Journal of Machine Learning Research 21 (2020) 1-71 Submitted 5/20; Published 5/20

Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients

Yuansi Chen YUANSI.CHEN@BERKELEY.EDU
Department of Statistics
University of California
Berkeley, CA 94720-1776, USA

Raaz Dwivedi RAAZ.RSK@BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
University of California
Berkeley, CA 94720-1776, USA

Martin J. Wainwright WAINWRIG@BERKELEY.EDU
Bin Yu BINYU@BERKELEY.EDU
Department of Statistics
Department of Electrical Engineering and Computer Sciences
University of California
Berkeley, CA 94720-1776, USA

Editor: Mohammad Emtiyaz Khan

Abstract

Hamiltonian Monte Carlo (HMC) is a state-of-the-art Markov chain Monte Carlo sampling algorithm for drawing samples from smooth probability densities over continuous spaces. We study the variant most widely used in practice, Metropolized HMC with the Störmer-Verlet or leapfrog integrator, and make two primary contributions. First, we provide a non-asymptotic upper bound on the mixing time of the Metropolized HMC with explicit choices of step-size and number of leapfrog steps. This bound gives a precise quantification of the faster convergence of Metropolized HMC relative to simpler MCMC algorithms such as the Metropolized random walk, or Metropolized Langevin algorithm. Second, we provide a general framework for sharpening mixing time bounds of Markov chains initialized at a substantial distance from the target distribution over continuous spaces. We apply this sharpening device to the Metropolized random walk and Langevin algorithms, thereby obtaining improved mixing time bounds from a non-warm initial distribution.

1. Introduction

Markov Chain Monte Carlo (MCMC) methods date back to the seminal work of Metropolis et al. (1953), and are the method of choice for drawing samples from high-dimensional distributions. They are widely used in practice, including in Bayesian statistics for exploring posterior distributions (Carpenter et al., 2017; Smith, 2014), in simulation-based methods for reinforcement learning, and in image synthesis in computer vision, among other areas. Since their origins in the 1950s, many MCMC algorithms have been introduced, applied

and studied; we refer the reader to the handbook by Brooks et al. (2011) for a survey of known results and contemporary developments.

There are a variety of MCMC methods for sampling from target distributions with smooth densities (Robert and Casella, 1999; Roberts et al., 2004; Roberts and Stramer, 2002; Brooks et al., 2011). Among them, the method of Hamiltonian Monte Carlo (HMC) stands out among practitioners: it is the default sampler for sampling from complex distributions in many popular software packages, including Stan (Carpenter et al., 2017), Mamba (Smith, 2014), and Tensorflow (Abadi et al., 2015). We refer the reader to the papers (Neal, 2011; Hoffman and Gelman, 2014; Durmus et al., 2017) for further examples and discussion of the HMC method. There are a number of variants of HMC, but the most popular choice involves combination of the leapfrog integrator with Metropolis-Hastings correction. Throughout this paper, we reserve the terminology HMC to refer to this particular Metropolized algorithm. The idea of using Hamiltonian dynamics in simulation first appeared in Alder and Wainwright (1959). Duane et al. (1987) introduced MCMC with Hamiltonian dynamics, and referred to it as Hybrid Monte Carlo. The algorithm was further refined by Neal (1994), and later re-christened in statistics community as Hamiltonian Monte Carlo. We refer the reader to Neal (2011) for an illuminating overview of the history of HMC and discussion of contemporary work.

1.1. Past work on HMC

While HMC enjoys fast convergence in practice, a theoretical understanding of this behavior remains incomplete. Some intuitive explanations are based on its ability to maintain a constant asymptotic accept-reject rate with large step-size (Creutz, 1988). Others suggest, based on intuition from the continuous-time limit of the Hamiltonian dynamics, that HMC is able to suppress random walk behavior using momentum (Neal, 2011). However, these intuitive arguments do not provide rigorous or quantitative justification for the fast convergence of the discrete-time HMC used in practice.

More recently, general asymptotic conditions under which HMC will or will not be geometrically ergodic have been established in some recent papers (Durmus et al., 2017; Livingstone et al., 2016). Other work has yielded some insight into the mixing properties of different variants of HMC, but it has focused mainly on *unadjusted* versions of the algorithm. Mangoubi and Smith (2017); Mangoubi and Vishnoi (2018) study versions of unadjusted HMC based on Euler discretization or leapfrog integrator (but omitting the Metropolis-Hastings step), and provide explicit bounds on the mixing time as a function of dimension d, condition number κ and error tolerance $\epsilon > 0$. Lee et al. (2018) studied an extended version of HMC that involves applying an ordinary differential equation (ODE) solver; they established bounds with sublinear dimension dependence, and even polylogarithmic for certain densities (e.g., those arising in Bayesian logistic regression). The mixing time for the same algorithm is further refined in the recent work by Chen and Vempala (2019). In a similar spirit, Lee and Vempala (2018a) studied the Riemannian variant of HMC (RHMC) with an ODE solver focusing on sampling uniformly from a polytope. While their result could be extended to log-concave sampling, the practical implementation for log-concave sampling of their ODE solver is unclear, and moreover requires a regularity condition on all the derivatives of density. It should be noted that such unadjusted HMC methods behave

differently from the Metropolized version most commonly used in practice. In the absence of the Metropolis-Hastings correction, the resulting Markov chain no longer converges to the correct target distribution, but instead exhibits a persistent bias even in the limit of infinite iterations. Consequently, analysis of such sampling methods requires controlling this bias; doing so leads to mixing times that scale polynomially in $1/\epsilon$, in sharp contrast with the $\log(1/\epsilon)$ that is typical for Metropolis-Hastings corrected methods.

Most closely related to our paper is the recent work by Bou-Rabee et al. (2018), which studies the same Metropolized HMC algorithm that we analyze in this paper. These authors use coupling methods to analyze HMC for a class of distributions that are strongly log-concave outside of a compact set. In the strongly log-concave case, they prove a mixing time bound that scales at least as $d^{3/2}$ in the dimension d. It should be noted that with a "warm" initialization, this dimension dependence grows more quickly than known bounds for the MALA algorithm (Dwivedi et al., 2018; Eberle, 2014), and so does not explain the superiority of HMC in practice.

In practice, it is known that Metropolized HMC is fairly sensitive to the choice of its parameters, namely the step-size η used in the discretization scheme, and the number of leapfrog steps K. At one extreme, taking a single leapfrog step $K = 1$, the algorithm reduces to the Metropolis adjusted Langevin algorithm (MALA). More generally, if too few leapfrog steps are taken, then HMC is likely to exhibit a random walk behavior similar to MALA. At the other extreme, if K is too large, the leapfrog steps tend to wander back to a neighborhood of the initial state, which leads to wasted computation as well as slower mixing (Betancourt et al., 2014). In terms of the step size η, choosing an overly large step size makes the discretization diverge from the underlying continuous dynamics, and causes the Metropolis acceptance probability to drop, hence slowing down the algorithm. On the other hand, an overly small choice of η does not allow the algorithm to explore the state space rapidly enough. While it is difficult to characterize the necessary and sufficient conditions on K and η to ensure fast convergence, many work suggest the choice of these two parameters based on the necessary conditions such as maintaining a constant acceptance rate (Chen et al., 2001). For instance, Beskos et al. (2013) showed that in the simplified scenario of target density with independent, identically distributed components, the number of leapfrog steps should scale as $d^{1/4}$ to achieve a constant acceptance rate. Besides, instead of setting the two parameters explicitly, various automatic strategies for tuning these two parameters have been proposed (Wang et al., 2013; Hoffman and Gelman, 2014; Wu et al., 2018). Despite being introduced via heuristic arguments and with additional computational cost, these methods, such as the No-U-Turn (NUTS) sampler (Hoffman and Gelman, 2014), have shown promising empirical evidence of its effectiveness on a wide range of simple target distributions.

1.2. Past work on mixing time dependency on initialization

Many proof techniques for the convergence of continuous-state Markov chains are inspired by the large body of work on discrete-state Markov chains; for instance, see the surveys (Lovász et al., 1993; Aldous and Fill, 2002) and references therein. Historically, much work has been devoted to improving the mixing time dependency on the initial distribution. For discrete-state Markov chains, Diaconis et al. (1996) were the first to show that the logarithmic

dependency of the mixing time of a Markov chain on the warmness parameter[1] of the starting distribution can be improved to double-logarithmic. This improvement—from logarithmic to doubly logarithmic—allows for a good bound on the mixing time even when starting distribution is not available. The innovation underlying this improvement is the use of log-Sobolev inequalities in place of the usual isoperimetric inequality. Later, closely related ideas such as average conductance (Lovász and Kannan, 1999; Kannan et al., 2006), evolving sets (Morris and Peres, 2005) and spectral profile (Goel et al., 2006) were shown to be effective for reducing dependence on initial conditions for discrete space chains. Thus far, only the notion of average conductance (Lovász and Kannan, 1999; Kannan et al., 2006) has been adapted to continuous-state Markov chains so as to sharpen mixing time analysis of the Ball walk (Lovász and Simonovits, 1990).

1.3. Our contributions

This paper makes two primary contributions. First, we provide a non-asymptotic upper bound on the mixing time of the Metropolized HMC algorithm for smooth densities (see Theorem 1). This theorem applies to the form of Metropolized HMC (based on the leapfrog integrator) that is most widely used in practice. To the best of our knowledge, Theorem 1 is the first rigorous confirmation of the faster non-asymptotic convergence of the Metropolized HMC as compared to MALA and other simpler Metropolized algorithms.[2] Other related works on HMC consider either its unadjusted version (without accept-reject step) with different integrators (Mangoubi and Smith, 2017; Mangoubi and Vishnoi, 2018) or the HMC based on an ODE solver (Lee et al., 2018; Lee and Vempala, 2018a). While the dimension dependency for these algorithms is usually better than MALA, they have polynomial dependence on the inverse error tolerance $1/\epsilon$ while MALA's mixing time scales as $\log(1/\epsilon)$. Moreover, our direct analysis of the Metropolized HMC with a leapfrog integrator provides explicit choices of the hyper-parameters for the sampler, namely, the step-size and the number of leapfrog updates in each step. Our theoretical choices of the hyper-parameters could potentially provide guidelines for parameter tuning in practical HMC implementations

Our second main contribution is formalized in Lemmas 3 and 4: we develop results based on the conductance profile in order to prove quantitative convergence guarantees general continuous state space Markov chains. Doing so involves non-trivial extensions of ideas from discrete state Markov chains to those in continuous state spaces. Our results not only enable us to establish the mixing time bounds for HMC with different classes of target distributions, but also allow simultaneous improvements on mixing time bounds of several Markov chains (for general continuous-state space) when the starting distribution is far from the stationary distribution. Consequentially, we improve upon previous mixing time bounds for Metropolized Random Walk (MRW) and MALA (Dwivedi et al., 2018), when the starting distribution is not *warm* with respect to the target distribution (see Theorem 5).

While this high-level road map is clear, a number of technical challenges arise en route in particular in controlling the conductance profile of HMC. The use of multiple gradient steps

1. See equation (4) for a formal definition.
2. As noted earlier, previous results by Bou-Rabee et al. (2018) on Metropolized HMC do not establish that it mixes more rapidly than MALA.

in each iteration of HMC helps it mix faster but also complicates the analysis; in particular, a key step is to control the overlap between the transition distributions of HMC chain at two nearby points; doing so requires a delicate argument (see Lemma 6 and Section 5.3 for further details).

