Logsmooth Gradient Concentration and Tighter Runtimes for Metropolized Hamiltonian Monte Carlo

Yin Tat Lee   YINTAT@UW.EDU
University of Washington

Ruoqi Shen   SHENR3@CS.WASHINGTON.EDU
University of Washington

Kevin Tian   KJTIAN@STANFORD.EDU
Stanford University

Abstract

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 Hamiltonian Monte Carlo (HMC) algorithm for sampling from a strongly logconcave distribution Dwivedi et al. (2018). 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 Dwivedi et al. (2018); Chen et al. (2019) 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; we also give evidence that this dependence is likely to be necessary for standard Metropolized first-order methods.

1. Introduction

Sampling from a high-dimensional logconcave distribution is a fundamental computational task, central to a variety of applications in disciplines such as statistics, machine learning, operations research, and theoretical computer science Andrieu et al. (2003); Brooks et al. (2011). The important problem of designing efficient samplers has received significant recent attention, due to the widespread use of Bayesian methods in modern large-scale machine learning algorithms Barber (2012), and the fact that directly computing posterior densities in these methods is often intractable. Correspondingly, recent research efforts have focused on giving convergence guarantees for Markov Chain Monte Carlo (MCMC) methods, for sampling from a variety of desirable structured distributions arising in both theory and practice.

   The specific problem we address in this paper is determining the mixing time of the well-studied Metropolized Hamiltonian Monte Carlo (HMC) algorithm[2], when sampling from a target distribu-

---

1. We use $\tilde{O}$ to hide logarithmic factors in problem parameters.
2. Metropolized HMC also refers to a family of algorithms which takes multiple *leapfrog* steps, see Algorithm 1. 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 absence of higher-derivative bounds past smoothness).

tion whose log-density is smooth and strongly concave. Indeed, as it is the default sampler implementation in a variety of popular packages Abadi (2016); Carpenter et al. (2017), understanding Metropolized HMC is of high practical import. Moreover, the specific setting we study, where the target distribution has a density proportional to $\exp(-f)$ for function f with quadratic upper and lower bounds, is commonplace in applications arising from multivariate Gaussians, logistic regression models, and structured mixture models Dwivedi et al. (2018). This setting is also of great theoretical interest because of its connection to a well-understood setting in convex optimization Nesterov (2003), 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. Ma et al. (2018); Talwar (2019) 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 Vempala (2005).

1.1. Previous work

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. Well-studied zeroth-order methods include the ball walk Lovász and Simonovits (1993, 1990), hit-and-run Bélisle et al. (1993); Lovász (1999); Lovász and Vempala (2006) and Metropolized random walk Mengersen et al. (1996); Roberts and Tweedie (1996). While possibly preferable in some cases due to weaker assumptions on both the class of target distributions and oracle access to the density, these methods typically have a large polynomial dependence on the dimension, and do not exploit the additional benefits afforded by having a finite condition number. For distributions with additional structure, sampling algorithms may wish to exploit said structure.

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 usually involves simulating a continuous Markov process whose stationary distribution is exactly the target distribution. The Langevin dynamics (LD) Gelfand and Mitter (1991), its underdamped variant (ULD), and Hamiltonian Monte Carlo (HMC) Kramers (1940); Neal et al. (2011) are among the most well-studied continuous-time Markov processes which converge to a specified logconcave measure, and sampling algorithms based on first-order information typically model their trajectories off one of these processes. 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 Dalalyan (2017b); Cheng et al. (2018); Durmus et al. (2019b); Mou et al. (2019); Mangoubi and Smith (2017); Lee et al. (2018); Chen and Vempala (2019); Shen and Lee (2019). 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 recent works Roberts et al. (1996); Roberts and

Tweedie (1996); Pillai et al. (2012); Bou-Rabee and Hairer (2013); Xifara et al. (2014); Dwivedi et al. (2018); Chen et al. (2019). Indeed, the standard discretizations of these methods are identical, which was observed in prior work (see Appendix A); we will refer to them both as Metropolized HMC. The works which inspired this study in particular were due to Dwivedi et al. (2018); Chen et al. (2019), 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*[3] of the negative log-density f. In the $\text{poly}(\epsilon^{-1})$ runtime regime, the recent work Durmus et al. (2019a) 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, Shen and Lee (2019) gives an algorithm that depends on $\kappa^{7/6}$ for Wasserstein-2 distance, but with better dependence on all other parameters (see Table 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 Bubeck (2015); 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 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 1; for a more complete discussion, we refer the reader to the excellent discussion in Dwivedi et al. (2018); Chen et al. (2019). 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.

1.2. Our contribution

Towards improving our understanding of Question 1, we show that there is an algorithm which runs Metropolized HMC (defined in Algorithm 1) for $O(\kappa d \log^3(\kappa d/\epsilon))$ iterations[5], 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,

---

3. 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.

4. With the exception of Durmus et al. (2019a), these results are measured in the 2-Wasserstein distance.

5. The precise statement of our algorithmic guarantee can be found as Theorem 13 of our paper.

Algorithm	Mixing Time	Metric
Langevin Diffusion Dalalyan (2017b)	κ^2/ϵ^2	W_2[4]
High-Order Langevin Diffusion Mou et al. (2019)	$\kappa^{19/4}/\epsilon^{1/2} + \kappa^{13/3}/\epsilon^{2/3}$	
HMC 1 (Collocation Method) Lee et al. (2018)	$\kappa^{1.5}/\epsilon$	
HMC 2 (Collocation Method) Chen and Vempala (2019)	$\kappa^{1.5}/\epsilon$	
ULD 1 (Euler Method) Cheng et al. (2018)	$\kappa^{1.5}/\epsilon$	
ULD 2 (Euler Method) Dalalyan and Riou-Durand (2018)	$\kappa^{1.5}/\epsilon + \kappa^2$	
ULD 3 (Random Midpoint Method) Shen and Lee (2019)	$\kappa^{7/6}/\epsilon^{1/3} + \kappa/\epsilon^{2/3}$	
Unadjusted Langevin Dynamics Durmus et al. (2019a)	$\kappa d^2/\epsilon^4$	TV
Metropolized HMC & MALA Chen et al. (2019)	$\kappa^{1.5}\sqrt{d} + \kappa d$	
Metropolized HMC & MALA (This paper)	κd	

Table 1: Mixing times for algorithms in the condition number regime of logconcave sampling.

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 Durmus et al. (2019a), 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 Dwivedi et al. (2018); Chen et al. (2019), and our algorithm does not require a warm start, as it explicitly bounds warmness dependence from a known starting distribution.

In Section 5, we also give preliminary evidence that for the equivalent first-order sampling methods of Metropolized HMC and Metropolis-adjusted Langevin dynamics, $\Omega(\kappa)$ iterations are necessary even for sampling from a quadratic. In particular, we show that if the step size is not sufficiently small, the chain can only move with exponentially small probability, which we combine with a lower bound on the mixing time of continuous-time HMC in Chen and Vempala (2019) for small step sizes.

