 3

B.1 Equivalence of HMC and Metropolis-adjusted Langevin dynamics

We briefly remark on the equivalence of Metropolized HMC and the Metropolis-adjusted Langevin dynamics algorithm (MALA), a well-studied algorithm since its introduction in [Bes94]. This equivalence was also commented on in [CDWY20]. The algorithm can be seen as a filtered discretization of the continuous-time Langevin dynamics,

$$dx_t = -\nabla f(x_t)dt + \sqrt{2}dW_t,$$

where W_t is Brownian motion. In short, the Metropolized HMC update is

$$v \sim \mathcal{N}(0, I),\ \widetilde{x} \leftarrow x + \eta v - \frac{\eta^2}{2}\nabla f(x),\ \text{accept with probability } \min\left\{1, \frac{\exp(-\mathrm{Ham}(\widetilde{x}, \widetilde{v}))}{\exp(-\mathrm{Ham}(x, v))}\right\}.$$

Similarly, the MALA update with step size h is

$$\widetilde{x} \sim \mathcal{N}(x - h\nabla f(x), 2hI),\ \text{accept with probability } \min\left\{1, \frac{\exp(-f(\widetilde{x}) - \|x - \widetilde{x} + h\nabla f(\widetilde{x})\|_2^2/4h)}{\exp(-f(x) - \|\widetilde{x} - x + h\nabla f(x)\|_2^2/4h)}\right\}.$$

It is clear that in HMC the distribution of $\widetilde{x}$ is

$$\widetilde{x} \sim \mathcal{N}\left(x - \frac{\eta^2}{2}\nabla f(x), \eta^2 I\right),$$

so it suffices to show for $h = \eta^2/2$,

$$\frac{\|\widetilde{x} - x + h\nabla f(x)\|_2^2 - \|x - \widetilde{x} + h\nabla f(\widetilde{x})\|_2^2}{4h} = \frac{1}{2}\left(\|v\|_2^2 - \|\widetilde{v}\|_2^2\right).$$

Indeed, the right hand side simplifies to

$$\frac{\eta}{2}\langle \nabla f(\widetilde{x}) + \nabla f(x), v\rangle - \frac{\eta^2}{8}\|\nabla f(\widetilde{x}) + \nabla f(x)\|_2^2,$$

and the left hand side is

$$\begin{aligned}&\frac{1}{2}\langle \nabla f(\widetilde{x}) + \nabla f(x), \widetilde{x} - x\rangle + \frac{h}{4}\left(\|\nabla f(x)\|_2^2 - \|\nabla f(\widetilde{x})\|_2^2\right)\\ =\ &\frac{1}{2}\left\langle \nabla f(\widetilde{x}) + \nabla f(x), \eta v - \frac{\eta^2}{2}\nabla f(x)\right\rangle + \frac{\eta^2}{8}\left(\|\nabla f(x)\|_2^2 - \|\nabla f(\widetilde{x})\|_2^2\right).\end{aligned}$$

Comparing coefficients shows the equivalence.

B.2 Improved concentration under Hessian log-Sobolev inequality

In this section, we show that the bound in Theorem 3 may be sharpened under a Hessian log-Sobolev inequality (LSI), which we define presently.

Definition 6 (Hessian log-Sobolev)**.** *We say density $d\pi^*/dx \propto \exp(-f(x))$ satisfies a* Hessian log-Sobolev inequality *if for all continuously differentiable $g : \mathbb{R}^d \to \mathbb{R}$, and for*

$$\mathrm{Ent}_{\pi^*}[g] \stackrel{\mathrm{def}}{=} \left(\int g(x)\log(g(x))\,d\pi^*(x)\right) - \left(\int g(x)d\pi^*(x)\right)\log\left(\int g(x)d\pi^*(x)\right),$$

we have

$$\mathrm{Ent}_{\pi^*}\left[g^2\right] \leq 2\int_{\mathbb{R}^d}\left\langle \left(\nabla^2 f(x)\right)^{-1}\nabla g(x), \nabla g(x)\right\rangle d\pi^*(x).$$

In general, this is a much more restrictive condition than Theorem 2; some sufficient conditions are given in [BL00]. We now show an improved concentration result under a Hessian LSI; the proof follows Herbst's argument, a framework developed in [Led99].

Theorem 15 (Gradient norm concentration under LSI)**.** *Suppose f is L-smooth and strongly convex, and $d\pi^*(x)/dx \propto \exp(-f(x))$ satisfies a Hessian log-Sobolev inequality. Then for all $c > 0$,*

$$\Pr_{\pi^*}\left[\|\nabla f(x)\| \geq \mathbb{E}_{\pi^*}[\|\nabla f\|] + c\sqrt{2L\log d}\right] \leq d^{-c^2}.$$

Proof. Denote $G \stackrel{\mathrm{def}}{=} \|\nabla f\|$, where we note $\nabla G = \frac{(\nabla^2 f)\nabla f}{\|\nabla f\|}$. Let $H(\lambda) \stackrel{\mathrm{def}}{=} \mathbb{E}_{\pi^*}[\exp(\lambda G)]$, such that $H'(\lambda) = \mathbb{E}_{\pi^*}[G\exp(\lambda G)]$. Then, for $g^2 = \exp(\lambda G)$,

$$H(\lambda) = \mathbb{E}_{\pi^*}\left[g^2\right],\ \lambda H'(\lambda) = \mathbb{E}_{\pi^*}\left[g^2\log g^2\right].$$

This in turn implies via the LSI that

$$\lambda H'(\lambda) - H(\lambda)\log H(\lambda) = \mathbb{E}_{\pi^*}\left[g^2\log g^2\right] - \mathbb{E}_{\pi^*}\left[g^2\right]\log\mathbb{E}_{\pi^*}\left[g^2\right] \leq 2\mathbb{E}_{\pi^*}\left[\|\nabla g\|^2_{(\nabla^2 f)^{-1}}\right]. \tag{B.1}$$

By smoothness and the definition of $g = \exp(\frac{1}{2}\lambda G)$, we may bound the right hand side:

$$\mathbb{E}_{\pi^*}\left[\|\nabla g\|^2_{(\nabla^2 f)^{-1}}\right] = \frac{\lambda^2}{4}\mathbb{E}_{\pi^*}\left[\|\nabla G\|^2_{(\nabla^2 f)^{-1}}\exp(\lambda G)\right] \leq \frac{\lambda^2 L}{4}H(\lambda). \tag{B.2}$$

In the last inequality we used our calculation of ∇G, and $\nabla^2 f \preceq LI_d$. Now, consider the function $K(\lambda) = \frac{1}{\lambda}\log H(\lambda)$. We handle the definition of $K(0)$ by a limiting argument (and $\log(1+x) \approx x$):

$$K(0) = \lim_{\lambda\to 0}\frac{1}{\lambda}\log\mathbb{E}_{\pi^*}\left[e^{\lambda G}\right] = \frac{H(\lambda) - H(0)}{\lambda} = H'(0) = \mathbb{E}_{\pi^*}[G].$$

We compute

$$K'(\lambda) = -\frac{1}{\lambda^2}\log H(\lambda) + \frac{H'(\lambda)}{\lambda H(\lambda)} = \frac{\lambda H'(\lambda) - H(\lambda)\log H(\lambda)}{\lambda^2 H(\lambda)}.$$

This, combined with (B.1) and (B.2) imply $K'(\lambda) \leq \frac{L}{2}$. Therefore, by integrating, we have

$$K(\lambda) \leq \mathbb{E}_{\pi^*}[G] + \frac{L\lambda}{2} \Rightarrow H(\lambda) = \exp(\lambda K(\lambda)) \leq \exp\left(\lambda \mathbb{E}_{\pi^*}[G] + \frac{L\lambda^2}{2}\right).$$

Finally, we have concentration:

$$\Pr_{\pi^*}[G \geq \mathbb{E}_{\pi^*}[G] + r] = \Pr_{\pi^*}[\exp(\lambda G) \geq \exp(\lambda \mathbb{E}_{\pi^*}[G] + \lambda r)] \leq \exp\left(-\lambda r + \frac{L\lambda^2}{2}\right),$$

where the last statement is by Markov. Choosing $r = c\sqrt{2L\log d}$, $\lambda = r/L$ yields the conclusion. □

B.3 Mixing time proofs

We prove various claims from Section 3.4; notation here is consistent with definitions in the body of the paper. All definitions will be with respect to some reversible random walk on $\mathbb{R}^d$ with transition distributions $\mathcal{T}_x$ and stationary distribution $d\pi^*$. We use $d\pi_k$ to denote the density after k steps.

B.3.1 Blocking conductance framework

In this section, we recall the *blocking conductance* framework of [KLM06]. This section is a restatement of their results for our purposes.

Preliminaries

Let $d\rho_k$ denote the "average" distribution density after k steps, e.g. $d\rho_k(x) = \frac{1}{k}\sum_{0\leq i<k} d\pi_i(x)$. Define the flow between two sets $S, T \subseteq \mathbb{R}^d$ by

$$Q(S,T) \stackrel{\text{def}}{=} \int_S \mathcal{T}_x(T) d\pi^*(x),\ Q(S) \stackrel{\text{def}}{=} Q(S, S^c).$$

For every $S \subseteq \mathbb{R}^d$ and $0 \leq x \leq \pi^*(S)$, define the conductance function by

$$\Psi(x, S^c) \stackrel{\text{def}}{=} \min_{T\subseteq S, \pi^*(T)=x} Q(T, S^c).$$

In other words, $\Psi(x, S^c)$ is the smallest amount of flow from subsets of S with stationary measure x to S^c. It is clear that $\Psi(x, S^c)$ is monotone increasing in x in the range $0 \leq$

$x \leq \pi^*(S)$, as choosing a subset of larger measure can only increase flow. By convention, for $1 \geq x \geq \pi^*(S)$,

$$\Psi(x, S^c) \stackrel{\text{def}}{=} \Psi(1 - x, S).$$

This definition clearly makes sense because $x \geq \pi^*(S) \Rightarrow 1 - x \leq \pi^*(S^c)$. Next, let the *spread* of $S \subseteq \mathbb{R}^d$ be defined as

$$\psi(S) \stackrel{\text{def}}{=} \int_0^{\pi^*(S)} \Psi(x, S^c) dx. \tag{B.3}$$

In other words, we can think of $\psi(S)$ as the worst-case flow between a subset of S and S^c, where the measure of the subset is averaged uniformly over $[0, \pi^*(S)]$. The spread enjoys the following useful property, which allows us to think of the spread as a notion of conductance.

Lemma 56. *For any set $S \subseteq \mathbb{R}^d$,*

$$\psi(S) \geq \frac{1}{4} Q(S)^2.$$

Proof. We claim first that for any $t \in [0, \pi^*(S)]$,

$$\psi(S) \geq (\pi^*(S) - t)\Psi(t, S^c). \tag{B.4}$$

To see this, we integrated only in the range $[t, \pi^*(S)]$, and used monotonicity of $\Psi(\cdot, S^c)$. Let $\gamma(S)$ denote the minimum measure of a subset R of S, such that $Q(R, S^c) = \frac{1}{2} Q(S, S^c)$. Note that this means any set T with measure $\pi^*(S) - \gamma(S)$ has flow $Q(T, S^c)$ at least

$$Q(S) - \frac{1}{2} Q(S) = \frac{1}{2} Q(S).$$

In (B.4), let $t = \pi^*(S) - \gamma(S)$, and let T be the subset which admits the value $\Psi(t, S^c)$, i.e. such that $Q(T, S^c) = \Psi(t, S^c)$ and $\pi^*(T) = t$. In particular, this implies

$$\psi(S) \geq \gamma(S) Q(T, S^c) \geq \frac{1}{2} \gamma(S) Q(S).$$

To show the conclusion, it suffices to show $\gamma(S) \geq Q(S)/2$. This is clear because if $\gamma(S) < Q(S)/2$, then any set of stationary measure $\gamma(S)$ could not absorb a flow of $Q(S)/2$ from the set S^c. □

The final definitions we will need are as follows. For an iteration k, let

$$g_k(x) \stackrel{\text{def}}{=} k \frac{d\rho_k}{d\pi^*}(x) = \sum_{0 \leq i < k} \frac{d\pi_i}{d\pi^*}(x).$$

A useful interpretation is that $\int_S g_k(x)d\pi^*(x)$ measures how many times the set S was visited on expectation over the first k iterations. Let $m_k : \mathbb{R}^d \to [0,1]$ be a measure-preserving map such that $g_k(m_k^{-1}(\cdot))$ is an increasing function. In other words, m_k orders the space $\mathbb{R}^d$ by their value g_k, such that for $0 \le s < t \le 1$, $g_k(m_k^{-1}(s)) \le g_k(m_k^{-1}(t))$. We define

$$q_k(x,y) = Q\left(m_k^{-1}([0,\min(x,y)]), m_k^{-1}([\max(x,y),1])\right). \tag{B.5}$$

In other words, for $x \le y$, q_k takes the set of measure x according to $d\rho_k/d\pi^*$ of least probability, and of measure $1-y$ of most probability, and measures the flow between them. For notational simplicity and when clear from context, we identify $x \in [0,1]$ with $m_k^{-1}(x)$, and similarly identify intervals. The following is then immediate:

$$\frac{d}{dx}q_k(x,y) = \begin{cases} \mathcal{T}_x([y,1]) & x < y \\ -\mathcal{T}_x([0,y]) & x \ge y \end{cases}. \tag{B.6}$$

Main claim

Here, we recall the main result of the blocking conductance framework in terms of mixing times.

Theorem 16. *Suppose the starting distribution π_0 is β-warm with respect to π^*. Let $h : [c, 1-c] \to \mathbb{R}_{\ge 0}$ satisfy, for some $c \in (0, \frac{1}{2})$, and some k,*

$$\int_c^{1-c} h(y)q_k(x,y)dy \ge 2x(1-x),\ \forall x \in [c,1-c]. \tag{B.7}$$

Then,

$$\|\rho_k - \pi^*\|_{\mathrm{TV}} \le \beta c + \frac{1}{k}\int_c^{1-c} h(x)dx.$$

We call a function h which satisfies (B.7) a c-mixweight function, and show how to construct such a function in Section B.3.2; they will be inversely related to standard notions of conductance. We first require the following helper results.

Lemma 57. *For any $t \in [0,1]$,*

$$\int_0^1 q_k(x,t)dg_k(x) = \pi_k([0,t]) - \pi_0([0,t]) = \pi_0([t,1]) - \pi_k([t,1]).$$

Proof. The equality $\pi_k([0,t]) - \pi_0([0,t]) = \pi_0([t,1]) - \pi_k([t,1])$ follows by definition, so we will simply show the first equality. Using (B.6) for $x \le t$ and integrating by parts,

$$\int_0^t g_k(x)\mathcal{T}_x([t,1])d\pi^*(x) = g_k(t)q(t,t) - \int_0^t q_k(x,t)dg_k(x).$$

Similarly,

$$\int_t^1 g_k(x)\mathcal{T}_x([0,t])d\pi^*(x) = g_k(t)q(t,t) + \int_t^1 q_k(x,t)dg_k(x).$$

Therefore, to derive the conclusion of the lemma, it suffices to show that

$$\int_t^1 g_k(x)\mathcal{T}_x([0,t])d\pi^*(x) - \int_0^t g_k(x)\mathcal{T}_x([t,1])d\pi^*(x) = \pi_k([0,t]) - \pi_0([0,t]).$$

By expanding the definition of g_k and telescoping, it suffices to show for all $0 \le i \le k-1$,

$$\int_t^1 \mathcal{T}_x([0,t])d\pi_i(x) - \int_0^t \mathcal{T}_x([t,1])d\pi_i(x) = \pi_{i+1}([0,t]) - \pi_i([0,t]).$$

This follows from

$$\pi_{i+1}([0,t]) - \pi_i([0,t]) = \left(\int_0^t (1 - \mathcal{T}_x([t,1]))\, d\pi_i(x) + \int_t^1 \mathcal{T}_x([0,t])d\pi_i(x)\right) - \int_0^t d\pi_i(x).$$

□

Next, let $t_0 \in [0,1]$ be such that $g_k(t_0) = k$, where we note that $\mathbb{E}_{\pi^*}[g_k] = k$ is the expected value.

Lemma 58.

$$\|\rho_k - \pi^*\|_{\mathrm{TV}} = \frac{1}{k}\int_0^{t_0} t dg_k(t) = \frac{1}{k}\int_{t_0}^1 (1-t)dg_k(t).$$

Proof. By the definition of t_0, we have that for all $t \le t_0$, $\frac{d\rho_k}{d\pi^*}(x) \le 1$, and for $t \ge t_0$, $\frac{d\rho_k}{d\pi^*}(x) \ge 1$. Therefore, the total variation distance is attained by the set $[0,t_0]$, i.e.

$$\|\rho_k - \pi^*\|_{\mathrm{TV}} = \pi^*([0,t_0]) - \rho_k([0,t_0]) = \int_0^{t_0}\left(1 - \frac{g_k(x)}{k}\right)d\pi^*(x). \tag{B.8}$$

Integrating by parts,

$$\|\rho_k - \pi^*\|_{\mathrm{TV}} = \left(1 - \frac{g_k(t_0)}{k}\right)t_0 + \frac{1}{k}\int_0^{t_0} t dg_k(t).$$

The first summand vanishes by the definition of t_0, so we attain the first equality in the lemma statement. The second equality follows from the same calculations, using the set $[t_0,1]$ which also attains the total variation distance, i.e. integrating by parts

$$\int_{t_0}^1 \left(\frac{g_k(x)}{k} - 1\right)d\pi^*(x) = \frac{1}{k}\int_{t_0}^1 (1-t)dg_k(t).$$

□

We also remark that for a β-warm start, it follows that every distribution π_i for $i \ge 0$ is also β-warm, as the warmness $d\pi_{i+1}/d\pi^*$ at a point is given by an average over the values $d\pi_i/d\pi^*$ of the prior iteration, and the conclusion follows by induction.

Lemma 59.

$$\|\rho_k - \pi^*\|_{\mathrm{TV}} \le \min\left(\beta c + \frac{1}{k}\int_c^{t_0} t dg_k(t),\ \beta c + \frac{1}{k}\int_{t_0}^{1-c}(1-t)dg_k(t)\right).$$

Proof. Recall in Lemma 58 we characterized

$$\begin{aligned}\|\rho_k - \pi^*\|_{\mathrm{TV}} &= \int_0^{t_0}\left(1 - \frac{g_k(x)}{k}\right)d\pi^*(x) \\ &= \int_0^{c}\left(1 - \frac{g_k(x)}{k}\right)d\pi^*(x) + \int_c^{t_0}\left(1 - \frac{g_k(x)}{k}\right)d\pi^*(x).\end{aligned}$$

Note that the first integral is at most c. The second integral is, integrating by parts,

$$\left(\frac{g_k(c)}{k} - 1\right)c + \frac{1}{k}\int_c^{t_0} t dg_k(t).$$

The first summand is bounded by $(\beta - 1)c$ by our earlier argument about the warmness at every iteration being bounded by β. Finally, the second half of the lemma statement follows by considering the other characterization of the total variation based on $[t_0, 1]$, e.g. bounding

$$\int_{t_0}^{1-c}\left(\frac{g_k(x)}{k} - 1\right)dx + \int_{1-c}^{1}\left(\frac{g_k(x)}{k} - 1\right)dx.$$

□

Proof of Theorem 16. First, if $t_0 \le c$, by (B.8), the total variation distance is at most $c\beta$ as desired. A similar conclusion follows if $t_0 \ge 1 - c$ from the other characterization of total variation in Lemma 59. We now consider when $t \in (c, 1-c)$. By Lemma 57, for all $y \in [c, 1-c]$,

$$1 \ge \pi_k([0, y]) \ge \int_c^{1-c} q_k(x, y) dg_k(x).$$

Multiplying by h and integrating over the range $[c, 1-c]$,

$$\int_c^{1-c} h(x)dx \ge \int_c^{1-c}\left(\int_c^{1-c} h(y)q_k(x, y)dy\right)dg_k(x) \ge \int_c^{1-c} 2x(1-x)dg_k(x).$$

The second inequality recalled the requirement (B.7). By combining this with half of Lemma 59,

$$\begin{aligned}\|\rho_k - \pi^*\|_{\mathrm{TV}} &\le \beta c + \frac{1}{k}\int_c^{t_0} x dg_k(x) \le \beta c + \frac{1}{2(1-t_0)k}\int_c^{t_0} 2x(1-x)dg_k(x) \\ &\le \beta c + \frac{1}{2(1-t_0)k}\int_c^{1-c} h(x)dx.\end{aligned}$$

By using the other half of Lemma 59, we may similarly conclude

$$\|\rho_k - \pi^*\|_{\mathrm{TV}} \le \beta c + \frac{1}{2t_0 k}\int_c^{1-c} h(x)dx.$$

The conclusion follows from combining these bounds, i.e. depending on if $t_0 \le \frac{1}{2}$ or $t_0 \ge \frac{1}{2}$. □

B.3.2 Mixweight functions

In this section, we propose a function h satisfying (B.7), and prove its correctness. First, we describe a useful sufficient condition.

Lemma 60. *Suppose $h : [c, 1-c] \to \mathbb{R}_{\geq 0}$ has $h(1-y) = h(y)$, and*

$$\int_c^{1-c} h(y)\Psi(y, S^c)dy \geq 2\pi^*(S)(1-\pi^*(S)),\ \forall S \subseteq \mathbb{R}^d : c \leq \pi^*(S) \leq \frac{1}{2}. \tag{B.9}$$

Then, h satisfies (B.7)*.*

Proof. Note that for $c \leq x \leq \frac{1}{2}$, choosing $S = m_k^{-1}([0, x])$ in (B.9) yields

$$2x(1-x) \leq \int_c^{1-c} h(y)\Psi\left(y, m_k^{-1}([x,1])\right) dy \leq \int_c^{1-c} h(y)q_k(x,y)dy.$$

The second inequality follows as for $x \geq y$, $q_k(x,y) \geq \Psi(y, m_k^{-1}([x,1]))$ by definition, and for $y \geq x$, we use symmetry of Ψ, q_k in their arguments. A similar argument holds for $x \geq \frac{1}{2}$ by symmetry. □

We now define our c-mixweight function h:

$$h(y) \stackrel{\text{def}}{=} \begin{cases} \max_{y \leq \pi^*(S) \leq \min\left\{2y, \frac{1}{2}\right\}} \frac{4\pi^*(S)}{\psi(S)} & y \leq \frac{1}{2} \\ h(1-y) & y \geq \frac{1}{2} \end{cases}. \tag{B.10}$$

In particular, note that for all $y \leq \frac{1}{2}$, by combining with Lemma 56,

$$h(y) \leq \max_{y \leq \pi^*(S) \leq 2y} \frac{16\pi^*(S)}{Q(S)^2}. \tag{B.11}$$

We will develop an upper bound on the ratio $\pi^*(S)/Q(S)^2$ for $\pi^*(S)$ which are "not too small" in the following section. We now prove correctness of the definition (B.10).

Lemma 61. *The function h defined in* (B.10) *satisfies* (B.7)*.*

Proof. Recall that it suffices to show that h satisfies (B.9). To this end, let S be some set such that $\pi^*(S) = x \in [2c, \frac{1}{2}]$. Then, recalling the definition of the spread $\psi(S)$ (B.3),

$$\begin{aligned} 2x &\leq \frac{4\pi^*(S)}{2\psi(S)} \left(\int_0^x \Psi(y, S^c)dy \right) \\ &\leq \frac{4\pi^*(S)}{\psi(S)} \int_{x/2}^x \Psi(y, S^c)dy \\ &\leq \int_{x/2}^x h(y)\Psi(y, S^c)h(y)dy \end{aligned}$$

$$\leq \int_c^{1-c} h(y)\Psi(y, S^c)dy.$$

In the second line, we used the monotonicity of $\Psi(\cdot, S^c)$ in the first argument; in the third line, we used the definition of h with the fact that $x \in [y, \min\{2y, \frac{1}{2}\}]$ for all $y \in [x/2, x]$. To handle the case $x \in [c, 2c]$,

$$\begin{aligned}\int_c^{1-c} h(y)\Psi(y, S^c)dy &\geq \int_x^{3x/2} h(y)\Psi(y, S^c)dy \\ &\geq \int_{x/2}^{x} h(y)\Psi(y, S^c)dy,\end{aligned}$$

where the second line is due to monotonicity, and the rest of the proof proceeds as before. □

Proof of Theorem 4. This follows from combining Theorem 16 with our particular choice of mixweight function given in (B.10), whose denominator we bound via (B.11). Because h is symmetric, it suffices to double the integration from c to $\frac{1}{2}$, and the bounds within the integral come from monotonicity of ϕ. □

B.3.3 Restricted conductance via total variation bounds

For $S \subseteq \mathbb{R}^d$ and $x \in \mathbb{R}^d$, we define $d(S, x) \stackrel{\text{def}}{=} \min_{y \in S} \|x - y\|$; for $S_1, S_2 \subseteq \mathbb{R}^d$, $d(S_1, S_2) \stackrel{\text{def}}{=} \min_{x \in S_2} d(S_1, x)$. The following isoperimetric inequality was given as Lemma 12 of [CDWY20].

Lemma 62 (Logarithmic isoperimetric inequality)**.** *Let π^* be any μ-strongly logconcave function. For any partition A_1, A_2, A_3 of $\mathbb{R}^d$ with $\pi^*(A_1) \leq \pi^*(A_2)$,*

$$\pi^*(A_3) \geq \frac{d(A_1, A_2)\sqrt{\mu}}{2}\pi^*(A_1)\log^{\frac{1}{2}}\left(1 + \frac{1}{\pi^*(A_1)}\right).$$

Lemma 63. *Let π^* be any μ-strongly logconcave function, and let $\delta\sqrt{\mu} < 1$ for some $\delta > 0$. Suppose for $\Omega \subset R^d$ with $\pi^*(\Omega) = 1 - s$, and all $x, y \in \Omega$ with $\|x - y\| \leq \delta$,*

$$\|\mathcal{T}_x - \mathcal{T}_y\|_{\text{TV}} \leq 1 - \alpha.$$

Then, for all $s \leq t \leq \frac{1}{2}$ and S with $\pi^*(S) = t$,

$$\frac{\pi^*(S)}{Q(S)^2} \leq \frac{16t}{\alpha^2}\left(\frac{\delta\sqrt{\mu}}{4}(t - s)\log^{\frac{1}{2}}\left(1 + \frac{1}{t}\right) - s\right)^{-2}.$$

In particular, if

$$s \leq \min\left(\frac{t}{2}, \frac{\delta\sqrt{\mu}t}{16}\sqrt{\log(3)}\right), \tag{B.12}$$

we have the simplified bound

$$\frac{\pi^*(S)}{Q(S)^2} \leq \frac{2^{16}}{\alpha^2\delta^2\mu t\log(1/t)}.$$

Proof. Let S have $\pi^*(S) = t$. Define the following three sets:

$$A_1 \stackrel{\text{def}}{=} \left\{x \in S \cap \Omega \text{mid} \mathcal{T}_x(S^c) < \frac{\alpha}{2}\right\},\ A_2 \stackrel{\text{def}}{=} \left\{x \in S^c \cap \Omega \text{mid} \mathcal{T}_x(S) < \frac{\alpha}{2}\right\},\ A_3 \stackrel{\text{def}}{=} (A_1 \cup A_2)^c.$$

Note that for any $x \in A_1$, $y \in A_2$, we have $\|\mathcal{T}_x - \mathcal{T}_y\|_{\text{TV}} > 1-\alpha$, and therefore $\|x-y\| > \delta$. Moreover, if $\pi^*(A_1) < \frac{1}{2}\pi^*(S)$,

$$Q(S) \geq \frac{\alpha}{4}(t-s).$$

Similarly, if $\pi^*(A_2) < \frac{1}{2}\pi^*(S^c \cap \Omega)$:

$$Q(S) \geq \frac{\alpha}{4}(1-t-s).$$

These bounds are subsumed by the third case, where $\pi^*(A_1) \geq \frac{1}{2}\pi^*(S)$, $\pi^*(A_2) \geq \frac{1}{2}\pi^*(S^c \cap \Omega)$. By Lemma 62, since we argued $d(A_1, A_2) > \delta$,

$$\begin{aligned}
\pi^*(A_3) &\geq \frac{\delta\sqrt{\mu}}{2}\min(\pi^*(A_1), \pi^*(A_2))\log^{\frac{1}{2}}\left(1 + \frac{1}{\min(\pi^*(A_1), \pi^*(A_2))}\right) \\
&\geq \frac{\delta\sqrt{\mu}}{4}\min\left(\pi^*(S\cap\Omega), \pi^*(S^c\cap\Omega)\right)\log^{\frac{1}{2}}\left(1 + \frac{1}{\min\left(\pi^*(S\cap\Omega), \pi^*(S^c\cap\Omega)\right)}\right) \\
&\geq \frac{\delta\sqrt{\mu}}{4}(t-s)\log^{\frac{1}{2}}\left(1+\frac{1}{t}\right).
\end{aligned}$$

This immediately implies

$$\pi^*(A_3 \cap \Omega) \geq \frac{\delta\sqrt{\mu}}{4}(t-s)\log^{\frac{1}{2}}\left(1+\frac{1}{t}\right) - s.$$

Finally, by the definition of stationary distribution,

$$\begin{aligned}
Q(S) &= \frac{1}{2}\left(\int_S \mathcal{T}_x(S^c)d\pi^*(x) + \int_{S^c}\mathcal{T}_x(S)d\pi^*(x)\right) \\
&\geq \frac{1}{2}\int_{A_3\cap\Omega}\frac{\alpha}{2}d\pi^*(x) = \frac{\alpha}{4}\pi^*(A_3\cap\Omega) \\
&\geq \frac{\alpha}{4}\left(\frac{\delta\sqrt{\mu}}{4}(t-s)\log^{\frac{1}{2}}\left(1+\frac{1}{t}\right) - s\right).
\end{aligned}$$

If (B.12) holds, we have the improved bound

$$Q(S) \geq \frac{\alpha\delta\sqrt{\mu}}{64}t\log^{\frac{1}{2}}\left(\frac{1}{t}\right).$$

□

B.3.4 Exponential convergence with a warm start

In this section, we give a simple reduction from a bound on the number of iterations it takes a Markov chain to attain constant total variation distance to the stationary distribution from a warm start, to the number of iterations it takes for the distance to decrease to ϵ, with logarithmic dependence on ϵ. Throughout, π^* is the stationary distribution of a Markov chain with transitions $\mathcal{T}$, and we let $\mathcal{T}^k\pi$ be the result of running k steps of the chain from starting distribution π. For specified π_0, we denote $\pi_k \stackrel{\text{def}}{=} \mathcal{T}^k\pi_0$. Sppose we have a bound of the following type.

Assumption 2. *$\exists T_{\text{mix}}$ such that for every π which is β/ϵ-warm with respect to π^*,*

$$\left\|\mathcal{T}^{T_{\text{mix}}}\pi - \pi^*\right\|_{\text{TV}} \leq \frac{1}{2e}.$$

We first recall some basic facts about the optimal *coupling* between two distributions π, ρ, which informally is the joint distribution μ with the prescribed marginals π and ρ which maximizes the probability that for $(x, y) \sim \mu$, $x = y$. For a reference, see [LPW09].

Fact 10. *Let μ be the optimal coupling between distributions π and ρ. The following hold.*

1. $\Pr_{(x,y)\sim\mu}[x \neq y] = \|\pi - \rho\|_{\text{TV}}$.

2. *Consider the marginal distribution of $(x, y) \sim \mu$ in the first variable, conditioned on $x \neq y$. It has a density proportional to $d\pi(x) - \min(d\pi(x), d\rho(x))$.*

The following result is well-known.

Lemma 64. *For any distribution π,*

$$\|\mathcal{T}\pi - \pi^*\|_{\text{TV}} \leq \|\pi - \pi^*\|_{\text{TV}}.$$

Proof. Consider the optimal coupling μ between π and π^*, and note that

$$\Pr_{(x,y)\sim\mu}[x \neq y] = \|\pi - \pi^*\|_{\text{TV}}.$$

It follows that the optimal coupling between μ' between $\mathcal{T}\pi$ and π^* has

$$\Pr_{(x,y)\sim\mu'}[x \neq y] \leq \Pr_{(x,y)=\mu}[x \neq y],$$

since $\mathcal{T}\pi^* = \pi^*$, and with probability $\Pr_{(x,y)\sim\mu}[x = y]$ the coupling μ' can keep x and y coupled. □

Lemma 65. *Under Assumption 2, letting π_0 be a β-warm start, and $k \geq T_{\mathrm{mix}} \log(\epsilon^{-1})$,*

$$\left\| \mathcal{T}^k \pi_0 - \pi^* \right\|_{\mathrm{TV}} \leq \epsilon.$$

Proof. Assume for the sake of contradiction that $\|\pi_k - \pi^*\|_{\mathrm{TV}} > \epsilon$; note that by Lemma 64, this implies that $\|\pi_i - \pi^*\|_{\mathrm{TV}} > \epsilon$ for all $i \leq k$. For any i, we denote μ_i to be the best coupling between π_i and π^*. Note that for any i, we can compute the marginal conditional distribution of the uncoupled set of π_i, under the coupling μ_i, by Fact 10:

$$\frac{d\tilde{\pi}_i}{d\pi^*}(x) \stackrel{\mathrm{def}}{=} \frac{\frac{d\pi_i}{d\pi^*}(x) - \min\left(\frac{d\pi_i}{d\pi^*}(x), 1\right)}{\int \left(\frac{d\pi_i}{d\pi^*}(x) - \min\left(\frac{d\pi_i}{d\pi^*}(x), 1\right)\right) d\pi^*(x)} \leq \frac{\frac{d\pi_i}{d\pi^*}(x)}{\|\pi_i - \pi^*\|_{\mathrm{TV}}} \leq \frac{\beta}{\epsilon}.$$

Here, we used the observation that if π_0 is β-warm, then so are all π_i for $i \geq 0$. Similarly, the conditional distribution of the uncoupled set of π^* under μ_i satisfies

$$\frac{d\tilde{\pi}_i^*}{d\pi^*}(x) \stackrel{\mathrm{def}}{=} \frac{1 - \min\left(\frac{d\pi_i}{d\pi^*}(x), 1\right)}{\int \left(1 - \min\left(\frac{d\pi_i}{d\pi^*}(x), 1\right)\right) d\pi^*(x)} \leq \frac{1}{\epsilon}.$$

This implies the conditional distributions $\tilde{\pi}_i$ and $\tilde{\pi}_i^*$ are both β/ϵ-warm with respect to π^* for any $i \leq k$. After T_{mix} iterations, the total variation distance between $\tilde{\pi}_i$ and $\tilde{\pi}_i^*$ is bounded by $1/e$ by Assumption 2 and the triangle inequality. Repeating this argument $\log(\epsilon^{-1})$ times implies that the measure of the uncoupled set decreases by at least a $1/e$ factor between iterations i and $i + T_{\mathrm{mix}}$, while $i \leq k$, so that the uncoupled set has measure at most ϵ by iteration k. Recalling that the measure of the uncoupled set is precisely the distance $\|\pi_k - \pi^*\|_{\mathrm{TV}}$ results in a contradiction. □

B.4 Total variation bounds

In this section, we prove the following lemma, which is the key step in lower bounding the conductance of one step of our algorithm.

Lemma 66. *For $\eta^2 \leq \frac{1}{20Ld \log \frac{\kappa}{\epsilon}}$, the Markov chain defined in Algorithm 3 satisfies*

$$\sup_{\|x-y\| \leq \eta} \|\mathcal{P}_x - \mathcal{P}_y\|_{\mathrm{TV}} \leq \frac{5}{8} \tag{B.13}$$

and, for Ω defined in (5.16)*,*

$$\sup_{x \in \Omega} \|\mathcal{P}_x - \mathcal{T}_x\|_{\mathrm{TV}} \leq \frac{1}{8}. \tag{B.14}$$

Proof. We first show (B.13). From any point $x \in \mathbb{R}^d$, let $\widetilde{x}$ be the proposed point according to Algorithm 3; we recall that the update is given by, for $v \sim \mathcal{N}(0, I_d)$,

$$\tilde{x} \leftarrow x + \eta v - \frac{\eta^2}{2}\nabla f(x) \Rightarrow \tilde{x} \sim \mathcal{N}\left(x - \frac{\eta^2}{2}\nabla f(x), \eta^2 I_d\right).$$

Therefore, recalling that the KL divergence d_{KL} between two Gaussians with covariance $\sigma^2 I_d$ and means μ_x, μ_y is $\|\mu_x - \mu_y\|^2 / 2\sigma^2$, Pinsker's inequality implies

$$\begin{aligned}\|\mathcal{P}_x - \mathcal{P}_y\|_{\mathrm{TV}} &\le \sqrt{\frac{1}{2} d_{KL}\left(\mathcal{P}_x, \mathcal{P}_y\right)} \\ &\le \frac{\left\|\left(x - \frac{\eta^2}{2}\nabla f(x)\right) - \left(y - \frac{\eta^2}{2}\nabla f(y)\right)\right\|}{2\eta} \\ &\le \frac{\left(1 + \frac{L\eta^2}{2}\right)\|x - y\|}{2\eta} \le \frac{5}{8},\end{aligned}$$

for $\|x - y\| \le \eta$ and $\eta^2 \le (2L)^{-1}$. The third line used the triangle inequality and ∇f is L-Lipschitz.