Table 1 provides an informal summary of our mixing time bounds of HMC and how they compare with known bounds for MALA when applied to log-concave target distributions. From the table, we see that Metropolized HMC takes fewer gradient evaluations than MALA to mix to the same accuracy for log-concave distributions. Note that our current analysis establishes logarithmic dependence on the target error ϵ for strongly-log-concave as well as for a sub-class of weakly log-concave distributions. [3]

	Strongly log-concave	**Weakly log-concave**	
Sampling algorithm	**Assumption (B)** ($\kappa \ll d$)	**Assumption (C)**	**Assumption (D)**
MALA (improved bound in Thm 5 in this paper)	$d\kappa \log \frac{1}{\epsilon}$ Dwivedi et al. (2018)	$\frac{d^2}{\epsilon^{\frac{3}{2}}} \log \frac{1}{\epsilon}$ Dwivedi et al. (2018)	$d^{\frac{3}{2}} \log \frac{1}{\epsilon}$ Mangoubi and Vishnoi (2019)
Metropolized HMC with leapfrog integrator [this paper]	$d^{\frac{11}{12}} \kappa \log \frac{1}{\epsilon}$ (Corollary 2)	$\frac{d^{\frac{11}{6}}}{\epsilon} \log \frac{1}{\epsilon}$ (Corollary 18)	$d^{\frac{4}{3}} \log \frac{1}{\epsilon}$ (Corollary 18)

Table 1: Comparisons of the number of gradient evaluations needed by MALA and Metropolized HMC with leapfrog integrator from a *warm start* to obtain an ϵ-accurate sample in TV distance from a log-concave target distribution on $\mathbb{R}^d$. The second column corresponds to strongly log-concave densities with condition number κ, and the third and fourth column correspond to weakly log-concave densities satisfying certain regularity conditions.

Organization: The remainder of the paper is organized as follows. Section 2 is devoted to background on the idea of Monte Carlo approximation, Markov chains and MCMC algorithms, and the introduction of the MRW, MALA and HMC algorithms. Section 3 contains our main results on mixing time of HMC in Section 3.2, followed by the general framework for obtaining sharper mixing time bounds in Section 3.3 and its application to MALA and MRW in Section 3.4. In Section 4, we describe some numerical experiments that we performed to explore the sharpness of our theoretical predictions in some simple scenarios. In Section 5, we prove Theorem 1 and Corollary 14, with the proofs of technical lemmas and other results deferred to the appendices. We conclude in Section 6 with a discussion of our results and future directions.

Notation: For two real-valued sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists a universal constant $c > 0$ such that $a_n \leq c b_n$. We write $a_n = \widetilde{O}(b_n)$ if $a_n \leq c_n b_n$, where c_n grows at most poly-logarithmically in n. We use $[K]$ to denote the integers from

3. For a comparison with previous results on unadjusted HMC or ODE based HMC refer to the discussion after Corollary 2 and Table 7 in Appendix D.2.

the set $\{1, 2, \ldots, K\}$. We denote the Euclidean norm on $\mathbb{R}^d$ as $\|\cdot\|_2$. We use $\mathcal{X}$ to denote the (general) state space of a Markov chain. We denote $\mathcal{B}(\mathcal{X})$ as the Borel σ-algebra of the state space $\mathcal{X}$. Throughout we use the notation c, c_1, c_2 to denote universal constants. For a function $f : \mathbb{R}^d \to \mathbb{R}$ that is three times differentiable, we represent its derivatives at $x \in \mathbb{R}^d$ by $\nabla f(x) \in \mathbb{R}^d$, $\nabla^2 f(x) \in \mathbb{R}^{d \times d}$ and $\nabla^3 f(x) \in \mathbb{R}^{d^3}$. Here

$$[\nabla f(x)]_i = \frac{\partial}{\partial x_i} f(x), \quad \left[\nabla^2 f(x)\right]_{i,j} = \frac{\partial^2}{\partial x_i \partial x_j} f(x), \quad \left[\nabla^3 f\right]_{i,j,k} = \frac{\partial^3}{\partial x_i \partial x_j \partial x_k} f(x).$$

Moreover for a square matrix A, we define its ℓ_2-operator norm $|||A|||_{\mathrm{op}} := \max_{\|v\|_2 = 1} \|Av\|_2$.

2. Background and problem set-up

In this section, we begin by introducing background on Markov chain Monte Carlo in Section 2.1, followed by definitions and terminology for Markov chains in Section 2.2. In Section 2.3, we describe several MCMC algorithms, including the Metropolized random walk (MRW), the Metropolis-adjusted Langevin algorithm (MALA), and the Metropolis-adjusted Hamiltonian Monte Carlo (HMC) algorithm. Readers familiar with the literature may skip directly to the Section 3, where we set up and state our main results.

2.1. Monte Carlo Markov chain methods

Consider a distribution Π^* equipped with a density $\pi^* : \mathcal{X} \to \mathbb{R}_+$, specified explicitly up to a normalization constant as follows

$$\pi^*(x) \propto e^{-f(x)}. \tag{1}$$

A standard computational task is to estimate the expectation of some function $g : \mathcal{X} \to \mathbb{R}$—that is, to approximate $\Pi^*(g) = \mathbb{E}_{\pi^*}[g(X)] = \int_{\mathcal{X}} g(x)\pi^*(x)dx$. In general, analytical computation of this integral is infeasible. In high dimensions, numerical integration is not feasible either, due to the well-known curse of dimensionality.

A Monte Carlo approximation to $\Pi^*(g)$ is based on access to a sampling algorithm that can generate i.i.d. random variables $Z_i \sim \pi^*$ for $i = 1, \ldots, N$. Given such samples, the random variable $\widehat{\Pi^*}(g) := \frac{1}{N}\sum_{i=1}^N g(Z_i)$ is an unbiased estimate of the quantity $\Pi^*(g)$, and has its variance proportional to $1/N$. The challenge of implementing such a method is drawing the i.i.d. samples Z_i. If π^* has a complicated form and the dimension d is large, it is difficult to generate i.i.d. samples from π^*. For example, rejection sampling (Gilks and Wild, 1992), which works well in low dimensions, fails due to the curse of dimensionality.

The Markov chain Monte Carlo (MCMC) approach is to construct a Markov chain on $\mathcal{X}$ that starts from some easy-to-simulate initial distribution μ_0, and converges to π^* as its stationary distribution. Two natural questions are: (i) methods for designing such Markov chains; and (ii) how many steps will the Markov chain take to converge close enough to the stationary distribution? Over the years, these questions have been the subject of considerable research; for instance, see the reviews by Tierney (1994); Smith and Roberts (1993); Roberts et al. (2004) and references therein. In this paper, we are particularly interested in comparing three popular Metropolis-Hastings adjusted Markov chains sampling algorithms (MRW, MALA, HMC). Our primary goal is to tackle the second question for HMC,

in particular via establishing its concrete non-asymptotic mixing time bound and thereby characterizing how HMC converges faster than MRW and MALA.

2.2. Markov chain basics

Let us now set up some basic notation and definitions on Markov chains that we use in the sequel. We consider *time-homogeneous* Markov chains defined on a measurable state space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ with a transition kernel $\Theta : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to \mathbb{R}_+$. By definition, the transition kernel satisfies the following properties:

$$\Theta(x, dy) \geq 0, \quad \text{for all } x \in \mathcal{X}, \quad \text{and} \quad \int_{y \in \mathcal{X}} \Theta(x, dy) dy = 1 \quad \text{for all } x \in \mathcal{X}.$$

The k-step transition kernel Θ^k is defined recursively as $\Theta^{k+1}(x, dy) = \int_{z \in \mathcal{X}} \Theta^k(x, dz)\Theta(z, dy)dz$.

The Markov chain is *irreducible* means that for all $x, y \in \mathcal{X}$, there is a natural number $k > 0$ such that $\Theta^k(x, dy) > 0$. We say that a Markov chain satisfies the *detailed balance condition* if

$$\pi^*(x)\Theta(x, dy)dx = \pi^*(y)\Theta(y, dx)dy \quad \text{for all } x, y \in \mathcal{X}. \tag{2}$$

Such a Markov chain is also called *reversible*. Finally, we say that a probability measure Π^* with density π^* on $\mathcal{X}$ is *stationary* (or *invariant*) for a Markov chain with the transition kernel Θ if

$$\int_{x \in \mathcal{X}} \pi^*(x)\Theta(y, dx) = \pi^*(y) \quad \text{for all } y \in \mathcal{X}.$$

Transition operator: We use $\mathcal{T}$ to denote the transition operator of the Markov chain on the space of probability measures with state space $\mathcal{X}$. In simple words, given a distribution μ_0 on the current state of the Markov chain, $\mathcal{T}(\mu_0)$ denotes the distribution of the next state of the chain. Mathematically, we have $\mathcal{T}(\mu_0)(A) = \int_{\mathcal{X}} \Theta(x, A)\mu_0(x)dx$ for any $A \in \mathcal{B}(\mathcal{X})$. In an analogous fashion, $\mathcal{T}^k$ stands for the k-step transition operator. We use $\mathcal{T}_x$ as the shorthand for $\mathcal{T}(\delta_x)$, the *transition distribution at* x; here δ_x denotes the Dirac delta distribution at $x \in \mathcal{X}$. Note that by definition $\mathcal{T}_x = \Theta(x, \cdot)$.

Distances between two distributions: In order to quantify the convergence of the Markov chain, we study the mixing time for a class of distances denoted $\mathcal{L}_\mathfrak{p}$ for $\mathfrak{p} \geq 1$. Letting Q be a distribution with density q, its $\mathcal{L}_\mathfrak{p}$-divergence with respect to the positive density ν is defined as

$$d_\mathfrak{p}(Q, \nu) = \left(\int_{\mathcal{X}} \left| \frac{q(x)}{\nu(x)} - 1 \right|^\mathfrak{p} \nu(x)dx \right)^{\frac{1}{\mathfrak{p}}}. \tag{3a}$$

Note that for $\mathfrak{p} = 2$, we get the χ^2-divergence. For $\mathfrak{p} = 1$, the distance $d_1(Q, \nu)$ represents two times the total variation distance between Q and ν. In order to make this distinction clear, we use $d_{\text{TV}}(Q, \nu)$ to denote the total variation distance.

Mixing time of a Markov chain: Consider a Markov chain with initial distribution μ_0, transition operator $\mathcal{T}$ and a target distribution Π^* with density π^*. Its $\mathcal{L}_\mathfrak{p}$ mixing time with respect to Π^* is defined as follows:

$$\tau_\mathfrak{p}(\epsilon; \mu_0) = \inf \left\{ k \in \mathbb{N} \mid d_\mathfrak{p}\left(\mathcal{T}^k(\mu_0), \Pi^*\right) \leq \epsilon \right\}. \tag{3b}$$

where $\epsilon > 0$ is an error tolerance. Since distance $d_\mathfrak{p}(Q, \Pi^*)$ increases as $\mathfrak{p}$ increases, we have

$$\tau_\mathfrak{p}(\epsilon; \mu_0) \leq \tau_{\mathfrak{p}'}(\epsilon; \mu_0) \quad \text{for any} \quad \mathfrak{p}' \geq \mathfrak{p} \geq 1. \tag{3c}$$

Warm initial distribution: We say that a Markov chain with state space $\mathcal{X}$ and stationary distribution Π^* has a β-*warm start* if its initial distribution μ_0 satisfies

$$\sup_{S \in \mathcal{B}(\mathcal{X})} \frac{\mu_0(S)}{\Pi^*(S)} \leq \beta, \tag{4}$$

where $\mathcal{B}(\mathcal{X})$ denotes the Borel σ-algebra of the state space $\mathcal{X}$. For simplicity, we say that μ_0 is a warm start if the warmness parameter β is a small constant (e.g., β does not scale with dimension d).

Lazy chain: We say that the Markov chain is ζ-*lazy* if at each iteration the chain is forced to stay at the previous iterate with probability ζ. We study $\frac{1}{2}$-lazy chains in this paper. In practice, one is not likely to use a lazy chain (since the lazy steps slow down the convergence rate by a constant factor); rather, it is a convenient assumption for theoretical analysis of the mixing rate up to constant factors.[4]

2.3. From Metropolized random walk to HMC

In this subsection, we provide a brief description of the popular algorithms used for sampling from the space $\mathcal{X} = \mathbb{R}^d$. We start with the simpler zeroth-order Metropolized random walk (MRW), followed by the single-step first-order Metropolis adjusted Langevin algorithm (MALA) and finally discuss the Hamiltonian Monte Carlo (HMC) algorithm.

2.3.1. MRW and MALA algorithms

One of the simplest Markov chain algorithms for sampling from a density of the form (1) defined on $\mathbb{R}^d$ is the Metropolized random walk (MRW). Given state $x_i \in \mathbb{R}^d$ at iterate i, it generates a new proposal vector $z_{i+1} \sim \mathcal{N}(x_i, 2\eta \mathbb{I}_d)$, where $\eta > 0$ is a step-size parameter.[5] It then decides to accept or reject z_{i+1} using a Metropolis-Hastings correction; see Algorithm 1 for the details. Note that the MRW algorithm uses information about the function f only via querying function values, but not the gradients.