The starting point of our analysis is the mixing time analysis framework for the HMC algorithm in Dwivedi et al. (2018); Chen et al. (2019). 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.

1.2.1. 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 Lee et al. (2018); Shen and Lee (2019); Vempala and Wibisono (2019) 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 the gradient norm. Indeed, this weaker bound is the bottleneck term in the analysis of Dwivedi et al. (2018); Chen et al. (2019), 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 Ledoux (1999).

1.2.2. 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. Sinclair and Jerrum (1989); Lovász and Simonovits (1993)). In particular, Lovász and Simonovits (1993) 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 Chen et al. (2019), because of the exponential warmness $\kappa^{d/2}$ of the starting distribution used, we require extensions in the theory of *average conductance* Lovász and Kannan (1999) 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 Chen et al. (2019) obtained this improved dependence on the warmness by generalizing the analysis of Goel et al. (2006) 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 nonconvex 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 28).

Blocking conductance. To mitigate the difficulty of poor small-set conductance due to the non-convexity of Ω, we use the *blocking conductance* analysis of Kannan et al. (2006), 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 $1 - \exp\left(-\sqrt{d}\right)$, 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 Dwivedi et al. (2018); Chen et al. (2019) resulted from the statement, with probability at most $\exp(-d)$, the gradient norm is bounded by $\sqrt{L\kappa d}$. We sharpen this by Corollary 6 to $\sqrt{L}d$, trading off a κ for a d, which suffices for our tighter runtime.

Boosting to high accuracy. Finally, the blocking conductance analysis of Kannan et al. (2006) 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 Kannan et al. (2006) is able to guarantee that the *average* density $d\rho_k \stackrel{\text{def}}{=} \frac{1}{k} \sum_{0 \le i < k} d\pi_i$ converges to $d\pi^*$ at a rate roughly $1/k$, with a factor depending on the average conductance[6]. In our case, we can show that in roughly $O(\kappa d)$ iterations of Algorithm 1, 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 Aldous and Diaconis (1986); Lovász and Winkler (1995) 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 Appendix C.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.

1.3. Discussion

While we obtain an improved upper bound on the runtime of Metropolized HMC, there are many questions which remain unanswered, and we believe are exciting to explore in the space of sampling algorithms and complexity. We state two here which we believe are natural.

Question 2 *Can we obtain a matching lower bound, or an improved upper bound, on the complexity of the Metropolized HMC algorithm in terms of the dependence on the dimension d?*

Question 3 *Can we obtain improved bounds (upper or lower) for the complexity of first-order sampling algorithms in general, in the condition number regime? For example, is $\Omega(\kappa)$ iterations always necessary, implying a separation with the optimization setting?*

2. Preliminaries

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

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

---

6. We note averaging has been observed to improve sampling accuracy in a different setting Durmus et al. (2019a); we leave as an interesting open direction whether this averaging is necessary for our method.

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 $\kappa \overset{\text{def}}{=} L/\mu \geq 1$ 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

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

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

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

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

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

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

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

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

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

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.

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 $x \in \mathbb{R}^d$, we define $\mathcal{P}_x$ to be the distribution of $\tilde{x}_k$ after one step of Algorithm 1 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 1 uses the subprocedure `Leapfrog`, which enjoys well-known reversible properties implying that $d\pi^*$ is stationary for Algorithm 1; we defer a proof to Appendix A.

Algorithm 1 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
end

Algorithm 2 Leapfrog: $\texttt{Leapfrog}(\eta, x, v)$

Input: Points $x, v \in \mathbb{R}^d$, step size η.
Output: Points $\tilde{x}, \tilde{v} \in \mathbb{R}^d$.
$v' \leftarrow v - \frac{\eta}{2}\nabla f(x)$.
$\tilde{x} \leftarrow x + \eta v'$.
$\tilde{v} \leftarrow v' - \frac{\eta}{2}\nabla f(\tilde{x})$.

---

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 Brascamp and Lieb (1976).

Theorem 4 (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 4 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 4; an exposition of this technique can be found in Ledoux (1999). We now state the concentration result.

Theorem 5 (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 $c > 0$,*

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

As an immediate corollary, we obtain the following.

Corollary 6 *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 Dalalyan (2017a); Vempala and Wibisono (2019); 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,

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

The last equality 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 4 holds, we can sharpen the bound in Corollary 6 to $O(d^{-c^2})$; we provide details in Appendix B. This sharpening is desirable for reasons related to the warmness of starting distributions for sampling from π^*, as will become clear in Section 4. However, the "Hessian log-Sobolev" inequality is strictly stronger than Theorem 4, and does not hold for general strongly logconcave distributions Bobkov and Ledoux (2000). Correspondingly, the concentration arguments derivable from Poincaré inequalities appear to be weaker Ledoux (1999): we find exploring the tightness of Corollary 6 to be an interesting open question.

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 Dalalyan (2017a); Dwivedi et al. (2018)

Lemma 7 (Initial warmness) *For $d\pi^* \propto \exp(-f(x))dx$ where f is L-smooth and μ-strongly convex with minimizer x^*[7], $\pi_0 = \mathcal{N}(x^*, L^{-1}I_d)$ 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. ∎

Let $d\pi_k$ be the density of x_k after running k steps of Algorithm 1, where x_0 is drawn from $\pi_0 = \mathcal{N}(x^*, L^{-1}I_d)$. Moreover, let $d\rho_k \stackrel{\text{def}}{=} \frac{1}{k}\sum_{0 \leq i < k} d\pi_i$ be the average density over the first k

---

7. 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 Dwivedi et al. (2018), and assume their exact knowledge for simplicity in this work.

iterations. In Section 4.1, we will show how to use the blocking conductance framework of Kannan et al. (2006) to obtain a bound on the number of iterations k required to obtain a constant-accuracy distribution. We then show in Section 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 13.

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 C. The first result is a restatement of the main result of Kannan et al. (2006), 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 8 (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*

$$\frac{\pi^*(S)}{Q(S)^2} \le \phi(t), \text{ for all } S \subseteq \mathbb{R}^d \text{ with } \pi^*(S) = t, \tag{4.1}$$

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,

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

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 Chen et al. (2019), we show in the following lemma that we can bound $\frac{\pi^*(S)}{Q(S)^2}$ when $\pi^*(S)$ is in some range.

Lemma 9 *Suppose for $\Omega \subset R^d$ with $\pi^*(\Omega) = 1 - s$, and all $x, y \in \Omega$ with $\|x - y\| \le \eta$,*

$$\|\mathcal{T}_x - \mathcal{T}_y\|_{\text{TV}} \le 1 - \alpha. \tag{4.2}$$

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 28. Note that in particular the lower range of t required by Theorem 8 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

$$\Omega \stackrel{\text{def}}{=} \left\{x \in \mathbb{R}^d \mid \|\nabla f(x)\|_2 \le 5\sqrt{L}d\log\frac{\kappa}{\epsilon}\right\}. \tag{4.3}$$

In Appendix D, we show the following.