Next, we show (B.14). From a point x, and for any proposed transition to $\widetilde{x} \neq x$, the proposal $\mathcal{P}_x$ places at least as much mass on $\widetilde{x}$ as $\mathcal{T}_x$, because the rejection probability is nonnegative; consequently, the set A maximizing $\mathcal{T}_x(A) - \mathcal{P}_x(A)$ is the singleton $A = \{x\}$, and the total variation distance is simply the probability $\mathcal{T}_x = x$, or

$$\begin{aligned}\|\mathcal{P}_x - \mathcal{T}_x\|_{\mathrm{TV}} &= 1 - \mathbb{E}_{v \sim \mathcal{N}(0, I_d)}\left[\min\left\{1, \exp\left(\mathrm{Ham}(x, v) - \mathrm{Ham}(\widetilde{x}, \widetilde{v})\right)\right\}\right] \\ &\le 1 - \mathbb{E}_{v \sim \mathcal{N}(0, I_d)}\left[\exp\left(\mathrm{Ham}(x, v) - \mathrm{Ham}(\widetilde{x}, \widetilde{v})\right)\right].\end{aligned}$$

Therefore, to show the desired $\|\mathcal{P}_x - \mathcal{T}_x\|_{\mathrm{TV}} \le 1/8$, it suffices to show that

$$\mathbb{E}_{v \sim \mathcal{N}(0, I_d)}\left[\exp\left(\mathrm{Ham}(x, v) - \mathrm{Ham}(\widetilde{x}, \widetilde{v})\right)\right] \ge \frac{7}{8}.$$

By the calculation

$$\frac{15}{16} \cdot \exp\left(-\frac{1}{16}\right) \ge \frac{7}{8},$$

it suffices to show that with probability $15/16$ over the randomness of v, $\mathrm{Ham}(x, v) - \mathrm{Ham}(\widetilde{x}, \widetilde{v}) \ge -1/16$. First, by a standard tail bound on the chi-squared distribution (Lemma 1 of [LM00]), we have

$$\Pr\left[\|v\|^2 \ge d + 2\sqrt{3d} + 6\right] \le \exp(-3) \le \frac{1}{16}.$$

Thus, assuming d is at least a sufficiently large constant, with probability at least $1/16$ over the randomness of v, we have $\|v\| \le \sqrt{2d}$. Finally, the conclusion follows from the claim

$$\mathrm{Ham}(\widetilde{x}, \widetilde{v}) - \mathrm{Ham}(x, v) \le \frac{1}{16}, \ \forall x \in \Omega, \ \|v\| \le \sqrt{2d},$$

which we now show. Recalling $\tilde{v} = v - \frac{\eta}{2}\left(\nabla f(\tilde{x}) + \nabla f(x)\right)$ and $\tilde{x} = x + \eta v - \frac{\eta^2}{2}\nabla f(x)$,

$$\begin{aligned}
\mathcal{H}\left(\tilde{x},\tilde{v}\right) - \mathcal{H}\left(x,v\right) &= -\frac{1}{2}\left\|v\right\|^2 + \frac{1}{2}\left\|\tilde{v}\right\|^2 - f(x) + f(\tilde{x}) \\
&\le -\frac{1}{2}\left\|v\right\|^2 + \frac{1}{2}\left\|\widetilde{v}\right\|^2 + \frac{1}{2}\left\langle \nabla f(\widetilde{x}) + \nabla f(x), \widetilde{x} - x\right\rangle + \frac{L}{4}\left\|\tilde{x} - x\right\|^2 \\
&= \frac{1}{2}\left\|v - \frac{\eta}{2}(\nabla f(x) + \nabla f(\widetilde{x}))\right\|^2 - \frac{1}{2}\left\|v\right\|^2 \\
&+ \frac{1}{2}\left\langle \nabla f(\widetilde{x}) + \nabla f(x), \eta v - \frac{\eta^2}{2}\nabla f(x)\right\rangle + \frac{L}{4}\left\|\tilde{x} - x\right\|^2 \\
&= -\frac{\eta}{2}\left\langle \nabla f(x) + \nabla f(\widetilde{x}), v\right\rangle + \frac{\eta^2}{8}\left\|\nabla f(x) + \nabla f(\widetilde{x})\right\|^2 \\
&+ \frac{1}{2}\left\langle \nabla f(\widetilde{x}) + \nabla f(x), \eta v - \frac{\eta^2}{2}\nabla f(x)\right\rangle + \frac{L}{4}\left\|\tilde{x} - x\right\|^2 \\
&= \frac{\eta^2}{8}\left\langle \nabla f(x) + \nabla f(\widetilde{x}), \nabla f(\widetilde{x}) - \nabla f(x)\right\rangle + \frac{L}{4}\left\|\tilde{x} - x\right\|^2 \\
&\le \frac{\eta^2 L}{8}\left\|x - \widetilde{x}\right\|\left\|\nabla f(x) + \nabla f(\widetilde{x})\right\| + \frac{L}{4}\left\|x - \widetilde{x}\right\|^2.
\end{aligned}$$

The second inequality followed from

$$f(\widetilde{x}) - f(x) \le \min\left(\left\langle \nabla f(\widetilde{x}), \widetilde{x} - x\right\rangle,\ \left\langle \nabla f(x), \widetilde{x} - x\right\rangle + \frac{L}{2}\left\|\widetilde{x} - x\right\|^2\right),$$

due to convexity and smoothness; the last inequality followed from smoothness and Cauchy-Schwarz, and every other line was by expanding the definitions. We now bound these two terms. First, since smoothness implies $\left\|\nabla f(x) + \nabla f(\widetilde{x})\right\| \le 2\left\|\nabla f(x)\right\| + L\left\|x - \widetilde{x}\right\|$,

$$\begin{aligned}
\frac{\eta^2 L}{8}\left\|x - \widetilde{x}\right\|\left\|\nabla f(x) + \nabla f(\widetilde{x})\right\| &\le \frac{\eta^2 L^2}{8}\left\|x - \widetilde{x}\right\|^2 + \frac{\eta^2 L}{4}\left\|\nabla f(x)\right\|\left\|x - \widetilde{x}\right\| \\
&\le \frac{L}{4}\left\|x - \widetilde{x}\right\|^2 + \frac{\eta^2 L}{4}\left\|\nabla f(x)\right\|\left\|x - \widetilde{x}\right\|
\end{aligned}$$

Here we used our choice of η. Next, since $\widetilde{x} - x = \eta v - \frac{\eta^2}{2}\nabla f(x)$, using the above bounds,

$$\begin{aligned}
\mathrm{Ham}(\widetilde{x},\widetilde{v}) - \mathrm{Ham}(x,v) &\le \frac{\eta^2 L}{8}\left\|x - \widetilde{x}\right\|\left\|\nabla f(x) + \nabla f(\widetilde{x})\right\| + \frac{L}{4}\left\|x - \widetilde{x}\right\|^2 \\
&\le \frac{L}{2}\left\|x - \widetilde{x}\right\|^2 + \frac{\eta^2 L}{4}\left\|\nabla f(x)\right\|\left\|x - \widetilde{x}\right\| \\
&\le L\eta^2\left\|v\right\|^2 + \frac{L\eta^4}{4}\left\|\nabla f(x)\right\|^2 + \frac{L\eta^3}{4}\left\|\nabla f(x)\right\|\left\|v\right\| + \frac{L\eta^4}{8}\left\|\nabla f(x)\right\|^2 \\
&\le \frac{9L\eta^2}{8}\left\|v\right\|^2 + \frac{L\eta^4}{2}\left\|\nabla f(x)\right\|^2 \le \frac{1}{16}.
\end{aligned}$$

We recalled $\left\|v\right\|^2 \le 2d$, $\left\|\nabla f(x)\right\|^2 \le 25Ld^2\log^2\frac{\kappa}{\epsilon}$, and the choice of η. □

Finally, we note that Lemma 4 immediately follows via the triangle inequality.

B.5 Deferred proofs

Lemma 67. *For any $C < 1$,*

$$\prod_{k=0}^{\infty} \left(\frac{1}{1 - \frac{C}{4^k}} \right)^{2^k} \leq \frac{1 + \sqrt{C}}{1 - \sqrt{C}}.$$

Proof. Define

$$V(C) \stackrel{\text{def}}{=} \prod_{k=1}^{\infty} \left(\frac{1}{1 - \frac{C}{4^k}} \right)^{2^k} \leq \frac{1 + \sqrt{C}}{1 - \sqrt{C}},$$

so we wish to bound $V(C)/(1 - C)$. It suffices to show $V(C) \leq (1 + \sqrt{C})^2$. Note that

$$\log V(C) = \sum_{k=1}^{\infty} 2^k \log \left(\frac{1}{1 - \frac{C}{4^k}} \right) = \sum_{k=1}^{\infty} 2^k \sum_{j=1}^{\infty} \frac{1}{j} \left(\frac{C}{4^k} \right)^j = \sum_{j=1}^{\infty} \frac{C^j}{j(2^{2j-1} - 1)}.$$

Thus, $\log V$ is a convex function in C. Note that $\log V(0) = 0$ and

$$\log V(1) \leq 1 + \sum_{j=2}^{\infty} \frac{1}{4^{j-1} j} = 1 + 4 \left(-\log \left(\frac{3}{4} \right) - \frac{1}{4} \right) \leq \log 4.$$

This implies $\log V(C) \leq C \log 4$, and the conclusion follows from $4^C \leq (1 + \sqrt{C})^2$ for $C \in [0, 1]$. □

Appendix C

DEFERRED CONTENTS FROM CHAPTER 4

C.1 Necessity of fixing a scale

We give a simple argument showing if the step size η of the HMC algorithm does not depend on the "scale" of the problem, namely the eigenvalues of the function Hessian (as opposed to scale-invariant quantities, e.g. the condition number κ and the dimension), then the task of proving lower bounds becomes much more trivial. In particular, we can adaptively pick a scale of the problem in response to the fixed η. This justifies the additional requirement in Theorems 6, 7, and 9 of the fixed scale $[1, \kappa]$, which we remark is a *strengthening* of an analogous scale-free lower bound.

Concretely, suppose we wished to prove the statement of Theorem 9 but only on functions with condition number κ (without specifying a range of eigenvalues). Then, for fixed η, K, consider

$$f(x) = \frac{\lambda}{2}x^2, \text{ where } \lambda \stackrel{\text{def}}{=} \frac{2\left(1 - \cos\left(\frac{\pi}{K}\right)\right)}{\eta^2}.$$

Clearly, $f : \mathbb{R} \to \mathbb{R}$ has condition number $1 \leq \kappa$ for any κ. Then, the proof of Proposition 5 applies to show that the HMC Markov chain cannot leave any symmetric set, because the coefficients encounter extremal points or zeroes of the Chebyshev polynomials.

C.2 HMC lower bounds beyond $\kappa\sqrt{d}$

Here, we analyze the behavior of HMC on the hard function (4.20). We will use this construction to demonstrate that when the number of steps K is small, we cannot improve either the relaxation time (Section C.2.1) or the mixing time (Section C.2.2) of MALA by more than roughly a $O(K)$ factor.

C.2.1 Relaxation time lower bound for small K

We first give a bound on the acceptance probability (4.6) for general HMC Markov chain. We expand the term $-\text{Ham}(x_K, v_K) + \text{Ham}(x_0, v_0)$ and extend the result given by Lemma 17.

Lemma 68. *For the iterates given by Fact 4, write* $\widetilde{x}_j \stackrel{\text{def}}{=} x_0 + \eta j v_0$ *for* $0 \le j \le K-1$. *Then, for a* κ*-smooth function* f,

$$-Ham(x_K, v_K) + Ham(x_0, v_0) \le \sum_{j=0}^{K-1} \left(-f(\widetilde{x}_{j+1}) + f(\widetilde{x}_j) + \frac{1}{2} \langle \eta v_0, \nabla f(\widetilde{x}_{j+1}) + \nabla f(\widetilde{x}_j) \rangle \right)$$
$$+\eta K \|v_0\|_2 \max_{0 \le j \le K} \|\nabla f(x_j) - \nabla f(\widetilde{x}_j)\|_2 + \frac{1}{2}\eta^2 K^2 \max_{0 \le j_1, j_2 \le K} \|\nabla f(\widetilde{x}_K) - \nabla f(x_{j_2})\|_2 \|\nabla f(x_{j_1})\|_2$$
$$+\frac{1}{2}\eta^2 K^2 \max_{0 \le j_1, j_2, j_3 \le K} \|\nabla f(x_{j_3})\|_2 \|\nabla f(x_{j_1}) - \nabla f(x_{j_2})\|_2 .$$

Proof. Expanding $\mathrm{Ham}(x_0, v_0) - \mathrm{Ham}(x_K, v_K)$ according to the definition of Ham, x_K and v_K,

$$\begin{aligned}
&\mathrm{Ham}(x_0, v_0) - \mathrm{Ham}(x_K, v_K) \\
=& -f(x_K) + f(x_0) - \frac{\left\| v_0 - \frac{\eta}{2}\nabla f(x_0) - \eta \sum_{j=1}^{K-1} \nabla f(x_j) - \frac{\eta}{2}\nabla f(x_K) \right\|_2^2}{2} + \frac{\|v_0\|_2^2}{2} \\
=& -f(x_K) + f(\widetilde{x}_K) - f(\widetilde{x}_K) + f(x_0) + \left\langle v_0, \frac{\eta}{2}\nabla f(x_0) + \eta \sum_{j=1}^{K-1} \nabla f(x_j) + \frac{\eta}{2}\nabla f(x_K) \right\rangle \\
&- \frac{1}{2} \left\| \frac{\eta}{2}\nabla f(x_0) + \eta \sum_{j=1}^{K-1} \nabla f(x_j) + \frac{\eta}{2}\nabla f(x_K) \right\|_2^2 \\
=& -f(x_K) + f(\widetilde{x}_K) + \sum_{j=0}^{K-1} (-f(\widetilde{x}_{j+1}) + f(\widetilde{x}_j)) + \left\langle \eta v_0, \frac{1}{2}\nabla f(\widetilde{x}_0) + \sum_{j=1}^{K-1} \nabla f(\widetilde{x}_j) + \frac{1}{2}\nabla f(\widetilde{x}_K) \right\rangle \\
&- \frac{1}{2} \left\| \frac{\eta}{2}\nabla f(x_0) + \eta \sum_{j=1}^{K-1} \nabla f(x_j) + \frac{\eta}{2}\nabla f(x_K) \right\|_2^2 \\
&+ \left\langle \eta v_0, \left(\frac{1}{2}\nabla f(x_0) + \sum_{j=1}^{K-1} \nabla f(x_j) + \frac{1}{2}\nabla f(x_K) \right) - \left(\frac{1}{2}\nabla f(\widetilde{x}_0) + \sum_{j=1}^{K-1} \nabla f(\widetilde{x}_j) + \frac{1}{2}\nabla f(\widetilde{x}_K) \right) \right\rangle \\
=& \sum_{j=0}^{K-1} \left(-f(\widetilde{x}_{j+1}) + f(\widetilde{x}_j) + \frac{1}{2} \langle \eta v_0, \nabla f(\widetilde{x}_{j+1}) + \nabla f(\widetilde{x}_j) \rangle \right) \\
&- f(x_K) + f(\widetilde{x}_K) - \frac{1}{2} \left\| \frac{\eta}{2}\nabla f(x_0) + \eta \sum_{j=1}^{K-1} \nabla f(x_j) + \frac{\eta}{2}\nabla f(x_K) \right\|_2^2 \\
&+ \left\langle \eta v_0, \left(\frac{1}{2}\nabla f(x_0) + \sum_{j=1}^{K-1} \nabla f(x_j) + \frac{1}{2}\nabla f(x_K) \right) - \left(\frac{1}{2}\nabla f(\widetilde{x}_0) + \sum_{j=1}^{K-1} \nabla f(\widetilde{x}_j) + \frac{1}{2}\nabla f(\widetilde{x}_K) \right) \right\rangle .
\end{aligned} \tag{C.1}$$

Now we bound the last two lines in the decomposition (C.1). For the second-to-last line

of (C.1), by convexity of f and the Cauchy-Schwarz inequality,

$$\begin{aligned}
&- f(x_K) + f(\widetilde{x}_K) - \frac{1}{2}\left\|\frac{\eta}{2}\nabla f(x_0) + \eta\sum_{j=1}^{K-1}\nabla f(x_j) - \frac{\eta}{2}\nabla f(x_K)\right\|_2^2 \\
\le &\left\langle \nabla f(\widetilde{x}_K), \frac{1}{2}K\eta^2\nabla f(x_0) + \eta^2\sum_{j=1}^{K-1}(K-j)\nabla f(x_j)\right\rangle - \frac{1}{2}\left\|\frac{\eta}{2}\nabla f(x_0) + \eta\sum_{j=1}^{K-1}\nabla f(x_j) + \frac{\eta}{2}\nabla f(x_K)\right\|_2^2 \\
\le &\frac{1}{2}\eta^2K^2\max_{0\le j_1,j_2,j_3\le K}\left(\nabla f(\widetilde{x}_K)^\top\nabla f(x_{j_1}) - \nabla f(x_{j_2})^\top\nabla f(x_{j_3})\right) \\
\le &\frac{1}{2}\eta^2K^2\left(\max_{0\le j_1,j_2\le K}\|\nabla f(\widetilde{x}_K) - \nabla f(x_{j_2})\|_2\,\|\nabla f(x_{j_1})\|_2 + \max_{0\le j_1,j_2,j_3\le K}\|\nabla f(x_{j_3})\|_2\,\|\nabla f(x_{j_1}) - \nabla f(x_{j_2})\|_2\right).
\end{aligned} \tag{C.2}$$

In the third line above, we used that the total "number of gradient inner products" for both terms is $\frac{1}{2}\eta^2K^2$, and took the largest such inner product difference.

Finally, for the last line of (C.1), by the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left\langle \eta v_0, \left(\frac{1}{2}\nabla f(x_0) + \sum_{j=1}^{K-1}\nabla f(x_j) + \frac{1}{2}\nabla f(x_K)\right) - \left(\frac{1}{2}\nabla f(\widetilde{x}_0) + \sum_{j=1}^{K-1}\nabla f(\widetilde{x}_j) + \frac{1}{2}\nabla f(\widetilde{x}_K)\right)\right\rangle \\
\le &\eta K\,\|v_0\|_2\max_{0\le j\le K}\|\nabla f(x_j) - \nabla f(\widetilde{x}_j)\|_2\,.
\end{aligned} \tag{C.3}$$

Combining (C.1), (C.2) and (C.3) proves the desired claim. □

We define a hard function $f_{\text{hard}} : \mathbb{R}^d \to \mathbb{R}$ that is κ-smooth and 1-strongly convex (note it is the same hard function as in Section 4.6, under the change of variable $h = \frac{\eta^2}{2}$). We will show it is hard to sample from the density proportional to $\exp(-f_{\text{hard}})$ when K is small.

$$f_{\text{hard}}(x) \stackrel{\text{def}}{=} \sum_{i\in[d]} f_i(x_i),\ \text{where } f_i(c) = \begin{cases} \frac{1}{2}c^2 & i = 1 \\ \frac{\kappa}{3}c^2 - \frac{\kappa\eta^2}{6}\cos\left(\frac{\sqrt{2}c}{\eta}\right) & 2 \le i \le d \end{cases}. \tag{C.4}$$

Lemma 69. *For $\eta^2 \le 1$, let $\widetilde{x}_j \stackrel{\text{def}}{=} x_0 + \eta j v_0$ for $0 \le j \le K-1$ and $v_0 \sim \mathcal{N}(0, \mathrm{id})$. Let $R^{(j)}$ be the random variable with given by $R^{(j)} = \sum_{i=1}^d R_i^{(j)}$ where*

$$R_i^{(j)} = -f_i([\widetilde{x}_{j+1}]_i) + f_i([\widetilde{x}_j]_i) + \frac{1}{2}\eta[v_0]_i\cdot(\nabla f_i([\widetilde{x}_{j+1}]_i) + \nabla f_i([\widetilde{x}_j]_i)).$$

Then,

$$\mathbb{E}_{v_0\sim\mathcal{N}(0,1)}\left[\sum_{j=0}^{K-1}R^{(j)}\right] \le -0.02\kappa\eta^2\sum_{i=2}^d\cos\frac{\sqrt{2}[x_0]_i}{\eta}. \tag{C.5}$$

and

$$\Pr\left[\sum_{j=0}^{K-1} R^{(j)} - \mathbb{E}\left[\sum_{j=0}^{K-1} R^{(j)}\right] \geq 10\eta^2 K\kappa\sqrt{d\log d}\right] \leq \frac{1}{d^5}. \tag{C.6}$$

Proof. In this proof, all expectations $\mathbb{E}$ are taken over $v_0 \sim \mathcal{N}(0, \mathrm{id})$, so we omit them. For $i = 1$,

$$\begin{aligned}\mathbb{E}\left[\sum_{j=0}^{K-1} R_i^{(j)}\right] &= \mathbb{E}\left[-\frac{1}{2}([x_0]_1 + \eta K[v_0]_1)^2 + \frac{1}{2}[x_0]_1^2 + \frac{1}{2}\sum_{j=0}^{K-1}\eta[v_0]_1(2[x_0]_1 + \eta(2j+1)[v_0]_1)\right] \\ &= \mathbb{E}\left[-\frac{1}{2}[x_0]_1^2 - \frac{1}{2}\eta^2K^2[v_0]_1^2 - \eta K[x_0]_1[v_0]_1 + \frac{1}{2}[x_0]_1^2 + \frac{1}{2}\eta^2K^2[v_0]_1^2 + \eta K[x_0]_1[v_0]_1\right] = 0.\end{aligned}$$

We bound each coordinate $2 \leq i \leq d$ separately.

$$\begin{aligned}&\mathbb{E}\left[\sum_{j=0}^{K-1} R_i^{(j)}\right] \\ =&\mathbb{E}\left[\sum_{j=0}^{K-1} -f_i([\widetilde{x}_{j+1}]_i) + f_i([\widetilde{x}_j]_i) + \frac{1}{2}\eta[v_0]_i\cdot(\nabla f_i([\widetilde{x}_{j+1}]_i) + \nabla f_i([\widetilde{x}_j]_i))\right] \\ =&-\frac{\kappa}{3}\mathbb{E}\left[([x_0]_i + \eta K[v_0]_i)^2 - [x_0]_i^2\right] + \frac{1}{3}\eta\kappa\mathbb{E}\left[[v_0]_i\cdot\left(2[x_0]_i + \eta\sum_{j=0}^{K-1}(2j+1)[v_0]_i\right)\right] \\ &+\frac{\kappa\eta^2}{6}\mathbb{E}\left[\sum_{j=0}^{K-1}\cos\frac{\sqrt{2}\,([x_0]_i + \eta(j+1)[v_0]_i)}{\eta} - \cos\frac{\sqrt{2}\,([x_0]_i + \eta j[v_0]_i)}{\eta}\right] \\ &+\frac{\sqrt{2}\eta^2\kappa}{12}\mathbb{E}\left[[v_0]_i\sum_{j=0}^{K-1}\left(\sin\frac{\sqrt{2}\,([x_0]_i + \eta j[v_0]_i)}{\eta} + \sin\frac{\sqrt{2}\,([x_0]_i + \eta(j+1)[v_0]_i)}{\eta}\right)\right] \\ =&-\frac{\kappa\eta^2}{6}\sum_{j=0}^{K-1}\exp(-j^2) - \exp(-(j+1)^2) - j\exp(-j^2) - (j+1)\exp(-(j+1)^2)\cos\frac{\sqrt{2}[x_0]_i}{\eta}\end{aligned}$$

The last line used the computation

$$\mathbb{E}\left[[v_0]_i\sin\frac{\sqrt{2}\,([x_0]_i + \eta j[v_0]_i)}{\eta}\right] = \sqrt{2}j\exp(-j^2)\cos\frac{\sqrt{2}[x_0]_i}{\eta},$$
$$\mathbb{E}\left[\cos\frac{\sqrt{2}\,([x_0]_i + \eta j[v_0]_i)}{\eta}\right] = \exp(-j^2)\cos\frac{\sqrt{2}[x_0]_i}{\eta}.$$

Next, we bound $\sum_{j=0}^{K-1}\left(\exp(-j^2) - \exp(-(j+1)^2) - j\exp(-j^2) - (j+1)\exp(-(j+1)^2)\right)$. For $j = 0$, $1 - \frac{2}{\exp(1)} \geq 0.264$. For $j = 1$, the negative terms have $-3\exp(-4) \geq -0.06$,

and the positive terms can only help this inequality. For the remaining terms,

$$\sum_{j=2}^{K-1}\left(\exp(-j^2)-\exp(-(j+1)^2)-j\exp(-j^2)-(j+1)\exp(-(j+1)^2)\right)$$
$$\geq \sum_{j=2}^{K-1}\left(-j\exp(-j^2)-(j+1)\exp(-(j+1)^2)\right)$$
$$\geq -2\sum_{j=2}^{K}\left(j\exp(-j^2)\right) \geq -2\frac{2}{\exp(4)}\frac{1}{1-2\exp(-5)} \geq -0.075.$$

The last inequality used the ratio between two consecutive terms is bounded by $\frac{j+1}{j}\exp(j^2-(j+1)^2) \leq 2\exp(-5)$. Summing over d coordinates proves (C.5).

Next, we prove the concentration property of $\sum_{j=0}^{K-1} R^{(j)}$. Let $\widetilde{x}_{j,s} = \widetilde{x}_j + s\eta v_0$, for $s \in [0,1]$ and $j = 0, ..., K-1$. By Lemma 11, we have

$$\sum_{j=0}^{K-1} R^{(j)} = \sum_{j=0}^{K-1} -\eta^2 \int_0^1 \left(\frac{1}{2}-s\right) v_0^\top \nabla^2 f(\widetilde{x}_{j,s}) v_0 ds.$$

For coordinate $1 \leq i \leq d$, $\left|\eta^2 \int_0^1 \left(\frac{1}{2}-s\right) f_i''([x_{j,s}]_i)ds\right| \leq \frac{\eta^2\kappa}{2}$ by smoothness. Then, the random variables $\sum_{j=0}^{K-1} R_i^{(j)} - \mathbb{E}\left[\sum_{j=0}^{K-1} R_i^{(j)}\right]$ for $1 \leq i \leq d$ are sub-exponential with parameter $\frac{\eta^2\kappa K}{2}$(for coordinates where the coefficient is negative, note the negation of a sub-exponential random variable is still sub-exponential). Hence, by Fact 3,

$$\Pr\left[\sum_{i\in[d]}\left(\sum_{k=0}^{K-1} R_i^{(j)} - \mathbb{E}\left[\sum_{k=0}^{K-1} R_i^{(j)}\right]\right) \geq 10\eta^2 K\kappa\sqrt{d\log d}\right] \leq \frac{1}{d^5}.$$

□

Now, we build a bad set Ω_{hard} with lower bounded measure that starting from a point $x_0 \in \Omega_{\text{hard}}$, such that with high probability, $-\mathbb{E}\left[\sum_{j=0}^{K-1} R^{(j)}\right]$ is very negative. Let $h = \frac{1}{2}\eta^2$ so that we may use the results from Section 4.4. We use the bad set Ω_{hard} defined in (4.21).

$$\Omega_{\text{hard}} = \Bigg\{ x \;\Big|\; |x_1| \leq 2, \forall 2 \leq i \leq d, \exists k_i \in \mathbb{Z}, |k_i| \leq \left\lfloor \frac{5}{\pi\sqrt{h\kappa}} \right\rfloor, \text{ such that}$$
$$-\frac{9}{20}\pi\sqrt{h} + 2\pi k_i\sqrt{h} \leq x_i \leq \frac{9}{20}\pi\sqrt{h} + 2\pi k_i\sqrt{h} \Bigg\}.$$

We restate Lemma 14 here, which lower bounds $\pi^*(\Omega_{\text{hard}})$ and bounds $\|\nabla f(x)\|_2$ for $x \in \Omega_{\text{hard}}$.

Lemma 14. *Let $h \leq \frac{1}{10000\pi^2\kappa}$. Let π^* have log-density $-f_{\text{hard}}$ (4.20). Then, $\pi^*(\Omega_{\text{hard}}) \geq \exp(-d)$. Moreover, for all $x \in \Omega_{\text{hard}}$, $\|\nabla f(x)\|_2 \leq 10\sqrt{\kappa d}$.*

We can further show the following, which is used to bound the remaining terms in Lemma 68.

Lemma 70. *Let $x_0 \in \Omega_{\text{hard}}$, $\eta K \le \frac{1}{100\sqrt{\kappa}\log d}$ and $d \ge 8$. Let let x_j for $1 \le j \le K-1$ be given by the iterates in Fact 4 and $\widetilde{x}_K = x_0 + \eta K v_0$. Then, with probability at least $1 - \frac{1}{d^5}$ over random $v_0 \sim \mathcal{N}(0, \mathrm{id})$, $\|v_0\|_2 \le 4\sqrt{d}\log d$ and for all $0 \le j \le K$, $\|\nabla f(x_j)\|_2 \le 11\sqrt{\kappa d}$ and $\|\nabla f(\widetilde{x}_K)\|_2 \le 11\sqrt{\kappa d}$.*

Proof. We first derive a bound on $v_0 \sim \mathcal{N}(0, \mathrm{id})$. By a standard Gaussian tail bound, for $d \ge 8$, with probability at least $1 - \frac{1}{d^5}$, $|[v_0]_i| \le 4\log d$ for all $1 \le i \le d$. Then, $\|v_0\|_2 \le \sqrt{16d(\log d)^2} = 4\sqrt{d}\log d$. Now, we prove the bound on $\|x_j - x_0\|_2$ and $\|\nabla f(x_j)\|_2$ using induction. First, $\|\nabla f(x_0)\| \le 11\sqrt{d\kappa}$ holds by Lemma 14. Assume for induction $\|\nabla f(x_k)\|_2 \le 11\sqrt{d\kappa}$ for $1 \le k < j$. Then,

$$\begin{aligned}\|x_j - x_0\|_2 &\le \left\| \eta j v_0 - \frac{\eta^2 j}{2}\nabla f(x_0) - \eta^2 \sum_{k=1}^{j-1}(j-k)\nabla f(x_k) \right\|_2 \\ &\le 4\eta j\sqrt{d}\log d + \eta^2 j^2 \cdot 11\sqrt{\kappa d} \le \sqrt{\frac{d}{\kappa}}.\end{aligned}$$

The last inequality used the assumption $\eta K \le \frac{1}{100\sqrt{\kappa}\log d}$. Since f is κ-smooth, we have

$$\|\nabla f(x_j)\|_2 \le \|\nabla f(x_0)\|_2 + \kappa \|x_j - x_0\|_2 \le 10\sqrt{\kappa d} + \kappa\sqrt{\frac{d}{\kappa}} \le 11\sqrt{\kappa d}.$$

This completes the induction step. Finally, we have

$$\|\nabla f(\widetilde{x}_K)\|_2 \le \|\nabla f(x_0)\|_2 + \kappa \|\eta K v_0\|_2 \le 10\sqrt{\kappa d} + 4\eta K \kappa\sqrt{d}\log d \le 11\sqrt{\kappa d},$$

where we used $\eta K \le \frac{1}{100\sqrt{\kappa}\log d}$. □

Lemma 71. *Let η and K satisfy $K \le \frac{\sqrt{d}}{10000\sqrt{\log d}}$, and $\eta K^3 \le \frac{1}{100000\sqrt{\kappa}\log d}$. For any $x_0 \in \Omega_{\text{hard}}$, let (x_K, v_K) be given by the iterates in Fact 4 and $v_0 \sim \mathcal{N}(0, \mathrm{id})$. With probability at least $1 - \frac{2}{d^5}$,*

$$-Ham(x_K, v_K) + Ham(x_0, v_0) \le -\Omega\left(\eta^2 \kappa d\right).$$

Proof. We first remark that the bound on ηK^3 implies we may apply Lemma 14 and Lemma 70. Next, for $x_0 \in \Omega_{\text{hard}}$, $\cos\frac{\sqrt{2}[x_0]_i}{\eta}$ is bounded away from 0 for all $2 \le i \le d$. By Lemma 69, when $K \le \frac{\sqrt{d}}{10000\sqrt{\log d}}$, with probability at least $1 - \frac{1}{d^5}$, $\sum_{j=0}^{K-1} R^{(j)} \le$

$-0.002\eta^2\kappa d$ (the expectation term dominates). By Lemma 70, with probability at least $1 - \frac{1}{d^5}$, the other terms in Lemma 68 have

$$\begin{aligned}
\eta K \|v_0\|_2 \max_{0\le j\le K} \|\nabla f(x_j) - \nabla f(\widetilde{x}_j)\|_2 + \frac{1}{2}\eta^2 K^2 \max_{0\le j_1,j_2\le K} \|\nabla f(\widetilde{x}_K) - \nabla f(x_{j_2})\|_2 \|\nabla f(x_{j_1})\|_2 \\
+\frac{1}{2}\eta^2 K^2 \max_{0\le j_1,j_2,j_3\le K} \|\nabla f(x_{j_3})\|_2 \|\nabla f(x_{j_1}) - \nabla f(x_{j_2})\|_2 \\
\le 4\eta K\sqrt{d}\log d\cdot\kappa\eta^2\left(K\|\nabla f(x_0)\|_2 + \sum_{j\in[K-1]}(K-j)\|\nabla f(x_j)\|_2\right) \\
+\eta^2K^2\cdot 11\sqrt{\kappa d}\cdot\kappa\left(\eta K\|v_0\|_2 + \eta^2 K\|\nabla f(x_0)\|_2 + \eta^2\sum_{j\in[K-1]}(K-j)\|\nabla f(x_j)\|_2\right) \\
\le 44\eta^3K^3\kappa^{1.5}d\log d + 44\eta^3K^3\kappa^{1.5}d\log d + 121\eta^4K^4\kappa^2 d \le 0.001\eta^2\kappa d.
\end{aligned}$$

The last inequality used the assumption $\eta \le \frac{1}{100000K^3\sqrt{\kappa}\log d}$. Combining the above bounds with Lemma 68 yields the claim. □

Proposition 16. *For $\eta^2K = O\left(\frac{\sqrt{\log d}}{\kappa\sqrt{d}}\right)$ and $K = O\left(d^{0.099}\right)$, there is a target density on $\mathbb{R}^d$ whose negative log-density is κ smooth, such that relaxation time of HMC is $\Omega\left(\frac{\kappa d}{K^2}\right)$.*

Proof. It is straightforward to check that such a range of η and K satisfies the assumptions of Lemma 71. Applying Lemma 71 with the hard function f_{hard}, the remainder of the proof follows analogously to that of Theorem 9. □

We give a brief discussion of the implications of Proposition 16. For $\eta^2K = \omega(\frac{\sqrt{\log d}}{\kappa\sqrt{d}})$, the proof of Theorem 9 rules out a polynomial relaxation time. In the remaining range, Proposition 16 implies that for small $K = O\left(d^{0.099}\right)$, the most we can improve the relaxation time of MALA (Theorem 7) by taking multiple steps in HMC is by a K^2 factor. Since each iteration takes K gradients, this is roughly an improvement of K in the query complexity, and strengthens Theorem 9 for small K.

C.2.2 Mixing time lower bound for small K

In this section, we first use prior results to narrow down the range of η we consider (assuming K is small). We then generalize the ideas of Section 4.5, our MALA mixing lower bound, to this setting.

Mixing time lower bound for large η. Suppose $K = O\left(d^{0.099}\right)$ throughout this section. The arguments of Section 4.6, specifically Proposition 5 and Lemma 23, imply

mixing time lower bounds for all $\eta K = \Omega(\frac{1}{\sqrt{\kappa}})$ (using the "boosting constants" argument of Section 4.6.2 for sufficiently large κ as necessary). For $\eta K = O(\frac{1}{\sqrt{\kappa}})$, the proof of Theorem 9 further implies mixing time lower bounds for all $\eta^2 K = \omega(\frac{\sqrt{\log d}}{\kappa\sqrt{d}})$. Hence, we can assume $\eta K = O(\frac{1}{\sqrt{\kappa}})$ and $\eta^2 K = O(\frac{\sqrt{\log d}}{\kappa\sqrt{d}})$.

Next, under the further assumption that $K = O\left(d^{0.099}\right)$, it is easy to check under the specified assumptions on η and K, the preconditions of Lemma 71 are met. This implies that we can rule out $\eta^2 = \omega(\frac{\log d}{\kappa d})$ for polynomial-time mixing. Thus, in the following discussion we assume

$$K = O\left(d^{0.099}\right),\ \eta^2 = O\left(\frac{\log d}{\kappa d}\right). \tag{C.7}$$

Mixing time lower bound for small η. Let $\pi^* = \mathcal{N}(0, \mathrm{id})$ be the standard d-dimensional multivariate Gaussian. We will let π_0 be the marginal distribution of π^* on the set

$$\Omega \stackrel{\text{def}}{=} \left\{x \mathrm{mid}\, \|x\|_2^2 \le \frac{1}{2} d\right\}.$$

Recall from Lemma 9 that π_0 is a $\exp(d)$-warm start. Our main proof strategy will be to show that for small η and K as in (C.7), after $T = O(\frac{\kappa d}{K^2 \log^3 d})$ iterations, with constant probability both of the following events happen: no rejections occur throughout the Markov chain, and $\|x_{t,K}\|_2^2 \le \frac{9}{10} d$ holds for all $t \in [T]$. Combining these two facts will demonstrate our total variation lower bound.

Lemma 72. *Let $\{x_{t,k}, v_{t,k}\}_{0 \le t < T, 0 \le k \le K}$ be the sub-iterates generated by the HMC Markov chain with step size $\eta^2 = O\left(\frac{\log d}{\kappa d}\right)$ and $\eta^2 K^2 \le 1$, for $T = O(\frac{\kappa d}{K^2 \log^3 d})$ and $x_0 \sim \pi_0$; we denote the actual HMC iterates by $\{x_t\}_{0 \le t < T}$. With probability at least $\frac{99}{100}$, both of the following events occur:*

1. *Throughout the Markov chain, $\|x_t\|_2 \le 0.9\sqrt{d}$.*

2. *Throughout the Markov chain, the Metropolis filter never rejected.*

Proof. Let $h = \frac{1}{2}\eta^2$. We inductively bound the failure probability of the above events in every iteration by $\frac{0.01}{T}$, which will yield the claim via a union bound. Take some iteration $t+1$, and note that by triangle inequality, and assuming all prior iterations did not reject,

$$\|x_{t+1,K}\|_2 \le \|x_{0,0}\|_2 + \eta K \left\|\sum_{s=0}^{t} v_{s,0}\right\| + \eta^2 K \sum_{s=0}^{t}\sum_{k=1}^{K} \|x_{s,k}\|_2 \le \|x_{0,0}\|_2 + 0.9\eta^2 K^2 T\sqrt{d} + \eta K \|G_t\|_2$$

$$\le 0.8\sqrt{d} + \eta K \|G_t\|_2 .$$

Here, we applied the inductive hypothesis on all $\|x_{s,k}\|_2$, the initial bound $\|x_{0,0}\|_2 \leq \sqrt{\frac{1}{2}d}$, and that $\eta^2 K^2 T = o(1)$ by assumption. We also defined $G_t = \sum_{s=0}^{t} v_{t,0}$, where $v_{t,0}$ is the random Gaussian used by HMC in iteration k; note that by independence, $G_t \sim \mathcal{N}(0, t+1)$. By Fact 2, with probability at least $\frac{1}{200T}$, $\|G_t\|_2 \leq 2\sqrt{Td}$, and hence $0.8\sqrt{d} + \eta K \|G_t\|_2 \leq 0.9\sqrt{d}$, as desired.

Next, we prove that with probability $\geq 1 - \frac{1}{200T}$, step t does not reject. This concludes the proof by union bounding over both events in iteration t, and then union bounding over all iterations. By Corollary 6 and the calculation in Lemma 20, when $\eta^2 K^2 \leq 1$, the accept probability is

$$\min\left(1, \exp\left(\frac{h}{4}\left(\left(2\alpha - \alpha^2\right)\|x_{t,0}\|_2^2 - \beta^2 \|v_{t,0}\|_2^2 - 2(1-\alpha)\beta\left\langle x_{t,0}, v_{t,0}\right\rangle\right)\right)\right),$$

for some $\alpha \in \left[0.8hK^2, hK^2\right]$ and $\beta \in \left[0.8\sqrt{2h}K, \sqrt{2h}K\right]$. We lower bound the argument of the exponential as follows. With probability at least $1 - d^{-5} \geq 1 - \frac{1}{400T}$, Facts 1 and 2 imply both of the events $\|v_{t,0}\|_2^2 \leq 2d$ and $\langle x_{t,0}, v_{t,0}\rangle \leq 10\sqrt{\log d}\,\|x_{t,0}\|_2$ occur. Conditional on these bounds, we compute (using $2\alpha \geq \alpha^2$ and the assumption $\|x_t\|_2 \leq 0.9\sqrt{d}$)

$$\left(2\alpha - \alpha^2\right)\|x_{t,0}\|_2^2 - \beta^2\|g\|_2^2 - 2(1-\alpha)\beta\left\langle x_{t,0}, g\right\rangle \geq -4hK^2 d - 40\sqrt{h}K\sqrt{d\log d} \geq -O(K^2 \log d).$$

Hence, the acceptance probability is at least

$$\exp\left(-O\left(\eta^2 K^2 \log d\right)\right) \geq 1 - \frac{1}{400T},$$

by our choice of T with $T\eta^2 K^2 \log d = o(1)$, concluding the proof. □

Proposition 17. *The HMC Markov chain with step size $\eta^2 = O\left(\frac{\log d}{\kappa d}\right)$ and $\eta^2 K^2 \leq 1$ requires $\Omega(\frac{\kappa d}{K^2 \log^3 d})$ iterations to reach total variation distance $\frac{1}{e}$ to π^*, starting from π_0.*

Proof. The proof is identical to Proposition 4, where we use Lemma 72 instead of Lemma 16. □

Appendix D

DEFERRED CONTENTS FROM CHAPTER 5

D.1 Discussion of inexactness tolerance

We briefly discuss the tolerance of our algorithm to approximation error in two places: computation of minimizers, and implementation of RGOs in the methods of Sections 5.3 and 5.5.

Inexact minimization. For all function classes considered in this work, there exist efficient optimization methods converging to a minimizer with logarithmic dependence on the target accuracy.