The Metropolis-adjusted Langevin algorithm (MALA) is a natural extension of the MRW algorithm: in addition to the function value $f(\cdot)$, it also assumes access to its gradient $\nabla f(\cdot)$ at any state $x \in \mathbb{R}^d$. Given state x_i at iterate i, it observes $(f(x_i), \nabla f(x_i))$ and then

4. Any lazy (time-reversible) chain is always aperiodic and admits a unique stationary distribution. For more details, see the survey (Vempala, 2005) and references therein.
5. The factor 2 in the step-size definition is a convenient notational choice so as to facilitate comparisons with other algorithms.

generates a new proposal $z_{i+1} \sim \mathcal{N}(x_i - \eta\nabla f(x_i), 2\eta\mathbb{I}_d)$, followed by a suitable Metropolis-Hastings correction; see Algorithm 2 for the details. The MALA algorithm has an interesting connection to the Langevin diffusion, a stochastic process whose evolution is characterized by the stochastic differential equation (SDE)

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dW_t. \tag{5}$$

The MALA proposal can be understood as the Euler-Maruyama discretization of the SDE (5).

Algorithm 1: Metropolized Random Walk (MRW)

Input: Step size $\eta > 0$ and a sample x_0 from a starting distribution μ_0
Output: Sequence $x_1, x_2, \ldots$
for $i = 0, 1, \ldots$ **do**
  Proposal step: *Draw* $z_{i+1} \sim \mathcal{N}(x_i, 2\eta\mathbb{I}_d)$
  Accept-reject step:
    compute $\alpha_{i+1} \leftarrow \min\left\{1, \frac{\exp(-f(z_{i+1}))}{\exp(-f(x_i))}\right\}$
    With probability α_{i+1} *accept* the proposal: $x_{i+1} \leftarrow z_{i+1}$
    With probability $1 - \alpha_{i+1}$ *reject* the proposal: $x_{i+1} \leftarrow x_i$
end

Algorithm 2: Metropolis adjusted Langevin algorithm (MALA)

Input: Step size η and a sample x_0 from a starting distribution μ_0
Output: Sequence $x_1, x_2, \ldots$
for $i = 0, 1, \ldots$ **do**
  Proposal step: *Draw* $z_{i+1} \sim \mathcal{N}(x_i - \eta\nabla f(x_i), 2\eta\mathbb{I}_d)$
  Accept-reject step:
    compute $\alpha_{i+1} \leftarrow \min\left\{1, \frac{\exp\left(-f(z_{i+1}) - \|x_i - z_{i+1} + \eta\nabla f(z_{i+1})\|_2^2/4\eta\right)}{\exp\left(-f(x_i) - \|z_{i+1} - x_i + \eta\nabla f(x_i)\|_2^2/4\eta\right)}\right\}$
    With probability α_{i+1} *accept* the proposal: $x_{i+1} \leftarrow z_{i+1}$
    With probability $1 - \alpha_{i+1}$ *reject* the proposal: $x_{i+1} \leftarrow x_i$
end

2.3.2. HMC sampling

The HMC sampling algorithm from the physics literature was introduced to the statistics literature by Neal; see his survey (Neal, 2011) for the historical background. The method is inspired by Hamiltonian dynamics, which describe the evolution of a state vector $q(t) \in \mathbb{R}^d$ and its momentum $p(t) \in \mathbb{R}^d$ over time t based on a Hamiltonian function $\mathcal{H} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ via Hamilton's equations:

$$\frac{dq}{dt}(t) = \frac{\partial\mathcal{H}}{\partial p}(p(t), q(t)), \quad \text{and} \quad \frac{dp}{dt}(t) = -\frac{\partial\mathcal{H}}{\partial q}(p(t), q(t)). \tag{6}$$

A straightforward calculation using the chain rule shows that the Hamiltonian remains invariant under these dynamics—that is, $\mathcal{H}(p(t), q(t)) = C$ for all $t \in \mathbb{R}$. A typical choice

of the Hamiltonian $\mathcal{H} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is given by

$$\mathcal{H}(p, q) = f(q) + \frac{1}{2} \|p\|_2^2 . \tag{7}$$

The ideal HMC algorithm for sampling is based on the continuous Hamiltonian dynamics; as such, it is not implementable in practice, but instead a useful algorithm for understanding. For a given time $T > 0$ and vectors $u, v \in \mathbb{R}^d$, let $q_T(u, v)$ denote the q-solution to Hamilton's equations at time T and with initial conditions $(p(0), q(0)) = (u, v)$. At iteration k, given the current iterate X_k, the ideal HMC algorithm generates the next iterate X_{k+1} via the update rule $X_{k+1} = q_T(p_k, X_k)$ where $p_k \sim N(0, \mathbb{I}_d)$ is a standard normal random vector, independent of X_k and all past iterates. It can be shown that with an appropriately chosen T, the ideal HMC algorithm converges to the stationary distribution π^* without a Metropolis-Hastings adjustment (see Neal (2011); Mangoubi and Vishnoi (2018) for the existence of such solution and its convergence).

However, in practice, it is impossible to compute an exact solution to Hamilton's equations. Rather, one must approximate the solution $q_T(p_k, X_k)$ via some discrete process. There are many ways to discretize Hamilton's equations other than the simple Euler discretization; see Neal (2011) for a discussion. In particular, using the leapfrog or Störmer-Verlet method for integrating Hamilton's equations leads to the Hamiltonian Monte Carlo (HMC) algorithm. It simulates the Hamiltonian dynamics for K steps via the leapfrog integrator. At each iteration, given previous state q_0 and fresh $p_0 \sim \mathcal{N}(0, \mathbb{I}_d)$, it runs the following updates for K times, for $0 \le k \le K - 1$,

$$p_{k+\frac{1}{2}} = p_k - \frac{\eta}{2} \nabla f(q_k) \tag{8a}$$

$$q_{k+1} = q_k + \eta p_{k+\frac{1}{2}} \tag{8b}$$

$$p_{k+1} = p_{k+\frac{1}{2}} - \frac{\eta}{2} \nabla f(q_{k+1}). \tag{8c}$$

Since discretizing the dynamics generates discretization error at each iteration, it is followed by a Metropolis-Hastings adjustment where the proposal (p_K, q_K) is accepted with probability

$$\min \left\{ 1, \frac{\exp\left(-\mathcal{H}(p_K, q_K)\right)}{\exp\left(-\mathcal{H}(p_0, q_0)\right)} \right\}. \tag{9}$$

See Algorithm 3 for a detailed description of the HMC algorithm with leapfrog integrator.

Remark: The HMC with leapfrog integrator can also be seen as a multi-step version of a simpler Langevin algorithm. Indeed, running the HMC algorithm with $K = 1$ is equivalent to the MALA algorithm after a re-parametrization of the step-size η. In practice, one also uses the HMC algorithm with a modified Hamiltonian, in which the quadratic term $\|p\|_2^2$ is replaced by a more general quadratic form $p^T M p$. Here M is a symmetric positive definite matrix to be chosen by the user; see Appendix D.1.1 for further discussion of this choice. In the main text, we restrict our analysis to the case $M = I$.

Algorithm 3: Metropolized HMC with leapfrog integrator

Input: Step size η, number of internal leapfrog updates K, and a sample x_0 from a starting distribution μ_0
Output: Sequence $x_1, x_2, \ldots$

for $i = 0, 1, \ldots$ **do**
  Proposal step:
  $q_0 \leftarrow x_i$
  Draw $p_0 \sim \mathcal{N}(0, \mathbb{I}_d)$
  for $k = 1, \ldots, K$ **do**
    $(p_k, q_k) \leftarrow \text{Leapfrog}(p_{k-1}, q_{k-1}, \eta)$
  end
  % q_K is now the new proposed state
  Accept-reject step:
    compute $\alpha_{i+1} \leftarrow \min\left\{1, \frac{\exp\left(-\mathcal{H}(p_K, q_K)\right)}{\exp\left(-\mathcal{H}(p_0, q_0)\right)}\right\}$
    With probability α_{i+1} *accept* the proposal: $x_{i+1} \leftarrow q_K$
    With probability $1 - \alpha_{i+1}$ *reject* the proposal: $x_{i+1} \leftarrow x_i$
end
Program `Leapfrog`$(p,\ q,\ \eta)$**:**
  $\tilde{p} \leftarrow p - \frac{\eta}{2}\nabla f(q)$
  $\tilde{q} \leftarrow q + \eta\tilde{p}$
  $\tilde{p} \leftarrow \tilde{p} - \frac{\eta}{2}\nabla f(\tilde{q})$
return $(\tilde{p}, \tilde{q})$

3. Main results

We now turn to the statement of our main results. We remind the readers that HMC refers to Metropolized HMC with leapfrog integrator, unless otherwise specified. We begin in Section 3.2 with our results for HMC: first, we derive the mixing time bounds for general target distributions in Theorem 1 and then apply that result to obtain concrete guarantees for HMC with strongly log-concave target distributions. We defer the discussion of weakly log-concave target distributions and perturbations of log-concave distributions to Appendix C.

In Section 3.3, we discuss the underlying results that are used to derive sharper mixing time bounds using conductance profile (see (Lemmas 3 and 4). In addition to being central to the proof of Theorem 1 in Section 5, these lemmas also allow us to sharpen mixing time guarantees for MALA and MRW (without much work). We state these improvements in Section 3.4.

3.1. Assumptions on the target distribution

In this section, we introduce some regularity notions and state the assumptions on the target distribution that our results in the next section rely on.

Regularity conditions: A function f is called:

$$L\text{-smooth}: \quad f(y) - f(x) - \nabla f(x)^\top (y-x) \le \frac{L}{2}\|x-y\|_2^2 \tag{10a}$$

$$m\text{-strongly convex}: \quad f(y) - f(x) - \nabla f(x)^\top (y-x) \ge \frac{m}{2}\|x-y\|_2^2 \tag{10b}$$

$$L_{\mathrm{H}}\text{-Hessian Lipschitz}: \quad \left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_{\mathrm{op}} \le L_{\mathrm{H}}\|x-y\|_2, \tag{10c}$$

where in all cases, the inequalities hold for all $x, y \in \mathbb{R}^d$.

A distribution Π with support $\mathcal{X} \subset \mathbb{R}^d$ is said to satisfy the *isoperimetric inequality* ($\mathfrak{a} = 0$) or the *log-isoperimetric inequality* ($\mathfrak{a} = \frac{1}{2}$) with constant $\psi_{\mathfrak{a}}$ if given any partition S_1, S_2, S_3 of $\mathcal{X}$, we have

$$\Pi(S_3) \ge \frac{1}{2\psi_{\mathfrak{a}}} \cdot d(S_1, S_2) \cdot \min\{\Pi(S_1), \Pi(S_2)\} \cdot \log^{\mathfrak{a}}\left(1 + \frac{1}{\min\{\Pi(S_1), \Pi(S_2\}}\right). \tag{10d}$$

where the distance between two sets S_1, S_2 is defined as $d(S_1, S_2) = \inf_{x\in S_1, y \in S_2}\{\|x-y\|_2\}$. For a distribution Π with density π and a given set Ω, its restriction to Ω is the distribution Π_Ω with the density $\pi_\Omega(x) = \frac{\pi(x)\mathbf{1}_\Omega(x)}{\Pi(\Omega)}$.

Assumptions on the target distribution: We introduce two sets of assumptions for the target distribution:

(A) We say that the target distribution Π^* is $(L, L_{\mathrm{H}}, s, \psi_{\mathfrak{a}}, M)$-*regular* if the negative log density f is L-smooth (10a) and has L_{H}-Lipschitz Hessian (10c), and there exists a convex measurable set Ω such that the distribution Π^*_Ω is $\psi_{\mathfrak{a}}$-isoperimetric (10d), and the following conditions hold:

$$\Pi^*(\Omega) \ge 1 - s \quad \text{and} \quad \|\nabla f(x)\|_2 \le M, \quad \text{for all } x \in \Omega. \tag{10e}$$

(B) We say that the target distribution Π^* is (L, L_{H}, m)-*strongly log-concave* if the negative log density is L-smooth (10a), m-strongly convex (10b), and L_{H}-Hessian-Lipschitz (10c). Moreover, we use $x^\star$ to denote the unique mode of Π^* whenever f is strongly convex.