Lemma 10 *For $\eta^2 \leq \frac{1}{20Ld\log\frac{\kappa}{\epsilon}}$ and all $x, y \in \Omega$ with $\|x - y\| \leq \eta$,*

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

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 11 *Let $\epsilon \in [0, 1]$, $\beta = \kappa^{d/2}$. From any β/ϵ-warm initial distribution π_0, running Algorithm 1 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 $\|\rho_k - \pi^*\|_{\mathrm{TV}} < (2e)^{-1}$.*

Proof Note that ρ_k as defined in the theorem statement is precisely the ρ_k of Theorem 8. Moreover, for the set Ω in (4.3), the probability $x \sim \pi^*$ is not in Ω is bounded via Corollary 6 by

$$s < 3d^{-5d\log_d(\kappa/\epsilon)} < (\kappa/\epsilon)^{-4d}. \tag{4.4}$$

For $\eta = \sqrt{\frac{1}{20Ld\log\frac{\kappa}{\epsilon}}}$ and $c \stackrel{\text{def}}{=} \epsilon/(4\beta e)$, $s \leq \frac{\eta\sqrt{\mu}t}{16}$ is satisfied for all t in the range $[c, \frac{1}{2}]$. Thus, we can apply Lemma 9 and conclude that (4.1) 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 8 apply, such that

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

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

Choosing k to be a sufficiently large multiple of $\kappa d\log\frac{\kappa}{\epsilon}\log\log\frac{\beta}{\epsilon}$, the guarantee follows. ∎

4.2. High-accuracy mixing

We now state a general framework for turning guarantees such as Proposition 11 into a ϵ-accuracy mixing bound guarantee, with logarithmic overhead in the quantity ϵ. We defer a more specific statement and proof to Appendix C.4.

Lemma 12 *Suppose there is a Markov chain with transitions given by $\widetilde{\mathcal{T}}$, and some nonnegative integer T_{mix}, such that for every π which is a β/ϵ-warm distribution with respect to π^*,*

$$\left\|\widetilde{\mathcal{T}}^{T_{\mathrm{mix}}}\pi - \pi^*\right\|_{\mathrm{TV}} \leq \frac{1}{2e}. \tag{4.5}$$

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

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

11

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 (4.5) iteratively. We can now state our main claim.

Theorem 13 (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 1 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 $\|\rho - \pi^*\|_{\text{TV}} \leq \epsilon$.

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 11. 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 1, 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 (4.5) with $T_{\text{mix}} = 1$, by the guarantees of Corollary 6. Thus, by running $\log(\epsilon^{-1})$ iterations of this Markov chain, we obtain the required guarantee. ∎

5. Step-size lower bound

We make progress towards showing that standard Metropolized first-order methods (i.e. Algorithm 1, which we recall is equivalent to the Metropolis-adjusted Langevin dynamics) have mixing times with at least a linear dependence on κ. Formally, consider the d-dimensional quadratic

$$f(x) = \frac{1}{2} x^\top D x,$$

where $D = \mathbf{diag}(\lambda_1, \ldots, \lambda_d)$, and $\lambda_i = \kappa$ for $i \in [d-1]$, $\lambda_d = 1$. This choice of f is 1-strongly convex and κ-smooth. We show that running Algorithm 1 with a step size much larger than $L^{-1/2}$, from a random point from the stationary distribution, will have an exponentially small accept probability. Moreover, Chen and Vempala (2019) shows that even the continuous-time HMC algorithm with step size $\eta = O(L^{-1/2})$ requires $\Omega(\kappa)$ iterations to converge.

Theorem 14 (Theorem 4 of Chen and Vempala (2019)) *For L-smooth, μ-strongly convex f, the relaxation time of ideal HMC for sampling $\sim \exp(-f(x))$ with step size $\eta = O(1/\sqrt{L})$ is $\Omega(\kappa)$.*

We defer the proof of the following to Appendix E.

Lemma 15 *In one step of Algorithm 1, suppose the starting point $x \sim \exp(-f(x))$ is drawn from the stationary distribution. For any step size $\eta = c\kappa^{-1/2}$, for $c > 40$, and with probability at least $1 - 3d^{-25}$ over the randomness of x, v, the accept probability satisfies*

$$\exp\left(-\mathcal{H}(\tilde{x}, \tilde{v}) + \mathcal{H}(x, v)\right) \leq \exp\left(-\Omega(c^6 d)\right).$$

References

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.

David J. Aldous and Persi Diaconis. Shuffling cards and stopping times. *American Mathematical Monthly*, 93:333–348, 1986.

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.

David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012. ISBN 0521518148.

Claude JP Bélisle, H Edwin Romeijn, and Robert L Smith. Hit-and-run algorithms for generating multivariate distributions. *Mathematics of Operations Research*, 18(2):255–266, 1993.

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.

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.

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.

Nawaf Bou-Rabee and Martin Hairer. Nonasymptotic mixing of the mala algorithm. *IMA Journal of Numerical Analysis*, 33(1):80–110, 2013.

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.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

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.

Yuansi Chen, Raaz Dwivedi, Martin J. Wainwright, and Bin Yu. Fast mixing of metropolized hamiltonian monte carlo: Benefits of multi-step gradients. *CoRR*, abs/1905.12247, 2019.

Zongchen Chen and Santosh S Vempala. Optimal convergence rate of hamiltonian monte carlo for strongly logconcave distributions. *arXiv preprint arXiv:1905.02313*, 2019.

Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped langevin MCMC: A non-asymptotic analysis. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, pages 300–323, 2018.

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, 2017a.

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, 2017b.

Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic langevin diffusions. *arXiv preprint arXiv:1807.09382*, 2018.

Alain Durmus, Szymon Majewski, and Blazej Miasojedow. Analysis of langevin monte carlo via convex optimization. *J. Mach. Learn. Res.*, 20:73:1–73:46, 2019a.

Alain Durmus, Eric Moulines, et al. High-dimensional bayesian inference via the unadjusted langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019b.

Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast! In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, pages 793–797, 2018.

Saul B Gelfand and Sanjoy K Mitter. Recursive stochastic algorithms for global optimization in rˆd. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

Sharad Goel, Ravi Montenegro, and Prasad Tetali. Mixing time bounds via the spectral profile. *Electronic Journal of Probability*, 11:1–26, 2006.

Ravi Kannan, László Lovász, and Ravi Montenegro. Blocking conductance and mixing in random walks. *Combinatorics, Probability & Computing*, 15(4):541–570, 2006.

Hendrik Anthony Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

Michel Ledoux. *Concentration of measure and logarithmic Sobolev inequalities*. Seminaire de probabilities XXXIII, 1999.

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.

David Asher Levin, Yuval Peres, and Elizabeth Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

László Lovász. Hit-and-run mixes fast. *Mathematical Programming*, 86(3):443–461, 1999.

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.

László Lovász and Miklós Simonovits. The mixing rate of markov chains, an isoperimetric inequality, and computing the volume. In *Proceedings [1990] 31st annual symposium on foundations of computer science*, pages 346–354. IEEE, 1990.