Specifically, for negative log-densities with condition number κ, accelerated gradient descent [Nes83] converges at a rate $O(\sqrt{\kappa})$ with logarithmic dependence on initial error and target accuracy (we implicitly assumed in stating our runtimes that one can attain initial error polynomial in problem parameters for negative log-densities; otherwise, there is additional logarithmic overhead in the quality of the initial point to optimization procedures). For composite functions $f_{\text{wc}} + f_{\text{oracle}}$ where f_{wc} has condition number κ, the FISTA method of [BT09] converges at the same rate with each iteration querying ∇f_{wc} and a proximal oracle for f_{oracle} once; typically, access to a proximal oracle is a weaker assumption than access to a restricted Gaussian oracle, so this is not restrictive. Finally, for minimizing finite sums with condition number κ, the algorithm of [All17] obtains a convergence rate linearly dependent on $n + \sqrt{n\kappa} \leq n + \kappa$; alternatively, [JZ13] has a dependence on $n + \kappa$. In all our final runtimes, these optimization rates do not constitute the bottleneck for oracle complexities.

The only additional difficulty our algorithms may present is if the function requiring minimization, say of the form $f_{\text{oracle}}(x) + \frac{1}{2\eta} \|x - y\|_2^2$ for some $y \in \mathbb{R}^d$ where we have computed the minimizer x^* to f_{oracle}, has $\|y - x^*\|_2^2$ very large (so the initial function error is bad). However, in all our settings y is drawn from a distribution with sub-Gaussian tails, so $\|y - x^*\|_2^2$ decays exponentially (whereas the complexity of first-order methods increases only logarithmically), negligibly affecting the expected oracle query complexity for our methods.

Finally, by solving the relevant optimization problems to high accuracy as a subroutine in each of our methods, and adjusting various distance bounds to the minimizer by constants (e.g. by expanding the radius in the definition of the sets Ω in Algorithm 8 and Section 5.6.2), this accomodates tolerance to inexact minimization and only affects all bounds throughout the paper by constants. The only other place that x^* is used in our algorithms is in initializing warm starts; tolerance to inexactness in our warmness calculations follows essentially identically to Section 3.2.1 of [DCWY19].

Inexact oracle implementation. Our algorithms based on restricted Gaussian oracle access are tolerant to total variation error inverse polynomial in problem parameters for the restricted Gaussian oracle for g. We discussed this at the end of Section 5.3, in the case of RGO use for our reduction framework. To see this in the case of the composite sampler in Section 5.5, we pessimistically handled the case where the sampler `YSample` for a quadratic restriction of f resulted in total variation error in the proof of Proposition 10, assuming that the error was incurred in every iteration. By accounting for similar amounts of error in calls to $\mathcal{O}$ (on the order of $\frac{\epsilon}{T}$, where T is the number of times an RGO was used), the bounds in our algorithm are only affected by constants.

D.2 Deferred proofs from Section 5.5

D.2.1 Deferred proofs from Section 5.5.2

Approximate rejection sampling

We first define the rejection sampling framework we will use, and prove various properties.

Definition 7 (Approximate rejection sampling)**.** *Let π be a distribution, with $\frac{d\pi}{dx}(x) \propto p(x)$. Suppose set Ω has $\pi(\Omega) = 1 - \epsilon'$, and distribution $\hat{\pi}$ with $\frac{d\hat{\pi}}{dx}(x) \propto \hat{p}(x)$ has for some $C \geq 1$,*

$$\frac{p(x)}{\hat{p}(x)} \leq C \text{ for all } x \in \Omega, \text{ and } \frac{\int \hat{p}(x)dx}{\int p(x)dx} \leq 1.$$

Suppose there is an algorithm $\mathcal{A}$ which draws samples from a distribution $\hat{\pi}'$, such that $\|\hat{\pi}' - \hat{\pi}\|_{\mathrm{TV}} \leq 1 - \delta$. We call the following scheme approximate rejection sampling: repeat independent runs of the following procedure until a point is outputted.

1. *Draw x via $\mathcal{A}$ until $x \in \Omega$.*

2. *With probability $\frac{p(x)}{C\hat{p}(x)}$, output x.*

Lemma 73. *Consider an approximate rejection sampling scheme with relevant parameters defined as in Definition 7, with* $2\delta \le \frac{1-\epsilon'}{C}$. *The algorithm terminates in at most*

$$\frac{1}{\frac{1-\epsilon'}{C} - 2\delta} \tag{D.1}$$

calls to $\mathcal{A}$ *in expectation, and outputs a point from a distribution* π' *with* $\|\pi' - \pi\|_{\mathrm{TV}} \le \epsilon' + \frac{2\delta C}{1-\epsilon'}$.

Proof. Define for notational simplicity normalization constants $Z \stackrel{\text{def}}{=} \int p(x)dx$ and $\hat{Z} \stackrel{\text{def}}{=} \int \hat{p}(x)dx$. First, we bound the probability any particular call to $\mathcal{A}$ returns in the scheme:

$$\begin{aligned}
\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x) &\ge \int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x) - \left| \int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} (d\hat{\pi}'(x) - d\hat{\pi}(x)) \right| \\
&= \int_{x\in\Omega} \frac{Z}{C\hat{Z}} d\pi(x) - \left| \int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} (d\hat{\pi}'(x) - d\hat{\pi}(x)) \right| \\
&\ge \frac{1-\epsilon'}{C} - \int_{x\in\Omega} |d\hat{\pi}'(x) - d\hat{\pi}(x)| \ge \frac{1-\epsilon'}{C} - 2\delta.
\end{aligned} \tag{D.2}$$

The second line followed by the definitions of Z and $\hat{Z}$, and the third followed by triangle inequality, the assumed lower bound on $Z/\hat{Z}$, and the total variation distance between $\hat{\pi}'$ and $\hat{\pi}$. By linearity of expectation and independence, this proves the first claim.

Next, we claim the output distribution is close in total variation distance to the conditional distribution of π restricted to Ω. The derivation of (D.2) implies

$$\begin{aligned}
&\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x) \ge \frac{1-\epsilon'}{C}, \ \left| \int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} (d\hat{\pi}'(x) - d\hat{\pi}(x)) \right| \le 2\delta, \\
&\qquad \implies 1 - \frac{2\delta C}{1-\epsilon'} \le \frac{\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)}{\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x)} \le 1 + \frac{2\delta C}{1-\epsilon'}.
\end{aligned} \tag{D.3}$$

Thus, the total variation of the true output distribution from π restricted to Ω is

$$\begin{aligned}
&\frac{1}{2} \int_{x\in\Omega} \left| \frac{d\pi(x)}{1-\epsilon'} - \frac{\frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)}{\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)} \right| \\
\le \frac{1}{2} \int_{x\in\Omega} \left| \frac{d\pi(x)}{1-\epsilon'} - \frac{\frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)}{\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x)} \right| &+ \frac{1}{2} \int_{x\in\Omega} \left| \frac{\frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)}{\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x)} - \frac{\frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)}{\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)} \right| \\
\le \frac{1}{2} \int_{x\in\Omega} \left| \frac{d\pi(x)}{1-\epsilon'} - \frac{\frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)}{\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x)} \right| &+ \frac{\delta C}{1-\epsilon'} = \frac{1}{2} \int_{x\in\Omega} \frac{d\pi(x)}{1-\epsilon'} \left| 1 - \frac{d\hat{\pi}'}{d\hat{\pi}}(x) \right| + \frac{\delta C}{1-\epsilon'}.
\end{aligned}$$

The first inequality was triangle inequality, and we bounded the second term by (D.3). To obtain the final equality, we used

$$\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x) = \int_{x\in\Omega} \frac{Z}{C\hat{Z}} d\pi(x) = \frac{(1-\epsilon')Z}{C\hat{Z}}$$

$$\implies \frac{\frac{p(x)}{C\hat{p}(x)}d\hat{\pi}'(x)}{\int_{x\in\Omega}\frac{p(x)}{C\hat{p}(x)}d\hat{\pi}(x)} = \frac{p(x)}{Z}\cdot\frac{\hat{Z}}{\hat{p}(x)}\cdot\frac{1}{1-\epsilon'}\cdot d\hat{\pi}'(x) = \frac{d\pi(x)}{1-\epsilon'}\cdot\frac{d\hat{\pi}'}{d\hat{\pi}}(x).$$

We now bound this final term. Observe that the given conditions imply that $\frac{d\pi}{d\hat{\pi}}(x)$ is bounded by C everywhere in Ω. Thus, expanding we have

$$\frac{1}{2}\int_{x\in\Omega}\frac{d\pi(x)}{1-\epsilon'}\left|1-\frac{d\hat{\pi}'}{d\hat{\pi}}(x)\right| \le \frac{C}{2(1-\epsilon')}\int_{x\in\Omega}|d\hat{\pi}(x)-d\hat{\pi}'(x)| \le \frac{\delta C}{1-\epsilon'}.$$

Finally, combining these guarantees, and the fact that restricting π to Ω loses ϵ' in total variation distance, yields the desired conclusion by triangle inequality. □

Corollary 17. *Let $\hat{\theta}(x)$ be an unbiased estimator for $\frac{p(x)}{\hat{p}(x)}$, and suppose $\hat{\theta}(x) \le C$ with probability 1 for all $x \in \Omega$. Then, implementing the procedure of Definition 7 with acceptance probability $\frac{\hat{\theta}(x)}{C}$ has the same runtime bound and total variation guarantee as given by Lemma 73.*

Proof. It suffices to take expectations over the randomness of $\hat{\theta}$ everywhere in the proof of Lemma 73. □

Distribution ratio bounds

We next show two bounds relating the densities of distributions π and $\hat{\pi}$. We first define the normalization constants of (5.15), (5.17) for shorthand, and then tightly bound their ratio.

Definition 8 (Normalization constants)**.** *We denote normalization constants of π and $\hat{\pi}$ by*

$$Z_\pi \stackrel{\text{def}}{=} \int_x \exp\left(-f(x)-g(x)\right)dx,$$
$$Z_{\hat{\pi}} \stackrel{\text{def}}{=} \int_{x,y}\exp\left(-f(y)-g(x)-\frac{1}{2\eta}\left\|y-x\right\|_2^2-\frac{\eta L^2}{2}\left\|x-x^*\right\|_2^2\right)dxdy.$$

Lemma 74 (Normalization constant bounds)**.** *Let Z_π and $Z_{\hat{\pi}}$ be as in Definition 8. Then,*

$$\left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}}\left(1+\frac{\eta L^2}{\mu}\right)^{-\frac{d}{2}} \le \frac{Z_{\hat{\pi}}}{Z_\pi} \le (2\pi\eta)^{\frac{d}{2}}.$$

Proof. For each x, by convexity we have

$$\begin{aligned}
&\int_y \exp\left(-f(y)-g(x)-\frac{1}{2\eta}\|y-x\|_2^2-\frac{\eta L^2}{2}\|x-x^*\|_2^2\right)dy \\
&\le \exp\left(-g(x)-\frac{\eta L^2}{2}\|x-x^*\|_2^2\right)\int_y \exp\left(-f(x)-\langle\nabla f(x),y-x\rangle-\frac{1}{2\eta}\|y-x\|_2^2\right)dy \\
&= \exp\left(-f(x)-g(x)-\frac{\eta L^2}{2}\|x-x^*\|_2^2\right)\int_y \exp\left(\frac{\eta}{2}\|\nabla f(x)\|_2^2-\frac{1}{2\eta}\|y-x+\eta\nabla f(x)\|_2^2\right)dy \\
&= (2\pi\eta)^{\frac{d}{2}}\exp\left(-f(x)-g(x)\right)\exp\left(\frac{\eta}{2}\|\nabla f(x)\|_2^2-\frac{\eta L^2}{2}\|x-x^*\|_2^2\right) \\
&\le (2\pi\eta)^{\frac{d}{2}}\exp\left(-f(x)-g(x)\right).
\end{aligned} \tag{D.4}$$

Integrating both sides over x yields the upper bound on $\frac{Z_{\hat\pi}}{Z_\pi}$. Next, for the lower bound we have a similar derivation. For each x, by smoothness

$$\begin{aligned}
&\int_y \exp\left(-f(y)-g(x)-\frac{1}{2\eta}\|y-x\|_2^2-\frac{\eta L^2}{2}\|x-x^*\|_2^2\right)dy \\
&\ge \exp\left(-f(x)-g(x)-\frac{\eta L^2}{2}\|x-x^*\|_2^2\right)\int_y \exp\left(\langle\nabla f(x),x-y\rangle-\frac{1+\eta L}{2\eta}\|y-x\|_2^2\right)dy \\
&= \exp\left(-f(x)-g(x)-\frac{\eta L^2}{2}\|x-x^*\|^2+\frac{\eta}{2(1+\eta L)}\|\nabla f(x)\|^2\right)\left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}} \\
&\ge \exp\left(-f(x)-g(x)-\frac{\eta L^2}{2}\|x-x^*\|_2^2\right)\left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}}.
\end{aligned}$$

Integrating both sides over x yields

$$\frac{Z_{\hat\pi}}{Z_\pi} \ge \left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}}\frac{\int_x \exp\left(-f(x)-g(x)-\frac{\eta L^2}{2}\|x-x^*\|_2^2\right)dx}{\int_x \exp\left(-f(x)-g(x)\right)dx} \ge \left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}}\left(1+\frac{\eta L^2}{\mu}\right)^{-\frac{d}{2}}.$$

The last inequality followed from Proposition 21, where we used $f+g$ is μ-strongly convex.

□

Lemma 75 (Relative density bounds)**.** *Let $\eta = \frac{1}{32L\kappa d\log(288\kappa/\epsilon)}$. For all $x\in\Omega$, as defined in (5.16), $\frac{d\pi}{d\hat\pi}(x)\le 2$. Here, $\frac{d\hat\pi}{dx}(x)$ denotes the marginal density of $\hat\pi$. Moreover, for all $x\in\mathbb{R}^d$, $\frac{d\pi}{d\hat\pi}(x)\ge\frac{1}{2}$.*

Proof. We first show the upper bound. By Lemma 74,

$$\begin{aligned}
\frac{d\pi}{d\hat\pi}(x) &= \frac{\exp\left(-f(x)-g(x)\right)}{\int_y \exp\left(-f(y)-g(x)-\frac{1}{2\eta}\|y-x\|_2^2-\frac{\eta L^2}{2}\|x-x^*\|_2^2\right)dy}\cdot\frac{Z_{\hat\pi}}{Z_\pi} \\
&\le \frac{\exp\left(-f(x)-g(x)\right)}{\int_y \exp\left(-f(y)-g(x)-\frac{1}{2\eta}\|y-x\|_2^2-\frac{\eta L^2}{2}\|x-x^*\|_2^2\right)dy}\cdot(2\pi\eta)^{\frac{d}{2}}.
\end{aligned} \tag{D.5}$$

We now bound the first term, for $x \in \Omega$. By smoothness, we have

$$\frac{\exp\left(-f(y)-g(x)\right)}{\exp\left(-f(x)-g(x)\right)} \geq \exp\left(\langle \nabla f(x), x-y\rangle - \frac{L}{2}\left\|y-x\right\|_2^2\right),$$

so applying this for each y,

$$\begin{aligned}
&\frac{\int_y \exp\left(-f(y)-g(x)-\frac{1}{2\eta}\left\|y-x\right\|_2^2-\frac{\eta L^2}{2}\left\|x-x^*\right\|_2^2\right)dy}{\exp\left(-f(x)-g(x)\right)} \\
&\geq \exp\left(-\frac{\eta L^2}{2}\left\|x-x^*\right\|_2^2\right)\int_y \exp\left(\langle \nabla f(x), x-y\rangle - \frac{1+\eta L}{2\eta}\left\|y-x\right\|_2^2\right)dy \\
&= \exp\left(-\frac{\eta L^2}{2}\left\|x-x^*\right\|_2^2+\frac{\eta}{2(1+\eta L)}\left\|\nabla f(x)\right\|_2^2\right)\int_y \exp\left(-\frac{1+\eta L}{2\eta}\left\|x-y-\frac{\eta}{1+\eta L}\nabla f(x)\right\|_2^2\right)dy \\
&\geq \exp\left(-\frac{\eta L^2}{2}\cdot\frac{16d\log(288\kappa/\epsilon)}{\mu}\right)\left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}} \geq \frac{3}{4}\left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}}.
\end{aligned}$$

In the last line, we used that $x \in \Omega$ implies $\left\|x-x^*\right\|_2^2 \leq \frac{16d\log(288\kappa/\epsilon)}{\mu}$, and the definition of η. Combining this bound with (D.5), we have the desired

$$\frac{d\pi}{d\hat{\pi}}(x) \leq \frac{4}{3}\left(1+\eta L\right)^{\frac{d}{2}} \leq 2.$$

Next, we consider the lower bound. By combining (D.4) with Lemma 74, we have the desired

$$\begin{aligned}
\frac{d\pi}{d\hat{\pi}}(x) &= \frac{\exp\left(-f(x)-g(x)\right)}{\int_y \exp\left(-f(y)-g(x)-\frac{1}{2\eta}\left\|y-x\right\|_2^2-\frac{\eta L^2}{2}\left\|x-x^*\right\|_2^2\right)dy}\cdot\frac{Z_{\hat{\pi}}}{Z_\pi} \\
&\geq (2\pi\eta)^{-\frac{d}{2}}\cdot\left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}}\left(1+\frac{\eta L^2}{\mu}\right)^{-\frac{d}{2}} = \left(\frac{1}{1+\eta L}\right)^{\frac{d}{2}}\left(1+\eta L\kappa\right)^{-\frac{d}{2}} \geq \frac{1}{2}.
\end{aligned}$$

□

Correctness of `Composite-Sample-Shared-Min`

Proposition 9. *Let* $\eta = \frac{1}{32L\kappa d\log(288\kappa/\epsilon)}$, *and assume* `Sample-Joint-Dist`$(f, g, x^*, \mathcal{O}, \delta)$ *samples within* δ *total variation of the* x*-marginal on* (5.17). `Composite-Sample-Shared-Min` *outputs a sample within total variation* ϵ *of* (5.15) *in an expected* $O(1)$ *calls to* `Sample-Joint-Dist`.

Proof. We remark that $\eta = \frac{1}{32L\kappa d\log(288\kappa/\epsilon)}$ is precisely the choice of η in `Sample-Joint-Dist` where $\delta = \epsilon/18$, as in `Composite-Sample-Shared-Min`. First, we may apply Fact 6 to conclude that the measure of set Ω with respect to the μ-strongly

logconcave density π is at least $1 - \epsilon/3$. The conclusion of correctness will follow from an appeal to Corollary 17, with parameters

$$C = 4,\ \epsilon' = \frac{\epsilon}{3},\ \delta = \frac{\epsilon}{18}.$$

Note that indeed we have $\epsilon' + \frac{2\delta C}{1-\epsilon'}$ is bounded by ϵ, as $1 - \epsilon' \geq \frac{2}{3}$. Moreover, the expected number of calls (D.1) is clearly bounded by a constant as well.

We now show that these parameters satisfy the requirements of Corollary 17. Define the functions

$$p(x) \stackrel{\text{def}}{=} \exp(-f(x) - g(x)),$$
$$\hat{p}(x) \stackrel{\text{def}}{=} (2\pi\eta)^{-\frac{d}{2}} \int_y \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y - x\|_2^2 - \frac{\eta L^2}{2}\|x - x^*\|_2^2\right) dy,$$

and observe that clearly the densities of π and $\hat{\pi}$ are respectively proportional to p and $\hat{p}$. Moreover, define $Z = \int p(x)dx$ and $\hat{Z} = \int \hat{p}(x)dx$. By comparing these definitions with Lemma 74, we have $Z = Z_\pi$ and $\hat{Z} = (2\pi\eta)^{-\frac{d}{2}} Z_{\hat{\pi}}$, so by the upper bound in Lemma 74, $\hat{Z}/Z \leq 1$. Next, we claim that the following procedure produces an unbiased estimator for $\frac{p(x)}{\hat{p}(x)}$.

1. Sample $y \sim \pi_x$, where $\frac{d\pi_x(y)}{dy} \propto \exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right)$
2. $\alpha \leftarrow \exp\left(f(y) - \langle \nabla f(x), y - x\rangle - \frac{L}{2}\|y - x\|_2^2 + g(x) + \frac{\eta L^2}{2}\|x - x^*\|_2^2\right)$
3. Output $\hat{\theta}(x) \leftarrow \exp\left(-f(x) - g(x) + \frac{\eta}{2(1+\eta L)}\|\nabla f(x)\|_2^2\right)(1 + \eta L)^{\frac{d}{2}}\alpha$

To prove correctness of this estimator $\hat{\theta}$, define for simplicity

$$Z_x \stackrel{\text{def}}{=} \int_y \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y - x\|_2^2 - \frac{\eta L^2}{2}\|x - x^*\|_2^2\right) dy.$$

We compute, using $\frac{d\pi_x(y)}{dy} = \frac{\exp(-f(y) - g(x) - \frac{1}{2\eta}\|y-x\|_2^2 - \frac{\eta L^2}{2}\|x - x^*\|_2^2)}{Z_x}$, that

$$\begin{aligned}
\mathbb{E}_{\pi_x}[\alpha] &= \int_y \exp\left(f(y) - \langle \nabla f(x), y - x\rangle - \frac{L}{2}\|y - x\|_2^2 + g(x) + \frac{\eta L^2}{2}\|x - x^*\|_2^2\right) d\pi_x(y) \\
&= \frac{1}{Z_x}\int_y \exp\left(-\langle \nabla f(x), y - x\rangle - \frac{L}{2}\|y - x\|_2^2 - \frac{1}{2\eta}\|y - x\|_2^2\right) dy \\
&= \frac{1}{Z_x}\exp\left(-\frac{\eta}{2(1+\eta L)}\|\nabla f(x)\|_2^2\right)\left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}}.
\end{aligned}$$

This implies that the output quantity

$$\hat{\theta}(x) = \exp\left(-f(x) - g(x) + \frac{\eta}{2(1+\eta L)}\|\nabla f(x)\|_2^2\right)(1 + \eta L)^{\frac{d}{2}}\alpha$$

is unbiased for $\frac{p(x)}{\hat{p}(x)} = \exp(-f(x)-g(x))Z_x^{-1}(2\pi\eta)^{\frac{d}{2}}$. Finally, note that for any y used in the definition of $\hat{\theta}(x)$, by using $f(y) - f(x) - \langle \nabla f(x), y - x\rangle - \frac{L}{2}\|y-x\|_2^2 \le 0$ via smoothness, we have

$$\begin{aligned}\hat{\theta}(x) &= \exp\left(-f(x) - g(x) + \frac{\eta}{2(1+\eta L)}\|\nabla f(x)\|_2^2\right)(1+\eta L)^{\frac{d}{2}}\alpha \\ &\le (1+\eta L)^{\frac{d}{2}}\exp\left(\frac{\eta}{2(1+\eta L)}\|\nabla f(x)\|_2^2 + \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) \\ &\le (1+\eta L)^{\frac{d}{2}}\exp\left(\eta L^2\|x-x^*\|_2^2\right) \le 4.\end{aligned}$$

Here, we used the definition of η and $L^2\|x-x^*\|_2^2 \le 16L\kappa d\log(288\kappa/\epsilon)$ by the definition of Ω. □

D.2.2 Deferred proofs from Section 5.5.3

Throughout this section, for error tolerance $\delta \in [0,1]$ which parameterizes `Sample-Joint-Dist`, we denote for shorthand a high-probability region Ω_δ and its radius R_δ by

$$\Omega_\delta \stackrel{\text{def}}{=} \{x \text{mid} \|x-x^*\|_2 \le R_\delta\}, \text{ for } R_\delta \stackrel{\text{def}}{=} 4\sqrt{\frac{d\log(16\kappa/\delta)}{\mu}}. \tag{D.6}$$

The following density ratio bounds hold within this region, by simply modifying Lemma 75.

Corollary 18. *Let $\eta = \frac{1}{32L\kappa d\log(16\kappa/\delta)}$, and let $\hat{\pi}$ be parameterized by this choice of η in* (5.17)*. For all $x \in \Omega_\delta$, as defined in* (D.6)*, $\frac{d\pi}{d\hat{\pi}}(x) \le 2$. Moreover, for all $x \in \mathbb{R}^d$, $\frac{d\pi}{d\hat{\pi}}(x) \ge \frac{1}{2}$.*

The following claim follows immediately from applying Fact 6.

Lemma 76. *With probability at least $1 - \frac{\delta^2}{8(1+\kappa)^d}$, $x \sim \hat{\pi}$ lies in Ω_δ.*

Finally, when clear from context, we overload $\hat{\pi}$ as a distribution on $x \in \mathbb{R}^d$ to be the x component marginal of the distribution (5.17), i.e. with density

$$\frac{d\hat{\pi}}{dx}(x) \propto \int_y \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y-x\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2\right)dy.$$

We first note that $\hat{\pi}$ is stationary for `Sample-Joint-Dist`; this follows immediately from Lemma 25. In Section D.2.2, we bound the *conductance* of the walk. We then use this bound in Section D.2.2 to bound the mixing time and overall complexity of `Sample-Joint-Dist`.

Conductance of `Sample-Joint-Dist`

We bound the conductance of this random walk, as a process on the iterates $\{x_k\}$, to show the final point has distribution close to the marginal of $\hat{\pi}$ on x. To do so, we break Proposition 12 into two pieces, which we will use in a more white-box manner to prove our conductance bound.

Definition 9 (Restricted conductance)**.** *Let a random walk with stationary distribution $\hat{\pi}$ on $x \in \mathbb{R}^d$ have transition densities $\mathcal{T}_x$, and let $\Omega \subseteq \mathbb{R}^d$. The Ω-restricted conductance, for $v \in (0, \frac{1}{2}\hat{\pi}(\Omega))$, is*

$$\Phi_\Omega(v) = \inf_{\hat{\pi}(S \cap \Omega) \in (0, v]} \frac{\mathcal{T}_S(S^c)}{\hat{\pi}(S \cap \Omega)}, \text{ where } \mathcal{T}_S(S^c) \stackrel{\text{def}}{=} \int_{x \in S} \int_{x' \in S^c} \mathcal{T}_x(x') d\hat{\pi}(x) dx'.$$

Proposition 18 (Lemma 1, [CDWY20])**.** *Let π_{start} be a β-warm start for $\hat{\pi}$, and let $x_0 \sim \pi_{\text{start}}$. For some $\delta > 0$, let $\Omega \subseteq \mathbb{R}^d$ have $\hat{\pi}(\Omega) \geq 1 - \frac{\delta^2}{2\beta^2}$. Suppose that a random walk with stationary distribution $\hat{\pi}$ satisfies the Ω-restricted conductance bound*

$$\Phi_\Omega(v) \geq \sqrt{B \log\left(\frac{1}{v}\right)}, \text{ for all } v \in \left[\frac{4}{\beta}, \frac{1}{2}\right].$$

Let x_K be the result of K steps of this random walk, starting from x_0. Then, for

$$K \geq \frac{64}{B} \log\left(\frac{\log \beta}{2\delta}\right),$$

the resulting distribution of x_K has total variation at most $\frac{\delta}{2}$ from $\hat{\pi}$.

We state a well-known strategy for lower bounding conductance, via showing the stationary distribution has good *isoperimetry* and that transition distributions of nearby points have large overlap.

Proposition 19 (Lemma 2, [CDWY20])**.** *Let a random walk with stationary distribution $\hat{\pi}$ on $x \in \mathbb{R}^d$ have transition distribution densities $\mathcal{T}_x$, and let $\Omega \subseteq \mathbb{R}^d$, and let $\hat{\pi}_\Omega$ be the conditional distribution of $\hat{\pi}$ on Ω. Suppose for any $x, x' \in \Omega$ with $\|x - x'\|_2 \leq \Delta$,*

$$\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\text{TV}} \leq \frac{1}{2}.$$

Also, suppose $\hat{\pi}_\Omega$ satisfies, for any partition S_1, S_2, S_3 of Ω, where $d(S_1, S_2)$ is the minimum Euclidean distance between points in S_1, S_2, the log-isoperimetric inequality

$$\hat{\pi}_\Omega(S_3) \geq \frac{1}{2\psi} d(S_1, S_2) \cdot \min\left(\hat{\pi}_\Omega(S_1), \hat{\pi}_\Omega(S_2)\right) \cdot \sqrt{\log\left(1 + \frac{1}{\min\left(\hat{\pi}_\Omega(S_1), \hat{\pi}_\Omega(S_2)\right)}\right)}. \quad \text{(D.7)}$$

Then, we have the bound for all $v \in (0, \frac{1}{2}]$

$$\Phi_\Omega(v) \geq \min\left(1, \frac{\Delta}{128\psi}\sqrt{\log\left(\frac{1}{v}\right)}\right).$$

To utilize Propositions 18 and 19, we prove the following bounds in Appendices D.3.1, D.3.2, and D.3.3.

Lemma 77 (Warm start)**.** *For* $\eta \leq \frac{1}{L\kappa d}$, π_{start} *defined in* (5.18) *is a* $2(1+\kappa)^{\frac{d}{2}}$*-warm start for* $\hat{\pi}$*.*

Lemma 78 (Transitions of nearby points)**.** *Suppose* $\eta L \leq 1$, $\eta L^2 R_\delta^2 \leq \frac{1}{2}$, *and* $400d^2\eta \leq R_\delta^2$*. For a point* x*, let* $\mathcal{T}_x$ *be the density of* x_k *after sampling according to Lines 6 and 7 of Algorithm 9 from* $x_{k-1} = x$*. For* $x, x' \in \Omega_\delta$ *with* $\|x - x'\|_2 \leq \frac{\sqrt{\eta}}{10}$*, for* Ω_δ *defined in* (D.6)*, we have* $\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\text{TV}} \leq \frac{1}{2}$*.*

Lemma 79 (Isoperimetry)**.** *Density* $\hat{\pi}$ *and set* Ω_δ *defined in* (5.17), (D.6) *satisfy* (D.7) *with* $\psi = 8\mu^{-\frac{1}{2}}$*.*

We note that the parameters of Algorithm 9 and the set Ω_δ in (D.6) satisfy all assumptions of Lemmas 77, 78, and 79. By combining these results in the context of Proposition 19, we see that the random walk satisfies the bound for all $v \in (0, \frac{1}{2}]$:

$$\Phi_{\Omega_\delta}(v) \geq \sqrt{\frac{\eta\mu}{2^{20}\cdot 100}\cdot\log\left(\frac{1}{v}\right)}.$$

Plugging this conductance lower bound, the high-probability guarantee of Ω_δ by Lemma 76, and the warm start bound of Lemma 77 into Proposition 18, we have the following conclusion.

Corollary 19 (Mixing time of ideal `Sample-Joint-Dist`)**.** *Assume that calls to* `YSample` *are exact in the implementation of* `Sample-Joint-Dist`*. Then, for any error parameter* δ*, and*

$$K \stackrel{\text{def}}{=} \frac{2^{26}\cdot 100}{\eta\mu}\log\left(\frac{d\log(16\kappa)}{4\delta}\right),$$

the distribution of x_K *has total variation at most* $\frac{\delta}{2}$ *from* $\hat{\pi}$*.*

Complexity of `Sample-Joint-Dist`

We first state a guarantee on the subroutine `YSample`, which we prove in Appendix D.3.4.

Lemma 80 (`YSample` guarantee)**.** *For $\delta \in [0,1]$, define R_δ as in* (D.6)*, and let $\eta = \frac{1}{32L\kappa d\log(16\kappa/\delta)}$. For any x with $\|x - x^*\|_2 \le \sqrt{\kappa d\log(16\kappa/\delta)} \cdot R_\delta$, Algorithm 14 (`YSample`) draws an exact sample y from the density proportional to $\exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right)$ in an expected 2 iterations.*

We also state a result due to [CDWY20], which bounds the mixing time of 1-step Metropolized HMC for well-conditioned distributions; this handles the case when $\|x - x^*\|_2$ is large in Algorithm 14.

Proposition 20 (Theorem 1, [CDWY20])**.** *Let π be a distribution on $\mathbb{R}^d$ whose negative log-density is convex and has condition number bounded by a constant. Then, Metropolized HMC from an explicit starting distribution mixes to total variation δ to the distribution π in $O(d\log(\frac{d}{\delta}))$ iterations.*

Proposition 10. *`Sample-Joint-Dist` outputs a point with distribution within δ total variation distance from the x-marginal of $\hat{\pi}$. The expected number of gradient queries per iteration is constant.*

Proof. Under an exact `YSample`, Corollary 19 shows the output distribution of `Sample-Joint-Dist` has total variation at most $\frac{\delta}{2}$ from $\hat{\pi}$. Next, the resulting distribution of the subroutine `YSample` is never larger than $\delta/(2Kd\log(\frac{d\kappa}{\delta}))$ in total variation distance away from an exact sampler. By running for K steps, and using the coupling characterization of total variation, it follows that this can only incur additional error $\delta/(2d\log(\frac{d\kappa}{\delta}))$, proving correctness (in fact, the distribution is always at most $O((d\log(d\kappa/\delta))^{-1})$ away in total variation from an exact `YSample`).

Next, we prove the guarantee on the expected gradient evaluations per iteration. Lemma 80 shows whenever the current iterate x_k has $\|x - x^*\|_2 \le \sqrt{\kappa d\log(16\kappa/\delta)} \cdot R_\delta$, the expected number of gradient evaluations is constant, and moreover Proposition 20 shows that the number of gradient evaluations is never larger than $O(d\log(\frac{d\kappa}{\delta}))$, where we use that the condition number of the log-density in (5.19) is bounded by a constant. Therefore, it suffices to show in every iteration $0 \le k \le K$, the probability $\|x_k - x^*\|_2 > \sqrt{\kappa d\log(16\kappa/\delta)} \cdot R_\delta$ is $O((d\log(d\kappa/\delta))^{-1})$. By the warmness assumption in Lemma 77, and the concentration bound in Fact 6, the probability x_0 does not satisfy this bound is negligible (inverse exponential in $\kappa d^2\log(\kappa/\delta)$). Since warmness is monotonically decreasing with an exact sampler,[1] and the accumulated error due to inexactness

[1]This fact is well-known in the literature, and a simple proof is that if a distribution is warm, then

of `YSample` is at most $O((d\log(d\kappa/\delta))^{-1})$ through the whole algorithm, this holds for all iterations. □

D.3 Mixing time ingredients

We now prove facts which are used in the mixing time analysis of `Sample-Joint-Dist`. Throughout this section, as in the specification of `Sample-Joint-Dist`, f and g are functions with properties as in (5.15), and share a minimizer x^*.

D.3.1 Warm start

We show that we obtain a warm start for the distribution $\hat{\pi}$ in algorithm `Sample-Joint-Dist` via one call to the restricted Gaussian oracle for g, by proving Lemma 77.

Lemma 77 (Warm start)**.** *For $\eta \le \frac{1}{L\kappa d}$, π_{start} defined in* (5.18) *is a $2(1+\kappa)^{\frac{d}{2}}$-warm start for $\hat{\pi}$.*

Proof. By the definitions of $\hat{\pi}$ and π_{start} in (5.17), (5.18), we wish to bound everywhere the quantity

$$\frac{d\pi_{\text{start}}}{d\hat{\pi}}(x) = \frac{Z_{\hat{\pi}}}{Z_{\text{start}}} \cdot \frac{\exp\left(-\frac{L}{2}\|x-x^*\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2 - g(x)\right)}{\int_y \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y-x\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) dy}. \tag{D.8}$$

Here, $Z_{\hat{\pi}}$ is as in Definition 8, and we let Z_{start} denote the normalization constant of π_{start}, i.e.

$$Z_{\text{start}} \stackrel{\text{def}}{=} \int_x \exp\left(-\frac{L}{2}\|x-x^*\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2 - g(x)\right) dx.$$

Regarding the first term of (D.8), the earlier derivation (D.4) showed

$$\int_y \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y-x\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) dy \le (2\pi\eta)^{\frac{d}{2}} \exp\left(-f(x) - g(x)\right).$$

Then, integrating, we can bound the ratio of the normalization constants

$$\begin{aligned}
\frac{Z_{\hat{\pi}}}{Z_{\pi_{\text{start}}}} &\le \frac{\int_x (2\pi\eta)^{\frac{d}{2}} \exp\left(-f(x) - g(x)\right) dx}{\int_x \exp\left(-\frac{L}{2}\|x-x^*\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2 - g(x)\right) dx} \\
&\le \frac{\int_x (2\pi\eta)^{\frac{d}{2}} \exp\left(-f(x^*) - \frac{\mu}{2}\|x-x^*\|_2^2 - g(x)\right) dx}{\int_x \exp\left(-\frac{L}{2}\|x-x^*\|_2^2 - \frac{\mu}{2}\|x-x^*\|_2^2 - g(x)\right) dx} \\
&\le (2\pi\eta)^{\frac{d}{2}} \exp\left(-f(x^*)\right) \left(1 + \frac{L}{\mu}\right)^{\frac{d}{2}}.
\end{aligned} \tag{D.9}$$

---

taking one step of the Markov chain induces a convex combination of warm point masses, and is thus also warm.

The second inequality followed from f is μ-strongly convex and $\eta L^2 \le \mu$ by assumption. The last inequality followed from Proposition 21, where we used $\frac{\mu}{2}\|x-x^*\|_2^2 + g(x)$ is μ-strongly convex. Next, to bound the second term of (D.8), notice first that

$$\frac{\exp\left(-\frac{L}{2}\|x-x^*\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2 - g(x)\right)}{\int_y \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y-x\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) dy} = \frac{\exp\left(-\frac{L}{2}\|x-x^*\|_2^2\right)}{\int_y \exp\left(-f(y) - \frac{1}{2\eta}\|y-x\|_2^2\right) dy}.$$

It thus suffices to lower bound $\exp\left(\frac{L}{2}\|x-x^*\|_2^2\right)\int_y \exp\left(-f(y) - \frac{1}{2\eta}\|y-x\|_2^2\right) dy$. We have

$$\begin{aligned}
&\exp\left(\frac{L}{2}\|x-x^*\|_2^2\right)\int_y \exp\left(-f(y) - \frac{1}{2\eta}\|y-x\|_2^2\right) dy \\
&\ge \exp\left(-f(x) + \frac{L}{2}\|x-x^*\|_2^2\right)\int_y \exp\left(-\langle \nabla f(x), y-x\rangle - \left(\frac{1}{2\eta}+\frac{L}{2}\right)\|y-x\|_2^2\right) dy \\
&= \exp\left(-f(x) + \frac{L}{2}\|x-x^*\|_2^2\right)\left(\frac{2\pi\eta}{1+L\eta}\right)^{\frac{d}{2}} \exp\left(\frac{\eta}{2(1+L\eta)}\|\nabla f(x)\|_2^2\right) \\
&\ge \exp(-f(x^*))\left(\frac{2\pi\eta}{1+L\eta}\right)^{\frac{d}{2}}
\end{aligned} \tag{D.10}$$

The first and third steps followed from L-smoothness of f, and the second applied the Gaussian integral (Fact 5). Combining the bounds in (D.9) and (D.10), (D.8) becomes

$$\frac{d\pi_{\text{start}}}{d\hat{\pi}}(x) \le \left(1+\frac{L}{\mu}\right)^{\frac{d}{2}}(1+L\eta)^{\frac{d}{2}} \le 2(1+\kappa)^{\frac{d}{2}},$$

where $x \in \mathbb{R}^d$ was arbitrary, which completes the proof. □

D.3.2 Transitions of nearby points

Here, we prove Lemma 78. Throughout this section, $\mathcal{T}_x$ is the density of x_k, according to the steps in Lines 6 and 7 of `Sample-Joint-Dist` (Algorithm 9) starting at $x_{k-1} = x$. We also define $\mathcal{P}_x$ to be the density of y_k, by just the step in Line 6. We first make a simplifying observation: by Observation 1, for any two points x, x', we have

$$\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\text{TV}} \le \|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\text{TV}}.$$

Thus, it suffices to understand $\|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\text{TV}}$ for nearby $x, x' \in \Omega_\delta$. Our proof of Lemma 78 combines two pieces: (1) bounding the ratio of normalization constants Z_x, $Z_{x'}$ of $\mathcal{P}_x$ and $\mathcal{P}_{x'}$ for nearby x, x' in Lemma 83 and (2) the structural result Proposition 22. To bound the normalization constant ratio, we state two helper lemmas. Lemma 81 characterizes facts about the minimizer of

$$f(y) + \frac{1}{2\eta}\|y-x\|_2^2. \tag{D.11}$$

Lemma 81. *Let f be convex with minimizer x^*, and y_x minimize* (D.11) *for a given x. Then,*

1. $\|y_x - y_{x'}\|_2 \leq \|x - x'\|_2$.
2. *For any x,* $\|y_x - x^*\|_2 \leq \|x - x^*\|_2$.
3. *For any x with $\|x - x^*\|_2 \leq R$,* $\|x - y_x\|_2 \leq \eta LR$.