Assumption (B) has appeared in several past papers on Langevin algorithms (Dalalyan, 2016; Dwivedi et al., 2018; Cheng and Bartlett, 2017) and the Lipschitz-Hessian condition (10c) has been used in analyzing Langevin algorithms with inaccurate gradients (Dalalyan and Karagulyan, 2019) as well as the unadjusted HMC algorithm (Mangoubi and Vishnoi, 2018). It is worth noting Assumption (A) is strictly weaker than Assumption (B), since it allows for distributions that are not log-concave. As we show in Lemma 15, Assumption (B) implies a version of Assumption (A); see Appendix B for details.

3.2. Mixing time bounds for HMC

We start with the mixing time bound for HMC applied to any distribution Π^* satisfying Assumption (A). Let HMC-(K, η) denote the $\frac{1}{2}$-lazy Metropolized HMC algorithm with η step size and K leapfrog steps in each iteration. Let $\tau_2^{\mathrm{HMC}}(\epsilon; \mu_0)$ denote the $\mathcal{L}_2$-mixing time (3b) for this chain with the starting distribution μ_0.

Theorem 1 *Consider an $(L, L_H, s, \psi_{\mathfrak{a}}, M)$-regular target distribution (cf. Assumption (A)) and a β-warm initial distribution μ_0. Then for any fixed target accuracy $\epsilon \in (0,1)$ such that $\epsilon^2 \geq 2\beta s$, there exist choices of the parameters (K, η) such that HMC-(K,η)chain with μ_0 start satisfies*

$$\tau_2^{HMC}(\epsilon; \mu_0) \leq \begin{cases} c \cdot \max\left\{\log \beta, \frac{\psi_{\mathfrak{a}}^2}{K^2\eta^2} \log\left(\frac{\log \beta}{\epsilon}\right)\right\} & \text{if } \mathfrak{a} = \frac{1}{2} \qquad \textit{[log-isoperimetric (10d)]} \\ c \cdot \frac{\psi_{\mathfrak{a}}^2}{K^2\eta^2} \log\left(\frac{\beta}{\epsilon}\right) & \text{if } \mathfrak{a} = 0 \qquad \textit{[isoperimetric (10d)]}. \end{cases}$$

See Section 5.2 for the proof, where we also provide explicit conditions on η and K in terms of the other parameters (cf. equation (26b)).

Theorem 1 covers mixing time bounds for distributions that satisfy isoperimetric or log-isoperimetric inequality provided that: (a) both the gradient and Hessian of the negative log-density are Lipschitz; and (b) there is a convex set that contains a large mass $(1-s)$ of the distribution. The mixing time only depends on two quantities: the log-isoperimetric (or isoperimetric) constant of the target distribution and the effective step-size $K^2\eta^2$. As shown in the sequel, these conditions hold for log-concave distributions as well as certain perturbations of them. If the distribution satisfies a log-isoperimetric inequality, then the mixing time dependency on the initialization warmness parameter β is relatively weak $O(\log\log\beta)$. On the other hand, when only an isoperimetric inequality (but not log-isoperimetric) is available, the dependency is relatively larger $O(\log\beta)$. In our current analysis, we can establish the ϵ-mixing time bounds up-to an error ϵ such that $\epsilon^2 \geq 2\beta s$. If mixing time bounds up to an arbitrary accuracy are desired, then the distribution needs to satisfy (10e) for arbitrary small s. For example, as we later show in Lemma 15, arbitary small s can be imposed for strongly log-concave densities (i.e., satisfying Assumption (B)).

Let us now derive several corollaries of Theorem 1. We begin with non-asymptotic mixing time bounds for HMC-(K,η) chain for strongly-log concave target distributions. Then we briefly discuss the corollaries for weakly log-concave target and non-log-concave target distributions and defer the precise statements to Appendix C. These results also provide a basis for comparison of our results with prior work.

3.2.1. Strongly log-concave target

We now state an explicit mixing time bound of HMC for a strongly log-concave distribution. We consider an (L, L_{H}, m)-strongly log-concave distribution (assumption (B)). We use $\kappa = L/m$ to denote the condition number of the distribution. Our result makes use of the following function

$$r(s) := 1 + \max\left\{\left(\frac{\log(1/s)}{d}\right)^{1/4}, \left(\frac{\log(1/s)}{d}\right)^{1/2}\right\}, \tag{11a}$$

and involves the step-size choices

$$\eta_{\mathrm{warm}} = \sqrt{\frac{1}{cL \cdot r\left(\frac{\epsilon^2}{2\beta}\right) d^{\frac{7}{6}}}}, \quad \text{and} \quad \eta_{\mathrm{feas}} = \sqrt{\frac{1}{cL \cdot r\left(\frac{\epsilon^2}{2\kappa^d}\right)} \min\left\{\frac{1}{d\kappa^{\frac{1}{2}}}, \frac{1}{d^{\frac{2}{3}}\kappa^{\frac{5}{6}}}, \frac{1}{d^{\frac{1}{2}}\kappa^{\frac{3}{2}}}\right\}}, \tag{11b}$$

With these definitions, we have the following:

Corollary 2 *Consider an (L, L_H, m)-strongly log-concave target distribution Π^* (cf. Assumption (B)) such that $L_H^{2/3} = O(L)$, and any error tolerance $\epsilon \in (0,1)$.*

(a) Suppose that $\kappa = O(d^{\frac{2}{3}})$ and $\beta = O\left(\exp\left(d^{\frac{2}{3}}\right)\right)$. Then with any β-warm initial distribution μ_0, hyper-parameters $K = d^{\frac{1}{4}}$ and $\eta = \eta_{warm}$, the HMC-(K,η) chain satisfies

$$\tau_2^{HMC}(\epsilon; \mu_0) \leq c\, d^{\frac{2}{3}}\, \kappa\, r\left(\frac{\epsilon^2}{2\beta}\right) \log\left(\frac{\log \beta}{\epsilon}\right). \tag{12a}$$

(b) With the initial distribution $\mu_\dagger = \mathcal{N}(x^\star, \frac{1}{L}\mathbb{I}_d)$, hyper-parameters $K = \kappa^{\frac{3}{4}}$ and $\eta = \eta_{feas}$, the HMC-(K,η) chain satisfies

$$\tau_2^{HMC}(\epsilon; \mu_\dagger) \leq c\, r\left(\frac{\epsilon^2}{2\kappa^d}\right) \max\left\{d\log\kappa, \max\left[d, d^{\frac{2}{3}}\kappa^{\frac{1}{3}}, d^{\frac{1}{2}}\kappa\right] \log\left(\frac{d\log\kappa}{\epsilon}\right)\right\}. \tag{12b}$$

See Appendix B for the proof. In the same appendix, we also provide a more refined mixing time of the HMC chain for a more general choice of hyper-parameters (see Corollary 14). In fact, as shown in the proof, the assumption $L_{\mathrm{H}}^{2/3} = O(L)$ is not necessary in order to control mixing; rather, we adopted it above to simplify the statement of our bounds. A more detailed discussion on the particular choice for step size η is provided in Appendix D.

MALA vs HMC—Warm start: Corollary 2 provides mixing time bounds for two cases. The first result (12a) implies that given a warm start (with constant β) for a well-conditioned strongly log concave distribution ($\kappa \ll d$), the ϵ-$\mathcal{L}_2$-mixing time[6] of HMC scales $\widetilde{O}(d^{\frac{2}{3}}\log(1/\epsilon))$. It is interesting to compare this guarantee with known bounds for the MALA algorithm; however, in order to do so in a fair way, we need to track the total number of gradient evaluation required by the HMC-(K,η) chain to mix. (Recall that each iteration of MALA uses only a single gradient evaluation.) For HMC to achieve accuracy ϵ, the total number of gradient evaluations is given by $K \cdot \tau_2^{\mathrm{HMC}}(\epsilon;\mu_0)$, which (in this case), scales as $\widetilde{O}(d^{\frac{11}{12}}\kappa\log(1/\epsilon))$. (This rate was also summarized in Table 1.) Note that the corresponding number of gradient evaluations for MALA (Theorem 1 in Dwivedi et al. (2018)) is $\widetilde{O}(d\kappa\log(1/\epsilon))$. As a result, we conclude that for this case, the upper bound for HMC is $d^{\frac{1}{12}}$ better than the known upper bound for MALA. We summarize the rates for this case in Table 2. Note that MRW is a zeroth order algorithm and does not make use of gradient information.

MALA vs HMC—Feasible start: In the second result (12b), we cover the case when a warm start is not available. In particular, we analyze the HMC chain with the feasible initial distribution $\mu_\dagger = \mathcal{N}(x^\star, \frac{1}{L}\mathbb{I}_d)$. Here $x^\star$ denotes the unique mode of the target distribution

6. Note that $r(\epsilon^2) \leq 6$ for $\epsilon \geq \frac{2}{e^{d/2}}$ and thus we can treat r as a small constant for a large range of ϵ. Otherwise, if ϵ needs to be extremely small, the results still hold with an extra $\log^{\frac{1}{2}}\left(\frac{1}{\epsilon}\right)$ dependency.

Sampling algorithm	Mixing time	#Gradient evaluations
MRW (Dwivedi et al., 2018, Theorem 2)	$d\kappa^2 \cdot \log \frac{1}{\epsilon}$	NA
MALA (Dwivedi et al., 2018, Theorem 1)	$d\kappa \cdot \log \frac{1}{\epsilon}$	$d\kappa \cdot \log \frac{1}{\epsilon}$
HMC-(K, η) [ours, Corollary 2]	$d^{\frac{2}{3}}\kappa \cdot \log \frac{1}{\epsilon}$	$d^{\frac{11}{12}}\kappa \cdot \log \frac{1}{\epsilon}$

Table 2: Summary of the ϵ-mixing time and the corresponding number of gradient evaluations for MRW, MALA and HMC from a *warm start* with an (L, L_{H}, m)-strongly-log-concave target. These statements hold under the assumption $L_{\mathrm{H}}^{2/3} = O(L)$, $\kappa = \frac{L}{m} \ll d$, and omit logarithmic terms in dimension.

and can be easily computed using an optimization scheme like gradient descent. It is not hard to show (see Corollary 1 in Dwivedi et al. (2018)) that for an L-smooth (10a) and m strongly log-concave target distribution (10b), the distribution $\mu_\dagger$ acts as a $\kappa^{d/2}$-warm start distribution. Once again, it is of interest to determine whether HMC takes fewer gradient steps when compared to MALA to obtain an ϵ-accurate sample. We summarize the results in Table 3 (where log factors are hidden) and note that HMC with $K = \kappa^{3/4}$ is faster than MALA for as long as κ is not too large. From the last column, we find that when $\kappa \ll d^{\frac{1}{2}}$, HMC is faster than MALA by a factor of $\kappa^{\frac{1}{4}}$ in terms of number of gradient evaluations.[7]

Sampling algorithm	Mixing time	# Gradient Evaluations	
		general κ	$\kappa \ll d^{\frac{1}{2}}$
MRW [ours, Theorem 5]	$d\kappa^2$	NA	NA
MALA [ours, Theorem 5]	$\max\left\{d\kappa, d^{\frac{1}{2}}\kappa^{\frac{3}{2}}\right\}$	$\max\left\{d\kappa, d^{\frac{1}{2}}\kappa^{\frac{3}{2}}\right\}$	$d\kappa$
HMC-(K, η) [ours, Corollary 2]	$\max\left\{d, d^{\frac{2}{3}}\kappa^{\frac{1}{3}}, d^{\frac{1}{2}}\kappa\right\}$	$\max\left\{d\kappa^{\frac{3}{4}}, d^{\frac{2}{3}}\kappa^{\frac{13}{12}}, d^{\frac{1}{2}}\kappa^{\frac{7}{4}}\right\}$	$d\kappa^{\frac{3}{4}}$

Table 3: Summary of the ϵ-mixing time and the corresponding number of gradient evaluations for MRW, MALA and HMC from the *feasible start* $\mu_\dagger = \mathcal{N}(x^\star, \frac{1}{L}\mathbb{I}_d)$ for an (L, L_{H}, m)-strongly-log-concave target. Here $x^\star$ denotes the unique mode of the target distribution. These statements hold uner the assumption $L_{\mathrm{H}} = O(L^{\frac{3}{2}})$, and hide the logarithmic factors in ϵ, d and $\kappa = L/m$.