László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random structures & algorithms*, 4(4):359–412, 1993.

László Lovász and Santosh Vempala. Hit-and-run from a corner. *SIAM Journal on Computing*, 35 (4):985–1005, 2006.

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.

Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I. Jordan. Sampling can be faster than optimization. *CoRR*, abs/1811.08413, 2018.

Oren Mangoubi and Aaron Smith. Rapid mixing of hamiltonian monte carlo on strongly log-concave distributions. *arXiv preprint arXiv:1708.07114*, 2017.

Kerrie L Mengersen, Richard L Tweedie, et al. Rates of convergence of the hastings and metropolis algorithms. *The annals of Statistics*, 24(1):101–121, 1996.

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.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course, volume I*. 2003.

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.

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.

Gareth O Roberts, Richard L Tweedie, et al. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

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.

Alistair Sinclair and Mark Jerrum. Approximate counting, uniform generation and rapidly mixing markov chains. *Inf. Comput.*, 82(1):93–133, 1989.

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.

Santosh Vempala. Geometric random walks: A survey. *MSRI Combinatorial and Computational Geometry*, 52:573–612, 2005.

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.

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.

Appendix A. Properties of Algorithm 1

A.1. Correctness of Algorithm 1

We state a well-known property about the reversibility of the `Leapfrog` step, which in turn allows us to prove Algorithm 1 is correct.

Lemma 16 *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 17 *$d\pi^*$ is a stationary distribution for the Markov chain defined by Algorithm 1.*

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 1 is equivalent to drawing v_k, negating it, and then running `Leapfrog`. Correctness for this equivalent algorithm follows by Lemma 16. ■

A.2. 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 Besag (1994). This equivalence was also commented on in Chen et al. (2019). 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

$$v \sim \mathcal{N}(0, I), \ \tilde{x} \leftarrow x + \eta v - \frac{\eta^2}{2}\nabla f(x), \ \text{accept with probability} \ \min\left\{1, \frac{\exp(-\mathcal{H}(\tilde{x}, \tilde{v}))}{\exp(-\mathcal{H}(x, v))}\right\}.$$

Similarly, the MALA update with step size h is

$$\tilde{x} \sim \mathcal{N}(x - h\nabla f(x), 2hI), \text{ accept with probability } \min\left\{1, \frac{\exp\left(-f(\tilde{x}) - \|x - \tilde{x} + h\nabla f(\tilde{x})\|_2^2/4h\right)}{\exp\left(-f(x) - \|\tilde{x} - x + h\nabla f(x)\|_2^2/4h\right)}\right\}.$$

It is clear that in HMC the distribution of $\tilde{x}$ is

$$\tilde{x} \sim \mathcal{N}\left(x - \frac{\eta^2}{2}\nabla f(x), \eta^2 I\right),$$

so it suffices to show for $h = \eta^2/2$,

$$\frac{\|\tilde{x} - x + h\nabla f(x)\|_2^2 - \|x - \tilde{x} + h\nabla f(\tilde{x})\|_2^2}{4h} = \frac{1}{2}\left(\|v\|_2^2 - \|\tilde{v}\|_2^2\right).$$

Indeed, the right hand side simplifies to

$$\frac{\eta}{2}\langle\nabla f(\tilde{x}) + \nabla f(x), v\rangle - \frac{\eta^2}{8}\|\nabla f(\tilde{x}) + \nabla f(x)\|_2^2,$$

and the left hand side is

$$\frac{1}{2}\langle\nabla f(\tilde{x}) + \nabla f(x), \tilde{x} - x\rangle + \frac{h}{4}\left(\|\nabla f(x)\|_2^2 - \|\nabla f(\tilde{x})\|_2^2\right)$$

$$= \frac{1}{2}\left\langle\nabla f(\tilde{x}) + \nabla f(x), \eta v - \frac{\eta^2}{2}\nabla f(x)\right\rangle + \frac{\eta^2}{8}\left(\|\nabla f(x)\|_2^2 - \|\nabla f(\tilde{x})\|_2^2\right).$$

Comparing coefficients shows the equivalence.

Appendix B. Concentration

Theorem 5 (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 $g(x) \stackrel{\text{def}}{=} \exp\left(\frac{1}{2}\lambda G(x)\right)$. 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. Dwivedi et al. (2018), 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 4, which implies (noting that the gradient of $\|\nabla f\|$ is $\nabla^2 f \frac{\nabla f}{\|\nabla f\|}$)

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

$$\leq \frac{L\lambda^2}{4}\mathbb{E}_{\pi^*}\left[\exp(\lambda G)\right].$$

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 $C = \frac{L\lambda^2}{4} < 1$, Bobkov and Ledoux (1997) 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 E. 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 $r = c\sqrt{L}\log d$,

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

$\blacksquare$

We also show that the bound in Theorem 5 may be sharpened under a Hessian log-Sobolev inequality (LSI), which we define presently.

Definition 18 (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*

$$\text{Ent}_{\pi^*}[g] \stackrel{\text{def}}{=} \left(\int g(x)\log(g(x))\,d\pi^*(x)\right) - \left(\int g(x)d\pi^*(x)\right)\log\left(\int g(x)d\pi^*(x)\right),$$

we have

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

19

In general, this is a much more restrictive condition than Theorem 4; some sufficient conditions are given in Bobkov and Ledoux (2000). We now show an improved concentration result under a Hessian LSI; the proof follows Herbst's argument, a framework developed in Ledoux (1999).

Theorem 19 (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 $c > 0$,*

$$\Pr_{\pi^*}\left[\|\nabla f(x)\| \geq \mathbb{E}_{\pi^*}\left[\|\nabla f\|\right] + c\sqrt{2L\log d}\right] \leq d^{-c^2}.$$

Proof Denote $G \stackrel{\text{def}}{=} \|\nabla f\|$, where we note $\nabla G = \frac{(\nabla^2 f)\nabla f}{\|\nabla f\|}$. Let $H(\lambda) \stackrel{\text{def}}{=} \mathbb{E}_{\pi^*}\left[\exp(\lambda G)\right]$, such that $H'(\lambda) = \mathbb{E}_{\pi^*}\left[G\exp(\lambda G)\right]$. 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

$$\lambda H'(\lambda) - H(\lambda)\log H(\lambda) = \mathbb{E}_{\pi^*}\left[g^2\log g^2\right] - \mathbb{E}_{\pi^*}\left[g^2\right]\log\mathbb{E}_{\pi^*}\left[g^2\right] \leq 2\mathbb{E}_{\pi^*}\left[\|\nabla g\|_{(\nabla^2 f)^{-1}}^2\right]. \tag{B.1}$$

By smoothness and the definition of $g = \exp\left(\frac{1}{2}\lambda G\right)$, we may bound the right hand side:

$$\mathbb{E}_{\pi^*}\left[\|\nabla g\|_{(\nabla^2 f)^{-1}}^2\right] = \frac{\lambda^2}{4}\mathbb{E}_{\pi^*}\left[\|\nabla G\|_{(\nabla^2 f)^{-1}}^2\exp(\lambda G)\right] \leq \frac{\lambda^2 L}{4}H(\lambda). \tag{B.2}$$