Proof. By optimality conditions in the definition of y_x,

$$\eta \nabla f(y_x) = x - y_x.$$

Fix two points x, x', and let $x_t \stackrel{\text{def}}{=} (1-t)x + tx'$. Letting $\mathbf{J}_x(y_x)$ be the Jacobian matrix of y_x,

$$\begin{aligned}\frac{d}{dt}\eta\nabla f(y_{x_t}) = \frac{d}{dt}(x_t - y_{x_t}) &\implies \eta\nabla^2 f(y_{x_t})\mathbf{J}_x(y_{x_t})(x'-x) = (\mathrm{id} - \mathbf{J}_x(y_{x_t}))(x'-x) \\ &\implies \mathbf{J}_x(y_{x_t})(x'-x) = (\mathrm{id} + \eta\nabla^2 f(y_{x_t}))^{-1}(x'-x).\end{aligned}$$

We can then compute

$$y_{x'} - y_x = \int_0^1 \frac{d}{dt} y_{x_t} dt = \int_0^1 \mathbf{J}_x(y_{x_t})(x'-x)dt = \int_0^1 (\mathrm{id} + \eta\nabla^2 f(y_{x_t}))^{-1}(x'-x)dt.$$

By triangle inequality and convexity of f, the first claim follows:

$$\|y_{x'} - y_x\|_2 \leq \int_0^1 \left\|(\mathrm{id} + \eta\nabla^2 f(y_{x_t}))^{-1}\right\|_2 \left\|x' - x\right\|_2 dt \leq \left\|x' - x\right\|_2.$$

The second claim follows from the first by $y_{x^*} = x^*$. The third claim follows from the second via

$$\|x - y_x\|_2 = \eta \|\nabla f(y_x)\|_2 \leq \eta L \|y_x - x^*\|_2 \leq \eta LR.$$

□

Next, Lemma 82 states well-known bounds on the integral of a well-conditioned function h.

Lemma 82. *Let h be a L_h-smooth, μ_h-strongly convex function and let y_h^* be its minimizer. Then*

$$\left(2\pi L_h^{-1}\right)^{\frac{d}{2}} \exp\left(-h(y_h^*)\right) \leq \int_y \exp\left(-h(y)\right) \leq \left(2\pi\mu_h^{-1}\right)^{\frac{d}{2}} \exp\left(-h(y_h^*)\right).$$

Proof. By smoothness and strong convexity,

$$\exp\left(-h(y_h^*) - \frac{L_h}{2}\|y - y_h^*\|_2^2\right) \le \exp(-h(y)) \le \exp\left(-h(y_h^*) - \frac{\mu_h}{2}\|y - y_h^*\|_2^2\right).$$

The result follows by Gaussian integrals, i.e. Fact 5. □

We now define the normalization constants of $\mathcal{P}_x$ and $\mathcal{P}_{x'}$:

$$\begin{aligned} Z_x &= \int_y \exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right) dy, \\ Z_{x'} &= \int_y \exp\left(-f(y) - \frac{1}{2\eta}\|y - x'\|_2^2\right) dy. \end{aligned} \tag{D.12}$$

We apply Lemma 81 and Lemma 82 to bound the ratio of Z_x and $Z_{x'}$.

Lemma 83. *Let f be μ-strongly convex and L-smooth. Let $x, x' \in \Omega_\delta$, for Ω_δ defined in* (D.6)*, and let $\|x - x'\|_2 \le \Delta$. Then, the normalization constants Z_x and $Z_{x'}$ in* (D.12) *satisfy*

$$\frac{Z_x}{Z_{x'}} \le 1.05 \exp\left(3LR\Delta + \frac{L\Delta^2}{2}\right).$$

Proof. First, applying Lemma 82 to Z_x and $Z_{x'}$ yields that the ratio is bounded by

$$\begin{aligned} \frac{Z_x}{Z_{x'}} &\le \frac{\exp\left(-f(y_x) - \frac{1}{2\eta}\|y_x - x\|_2^2\right)\left(2\pi\left(\mu + \frac{1}{\eta}\right)^{-1}\right)^{\frac{d}{2}}}{\exp\left(-f(y_{x'}) - \frac{1}{2\eta}\|y_{x'} - x\|_2^2\right)\left(2\pi\left(L + \frac{1}{\eta}\right)^{-1}\right)^{\frac{d}{2}}} \\ &\le 1.05 \exp\left(f(y_{x'}) - f(y_x) + \frac{1}{2\eta}\left(\|y_{x'} - x'\|_2^2 - \|y_x - x\|_2^2\right)\right). \end{aligned}$$

Here, we used the bound for $\eta^{-1} \ge 32Ld$ that

$$\left(\frac{L + \frac{1}{\eta}}{\mu + \frac{1}{\eta}}\right)^{d/2} \le 1.05.$$

Regarding the remaining term, recall x, x' both belong to Ω_δ, and $\|x - x'\|_2 \le \Delta$. We have

$$\begin{aligned} &f(y_{x'}) - f(y_x) + \frac{1}{2\eta}\left(\|y_{x'} - x'\|_2^2 - \|y_x - x\|_2^2\right) \\ &\le \langle \nabla f(y_x), y_{x'} - y_x\rangle + \frac{L}{2}\|y_{x'} - y_x\|_2^2 + \frac{1}{2\eta}\langle y_{x'} - x' + y_x - x, y_{x'} - y_x + x - x'\rangle \\ &\le LR\Delta + \frac{L\Delta^2}{2} + \frac{1}{2\eta}\left(\|y_x - x\|_2 + \|y_{x'} - x'\|_2\right)\left(\|y_{x'} - y_x\|_2 + \|x' - x\|_2\right) \\ &\le LR\Delta + \frac{L\Delta^2}{2} + \frac{2\eta LR}{2\eta}\left(\|y_{x'} - y_x\|_2 + \|x' - x\|_2\right) \le 3LR\Delta + \frac{L\Delta^2}{2}. \end{aligned}$$

The first inequality was smoothness and expanding the difference of quadratics. The second was by $\|\nabla f(y_x)\|_2 \le L\|y_x - x^*\|_2 \le LR$ and $\|y_{x'} - y_x\|_2 \le \Delta$, where we used the first and second parts of Lemma 81; we also applied Cauchy-Schwarz and triangle inequality. The third used the third part of Lemma 81. Finally, the last inequality was by the first part of Lemma 81 and $\|x' - x\|_2 \le \Delta$. □

We now are ready to prove Lemma 78.

Lemma 78 (Transitions of nearby points)**.** *Suppose $\eta L \le 1$, $\eta L^2 R_\delta^2 \le \frac{1}{2}$, and $400d^2\eta \le R_\delta^2$. For a point x, let $\mathcal{T}_x$ be the density of x_k after sampling according to Lines 6 and 7 of Algorithm 9 from $x_{k-1} = x$. For $x, x' \in \Omega_\delta$ with $\|x - x'\|_2 \le \frac{\sqrt{\eta}}{10}$, for Ω_δ defined in* (D.6)*, we have $\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\mathrm{TV}} \le \frac{1}{2}$.*

Proof. First, by Observation 1, it suffices to show $\|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\mathrm{TV}} \le \frac{1}{2}$. Pinsker's inequality states

$$\|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2} d_{\mathrm{KL}}\left(\mathcal{P}_x, \mathcal{P}_{x'}\right)},$$

where d_{KL} is KL-divergence, so it is enough to show $d_{\mathrm{KL}}\left(\mathcal{P}_x, \mathcal{P}_{x'}\right) \le \frac{1}{2}$. Notice that

$$d_{\mathrm{KL}}\left(\mathcal{P}_x, \mathcal{P}_{x'}\right) = \log\left(\frac{Z_{x'}}{Z_x}\right) + \int_y \mathcal{P}_x(y) \log\left(\frac{\exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right)}{\exp\left(-f(y) - \frac{1}{2\eta}\|y - x'\|_2^2\right)}\right) dy.$$

By Lemma 83, the first term satisfies, for $\Delta \stackrel{\mathrm{def}}{=} \frac{\sqrt{\eta}}{10}$,

$$\log\left(\frac{Z_{x'}}{Z_x}\right) \le 3LR\Delta + \frac{L\Delta^2}{2} + \log(1.05).$$

To bound the second term, we have

$$\begin{aligned}
\int_y \mathcal{P}_x(y) \log\left(\frac{\exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right)}{\exp\left(-f(y) - \frac{1}{2\eta}\|y - x'\|_2^2\right)}\right) dy &= \frac{1}{2\eta}\int_y \mathcal{P}_x(y)\left(\|y - x'\|_2^2 - \|y - x\|_2^2\right) dy \\
&= \frac{1}{2\eta}\int_y \mathcal{P}_x(y)\left\langle x - x', 2\left(y - x\right) + \left(x - x'\right)\right\rangle dy \\
&\le \frac{\Delta^2}{2\eta} + \frac{\Delta}{\eta}\left\|\int_y y\mathcal{P}_x(y)dy - x\right\|_2.
\end{aligned}$$

Here, the second line was by expanding and the third line was by $\|x - x'\|_2 \le \Delta$ and Cauchy-Schwarz. By Proposition 22, $\left\|\int_y y\mathcal{P}_x(y)dy - x\right\|_2 \le 2\eta LR$, where by assumption the parameters satisfy the conditions of Proposition 22. Then, combining the two bounds, we have

$$d_{\mathrm{KL}}\left(\mathcal{P}_x, \mathcal{P}_{x'}\right) \le 3LR\Delta + \frac{L\Delta^2}{2} + \frac{\Delta^2}{2\eta} + 2LR\Delta + \log(1.05) = 5LR\Delta + \frac{L\Delta^2}{2} + \frac{\Delta^2}{2\eta} + \log(1.05).$$

When $\Delta = \frac{\sqrt{\eta}}{10}$, $\eta L \le 1$, and $\eta L^2 R^2 \le \frac{1}{2}$, we have the desired

$$d_{\mathrm{KL}}\left(\mathcal{P}_x, \mathcal{P}_{x'}\right) \le \frac{\sqrt{\eta} L R}{2} + \frac{L\eta}{200} + \frac{1}{200} + \log(1.05) \le \frac{1}{2}.$$

□

D.3.3 Isoperimetry

In this section, we prove Lemma 79, which asks to show that $\hat{\pi}_{\Omega_\delta}$ satisfies a log-isoperimetric inequality (D.7). Here, we define $\hat{\pi}_{\Omega_\delta}$ to be the conditional distribution of the $\hat{\pi}$ x-marginal on set Ω_δ. We recall this means that for any partition S_1, S_2, S_3 of Ω_δ,

$$\hat{\pi}_{\Omega_\delta}(S_3) \ge \frac{1}{2\psi} d(S_1, S_2) \cdot \min\left(\hat{\pi}_{\Omega_\delta}(S_1), \hat{\pi}_{\Omega_\delta}(S_2)\right) \cdot \sqrt{\log\left(1 + \frac{1}{\min\left(\hat{\pi}_{\Omega_\delta}(S_1), \hat{\pi}_{\Omega_\delta}(S_2)\right)}\right)}.$$

The following fact was shown in [CDWY20].

Lemma 84 ([CDWY20], Lemma 11)**.** *Any μ-strongly logconcave distribution π satisfies the log-isoperimetric inequality* (D.7) *with $\psi = \mu^{-\frac{1}{2}}$.*

Observe that π_{Ω_δ}, the restriction of π to the convex set Ω_δ, is μ-strongly logconcave by the definition of π (5.15), so it satisfies a log-isoperimetric inequality. We now combine this fact with the relative density bounds Lemma 75 to prove Lemma 79.

Lemma 79 (Isoperimetry)**.** *Density $\hat{\pi}$ and set Ω_δ defined in* (5.17)*,* (D.6) *satisfy* (D.7) *with $\psi = 8\mu^{-\frac{1}{2}}$.*

Proof. Fix some partition S_1, S_2, S_3 of Ω_δ, and without loss of generality let $\hat{\pi}_{\Omega_\delta}(S_1) \le \hat{\pi}_{\Omega_\delta}(S_2)$. First, by applying Corollary 18, which shows $\frac{d\pi}{d\hat{\pi}}(x) \in [\frac{1}{2}, 2]$ everywhere in Ω_δ, we have the bounds

$$\frac{1}{2}\pi_{\Omega_\delta}(S_1) \le \hat{\pi}_{\Omega_\delta}(S_1) \le 2\pi_{\Omega_\delta}(S_1),\ \frac{1}{2}\pi_{\Omega_\delta}(S_2) \le \hat{\pi}_{\Omega_\delta}(S_2) \le 2\pi_{\Omega_\delta}(S_2),\ \text{and } \hat{\pi}_{\Omega_\delta}(S_3) \ge \frac{1}{2}\pi_{\Omega_\delta}(S_3).$$

Therefore, we have the sequence of conclusions

$$\begin{aligned}
\hat{\pi}_{\Omega_\delta}(S_3) &\ge \frac{1}{2}\pi_{\Omega_\delta}(S_3) \\
&\ge \frac{d(S_1, S_2)\sqrt{\mu}}{4} \cdot \min\left(\pi_{\Omega_\delta}(S_1), \pi_{\Omega_\delta}(S_2)\right) \cdot \sqrt{\log\left(1 + \frac{1}{\min\left(\pi_{\Omega_\delta}(S_1), \pi_{\Omega_\delta}(S_2)\right)}\right)} \\
&\ge \frac{d(S_1, S_2)\sqrt{\mu}}{8} \cdot \hat{\pi}_{\Omega_\delta}(S_1) \cdot \sqrt{\log\left(1 + \frac{1}{2\hat{\pi}_{\Omega_\delta}(S_1)}\right)}
\end{aligned}$$

$$\geq \frac{d(S_1, S_2)\sqrt{\mu}}{16} \cdot \hat{\pi}_{\Omega_\delta}(S_1) \cdot \sqrt{\log\left(1 + \frac{1}{\hat{\pi}_{\Omega_\delta}(S_1)}\right)}.$$

Here, the second line was by applying Lemma 84 to the μ-strongly logconcave distribution π_{Ω_δ}, and the final line used $\sqrt{\log(1+\alpha)} \leq 2\sqrt{\log(1+\frac{\alpha}{2})}$ for all $\alpha > 0$. □

D.3.4 Correctness of `YSample`

In this section, we show how we can sample y efficiently in the alternating scheme of the algorithm `Sample-Joint-Dist`, within an extremely high probability region. Specifically, for any x with $\|x - x^*\|_2 \leq \sqrt{\kappa d \log(16\kappa/\delta)} \cdot R_\delta$, where R_δ is defined in (D.6), we give a method for implementing

$$\text{draw } y \propto \exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right) dy.$$

The algorithm is Algorithm 14, which is a simple rejection sampling scheme.

Algorithm 14 `YSample`(f, x, η, δ)

Input: L-smooth, μ-strongly convex $f : \mathbb{R}^d \to \mathbb{R}$ with minimizer x^*, $\eta > 0$, $\delta \in [0, 1]$, $x \in \mathbb{R}^d$.

Output: If $\|x - x^*\|_2 \leq \sqrt{\kappa d \log(16\kappa/\delta)} \cdot R_\delta$, return exact sample from distribution with density $\propto \exp(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2)$ (see (D.6) for definition of R_δ). Otherwise, return sample within δ TV from distribution with density $\propto \exp(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2)$.

if $\|x - x^*\|_2 \leq \sqrt{\kappa d \log(16\kappa/\delta)} \cdot R_\delta$ **then**
  while true do
    Draw $y \sim \mathcal{N}(x - \eta\nabla f(x), \eta\,\mathrm{id})$
    $\tau \sim \mathrm{Unif}[0, 1]$
    if $\tau \leq \exp(f(x) + \langle \nabla f(x), y - x\rangle - f(y))$ **then**
      return y
    end if
  end while
end if
return Sample x within TV δ from density $\propto \exp(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2)$ using [CDWY20]

We recall that we gave guarantees on rejection sampling procedures in Lemma 29 (an

"exact" version of Lemma 73 and Corollary 17). We now prove Lemma 80 via a direct application of Lemma 29.

Lemma 80 (`YSample` guarantee)**.** *For $\delta \in [0,1]$, define R_δ as in* (D.6)*, and let $\eta = \frac{1}{32L\kappa d\log(16\kappa/\delta)}$. For any x with $\|x-x^*\|_2 \le \sqrt{\kappa d\log(16\kappa/\delta)}\cdot R_\delta$, Algorithm 14 (`YSample`) draws an exact sample y from the density proportional to $\exp\left(-f(y)-\frac{1}{2\eta}\|y-x\|_2^2\right)$ in an expected 2 iterations.*

Proof. For $\|x-x^*\|_2 \le \sqrt{\kappa d\log(16\kappa/\delta)}\cdot R_\delta$, `YSample` is a rejection sampling scheme with

$$p(y) = \exp\left(-f(y)-\frac{1}{2\eta}\|y-x\|_2^2\right),\ \hat{p}(y) = \exp\left(-f(x)-\langle\nabla f(x), y-x\rangle - \frac{1}{2\eta}\|y-x\|_2^2\right).$$

It is clear that $p(y)\le\hat{p}(y)$ everywhere by convexity of f, so we may choose $C=1$. To bound the expected number of iterations and obtain the desired conclusion, Lemma 29 requires a bound on

$$\frac{\int_y \exp\left(-f(x)-\langle\nabla f(x),y-x\rangle-\frac{1}{2\eta}\|y-x\|_2^2\right)dy}{\int_y \exp\left(-f(y)-\frac{1}{2\eta}\|y-x\|_2^2\right)dy}, \tag{D.13}$$

the ratio of the normalization constants of $\hat{p}$ and p. First, by Fact 5,

$$\int_y \exp\left(-f(x)-\langle\nabla f(x),y-x\rangle-\frac{1}{2\eta}\|y-x\|_2^2\right)dy = \exp\left(-f(x)+\frac{\eta}{2}\|\nabla f(x)\|_2^2\right)(2\pi\eta)^{\frac{d}{2}}.$$

Next, by smoothness and Fact 5 once more,

$$\begin{aligned}\int_y \exp\left(-f(y)-\frac{1}{2\eta}\|y-x\|_2^2\right)dy &\ge \int_y \exp\left(-f(x)-\langle\nabla f(x),y-x\rangle-\frac{1+\eta L}{2\eta}\|y-x\|_2^2\right)dy\\ &= \exp\left(-f(x)+\frac{\eta}{2(1+\eta L)}\|\nabla f(x)\|_2^2\right)\left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}}.\end{aligned}$$

Taking a ratio, the quantity in (D.13) is bounded above by

$$\begin{aligned}\exp\left(\left(\frac{\eta}{2}-\frac{\eta}{2(1+\eta L)}\right)\|\nabla f(x)\|_2^2\right)(1+\eta L)^{\frac{d}{2}} &\le 1.5\exp\left(\frac{\eta^2 L}{2(1+\eta L)}\|\nabla f(x)\|_2^2\right)\\ &\le 1.5\exp\left(\frac{\eta^2L^3}{2}\cdot\left(\frac{16\kappa d^2\log^2(16\kappa/\delta)}{\mu}\right)\right)\le 2.\end{aligned}$$

The first inequality was $(1+\eta L)^{\frac{d}{2}}\le 1.5$, the second used smoothness and the assumed bound on $\|x-x^*\|_2$, and the third again used our choice of η. □

D.4 Structural results

Here, we prove two structural results about distributions whose negative log-densities are small perturbations of a quadratic, which obtain tighter concentration guarantees

compared to naive bounds on strongly logconcave distributions. They are used in obtaining our bounds in Section D.3 (and for the warm start bounds in Section 5.4), but we hope both the statements and proof techniques are of independent interest to the community. Our first structural result is a bound on normalization constant ratios, used throughout the paper.

Proposition 21. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex with minimizer x^*, and let $\lambda > 0$. Then,*

$$\frac{\int \exp(-f(x))dx}{\int \exp\left(-f(x) - \frac{1}{2\lambda}\|x - x^*\|_2^2\right)dx} \le \left(1 + \frac{1}{\mu\lambda}\right)^{\frac{d}{2}}.$$

Proof. Define the function

$$R(\alpha) \stackrel{\text{def}}{=} \frac{\int \exp\left(-f(x) - \frac{1}{2\lambda\alpha}\|x - x^*\|_2^2\right)dx}{\int \exp\left(-f(x) - \frac{1}{2\lambda}\|x - x^*\|_2^2\right)dx}.$$

Let $d\pi_\alpha(x)$ be the density proportional to $\exp\left(-f(x) - \frac{1}{2\lambda\alpha}\|x - x^*\|_2^2\right)dx$. We compute

$$\begin{aligned}
\frac{d}{d\alpha}R(\alpha) &= \int \frac{\exp\left(-f(x) - \frac{1}{2\lambda\alpha}\|x - x^*\|_2^2\right)}{\int \exp\left(-f(x) - \frac{1}{2\lambda}\|x - x^*\|_2^2\right)dx} \frac{1}{2\lambda\alpha^2}\|x - x^*\|_2^2\, dx \\
&= \frac{R(\alpha)}{2\lambda\alpha^2}\int \frac{\exp\left(-f(x) - \frac{1}{2\lambda\alpha}\|x - x^*\|_2^2\right)\|x - x^*\|_2^2}{\int \exp\left(-f(x) - \frac{1}{2\lambda\alpha}\|x - x^*\|_2^2\right)dx}dx \\
&= \frac{R(\alpha)}{2\lambda\alpha^2}\int \|x - x^*\|_2^2\, d\pi_\alpha(x) \le \frac{R(\alpha)}{2\alpha}\cdot\frac{d}{\mu\lambda\alpha + 1}.
\end{aligned}$$

Here, the last inequality was by Fact 8, using the fact that the function $f(x) + \frac{1}{2\lambda\alpha}\|x - x^*\|_2^2$ is $\mu + \frac{1}{\lambda\alpha}$-strongly convex. Moreover, note that $R(1) = 1$, and

$$\frac{d}{d\alpha}\log\left(\frac{\alpha}{\mu\lambda\alpha + 1}\right) = \frac{1}{\alpha} - \frac{\mu\lambda}{\mu\lambda\alpha + 1} = \frac{1}{\mu\lambda\alpha^2 + \alpha}.$$

Solving the differential inequality

$$\frac{d}{d\alpha}\log(R(\alpha)) = \frac{dR(\alpha)}{d\alpha}\cdot\frac{1}{R(\alpha)} \le \frac{d}{2}\cdot\frac{1}{\mu\lambda\alpha^2 + \alpha},$$

we obtain the bound for any $\alpha \ge 1$ (since $\log(R(1)) = 0$)

$$\log(R(\alpha)) \le \frac{d}{2}\log\left(\frac{\mu\lambda\alpha + \alpha}{\mu\lambda\alpha + 1}\right) \implies R(\alpha) \le \left(\frac{\mu\lambda\alpha + \alpha}{\mu\lambda\alpha + 1}\right)^{\frac{d}{2}} \le \left(1 + \frac{1}{\mu\lambda}\right)^{\frac{d}{2}}.$$

Taking a limit $\alpha \to \infty$ yields the conclusion. □

Our second structural result uses a similar proof technique to show that the mean of a bounded perturbation f of a Gaussian is not far from its mode, as long as the gradient of the mode is small. We remark that one may directly apply strong logconcavity, i.e. a variant of Fact 8, to obtain a weaker bound by roughly a $\sqrt{d}$ factor, which would result in a loss of $\Omega(d)$ in the guarantees of Theorem 11. This tighter analysis is crucial in our improved mixing time result.

Before stating the bound, we apply Fact 7 to the convex functions $h(x) = (\theta^\top x)^2$ and $h(x) = \|x\|_2^4$ to obtain the following conclusions which will be used in the proof of Proposition 22.

Corollary 20. *Let π be a μ-strongly logconcave density. Then,*

1. $\mathbb{E}_\pi[(\theta^\top(x - \mathbb{E}_\pi[x]))^2] \le \mu^{-1}$, *for all unit vectors θ.*

2. $\mathbb{E}_\pi[\|x - \mathbb{E}_\pi[x]\|_2^4] \le 3d^2\mu^{-2}$.

Proposition 22. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and convex with minimizer x^*, let $x \in \mathbb{R}^d$ with $\|x - x^*\|_2 \le R$, and let $d\pi_\eta(y)$ be the density proportional to $\exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right) dy$. Suppose that $\eta \le \min\left(\frac{1}{2L^2R^2}, \frac{R^2}{400d^2}\right)$. Then,*

$$\left\|\mathbb{E}_{\pi_\eta}[y] - x\right\|_2 \le 2\eta LR.$$

Proof. Define a family of distributions π^α for $\alpha \in [0, 1]$, with

$$d\pi^\alpha(y) \propto \exp\left(-\alpha\left(f(y) - f(x) - \langle \nabla f(x), y - x\rangle\right) - f(x) - \langle \nabla f(x), y - x\rangle - \frac{1}{2\eta}\|y - x\|_2^2\right) dy.$$

In particular, $\pi^1 = \pi_\eta$, and π^0 is a Gaussian with mean $x - \eta\nabla f(x)$. We define $\bar{y}_\alpha \stackrel{\text{def}}{=} \mathbb{E}_{\pi_\alpha}[y]$, and

$$y_\alpha^* \stackrel{\text{def}}{=} \text{argmin}_y \left\{\alpha\left(f(y) - f(x) - \langle \nabla f(x), y - x\rangle\right) + f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2\eta}\|y - x\|_2^2\right\}.$$

Define the function $D(\alpha) \stackrel{\text{def}}{=} \|\bar{y}_\alpha - x\|_2$, such that we wish to bound $D(1)$. First, by smoothness

$$D(0) = \|\mathbb{E}_{\pi_0}[y] - x\|_2 = \|\eta\nabla f(x)\|_2 \le \eta LR.$$

Next, we observe

$$\frac{d}{d\alpha}D(\alpha) = \left\langle \frac{\bar{y}_\alpha - x}{\|\bar{y}_\alpha - x\|_2}, \frac{d\bar{y}_\alpha}{d\alpha}\right\rangle \le \left\|\frac{d\bar{y}_\alpha}{d\alpha}\right\|_2.$$

In order to bound $\left\|\frac{d\bar{y}_\alpha}{d\alpha}\right\|_2$, fix a unit vector θ. We have

$$\begin{aligned}\left\langle \frac{d\bar{y}_\alpha}{d\alpha}, \theta \right\rangle &= \frac{d}{d\alpha}\left\langle \int (y-x)d\pi^\alpha(y), \theta \right\rangle \\ &= \int \langle y-x, \theta\rangle \left(f(x) + \langle \nabla f(x), y-x\rangle - f(y)\right) d\pi^\alpha(y) \\ &\leq \sqrt{\int (\langle y-x,\theta\rangle)^2 d\pi^\alpha(y)} \sqrt{\int (f(x) + \langle \nabla f(x), y-x\rangle - f(y))^2 d\pi^\alpha(y)} \\ &\leq \sqrt{\int (\langle y-x,\theta\rangle)^2 d\pi^\alpha(y)} \sqrt{\int \frac{L^2}{4}\|y-x\|_2^4 \, d\pi^\alpha(y)}.\end{aligned} \tag{D.14}$$

The third line was Cauchy-Schwarz and the last line used smoothness and convexity, i.e.

$$-\frac{L}{2}\|y-x\|_2^2 \leq f(x) + \langle \nabla f(x), y-x\rangle - f(y) \leq 0.$$

We now bound these terms. First,

$$\begin{aligned}\int (\langle y-x,\theta\rangle)^2 d\pi^\alpha(y) &\leq 2\int (\langle y-\bar{y}_\alpha,\theta\rangle)^2 d\pi^\alpha(y) + 2\int (\langle \bar{y}_\alpha - x,\theta\rangle)^2 d\pi^\alpha(y) \\ &\leq 2\eta + 2\|\bar{y}_\alpha - x\|_2^2 = 2\eta + 2D(\alpha)^2.\end{aligned} \tag{D.15}$$

Here, we applied the first part of Corollary 20, as π^α is η^{-1}-strongly logconcave, and the definition of $D(\alpha)$. Next, using for any $a, b \in \mathbb{R}^d$, $\|a+b\|_2^4 \leq (\|a\|_2 + \|b\|_2)^4 \leq 16\|a\|_2^4 + 16\|b\|_2^4$, we have

$$\begin{aligned}\int \frac{L^2}{4}\|y-x\|_2^4 d\pi^\alpha(y) &\leq \int 4L^2\|y-\bar{y}_\alpha\|_2^4 d\pi^\alpha(y) + \int 4L^2\|x-\bar{y}_\alpha\|_2^4 d\pi^\alpha(y) \\ &\leq 12L^2d^2\eta^2 + 4L^2D(\alpha)^4.\end{aligned} \tag{D.16}$$

Here, we used the second part of Corollary 20. Maximizing (D.14) over θ, and applying (D.15), (D.16),

$$\begin{aligned}\frac{d}{d\alpha}D(\alpha) \leq \left\|\frac{d\bar{y}_\alpha}{d\alpha}\right\|_2 &\leq \sqrt{8L^2(\eta + D(\alpha)^2)(3d^2\eta^2 + D(\alpha)^4)} \\ &\leq 4L(\sqrt{\eta} + D(\alpha)) \cdot \max(2\eta d, D(\alpha)^2).\end{aligned} \tag{D.17}$$

Assume for contradiction that $D(1) > 2\eta LR$, violating the conclusion of the proposition. By continuity of D, there must have been some $\bar{\alpha} \in (0,1)$ where $D(\bar{\alpha}) = 2\eta LR$, and for all $0 \leq \alpha < \bar{\alpha}$, $D(\alpha) < 2\eta LR$. By the mean value theorem, there then exists $0 \leq \hat{\alpha} \leq \bar{\alpha}$ such that

$$\frac{dD(\hat{\alpha})}{d\alpha} = \frac{D(\bar{\alpha}) - D(0)}{\bar{\alpha}} > \eta LR.$$

On the other hand, by our assumption that $2\eta L^2R^2 \leq 1$, for any $d \geq 1$ it follows that

$$2\eta d \geq 4\eta^2L^2R^2 > D(\hat{\alpha})^2, \ \sqrt{2\eta} \geq 2\eta LR > D(\hat{\alpha}).$$

Then, plugging these bounds into (D.17) and using $\sqrt{\eta} + D(\hat{\alpha}) \leq \frac{5}{2}\sqrt{\eta}$ as $\sqrt{2} \leq \frac{3}{2}$,

$$\frac{d}{d\alpha} D(\hat{\alpha}) \leq 4L \cdot \frac{5}{2}\sqrt{\eta} \cdot 2\eta d = 20\sqrt{\eta}\frac{d}{R} \cdot \eta LR \leq \eta LR.$$

We used $\eta \leq \frac{R^2}{400d^2}$ in the last inequality. This is a contradiction, implying $D(1) \leq 2\eta LR$. □

Appendix E

DEFERRED CONTENTS FROM CHAPTER 6

E.1 Information-theoretic lower bound

In this section, we show that prior information-theoretic lower bounds from [DJWW15] and [GLL22] can be straightforwardly extended to the settings studied by this paper to show that the value oracle complexities used by our algorithms in Sections 6.3 and 6.5 are near-optimal. We first recall some notation from prior work and summarize previous results we will leverage.

Setup. We consider the setting of stochastic optimization where there is a distribution over distributions $\{\mathcal{P}_v\}_v$ indexed by v. An index v is randomly selected, and we consider algorithms interacting with $\mathcal{P}_v$ in one of two different ways. Letting $k \in \mathbf{N}$ and $\mathcal{X} \subset \mathbb{R}^d$, [DJWW15] defined a family of algorithms $\mathbb{A}_k$ such that $\mathcal{A} \in \mathbb{A}_k$ can (adaptively) query a sequence of k values $f(x; s)$ where $x \in \mathcal{X}$ and s is a fresh random sample from $\mathcal{P}_v$. The follow-up work [GLL22] defined another family of algorithms $\mathbb{B}_k$ which takes as input a dataset $\mathcal{D} = \{s_i\}_{i\in[n]}$ and can (adaptively) query a sequence of k values $f(x; s)$ where $x \in \mathcal{X}$ and $s \in \mathcal{D}$. These algorithm families model the SCO and ERM problems stated in Problem 2, without the privacy requirement. In a slight abuse of notation, we denote the output of an algorithm $\mathcal{A} \in \mathbb{A}_k \cup \mathbb{B}_k$ in a SCO or ERM problem corresponding to a distribution $\mathcal{P}$ by $\mathcal{A}(\mathcal{P})$, where $\mathcal{A} \in \mathbb{B}_k$ also depends on the dataset received.

Both [DJWW15, GLL22] let v be drawn uniformly at random from $\mathcal{V} \stackrel{\text{def}}{=} \{-1, 1\}^d$ and let

$$\mathcal{P}_v \stackrel{\text{def}}{=} \mathcal{N}\left(\kappa v, \sigma^2 \operatorname{id}_d\right), \; f(x; s) \stackrel{\text{def}}{=} \langle s, x\rangle$$

for parameters κ, σ to be chosen. We fix this notation throughout this section. For any algorithm $\mathcal{A} \in \mathbb{A}_k \cup \mathbb{B}_k$ corresponding to a set $\mathcal{X}$ and a distribution $\mathcal{P}$, we define the optimality gap

$$\epsilon_k(\mathcal{A}, \mathcal{X}, \mathcal{P}) \stackrel{\text{def}}{=} \mathbb{E}\left[\mathbb{E}_{s\sim\mathcal{P}} f(\mathcal{A}(\mathcal{P}); s)\right] - \min_{x\in\mathcal{X}} \mathbb{E}_{s\sim\mathcal{P}} f(x; s),$$

where the first outer expectation is over any randomness in $\mathcal{A}$, as well as in the samples

used. We also define the minimax risk over a family of distributions P,

$$\epsilon_k^\star(\mathbb{A}_k \cup \mathbb{B}_k, P, \mathcal{X}) \stackrel{\text{def}}{=} \inf_{\mathcal{A} \in \mathbb{A}_k \cup \mathbb{B}_k} \sup_{\mathcal{P} \in P} \epsilon_k\left(\mathcal{A}, \mathcal{P}, \mathcal{X}\right).$$

For $p \in [1, 2]$, we let $P_{G,p}$ denote the family of distributions $\mathcal{P}$ over vectors s such that

$$\mathbb{E}_{s \sim \mathcal{P}} \|s\|_q^2 \le G^2, \text{ where } \frac{1}{p} + \frac{1}{q} = 1.$$

Our lower bounds in this section will be on $\epsilon_k^\star(\mathbb{A}_k \cup \mathbb{B}_k, P_{G,p}, \mathcal{X})$, where $\mathcal{X}$ is a scaled ℓ_p ball. The family $P_{G,p}$ induces random linear functions $\langle s, \cdot \rangle$ with gradient s, and hence $\mathcal{P} \in P_{G,p}$ implies that the induced function $\mathbb{E}_{s \sim \mathcal{P}} \langle s, \cdot \rangle$ has a bounded-variance gradient oracle in the ℓ_p norm via queries to $\mathcal{P}$. We use the following facts from prior work in our proofs.

Lemma 85 (Section 5.1, [DJWW15])**.** *Let $\mathcal{X}$ be the ℓ_p ball of diameter D for $p \in [1, 2]$. For any $v \in \mathcal{V}$ and $x \in \mathcal{X}$, letting $x_v^\star \stackrel{\text{def}}{=} \min_{x \in \mathcal{X}} \mathbb{E}_{s \sim \mathcal{P}_v} f(x; s)$, and letting $\mathbb{1}(\text{sign}(a) = \text{sign}(b))$ be the 0-1 function which is 1 if and only if the signs of a and b agree,*

$$\mathbb{E}_{s \sim \mathcal{P}_v}\left[f(x; s)\right] - \mathbb{E}_{s \sim \mathcal{P}_v}\left[f(x_v^\star; s)\right] \ge \frac{(1 - \frac{1}{p})\kappa D}{2d^{\frac{1}{p}}} \sum_{j \in [d]} \mathbb{1}\left(\text{sign}(x_j) = \text{sign}(v_j)\right).$$

Lemma 85 shows that it suffices to lower bound the expected Hamming distance between the signs of an estimate x and a randomly sampled $-v$. Such a lower bound was given in [DJWW15, GLL22] for estimates returned by $\mathcal{A} \in \mathbb{A}_k \cup \mathbb{B}_k$ via information-theoretic arguments.

Lemma 86 (Section 5.1, [DJWW15], Lemma 7.4, [GLL22])**.** *Let $\mathcal{X}$ be the ℓ_p ball of diameter D, and let $\mathcal{A} \in \mathbb{A}_k \cup \mathbb{B}_k$ be parameterized by $\mathcal{X}$ and $\mathcal{P}_v$. Then*

$$\mathbb{E}_{v \sim_{\text{unif.}} \mathcal{V}}\left[\sum_{j \in [d]} \mathbb{1}(\text{sign}(\mathcal{A}(\mathcal{P}_v)_j) = \text{sign}(v_j))\right] \ge \frac{d}{2}\left(1 - \frac{\kappa\sqrt{k}}{\sigma\sqrt{d}}\right).$$

To lower bound the oracle query complexity of our sampler we use the following standard result.

Lemma 87 ([DKL18], Corollary 1)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be compact and convex, $f : \mathcal{X} \to \mathbb{R}$ be convex, $k > 0$, and π be the density over $\mathcal{X}$ proportional to $\exp(-kf)$. Then,*

$$\mathbb{E}_{x \sim \pi}[f(x)] - \min_{x \in \mathcal{X}} f(x) \le \frac{d}{k}.$$

Lower bounds. We now state three lower bounds generalizing results from [DJWW15, GLL22]. Our results follow straightforwardly from Lemmas 85, 86, and 87 with appropriate parameters.