Metropolized HMC vs Unadjusted HMC: There are many recent results on the 1-Wasserstein distance mixing of unadjusted versions of HMC (for instance, see the pa-

7. It is worth noting that for the feasible start $\mu_\dagger$, the mixing time bounds for MALA and MRW in our prior work Dwivedi et al. (2018) were loose by a factor d when compared to the tighter bounds in Theorem 5 derived later in this paper.

pers Mangoubi and Vishnoi (2018); Lee et al. (2018)). For completeness, we compare our results with them in the Appendix D.2; in particular, see Table 7 for a comparative summary.) We remark that comparisons of these different results is tricky for two reasons: (a) The 1-Wasserstein distance and the total variation distance are not strictly comparable, and, (b) the unadjusted HMC results always have a polynomial dependence on the error parameter ϵ while our results for Metropolized HMC have a superior logarithmic dependence on ϵ. Nonetheless, the second difference between these chains has a deeper consequence, upon which we elaborate further in Appendix D.2. On one hand, the unadjusted chains have better mixing time in terms of scaling with d, if we fix ϵ or view it as independent of d. On the other hand, when such chains are used to estimate certain higher-order moments, the polynomial dependence on ϵ might become the bottleneck and Metropolis-adjusted chains would become the method of choice.

Ill-conditioned target distributions: In order to keep the statement of Corollary 2 simple, we stated the mixing time bounds of HMC-(K, η)-chain only for a particular choice of (K, η). In our analysis, this choice ensures that HMC is better than MALA only when condition number κ is small. For Ill-conditioned distributions, i.e., when κ is large, finer tuning of HMC-(K, η)-chain is required. In Appendices B and D (see Table 4 for the hyperparameter choices), we show that HMC is strictly better than MALA as long as $\kappa \leq d$ and as good as MALA when $\kappa \geq d$.

Beyond strongly-log-concave: The proof of Corollary 2 is based on the fact that (L, L_{H}, m)-strongly-log-concave distribution is in fact an $(L, L_{\mathrm{H}}, s, \psi_{1/2}, M_s)$ -regular distribution for any $s \in (0, 1)$. Here $\psi_{1/2} = 1/\sqrt{m}$ is fixed and the bound on the gradient $M_s = r(s)\sqrt{d/m}$ depends on the choice of s. The result is formally stated in Lemma 15 in Appendix B. Moreover, in Appendix C, we discuss the case when the target distribution is weakly log concave (under a bounded fourth moment or bounded covariance matrix assumption) or a perturbation of log-concave distribution. See Corollary 18 for specific details where we provide explicit expressions for the rates that appear in third and fourth columns of Table 1.

3.3. Mixing time bounds via conductance profile

In this section, we discuss the general results that form the basis of the analysis in this paper. A standard approach to controlling mixing times is via worst-case conductance bounds. This method was introduced by Jerrum and Sinclair (1988) for discrete space chains and then extended to the continuous space settings by Lovász and Simonovits (1993), and has been thoroughly studied. See the survey (Vempala, 2005) and the references therein for a detailed discussion of conductance based methods for continuous space Markov chains.

Somewhat more recent work on discrete state chains has introduced more refined methods, including those based on the conductance profile (Lovász and Kannan, 1999), the spectral and conductance profile (Goel et al., 2006), as well as the evolving set method (Morris and Peres, 2005). Here we extend one of the conductance profile techniques from the paper by Goel et al. (2006) from discrete state to continuous state chains, albeit with several appropriate modifications suited for the general setting.

We first introduce some background on the conductance profile. Given a Markov chain with transition probability $\Theta : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to \mathbb{R}$, its stationary *flow* $\phi : \mathcal{B}(\mathcal{X}) \to \mathbb{R}$ is defined as

$$\phi(S) = \int_{x \in S} \Theta(x, S^c)\pi^*(x)dx \quad \text{for any } S \in \mathcal{B}(\mathcal{X}). \tag{13}$$

Given a set $\Omega \subset \mathcal{X}$, the Ω-*restricted conductance profile* is given by

$$\Phi_\Omega(v) = \inf_{\Pi^*(S \cap \Omega) \in (0,v]} \frac{\phi(S)}{\Pi^*(S \cap \Omega)} \qquad \text{for any } v \in \big(0,\ \Pi^*(\Omega)/2\big]. \tag{14}$$

(The classical conductance constant Φ is a special case; it can be expressed as $\Phi = \Phi_{\mathcal{X}}(\frac{1}{2})$.) Moreover, we define the *truncated extension* $\tilde{\Phi}_\Omega$ of the function Φ_Ω to the positive real line as

$$\tilde{\Phi}_\Omega(v) = \begin{cases} \Phi_\Omega(v), & v \in \left(0, \frac{\Pi^*(\Omega)}{2}\right] \\ \Phi_\Omega(\Pi^*(\Omega)/2), & v \in \left[\frac{\Pi^*(\Omega)}{2}, \infty\right). \end{cases} \tag{15}$$

In our proofs we use the conductance profile with a suitably chosen set Ω.

Smooth chain assumption: We say that the Markov chain satisfies the *smooth chain assumption* if its transition probability function $\Theta : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to \mathbb{R}_+$ can be expressed in the form

$$\Theta(x, dy) = \theta(x, y)dy + \alpha_x \delta_x(dy) \quad \text{for all } x, y \in \mathcal{X}, \tag{16}$$

where θ is the transition kernel satisfying $\theta(x, y) \geq 0$ for all $x, y \in \mathcal{X}$. Here δ_x denotes the Dirac-delta function at x and consequently, α_x denotes the one-step probability of the chain to stay at its current state x. Note that the three algorithms discussed in this paper (MRW, MALA and HMC) all satisfy the smooth chain assumption (16). Throughout the paper, when dealing with a general Markov chain, we assume that it satisfies the smooth chain assumption.

Mixing time via conductance profile: We now state our Lemma 3 that provides a control on the mixing time of a Markov chain with continuous-state space in terms of its restricted conductance profile. We show that this control (based on conductance profile) allows us to have a better initialization dependency than the usual conductance based control (see Lovász and Simonovits (1990, 1993); Dwivedi et al. (2018)). This method for sharpening the dependence is known for discrete-state Markov chains; to the best of our knowledge, the following lemma is the first statement and proof of an analogous sharpening for continuous state space chains:

Lemma 3 *Consider a reversible, irreducible, ζ-lazy and smooth Markov chain* (16) *with stationary distribution Π^*. Then for any error tolerance ϵ, and a β-warm distribution μ_0, given a set Ω such that $\Pi^*(\Omega) \geq 1 - \frac{\epsilon^2}{2\beta^2}$, the ϵ-$\mathcal{L}_2$ mixing time of the chain is bounded as*

$$\tau_2(\epsilon; \mu_0) \leq \int_{4/\beta}^{8/\epsilon^2} \frac{4\, dv}{\zeta \cdot v\tilde{\Phi}^2_\Omega(v)}, \tag{17}$$

where $\tilde{\Phi}_\Omega$ denotes the truncated Ω-restricted conductance profile (15).

See Appendix A.1 for the proof, which is based on an appropriate generalization of the ideas used by Goel et al. (2006) for discrete state chains.

The standard conductance based analysis makes use of the worst-case conductance bound for the chain. In contrast, Lemma 3 relates the mixing time to the conductance profile, which can be seen as point-wise conductance. We use the Ω-restricted conductance profile to state our bounds, because often a Markov chain has poor conductance only in regions that have very small probability under the target distribution. Such a behavior is not disastrous as it does not really affect the mixing of the chain up to a suitable tolerance. Given the bound (17), we can derive mixing time bound for a Markov chain readily if we have a bound on the Ω-restricted conductance profile Φ_Ω for a suitable Ω. More precisely, if the Ω-restricted conductance profile Φ_Ω of the Markov chain is bounded as

$$\Phi_\Omega(v) \geq \sqrt{B \log\left(\frac{1}{v}\right)} \quad \text{for } v \in \left[\frac{4}{\beta}, \frac{1}{2}\right],$$

for some $\beta > 0$ and Ω such that $\Pi^*(\Omega) \geq 1 - \frac{\epsilon^2}{2\beta^2}$. Then with a β-warm start, Lemma 3 implies the following useful bound on the mixing time of the ζ-lazy Markov chain:

$$\tau_2(\epsilon; \mu_0) \leq \frac{32}{\zeta B} \log\left(\frac{\log \beta}{2\epsilon}\right). \tag{18}$$

We now relate our result to prior work based on conductance profile.

Prior work: For discrete state chains, a result similar to Lemma 3 was already proposed by Lovász and Kannan (Theorem 2.3 in Lovász and Kannan (1999)). Later on, Morris and Peres (2005) and Goel et al. (2006) used the notion of evolving sets and spectral profile respectively to sharpen the mixing time bounds based on average conductance for discrete-state space chains. In the context of continuous state space chains, Lovász and Kannan claimed in their original paper (Lovász and Kannan, 1999) that a similar result should hold for general state space chain as well, although we were unable to find any proof of such a general result in that or any subsequent work. Nonetheless, in a later work an average conductance based bound was used by Kannan et al. to derive faster mixing time guarantees for uniform sampling on bounded convex sets for ball walk (see Section 4.3 in Kannan et al. (2006)). Their proof technique is not easily extendable to more general distributions including the general log-concave distributions in $\mathbb{R}^d$. Instead, our proof of Lemma 3 for general state space chains proceeds by an appropriate generalization of the ideas based on the spectral profile by Goel et al. (2006) (for discrete state chains).

Lower bound on conductance profile: Given the bound (18), it suffices to derive a lower bound on the conductance profile Φ_Ω of the Markov chain with a suitable choice of the set Ω. We now state a lower bound for the restricted-conductance profile of a general state space Markov chain that comes in handy for this task. We note that a closely related logarithmic-Cheeger inequality was used for sampling from uniform distribution of a convex body (Kannan et al., 2006) and for sampling from log-concave distributions (Lee and Vempala, 2018b) without explicit constants. Since we would like to derive a non-asymptotic mixing rate, we re-derive an explicit form of their result.

Let scalars $s \in (0, 1/2]$, $\omega \in (0, 1)$ and $\Delta > 0$ be given and let $\mathcal{T}_x$ denote the one-step transition distribution of the Markov chain at point x. Suppose that that chain satisfies

$$d_{\mathrm{TV}}(\mathcal{T}_x, \mathcal{T}_y) \leq 1 - \omega \qquad \text{whenever } x, y \in \Omega \text{ and } \|x - y\|_2 \leq \Delta. \tag{19}$$

Lemma 4 *For a given target distribution Π^*, let Ω be a convex measurable set such that the distribution Π^*_Ω satisfies the isoperimetry (or log-isoperimetry) condition* (10d) *with $\mathfrak{a} = 0$ (or $\mathfrak{a} = \frac{1}{2}$ respectively). Then for any Markov chain satisfying the condition* (19)*, we have*

$$\Phi_\Omega(v) \geq \frac{\omega}{4} \cdot \min\left\{1, \frac{\Delta}{16\psi_{\mathfrak{a}}} \cdot \log^{\mathfrak{a}}\left(1 + \frac{1}{v}\right)\right\}, \quad \textit{for any } v \in \left[0, \frac{\Pi^*(\Omega)}{2}\right]. \tag{20}$$

See Appendix A.2 for the proof; the extra logarithmic term comes from the logarithmic isoperimetric inequality ($\mathfrak{a} = \frac{1}{2}$).

Faster mixing time bounds: For any target distribution satisfying a logarithmic isoperimetric inequality (including the case of a strongly log-concave distribution), Lemma 4 is a strict improvement of the conductance bounds derived in previous works (Lovász, 1999; Dwivedi et al., 2018). Given this result, suppose that we can find a convex set Ω such that $\Pi^*(\Omega) \approx 1$ and the conditions of Lemma 4 are met, then with a β-warm start μ_0, a direct application of the bound (18) along with Lemma 4 implies the following bound:

$$\tau_2(\epsilon; \mu_0) \leq O\left(\frac{1}{\omega^2 \Delta^2} \log \frac{\log \beta}{\epsilon}\right). \tag{21}$$

Results known from previous work for continuous state Markov chains scale like $\frac{\log(\beta/\epsilon)}{\omega^2\Delta^2}$; for instance, see Lemma 6 in Chen et al. (2018). In contrast, the bound (21) provides an additional logarithmic factor improvement in the factor β. Such an improvement also allows us to derive a sharper dependency on dimension d for the mixing time for sampling algorithms other than HMC as we now illustrate in the next section.