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 $\log(1+x) \approx x$):

$$K(0) = \lim_{\lambda \to 0}\frac{1}{\lambda}\log\mathbb{E}_{\pi^*}\left[e^{\lambda G}\right] = \frac{H(\lambda) - H(0)}{\lambda} = H'(0) = \mathbb{E}_{\pi^*}\left[G\right].$$

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

$$K(\lambda) \leq \mathbb{E}_{\pi^*}\left[G\right] + \frac{L\lambda}{2} \Rightarrow H(\lambda) = \exp(\lambda K(\lambda)) \leq \exp\left(\lambda\mathbb{E}_{\pi^*}\left[G\right] + \frac{L\lambda^2}{2}\right).$$

Finally, we have concentration:

$$\Pr_{\pi^*}\left[G \geq \mathbb{E}_{\pi^*}\left[G\right] + r\right] = \Pr_{\pi^*}\left[\exp(\lambda G) \geq \exp(\lambda\mathbb{E}_{\pi^*}\left[G\right] + \lambda r)\right] \leq \exp\left(-\lambda r + \frac{L\lambda^2}{2}\right),$$

where the last statement is by Markov. Choosing $r = c\sqrt{2L\log d}$, $\lambda = r/L$ yields the conclusion. $\blacksquare$

Appendix C. Mixing time proofs

We prove various claims from Section 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.

C.1. Blocking conductance framework

In this section, we recall the *blocking conductance* framework of Kannan et al. (2006). This section is a restatement of their results for our purposes.

C.1.1. 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

$$\psi(S) \stackrel{\text{def}}{=} \int_0^{\pi^*(S)} \Psi(x, S^c) dx. \tag{C.1}$$

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 20 *For any set $S \subseteq \mathbb{R}^d$,*

$$\psi(S) \geq \frac{1}{4} Q(S)^2.$$

Proof We claim first that for any $t \in [0, \pi^*(S)]$,

$$\psi(S) \geq (\pi^*(S) - t)\Psi(t, S^c). \tag{C.2}$$

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 (C.2), 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

$$g_k(x) \overset{\text{def}}{=} k\frac{d\rho_k}{d\pi^*}(x) = \sum_{0 \leq i < k} \frac{d\pi_i}{d\pi^*}(x).$$

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 $0 \leq s < t \leq 1$, $g_k(m_k^{-1}(s)) \leq g_k(m_k^{-1}(t))$. We define

$$q_k(x, y) = Q\left(m_k^{-1}([0, \min(x, y)]), m_k^{-1}([\max(x, y), 1])\right). \tag{C.3}$$

In other words, for $x \leq y$, 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:

$$\frac{d}{dx}q_k(x, y) = \begin{cases} \mathcal{T}_x([y, 1]) & x < y \\ -\mathcal{T}_x([0, y]) & x \geq y \end{cases}. \tag{C.4}$$

C.1.2. MAIN CLAIM

Here, we recall the main result of the blocking conductance framework in terms of mixing times.

Theorem 21 *Suppose the starting distribution π_0 is β-warm with respect to π^*. Let $h : [c, 1-c] \to \mathbb{R}_{\geq 0}$ satisfy, for some $c \in (0, \frac{1}{2})$, and some k,*

$$\int_c^{1-c} h(y)q_k(x, y)dy \geq 2x(1 - x), \ \forall x \in [c, 1 - c]. \tag{C.5}$$

Then,

$$\|\rho_k - \pi^*\|_{\text{TV}} \leq \beta c + \frac{1}{k}\int_c^{1-c} h(x)dx.$$

We call a function h which satisfies (C.5) a c-mixweight function, and show how to construct such a function in Section C.2; they will be inversely related to standard notions of conductance. We first require the following helper results.

Lemma 22 *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 (C.4) 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 23

$$\|\rho_k - \pi^*\|_{\mathrm{TV}} = \frac{1}{k} \int_0^{t_0} t dg_k(t) = \frac{1}{k} \int_{t_0}^1 (1 - t)dg_k(t).$$

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.

$$\|\rho_k - \pi^*\|_{\mathrm{TV}} = \pi^*([0, t_0]) - \rho_k([0, t_0]) = \int_0^{t_0} \left(1 - \frac{g_k(x)}{k} \right) d\pi^*(x). \tag{C.6}$$

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 $i \geq 0$ 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 24

$$\|\rho_k - \pi^*\|_{\text{TV}} \leq \min \left(\beta c + \frac{1}{k} \int_c^{t_0} t dg_k(t), \ \beta c + \frac{1}{k} \int_{t_0}^{1-c} (1-t) dg_k(t) \right).$$

Proof Recall in Lemma 23 we characterized

$$\|\rho_k - \pi^*\|_{\text{TV}} = \int_0^{t_0} \left(1 - \frac{g_k(x)}{k} \right) d\pi^*(x)$$

$$= \int_0^c \left(1 - \frac{g_k(x)}{k} \right) d\pi^*(x) + \int_c^{t_0} \left(1 - \frac{g_k(x)}{k} \right) d\pi^*(x).$$

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 [Proof of Theorem 21] First, if $t_0 \leq c$, by (C.6), the total variation distance is at most $c\beta$ as desired. A similar conclusion follows if $t_0 \geq 1 - c$ from the other characterization of total variation in Lemma 24. We now consider when $t \in (c, 1-c)$. By Lemma 22, for all $y \in [c, 1-c]$,

$$1 \geq \pi_k([0, y]) \geq \int_c^{1-c} q_k(x, y) dg_k(x).$$

Multiplying by h and integrating over the range $[c, 1-c]$,

$$\int_c^{1-c} h(x) dx \geq \int_c^{1-c} \left(\int_c^{1-c} h(y) q_k(x, y) dy \right) dg_k(x) \geq \int_c^{1-c} 2x(1-x) dg_k(x).$$

24

The second inequality recalled the requirement (C.5). By combining this with half of Lemma 24,

$$\|\rho_k - \pi^*\|_{\mathrm{TV}} \leq \beta c + \frac{1}{k} \int_c^{t_0} x\, dg_k(x) \leq \beta c + \frac{1}{2(1-t_0)k} \int_c^{t_0} 2x(1-x)\, dg_k(x)$$

$$\leq \beta c + \frac{1}{2(1-t_0)k} \int_c^{1-c} h(x)\, dx.$$

By using the other half of Lemma 24, we may similarly conclude

$$\|\rho_k - \pi^*\|_{\mathrm{TV}} \leq \beta c + \frac{1}{2t_0 k} \int_c^{1-c} h(x)\, dx.$$

The conclusion follows from combining these bounds, i.e. depending on if $t_0 \leq \frac{1}{2}$ or $t_0 \geq \frac{1}{2}$. ∎

C.2. Mixweight functions

In this section, we propose a function h satisfying (C.5), and prove its correctness. First, we describe a useful sufficient condition.