Proposition 23 (Minimax risk lower bound, $P_{G,p}$)**.** *Let $G, D > 0$, and let $p \in [1, 2]$, $q \geq 2$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Let $\mathcal{X}$ be the ℓ_p ball of diameter D. Then,*

$$\epsilon_k^\star\left(\mathbb{A}_k \cup \mathbb{B}_k, P_{G,p}, \mathcal{X}\right) = \Omega\left(GD \max\left(1 - \frac{1}{p}, \frac{1}{\log d}\right) \min\left(1, \sqrt{\frac{d}{k \log d}}\right)\right).$$

Proof. Throughout the proof, let $\kappa = \frac{\sigma\sqrt{d}}{2\sqrt{k}}$, and let

$$\sigma = \frac{G d^{-\frac{1}{q}}}{\sqrt{\frac{d}{k} + 4 \log d}}. \tag{E.1}$$

By well-known bounds on the expected maximum of d standard Gaussians, we have

$$\begin{aligned}
\mathbb{E}_{s \sim \mathcal{P}_v}\left[\|s\|_q^2\right] &\leq 2\kappa^2 \|v\|_q^2 + 2\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 \mathrm{id}_d)}\left[\|u\|_q^2\right] \\
&\leq 2\kappa^2 d^{\frac{2}{q}} + 2d^{\frac{2}{q}} \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 \mathrm{id}_d)}\left[\|u\|_\infty^2\right] \\
&\leq \sigma^2 d^{\frac{2}{q}}\left(\frac{d}{k} + 4 \log d\right) \leq G^2.
\end{aligned}$$

Hence, $\mathcal{P}_v \in P_{G,p}$ for all $v \in \mathcal{V}$, so it suffices to lower bound $\epsilon_k(\mathcal{A}, \mathcal{P}_v, \mathcal{X})$. Combining Lemmas 85 and 86 with our choices of parameters,

$$\epsilon_k(\mathcal{A}, \mathcal{P}_v, \mathcal{X}) \geq \frac{(1 - \frac{1}{p})\kappa D d^{1 - \frac{1}{p}}}{8} = \Omega\left(GD\left(1 - \frac{1}{p}\right) \min\left(1, \sqrt{\frac{d}{k \log d}}\right)\right).$$

The conclusion then follows because for $p \leq 1 + \frac{1}{\log d}$, choosing a larger value of p only affects problem parameters by constant factors by norm conversions. □

We give a slight extension of Proposition 23 for the family $\overline{P}_{G,p}$ of distributions over linear functions $\langle s, \cdot \rangle$, where s is required to satisfy $\|s\|_q \leq G$ with probability 1, by simply truncating a draw from $\mathcal{P}_v$. This family is compatible with the setting in Problem 2.

Corollary 21 (Minimax risk lower bound, $\overline{P}_{G,p}$)**.** *In the setting of Proposition 23,*

$$\epsilon_k^\star\left(\mathbb{A}_k \cup \mathbb{B}_k, \overline{P}_{G,p}, \mathcal{X}\right) = \Omega\left(GD \max\left(1 - \frac{1}{p}, \frac{1}{\log d}\right) \min\left(1, \sqrt{\frac{d}{k \log(dk)}}\right)\right).$$

Proof. We define a distribution $\overline{\mathcal{P}}_v$ as follows: first $s \sim \mathcal{P}_v$, and then if $\|s\|_q \geq G$, we set $s \leftarrow 0$. By adjusting the logarithmic term in (E.1) to be $O(\log(dk))$, with probability

at most $\text{poly}((dk)^{-1})$, all k draws from $\mathcal{P}_v$ and $\overline{\mathcal{P}}_v$ used are identical by a union bound. Further, due to problem constraints the function error is always at most GD. So, the risk is affected by at most $GD \cdot \text{poly}((dk)^{-1})$. □

Corollary 21 shows that when β in Assumption 1 is polynomially bounded, the value oracle complexities used by Theorem 14 for both DP-SCO and DP-ERM are optimal up to logarithmic factors for the expected excess risk bounds they produce, even without the requirement of privacy. Finally, we show that the value oracle complexity of our sampler in Theorem 13 is also near-optimal.

Corollary 22. *In the setting of Proposition 23, let $r : \mathcal{X} \to \mathbb{R}$ be 1-strongly convex in $\|\cdot\|_p$ with additive range $O(D^2 \min(\log d, \frac{1}{p-1}))$. Let $\mathcal{I}$ be a distribution over i such that all $f_i : \mathcal{X} \to \mathbb{R}$ are G-Lipschitz in $\|\cdot\|_p$, and let $F \stackrel{\text{def}}{=} \mathbb{E}_{i\sim\mathcal{I}} f_i$. No algorithm using $o(\frac{G^2}{\mu}\log^{-4} d)$ value oracle queries to some f_i samples within total variation*

$$o\left(\min\left(\frac{1}{\log d}, \sqrt{\frac{d}{k\log^3(dk)}}\right)\right)$$

of the density proportional to $\exp(-F - \mu r(x))\mathbb{1}_{\mathcal{X}}(x)$.

Proof. Assume for contradiction that $\mathcal{A}$ is an algorithm satisfying the stated criterion using $k = o(\frac{G^2}{\mu}\log^{-4} d)$ value oracle queries, and let F be minimized by $x^\star \in \mathcal{X}$. We choose

$$\mu = \frac{d}{D^2 \min(\log d, \frac{1}{p-1})}.$$

Lemma 87 then shows that the sampled x satisfies

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{A}}\left[F(x)\right] - F(x^\star) &\le \mu\left(r(x^\star) - r(x)\right) + d + GD \cdot o\left(\min\left(\frac{1}{\log d}, \sqrt{\frac{d}{k\log^3(dk)}}\right)\right)\\ &= O(d) + o\left(\frac{GD}{\log d}\min\left(1, \sqrt{\frac{d}{k\log(dk)}}\right)\right).\end{aligned}$$

For the given values of k and μ, this contradicts Corollary 21. □

Corollary 22 implies that for samplers with value query complexity depending polylogarithmically on the total variation distance, $\frac{G^2}{\mu}$ queries are required (up to polylogarithmic factors). This applies to the setting of our sampler in Theorem 13; we also note that the LLT-based regularizers we use in our ℓ_p applications (Section 6.5.2) satisfy the additive range bound in Corollary 22.

E.2 Lower bound on the range of $\psi_{1,1}$

In this section, we provide a lower bound on the range of $\psi_{1,1}$ (6.25) which grows with the dimension d, demonstrating non-scale invariance of our family of LLTs. Recall that $\psi_{1,1}(x)$ is defined by

$$\psi_{1,1}(x) \stackrel{\text{def}}{=} \log\left(\int \exp\left(\langle x, y\rangle - \|y\|_\infty^2\right) \mathrm{d}y\right).$$

Lemma 88. *The additive range of* $\psi_{1,1}$ *over* $\{x \in \mathbb{R}^d mid \|x\|_1 \leq 1\}$ *is* $\Omega(\sqrt{d})$.

Proof. Throughout the proof denote for simplicity $\psi := \psi_{1,1}$ and let

$$\mathcal{D}_x^\varphi(y) \propto \exp\left(\langle x, y\rangle - \|y\|_\infty^2\right).$$

Then, following (6.26), we can write $\psi(x) - \psi(0)$ as

$$\psi(x) - \psi(0) = \log\left[\mathbb{E}_{y\sim\mathcal{D}_0^\varphi} \exp(\langle x, y\rangle)\right],$$

where $\mathcal{D}_0^\varphi \propto \exp(-\|y\|_\infty^2)$. Let π be the probability density on $\mathbb{R}_{\geq 0}$ such that

$$\mathrm{d}\pi(r) \propto r^{d-1}\exp(-r^2)\mathrm{d}r.$$

Here, $\mathrm{d}\pi(r)$ is the density of the scalar quantity $r = \|y\|_\infty$ for $y \sim \mathcal{D}_0^\varphi$. Note that the distribution of y conditioned on $\|y\|_\infty = r$ is uniform over the surface of the ℓ_∞ ball, where one random coordinate is set to $\pm r$, and the remaining coordinates are uniform on a $d-1$ dimensional hypercube with side length r. We denote this distribution as $\mathcal{P}_r$, and write

$$\begin{aligned}\mathbb{E}_{y\sim\mathcal{D}_0^\varphi} \exp(\langle x, y\rangle) &= \mathbb{E}_{r\sim\pi}\left[\mathbb{E}_{y\sim\mathcal{P}_r} \exp(\langle x, y\rangle)\right] \\ &= \mathbb{E}_{r\sim\pi}\left[\frac{1}{d}\sum_{i^\star\in[d]}\frac{1}{2}\sum_{y_{i^\star}\in\{-r,r\}}\exp(x_{i^\star}y_{i^\star})\prod_{i\neq i^*}\int_{-r}^{r}\frac{1}{2r}\exp(x_i y_i)\mathrm{d}y_i\right].\end{aligned}$$

Let $x = e_1$ and $g_{i^\star}^{(r)} = \exp(x_{i^\star}r)\prod_{i\neq i^\star}\int_{-r}^{r}\frac{1}{2r}\exp(x_i y_i)\mathrm{d}y_i$. Then,

$$\mathbb{E}_{y\sim\mathcal{D}_0^\varphi} \exp(\langle x, y\rangle) \geq \frac{1}{2d}\sum_{i^\star\in[d]}\mathbb{E}_{r\sim\pi(r)} g_{i^\star}^{(r)}$$

since this drops terms where $y_{i^\star} = -r$. When $i^\star = 1$, we have $g_{i^\star}^{(r)} = \exp(r)$. When $i^\star \neq 1$, we have

$$g_{i^\star}^{(r)} = \int_{-r}^{r}\frac{1}{2r}\exp(y_1)\mathrm{d}y_1 = \frac{1}{2r}\left(\exp(r) - \exp(-r)\right).$$

Now, consider $r_1 = \sqrt{\frac{d-1}{2}}$. For any $r \le r_1$, $\frac{\mathrm{d}}{\mathrm{d}r}[(d-1)\log r - r^2] = \frac{d-1}{r} - 2r \ge 0$. Thus, we have

$$I := \int_0^{\frac{1}{2}r_1} \exp((d-1)\log r - r^2)\mathrm{d}r \le \int_{\frac{1}{2}r_1}^{r_1} \exp((d-1)\log r - r^2)\mathrm{d}r. \tag{E.2}$$

Letting $Z \overset{\text{def}}{=} \int_0^\infty \exp((d-1)\log r - r^2)\mathrm{d}r$, (E.2) shows that

$$\int_{\frac{1}{2}r_1}^{\infty} \exp((d-1)\log r - r^2)\mathrm{d}r = Z - I \ge Z - \frac{1}{2}Z = \frac{1}{2}Z.$$

Then, for all $i^\star \in [d]$,

$$\begin{aligned}
\mathbb{E}_{r\sim\pi} g_{i^\star} &= \frac{\int_0^\infty \exp((d-1)\log r - r^2) g_{i^\star}^{(r)} \mathrm{d}r}{Z} \\
&\ge \frac{\int_{\frac{1}{2}r_1}^\infty \exp((d-1)\log r - r^2) g_{i^\star}^{(r)} \mathrm{d}r}{Z} \\
&\ge \frac{2\int_{\frac{1}{2}r_1}^\infty \exp((d-1)\log r - r^2) g_{i^\star}^{(r)} \mathrm{d}r}{\int_{\frac{1}{2}r_1}^\infty \exp((d-1)\log r - r^2)\mathrm{d}r} \\
&\ge 2 \min_{r \ge r_1} \exp(r - \log(4r)) = 2\exp(r_1 - \log(4r_1)).
\end{aligned}$$

The fourth step follows from $g_{i^\star}^{(r)} \ge \frac{1}{4r}\exp(r)$ for $r \ge r_1$. The last step follows from $r - \log 4r$ increases on $r \ge r_1$. Combining with $\mathbb{E}_{y\sim\mathcal{P}_0}\exp(\langle x, y\rangle) \ge \frac{1}{2d}\sum_{i^\star\in[d]} \mathbb{E}_{r\sim\pi(r)} g_{i^\star}$,

$$\psi(x) - \psi(0) = \log \mathbb{E}_{y\sim\mathcal{P}_0}\exp(\langle x, y\rangle) \ge \log\left(\frac{d-1}{d}\exp(r_1 - \log(4r_1))\right) = \Omega(\sqrt{d}).$$

□

E.3 Deferred proofs from Section 6.4

Lemma 49. *For λ defined in* (6.23)*,*

$$\mathbb{E}\left[|\lambda| \mathbb{1}_{\rho\notin[0,2]}\right] \le \frac{\delta}{4}.$$

Proof. Clearly, it suffices to show $\mathbb{E}|\lambda| \le \frac{\delta}{4}$. Define random variables,

$$\Delta_i \overset{\text{def}}{=} |f_i(x_2) - f_i(x_1)|,\ \Delta \overset{\text{def}}{=} \mathbb{E}_{i\sim\mathcal{I}}\Delta_i,$$

whose randomness comes from $x_1, x_2 \sim \gamma_y$. By definition,

$$\mathbb{E}|\lambda| = \sum_{b>H} \frac{1}{b!}\mathbb{E}_{x_1,x_2\sim\gamma}[\Delta]^B.$$

Define $\Phi(t) := \sum_{b>H} \frac{t^b}{b!}$. For $H = \lceil 10 \log \frac{1}{\delta} \rceil$, it is straightforward to check $\Phi(t) \le \frac{\delta}{16}$ for any $|t| \le 1$, and for all nonnegative t, $\Phi(t) \le \exp(t)$. Hence, letting p_Δ be the density of Δ,

$$\begin{aligned}\mathbb{E}|\lambda| &\le \frac{\delta}{16} + \mathbb{E}[\mathbb{1}_{\Delta>1} e^{\Delta}] \le \frac{\delta}{16} + \int_1^\infty \exp\left(\lceil \Delta \rceil\right) p_\Delta(\Delta) \mathrm{d}\Delta \\ &\le \frac{\delta}{16} + \sum_{k \ge 1} \exp(k+1) \Pr_{x_1, x_2 \sim \gamma}[\Delta \ge k].\end{aligned} \tag{E.3}$$

It now suffices to bound on $\Pr[\Delta \ge k]$. Define a function $h_{x_1,x_2}(k) := \Pr_{i \sim \mathcal{I}}[|f_i(x_1) - f_i(x_2)| \ge k]$. Since each f_i is G-Lipschitz, and γ_y is $\frac{1}{12\eta}$-strongly logconcave in by Lemma 40, by Lemma 47:

$$\mathbb{E}_{x_1,x_2}[h_{x_1,x_2}(k)] = \Pr_{x_1,x_2,i \sim \mathcal{I}}[|f_i(x_1) - f_i(x_2)| \ge k] \le 4 \exp\left(-\frac{k^2}{96\eta G^2}\right),$$

and so by Markov's inequality we have

$$\Pr_{x_1,x_2}[h_{x_1,x_2}(k) \ge e^{-t}] \le 4 \exp\left(t - \frac{k^2}{96\eta G^2}\right). \tag{E.4}$$

For fixed x_1, x_2, as each f_i is G-Lipschitz in $\|\cdot\|_{\mathcal{X}}$, $|f_i(x_1) - f_i(x_2)| \le G \|x_1 - x_2\|_{\mathcal{X}}$, and hence

$$\mathbb{E}_{i \sim \mathcal{I}}[|f_i(x_1) - f_i(x_2)|] \le \min_{k \ge 0} k + h_{x_1,x_2}(k) \cdot G \|x_1 - x_2\|_{\mathcal{X}}.$$

This then shows that if for some k, $h_{x_1,x_2}(k) \le \exp(-\frac{k^2}{192\eta G^2})$,

$$\mathbb{E}_{i \sim \mathcal{I}}[|f_i(x_1) - f_i(x_2)|] \le k + \exp\left(-\frac{k^2}{192\eta G^2}\right) \cdot G \|x_1 - x_2\|_{\mathcal{X}},$$

which implies via (E.4) that

$$\begin{aligned}&\Pr_{x_1,x_2}\left[\Delta \ge k + \exp\left(-\frac{k^2}{192\eta G^2}\right) \cdot G \|x_1 - x_2\|_{\mathcal{X}}\right] \\ \le\ &\Pr_{x_1,x_2}\left[h_{x_1,x_2}(k) \ge \exp\left(-\frac{k^2}{192\eta G^2}\right)\right] \le 4 \exp\left(-\frac{k^2}{192\eta G^2}\right).\end{aligned} \tag{E.5}$$

Further, since $\|x_1 - \mathbb{E}x_1\|_{\mathcal{X}}$ is a 1-Lipschitz function in x_1 with a nonnegative mean, by Lemma 47,

$$\Pr\left[\|x_1 - x_2\|_{\mathcal{X}} \ge k\right] \le 2 \Pr\left[\|x_1 - \mathbb{E}x_1\|_{\mathcal{X}} \ge k\right] \le 2 \exp\left(-\frac{k^2}{96\eta G^2}\right). \tag{E.6}$$

Combining (E.5) and (E.6),

$$\begin{aligned}\Pr_{x_1,x_2}[\Delta \ge 2k] &= \Pr_{x_1,x_2}\left[\Delta \ge 2k \wedge \|x_1 - x_2\|_{\mathcal{X}} \ge \frac{k}{G}\right] + \Pr_{x_1,x_2}\left[\Delta \ge 2k \wedge \|x_1 - x_2\|_{\mathcal{X}} \le \frac{k}{G}\right] \\ &\le 2 \exp\left(-\frac{k^2}{96\eta G^2}\right) + \Pr_{x_1,x_2}\left[\Delta \ge k + \exp\left(-\frac{k^2}{192\eta G^2}\right) G \|x_1 - x_2\|_{\mathcal{X}}\right] \\ &\le 6 \exp\left(-\frac{k^2}{192\eta G^2}\right).\end{aligned} \tag{E.7}$$

Plugging (E.7) into (E.3), and using $\eta^{-1} \geq 10^4 G^2 \log \frac{1}{\delta}$, we have the desired

$$\mathbb{E}(|\lambda| \mathbb{1}_{\rho \notin [0,2]}) \leq \frac{\delta}{16} + \sum_{k=1}^{\infty} 6 \exp\left(k - \frac{k^2}{768 \eta G^2}\right) \leq \frac{\delta}{4}.$$

□

Lemma 50. *For σ defined in* (6.23)*,*

$$\mathbb{E}\left[|\sigma| \mathbb{1}_{\rho \notin [0,2]}\right] \leq \frac{\delta}{4}.$$

Proof. We begin by bounding, analogously to (E.3),

$$\mathbb{E}[|\sigma| \mathbb{1}_{\rho \notin [0,2]}] \leq 2^H \Pr[\rho \notin [0,2]] + \sum_{k \geq 1} \Pr\left[|\sigma| > 2^{kH}\right] 2^{(k+1)H}. \tag{E.8}$$

Recall when $a \leq H$, $|\mathcal{J}| \leq \frac{1}{2} H^2$. By a union bound over Lemma 47,

$$\Pr_{x_1, x_2}\left[|f_i(x_1) - f_i(x_2)| \geq \frac{2^k}{3} \; \forall i \in \mathcal{J}\right] \leq H^2 \exp\left(-\frac{4^k}{864 \eta G^2}\right).$$

If for each $i \in \mathcal{J}$, $|f_i(x_1) - f_i(x_2)| \leq \frac{2^k}{3}$, we have for $k \geq 1$

$$|\sigma| = \sum_{b=0}^{H} \mathbb{1}_{a \geq b} \prod_{i \in [b]} (f_{j_{i,b}}(x_2) - f_{j_{i,b}}(x_1)) \leq 1 + \sum_{b=1}^{H} \left(\frac{2^k}{3}\right)^b \leq 2^{kH},$$

which implies that $\Pr[|\sigma| \geq 2^{kH}] \leq H^2 \exp(-\frac{4^k}{864 \eta G^2})$ and hence using our choice of $\eta \leq \frac{1}{500 G^2 H}$,

$$\begin{aligned} \sum_{k=1}^{\infty} 2^{(k+1)H} \Pr\left[|\sigma| > 2^{kH}\right] &\leq \sum_{k=1}^{\infty} 2^{(k+1)H} H^2 \exp\left(-\frac{4^k}{864 \eta G^2}\right) \\ &\leq \sum_{k=1}^{\infty} 2^{4kH} \exp(-2 \cdot 4^k H) \leq \sum_{k=1}^{\infty} 2^{-kH} \leq \frac{\delta}{8}. \end{aligned} \tag{E.9}$$

It remains to bound $\Pr[\rho \notin [0,2]]$. Recall $\Pr[a > H] \leq \frac{1}{H!}$ so since $a \leq H \implies \sigma = \rho$, $\Pr[\rho \notin [0,2]] \leq \frac{1}{H!} + \Pr[\sigma \notin [0,2]]$. Next, by a union bound over Lemma 47 and $\frac{1}{2} H^2$ indices in $\mathcal{J}$,

$$\Pr_{x_1, x_2}\left[|f_i(x_1) - f_i(x_2)| \geq \frac{1}{2} \; \forall i \in \mathcal{I}\right] \leq 2H^2 \exp\left(-\frac{1}{384 \eta G^2}\right).$$

Under the event that $|f_i(x_1) - f_i(x_2)| \leq \frac{1}{2}$ for all $i \in \mathcal{I}$, $0 \leq \sigma \leq 2$ by definition. Hence we know $\Pr[\sigma \notin [0,2]] \leq 2H^2 \exp(-\frac{1}{384 \eta G^2})$ and by our setting that $H > 10 \log \frac{1}{\delta}$, we have

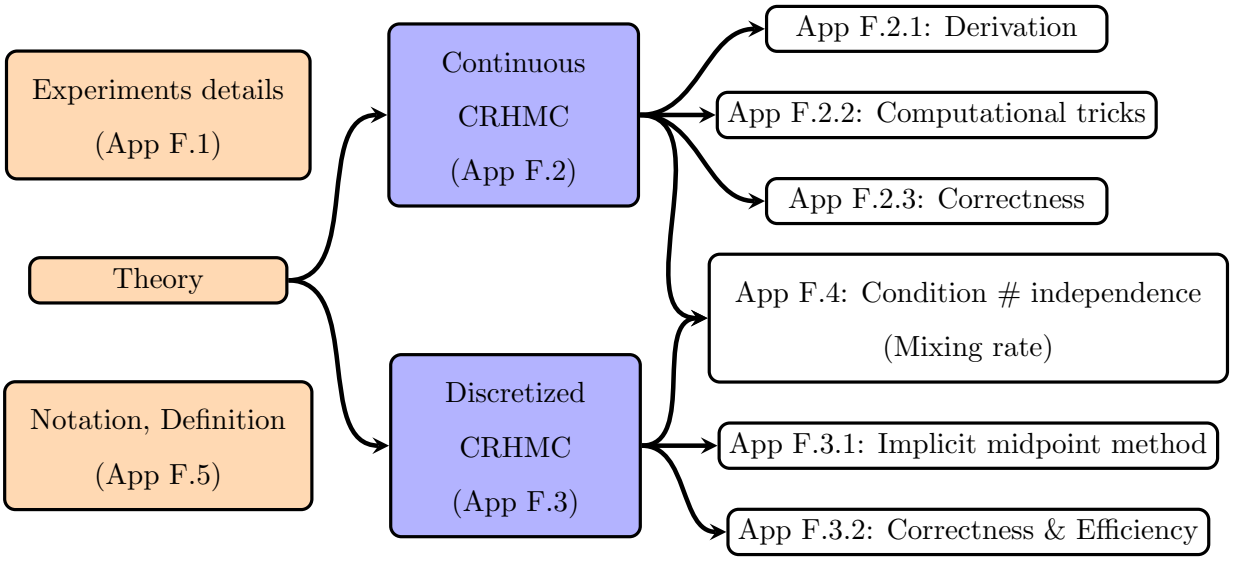
$$\Pr[\rho \notin [0,2]] \cdot 2^H \leq 2^H \left(2H^2 \exp\left(-\frac{1}{384 \eta G^2}\right) + \frac{1}{H!}\right) \leq \frac{\delta}{8}. \tag{E.10}$$

Combining (E.8), (E.9) and (E.10) completes the proof. □

Appendix F

DEFERRED CONTENTS FROM CHAPTER 7

F.1 Additional Experiment Details



Dataset. We summarize in Table F.1 the dataset used in experiments. If a model is unbounded, we make it bounded by setting $l = \max(l, -10^7)$ and $u = \min(u, 10^7)$. As existing packages require full-dimensional representations of polytopes (i.e., $\{x : A'x \leq b'\}$), we transformed all constraint-based models to prepare instances for them as follows: (1) first preprocess each model by removing redundant constraints and appropriately scaling it, (2) find its corresponding full-dimensional description, and (3) round it via the maximum volume ellipsoid (MVE) algorithm making the polytope more amenable to sampling. We note that a full-dimensional polytope can be transformed into a constraint-based polytope and vice versa, so CRHMC can be run on either representation.

Preprocessing. We preprocessed each constrained-based model prior to sampling. This preprocessing consists mainly of simplifying polytopes, scaling properly for numerical stability, and finding a feasible starting point. To simplify a given polytope, we check if $l_i = u_i$ for each $i \in [n]$ and then incorporate such variables x_i into $Ax = b$. Any dense

Bio Model	Full-dim	Consts (m)	Vars (n)	nnz
ecoli	24	72	95	291
cardiac_mit	12	230	220	228
Aci_D21	103	856	851	1758
Aci_MR95	123	917	994	2859
Abi_49176	157	952	1069	2951
Aci_20731	164	1009	1090	2946
Aci_PHEA	328	1319	1561	4640
iAF1260	572	1668	2382	6368
iJO1366	590	1805	2583	7284
Recon1	932	2766	3742	8717
Recon2	2430	5063	7440	19791
Recon3	5335	8399	13543	48187

LP Model	Full-dim	Consts (m)	Vars (n)	nnz
israel	142	174	316	2519
gfrd_pnc	544	616	1160	2393
25fv47	1056	821	1876	10566
pilot_ja	1002	940	2267	11886
sctap2	1410	1090	2500	7334
ship08l	2700	778	4363	9434
cre_a	3703	3516	7248	17368
woodw	4656	1098	8418	23158
80bau3b	9233	2262	12061	22341
ken_18	49896	105127	154699	295946

Table F.1: Constraint-based models. Each constraint-based model has a form of $\{x \in \mathbb{R}^n : Ax = b, l \leq x \leq u\}$ for $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$ and $l, u \in \mathbb{R}^n$, where the rows and columns correspond to constraints and variables respectively. The full-dimension of each model is obtained by transforming its degenerate subspace to a full dimensional representation (i.e., $A'x \leq b'$), and we count the number of nonzero (nnz) entries of a preprocessed matrix A.

column is split into several columns with less non-zero entries by introducing additional variables. Then we remove dependent rows of A by the Cholesky decomposition. Then we find the Dikin ellipsoid of the polytope. If the width along some axis is smaller than a preset tolerance, then we fix variables in such directions, reducing columns of A. Lastly, we run the primal-dual interior-point method with the log-barrier to find an analytic center of the polytope, which will be used as a starting point in sampling. When finding the analytic center of the simplified polytope, if a coordinate of the analytic center is too close to a boundary (to be precise, smaller than a preset tolerance boundary 10^{-8}), then we assume that the inequality constraint (either $x_i \leq u_i$ or $l_i \leq x_i$) is tight, and we collapse such a variable by moving it into the constraints $Ax = b$. We go back to the step for removing dependent rows and repeat until no more changes are made to A. Along with simplification, we keep rescaling A, b, l, u for numerical stability.

Coordinate Hit-and-Run (CDHR). We briefly explain how CHRR works. First, rounding via the MVE algorithm finds the maximum volume ellipsoid inscribed in the polytope and applies, to the polytope, an affine transformation that makes this ellipsoid a unit ball. This procedure puts a possibly highly-skewed polytope into John's position, which guarantees that the polytope contains a unit ball and is contained in a ball of radius n. This position still has a beneficial effect on sampling in practice in the sense that the random walk can converge in fewer steps. After the transformation, the random walk based on Coordinate Hit-and-Run (CHAR) chooses a random coordinate and moves to a random point on the line through the current point along the chosen coordinate.

When running CHRR and CDHR, we recorded a sample every n^2 steps. The mixing rate (i.e., the number of steps required to get a sample from a target distribution) of Hit-and-Run (HAR), a general version of CHAR choosing a random direction (unit vector) instead of a random coordinate, is $O^*(n^2R^2)$ for a polytope P with $B_n \subseteq P \subseteq R \cdot B_n$, where B_n is the unit ball in $\mathbb{R}^n$ [LV06a]. It was proved only recently that CHAR mixes in $O^*(n^9R^2)$ steps on such a polytope [LV21, NS21]. Even though this bound is not as tight as the mixing-rate bound for HAR, it was reported in [HCT$^+$17] that CHRR mixes in the same number of steps as HAR empirically. Moreover, the per-step complexity of CHAR can be n times faster than that of HAR, so CHAR brings a significant speed-up in practice.

Comparison Setup. We set the parameters of CRHMC to values in `default_options.m` in the experiments. For the competitors, we proceeded with the following additional steps for fair comparison. First, as the VolEsti package does not support the MVE rounding, we rounded each polytope by the MVE algorithm in the CHRR package and then transformed the rounded polytope so that the R interface can read the data file. Next, we limited all algorithms to a single core, since the R interface uses a single core as a default whereas MATLAB uses as many available cores as possible.

F.1.1 Polytope Definition

Hypercube. The n-dimensional hypercube is defined by $\{x \in \mathbb{R}^n : -\frac{1}{2} \leq x_i \leq \frac{1}{2} \text{ for all } i \in [n]\}$. Note that it has no equality constraint and its full-dimension is n.

Simplex. The n-dimensional simplex is defined by $\{x \in \mathbb{R}^n : 0 \leq x_i \text{ for all } i \in [n], \sum_{i=1}^n x_i = 1\}$. Note that its full-dimension is $n-1$.

Birkhoff Polytope. The n^{th} Birkhoff polytope B_n is the set of all doubly stochastic $n \times n$ matrices (or the convex hull of all permutation matrices), which is defined as

$$B_n = \{(X_{ij})_{i,j\in[n]} : \sum_j X_{ij} = 1 \text{ for all } i \in [n], \sum_i X_{ij} = 1 \text{ for all } j \in [n], \text{ and } X_{ij} \geq 0\}.$$

Namely, B_n is defined in a constrained $\mathbb{R}^{n^2}$-dimensional space, and its full-dimension is $n^2 - (2n-1) = (n-1)^2$. We ran CRHMC on $B_{\sqrt{n}}$ to examine its efficiency on (roughly) n-dimensional Birkhoff polytope.

F.2 Deferred details of CRHMC

In this section, we present all technical details behind an *idealized* version of our algorithm, CRHMC, together with correctness of CRHMC. Subsequently in Appendix F.3, we provide details on a *discretized* version of CRHMC.

F.2.1 Deferred details of Section 7.2.1

Recall that in Section 7.2.1 we mention that the following constrained Hamiltonian satisfies the Hamiltonian ODE $\left(\frac{dx}{dt} = \frac{\partial H(x,v)}{\partial v},\ \frac{dv}{dt} = -\frac{\partial H(x,v)}{\partial x}\right)$:

$$H(x,v) = \overline{H}(x,v) + \lambda(x,v)^\top c(x) \quad \text{with} \quad \overline{H}(x,v) = f(x) + \frac{1}{2} v^\top M(x)^\dagger v + \log \text{pdet}(M(x))$$

where

$$\lambda(x,v) = (Dc(x)Dc(x)^\top)^{-1}\left(D^2c(x)[v, \frac{dx}{dt}] - Dc(x)\frac{\partial \overline{H}(x,v)}{\partial x}\right).$$

Lemma 15. *Consider the constrained Hamiltonian defined by (7.7) with* $\text{Range}(M(x)) = \text{Null}(Dc(x))$ *and*

$$\lambda(x_t,v_t) = (Dc(x_t)Dc(x_t)^\top)^{-1}\left(D^2c(x_t)[v_t, \frac{dx_t}{dt}] - Dc(x_t)\frac{\partial \overline{H}(x_t,v_t)}{\partial x}\right).$$

When the initial point satisfies $c(x_0) = 0$, the ODE solution of (7.4) satisfies $c(x_t) = 0$ and $Dc(x_t)v_t = Dc(x_0)v_0$ for all t.

Proof. First we compute

$$\frac{d}{dt}c(x_t) = Dc(x_t)\cdot\frac{dx_t}{dt} = Dc(x_t)\cdot\frac{\partial H(x_t,v_t)}{\partial v_t}$$

$$= Dc(x_t)M(x_t)^\dagger v + Dc(x_t)D_v\lambda(x_t, v_t)^\top c(x_t)$$
$$= Dc(x_t)D_v\lambda(x_t, v_t)^\top c(x_t)$$

where we used $\mathrm{Range}(M(x)^\dagger) = \mathrm{Range}(M(x)) = \mathrm{Null}(Dc(x))$. Since $c(x_0) = 0$, by the uniqueness of the ODE solution, we have that $c(x_t) = 0$ for all t. Next we compute

$$\begin{aligned}\frac{dv_t}{dt} &= -\frac{\partial H(x_t, v_t)}{\partial x} \\ &= -\frac{\partial \overline{H}(x_t, v_t)}{\partial x} - Dc(x_t)^\top \lambda(x_t, v_t) - D_x\lambda(x_t, v_t)^\top c(x_t) \\ &= -\frac{\partial \overline{H}(x_t, v_t)}{\partial x} - Dc(x_t)^\top \lambda(x_t, v_t)\end{aligned}$$

where we used $c(x_t) = 0$. Hence, we have

$$\begin{aligned}\frac{d}{dt}Dc(x_t)v_t &= D^2c(x_t)[v_t, \frac{dx_t}{dt}] + Dc(x_t)\frac{dv_t}{dt} \\ &= D^2c(x_t)[v_t, \frac{dx_t}{dt}] - Dc(x_t)\frac{\partial \overline{H}(x_t, v_t)}{\partial x} - Dc(x_t)Dc(x_t)^\top \lambda(x_t, v_t).\end{aligned}$$

By setting $\lambda(x_t, v_t) = (Dc(x_t)Dc(x_t)^\top)^{-1}(D^2c(x_t)[v_t, \frac{dx_t}{dt}] - Dc(x_t)\frac{\partial \overline{H}(x_t, v_t)}{\partial x})$, we have $\frac{d}{dt}Dc(x_t)v_t = 0$ and $Dc(x_t)v_t = Dc(x_0)v_0$ for all t (i.e., $v_t \in \mathrm{Null}(Dc(x_t))$ during Step 2). □

F.2.2 Deferred details of Section 7.2.2

In Section 7.2.2, we mention that a naive algorithm computing $\partial H/\partial x$ and $\partial H/\partial v$ is bound to face the following challenges, especially in high-dimensional regime, and briefly explain how we address each of them. In this section, we give full details on our computational tricks.

1. Computation of the pseudo-inverse and its derivatives takes $O(n^3)$, except for very special matrices $\Longrightarrow$ Find equivalent formulas (Appendix F.2.2).

2. The Lagrangian term in the constrained Hamiltonian entails extra computation such as $D^2c(x)$ $\Longrightarrow$ Simplify the constrained Hamiltonian (Appendix F.2.2).

3. A naive approach to computing leverage scores in $\partial H/\partial x$ results in a very dense matrix $\Longrightarrow$ Track sparsity pattern (Appendix F.2.2).

Avoiding pseudo-inverse and pseudo-determinant

We start with a formula for $M(x)^\dagger$.

Lemma 16. *Let* $M(x) = Q(x) \cdot g(x) \cdot Q(x)$ *where* $Q(x) = I - Dc(x)^\top (Dc(x) \cdot Dc(x)^\top)^{-1} Dc(x)$ *is the orthogonal projection to the null space of* $Dc(x)$*. Then,* $Dc(x) \cdot M(x)^\dagger = 0$ *and* $M(x)^\dagger = g(x)^{-\frac{1}{2}} \cdot (I - P(x)) \cdot g(x)^{-\frac{1}{2}}$ *with*

$$P(x) = g(x)^{-\frac{1}{2}} \cdot Dc(x)^\top (Dc(x) \cdot g(x)^{-1} \cdot Dc(x)^\top)^{-1} Dc(x) \cdot g(x)^{-\frac{1}{2}}.$$

Proof. Recall that $\text{Range}(M(x)^\dagger) = \text{Range}(M(x))$. Hence, for any $u \in \mathbb{R}^n$, we have that $M(x)^\dagger u \in \text{Range}(M(x))$. Since $\text{Range}(M(x)) \subseteq \text{Range}(Q(x))$ and $\text{Range}(Q(x)) = \text{Null}(Dc(x))$ due to the definition of the orthogonal projection $Q(x)$, it follows that $Dc(x) \cdot M(x)^\dagger u = 0$ for all u.

For the formula of $M(x)^\dagger$, we simplify the notation by ignoring the parameter x. Let $N = g^{-\frac{1}{2}} P g^{-\frac{1}{2}}$ and $J = Dc(x)$. The goal is to prove that $M^\dagger = N$. First, we show some basic identities about Q and N:

$$\begin{aligned}
QN =& Qg^{-\frac{1}{2}}(I - g^{-\frac{1}{2}} J^\top (Jg^{-1}J^\top)^{-1} J g^{-\frac{1}{2}}) g^{-\frac{1}{2}} \\
=& (I - J^\top (JJ^\top)^{-1} J)(g^{-1} - g^{-1} J^\top (Jg^{-1}J^\top)^{-1} J g^{-1}) \\
=& g^{-1} - J^\top (JJ^\top)^{-1} J g^{-1} \\
& - (g^{-1} J^\top (Jg^{-1}J^\top)^{-1} J g^{-1} - J^\top (JJ^\top)^{-1} J g^{-1} J^\top (Jg^{-1}J^\top)^{-1} J g^{-1}) \\
=& N. \qquad \text{(F.1)}
\end{aligned}$$

Similarly, we have $NQ = N$, $QgN = Q$, and $NgQ = Q$. To prove that $M^\dagger = N$, we need to check that MN and NM are symmetric, $MNM = M$, and $NMN = N$.

For symmetry of MN and NM, we note that $MN = QgQN = QgN = Q$ and $NM = NQgQ = NgQ = Q$. For the formula of MNM and NMN, we note that that Q is a projection matrix and hence

$$\begin{aligned}
MNM = QM = QQgQ = QgQ = M, \\
NMN = QN = N.
\end{aligned}$$

Therefore, we have $M^\dagger = N$. □

Another bottleneck of the algorithm is to compute $\log \text{pdet} M(x)$. The next lemma shows a simpler formula that can take advantage of sparse Cholesky decomposition.

Lemma 17. *We have that*

$$\log \operatorname{pdet}(M(x)) = \log\det g(x) + \log\det\Big(Dc(x)\cdot g(x)^{-1}\cdot Dc(x)^\top\Big) - \log\det\Big(Dc(x)\cdot Dc(x)^\top\Big).$$

Proof. We simplify the notation by ignoring the parameter x and letting $J = Dc(x)$. Let

$$\begin{aligned} f_1(g) &= \log\operatorname{pdet}(Q\cdot g\cdot Q),\\ f_2(g) &= \log\det g + \log\det Jg^{-1}J^\top - \log\det JJ^\top. \end{aligned}$$

Clearly, $f_1(I) = f_2(I) = 0$, and hence it suffices to prove that their derivatives are the same.

Note that $\operatorname{Range}(Q\cdot g\cdot Q) = \operatorname{Null}(J)$ and $\operatorname{Range}(J^\top)$ is the orthogonal complement of $\operatorname{Null}(J)$. Since $J^\top(JJ^\top)^{-1}J$ is the orthogonal projection to $\operatorname{Range}(J^\top)$, all of its eigenvectors in $\operatorname{Range}(J^\top)$ have eigenvalue 1 and all the rest in $\operatorname{Null}(J)$ have eigenvalue 0. Therefore, by padding eigenvalue 1 on $\operatorname{Range}(J^\top) = \operatorname{Null}(J)^\perp = \operatorname{Range}(QgQ)^\perp$, we have

$$\begin{aligned} \operatorname{pdet}(Q\cdot g\cdot Q) &= \det(Q\cdot g\cdot Q + J^\top(JJ^\top)^{-1}J)\\ &= \det(Q\cdot g\cdot Q + (I-Q)). \end{aligned}$$

Using $D\log\det A(g)[u] = \operatorname{Tr}(A(g)^{-1}DA(g)[u])$, the directional derivative of f_1 on direction u is

$$Df_1(g)[u] = \operatorname{Tr}\big((Q\cdot g\cdot Q + (I-Q))^{-1}Q\cdot u\cdot Q\big).$$

Let $N = (Q\cdot g\cdot Q)^\dagger$. As shown in the proof of Lemma 16, we have $NQ = QN = N$ and $QgN = Q$. By using these identities, we can manually check that $(Q\cdot g\cdot Q + (I-Q))^{-1} = N + (I-Q)$. Hence,

$$\begin{aligned} Df_1(g)[u] &= \operatorname{Tr}\left((N + (I-Q))Q\cdot u\cdot Q\right) = \operatorname{Tr}(NuQ)\\ &= \operatorname{Tr}(QNu) = \operatorname{Tr}(Nu) \end{aligned}$$

where we used idempotence of the projection matrix Q (i.e., $Q^2 = Q$).