3.4. Improved warmness dependency for MALA and MRW

As discussed earlier, the bound (21) helps derive a $\frac{\log\log\beta}{\log\beta}$ factor improvement in the mixing time bound from a β-warm start in comparison to earlier conductance based results. In many settings, a suitable choice of initial distribution has a warmness parameter that scales exponentially with dimension d, e.g., $\beta = O(e^d)$. For such cases, this improvement implies a gain of $O(\frac{d}{\log d})$ in mixing time bounds. As already noted the distribution $\mu_\dagger = \mathcal{N}(x^*, \frac{1}{L}\mathbb{I}_d)$ is a feasible starting distribution[8] whose warmness scales exponentially with dimension d. We now state sharper mixing time bounds for MALA and MRW with $\mu_\dagger$ as the starting distribution. In the result, we use c_1, c_2 to denote positive universal constants.

Theorem 5 *Assume that the target distribution Π^* satisfies the conditions* (10a) *and* (10b) *(i.e., the negative log-density is L-smooth and m-strongly convex). Then given the initial*

8. See Section 3.2 of the paper by Dwivedi et al. (2018), where the authors show that computing x^* is not expensive and even approximate estimates of x^* and L are sufficient to provide a feasible starting distribution.

distribution $\mu_\dagger = \mathcal{N}(x^*, \frac{1}{L}\mathbb{I}_d)$*, the* $\frac{1}{2}$*-lazy versions of MRW and MALA (Algorithms 1 and 2) with step sizes*

$$\eta_{MRW} = c_2 \cdot \frac{1}{Ld\kappa}, \quad \textit{and} \quad \eta_{MALA} = c_1 \cdot \frac{1}{Ld \cdot \max\left\{1, \sqrt{\kappa/d}\right\}} \tag{22}$$

respectively, satisfy the mixing time bounds

$$\tau_2^{MRW}(\epsilon; \mu_0) = O\left(d\kappa^2 \log \frac{d}{\epsilon}\right), \quad \textit{and} \tag{23a}$$

$$\tau_2^{MALA}(\epsilon; \mu_0) = O\left(d\kappa \log \frac{d}{\epsilon} \cdot \max\left\{1, \sqrt{\frac{\kappa}{d}}\right\}\right). \tag{23b}$$

The proof is omitted as it directly follows from the conductance profile based mixing time bound in Lemma 3, Lemma 4 and the overlap bounds for MALA and MRW provided in Dwivedi et al. (2018). Theorem 5 states that the mixing time bounds for MALA and MRW with the feasible distribution $\mu_\dagger$ as the initial distribution scale as $\widetilde{O}(d\kappa \log(1/\epsilon))$ and $\widetilde{O}(d\kappa^2 \log(1/\epsilon))$. Once again, we note that in light of the inequality (3c) we obtain same bounds for the number of steps taken by these algorithms to mix within ϵ total-variation distance of the target distribution Π^*. Consequently, our results improve upon the previously known mixing time bounds for MALA and MRW (Dwivedi et al., 2018) for strongly log-concave distributions. With $\mu_\dagger$ as the initial distribution, the authors had derived bounds of order $\widetilde{O}(d^2\kappa \log(1/\epsilon))$ and $\widetilde{O}(d^2\kappa^2 \log(1/\epsilon))$ for MALA and MRW respectively (cf. Corollary 1 in Dwivedi et al. (2018)). However, the authors stated that their numerical experiments suggested a better dependency on the dimension for the mixing time. Indeed the mixing time bounds from Theorem 5 are smaller by a factor of $\frac{d}{\log d}$, compared to their bounds for both of these chains thereby resolving their open question. Nonetheless, it is still an open question how to establish a lower bound on the mixing time of these sampling algorithms.

4. Numerical experiments

In this section, we numerically compare HMC with MALA and MRW to verify that our suggested step-size and leapfrog steps lead to faster convergence for the HMC algorithm. We adopt the step-size choices for MALA and MRW given in Dwivedi et al. (2018), whereas the choices for step-size and leapfrog steps for HMC are taken from Corollary 14 in this paper. Since we do experiments with multivariate Gaussian distribution, the Hessian-Lipschitz constant L_{H} is always zero. So we also experiment with the step-size and leapfrog steps choice suggested in Appendix D.1.1. When L_{H} is small, HMC can be run with much larger step-size and much larger number of leapfrog steps.

In this simulation, we check the dimension d dependency and condition number κ dependency in the multivariate Gaussian case under our step-size choices. We consider sampling from the multivariate Gaussian distribution with density

$$\Pi^*(x) \propto e^{-\frac{1}{2}x^\top \Sigma^{-1} x}, \tag{24}$$

for some covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. The log density (disregarding constants) and its deriviatives are given by

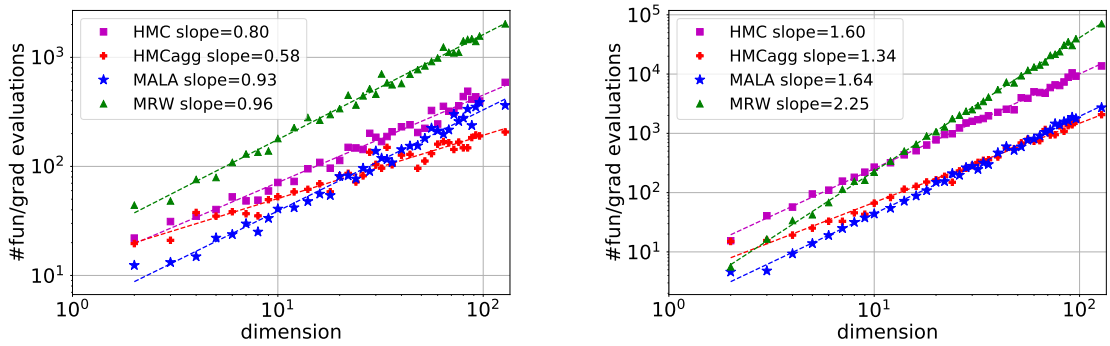
$$f(x) = \frac{1}{2}x^\top \Sigma^{-1} x, \quad \nabla f(x) = \Sigma^{-1}x, \quad \text{and} \quad \nabla^2 f(x) = \Sigma^{-1}.$$

Consequently, the function f is strongly convex with parameter $m = 1/\lambda_{\max}(\Sigma)$ and smooth with parameter $L = 1/\lambda_{\min}(\Sigma)$. For convergence diagnostics, we use the error in quantiles along different directions. Using the exact quantile information for each direction for Gaussians, we measure the error in the 75% quantile of the sample distribution and the true distribution in the *least favorable direction*, i.e., along the eigenvector of Σ corresponding to the eigenvalue $\lambda_{\max}(\Sigma)$. The *quantile mixing time* is defined as the smallest iteration when this error falls below δ. We use $\mu_0 = \mathcal{N}\left(0, L^{-1}\mathbb{I}_d\right)$ as the initial distribution. To make the comparison with MRW and MALA fair, we compare the number of function or gradient evaluations instead of number of iterations. For HMC, the number of gradient evaluations is K times the number of outer-loop iterations.

For every simulation, the parameters for HMC-(K, η) are chosen according to the warm start case in Corollary 2 with $K = 4 \cdot d^{1/4}$, and for MRW and MALA are chosen according to the paper Dwivedi et al. (2018). We also added HMCagg which means that the larger step-sizes are chosen according to Appendix D.1.1 with $K = 4 \cdot d^{1/8}\kappa^{1/4}$, by taking into account that L_{H} is zero for Gaussian distribution. We simulate 10 independent runs of the three chains each with 100 samples to determine the quantile mixing time. The quantile mixing time is used to calculate the number of funtion/gradient evaluations. The final number of function/gradient evaluations for each walk is the average of that over these 10 independent runs.

(a) Dimension dependency for fixed κ: We fix the condition number to be $\kappa = 4$. The Hessian Σ in the multivariate Gaussian distribution is chosen to be diagonal and the square roots of its eigenvalues are linearly spaced between 1.0 to 2.0. To estimate the dimension dependency, we vary dimension d from 2 to 128. Figure 1 (a) shows the dependency of the number of function/gradient evaluations as a function of dimension d for the four random walks in log-log scale. To examine the dimension dependency, we perform linear regression for the number of function/gradient evaluations with respect to dimensions in the log-log scale. The least-squares fits of the slopes for HMC, HMCagg, MALA and MRW are $0.80(\pm 0.12)$, $0.58(\pm 0.15)$, $0.93(\pm 0.13)$ and $0.96(\pm 0.10)$, respectively. Standard errors of the regression coefficient is reported in parentheses. The corresponding theoretical slopes (seen from Table 2 and Appendix D.1.1) are $0.92, 0.63, 1.0, 1.0$ respectively.

(b) Dimension dependency for $\kappa = d^{2/3}$: Here we construct densities with condition number $\kappa = d^{2/3}$ while still varying the dimension d from 2 to 128. This is done by choosing the Hessian Σ in the multivariate Gaussian distribution to be diagonal and to have the square roots of its eigenvalues linearly spaced between 1.0 to $d^{1/3}$. Figure 1 (b) shows the dependency of the number of function/gradient evaluations as a function of dimension d for the four random walks in log-log scale. In order to estimate the exponent α in the dimension dependency d^α, we perform a linear regression of the log mixing time on the log dimension; doing so yields estimated exponents $\widehat{\alpha}$ of $1.60(\pm 0.09)$, $1.34(\pm 0.17)$, $1.64(\pm 0.11)$ and $2.25(\pm 0.08)$ for HMC, HMCagg, MALA and MRW, respectively. Standard errors of the



(a) $\kappa = 4$ (b) $\kappa = d^{\frac{2}{3}}$

Figure 1: Average number of function/gradient evaluations as a function of dimension for four random walks on multivariate Gaussian density (24) where the covariance has a condition number κ that is (a) constant 4 and (b) scales with dimension d. With suggested step-size and leapfrog steps in Corollary 2, the number of function/gradient evaluations of HMC has a smaller dimension dependency than that of MALA or MRW. Since the target distributon is Gaussian and the Hessian-Lipschitz constant L_{H} is zero, larger step-size and larger number of leapfrog steps can be chosen according to Appendix D.1.1. The plots does show that HMCagg with larger step-size and larger number of leapfrog steps uses smaller number of function/gradient evaluations to achieve the same quantile mixing.

regression coefficient is reported in parentheses. The theoretical guarantees given in Table 5 (for HMC), in Table 6 (for HMCagg) and in Table 2 (for MALA and MRW) correspond to the exponents of 1.58, 1.46, 1.67 and 2.33 for the four algorithms respectively.

5. Proofs

This section is devoted primarily to the proof of Theorem 1. In order to do so, we begin with the mixing time bound based on the conductance profile from Lemma 3. We then seek to apply Lemma 4 in order derive a bound on the conductance profile itself. However, in order to do so, we need to derive bound on the overlap between the proposal distributions of HMC at two nearby points and show that the Metropolis-Hastings step only modifies the proposal distribution by a relatively small amount. This control is provided by Lemma 6, stated in Section 5.1. We use it to prove Theorem 1 in Section 5.2. Finally, Section 5.3 is devoted to the proof of Lemma 6.

5.1. Overlap bounds for HMC

In this subsection, we derive two important bounds for the Metropolized HMC chain: (1) first, we quantify the overlap between proposal distributions of the chain for nearby points, and, (2) second, we show that the distortion in the proposal distribution introduced by

the Metropolis-Hastings accept-reject step can be controlled if an appropriate step-size is chosen. Putting the two pieces together enables us to invoke Lemma 4 to prove Theorem 1.