Lemma 25 *Suppose $h : [c, 1-c] \to \mathbb{R}_{\geq 0}$ has $h(1-y) = h(y)$, and*

$$\int_c^{1-c} h(y)\Psi(y, S^c)\, dy \geq 2\pi^*(S)(1 - \pi^*(S)), \quad \forall S \subseteq \mathbb{R}^d : c \leq \pi^*(S) \leq \frac{1}{2}. \qquad (\text{C.7})$$

Then, h satisfies (C.5).

Proof Note that for $c \leq x \leq \frac{1}{2}$, choosing $S = m_k^{-1}([0, x])$ in (C.7) 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:

$$h(y) \stackrel{\text{def}}{=} \begin{cases} \max_{y \leq \pi^*(S) \leq \min\{2y, \frac{1}{2}\}} \frac{4\pi^*(S)}{\psi(S)} & y \leq \frac{1}{2} \\ h(1-y) & y \geq \frac{1}{2} \end{cases}. \qquad (\text{C.8})$$

In particular, note that for all $y \leq \frac{1}{2}$, by combining with Lemma 20,

$$h(y) \leq \max_{y \leq \pi^*(S) \leq 2y} \frac{16\pi^*(S)}{Q(S)^2}. \qquad (\text{C.9})$$

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 (C.8).

25

Lemma 26 *The function h defined in* (C.8) *satisfies* (C.5).

Proof Recall that it suffices to show that h satisfies (C.7). 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 $\psi(S)$ (C.1),

$$
\begin{aligned}
2x &\le \frac{4\pi^*(S)}{2\psi(S)} \left(\int_0^x \Psi(y, S^c) dy \right) \\
&\le \frac{4\pi^*(S)}{\psi(S)} \int_{x/2}^x \Psi(y, S^c) dy \\
&\le \int_{x/2}^x h(y)\Psi(y, S^c)h(y) dy \\
&\le \int_c^{1-c} h(y)\Psi(y, S^c) dy.
\end{aligned}
$$

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 $x \in [c, 2c]$,

$$
\begin{aligned}
\int_c^{1-c} h(y)\Psi(y, S^c) dy &\ge \int_x^{3x/2} h(y)\Psi(y, S^c) dy \\
&\ge \int_{x/2}^x h(y)\Psi(y, S^c) dy,
\end{aligned}
$$

where the second line is due to monotonicity, and the rest of the proof proceeds as before. ∎

Proof [Proof of Theorem 8] This follows from combining Theorem 21 with our particular choice of mixweight function given in (C.8), whose denominator we bound via (C.9). 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. ∎

C.3. Restricted conductance via total variation bounds

For $S \subseteq \mathbb{R}^d$ and $x \in \mathbb{R}^d$, we define $d(S, x) \overset{\text{def}}{=} \min_{y \in S} \|x - y\|$; for $S_1, S_2 \subseteq \mathbb{R}^d$, $d(S_1, S_2) \overset{\text{def}}{=} \min_{x \in S_2} d(S_1, x)$. The following isoperimetric inequality was given as Lemma 12 of Chen et al. (2019).

Lemma 27 (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 $\pi^*(A_1) \le \pi^*(A_2)$,*

$$
\pi^*(A_3) \ge \frac{d(A_1, A_2)\sqrt{\mu}}{2} \pi^*(A_1) \log^{\frac{1}{2}} \left(1 + \frac{1}{\pi^*(A_1)} \right).
$$

Lemma 28 *Let π^* be any μ-strongly logconcave function, and let $\delta\sqrt{\mu} < 1$ for some $\delta > 0$. Suppose for $\Omega \subset \mathbb{R}^d$ with $\pi^*(\Omega) = 1 - s$, and all $x, y \in \Omega$ with $\|x - y\| \le \delta$,*

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

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

$$s \leq \min\left(\frac{t}{2}, \frac{\delta\sqrt{\mu}t}{16}\sqrt{\log(3)} \right), \tag{C.10}$$

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:

$$A_1 \overset{\text{def}}{=} \left\{ x \in S \cap \Omega \mid \mathcal{T}_x(S^c) < \frac{\alpha}{2} \right\}, \ A_2 \overset{\text{def}}{=} \left\{ x \in S^c \cap \Omega \mid \mathcal{T}_x(S) < \frac{\alpha}{2} \right\}, \ A_3 \overset{\text{def}}{=} (A_1 \cup A_2)^c.$$

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 27, since we argued $d(A_1, A_2) > \delta$,

$$\begin{aligned}
\pi^*(A_3) &\geq \frac{\delta\sqrt{\mu}}{2} \min(\pi^*(A_1), \pi^*(A_2)) \log^{\frac{1}{2}}\left(1 + \frac{1}{\min(\pi^*(A_1), \pi^*(A_2))}\right) \\
&\geq \frac{\delta\sqrt{\mu}}{4} \min(\pi^*(S \cap \Omega), \pi^*(S^c \cap \Omega)) \log^{\frac{1}{2}}\left(1 + \frac{1}{\min(\pi^*(S \cap \Omega), \pi^*(S^c \cap \Omega))}\right) \\
&\geq \frac{\delta\sqrt{\mu}}{4}(t-s) \log^{\frac{1}{2}}\left(1 + \frac{1}{t}\right).
\end{aligned}$$

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 (C.10) holds, we have the improved bound

$$Q(S) \geq \frac{\alpha \delta \sqrt{\mu}}{64} t \log^{\frac{1}{2}} \left(\frac{1}{t} \right).$$

∎

C.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 29 $\exists T_{\text{mix}}$ *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 Levin et al. (2009).

Fact 30 *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 31 *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 32 *Under Assumption 29, 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 31, 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 30:

$$\frac{d\tilde{\pi}_i}{d\pi^*}(x) \stackrel{\text{def}}{=} \frac{\frac{d\pi_i}{d\pi^*}(x) - \min\left(\frac{d\pi_i}{d\pi^*}(x), 1\right)}{\int \left(\frac{d\pi_i}{d\pi^*}(x) - \min\left(\frac{d\pi_i}{d\pi^*}(x), 1\right)\right) d\pi^*(x)} \leq \frac{\frac{d\pi_i}{d\pi^*}(x)}{\|\pi_i - \pi^*\|_{\mathrm{TV}}} \leq \frac{\beta}{\epsilon}.$$

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

$$\frac{d\tilde{\pi}_i^*}{d\pi^*}(x) \stackrel{\text{def}}{=} \frac{1 - \min\left(\frac{d\pi_i}{d\pi^*}(x), 1\right)}{\int \left(1 - \min\left(\frac{d\pi_i}{d\pi^*}(x), 1\right)\right) d\pi^*(x)} \leq \frac{1}{\epsilon}.$$

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 29 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. ∎