On the other hand, we have

$$\begin{aligned} Df_2(g)[u] &= \operatorname{Tr}(g^{-1}u) - \operatorname{Tr}\Big((Jg^{-1}J^\top)^{-1}(Jg^{-1}ug^{-1}J^\top)\Big)\\ &= \operatorname{Tr}\Big((g^{-1} - g^{-1}J^\top(Jg^{-1}J^\top)^{-1}Jg^{-1})u\Big)\\ &= \operatorname{Tr}(Nu) \end{aligned}$$

where we used the alternative formula of N in Lemma 16. This shows that the derivative of f_1 equals to that of f_2 at any point $g \succ 0$. Since the set of positive definite matrices is connected and $f_1(I) = f_2(I)$, this implies that $f_1(g) = f_2(g)$ for all $g \succ 0$. □

Combining Lemma 16 and Lemma 17, we have the following formula of the Hamiltonian.

$$\begin{aligned}\overline{H}(x,v) =& H_0(x,v) + \lambda(x,v)^\top c(x),\\ H_0(x,v) =& f(x) + \frac{1}{2}v^\top g(x)^{-\frac{1}{2}}\left(I - g(x)^{-\frac{1}{2}}\cdot Dc(x)^\top (Dc(x)\cdot g(x)^{-1}\cdot Dc(x)^\top)^{-1} Dc(x)\cdot g(x)^{-\frac{1}{2}}\right) g(x)^{-\frac{1}{2}}v\\ &+ \frac{1}{2}\left(\log\det g(x) + \log\det\left(Dc(x)\cdot g(x)^{-1}\cdot Dc(x)^\top\right) - \log\det\left(Dc(x)\cdot Dc(x)^\top\right)\right).\end{aligned}$$

Simplification for subspace constraints

For the case $c(x) = Ax - b$, the constrained Hamiltonian is

$$\overline{H}(x,v) = f(x) + \frac{1}{2}v^\top g^{-\frac{1}{2}}(I-P)\,g^{-\frac{1}{2}}v + \frac{1}{2}\left(\log\det g + \log\det Ag^{-1}A^\top - \log\det AA^\top\right) + \lambda^\top c \tag{F.2}$$

where $P = g^{-\frac{1}{2}}A^\top(Ag^{-1}A^\top)^{-1}Ag^{-\frac{1}{2}}$. The following lemma shows that the dynamics corresponding to $\overline{H}$ above is equivalent to a simpler Hamiltonian. The key observation is that the algorithm only needs to know $x(h)$ in the HMC dynamics, and not $v(h)$. Thus we can replace $\overline{H}$ by any other H that produces the same $x(h)$.

Lemma 18. *The Hamiltonian dynamics of x corresponding to (F.2) is same as the dynamics of x corresponding to*

$$H(x,v) = f(x) + \frac{1}{2}v^\top g^{-\frac{1}{2}}(I-P)\,g^{-\frac{1}{2}}v + \frac{1}{2}\left(\log\det g + \log\det Ag^{-1}A^\top\right) \tag{F.3}$$

where $P = g^{-\frac{1}{2}}A^\top(Ag^{-1}A^\top)^{-1}Ag^{-\frac{1}{2}}$. Furthermore, we have

$$\frac{dx}{dt} = g^{-\frac{1}{2}}(I-P)\,g^{-\frac{1}{2}}v, \tag{F.4}$$

$$\frac{dv}{dt} = -\nabla f(x) + \frac{1}{2}Dg\left[\frac{dx}{dt},\frac{dx}{dt}\right] - \frac{1}{2}\mathrm{Tr}(g^{-\frac{1}{2}}(I-P)\,g^{-\frac{1}{2}}Dg). \tag{F.5}$$

Proof. Note that the dynamics of x corresponding to (F.2) is given by

$$\begin{aligned}\frac{dx}{dt} = \frac{\partial\overline{H}}{\partial v} &= g^{-\frac{1}{2}}(I-P)\,g^{-\frac{1}{2}}v + (D_v\lambda)^\top c\\ &= g^{-\frac{1}{2}}(I-P)\,g^{-\frac{1}{2}}v\end{aligned} \tag{F.6}$$

where we used that $c(x) = 0$ (Lemma 15).

Now let us compute the dynamics of v. Note that

$$v^\top g^{-\frac{1}{2}}(I-P)\,g^{-\frac{1}{2}}v = v^\top g^{-1}v - v^\top g^{-1}A^\top(A\cdot g^{-1}\cdot A^\top)^{-1}Ag^{-1}v.$$

Hence, we have

$$\begin{aligned}
&D_x\left(\frac{1}{2}v^\top g^{-\frac{1}{2}}\left(I-P\right)g^{-\frac{1}{2}}v\right)\\
=&-\frac{1}{2}v^\top g^{-1}\cdot Dg\cdot g^{-1}v+v^\top g^{-1}\cdot Dg\cdot g^{-1}A^\top(A\cdot g^{-1}\cdot A^\top)^{-1}Ag^{-1}v\\
&-\frac{1}{2}v^\top g^{-1}A^\top(A\cdot g^{-1}\cdot A^\top)^{-1}A\cdot g^{-1}\cdot Dg\cdot g^{-1}\cdot A^\top(A\cdot g^{-1}\cdot A^\top)^{-1}Ag^{-1}v\\
=&-\frac{1}{2}v^\top g^{-\frac{1}{2}}(I-P)g^{-\frac{1}{2}}\cdot Dg\cdot g^{-\frac{1}{2}}(I-P)g^{-\frac{1}{2}}v\\
=&-\frac{1}{2}Dg\left[\frac{dx}{dt},\frac{dx}{dt}\right],
\end{aligned}$$

where we used $\frac{dx}{dt}=g^{-\frac{1}{2}}(I-P)g^{-\frac{1}{2}}v$ in (F.6). Therefore, it follows that

$$\begin{aligned}
\frac{dv}{dt}&=-\frac{\partial\overline{H}}{\partial x}-(D_v\lambda)^\top c-A^\top\lambda && \text{(F.7)}\\
&=-\nabla f(x)+\frac{1}{2}Dg\left[\frac{dx}{dt},\frac{dx}{dt}\right]-\frac{1}{2}\mathrm{Tr}(g^{-1}Dg) && \text{(F.8)}\\
&\quad+\frac{1}{2}\mathrm{Tr}\left((Ag(x)^{-1}A^\top)^{-1}Ag(x)^{-1}\cdot Dg\cdot g(x)^{-1}A^\top\right)-A^\top\lambda\\
&=-\nabla f(x)+\frac{1}{2}Dg\left[\frac{dx}{dt},\frac{dx}{dt}\right]-\frac{1}{2}\mathrm{Tr}(g^{-\frac{1}{2}}\left(I-P\right)g^{-\frac{1}{2}}Dg)-A^\top\lambda
\end{aligned}$$

where we used that $c=0$ again in the second equality.

Recall that $\frac{dx}{dt}=g^{-\frac{1}{2}}\left(I-P\right)g^{-\frac{1}{2}}v$. In this formula, let us perturb v by $A^\top y$ for any y as follows.

$$\begin{aligned}
(I-P)\,g^{-\frac{1}{2}}(v+A^\top y)&=(I-P)\,g^{-\frac{1}{2}}v+\left(I-g^{-\frac{1}{2}}A^\top(Ag^{-1}A^\top)^{-1}Ag^{-\frac{1}{2}}\right)g^{-\frac{1}{2}}A^\top y\\
&=(I-P)\,g^{-\frac{1}{2}}v+(g^{-\frac{1}{2}}A^\top y-g^{-\frac{1}{2}}A^\top(Ag^{-1}A^\top)^{-1}(Ag^{-1}A^\top)y)\\
&=(I-P)\,g^{-\frac{1}{2}}v+(g^{-\frac{1}{2}}A^\top y-g^{-\frac{1}{2}}A^\top y)\\
&=(I-P)\,g^{-\frac{1}{2}}v.
\end{aligned}$$

Hence, removing $A^\top\lambda$ from $\frac{dv}{dt}$ in (F.7) does not change the dynamics of x, and thus we have the new dynamics given simply by (F.4) and (F.5). By repeating this proof, one can check that the simplified Hamiltonian (F.3) also yields (F.4) and (F.5). □

Efficient Computation of Leverage Score

In this section, we discuss how we efficiently compute the diagonal entries of $A^\top(Ag^{-1}A^\top)^{-1}A$. Our idea is based on the fact that certain entries of $(Ag^{-1}A^\top)^{-1}$ can be computed as fast as computing sparse Cholesky decomposition of $Ag^{-1}A^\top$ [Tak73, CD95], which can be $O(n)$ time faster than computing $(Ag^{-1}A^\top)^{-1}$ in many settings.

For simplicity, we focus on the case $g(x)$ as a diagonal matrix, since we use the log-barrier $\phi(x) = -\sum_{i=1}^{m}(\log(x_i - l_i) + \log(u_i - x_i))$ in implementation. We first note that we maintain a "sparsity pattern" $\mathrm{sp}(M)$ of a sparse matrix M so that we handle only these entries in downstream tasks. The sparsity pattern indicates "candidates" of nonzero entries of a matrix (i.e., $\mathrm{sp}(M) \supseteq nnz(M) = \{(i,j) : M_{ij} \neq 0\}$). For instance, it is obvious that $\mathrm{sp}(cc^\top) = \{(i,j) : c_i c_j \neq 0\} = nnz(cc^\top)$ for a column vector c and that $\mathrm{sp}(Ag^{-1}A^\top) = \bigcup_{i\in[n]} \mathrm{sp}(A_i A_i^\top)$ follows from the equality $Ag^{-1}A^\top = \sum_{i=1}^{n}(Ag^{-\frac{1}{2}})_i(Ag^{-\frac{1}{2}})_i^\top$, where M_i denote the i^{th} column of M (See Theorem 2.1 in [Dav06]). Then we compute the Cholesky decomposition to obtain a sparse triangular matrix L such that $LL^\top = Ag^{-1}A^\top$ with a property $\mathrm{sp}(Ag^{-1}A^\top) \subseteq \mathrm{sp}(L^\top) \cup \mathrm{sp}(L)$ (See Theorem 4.2 in [Dav06]).

Once the sparsity pattern of L is identified, we compute $S := (Ag^{-1}A^\top)^{-1}|_{\mathrm{sp}(L)}$, the restriction of S to $\mathrm{sp}(L)$, that is, the inverse matrix S is computed only for entries in $\mathrm{sp}(L)$. [Tak73, CD95] showed that this matrix S can be computed as fast as the Cholesky decomposition of $Ag^{-1}A^\top$.

For completeness, we explain how they compute S efficiently. Let $L_0 D L_0^\top$ be the LDL decomposition of $Ag^{-1}A^\top$ such that the diagonals of L_0 is one and so $L = L_0 D^{\frac{1}{2}}$, and it easily follows that

$$S = D^{-1}L_0^{-1} + (I - L_0^\top)S = D^{-\frac{1}{2}}L^{-1} + (I - L^\top D^{-\frac{1}{2}})^{-1}S.$$

Since $D^{-1}L_0^{-1}$ is lower triangular and $I - L_0^\top$ is strictly upper triangular, symmetry of S implies that S can be computed from the bottom row to the top row one by one. We note that the computation of S on any entry in $\mathrm{sp}(L)$ only requires previously computed S on entries in $\mathrm{sp}(L)$, due to the sparsity pattern of $I - L^\top D^{-\frac{1}{2}}$. [Tak73, CD95] showed that the total cost of computing S is $O(\sum_{i=1}^{n} n_i^2)$ for backward substitution, where n_i is the number of nonzeros in the i^{th} column of L. This exactly matches the cost of computing L. In our experiments, for many sparse matrices A, we found that $O(\sum_{i=1}^{n} n_i^2)$ is roughly $O(n^{1.5})$ and it is much faster than dense matrix inverse.

We have presented methods to save computational cost, avoiding full computation of the inverse $(Ag^{-1}A^\top)^{-1}$. This attempt is justified by the fact that only entries of $Ag^{-1}A^\top$ in $\mathrm{sp}(L) \cup \mathrm{sp}(L^\top)$ matter in computing $\mathrm{diag}(A^\top SA) = \mathrm{diag}(A^\top(Ag^{-1}A^\top)^{-1}A)$.

Lemma 19. *Computation of* $\mathrm{diag}(A^\top(Ag^{-1}A^\top)^{-1}A)$ *involves accessing only entries of* $(Ag^{-1}A^\top)^{-1}$ *in* $sp(Ag^{-1}A^\top)$.

Proof. Let $M := (Ag^{-1}A^\top)^{-1} \in \mathbb{R}^{m\times m}$, $\sigma_i := (A^\top(Ag^{-1}A^\top)^{-1}A)_{ii}$ for $i \in [n]$, and a_i be the i^{th} column of A. Observe that

$$\sigma_i = a_i^\top (Ag^{-1}A^\top)^{-1}a_i = \mathrm{Tr}(a_i^\top M a_i) = \mathrm{Tr}(M a_i a_i^\top).$$

As the entries of M only in $\mathrm{sp}(a_i a_i^\top)$ matter when computing the trace, we have that all the entries of M used for computing σ_i for all $i \in [n]$ are included in $\bigcup_{i=1}^n \mathrm{sp}(a_i a_i^\top) = \mathrm{sp}(Ag^{-1}A^\top)$. □

Now let us divide the diagonals of S by 2. Then we have $(Ag^{-1}A^\top)^{-1}|_{\mathrm{sp}(L)\cup\mathrm{sp}(L^\top)} = S + S^\top$ and thus

$$\begin{aligned}\mathrm{diag}(A^\top(Ag^{-1}A^\top)^{-1}A) &= \mathrm{diag}(A^\top(Ag^{-1}A^\top)^{-1}|_{\mathrm{sp}(L)\cup\mathrm{sp}(L^\top)}A)\\ &= \mathrm{diag}(A^\top SA + A^\top S^\top A) = 2\cdot\mathrm{diag}(A^\top SA)\end{aligned}$$

and the last term can be computed efficiently using S. In our experiment, the cost of computing leverage score is roughly twice the cost of computing Cholesky decomposition in all datasets.

Finally, we discuss another approach to compute leverage score with the same asymptotic complexity. We consider the function

$$V(g) = \log\det Ag^{-1}A^\top$$

where g is a sparse matrix $g \in \mathbb{R}^{\mathrm{sp}(g)}$ and V is defined only on $\mathbb{R}^{\mathrm{sp}(g)}$. Note that $V(g)$ can be computed using Cholesky decomposition of $A^\top g^{-1}A^\top$ and multiplying the diagonal of the decomposition. Next, we note that

$$\nabla V(g) = -(g^{-1}A^\top(Ag^{-1}A^\top)^{-1}Ag^{-1})|_{\mathrm{sp}(g)}.$$

Hence, we can compute leverage score by first computing $\nabla V(g)$ via automatic differentiation, and the time complexity of computing ∇V is only a small constant factor more than the time complexity of computing V [GW08]. The only problem with this approach is that the Cholesky decomposition algorithm is an algorithm involving a large loop and sparse operations and existing automatic differentiation packages are not efficient to differentiate such functions.

F.2.3 Stationarity of CRHMC

Now, the ideal CRHMC (or the continuous CRHMC) is the same as Algorithm 13 with the simplified constrained Hamiltonian H in place of the unconstrained Hamiltonian. In

this section, we prove that the Markov chain defined by the ideal CRHMC projected to x satisfies detailed balance with respect to its target distribution proportional to $e^{-f(x)}$ subject to $c(x) = 0$, leading to the target distribution being stationary.

To this end, we introduce a few notations here. Let $\mathcal{M} = \{x \in \mathbb{R}^n : c(x) = 0\}$ be a manifold in $\mathbb{R}^n$ and $\pi(x)$ be a desired distribution on $\mathcal{M}$ proportional to $e^{-f(x)}$ satisfying $\int_{\mathcal{M}} \pi(x)dx = 1$ (to be precise, the Radon-Nikodym derivative of π w.r.t. the Hausdorff measure on the manifold $\mathcal{M}$ is proportional to $e^{-f(x)}$). We denote the set of velocity v at $x \in \mathcal{M}$ (i.e., cotangent space) by $\mathcal{T}_x\mathcal{M} = \text{Null}(Dc(x)) = \{v \in \mathbb{R}^n : Dc(x)M(x)^\dagger v = 0\}$. Let T_h be the map sending (x, v) to $(x', v') = (x(h), y(h))$ in the Hamiltonian ODE (Step 2 of Algorithm 13) and define $F_{x,h}(v) := (\pi_1 \circ T_h)(x, v) = x'$, where $\pi_1(x, v) := x$ is the projection to the position space x. For a matrix A, we denote by $|A|$ the absolute value of its determinant $|\det(A)|$.

Note that we check the detailed balance of the induced chain on the "original (x)" space without moving to the "phase (x, v)" space, unlike Brubaker's proof [BSU12].

Theorem 20. *For $x, x' \in \mathcal{M}$, let $\mathbf{P}_x(x')$ be the probability density of the one-step distribution to x' starting at x in CRHMC (i.e., transition kernel from x to x'). It satisfies detailed balance with respect to the desired distribution π (i.e., $\pi(x)\mathbf{P}_x(x') = \pi(x')\mathbf{P}_{x'}(x)$).*

Proof. Fix x and x' in $\mathcal{M}$. Let C_1 be the normalization constant of $e^{-f(x)}$ (i.e., $\pi(x) = C_1 e^{-f(x)}$). The transition kernel $\mathbf{P}_x(x')$ is characterized as the pushforward by $F_{x,h}$ of the probability measure $v \sim \mathcal{N}(0, M(x))$ on $\mathcal{T}_x\mathcal{M}$, so it follows that

$$\mathbf{P}_x(x') = C_2 \int_{V_x} \frac{e^{-\frac{1}{2}\log \text{pdet}(M(x)) - \frac{1}{2}v^\top M(x)^\dagger v}}{|DF_{x,h}(v)|} dv,$$

where C_2 is the normalization constant of $e^{-\frac{1}{2}\log \text{pdet}(M(x)) - \frac{1}{2}v^\top M(x)^\dagger v}$ and $V_x = \{v \in \mathcal{T}_x\mathcal{M} : F_{x,h}(v) = x'\}$ is the set of velocity in cotangent space at x such that the Hamiltonian ODE with step size h sends (x, v) to (x', v'). (Further details for deducing the 1-step distribution can be found in Lemma 10 of [LV18]) As $c(x) = 0$ for $x \in \mathcal{M}$, it follows that

$$\begin{aligned}
&\pi(x)\mathbf{P}_x(x') \\
&= C_1 C_2 \int_{V_x} \frac{e^{-f(x) - \frac{1}{2}\log \text{pdet}(M(x)) - \frac{1}{2}v^\top M(x)^\dagger v - \lambda(x,v)^\top c(x)}}{|DF_{x,h}(v)|} dv = C_1 C_2 \int_{V_x} \frac{e^{-H(x,v)}}{|DF_{x,h}(v)|} dv.
\end{aligned}$$

Going forward, we use three important properties of the Hamiltonian dynamics including reversibility, Hamiltonian preservation, and volume preservation, which still hold for the constrained Hamiltonian H. Due to reversibility $T_{-h}(x', v') = (x, v)$, we can write

$$\pi(x')\mathbf{P}_{x'}(x) = C_1C_2\int_{V_{x'}} \frac{e^{-H(x',v')}}{|DF_{x',-h}(v')|}dv',$$

where $V_{x'} = \{v' \in \mathcal{T}_{x'}\mathcal{M} : F_{x',-h}(v') = x\}$ is the counterpart of V_x. From reversibility $T_{-h} \circ T_h = I$, the inverse function theorem implies $DT_{-h} = (DT_h)^{-1}$. Now let us denote

$$DT_h(x,v) = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \quad \& \quad DT_{-h}(x',v') = \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix},$$

where each entry is a block matrix with the same size. Note that $DF_{x,h}(v) = B$ and $DF_{x',-h}(v') = B'$ hold by the definition of Jacobian. Together with $DT_{-h} = (DT_h)^{-1}$, a formula for the inverse of a block matrix results in

$$|DF_{x',-h}(v')| = |B'| = \frac{|B|}{|D||A - BD^{-1}C|} = \frac{|B|}{|DT_h(x,v)|} = |B| = |DF_{x,h}(v)|,$$

where we use the property of volume preservation in the fourth equality (i.e., $|DT_h(x,v)| = 1$). Finally, the property of Hamiltonian preservation implies $H(x,v) = H(x',v')$ and thus

$$\int_{V_{x'}} \frac{e^{-H(x',v')}}{|DF_{x',-h}(v')|}dv' = \int_{V_x} \frac{e^{-H(x,v)}}{|DF_{x,h}(v)|}dv.$$

Therefore, $\pi(x)\mathbf{P}_x(x') = \pi(x')\mathbf{P}_{x'}(x)$ holds. □

Similar reasoning as Theorem 3 and Lemma 1 in [BSU12] gives π-irreducibility and aperiodicity of the process, so CRHMC converges to the unique stationary distribution $\pi \propto e^{-f(x)}$.

F.3 Discretization

We discuss how to implement our Hamiltonian dynamics using the implicit midpoint method in Section F.3.1 and present theoretical guarantees of correctness and efficiency of the discretized CRHMC in Section F.3.2.

F.3.1 Discretized CRHMC based on Implicit Midpoint Integrator

In our algorithm, we discretize the Hamiltonian process into steps of step size h and run the process for T iterations (see Algorithm 15). Rather than resampling the velocity at every step, we may change the velocity more gradually, using the following update:

$$v' \leftarrow \sqrt{\beta}v + \sqrt{1-\beta}z$$

where $z \sim \mathcal{N}(0, M(x))$ and β is a parameter. We note that this step is time-reversible, i.e., $\mathbf{P}(v|x)\mathbf{P}(v \to v') = \mathbf{P}(v'|x)\mathbf{P}(v' \to v)$ (see Theorem 22). Starting from $(x^{(0)}, v^{(0)})$, let $(x^{(t)}, v^{(t)})$ be the point obtained after iteration t. In the beginning of each iteration, we compute the Cholesky decomposition of $Ag(x)^{-1}A^\top$ for later use and resample the velocity with momentum. As noted previously in Lemma 18, for $c(x) = Ax - b$ we can just use the simplified Hamiltonian in (F.3),

$$H(x,v) = f(x) + \frac{1}{2}v^\top g(x)^{-\frac{1}{2}}\left(I - P(x)\right)g(x)^{-\frac{1}{2}}v + \frac{1}{2}\left(\log\det g(x) + \log\det Ag(x)^{-1}A^\top\right)$$

instead of the constrained Hamiltonian $H + \lambda^\top c$. We solve the Hamiltonian dynamics for H by the implicit midpoint method, which we will discuss below, and then use a Metropolis filter on H to ensure the distribution is correct.

Implicit Midpoint Method. For general Riemannian manifolds, explicit integrators such as the leapfrog method (LM) are not symplectic, unlike IMM. LM is symplectic when the Hamiltonian equations are separable (i.e., each of dx/dt and dv/dt is a function of either x or v only). However, in the general Riemannian manifold setting, where dx/dt depends on position x due to mass matrices (which is $g(x)$ in our paper) as well as velocity v, the Hamiltonian is no longer separable, which prevents us from using LM. We refer interested readers to Section 3 and Section 4.1 in [CBMR19].

We now elaborate on how the implicit midpoint integrator works (see Algorithm 16), which is symplectic (so measure-preserving) and reversible [HHIL06]. Let us write $H(x,v) = \overline{H}_1(x,v) + \overline{H}_2(x,v)$, where

$$\begin{aligned}
\overline{H}_1(x,v) &= f(x) + \frac{1}{2}\left(\log\det g(x) + \log\det Ag(x)^{-1}A^\top\right), \\
\overline{H}_2(x,v) &= \frac{1}{2}v^\top g(x)^{-\frac{1}{2}}\left(I - P(x)\right)g(x)^{-\frac{1}{2}}v.
\end{aligned}$$

Starting from (x_0, v_0), in the first step of the integrator, we run the process on the Hamiltonian $\overline{H}_1$ with step size $\frac{h}{2}$ to get $(x_{1/3}, v_{1/3})$, and this discretization leads to $x_{1/3} = x_0 + \frac{h}{2}\frac{\partial \overline{H}_1}{\partial v}(x_0, v_0)$ and $v_{1/3} = v_0 - \frac{h}{2}\frac{\partial \overline{H}_1}{\partial x}(x_0, v_0)$. Note that $x_{1/3} = x_0$ due to $\frac{\partial \overline{H}_1}{\partial v} = 0$. In the second step of the integrator, we run the process on $\overline{H}_2$ with step size h by solving

$$\begin{aligned}
x_{\frac{2}{3}} &= x_{\frac{1}{3}} + h\frac{\partial \overline{H}_2}{\partial v}\left(\frac{x_{\frac{1}{3}} + x_{\frac{2}{3}}}{2}, \frac{v_{\frac{1}{3}} + v_{\frac{2}{3}}}{2}\right), \\
v_{\frac{2}{3}} &= v_{\frac{1}{3}} - h\frac{\partial \overline{H}_2}{\partial x}\left(\frac{x_{\frac{1}{3}} + x_{\frac{2}{3}}}{2}, \frac{v_{\frac{1}{3}} + v_{\frac{2}{3}}}{2}\right).
\end{aligned}$$

To this end, starting from $x_{2/3} = x_{1/3}$ and $v_{2/3} = v_{1/3}$, we apply $x_{2/3} \leftarrow x_{1/3} + h\frac{\partial \overline{H}_2}{\partial v}\left(\frac{x_{1/3}+x_{2/3}}{2}, \frac{v_{1/3}+v_{2/3}}{2}\right)$ and $v_{2/3} \leftarrow v_{1/3} - h\frac{\partial \overline{H}_2}{\partial x}\left(\frac{x_{1/3}+x_{2/3}}{2}, \frac{v_{1/3}+v_{2/3}}{2}\right)$ iteratively with the following subroutine for computing $\frac{\partial \overline{H}_2}{\partial v}$ and $\frac{\partial \overline{H}_2}{\partial x}$. According to Lemma 18, this computation involves solving $g(x)^{-1}A^\top \left(Ag(x)^{-1}A^\top\right)^{-1} Ag(x)^{-1}v$ for some v and x. To compute $\left(Ag(x)^{-1}A^\top\right)^{-1} Ag(x)^{-1}v$, we use the Newton's method, which iteratively computes $\nu \leftarrow \nu + M^{-1}Ag(x)^{-1}\left(v - A^\top\nu\right)$ for some M. Note that the Newton's method guarantees that ν converges to $M^{-1}Ag(x)^{-1}v$ if M is invertible. Here, we choose $M = Ag(x^{(t)})^{-1}A^\top$ to ensure fast convergence. Since we have already computed the Cholesky decomposition of M in the beginning, $M^{-1}Ag(x)^{-1}\left(v - A^\top\nu\right)$ can be computed efficiently by backward and forward substitution. In the third step of the integrator, we run the process on the Hamiltonian $\overline{H}_1$ with step size $\frac{h}{2}$ again to get (x_1, v_1), which results in $x_1 = x_{2/3}$ and $v_1 = v_{2/3} - \frac{h}{2}\frac{\partial \overline{H}_1}{\partial x}(x_1, v_{2/3})$.

We note that CRHMC is affine-invariant and provably independent of condition number (Theorem 29), and thus the step size and momentum only need to depend on the dimension. In practice, we set the momentum to roughly $1 - h$, and for the step size h, we decrease it until the acceptance probability is close enough to 1 during the warm-up phase. Empirically, we found that the step size stays between 0.05 and 0.2 in practice even for high dimensional ill-conditioned polytopes. This step size is remarkable, given that for these instances a standard package like STAN ends up selecting a small step size like 10^{-8} and thus fails to converge.

Putting Algorithm 15 and Algorithm 16 together, we obtain discretization of constrained Riemannian Hamiltonian Monte Carlo algorithm.

F.3.2 Theoretical Guarantees

In terms of efficiency, we first show that one iteration of Algorithm 15 incurs the cost of solving a few Cholesky decomposition and $O(K)$ sparse triangular systems. We also show in Lemma 23 that the implicit midpoint integrator converges to the solution of Eq. (F.9) in logarithmically many iterations. Regarding correctness, Theorem 22 and Lemma 23 together show that the discretized CRHMC (Algorithm 15) converges to the stationary distribution indeed (see Remark 24).

Theorem 21. *The cost of each iteration of Algorithm 15 is solving $O(1)$ Cholesky decomposition and $O(K)$ triangular systems, where K is the number of iterations in Algorithm 16.*

Algorithm 15 `Discretized Constrained Riemannian Hamiltonian Monte Carlo with Momentum`

Input: Initial point $x^{(0)}$, velocity $v^{(0)}$, record frequency T, step size h, ODE steps K

for $t = 1, 2, \cdots, T$ **do**

Let $\overline{v} = v^{(t-1)}$ and $x = x^{(t-1)}$.

Step 1: Resample v with momentum

Let $z \sim \mathcal{N}(0, M(x))$. Update $\overline{v}$:

$$v \leftarrow \sqrt{\beta}\overline{v} + \sqrt{1-\beta}z.$$

Step 2: Solve $\frac{dx}{dt} = \frac{\partial H(x,v)}{\partial v}, \frac{dv}{dt} = -\frac{\partial H(x,v)}{\partial x}$ via the implicit midpoint method

Use `Implicit Midpoint Method`(x, v, h, K) to find (x', v') such that

$$\begin{aligned} v_{\frac{1}{3}} &= v - \frac{h}{2}\frac{\partial \overline{H}_1(x,v)}{\partial x}, \\ x' &= x + h\frac{\partial \overline{H}_2(\frac{x+x'}{2}, \frac{v_{1/3}+v_{2/3}}{2})}{\partial v},\ v_{\frac{2}{3}} = v_{\frac{1}{3}} - h\frac{\partial \overline{H}_2(\frac{x+x'}{2}, \frac{v_{1/3}+v_{2/3}}{2})}{\partial x}, \\ v' &= v_{\frac{2}{3}} - \frac{h}{2}\frac{\partial \overline{H}_1(x', v_{\frac{2}{3}})}{\partial x}. \end{aligned} \tag{F.9}$$

Step 3: Filter

With probability $\min\left\{1, \frac{e^{-H(x',v')}}{e^{-H(x,v)}}\right\}$, set $x^{(t)} \leftarrow x'$ and $v^{(t)} \leftarrow v'$.

Otherwise, set $x^{(t)} \leftarrow x$ and $v^{(t)} \leftarrow -v$.

end for

Output: $x^{(T)}$

Proof. We first solve the Cholesky decomposition to get $L_{t-1}L_{t-1}^\top = Ag(x^{(t-1)})^{-1}A^\top$ at the beginning of iteration. Recall that

$$\begin{aligned} H(x,v) &= \overline{H}_1(x,v) + \overline{H}_2(x,v) \\ &= \left(f(x) + \frac{1}{2}(\log\det g(x) + \log\det Ag(x)^{-1}A^\top)\right) \\ &\quad + \left(\frac{1}{2}v^\top g(x)^{-\frac{1}{2}}\left(I - g(x)^{-\frac{1}{2}}A^\top(Ag(x)^{-1}A^\top)^{-1}Ag(x)^{-\frac{1}{2}}\right)g(x)^{-\frac{1}{2}}v\right). \end{aligned}$$

The value of $H(x^{(t-1)}, v^{(t-1)})$ should be computed later for the filter step and can be efficiently computed by the given $L_{t-1}L_{t-1}^\top = Ag(x^{(t-1)})^{-1}A^\top$and solving two sparse tri-

Algorithm 16 `Implicit Midpoint Method`

Input: Initial point $x^{(0)}$, velocity $v^{(0)}$, record frequency T, step size h, ODE steps K

Set $x_{\frac{1}{3}} \leftarrow x$ and $v_{\frac{1}{3}} \leftarrow v - \frac{h}{2}\frac{\partial \overline{H}_1(x,v)}{\partial x}$.

Set $\nu \leftarrow 0$.

for $k = 1, 2, \cdots, K$ **do**

Let $x_{\text{mid}} \leftarrow \frac{1}{2}\left(x_{\frac{1}{3}} + x_{\frac{2}{3}}\right)$ and $v_{\text{mid}} \leftarrow \frac{1}{2}\left(v_{\frac{1}{3}} + v_{\frac{2}{3}}\right)$

Set $\nu \leftarrow \nu + \left(LL^\top\right)^{-1} Ag(x_{\text{mid}})^{-1}\left(v_{\text{mid}} - A^\top\nu\right)$

Set $x_{\frac{2}{3}} \leftarrow x_{\frac{1}{3}} + hg(x_{\text{mid}})^{-1}\left(v_{\text{mid}} - A^\top\nu\right)$ and $v_{\frac{2}{3}} \leftarrow v_{\frac{1}{3}} + \frac{h}{2}Dg(x_{\text{mid}})\left[g(x_{\text{mid}})^{-1}\left(v_{\text{mid}} - A^\top\nu\right), g(x_{\text{mid}})^{-1}\left(v_{\text{mid}} - A^\top\nu\right)\right]$

Set $x_1 \leftarrow x_{\frac{2}{3}}$ and $v_1 \leftarrow v_{\frac{2}{3}} - \frac{h}{2}\frac{\partial \overline{H}_1}{\partial x}(x_{\frac{2}{3}}, v_{\frac{2}{3}})$.

end for

Output: x_1, v_1

angular systems (i.e., $L_{t-1}^{-\top}(L_{t-1}^{-1}(Ag(x)^{-\frac{1}{2}}))$). We need the same cost (i.e., Cholesky decomposition and solving two triangular systems) for the value of $H(x', v')$, where (x', v') is the output of Algorithm 16. We note that L inherits sparsity of A and thus each triangular system can be solved efficiently by backward and forward substitution.

In the implicit midpoint method, one main component is computation of $\frac{\partial \overline{H}_1(x,v)}{\partial x}$ in Step 1 and $\frac{\partial \overline{H}_1}{\partial x}(x_{\frac{2}{3}}, v_{\frac{2}{3}})$ in Step 3 due to leverage scores. As seen in Section F.2.2, the cost for these computations is within a constant factor of solving the Cholesky decomposition for $Ag(x^{(t-1)})^{-1}A^\top$ and $Ag(x_{\frac{2}{3}})^{-1}A^\top$. Another component is solving $O(K)$ triangular systems to update ν in Step 2.

Adding up all these costs, each iteration of Algorithm 15 only requires solving $O(1)$ Cholesky decomposition and $O(K)$ sparse triangular systems. □

Theorem 22. *The Markov chain defined by Algorithm 15 projected to x has a stationary density proportional to $\exp(-f(x))$, and is irreducible and aperiodic. Therefore, this Markov chain converges to the stationary distribution.*

Proof. Each iteration consists of two stages: resampling velocity with momentum in Step 1 (i.e., $(x, \overline{v})$ to (x, v)) and solving ODE followed by the filter in Step 2 and 3 (i.e., (x, v) to (x', v')). To prove the claim, we show that Step 1 is time-reversible with respect to the conditional distribution $\pi(v|x)$ and that Step 2 followed by Step 3 is also time-reversible with respect to $\pi(x, v)$.

We begin with the first part. We have $\pi(\overline{v}|x) = \mathcal{N}(0, M(x))$ due to the definition of H. Since $\overline{v}|x \sim \mathcal{N}(0, M(x))$ and $z \sim \mathcal{N}(0, M(x))$ are independent Gaussians, the update rule $v = \sqrt{\beta}\overline{v} + \sqrt{1-\beta}z$ implies $\pi(v|x) = \mathcal{N}(0, M(x))$. Let $\mathbf{P}(z)$ be the probability density and C be the normalization constant for Gaussian $\mathcal{N}(0, M(x))$. Then, the time-reversibility w.r.t. $\pi(v|x)$ is immediate from the following computation:

$$\begin{aligned}
\pi(\overline{v}|x)\mathbf{P}(\overline{v} \to v) &= C^2 \exp(-\frac{1}{2}\overline{v}^\top M^\dagger \overline{v}) \cdot \exp(-\frac{1}{2}\frac{(v-\sqrt{\beta}\overline{v})^\top M^\dagger (v-\sqrt{\beta}\overline{v})}{1-\beta}) \\
&= C^2 \exp\left(-\frac{1}{2}\left(\overline{v}^\top M^\dagger \overline{v} + \frac{v^\top M^\dagger v}{1-\beta} + \frac{\beta \overline{v}^\top M^\dagger \overline{v}}{1-\beta} - \frac{\sqrt{\beta}}{1-\beta}(\overline{v}^\top M^\dagger v + v^\top M^\dagger \overline{v})\right)\right) \\
&= C^2 \exp\left(-\frac{1}{2}\left(\frac{v^\top M^\dagger v}{1-\beta} + \frac{\overline{v}^\top M^\dagger \overline{v}}{1-\beta} - \frac{\sqrt{\beta}}{1-\beta}(\overline{v}^\top M^\dagger v + v^\top M^\dagger \overline{v})\right)\right), \\
\pi(v|x)\mathbf{P}(v \to \overline{v}) &= C^2 \exp(-\frac{1}{2}v^\top M^\dagger v) \cdot \exp(-\frac{1}{2}\frac{(\overline{v}-\sqrt{\beta}v)^\top M^\dagger (\overline{v}-\sqrt{\beta}v)}{1-\beta}) \\
&= C^2 \exp\left(-\frac{1}{2}\left(v^\top M^\dagger v + \frac{\overline{v}^\top M^\dagger \overline{v}}{1-\beta} + \frac{\beta v^\top M^\dagger v}{1-\beta} - \frac{\sqrt{\beta}}{1-\beta}(\overline{v}^\top M^\dagger v + v^\top M^\dagger \overline{v})\right)\right) \\
&= C^2 \exp\left(-\frac{1}{2}\left(\frac{v^\top M^\dagger v}{1-\beta} + \frac{\overline{v}^\top M^\dagger \overline{v}}{1-\beta} - \frac{\sqrt{\beta}}{1-\beta}(\overline{v}^\top M^\dagger v + v^\top M^\dagger \overline{v})\right)\right) \\
\implies \quad & \pi(\overline{v}|x)\mathbf{P}(\overline{v} \to v) = \pi(v|x)\mathbf{P}(v \to \overline{v}).
\end{aligned}$$

The second part follows from a stronger statement due to symmetry of v in $H(x, v)$: In the space where (x, v) and $(x, -v)$ are identified, the Markov chain defined by Step 2 and 3 satisfies detailed balance with respect the density $\pi([x, v])$ proportional to $\exp(-H(x, v))$, where $[x, v]$ denotes the identified point for (x, v) and $(x, -v)$. Consider the pairs $[x, v] = \{(x, v), (x, -v)\}$ and $[x', v'] = \{(x', v'), (x', -v')\}$ where in Step 2 (x, v) goes to (x', v') and $(x', -v')$ goes to $(x, -v)$ due to reversibility of the implicit midpoint method. We now verify that the filtering probability is the same in either direction, using the measure-preserving property of Step 2

$$\begin{aligned}
\pi(x, v)\mathbf{P}\left((x, v) \to (x', v')\right) &= \pi(x, v) \min\left\{1, \frac{\pi(x', v')}{\pi(x, v)}\right\} \\
&= \min\left\{\pi(x, v), \pi(x', v')\right\} \\
&= \min\left\{\pi(x, -v), \pi(x', -v')\right\} \\
&= \pi(x', -v') \min\left\{1, \frac{\pi(x, -v)}{\pi(x', -v')}\right\} \\
&= \pi(x', -v')\mathbf{P}\left((x', -v') \to (x, -v)\right).
\end{aligned}$$

Therefore, for any two pairs $[x, v]$ and $[x', v']$, we have $\pi([x, v])\mathbf{P}\left([x, v] \to [x', v']\right) = \pi([x', v']\mathbf{P}\left([x', v'] \to [x, v]\right)$, and thus this detailed balance implies that the target density is stationary.