In order to do so, we begin with some notation. Let $\mathcal{T}$ denote the transition operator of the HMC chain with leapfrog integrator taking step-size η and number of leapfrog updates K. Let $\mathcal{P}_x$ denote the proposal distribution at $x \in \mathcal{X}$ for the chain before the accept-reject step and the lazy step. Let $\mathcal{T}_x^{\text{before-lazy}}$ denote the corresponding transition distribution after the proposal and the accept-reject step, before the lazy step. By definition, we have

$$\mathcal{T}_x(A) = \zeta\delta_x(A) + (1-\zeta)\mathcal{T}_x^{\text{before-lazy}}(A) \qquad \text{for any measurable set } A \in \mathcal{B}(\mathcal{X}). \tag{25}$$

Our proofs make use of the Euclidean ball $\mathcal{R}_s$ defined in equation (29). At a high level, the HMC chain has bounded gradient inside the ball $\mathcal{R}_s$ for a suitable choice of s, and the gradient of the log-density gets too large outside such a ball making the chain unstable in that region. However, since the target distribution has low mass in that region, the chain's visit to the region outside the ball is a rare event and thus we can focus on the chain's behavior inside the ball to analyze its mixing time.

In the next lemma, we state the overlap bounds for the transition distributions of the HMC chain. For a fixed univeral constant c, we require

$$K^2\eta^2 \leq \frac{1}{4\max\left\{d^{\frac{1}{2}}L, d^{\frac{2}{3}}L_{\text{H}}^{\frac{2}{3}}\right\}}, \qquad \text{and} \tag{26a}$$

$$\eta^2 \leq \frac{1}{cL}\min\left\{\frac{1}{K^2}, \frac{1}{Kd^{\frac{1}{2}}}, \frac{1}{K^{\frac{2}{3}}d^{\frac{1}{3}}\left(\frac{M^2}{L}\right)^{\frac{1}{3}}}, \frac{1}{K\frac{M}{L^{\frac{1}{2}}}}, \frac{1}{K^{\frac{2}{3}}d}\frac{L}{L_{\text{H}}^{\frac{2}{3}}}, \frac{1}{K^{\frac{4}{3}}\frac{M}{L^{\frac{1}{2}}}}\left(\frac{L}{L_{\text{H}}^{\frac{2}{3}}}\right)^{\frac{1}{2}}\right\}. \tag{26b}$$

Lemma 6 *Consider a $(L, L_H, s, \psi_{\mathfrak{a}}, M)$-regular target distribution (cf. Assumption (A)) with Ω the convex measurable set satisfying* (10e)*. Then with the parameters (K, η) satisfying $K\eta \leq \frac{1}{4L}$ and condition* (26a)*, the HMC-(K, η) chain satisfies*

$$\sup_{\|q_0 - \tilde{q}_0\|_2 \leq \frac{K\eta}{4}} d_{TV}(\mathcal{P}_{q_0}, \mathcal{P}_{\tilde{q}_0}) \leq \frac{1}{2}. \tag{27a}$$

If, in addition, condition (26b) *holds, then we have*

$$\sup_{x\in\Omega} d_{TV}\left(\mathcal{P}_x, \mathcal{T}_x^{\textit{before-lazy}}\right) \leq \frac{1}{8}. \tag{27b}$$

See Appendix 5.3 for the proof.

Lemma 6 is crucial to the analysis of HMC as it enables us to apply the conductance profile based bounds discussed in Section 3.3. It reveals two important properties of the Metropolized HMC. First, from equation (27a), we see that proposal distributions of HMC at two different points are close if the two points are close. This is proved by controlling the KL-divergence of the two proposal distributions of HMC via change of variable formula. Second, equation (27b) shows that the accept-reject step of HMC is well behaved inside Ω provided the gradient is bounded by M.

5.2. Proof of Theorem 1

We are now equipped to prove our main theorem. In order to prove Theorem 1, we begin by using Lemma 4 and Lemma 6 to derive an explicit bound for on the HMC conductance profile. Given the assumptions of Theorem 1, conditions (26a) and (26b) hold, enabling us to invoke Lemma 6 in the proof.

Define the function $\Psi_\Omega : [0,1] \mapsto \mathbb{R}_+$ as

$$\Psi_\Omega(v) = \begin{cases} \frac{1}{32} \cdot \min\left\{1, \frac{K\eta}{64\psi_{\mathfrak{a}}} \log^{\mathfrak{a}}\left(\frac{1}{v}\right)\right\} & \text{if } v \in \left[0, \frac{1-s}{2}\right]. \\ \frac{K\eta}{2048\psi_{\mathfrak{a}}}, & \text{if } v \in \left(\frac{1-s}{2}, 1\right]. \end{cases} \tag{28}$$

This function acts as a lower bound on the truncated conductance profile. Define the Euclidean ball

$$\mathcal{R}_s = \mathbb{B}\left(x^\star, r(s)\sqrt{\frac{d}{m}}\right), \tag{29}$$

and consider a pair $(x, y) \in \mathcal{R}_s$ such that $\|x - y\|_2 \leq \frac{1}{4}K\eta$. Invoking the decomposition (25) and applying triangle inequality for ζ-lazy HMC, we have

$$\begin{aligned} d_{\mathrm{TV}}\left(\mathcal{T}_x, \mathcal{T}_y\right) &\leq \zeta + (1-\zeta)\, d_{\mathrm{TV}}\left(\mathcal{T}_x^{\text{before-lazy}}, \mathcal{T}_y^{\text{before-lazy}}\right) \\ &\leq \zeta + (1-\zeta)\left(d_{\mathrm{TV}}\left(\mathcal{T}_x^{\text{before-lazy}}, \mathcal{P}_y\right) + d_{\mathrm{TV}}\left(\mathcal{P}_x, \mathcal{P}_y\right) + d_{\mathrm{TV}}\left(\mathcal{P}_x, \mathcal{T}_y^{\text{before-lazy}}\right)\right) \\ &\overset{(i)}{\leq} \zeta + (1-\zeta)\left(\frac{1}{4} + \frac{1}{2} + \frac{1}{4}\right) \\ &= 1 - \frac{1-\zeta}{4}, \end{aligned}$$

where step (i) follows from the bounds (27a) and (27b) from Lemma 6. For $\zeta = \frac{1}{2}$, substituting $\omega = \frac{1}{8}$, $\Delta = \frac{1}{4}K\eta$ and the convex set $\Omega = \mathcal{R}_s$ into Lemma 4, we obtain that

$$\Phi_\Omega(v) \geq \frac{1}{32} \cdot \min\left\{1, \frac{K\eta}{64\psi_{\mathfrak{a}}} \log^{\mathfrak{a}}\left(1 + \frac{1}{v}\right)\right\}, \quad \text{for } v \in \left[0, \frac{1-s}{2}\right].$$

Here $\mathfrak{a}$ equals to $\frac{1}{2}$ or 0, depending on the assumption (10d). By the definition of the truncated conductance profile (15), we have that $\tilde{\Phi}_\Omega(v) \geq \frac{K\eta}{2048\psi_{\mathfrak{a}}}$ for $v \in \left[\frac{1-s}{2}, 1\right]$. As a consequence, Ψ_Ω is effectively a lower bound on the truncated conductance profile. Note that the assumption (A) ensures the existence of Ω such that $\Pi^*(\Omega) \geq 1 - s$ for $s = \frac{\epsilon^2}{2\beta^2}$. Putting the pieces together and applying Lemma 3 with the convex set Ω concludes the proof of the theorem.

5.3. Proof of Lemma 6

In this subsection, we prove the two main claims (27a) and (27b) in Lemma 6. Before going into the claims, we first provide several convenient properties about the HMC proposal.

5.3.1. Properties of the HMC proposal

Recall the Hamiltonian Monte Carlo (HMC) with leapfrog integrator (8c). Using an induction argument, we find that the final states in one iteration of K steps of the HMC chain, denoted by q_K and p_K satisfy

$$p_K = p_0 - \frac{\eta}{2}\nabla f(q_0) - \sum_{j=1}^{K-1} \nabla f(q_j) - \frac{\eta}{2}\nabla f(q_K), \tag{30a}$$

$$\text{and} \quad q_K = q_0 + K\eta p_0 - \frac{K\eta^2}{2}\nabla f(q_0) - \eta^2 \sum_{j=1}^{K-1} (K-j)\,\nabla f(q_j). \tag{30b}$$

It is easy to see that for $k \in [K]$, q_k can be seen as a function of the initial state q_0 and p_0. We denote this function as the *forward mapping* F,

$$q_k =: F_k(p_0, q_0) \quad \text{and} \quad q_K =: F_K(p_0, q_0) =: F\left(p_0, q_0\right) \tag{30c}$$

where we introduced the simpler notation $F := F_K$ for the final iterate. The forward mappings F_k and F are deterministic functions that only depends on the gradient ∇f, the number of leapfrog updates K and the step size η.

Denote $\mathbf{J}_x F$ as the Jacobian matrix of the forward mapping F with respect to the first variable. By definition, it satisfies

$$\left[\mathbf{J}_x F(x, q_0)\right]_{ij} = \frac{\partial}{\partial x_j}\left[F(x, q_0)\right]_i, \quad \text{for all} \quad i, j \in [d]\,. \tag{30d}$$

Similarly, denote $\mathbf{J}_y F$ as the Jacobian matrix of the forward mapping F with respect to the second variable. The following lemma characterizes the eigenvalues of the Jacobian $\mathbf{J}_x F$.

Lemma 7 *Suppose the log density f is L-smooth. For the number of leapfrog steps and step-size satisfying $K^2\eta^2 \le \frac{1}{4L}$, we have*

$$\left\|\!\left| K\eta\mathbb{I}_d - \mathbf{J}_x F(x, y) \right|\!\right\|_2 \le \frac{1}{8} K\eta, \quad \textit{for all} \quad x, y \in \mathcal{X} \textit{ and } i \in [d]\,.$$

Also all eigenvalues of $\mathbf{J}_x F(x, y)$ have absolute value greater or equal to $\frac{7}{8}K\eta$.

See Appendix A.3.1 for the proof.

Since the Jacobian is invertible for $K^2\eta^2 \le \frac{1}{4L}$, we can define the inverse function of F with respect to the first variable as the backward mapping G. We have

$$F\left(G(x, y), y\right) = x, \quad \text{for all} \quad x, y \in \mathcal{X}. \tag{31}$$

Moreover as a direct consequence of Lemma 7, we obtain that the magnitude of the eigenvalues of the Jacobian matrix $\mathbf{J}_x G(x, y)$ lies in the interval $\left[\frac{8}{9K\eta}, \frac{8}{7K\eta}\right]$. In the next lemma, we state another set of bounds on different Jacobian matrices:

Lemma 8 *Suppose the log density f is L-smooth. For the number of leapfrog steps and step-size satisfying $K^2\eta^2 \leq \frac{1}{4L}$, we have*

$$\left\|\!\left|\mathbf{J}_y G(x,y)\right|\!\right\|_2 \leq \frac{4}{3K\eta}, \quad \textit{for all} \quad x, y \in \mathcal{X}, \quad \textit{and} \tag{32a}$$

$$\left\|\frac{\partial F_k(G(x,y),y)}{\partial y}\right\|_2 \leq 3, \qquad \textit{for all} \quad k \in [K]\,. \tag{32b}$$

See Appendix A.3.2 for the proof.

Next, we would like to obtain a bound on the quantity $\frac{\partial \log\det \mathbf{J}_x G(x,q_0)}{\partial y}$. Applying the chain rule, we find that

$$\frac{\partial \log\det \mathbf{J}_x G(x,q_0)}{\partial y} = \begin{bmatrix} \text{trace}\left([\mathbf{J}_x G(x,q_0)]^{-1}\mathbf{J}_{xy_1} G(x,q_0)\right) \\ \vdots \\ \text{trace}\left([\mathbf{J}_x G(x,q_0)]^{-1}\mathbf{J}_{xy_d} G(x,q_0)\right) \end{bmatrix}. \tag{33}$$

Here $\mathbf{J}_{xy} G(x,q_0)$ is a third order tensor and we use $\mathbf{J}_{xy_l} G(x,q_0)$ to denote the matrix corresponding to the l-th slice of the tensor which satisfies

$$\left[\mathbf{J}_{xy_l} G(x,q_0)\right]_{ij} = \frac{\partial\partial}{\partial x_j y_l}\left[F(x,q_0)\right]_i, \qquad \text{for all} \quad i,j,l \in [d]\,.$$

Lemma 9 *Suppose the log density f is L-smooth and L_H-Hessian Lipschitz. For the number of leapfrog steps and step-size satisfying $K^2\eta^2 \leq \frac{1}{4L}$, we have*

$$\left\|\frac{\partial \log\det \mathbf{J}_x G(x,q_0)}{\partial y}\right\|_2 = \left\|\begin{bmatrix} \textit{trace}\left([\mathbf{J}_x G(x,q_0)]^{-1}\mathbf{J}_{xy_1} G(x,q_0)\right) \\ \vdots \\ \textit{trace}\left([\mathbf{J}_x G(x,q_0)]^{-1}\mathbf{J}_{xy_d} G(x,q_0)\right) \end{bmatrix}\right\|_2 \leq 2dK^2\eta^2 L_H.$$

See Appendix A.3.3 for the proof.