Appendix D. 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 33 *For $\eta^2 \leq \frac{1}{20 L d \log \frac{\kappa}{\epsilon}}$, the Markov chain defined in Algorithm 1 satisfies*

$$\sup_{\|x-y\| \leq \eta} \|\mathcal{P}_x - \mathcal{P}_y\|_{\mathrm{TV}} \leq \frac{5}{8} \tag{D.1}$$

and, for Ω defined in (4.3),

$$\sup_{x \in \Omega} \|\mathcal{P}_x - \mathcal{T}_x\|_{\mathrm{TV}} \leq \frac{1}{8}. \tag{D.2}$$

Proof We first show (D.1). From any point $x \in \mathbb{R}^d$, let $\tilde{x}$ be the proposed point according to Algorithm 1; 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

$$
\begin{aligned}
\|\mathcal{P}_x - \mathcal{P}_y\|_{\mathrm{TV}} &\leq \sqrt{\frac{1}{2} d_{KL}\left(\mathcal{P}_x, \mathcal{P}_y\right)} \\
&\leq \frac{\left\|\left(x - \frac{\eta^2}{2}\nabla f(x)\right) - \left(y - \frac{\eta^2}{2}\nabla f(y)\right)\right\|}{2\eta} \\
&\leq \frac{\left(1 + \frac{L\eta^2}{2}\right)\|x - y\|}{2\eta} \leq \frac{5}{8},
\end{aligned}
$$

for $\|x - y\| \leq \eta$ and $\eta^2 \leq (2L)^{-1}$. The third line used the triangle inequality and ∇f is L-Lipschitz.

Next, we show (D.2). From a point x, and for any proposed transition to $\tilde{x} \neq x$, the proposal $\mathcal{P}_x$ places at least as much mass on $\tilde{x}$ 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

$$
\begin{aligned}
\|\mathcal{P}_x - \mathcal{T}_x\|_{\mathrm{TV}} &= 1 - \mathbb{E}_{v \sim \mathcal{N}(0, I_d)}\left[\min\left\{1, \exp\left(\mathcal{H}(x, v) - \mathcal{H}(\tilde{x}, \tilde{v})\right)\right\}\right] \\
&\leq 1 - \mathbb{E}_{v \sim \mathcal{N}(0, I_d)}\left[\exp\left(\mathcal{H}(x, v) - \mathcal{H}(\tilde{x}, \tilde{v})\right)\right].
\end{aligned}
$$

Therefore, to show the desired $\|\mathcal{P}_x - \mathcal{T}_x\|_{\mathrm{TV}} \leq 1/8$, it suffices to show that

$$
\mathbb{E}_{v \sim \mathcal{N}(0, I_d)}\left[\exp\left(\mathcal{H}(x, v) - \mathcal{H}(\tilde{x}, \tilde{v})\right)\right] \geq \frac{7}{8}.
$$

By the calculation

$$
\frac{15}{16} \cdot \exp\left(-\frac{1}{16}\right) \geq \frac{7}{8},
$$

it suffices to show that with probability $15/16$ over the randomness of v, $\mathcal{H}(x, v) - \mathcal{H}(\tilde{x}, \tilde{v}) \geq -1/16$. First, by a standard tail bound on the chi-squared distribution (Lemma 1 of Laurent and Massart (2000)), we have

$$
\Pr\left[\|v\|^2 \geq d + 2\sqrt{3d} + 6\right] \leq \exp(-3) \leq \frac{1}{16}.
$$

Thus, assuming d is at least a sufficiently large constant, with probability at least $1/16$ over the randomness of v, we have $\|v\| \leq \sqrt{2d}$. Finally, the conclusion follows from the claim

$$
\mathcal{H}(\tilde{x}, \tilde{v}) - \mathcal{H}(x, v) \leq \frac{1}{16}, \ \forall x \in \Omega, \ \|v\| \leq \sqrt{2d},
$$

which we now show. Recalling $\tilde{v} = v - \frac{\eta}{2}\left(\nabla f(\tilde{x}) + \nabla f(x)\right)$ and $\tilde{x} = x + \eta v - \frac{\eta^2}{2}\nabla f(x)$,

$$
\begin{aligned}
\mathcal{H}\left(\tilde{x}, \tilde{v}\right) - \mathcal{H}\left(x, v\right) &= -\frac{1}{2}\left\|v\right\|^2 + \frac{1}{2}\left\|\tilde{v}\right\|^2 - f(x) + f(\tilde{x}) \\
&\leq -\frac{1}{2}\|v\|^2 + \frac{1}{2}\|\tilde{v}\|^2 + \frac{1}{2}\left\langle\nabla f(\tilde{x}) + \nabla f(x), \tilde{x} - x\right\rangle + \frac{L}{4}\left\|\tilde{x} - x\right\|^2 \\
&= \frac{1}{2}\left\|v - \frac{\eta}{2}(\nabla f(x) + \nabla f(\tilde{x}))\right\|^2 - \frac{1}{2}\|v\|^2 \\
&\quad + \frac{1}{2}\left\langle\nabla f(\tilde{x}) + \nabla f(x), \eta v - \frac{\eta^2}{2}\nabla f(x)\right\rangle + \frac{L}{4}\left\|\tilde{x} - x\right\|^2 \\
&= -\frac{\eta}{2}\left\langle\nabla f(x) + \nabla f(\tilde{x}), v\right\rangle + \frac{\eta^2}{8}\left\|\nabla f(x) + \nabla f(\tilde{x})\right\|^2 \\
&\quad + \frac{1}{2}\left\langle\nabla f(\tilde{x}) + \nabla f(x), \eta v - \frac{\eta^2}{2}\nabla f(x)\right\rangle + \frac{L}{4}\left\|\tilde{x} - x\right\|^2 \\
&= \frac{\eta^2}{8}\left\langle\nabla f(x) + \nabla f(\tilde{x}), \nabla f(\tilde{x}) - \nabla f(x)\right\rangle + \frac{L}{4}\left\|\tilde{x} - x\right\|^2 \\
&\leq \frac{\eta^2 L}{8}\|x - \tilde{x}\|\|\nabla f(x) + \nabla f(\tilde{x})\| + \frac{L}{4}\|x - \tilde{x}\|^2.
\end{aligned}
$$

The second inequality followed from

$$
f(\tilde{x}) - f(x) \leq \min\left(\left\langle\nabla f(\tilde{x}), \tilde{x} - x\right\rangle, \ \left\langle\nabla f(x), \tilde{x} - x\right\rangle + \frac{L}{2}\|\tilde{x} - x\|^2\right),
$$

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 $\|\nabla f(x) + \nabla f(\tilde{x})\| \leq 2\|\nabla f(x)\| + L\|x - \tilde{x}\|$,

$$
\begin{aligned}
\frac{\eta^2 L}{8}\|x - \tilde{x}\|\|\nabla f(x) + \nabla f(\tilde{x})\| &\leq \frac{\eta^2 L^2}{8}\|x - \tilde{x}\|^2 + \frac{\eta^2 L}{4}\|\nabla f(x)\|\|x - \tilde{x}\| \\
&\leq \frac{L}{4}\|x - \tilde{x}\|^2 + \frac{\eta^2 L}{4}\|\nabla f(x)\|\|x - \tilde{x}\|
\end{aligned}
$$