Its irreducibility is implied by the non-zero lower bound on the conductance of the discretized CRHMC (Theorem 29). To see this, let A and B be two subsets of positive measure such that one subset is not reachable from another in infinitely many steps. Take the set R of reachable points from A via running the Markov chain, and note that R and $R^c(\supseteq B)$ have non-zero measures. However, the non-zero conductance, meaning that there must be a positive probability of stepping out of R, which contradicts the definition of R. Now for aperiodicity, as assumed at the beginning of the mixing rate proof (Appendix F.4), we consider a lazy version of the discretized CRHMC instead, which makes the chain stay where it is at with probability $1/2$ at each iteration, which prevents potential periodicity of the process. Note that this modification worsens the mixing rate only by a factor of 2.

Putting these three together, we can show that the discretized CRHMC converges to the target distribution. □

Now we show in Lemma 23 that the implicit midpoint method (Algorithm 16) converges to the solution of (F.9) in logarithmically many iterations. To show the convergence of Algorithm 16, we denote by $\mathcal{T}$ the map induced by one iteration of Step 2.

Definition 10. *Let*

$$\mathcal{T}(x,v,\nu)=\begin{pmatrix} x_{\frac{1}{3}}+hg(x_{mid})^{-1}(v_{mid}-A^\top\lambda_1) \\ v_{\frac{1}{3}}+\frac{h}{2}Dg(x_{mid})[g(x_{mid})^{-1}(v_{mid}-A^\top\lambda_1),g(x_{mid})^{-1}(v_{mid}-A^\top\lambda_1)] \\ \lambda_1 \end{pmatrix},$$

where $x_{mid}=\frac{1}{2}(x_{\frac{1}{3}}+x)$, $v_{mid}=\frac{1}{2}(v_{\frac{1}{3}}+v)$, *and* $\lambda_1=\nu+(LL^\top)^{-1}Ag(x_{mid})^{-1}\left(v_{mid}-A^\top\nu\right)$. *Let* $(x^*_{\frac{2}{3}},v^*_{\frac{2}{3}},\nu^*)$ *be the fixed point of* $\mathcal{T}$.

We assume that g is given by the Hessian of a highly self-concordant barrier ϕ (see F.5.2). Note that the log-barrier is highly self-concordant. We can show that for small enough step size h, Algorithm 16 can solve (F.9) to δ-accuracy in logarithmically many iterations.

Lemma 23. *Suppose* $g(x)=\nabla^2\phi(x)$ *for some highly self-concordant barrier* ϕ. *For any input* $(x_{\frac{1}{3}},v_{\frac{1}{3}})$, *let* $(x^{(k)}_{\frac{2}{3}},v^{(k)}_{\frac{2}{3}},\nu^{(k)})$ *be points obtained after* k *iterations in Step 2 of Algorithm 16. Let* $(\widetilde{x}_{\frac{2}{3}},\widetilde{v}_{\frac{2}{3}})$ *be the solution for* $(x_{\frac{2}{3}},v_{\frac{2}{3}})$ *in the following equation*

$$x_{\frac{2}{3}}=x_{\frac{1}{3}}+h\frac{\partial\overline{H}_2}{\partial v}\left(\frac{x_{\frac{1}{3}}+x_{\frac{2}{3}}}{2},\frac{v_{\frac{1}{3}}+v_{\frac{2}{3}}}{2}\right),\ v_{\frac{2}{3}}=v_{\frac{1}{3}}-h\frac{\partial\overline{H}_2}{\partial x}\left(\frac{x_{\frac{1}{3}}+x_{\frac{2}{3}}}{2},\frac{v_{\frac{1}{3}}+v_{\frac{2}{3}}}{2}\right).$$

Let $\|x\|_A := \sqrt{x^\top Ax}$ *for a matrix* A*. For any* (x, v, ν)*, define the norm*

$$\|(x, v, \lambda)\| := \|x\|_{g(x_{\frac{1}{3}})} + \|v\|_{g(x_{\frac{1}{3}})^{-1}} + h\|A^\top \nu\|_{g(x_{\frac{1}{3}})^{-1}}.$$

If $\left\|(x^{(0)}_{\frac{2}{3}}, v^{(0)}_{\frac{2}{3}}, \nu^{(0)}) - (\widetilde{x}_{\frac{2}{3}}, \widetilde{v}_{\frac{2}{3}}, \nu^*)\right\| \le r$ *with* $h \le r \le \min(\frac{1}{10}, \frac{\sqrt{h}}{4}, \frac{\|v^*\|_{g(x_0)^{-1}}}{4})$*, then*

$$\left\|(x^{(L)}_{\frac{2}{3}}, v^{(L)}_{\frac{2}{3}}, \nu^{(L)}) - (\widetilde{x}_{\frac{2}{3}}, \widetilde{v}_{\frac{2}{3}}, \nu^*)\right\| \le \delta$$

for some $L = O\left(\log_{1/C} \frac{r}{\delta}\right)$*, where* $C = O_n(h)$ *is the Lipschitz constant of the map* $\mathcal{T}$*.*

Proof. Since $(x^*_{\frac{2}{3}}, v^*_{\frac{2}{3}}, \nu^*)$ is the fixed point of $\mathcal{T}$ (i.e., $\nu^* = \lambda_1$), we have

$$\nu^* = \nu^* + (LL^\top)^{-1} Ag(x_{\text{mid}})^{-1} \left(v_{\text{mid}} - A^\top \nu^*\right)$$

and thus $Ag(x_{\text{mid}})^{-1} v_{\text{mid}} = Ag(x_{\text{mid}})^{-1} A^\top \nu^*$. For invertible $Ag(x_{\text{mid}})^{-1} A^\top$, we have

$$\nu^* = \left(Ag(x_{\text{mid}})^{-1} A^\top\right)^{-1} Ag(x_{\text{mid}})^{-1} v_{\text{mid}}.$$

Similarly by using the definition of the fixed point and this new formula for ν^*,

$$\begin{aligned} x^*_{\frac{2}{3}} &= x_{\frac{1}{3}} + hg(x_{\text{mid}})^{-1} v_{\text{mid}} - hg(x_{\text{mid}})^{-1} A^\top \nu^* \\ &= x_{\frac{1}{3}} + hg(x_{\text{mid}})^{-1} v_{\text{mid}} - hg(x_{\text{mid}})^{-1} A^\top \left(Ag(x_{\text{mid}})^{-1} A^\top\right)^{-1} Ag(x_{\text{mid}})^{-1} v_{\text{mid}} \\ &= x_{\frac{1}{3}} + h \frac{\partial \overline{H}_2}{\partial v}(x_{\text{mid}}, v_{\text{mid}}) \end{aligned}$$

and

$$\begin{aligned} v^*_{\frac{2}{3}} &= v_{\frac{1}{3}} + \frac{h}{2} Dg(x_{\text{mid}})[g(x_{\text{mid}})^{-1}(v_{\text{mid}} - A^\top \nu^*), g(x_{\text{mid}})^{-1}(v_{\text{mid}} - A^\top \nu^*)] \\ &= v_{\frac{1}{3}} - h \frac{\partial \overline{H}_2}{\partial x}(x_{\text{mid}}, v_{\text{mid}}) \end{aligned}$$

which shows that $(x^*_{\frac{2}{3}}, v^*_{\frac{2}{3}})$ is exactly the solution for (x, v) in the equation

$$x = x_{\frac{1}{3}} + h \frac{\partial \overline{H}_2}{\partial v}\left(\frac{x_{\frac{1}{3}} + x}{2}, \frac{v_{\frac{1}{3}} + v}{2}\right), \; v = v_{\frac{1}{3}} - h \frac{\partial \overline{H}_2}{\partial x}\left(\frac{x_{\frac{1}{3}} + x}{2}, \frac{v_{\frac{1}{3}} + v}{2}\right).$$

Next, we show that the iterations in Step 2 converges to $(x^*_{\frac{2}{3}}, v^*_{\frac{2}{3}}, \nu^*)$. If $\left\|(x^{(0)}_{\frac{2}{3}}, v^{(0)}_{\frac{2}{3}}, \nu^{(0)}) - (x^*_{\frac{2}{3}}, v^*_{\frac{2}{3}}, \nu^*)\right\| \le r$ for some $C = O_n(h)$, we have

$$\begin{aligned} \left\|(x^{(\ell)}_{\frac{2}{3}}, v^{(\ell)}_{\frac{2}{3}}, \nu^{(\ell)}) - (x^*_{\frac{2}{3}}, v^*_{\frac{2}{3}}, \nu^*)\right\| &= \left\|\mathcal{T}(x^{(\ell-1)}_{\frac{2}{3}}, v^{(\ell-1)}_{\frac{2}{3}}, \nu^{(\ell)}) - \mathcal{T}(x^*_{\frac{2}{3}}, v^*_{\frac{2}{3}}, \nu^*)\right\| \\ &\le C \left\|(x^{(\ell-1)}_{\frac{2}{3}}, v^{(\ell-1)}_{\frac{2}{3}}, \nu^{(\ell-1)}) - (x^*_{\frac{2}{3}}, v^*_{\frac{2}{3}}, \nu^*)\right\| \end{aligned}$$

$$\leq C^{\ell} \left\| (x^{(0)}_{\frac{2}{3}}, v^{(0)}_{\frac{2}{3}}, \nu^{(0)}) - (x^{*}_{\frac{2}{3}}, v^{*}_{\frac{2}{3}}, \nu^{*}) \right\|,$$

where the first equality follows from $(x^*_{\frac{2}{3}}, v^*_{\frac{2}{3}}, \nu^*)$ is the fixed point of $\mathcal{T}$ and the second inequality follows from Lemma 26. Therefore, we have $\left\| (x^{(L)}_{\frac{2}{3}}, v^{(L)}_{\frac{2}{3}}, \nu^{(L)}) - (x^{*}_{\frac{2}{3}}, v^{*}_{\frac{2}{3}}, \nu^{*}) \right\| \leq \delta$ for $L = O\left(\log_C \frac{r}{\delta}\right)$. □

Remark 24. *Lemma 23 shows that Algorithm 16 converges to the solution of (F.9) in logarithmically many iterations for small enough step size h. In Step 1 of Algorithm 15, v is resampled so that every iteration of Algorithm 15 is a non-degenerate map. Then, the total variation distance between the distributions generated by solving (F.9) using Algorithm 16 and solving (F.9) exactly in one iteration of Algorithm 15 can be bounded by error due to Algorithm 16. Theorem 22 shows that the process will converge to the exact stationary distribution. Therefore, in order for the accumulated error of Algorithm 15 to remain bounded for polynomially many steps, it suffices to run logarithmically many iterations in Algorithm 16. Any small bias due to the numerical error in the ODE computation is corrected by the filter, and maintaining as small error as possible is important to keep the acceptance probability high.*

F.3.3 Deferred Proof

Lemma 25 ([KLSV22c], Lemma 28)**.** *Suppose $g(x) = \nabla^2\phi(x)$ for some highly self-concordance barrier ϕ. Then, we have that*

- $(1 - \|y - x\|_{g(x)})^2 g(x) \preceq g(y) \preceq \frac{1}{(1-\|y-x\|_{g(x)})^2} g(x)$.
- $\|Dg(x)[v, v]\|_{g(x)^{-1}} \leq 2\|v\|^2_{g(x)}$.
- $\|Dg(x)[v, v] - Dg(y)[v, v]\|_{g(x)^{-1}} \leq \frac{6}{(1-\|y-x\|_{g(x)})^3} \|v\|^2_{g(x)} \|y - x\|_{g(x)}$.
- $\|Dg(x)[v, v] - Dg(x)[w, w]\|_{g(x)^{-1}} \leq 2 \|v - w\|_{g(x)} \|v + w\|_{g(x)}$.

Lemma 26. *Let $g(x) = \nabla^2\phi(x)$ for some highly self-concordance barrier ϕ. Given x_0, v_0 and L such that $LL^\top = Ag(x_0)^{-1}A^\top$, consider the map*

$$\mathcal{T}(x, v, \lambda) = \begin{pmatrix} x_0 + hg(x_{1/2})^{-1}(v_{1/2} - A^\top\lambda_1) \\ v_0 + \frac{h}{2} Dg(x_{1/2})[g(x_{1/2})^{-1}(v_{1/2} - A^\top\lambda_1), g(x_{1/2})^{-1}(v_{1/2} - A^\top\lambda_1)] \\ \lambda_1 \end{pmatrix}$$

where $x_{1/2} = (x_0+x)/2$, $v_{1/2} = (v_0+v)/2$ *and* $\lambda_1 = \lambda+(LL^\top)^{-1}Ag(x_{1/2})^{-1}\left(v_{1/2} - A^\top\lambda\right)$. *Let* (x^*, v^*, λ^*) *be a fixed point of* $\mathcal{T}$. *For any* x, v, λ, *we define the norm*

$$\|(x, v, \lambda)\| = \|x\|_{g(x_0)} + \|v\|_{g(x_0)^{-1}} + h\|A^\top\lambda\|_{g(x_0)^{-1}}.$$

Let $\Omega = \{(x, v, \lambda) : \|(x, v, \lambda) - (x^*, v^*, \lambda^*)\| \le r\}$ *with* $h \le r \le \min(\frac{1}{10}, \frac{\sqrt{h}}{4}, \frac{\|v^*\|_{g(x_0)^{-1}}}{4})$. *Suppose that* $(x_0, v_0, 0) \in \Omega$. *Then, for any* $(x, v, \lambda), (\overline{x}, \overline{v}, \overline{\lambda}) \in \Omega$, *we have*

$$\|\mathcal{T}(x, v, \lambda) - \mathcal{T}(\overline{x}, \overline{v}, \overline{\lambda})\| \le C\|(x, v, \lambda) - (\overline{x}, \overline{v}, \overline{\lambda})\|$$

where $C = (\frac{3r}{h} + \|v^*\|_{g(x_0)^{-1}})(400r + 18h\|v^*\|_{g(x_0)^{-1}})$.

Remark 27. *Note that we should think* $r = \Theta_n(h)$ *because that is the distance between* $(x_0, v_0, 0)$ *and* (x^*, v^*, λ^*). *In that case, the Lipschitz constant of* $\mathcal{T}$ *is* $O_n(h\|v^*\|^2_{g(x_0)^{-1}}) = O_n(h)$. *Hence, if the step size* h *is small enough, then* $\mathcal{T}$ *is a contractive mapping. In practice, we can take* h *close to a constant because* g *is decomposable into barriers in each dimension and the bound can be improved using this.*

Proof. We use $\mathcal{T}(x, v, \lambda)_x$ to denote the x component of $\mathcal{T}(x, v, \lambda)$ and similarly for $\mathcal{T}(x, v, \lambda)_v$ and $\mathcal{T}(x, v, \lambda)_\lambda$. For simplicity, we write $g_0 = g(x_0)$, $g_{1/2} = g(x_{1/2})$ and $\overline{g}_{1/2} = g(\overline{x_{1/2}})$. By the assumption, we have that

$$\|x - x_0\|_{g_0} \le \|x - x^*\|_{g_0} + \|x^* - x_0\|_{g_0} \le 2r.$$

Similarly, $\|\overline{x} - x_0\|_{g_0} \le 2r$.

We first bound $\mathcal{T}(x, v, \lambda)_\lambda$. Note that

$$\mathcal{T}(\overline{x}, \overline{v}, \overline{\lambda})_\lambda - \mathcal{T}(x, v, \lambda)_\lambda = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4$$

where

$$\begin{aligned}
\alpha_1 &= (I - (LL^\top)^{-1}Ag_0^{-1}A^\top)(\overline{\lambda} - \lambda),\\
\alpha_2 &= (LL^\top)^{-1}Ag_0^{-1}(\overline{v_{1/2}} - v_{1/2}),\\
\alpha_3 &= (LL^\top)^{-1}A(g_{1/2}^{-1} - g_0^{-1})((\overline{v_{1/2}} - A^\top\overline{\lambda}) - (v_{1/2} - A^\top\lambda)),\\
\alpha_4 &= (LL^\top)^{-1}A(\overline{g}_{1/2}^{-1} - g_{1/2}^{-1})(\overline{v_{1/2}} - A^\top\overline{\lambda}).
\end{aligned}$$

Using that $LL^\top = Ag(x_0)^{-1}A^\top$, we have $\alpha_1 = 0$. For α_2, we have

$$\|A^\top\alpha_2\|^2_{g_0^{-1}} = (\overline{v_{1/2}} - v_{1/2})^\top g_0^{-1}A^\top(LL^\top)^{-1}Ag_0^{-1}A^\top(L^\top L)^{-1}Ag_0^{-1}(\overline{v_{1/2}} - v_{1/2})$$

$$
\begin{aligned}
&= (\overline{v_{1/2}} - v_{1/2})^\top g_0^{-1} A^\top (A g_0^{-1} A^\top)^{-1} A g_0^{-1} (\overline{v_{1/2}} - v_{1/2}) \\
&\le (\overline{v_{1/2}} - v_{1/2})^\top g_0^{-1} (\overline{v_{1/2}} - v_{1/2}) \\
&= \frac{1}{4} \|\overline{v} - v\|_{g_0^{-1}}^2
\end{aligned}
$$

where we use $LL^\top = Ag(x_0)^{-1}A^\top$ and $g_0^{-1/2} A^\top (Ag_0^{-1}A^\top)^{-1} Ag_0^{-1/2} = B^\top (BB^\top)^{-1} B \preceq I$ for $B = Ag_0^{-1/2}$. For α_3, by self-concordance of g (Lemma 25) and $\|x - x_0\|_{g_0} \le 2r$, we have

$$
(1-r)^2 g_0 \preceq g_{1/2} \preceq \frac{1}{(1-r)^2} g_0 \tag{F.10}
$$

and hence $(g_0^{1/2}(g_{1/2}^{-1} - g_0^{-1}) g_0^{1/2})^2 \preceq ((1-r)^{-2} - 1)^2 I$. Using this and $P = g_0^{-1/2} A^\top (Ag_0^{-1}A^\top)^{-1} Ag_0^{-1/2} \preceq I$, we have

$$
\begin{aligned}
\|A^\top \alpha_3\|_{g_0^{-1}} &= \|g_0^{1/2}(g_{1/2}^{-1} - g_0^{-1})((\overline{v_{1/2}} - A^\top \overline{\lambda}) - (v_{1/2} - A^\top \lambda))\|_P \\
&\le \|g_0^{1/2}(g_{1/2}^{-1} - g_0^{-1})((\overline{v_{1/2}} - A^\top \overline{\lambda}) - (v_{1/2} - A^\top \lambda))\|_2 \\
&\le ((1-r)^{-2} - 1) \|g_0^{-1/2}((\overline{v_{1/2}} - A^\top \overline{\lambda}) - (v_{1/2} - A^\top \lambda))\|_2 \\
&\le ((1-r)^{-2} - 1)(\frac{1}{2}\|\overline{v} - v\|_{g_0^{-1}} + \|A^\top(\overline{\lambda} - \lambda)\|_{g_0^{-1}}).
\end{aligned}
$$

Using $r \le 1/10$, we have

$$
\|A^\top \alpha_3\|_{g_0^{-1}} \le 1.2r \|\overline{v} - v\|_{g_0^{-1}} + 2.4r \|A^\top(\overline{\lambda} - \lambda)\|_{g_0^{-1}}.
$$

For α_4, similarly, we have

$$
\begin{aligned}
\|A^\top \alpha_4\|_{g_0^{-1}} &\le ((1 - 0.5\|\overline{x} - x\|_{g_{1/2}})^{-2} - 1) \|\overline{v_{1/2}} - A^\top \overline{\lambda}\|_{g_0^{-1}} \\
&\le ((1 - 0.6\|\overline{x} - x\|_{g_0})^{-2} - 1) \|\overline{v_{1/2}} - A^\top \overline{\lambda}\|_{g_0^{-1}} \\
&\le 1.5 \|\overline{x} - x\|_{g_0} \|\overline{v_{1/2}} - A^\top \overline{\lambda}\|_{g_0^{-1}}
\end{aligned}
$$

where we used $g_{1/2} \preceq 1.2 g_0$ (by (F.10)) in the second inequality and $\|\overline{x} - x\|_{g_0} \le \|\overline{x} - x^*\|_{g_0} + \|x - x^*\|_{g_0} \le \frac{1}{5}$ at the end. Combining everything, we have

$$
\begin{aligned}
&\|A^\top(\mathcal{T}(\overline{x}, \overline{v}, \overline{\lambda})_\lambda - \mathcal{T}(x, v, \lambda)_\lambda)\|_{g_0^{-1}} = \|A^\top(\overline{\lambda}_1 - \lambda_1)\|_{g_0^{-1}} \\
&\quad \le 0.7 \|\overline{v} - v\|_{g_0^{-1}} + 2.4r \|A^\top(\overline{\lambda} - \lambda)\|_{g_0^{-1}} + 1.5 \|\overline{x} - x\|_{g_0} \|\overline{v_{1/2}} - A^\top \overline{\lambda}\|_{g_0^{-1}}.
\end{aligned} \tag{F.11}
$$

Now we bound $\mathcal{T}(x, v, \lambda)_x$. Note that

$$
\mathcal{T}(\overline{x}, \overline{v}, \overline{\lambda})_x - \mathcal{T}(x, v, \lambda)_x = h\beta_1 + h\beta_2
$$

where

$$\begin{aligned}\beta_1 =& g_{1/2}^{-1}((\overline{v}_{1/2} - A^\top \overline{\lambda}_1) - (v_{1/2} - A^\top \lambda_1)),\\ \beta_2 =& (\overline{g}_{1/2}^{-1} - g_{1/2}^{-1})(\overline{v}_{1/2} - A^\top \overline{\lambda}_1).\end{aligned}$$

By a proof similar to above, we have

$$\begin{aligned}\|\beta_1\|_{g_0} &\le 1.2(\|\overline{v}_{1/2} - v_{1/2}\|_{g_0^{-1}} + \|A^\top(\overline{\lambda}_1 - \lambda_1)\|_{g_0^{-1}}),\\ \|\beta_2\|_{g_0} &\le 0.6\|\overline{x} - x\|_{g_0}\|\overline{v}_{1/2} - A^\top \overline{\lambda}_1\|_{g_0^{-1}}.\end{aligned}$$

and thus

$$\begin{aligned}&\|\mathcal{T}(\overline{x}, \overline{v}, \overline{\lambda})_x - \mathcal{T}(x, v, \lambda)_x\|_{g_0}\\ &\quad\le 0.6h\|\overline{v} - v\|_{g_0^{-1}} + 1.2h\|A^\top(\overline{\lambda}_1 - \lambda_1)\|_{g_0^{-1}} + 0.6h\|\overline{x} - x\|_{g_0}\|\overline{v}_{1/2} - A^\top \overline{\lambda}_1\|_{g_0^{-1}}.\end{aligned}$$

Finally, we bound $\mathcal{T}(x, v, \lambda)_v$. We split the term

$$\mathcal{T}(\overline{x}, \overline{v}, \overline{\lambda})_v - \mathcal{T}(x, v, \lambda)_v = \frac{h}{2}\gamma_1 + \frac{h}{2}\gamma_2$$

where

$$\begin{aligned}\gamma_1 =& Dg(x_{1/2})[\overline{g}_{1/2}^{-1}(\overline{v}_{1/2} - A^\top \overline{\lambda}_1), \overline{g}_{1/2}^{-1}(\overline{v}_{1/2} - A^\top \overline{\lambda}_1)]\\ &- Dg(x_{1/2})[g_{1/2}^{-1}(v_{1/2} - A^\top \lambda_1), g_{1/2}^{-1}(v_{1/2} - A^\top \lambda_1)],\\ \gamma_2 =& Dg(\overline{x}_{1/2})[\overline{g}_{1/2}^{-1}(\overline{v}_{1/2} - A^\top \overline{\lambda}_1), \overline{g}_{1/2}^{-1}(\overline{v}_{1/2} - A^\top \overline{\lambda}_1)]\\ &- Dg(x_{1/2})[\overline{g}_{1/2}^{-1}(\overline{v}_{1/2} - A^\top \overline{\lambda}_1), \overline{g}_{1/2}^{-1}(\overline{v}_{1/2} - A^\top \overline{\lambda}_1)].\end{aligned}$$

Let $\overline{\eta} = \overline{g}_{1/2}^{-1}(\overline{v}_{1/2} - A^\top \overline{\lambda}_1)$ and $\eta = g_{1/2}^{-1}(v_{1/2} - A^\top \lambda_1)$. For γ_1, we have that

$$\begin{aligned}&\|Dg(x_{1/2})[\overline{\eta}, \overline{\eta}] - Dg(x_{1/2})[\eta, \eta]\|_{g_{1/2}^{-1}}\\ &\quad\le 2\|Dg(x_{1/2})[\overline{\eta} - \eta, \overline{\eta}]\|_{g_{1/2}^{-1}} + \|Dg(x_{1/2})[\overline{\eta} - \eta, \overline{\eta} - \eta]\|_{g_{1/2}^{-1}}\\ &\quad\le 4\|\overline{\eta} - \eta\|_{g_{1/2}}\|\overline{\eta}\|_{g_{1/2}} + 2\|\overline{\eta} - \eta\|_{g_{1/2}}^2\end{aligned}$$

where we use Lemma 25. Using $g_{1/2} \preceq 1.2g_0$ (by (F.10)),

$$\begin{aligned}\|\gamma_1\|_{g_0^{-1}} \le& 4\|(\overline{v}_{1/2} - A^\top \overline{\lambda}_1) - (v_{1/2} - A^\top \lambda_1)\|_{g_0^{-1}}\|\overline{v}_{1/2} - A^\top \overline{\lambda}_1\|_{g_0^{-1}}\\ &+ 2\|(\overline{v}_{1/2} - A^\top \overline{\lambda}_1) - (v_{1/2} - A^\top \lambda_1)\|_{g_0^{-1}}^2.\end{aligned}$$

For γ_2, we use Lemma 25 and get

$$\|\gamma_2\|_{g_0^{-1}} \le \frac{4}{(1 - 0.6\|\overline{x} - x\|_{g_0})^3}\|\overline{v}_{1/2} - A^\top \overline{\lambda}_1\|_{g_0^{-1}}^2\|\overline{x} - x\|_{g_0}$$

$$\leq 6\|\overline{v}_{1/2} - A^\top \overline{\lambda}_1\|^2_{g_0^{-1}} \|\overline{x} - x\|_{g_0}.$$

Combining everything, we have

$$\begin{aligned}
&\|\mathcal{T}(\overline{x}, \overline{v}, \overline{\lambda})_v - \mathcal{T}(x, v, \lambda)_v\|_{g_0^{-1}} \\
&\quad \leq 2h\|(\overline{v}_{1/2} - A^\top \overline{\lambda}_1) - (v_{1/2} - A^\top \lambda_1)\|_{g_0^{-1}} \|\overline{v}_{1/2} - A^\top \overline{\lambda}_1\|_{g_0^{-1}} \\
&\qquad + h\|(\overline{v}_{1/2} - A^\top \overline{\lambda}_1) - (v_{1/2} - A^\top \lambda_1)\|^2_{g_0^{-1}} \\
&\qquad + 3h\|\overline{v}_{1/2} - A^\top \overline{\lambda}_1\|^2_{g_0^{-1}} \|\overline{x} - x\|_{g_0}
\end{aligned}$$

□

Combining the bounds for $\mathcal{T}_\lambda, \mathcal{T}_x, \mathcal{T}_v$, we have

$$\begin{aligned}
&\|\mathcal{T}(x, v, \lambda) - \mathcal{T}(\overline{x}, \overline{v}, \overline{\lambda})\| \\
&\leq 0.7h\|\overline{v} - v\|_{g_0^{-1}} + 2.4rh\|A^\top(\overline{\lambda} - \lambda)\|_{g_0^{-1}} + 1.5h\|\overline{x} - x\|_{g_0} \|\overline{v_{1/2}} - A^\top \overline{\lambda}\|_{g_0^{-1}} \\
&\quad + 0.6h\|\overline{v} - v\|_{g_0^{-1}} + 1.2h\|A^\top(\overline{\lambda}_1 - \lambda_1)\|_{g_0^{-1}} + 0.6h\|\overline{x} - x\|_{g_0} \|\overline{v}_{1/2} - A^\top \overline{\lambda}_1\|_{g_0^{-1}} \\
&\quad + 2h\|(\overline{v}_{1/2} - A^\top \overline{\lambda}_1) - (v_{1/2} - A^\top \lambda_1)\|_{g_0^{-1}} \|\overline{v}_{1/2} - A^\top \overline{\lambda}_1\|_{g_0^{-1}} \\
&\quad + h\|(\overline{v}_{1/2} - A^\top \overline{\lambda}_1) - (v_{1/2} - A^\top \lambda_1)\|^2_{g_0^{-1}} \\
&\quad + 3h\|\overline{v}_{1/2} - A^\top \overline{\lambda}_1\|^2_{g_0^{-1}} \|\overline{x} - x\|_{g_0}.
\end{aligned}$$

To simplify the terms, we note that

$$\begin{aligned}
\|\overline{v}_{1/2} - A^\top \overline{\lambda}_1\|_{g_0^{-1}} \leq& \|\overline{v}_{1/2} - A^\top \overline{\lambda}\|_{g_0^{-1}} + \|A^\top (LL^\top)^{-1} A \overline{g}^{-1}_{1/2} \left(\overline{v}_{1/2} - A^\top \overline{\lambda}\right)\|_{g_0^{-1}} \\
=& \|\overline{v}_{1/2} - A^\top \overline{\lambda}\|_{g_0^{-1}} + \|g_0^{1/2} \overline{g}^{-1}_{1/2} \left(\overline{v}_{1/2} - A^\top \overline{\lambda}\right)\|_P \\
\leq& \|\overline{v}_{1/2} - A^\top \overline{\lambda}\|_{g_0^{-1}} + \|g_0^{1/2} \overline{g}^{-1}_{1/2} \left(\overline{v}_{1/2} - A^\top \overline{\lambda}\right)\|_2 \\
\leq& 3\|\overline{v}_{1/2} - A^\top \overline{\lambda}\|_{g_0^{-1}}.
\end{aligned}$$

Using this and simplifying, we have

$$\begin{aligned}
&\|\mathcal{T}(x, v, \lambda) - \mathcal{T}(\overline{x}, \overline{v}, \overline{\lambda})\| \\
\leq& 1.3h\|\overline{v} - v\|_{g_0^{-1}} + 2.4rh\|A^\top(\overline{\lambda} - \lambda)\|_{g_0^{-1}} + 3.3h\|\overline{x} - x\|_{g_0} \|\overline{v_{1/2}} - A^\top \overline{\lambda}\|_{g_0^{-1}} \\
&\quad + 1.2h\|A^\top(\overline{\lambda}_1 - \lambda_1)\|_{g_0^{-1}} \\
&\quad + 6h(\frac{1}{2}\|\overline{v} - v\|_{g_0^{-1}} + \|A^\top(\overline{\lambda}_1 - \lambda_1)\|_{g_0^{-1}}) \|\overline{v}_{1/2} - A^\top \overline{\lambda}\|_{g_0^{-1}} \\
&\quad + h\|\overline{v} - v\|^2_{g_0^{-1}} + 2h\|A^\top(\overline{\lambda}_1 - \lambda_1)\|^2_{g_0^{-1}}
\end{aligned}$$

$$+ 27h\|\overline{v}_{1/2} - A^\top\overline{\lambda}\|_{g_0^{-1}}^2\|\overline{x} - x\|_{g_0}.$$

Next, we note that

$$\begin{aligned}\|\overline{v_{1/2}} - A^\top\overline{\lambda}\|_{g_0^{-1}} \leq& \frac{1}{2}\|\overline{v} - v^*\|_{g_0^{-1}} + \frac{1}{2}\|v_0 - v^*\|_{g_0^{-1}} + \|v^*\|_{g_0^{-1}} \\ &+ \frac{1}{2}\|A^\top\overline{\lambda} - A^\top\lambda^*\|_{g_0^{-1}} + \frac{1}{2}\|A^\top\overline{\lambda} - A^\top\lambda^*\|_{g_0^{-1}} + \|A^\top\lambda^*\|_{g_0^{-1}} \\ \leq& \frac{1}{2}r + \frac{1}{2}r + \|v^*\|_{g_0^{-1}} + \frac{r}{2h} + \frac{r}{2h} + \frac{r}{h} \leq \frac{3r}{h} + \|v^*\|_{g_0^{-1}}\end{aligned}$$

Using this, (F.11), $h \leq r$, $r^2 \leq \frac{h}{16}$, $r \leq \|v^*\|_{g_0^{-1}}/4$, we have

$$\begin{aligned}\|A^\top(\overline{\lambda}_1 - \lambda_1)\|_{g_0^{-1}} \leq& \|\overline{v} - v\|_{g_0^{-1}} + 3r\|A^\top(\overline{\lambda} - \lambda)\|_{g_0^{-1}} + (\frac{5r}{h} + 2\|v^*\|_{g_0^{-1}})\|\overline{x} - x\|_{g_0} \\ \leq& r + \frac{3r^2}{h} + \frac{5r^2}{h} + 2r\|v^*\|_{g_0^{-1}} \leq \frac{8r^2}{h} + 2r\|v^*\|_{g_0^{-1}} \leq 1\end{aligned}$$

Hence, we can further simplify it to

$$\begin{aligned}&\|\mathcal{T}(x, v, \lambda) - \mathcal{T}(\overline{x}, \overline{v}, \overline{\lambda})\| \\ \leq& 2.3h\|\overline{v} - v\|_{g_0^{-1}} + 2.4rh\|A^\top(\overline{\lambda} - \lambda)\|_{g_0^{-1}} + 3.3h\|\overline{x} - x\|_{g_0}\|\overline{v_{1/2}} - A^\top\overline{\lambda}\|_{g_0^{-1}} \\ &+ 3.2h\|A^\top(\overline{\lambda}_1 - \lambda_1)\|_{g_0^{-1}} \\ &+ 6h(\frac{1}{2}\|\overline{v} - v\|_{g_0^{-1}} + \|A^\top(\overline{\lambda}_1 - \lambda_1)\|_{g_0^{-1}})\|\overline{v}_{1/2} - A^\top\overline{\lambda}\|_{g_0^{-1}} \\ &+ 27h\|\overline{v}_{1/2} - A^\top\overline{\lambda}\|_{g_0^{-1}}^2\|\overline{x} - x\|_{g_0} \\ \leq& (\frac{3r}{h} + \|v^*\|_{g_0^{-1}})(6h\|\overline{v} - v\|_{g_0^{-1}} + 9h\|A^\top(\overline{\lambda}_1 - \lambda_1)\|_{g_0^{-1}} + 31h\|\overline{x} - x\|_{g_0}) \\ &+ 2.4rh\|A^\top(\overline{\lambda} - \lambda)\|_{g_0^{-1}}\end{aligned}$$

where we used $\|\overline{v}_{1/2} - A^\top\overline{\lambda}\|_{g_0^{-1}} \leq \frac{3r}{h} + \|v^*\|_{g_0^{-1}}$ and $r \geq h$. Using the bound on $\|A^\top(\overline{\lambda}_1 - \lambda_1)\|_{g_0^{-1}}$, we have

$$\begin{aligned}&\|\mathcal{T}(x, v, \lambda) - \mathcal{T}(\overline{x}, \overline{v}, \overline{\lambda})\| \\ \leq& (\frac{3r}{h} + \|v^*\|_{g_0^{-1}})(15h\|\overline{v} - v\|_{g_0^{-1}} + 27rh\|A^\top(\overline{\lambda} - \lambda)\|_{g_0^{-1}} + 9h(\frac{36r}{h} + 2\|v^*\|_{g_0^{-1}})\|\overline{x} - x\|_{g_0}) \\ &+ 2.4rh\|A^\top(\overline{\lambda} - \lambda)\|_{g_0^{-1}} \\ \leq& (\frac{3r}{h} + \frac{1}{4r})(15h\|\overline{v} - v\|_{g_0^{-1}} + 30rh\|A^\top(\overline{\lambda} - \lambda)\|_{g_0^{-1}} + 9h(\frac{36r}{h} + 2\|v^*\|_{g_0^{-1}})\|\overline{x} - x\|_{g_0}) \\ \leq& (\frac{3r}{h} + \|v^*\|_{g_0^{-1}})(400r + 18h\|v^*\|_{g_0^{-1}})\|(x, v, \lambda) - (\overline{x}, \overline{v}, \overline{\lambda})\|.\end{aligned}$$

F.4 Condition Number Independence via Self-concordant Barrier

In this section, we analyze the convergence rates of the ideal CRHMC and discretized RHMC in our setting respectively, showing that both are independent of condition num-

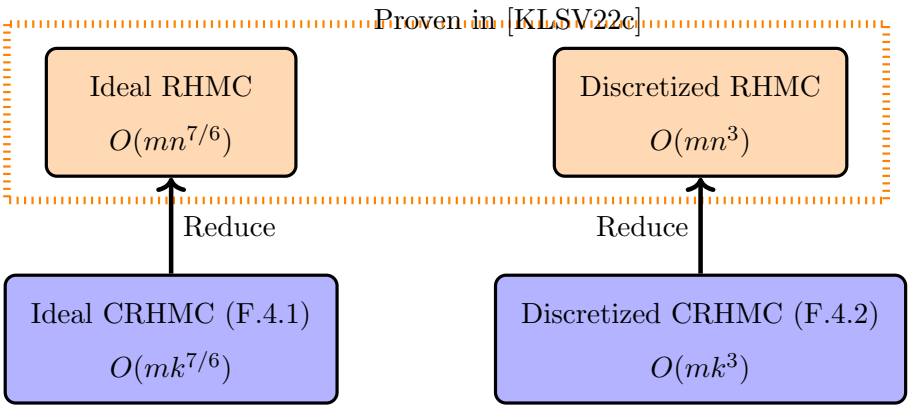


Figure F.1: Proof outline for the mixing rates of CRHMC

bers. We only show the case when f **is linear**,

$$\pi(x) \propto e^{-f(x)} = e^{-\alpha^\top x}, \text{ for some } \alpha \in \mathbb{R}^n.$$

However, recall that **all logconcave densities** can be reduced to this linear case (see (7.2)). We also focus on when a manifold $\mathcal{M}$ is a polytope in the form of $\{x \in \mathbb{R}^n : A'x \geq b',\ Ax = 0\}$ for full-rank $A' \in \mathbb{R}^{m\times n}, A \in \mathbb{R}^{p\times n}$ and $b' \in \mathbb{R}^m$, with the Riemannian metric induced by the Hessian of the logarithmic barrier of the polytope. For simplicity, we consider the case when there is no momentum (i.e., $\beta = 0$) in Algorithm 15. In addition, we consider a *lazy* version of Algorithm 15 to avoid a uniqueness issue of the stationary distribution of the Markov chain. The lazy version of the Markov chain, at each step, does nothing with probability $\frac{1}{2}$ (in other words, stays at where it is and does not move). Note that this change for the purpose of proof worsens the mixing rate only by a factor of 2.

In this setting, we show that the mixing rates of the ideal CRHMC and the discretized CRHMC (Algorithm 15) are $O\left(mk^{7/6}\log^2\frac{\Lambda}{\epsilon}\right)$ and $O\left(mk^3\log^3\frac{\Lambda}{\epsilon}\right)$, where Λ is a warmness parameter and k is the dimension of the constrained space defined by $\{x \in \mathbb{R}^n : Ax = 0\}$. We remark that our algorithm is actually independent of condition number (i.e., no dependency on $\|\alpha\|_2$ and the geometry of polytope). This is the key reason that our sampler is much more efficient for skewed instances than previous samplers.

We first shed light on how RHMC and CRHMC can be related (see Figure F.1), establishing a correspondence between RHMC (full-dimensional space) and CRHMC (constrained space). This connection enables us to refer to the mixing rates of the ideal RHMC and discretized RHMC proven in [KLSV22c]. To be precise, we prove in Section F.4.1 the

following theorem on the mixing rate of the ideal CRHMC, which can solve the Hamiltonian equations accurately without any error.

Theorem 28 (Mixing rate of ideal CRHMC)**.** *Let π_T be the distribution obtained after T iterations of the ideal CRHMC on a convex body $\mathcal{M} = \{x \in \mathbb{R}^n : A'x \geq b',\, Ax = 0\}$. Let $\Lambda = \sup_{S \subseteq \mathcal{M}} \frac{\pi_0(S)}{\pi(S)}$ be the warmness of the initial distribution π_0. For any $\epsilon > 0$, there exists $T = O\left(mk^{7/6} \log^2 \frac{\Lambda}{\epsilon}\right)$ such that $\|\pi_T - \pi\|_{\mathrm{TV}} \leq \epsilon$, where k is the dimension of the constrained space defined by $\{x \in \mathbb{R}^n : Ax = 0\}$.*

We then prove in Section F.4.2 the convergence rate of the discretized RHMC (Algorithm 15).

Theorem 29 (Mixing rate of discretized CRHMC)**.** *Let π_T be the distribution obtained after T iterations of Algorithm 15 on a convex body $\mathcal{M} = \{x \in \mathbb{R}^n : A'x \geq b',\, Ax = 0\}$. Let $\Lambda = \sup_{S \subseteq \mathcal{M}} \frac{\pi_0(S)}{\pi(S)}$ be the warmness of the initial distribution π_0. For any $\epsilon > 0$, there exists $T = O\left(mk^3 \log^3 \frac{\Lambda}{\epsilon}\right)$ such that $\|\pi_T - \pi\|_{\mathrm{TV}} \leq \epsilon$, where k is the dimension of the constrained space defined by $\{x \in \mathbb{R}^n : Ax = 0\}$.*

We believe there is room for improvement on the n-dependence via a more careful analysis.

F.4.1 Convergence rate of ideal CRHMC

Lee and Vempala [LV18] first analyzed Riemannian Hamiltonian Monte Carlo (RHMC) on n-dimensional polytopes embedded in $\mathbb{R}^n$, with an invertible metric induced by the Hessian of the logarithmic barrier of the polytopes. They bounded the mixing rate in terms of *smoothness* parameters that depend on the manifold. In particular for uniform sampling, they showed that the convergence rate of RHMC is $O(mn^{2/3})$. Subsequently, [KLSV22c] extended their analysis to exponential densities and further analyzed the convergence rate of RHMC discretized by the implicit midpoint method, showing that the mixing rates are $O(mn^{7/6})$ and $O(mn^3)$, respectively.

However, our metric $M(x)$ defined for the constrained space could be singular in the underlying space $\mathbb{R}^n$, so we cannot directly refer to any theoretical results from [LV18, KLSV22c]. To address this challenge, we establish a formalism that allows us to reduce the ideal CRHMC to the ideal RHMC, obtaining the mixing rate through this reduction.

Even though our convex body $\mathcal{M} = \{x \in \mathbb{R}^n : A'x \geq b',\, Ax = 0\}$ of dimension k is embedded in $\mathbb{R}^n$, we can handle it with an invertible metric $\overline{g}$ on $\mathcal{M}$ properly defined as

if it is embedded in $\mathbb{R}^k$. To this end, we use $\{u_1, ..., u_k\}$ to denote an orthonormal basis of the constrained space (which is the null space of A) and extend it to an orthonormal basis of $\mathbb{R}^n$ denoted by $\{u_1, ..., u_k, ..., u_n\}$. We also define two matrices $U_k \in \mathbb{R}^{n\times k}$ and $U \in \mathbb{R}^{n\times n}$ by

$$U_k = \begin{bmatrix} u_1 & \cdots & u_k \end{bmatrix} \quad \& \quad U = \begin{bmatrix} U_k & u_{k+1} & \cdots & u_n \end{bmatrix}.$$

Using this orthonormal basis $\{u_1, ..., u_k\}$, we can consider a new coordinate system $y = (y_1, ..., y_k) \in \mathbb{R}^k$ on the k-dimensional manifold $\mathcal{M}$. Moreover, there exists one-to-one correspondence between y and x; for any $x \in \mathcal{M}$ there is a unique y such that $x = U_k y$, and we can recover this y by multiplying $U_k^\top$ (i.e., $y = U_k^\top x$).

Let us define the invertible local metric $\overline{g}$ at $y \in \mathcal{M}$ by

$$\overline{g}(y)(u_i, u_j) \stackrel{\text{def}}{=} g(x)(u_i, u_j) \quad \text{for } i, j \le k.$$

With abuse of notations, we also use $\overline{g}(y)$ to denote the $k \times k$ matrix with its (i, j)-entry being $\overline{g}(y)(u_i, u_j)$. We first establish relationships between $\overline{g}(y)$ (and its inverse $\overline{g}^{-1}$) and $M(x)$ (and its pseudoinverse $W \stackrel{\text{def}}{=} M(x)^\dagger$). We recall that for the orthogonal projection Q to the null space of A

$$M(x) = Q^\top g(x) Q, \; W(x) = g(x)^{-\frac{1}{2}}(I - P(x))g(x)^{-\frac{1}{2}}.$$

Lemma 30. *We have* $\overline{g}(y) = U_k^\top M(x) U_k = U_k^\top g(x) U_k$ *and* $\overline{g}(y)^{-1} = U_k^\top W(x) U_k$.

Proof. It is immediate from the definition of $\overline{g}$ that $\overline{g}(y) = U_k^\top g(x) U_k$. Since the quadratic forms of $M(x)$ and $g(x)$ agree on the constrained space, we also have $\overline{g}(y) = U_k^\top M(x) U_k$.

For $\overline{g}^{-1}$, we define two matrices $P_k \in \mathbb{R}^{n\times k}$ and $P_r \in \mathbb{R}^{n\times(n-k)}$ by

$$P_k = \begin{bmatrix} I_k \\ 0_{(n-k)\times k} \end{bmatrix}, \; P_r = \begin{bmatrix} 0_{k\times(n-k)} \\ I_{n-k} \end{bmatrix}$$

where $0_{(n-k)\times k}$ is the zero matrix of size $(n-k) \times k$, I_k is the identity matrix of size $k \times k$ and so on. Due to $U_k = U P_k$, the upper-left $k \times k$ submatrix of $g'(x) := U^\top g(x) U$ is exactly $\overline{g}(y)$. Let us represent the inverse of g' in the form of block matrix:

$$g'(x)^{-1} = \begin{bmatrix} B_1 & B_2 \\ B_2^\top & B_3 \end{bmatrix},$$

for $B_1 \in \mathbb{R}^{k\times k}$, $B_2 \in \mathbb{R}^{k\times(n-k)}$ and $B_3 \in \mathbb{R}^{(n-k)\times(n-k)}$. Using the formula of the inverse of block matrices (see App. F.5.3),

$$\overline{g}(y)^{-1} = B_1 - B_2 B_3^{-1} B_2^\top.$$

It is straightforward to check

$$\begin{aligned} B_1 &= P_k^\top g'(x)^{-1} P_k = P_k^\top U^\top g(x)^{-1} U P_k \\ &= U_k^\top g(x)^{-1} U_k, \\ B_2 &= P_k^\top g'(x)^{-1} P_r = U_k^\top g(x)^{-1} U_r, \\ B_3 &= P_r^\top g'(x)^{-1} P_r = U_r^\top g(x)^{-1} U_r, \end{aligned}$$

for $U_r = \begin{bmatrix} u_{k+1} & \cdots & u_n \end{bmatrix} \in \mathbb{R}^{n\times(n-k)}$. Therefore,

$$\begin{aligned} \overline{g}(y)^{-1} &= U_k^\top g(x)^{-1} U_k - U_k^\top g(x)^{-1} U_r (U_r^\top g(x)^{-1} U_r)^{-1} U_r^\top g(x)^{-1} U_k \\ &= U_k^\top \left(g(x)^{-1} - g(x)^{-1} U_r (U_r^\top g(x)^{-1} U_r)^{-1} U_r^\top g(x)^{-1} \right) U_k. \end{aligned}$$

Since $g(x)^{-\frac{1}{2}} U_r (U_r^\top g(x)^{-1} U_r)^{-1} U_r^\top g(x)^{-\frac{1}{2}}$ is the orthogonal projection to the row space of $U_r^\top g(x)^{-\frac{1}{2}}$ and this row space is the same with the row space of $Ag(x)^{-\frac{1}{2}}$, the uniqueness of orthogonal projection matrices implies

$$g(x)^{-\frac{1}{2}} A^\top (Ag(x)^{-1} A^\top)^{-1} Ag(x)^{-\frac{1}{2}} = g(x)^{-\frac{1}{2}} U_r (U_r^\top g(x)^{-1} U_r)^{-1} U_r^\top g(x)^{-\frac{1}{2}}.$$

Therefore,

$$\begin{aligned} \overline{g}(y)^{-1} &= U_k^\top \left(g(x)^{-1} - g(x)^{-1} A^\top (Ag(x)^{-1} A^\top)^{-1} Ag(x)^{-1} \right) U_k \\ &= U_k^\top \left(g(x)^{-\frac{1}{2}} \left(I - g(x)^{-\frac{1}{2}} A^\top (Ag(x)^{-1} A^\top)^{-1} Ag(x)^{-\frac{1}{2}} \right) g(x)^{-\frac{1}{2}} \right) U_k \\ &= U_k^\top \left(g(x)^{-\frac{1}{2}} (I - P(x)) g(x)^{-\frac{1}{2}} \right) U_k \\ &= U_k^\top W(x) U_k. \end{aligned}$$

□

We can now view the ideal CRHMC with the metric $M(x)$ as the ideal RHMC with the metric $\overline{g}$ on the k-dimensional manifold. Note that we have to ensure that the local metric $\overline{g}$ is also induced by the Hessian of a logarithmic barrier, in order to refer to results from it.

Lemma 31. *Let $\overline{A'} = A'U_k$ and $\psi(y)$ be the logarithmic barrier of the k-dimensional polytope defined by $\{y \in \mathbb{R}^k : \overline{A'} y \leq b'\}$. Then $\nabla_y^2 \psi(y) = \overline{g}(y)$.*

Proof. Observe that $\{y \in \mathbb{R}^k : \overline{A'} y \geq b'\}$ is the new representation of $\mathcal{M} = \{x \in \mathbb{R}^n : A'x \geq b', Ax = 0\}$ in the y-coordinate system. Due to $\overline{A'} y = Ax$, we have $S_x = \mathrm{Diag}(A'x - b') =$

$\mathrm{Diag}(\overline{A'}y - b) = S_y$. For the logarithmic barrier $\phi(x)$ of $\{x \in \mathbb{R}^n : A'x \geq b\}$, direct computation results in

$$\begin{aligned}\nabla_y^2 \psi(y) &= \overline{A'}^\top S_y^{-2} \overline{A'} = U_k^\top A'^\top S_x^{-2} A' U_k = U_k^\top \nabla_x^2 \phi(x) U_k \\ &= U_k^\top g(x) U_k = \overline{g}(y),\end{aligned}$$

where we used $\nabla_x^2 \phi(x) = A'^\top S_x^{-2} A'$ in the third equality and Lemma 30 in the last equality. □

Most importantly, we prove that this ideal RHMC on the k-dimensional manifold with the metric $\overline{g}(y)$ is equivalent to the ideal CRHMC with the metric $M(x)$.

Lemma 32. *The dynamics (x, v) and (y, u) of the ideal CRHMC in $\mathbb{R}^n$ and the ideal RHMC in $\mathbb{R}^k$ are equivalent in a sense that the Hamiltonian equations for (x, v) can be obtained by lifting up the Hamiltonian equations for (y, u) from $\mathbb{R}^k$ to $\mathbb{R}^n$ via multiplying U_k. That is, when we lift up the dynamics (y, u) in $\mathbb{R}^k$ to the dynamics $(\overline{x}, \overline{v})$ in $\mathbb{R}^n$ defined by $\overline{x} = U_k y$ and $\overline{v} = U_k u$, it follows that*

$$\frac{dx}{dt} = \frac{d\overline{x}}{dt}, \frac{dv}{dt} = \frac{d\overline{v}}{dt} \text{ and } v, \overline{v} \sim \mathbf{N}(0, M(x)).$$

Proof. We first recall from the proof of Lemma 18 that the Hamiltonian equations of (x, v) are

$$\begin{aligned}\frac{dx}{dt} &= W(x)v, \\ \frac{dv}{dt} &= -\nabla_x f(x) - \frac{1}{2}\mathrm{Tr}\left[W(x)Dg(x)\right] + \frac{1}{2}Dg(x)\left[\frac{dx}{dt}, \frac{dx}{dt}\right] - A^\top \lambda(x, v) \\ &= -\nabla_x f(x) - \frac{1}{2}\mathrm{Tr}\left[W(x)Dg(x)\right] + \frac{1}{2}Dg(x)\left[\frac{dx}{dt}, \frac{dx}{dt}\right] \\ &\quad - A^\top \left(AA^\top\right)^{-1} A\left(-\nabla_x f(x) - \frac{1}{2}\mathrm{Tr}\left[W(x)Dg(x)\right] + \frac{1}{2}Dg(x)\left[\frac{dx}{dt}, \frac{dx}{dt}\right]\right) \\ &= (I - A^\top \left(AA^\top\right)^{-1} A)\left(-\nabla_x f(x) - \frac{1}{2}\mathrm{Tr}\left[W(x)Dg(x)\right] + \frac{1}{2}Dg(x)\left[\frac{dx}{dt}, \frac{dx}{dt}\right]\right) \\ &= U_k U_k^\top \left(-\nabla_x f(x) - \frac{1}{2}\mathrm{Tr}\left[W(x)Dg(x)\right] + \frac{1}{2}Dg(x)\left[\frac{dx}{dt}, \frac{dx}{dt}\right]\right),\end{aligned}$$

where the last equality follows from that $U_k U_k^\top$ is the orthogonal projection to the null space of A.

From Lemma 7 of [LV18], the Hamiltonian equations of (y, u) are

$$\frac{dy}{dt} = \overline{g}(y)^{-1} u,$$

$$\frac{du}{dt} = -\nabla_y f(U_k y) - \frac{1}{2}\mathrm{Tr}\left[\overline{g}(y)^{-1} D\overline{g}(y)\right] + \frac{1}{2} D\overline{g}(y)\left[\frac{dy}{dt}, \frac{dy}{dt}\right].$$

From the definitions of $(\overline{x}, \overline{v})$, we have

$$\begin{aligned}
\frac{d\overline{x}}{dt} &= U_k \overline{g}(y)^{-1} u = U_k U_k^\top W(\overline{x}) U_k u = U_k U_k^\top W(\overline{x})\overline{v} \\
&\underset{(*)}{=} W(\overline{x})\overline{v}, \\
\frac{d\overline{v}}{dt} &= U_k \left(-\nabla_y f(U_k y) - \frac{1}{2}\mathrm{Tr}\left[\overline{g}(y)^{-1} D\overline{g}(y)\right] + \frac{1}{2} D\overline{g}(y)\left[\frac{dy}{dt}, \frac{dy}{dt}\right]\right),
\end{aligned}$$

where $(*)$ follows from that $W(x)\overline{v}$ is already in the constrained space (i.e., the null space of A). Let us examine each term in $d\overline{v}/dt$ separately.

$$\begin{aligned}
\nabla_y f(U_k y) &= U_k^\top \nabla_{\overline{x}} f(\overline{x}), \\
\mathrm{Tr}\left[\overline{g}(y)^{-1} D_y \overline{g}(y)\right] &= \mathrm{Tr}\left[U_k^\top W(\overline{x}) U_k \cdot D_y \left(U_k^\top g(U_k y) U_k\right)\right] \\
&= \mathrm{Tr}\left[U_k U_k^\top W(\overline{x}) U_k U_k^\top (U_k^\top D_{\overline{x}} g(\overline{x}))\right] \\
&= \mathrm{Tr}\left[W(\overline{x}) U_k U_k^\top (U_k^\top D_{\overline{x}} g(\overline{x}))\right] \\
&= \mathrm{Tr}\left[(U_k U_k^\top W(\overline{x}))^\top (U_k^\top D_{\overline{x}} g(\overline{x}))\right] \\
&= \mathrm{Tr}\left[W(\overline{x})(U_k^\top D_{\overline{x}} g(\overline{x}))\right] \\
&= U_k^\top \mathrm{Tr}\left[W(\overline{x}) D_{\overline{x}} g(\overline{x})\right], \\
D_y \overline{g}(y)\left[\frac{dy}{dt}, \frac{dy}{dt}\right] &= U_k^\top (U_k^\top D_{\overline{x}} g(\overline{x})) U_k \left[\frac{dy}{dt}, \frac{dy}{dt}\right] \\
&= \left(\overline{v}^\top W(\overline{x}) U_k\right) U_k^\top (U_k^\top D_{\overline{x}} g(\overline{x})) U_k \left(U_k^\top W(\overline{x})\overline{v}\right) \\
&= \overline{v}^\top W(\overline{x})(U_k^\top D_{\overline{x}} g(\overline{x})) W(\overline{x})\overline{v} \\
&= U_k^\top D_{\overline{x}} g(\overline{x})\left[\frac{d\overline{x}}{dt}, \frac{d\overline{x}}{dt}\right].
\end{aligned}$$

Putting all these together, the Hamiltonian equations of $(\overline{x}, \overline{v})$ can be written as

$$\begin{aligned}
\frac{d\overline{x}}{dt} &= W(\overline{x})\overline{v}, \\
\frac{d\overline{v}}{dt} &= U_k U_k^\top \left(-\nabla f(\overline{x}) - \frac{1}{2}\mathrm{Tr}\left[W(\overline{x}) Dg(\overline{x})\right] + \frac{1}{2} Dg(\overline{x})\left[\frac{d\overline{x}}{dt}, \frac{d\overline{x}}{dt}\right]\right).
\end{aligned}$$

Therefore, the Hamiltonian equations of (x, v) and $(\overline{x}, \overline{v})$ are exactly the same. In addition, $\overline{v} = U_k u$ leads to $\overline{v} \sim \mathbf{N}(0, U_k \overline{g}(y) U_k^\top) = \mathbf{N}(0, M(x))$. □

Using these three lemmas, we conclude that the dynamics of the ideal CRHMC on the constrained space is equivalent to that of the ideal RHMC on the corresponding k-dimensional polytope. Therefore, Theorem 28 immediately follows from Corollary 3 in [KLSV22c].

Corollary 23. *Let π be a target distribution on a polytope with m constraints in $\mathbb{R}^n$ such that $\frac{d\pi}{dx} \sim e^{-\alpha^\top x}$ for $\alpha \in \mathbb{R}^n$. Let $\mathcal{M}$ be the Hessian manifold of the polytope induced by the logarithmic barrier of the polytope. Let $\Lambda = \sup_{S \subset \mathcal{M}} \frac{\pi_0(S)}{\pi(S)}$ be the warmness of the initial distribution π_0. Let π_T be the distribution obtained after T iterations of the ideal RHMC on $\mathcal{M}$. For any $\varepsilon > 0$ and step size $h = O\left(\frac{1}{n^{7/12} \log^{1/2} \frac{\Lambda}{\varepsilon}}\right)$, there exists $T = O\left(mn^{7/6} \log^2 \frac{\Lambda}{\varepsilon}\right)$ such that $\|\pi_T - \pi\|_{TV} \leq \varepsilon$.*

F.4.2 Convergence rate of discretized CRHMC

We attempt to demonstrate a similar reduction of the discretized CRHMC. However, it is trickier than that of the ideal RHMC, since Algorithm 15 uses the simplified Hamiltonian, which omits the Lagrangian term $c(x)^\top \lambda(x, v)$, in place of the full Hamiltonian.

We look into the reduction in two steps. First of all, we show in Section F.4.2 that the dynamics of x is the same under the discretized CRHMC via IMM with the simplified Hamiltonian, and the discretized CRHMC via IMM with the full Hamiltonian, and that the acceptance probabilities are the same as well. Next in Section F.4.2, we show a correspondence between the discretized CRHMC via IMM and the discretized RHMC via IMM, just as we did for the ideal case in Section F.4.1.

Simplified Hamiltonian and full Hamiltonian in constrained space

We recall that the simplified Hamiltonian $\overline{H}(x, v)$ is the sum of two parts defined by

$$\begin{aligned}
\overline{H}_1(x, v) &= f(x) + \frac{1}{2} \log \mathrm{pdet} M(x) \\
&= f(x) + \frac{1}{2} \left(\log\det g(x) + \log\det A g(x)^{-1} A^\top - \log\det AA^\top \right), \\
\overline{H}_2(x, v) &= \frac{1}{2} v^\top W(x) v.
\end{aligned}$$

The full Hamiltonian $H(x, v)$ with the Lagrangian term $c(x)^\top \lambda(x, v)$ can be also written as the sum of two parts defined by

$$\begin{aligned}
H_1(x, v) &= f(x) + \frac{1}{2} \left(\log\det g(x) + \log\det A g(x)^{-1} A^\top - \log\det AA^\top \right) \\
&\quad - x^\top A^\top (AA^\top)^{-1} A \left(\nabla f(x) + \frac{1}{2} \mathrm{Tr}\left[W(x) Dg(x) \right] \right), \\
H_2(x, v) &= \frac{1}{2} v^\top W(x) v + \frac{1}{2} x^\top A^\top (AA^\top)^{-1} A Dg(x) \left[\frac{dx}{dt}, \frac{dx}{dt} \right],
\end{aligned}$$

and IMM with this Hamiltonian is implemented as in (F.9). We note from the proof of Lemma 32 that

$$\frac{\partial H_1}{\partial x}(x,v) = U_k U_k^\top \frac{\partial \overline{H}_1}{\partial x}(x,v),$$
$$\frac{\partial H_2}{\partial x}(x,v) = U_k U_k^\top \frac{\partial \overline{H}_2}{\partial x}(x,v),$$
$$\frac{\partial H_2}{\partial v}(x,v) = U_k U_k^\top \frac{\partial \overline{H}_2}{\partial v}(x,v).$$

Lemma 33. *For step size h, let $(\overline{x},\overline{v})$ and (x,v) be the outputs of IMM with the simplified Hamiltonian and with the full Hamiltonian starting from (x_0,v_0), respectively. Then $\overline{x}=x$, and $\frac{e^{-\overline{H}(\overline{x},\overline{v})}}{e^{-\overline{H}(x_0,v_0)}} = \frac{e^{-H(x,v)}}{e^{-H(x_0,v_0)}}$.*

Proof. We use $x_{\frac{1}{3}}(=x_0), x_{\frac{2}{3}}(=x)$ and $v_{\frac{1}{3}}, v_{\frac{2}{3}}, v$ to denote the points obtained during one step of IMM with the full Hamiltonian. We similarly define $\overline{x}_i$ and $\overline{v}_i$ for $i=\frac{1}{3},\frac{2}{3}$. As (x_0,v_0) is a starting point, $U_kU_k^\top x_0 = x_0$ and $U_kU_k^\top v_0 = v_0$. Due to $\mathrm{Null}(W(x)) = \mathrm{row}(A)$, we have that $W(z)U_kU_k^\top w = W(z)w$. By comparing the first step of IMM for each Hamiltonian,

$$U_kU_k^\top \overline{v}_{\frac{1}{3}} = v_0 - \frac{h}{2}U_kU_k^\top \frac{\partial \overline{H}_1}{\partial x}(x_0,v_0) = v_0 - \frac{h}{2}\frac{\partial H_1}{\partial x}(x_0,v_0) = v_{\frac{1}{3}},$$

and thus $U_kU_k^\top \overline{v}_{\frac{1}{3}} = v_{\frac{1}{3}}$.

From the second step of IMM, $\overline{x}_{\frac{2}{3}}$ is already in the null space of A. For $x_{\frac{1}{3}} = \overline{x}_{\frac{1}{3}} = x_0, \overline{x}_{\mathrm{mid}} = (\overline{x}_{\frac{1}{3}} + \overline{x}_{\frac{2}{3}})/2$ and $\overline{v}_{\mathrm{mid}} = (\overline{v}_{\frac{1}{3}} + \overline{v}_{\frac{2}{3}})/2$, the second step of IMM with the simplified Hamiltonian is

$$\begin{aligned}
\overline{x}_{\frac{2}{3}} &= x_{\frac{1}{3}} + hU_kU_k^\top \frac{\partial \overline{H}_2}{\partial v}(\overline{x}_{\mathrm{mid}},\overline{v}_{\mathrm{mid}}) = x_{\frac{1}{3}} + hU_kU_k^\top W(\overline{x}_{\mathrm{mid}})\overline{v}_{\mathrm{mid}} \\
&= x_{\frac{1}{3}} + hU_kU_k^\top W(\overline{x}_{\mathrm{mid}})U_kU_k^\top \overline{v}_{\mathrm{mid}} = x_{\frac{1}{3}} + hU_kU_k^\top \frac{\partial \overline{H}_2}{\partial v}(\overline{x}_{\mathrm{mid}}, U_kU_k^\top \overline{v}_{\mathrm{mid}}) \\
&= x_{\frac{1}{3}} + h\frac{\partial H_2}{\partial v}(\overline{x}_{\mathrm{mid}}, U_kU_k^\top \overline{v}_{\mathrm{mid}}), \\
U_kU_k^\top \overline{v}_{\frac{2}{3}} &= v_{\frac{1}{3}} + hU_kU_k^\top \frac{\partial \overline{H}_2}{\partial x}(\overline{x}_{\mathrm{mid}},\overline{v}_{\mathrm{mid}}) \\
&= v_{\frac{1}{3}} + hU_kU_k^\top \left(-\frac{1}{2}Dg(\overline{x}_{\mathrm{mid}})\left[W(\overline{x}_{\mathrm{mid}})\overline{v}_{\mathrm{mid}}, W(\overline{x}_{\mathrm{mid}})\overline{v}_{\mathrm{mid}}\right]\right) \\
&= v_{\frac{1}{3}} + hU_kU_k^\top \left(-\frac{1}{2}Dg(\overline{x}_{\mathrm{mid}})\left[W(\overline{x}_{\mathrm{mid}})U_kU_k^\top \overline{v}_{\mathrm{mid}}, W(\overline{x}_{\mathrm{mid}})U_kU_k^\top \overline{v}_{\mathrm{mid}}\right]\right) \\
&= v_{\frac{1}{3}} + hU_kU_k^\top \frac{\partial \overline{H}_2}{\partial x}(\overline{x}_{\mathrm{mid}}, U_kU_k^\top \overline{v}_{\mathrm{mid}}) \\
&= v_{\frac{1}{3}} + h\frac{\partial H_2}{\partial x}(\overline{x}_{\mathrm{mid}}, U_kU_k^\top \overline{v}_{\mathrm{mid}}).
\end{aligned}$$

Note that $U_k U_k^\top \overline{v}_{\text{mid}} = (U_k U_k^\top \overline{v}_{\frac{2}{3}} + v_{\frac{1}{3}})/2$ and $\overline{x}_{\text{mid}} = (\overline{x}_{\frac{2}{3}} + x_{\frac{1}{3}})/2$. Since the solution of this second step is characterized as a unique fixed-point, it follows that $(\overline{x}_{\frac{2}{3}}, U_k U_k^\top \overline{v}_{\frac{2}{3}}) = (x_{\frac{2}{3}}, v_{\frac{2}{3}})$ and so $\overline{x} = x$. In the same way we analyzed $\overline{v}_{\frac{2}{3}}$, we can obtain that $U_k U_k^\top \overline{v} = v$.

We now compare the acceptance probabilities. We clearly have $\overline{H}(x_0, v_0) = H(x_0, v_0)$ due to $c(x_0) = 0$ and have $\overline{H}_1(\overline{x}, \overline{v}) = H_1(x, v)$ due to $\overline{x} = x$. For $\overline{H}_2$,

$$\overline{v}^\top W(\overline{x})\overline{v} = \overline{v}^\top U_k U_k^\top W(x) U_k U_k^\top \overline{v} = v^\top W(x) v,$$

and so $\overline{H}_2(\overline{x}, \overline{v}) = H_2(x, v)$. □

CRHMC and RHMC discretized by IMM

In this section, we show that there is a correspondence between the dynamics of CRHMC discretized by IMM and that of RHMC discretized by IMM.

Lemma 34. *The discretized CRHMC via IMM in $\mathbb{R}^n$ and the discretized RHMC via IMM in $\mathbb{R}^k$ are equivalent. That is, the output (x_1, v_1) given by the discretized CRHMC starting from (x, v) is the same with $(U_k y_1, U_k u_1)$, where (y_1, u_1) is the output of the discretized RHMC starting from (y, u) satisfying $(x, v) = (U_k y, U_k u)$. Moreover, the acceptance probabilities are the same due to*

$$\frac{e^{-H^c(x_1, v_1)}}{e^{-H^c(x, v)}} = \frac{e^{-H^r(y_1, u_1)}}{e^{-H^r(y, u)}},$$

where $H^c(x, v)$ and $H^r(y, u)$ are the Hamiltonians of CRHMC and RHMC respectively.

Proof. We first recall that $H^c(x, v)$ can be rewritten as the sum of two parts defined by

$$\begin{aligned} H_1^c(x, v) &= f(x) + \frac{1}{2}\log \text{pdet} M(x) \\ &\quad - x^\top A^\top (AA^\top)^{-1} A \left(\nabla f(x) + \frac{1}{2}\text{Tr}\left[W(x) Dg(x)\right] \right), \\ H_2^c(x, v) &= \frac{1}{2} v^\top W(x) v + \frac{1}{2} x^\top A^\top (AA^\top)^{-1} A Dg(x) \left[\frac{dx}{dt}, \frac{dx}{dt}\right]. \end{aligned}$$

Similarly for H^r, we can represent it by the sum of two parts defined by

$$\begin{aligned} H_1^r(y, u) &= f(U_k y) + \frac{1}{2}\log\det \overline{g}(y), \\ H_2^r(y, u) &= \frac{1}{2} u^\top \overline{g}(y)^{-1} u. \end{aligned}$$

For the first claim, we need to show that each step of IMM for RHMC and CRHMC is equivalent, thus it suffices to check that for any $(y, u) \in \mathbb{R}^k \times \mathbb{R}^k$

$$\frac{\partial H_1^c(U_k y, U_k u)}{\partial x} = U_k \frac{\partial H_1^r(y, u)}{\partial y},$$

$$\frac{\partial H_2^c(U_k y, U_k u)}{\partial x} = U_k \frac{\partial H_2^r(y,u)}{\partial y},$$
$$\frac{\partial H_2^c(U_k y, U_k u)}{\partial v} = U_k \frac{\partial H_2^r(y,u)}{\partial u}.$$

These computations were already checked in the proof of Lemma 32.

For the second claim, we note that the Lagrangian term vanishes due to $c(x) = c(U_k y) = 0$. Then the second claim follows from

$$\begin{aligned}\log\det \overline{g}(y') &= \log\det U_k^\top M(U_k y')U_k \quad \text{(Lemma 30)}\\ &= \log \mathrm{pdet} M(U_k y'),\\ u'^\top \overline{g}(y')^{-1} u' &= u'^\top U_k^\top W(U_k y')U_k u'. \quad \text{(Lemma 30)}\end{aligned}$$

□

The previous two lemmas imply that the dynamics of the discretized CRHMC via IMM on the constrained space is equivalent to that of the discretized RHMC via IMM on the corresponding k-dimensional polytope. Therefore, Theorem 29 follows from Corollary 4 in [KLSV22c].

Corollary 24. *Let π be a target distribution on a polytope with m constraints in $\mathbb{R}^n$ such that $\frac{d\pi}{dx} \sim e^{-\alpha^\top x}$ for $\alpha \in \mathbb{R}^n$. Let $\mathcal{M}$ be the Hessian manifold of the polytope induced by the logarithmic barrier of the polytope. Let $\Lambda = \sup_{S\subset\mathcal{M}} \frac{\pi_0(S)}{\pi(S)}$ be the warmness of the initial distribution π_0. Let π_T be the distribution obtained after T iterations of RHMC discretized by IMM on $\mathcal{M}$. For any $\varepsilon > 0$ and step size $h = O\left(\frac{1}{n^{3/2}\log\frac{\Lambda}{\varepsilon}}\right)$, there exists $T = O\left(mn^3 \log^3 \frac{\Lambda}{\varepsilon}\right)$ such that $\|\pi_T - \pi\|_{TV} \le \varepsilon$.*

F.5 Missing Notations and Definitions

F.5.1 Notations

- We use $\mathcal{N}(\mu, \Sigma)$ to denote Gaussian distribution with mean μ and covariance Σ.
- We use $\mathrm{Null}(A)$ and $\mathrm{Range}(A)$ to denote the null space and image space of a matrix or linear operator A.
- We use $\nabla^2 f \in \mathbb{R}^{n\times n}$ to denote the Hessian of a function $f : \mathbb{R}^n \to \mathbb{R}$.
- We use $\|\cdot\|$ to denote ℓ_2-norm unless specified otherwise, and define $\|x\|_A := \sqrt{x^\top A x}$ for a vector $x \in \mathbb{R}^n$ and a matrix $A \in \mathbb{R}^{n\times n}$.

- We use ∂K to denote the boundary of the set K.

- For a matrix $g(x)$ with $x \in \mathbb{R}^n$, we use $Dg(x)$ to denote the derivative of $g(x)$ with respect to x. This can be thought of as the $n \times n \times n$ tensor such that $(Dg(x))(i,j,k) = \frac{\partial (g(x))_{ij}}{\partial x_k}$. In other words, $(Dg(x))(\cdot,\cdot,k)$ is the matrix, each of entries is the derivative of $g(x)$ with respect to x_k. In addition, for a vector $v \in \mathbb{R}^n$, $Dg(x)[v,v]$ is a vector in $\mathbb{R}^n$ such that $(Dg(x)[v,v])_i = v^\top Dg(x)(\cdot,\cdot,i)v$.

- For a matrix A of size $n \times n$, we use $A \cdot Dg(x)$ to denote a $n \times n \times n$ tensor such that $(A \cdot Dg(x))(\cdot,\cdot,i) = A \cdot (Dg(x))(\cdot,\cdot,i)$. We use $\mathrm{Tr}(A \cdot Dg(x))$ to denote a vector in $\mathbb{R}^n$ such that $(\mathrm{Tr}(A \cdot Dg(x)))_i = \mathrm{Tr}((A \cdot Dg(x))(\cdot,\cdot,i))$.

F.5.2 Definitions

Convex body. A convex body is a compact and convex set.

Isotropy. A random variable X is said to be in isotropic position if $\mathbb{E}X = 0$ and $\mathbb{E}XX^\top = I$.

Pseudo-inverse. For a matrix $A \in \mathbb{R}^{m \times n}$, it is well known that there always exists the unique pseudo-inverse matrix $A^\dagger$ that satisfies the following conditions:

1. $A^\dagger A A^\dagger = A^\dagger$.
2. $A A^\dagger A = A$.
3. $AA^\dagger$ and $A^\dagger A$ are symmetric.

It is also well known that $\mathrm{Null}(A^\dagger) = \mathrm{Null}(A^\top)$ and $\mathrm{Range}(A^\dagger) = \mathrm{Range}(A^\top)$.

Pseudo-determinant. For a square matrix A, its pseudo-determinant $\mathrm{pdet}(A)$ is defined as the product of non-zero eigenvalues of A.

Leverage score. For a matrix $A \in \mathbb{R}^{m \times n}$, the leverage score of the i^{th} row is $(A(A^\top A)^\dagger A^\top)_{ii}$ for $i \in [m]$. When A is full-rank, it is simply $(A(A^\top A)^{-1} A^\top)_{ii}$.

Log-barrier & Dikin ellipsoid. For a polytope $P = \{x \in \mathbb{R}^n : Ax \leq b\}$ where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, let us denote the i^{th} row of A by a_i and the i^{th} row of b by b_i. The log-barrier of P is defined by

$$\phi(x) = -\sum_{i=1}^{m} \log(b_i - a_i^\top x).$$

For $x \in P$, the Dikin ellipsoid at x is defined by $D(x) := \{y \in \mathbb{R}^n : (y-x)^\top \nabla^2\phi(x)(y-x) \leq 1\}$. The Dikin ellipsoid is always contained in P.

Analytic center. The analytic center x_{ac} of the polytope P is the point minimizing the log-barrier (i.e., $x_{ac} = \arg\min \phi(x)$).

Self-concordant function. A function $f : \mathbb{R}^n \to \mathbb{R}$ is self-concordant if it satisfies $\left|D^3 f(x)[h,h,h]\right| \leq 2\left(D^2 f(x)[h,h]\right)^{3/2}$ for all $h \in \mathbb{R}^n$ and $x \in \mathbb{R}^n$.

Highly self-concordant function. A barrier ϕ is called *highly self-concordant* if it satisfies for all $h \in \mathbb{R}^d$ and $x \in \mathbb{R}^d$

$$\left|D^3\phi(x)[h,h,h]\right| \leq 2\left(D^2\phi(x)[h,h]\right)^{3/2} \quad \text{and} \quad \left|D^4\phi(x)[h,h,h,h]\right| \leq 6\left(D^2\phi(x)[h,h]\right)^2.$$

Total variation. For two probability distributions P and Q on support K, the total variation distance of P and Q is

$$\|P - Q\|_{\mathrm{TV}} \stackrel{\text{def}}{=} \sup_{A\subseteq K} (P(A) - Q(A)).$$

F.5.3 Details

Inverse and Determinant of block matrix. For a square matrix $M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$ with blocks A, B, C, D of same size, if D and $A - BD^{-1}C$ are invertible, then its inverse and determinant can be computed by

$$M^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix},$$

$$\det(M) = \det(D)\det(A - BD^{-1}C).$$

Orthogonal projection. Let $S = \{x \in \mathbb{R}^n : Ax = b\}$ for $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, and x_0 be a point in S. Thus $S - x_0$ is the null space of A, due to $A(x - x_0) = 0$. The orthogonal projection P to this null space is

$$P = I - A^\top (AA^\top)^{-1} A.$$

Note that the range of P always lies in the null space because

$$A(Pv) = A(I - A^\top (AA^\top)^{-1} A)v = Av - AA^\top (AA^\top)^{-1} Av = Av - Av = 0.$$

$I - P$ is also an orthogonal projection matrix, and eigenvalues of orthogonal projection matrices are either 0 or 1.

Matrix calculus. Let $U(x)$ be a $n \times n$ matrix with a parameter $x \in \mathbb{R}^n$.

$$\frac{\partial U^{-1}(x)}{\partial x_i} = -U(x) \frac{\partial U(x)}{\partial x_i} U(x).$$

Hence using the notation Dg, we can write in a more compact way as

$$DU^{-1}(x) = -U(x) DU(x) U(x).$$

For $\log\det$,

$$\frac{\partial \log\det U(x)}{\partial x_i} = \operatorname{Tr}\left(U^{-1}(x) \frac{\partial U(x)}{\partial x_i} \right).$$

In other words,

$$D(\log\det U(x)) = \operatorname{Tr}(U^{-1}(x) DU(x)).$$

Cholesky decomposition. For a symmetric positive definite matrix A, there exists a lower triangular matrix L such that $LL^\top = A$.

Newton's method. For f convex and twice differentiable in $\mathbb{R}^n$, consider an unconstrained convex optimization $\min_x f(x)$. Given a starting point $x_0 \in \mathbb{R}^n$, the Newton's method repeats

$$x_i = x_{i-1} - (\nabla^2 f(x_{i-1}))^{-1} \nabla f(x_{i-1}) \quad \forall i \in \mathbb{N}$$

to solve the optimization problem.