Here we used our choice of η. Next, since $\tilde{x} - x = \eta v - \frac{\eta^2}{2}\nabla f(x)$, using the above bounds,

$$
\begin{aligned}
\mathcal{H}(\tilde{x}, \tilde{v}) - \mathcal{H}(x, v) &\leq \frac{\eta^2 L}{8}\|x - \tilde{x}\|\|\nabla f(x) + \nabla f(\tilde{x})\| + \frac{L}{4}\|x - \tilde{x}\|^2 \\
&\leq \frac{L}{2}\|x - \tilde{x}\|^2 + \frac{\eta^2 L}{4}\|\nabla f(x)\|\|x - \tilde{x}\| \\
&\leq L\eta^2\|v\|^2 + \frac{L\eta^4}{4}\|\nabla f(x)\|^2 + \frac{L\eta^3}{4}\|\nabla f(x)\|\|v\| + \frac{L\eta^4}{8}\|\nabla f(x)\|^2 \\
&\leq \frac{9L\eta^2}{8}\|v\|^2 + \frac{L\eta^4}{2}\|\nabla f(x)\|^2 \leq \frac{1}{16}.
\end{aligned}
$$

We recalled $\|v\|^2 \leq 2d$, $\|\nabla f(x)\|^2 \leq 25Ld^2\log^2\frac{\kappa}{\epsilon}$, and the choice of η. $\blacksquare$

Finally, we note that Lemma 10 immediately follows via the triangle inequality.

31

Appendix E. Deferred proofs

Lemma 34 *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

$$V(C) \overset{\text{def}}{=} \prod_{k=1}^{\infty} \left(\frac{1}{1 - \frac{C}{4^k}} \right)^{2^k} \leq \frac{1 + \sqrt{C}}{1 - \sqrt{C}},$$

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 $C \in [0, 1]$. ∎

Here, we restate the claim that our algorithm accepts with exponentially small probability if the step size is too large, and give the proof.

Lemma 15 *In one step of Algorithm 1, suppose the starting point $x \sim \exp(-f(x))$ is drawn from the stationary distribution. For any step size $\eta = c\kappa^{-1/2}$, for $c > 40$, and with probability at least $1 - 3d^{-25}$ over the randomness of x, v, the accept probability satisfies*

$$\exp\left(-\mathcal{H}(\tilde{x}, \tilde{v}) + \mathcal{H}(x, v)\right) \leq \exp\left(-\Omega(c^6 d)\right).$$

Proof Recalling $\tilde{v} = v - \frac{\eta}{2}Dx - \frac{\eta}{2}D\tilde{x}$ and $\tilde{x} = x + \eta v - \frac{\eta^2}{2}Dx$, we compute

$$\begin{aligned}
\mathcal{H}(\tilde{x}, \tilde{v}) - \mathcal{H}(x, v) &= \frac{1}{2}\|\tilde{v}\|^2 - \frac{1}{2}\|v\|^2 + \frac{1}{2}\tilde{x}^\top D\tilde{x} - \frac{1}{2}x^\top Dx \\
&= \frac{1}{2}\left\| v - \frac{\eta}{2}Dx - \frac{\eta}{2}D\tilde{x} \right\|^2 - \frac{1}{2}\|v\|^2 + \frac{1}{2}\langle D(x + \tilde{x}), \tilde{x} - x \rangle \\
&= -\frac{\eta}{2}\langle v, D(x + \tilde{x}) \rangle + \frac{\eta^2}{8}\|D(x + \tilde{x})\|^2 + \frac{1}{2}\left\langle D(x + \tilde{x}), \eta v - \frac{\eta^2}{2}Dx \right\rangle \\
&= \frac{\eta^2}{8}\left(\tilde{x}^\top D^2\tilde{x} - x^\top D^2 x \right).
\end{aligned}$$

Let x_i be the ith coordinate of x. For $i \in [d-1]$, $\lambda_i = \kappa$, and

$$
\begin{aligned}
\tilde{x}_i^2 - x_i^2 &= 2x_i \left(\eta v_i - \frac{\eta^2}{2} \lambda_i x_i \right) + \left(\eta v_i - \frac{\eta^2}{2} \lambda_i x_i \right)^2 \\
&= 2\eta x_i v_i - \eta^2 \lambda_i x_i^2 + \eta^2 v_i^2 + \frac{\eta^4}{4} \lambda_i^2 x_i^2 - \eta^3 \lambda_i x_i v_i \\
&\geq - \left(\eta^2 v_i^2 + x_i^2 \right) - \eta^2 \lambda_i x_i^2 + \eta^2 v_i^2 + \frac{\eta^4}{4} \lambda_i^2 x_i^2 - \left(\frac{\eta^4}{8} \lambda_i^2 x_i^2 + 2\eta^2 v_i^2 \right) \\
&= \left(\frac{\eta^4}{8} \lambda_i^2 - \eta^2 \lambda_i - 1 \right) x_i^2 - 2\eta^2 v_i^2 \geq \frac{\eta^4}{16} \lambda_i^2 x_i^2 - 2\eta^2 v_i^2,
\end{aligned}
$$

where the last step follows from our assumption that $\eta^2 \kappa \geq 20$. Then, letting $\hat{x}$ be the vector containing the first $d-1$ entries of the vector x,

$$
\begin{aligned}
\mathcal{H}(\tilde{x}, \tilde{v}) - \mathcal{H}(x, v) &\geq \frac{\eta^2}{8} \left(\kappa^2 \sum_{i \in [d-1]} \left(\frac{\eta^4}{16} \kappa^2 x_i^2 - 2\eta^2 v_i^2 \right) + \left(\frac{\eta^4}{8} - \eta^2 - 1 \right) x_d^2 - 2\eta^2 v_d^2 \right) \\
&\geq \frac{\eta^6 \kappa^4}{128} \|\hat{x}\|^2 - \frac{\eta^4 \kappa^2}{4} \|v\|^2 + \frac{\eta^2}{8} \left(\frac{\eta^4}{8} - \eta^2 - 1 \right) x_d^2.
\end{aligned}
$$

$$(\text{E.1})$$

By standard tail bounds on the chi-squared distribution (Lemma 1 of Laurent and Massart (2000)), with probability at least $1 - 3d^{-25}$, all of the following hold:

$$
\kappa \|\hat{x}\|^2 \geq (d-1) - 10\sqrt{d \log d}, \ \|v\|^2 \geq d + 10\sqrt{d \log d} + 50 \log d, \ x_d^2 \leq 10 \log d.
$$

Plugging these bounds into (E.1), noting that the dominant term $\eta^6 \kappa^3$ behaves as c^6, and using that if $\eta > 10$ then the third term is nonnegative, we see that for sufficiently large d, the probability the chain can move is at most $\exp\left(-\Omega(c^6 d)\right)$. ∎