As a direct consequence of the equation (30b) at k-th step of leapfrog updates, we obtain the following two bounds for the difference between successive F_k terms that come in handy later in our proofs.

Lemma 10 *Suppose that the log density f is L-smooth. For the number of leapfrog steps and step-size satisfying $K^2\eta^2 \leq \frac{1}{4L}$, we have*

$$\|F_k(p_0,q_0) - q_0\|_2 \leq 2k\eta\,\|p_0\|_2 + 2k^2\eta^2\,\|\nabla f(q_0)\|_2 \quad \textit{for } k \in [K]\,, \ \textit{and} \tag{34a}$$

$$\|F_{k+1}(p_0,q_0) - F_k(p_0,q_0)\|_2 \leq 2\eta\,\|p_0\|_2 + 2(k+1)\eta^2\,\|\nabla f(q_0)\|_2 \ \textit{ for } k \in [K-1]\,. \tag{34b}$$

See Appendix A.3.4 for the proof.

We now turn to the proof the two claims in Lemma 6. Note that the claim (27a) states that the proposal distributions at two close points are close; the claim (27b) states that the proposal distribution and the transition distribution are close.

5.3.2. Proof of claim (27a) in Lemma 6

In order to bound the distance between proposal distributions of nearby points, we prove the following stronger claim: For a L-smooth L_{H}-Hessian-Lipschitz target distribution, the proposal distribution of the HMC algorithm with step size η and leapfrog steps K such that $K\eta \le \frac{1}{4L}$ satisfies

$$d_{\mathrm{TV}}\left(\mathcal{P}_{q_0}, \mathcal{P}_{\tilde{q}_0}\right) \le \left(\frac{2\left\|q_0 - \tilde{q}_0\right\|_2^2}{K^2\eta^2} + 3\sqrt{d}K\eta L\left\|q_0 - \tilde{q}_0\right\|_2 + 4dK^2\eta^2 L_{\mathrm{H}}\left\|q_0 - \tilde{q}_0\right\|_2\right)^{1/2}, \quad (35)$$

for all $q_0, \tilde{q}_0 \in \mathbb{R}^d$. Then for any two points $q_0, \tilde{q}_0$ such that $\|q_0 - \tilde{q}_0\|_2 \le \frac{1}{4}K\eta$, under the condition (26a), i.e., $K^2\eta^2 \le \frac{1}{4\max\left\{d^{\frac{1}{2}}L, d^{\frac{2}{3}}L_{\mathrm{H}}^{\frac{2}{3}}\right\}}$, we have

$$d_{\mathrm{TV}}\left(\mathcal{P}_{q_0}, \mathcal{P}_{\tilde{q}_0}\right) \le \left(\frac{1}{8} + \frac{3}{64} + \frac{1}{64}\right)^{1/2} \le \frac{1}{2},$$

and the claim (27a) follows.

The proof of claim (35) involves the following steps: (1) we make use of the update rules (30b) and change of variable formula to obtain an expression for the density of q_n in terms of q_0, (2) then we use Pinsker's inequality and derive expressions for the KL-divergence between the two proposal distributions, and (3) finally, we upper bound the KL-divergence between the two distributions using different properties of the forward mapping F from Appendix 5.3.1.

According to the update rule (30b), the proposals from two initial points q_0 and $\tilde{q}_0$ satisfy respectively

$$q_K = F(p_0, q_0), \quad \text{and} \quad \tilde{q}_K = F(\tilde{p}_0, \tilde{q}_0),$$

where p_0 and $\tilde{p}_0$ are independent random variable from Gaussian distribution $\mathcal{N}(0, \mathbb{I}_d)$.

Denote ρ_{q_0} as the density function of the proposal distribution $\mathcal{P}_{q_0}$. For two different initial points q_0 and $\tilde{q}_0$, the goal is to bound the total variation distance between the two proposal distribution, which is by definition

$$d_{\mathrm{TV}}\left(\mathcal{P}_{q_0}, \mathcal{P}_{\tilde{q}_0}\right) = \frac{1}{2}\int_{x\in\mathcal{X}} \left|\rho_{q_0}(x) - \rho_{\tilde{q}_0}(x)\right| dx. \quad (36)$$

Given q_0 fixed, the random variable q_K can be seen as a transformation of the Gaussian random variable p_0 through the function $F(\cdot, q_0)$. When F is invertible, we can use the change of variable formula to obtain an explicit expression of the density ρ_{q_0}:

$$\rho_{q_0}(x) = \varphi\left(G(x, q_0)\right) \det\left(\mathbf{J}_x G(x, q_0)\right), \quad (37)$$

where φ is the density of the standard Gaussian distribution $\mathcal{N}(0, \mathbb{I}_d)$. Note that even though explicit, directly bounding the total variation distance (36) using the complicated density expression (37) is difficult. We first use Pinsker's inequality (Cover and Thomas, 1991) to give an upper bound of the total variance distance in terms of KL-divergence

$$d_{\mathrm{TV}}\left(\mathcal{P}_{q_0}, \mathcal{P}_{\tilde{q}_0}\right) \le \sqrt{2\mathrm{KL}(\mathcal{P}_{q_0} \parallel \mathcal{P}_{\tilde{q}_0})}, \quad (38)$$

and then upper bound the KL-divergence. Plugging the density (37) into the KL-divergence formula, we obtain that

$$\begin{aligned}
\mathrm{KL}(\mathcal{P}_{q_0} \parallel \mathcal{P}_{\tilde{q}_0}) &= \int_{\mathbb{R}^d} \rho_{q_0}(x) \log\left(\frac{\rho_{q_0}(x)}{\rho_{\tilde{q}_0}(x)}\right) dx \\
&= \int_{\mathbb{R}^d} \rho_{q_0}(x) \left[\log\left(\frac{\varphi\left(G(x,q_0)\right)}{\varphi\left(G(x,\tilde{q}_0)\right)}\right) + \log\det \mathbf{J}_x G(x,q_0) - \log\det \mathbf{J}_x G(x,\tilde{q}_0)\right] dx \\
&= \underbrace{\int_{\mathbb{R}^d} \rho_{q_0}(x) \left[\frac{1}{2}\left(-\|G(x,q_0)\|_2^2 + \|G(x,\tilde{q}_0)\|_2^2\right)\right] dx}_{T_1} \\
&\qquad + \underbrace{\int_{\mathbb{R}^d} \rho_{q_0}(x) \left[\log\det \mathbf{J}_x G(x,q_0) - \log\det \mathbf{J}_x G(x,\tilde{q}_0)\right] dx}_{T_2} \qquad (39)
\end{aligned}$$

We claim the following bounds on the terms T_1 and T_2:

$$|T_1| \leq \frac{8}{9}\frac{\|q_0 - \tilde{q}_0\|_2^2}{K^2\eta^2} + \frac{3}{2}\sqrt{d}K\eta L \|q_0 - \tilde{q}_0\|_2, \quad \text{and} \qquad (40a)$$

$$|T_2| \leq 2dK^2\eta^2 L_{\mathrm{H}} \|q_0 - \tilde{q}_0\|_2, \qquad (40b)$$

where the bound on T_2 follows readily from Lemma 9:

$$\begin{aligned}
|T_2| &= \left| \int \rho_{q_0}(x) \left[\log\det \mathbf{J}_x G(x,q_0) - \log\det \mathbf{J}_x G(x,\tilde{q}_0)\right] dx \right| \\
&\leq \left\| \frac{\partial \log\det \mathbf{J}_x G(x,q_0)}{\partial y} \right\|_2 \|q_0 - \tilde{q}_0\|_2 \\
&\leq 2dK^2\eta^2 L_{\mathrm{H}} \|q_0 - \tilde{q}_0\|_2 . \qquad (41)
\end{aligned}$$

Putting together the inequalities (38), (39), (40a) and (40b) yields the claim (35).

It remains to prove the bound (40a) on T_1.

Proof of claim (40a)**:** For the term T_1, we observe that

$$\frac{1}{2}\left(\|G(x,\tilde{q}_0)\|_2^2 - \|G(x,q_0)\|_2^2\right) = \frac{1}{2}\|G(x,q_0) - G(x,\tilde{q}_0)\|_2^2 - \left(G(x,q_0) - G(x,\tilde{q}_0)\right)^\top G(x,q_0).$$

The first term on the RHS can be bounded via the Jacobian of G with respect to the second variable. Applying the bound (32a) from Lemma 8, we find that

$$\|G(x,q_0) - G(x,\tilde{q}_0)\|_2 \leq |||\mathbf{J}_y G(x,y)|||_2 \|q_0 - \tilde{q}_0)\|_2 \leq \frac{4}{3K\eta}\|q_0 - \tilde{q}_0)\|_2 . \qquad (42)$$

For the second part, we claim that there exists a deterministic function C of q_0 and $\tilde{q}_0$ and independent of x, such that

$$\|G(x,q_0) - G(x,\tilde{q}_0) - C(q_0,\tilde{q}_0)\|_2 \leq \frac{3}{2}K\eta L \|q_0 - \tilde{q}_0\|_2 . \qquad (43)$$

Assuming the claim (43) as given at the moment, we can further decompose the second part of T_1 into two parts:

$$\left(G(x,q_0)-G(x,\tilde{q}_0)\right)^\top G(x,q_0) = \left(G(x,q_0)-G(x,\tilde{q}_0)-C(q_0,\tilde{q}_0)\right)^\top G(x,q_0) + C(q_0,\tilde{q}_0)^\top G(x,q_0) \tag{44}$$

Applying change of variables along with equation (37), we find that

$$\int \rho_{q_0}(x) G(x,q_0) dx = \int \varphi(x) x dx = 0.$$

Furthermore, we also have

$$\begin{aligned}\int_{x\in\mathcal{X}} \rho_{q_0}(x) \left\|G(x,q_0)\right\|_2 dx &= \int_{x\in\mathcal{X}} \varphi(x) \left\|x\right\|_2 dx \\ &\overset{(i)}{\leq} \left[\left(\int_{x\in\mathcal{X}} \varphi(x)\left\|x\right\|_2^2 dx\right)\left(\int_{x\in\mathcal{X}}\varphi(x)dx\right)\right]^{1/2} = \sqrt{d},\end{aligned}$$

where step (i) follows from Cauchy-Schwarz's inequality. Combining the inequalities (42), (43) and (44) together, we obtain the following bound on term T_1:

$$\begin{aligned}|T_1| &= \left|\int \rho_{q_0}(x)\left[-\frac{1}{2}\left\|G(x,q_0)\right\|_2^2 + \frac{1}{2}\left\|G(x,\tilde{q}_0)\right\|_2^2\right] dx\right| \\ &\leq \frac{1}{2}\left|\int \rho_{q_0}(x)\left\|G(x,q_0)-G(x,\tilde{q}_0)\right\|_2^2 dx\right| \\ &\qquad + \left|\int \rho_{q_0}(x)\left\|G(x,q_0)-G(x,\tilde{q}_0)-C(q_0,\tilde{q}_0)\right\|_2\left\|G(x,q_0)\right\|_2 dx\right| \\ &\leq \frac{8}{9}\frac{\left\|q_0-\tilde{q}_0\right\|_2^2}{K^2\eta^2} + \frac{3}{2}\sqrt{d}K\eta\left\|q_0-\tilde{q}_0\right\|_2, \end{aligned} \tag{45}$$

which yields the claimed bound on T_1.

We now prove our earlier claim (43).

Proof of claim (43)**:** For any pair of states q_0 and $\tilde{q}_0$, invoking the definition (31) of the map $G(x,\cdot)$, we obtain the following implicit equations:

$$x = q_0 + K\eta G(x,q_0) - K\frac{\eta^2}{2}\nabla f(q_0) - \eta^2\sum_{j=1}^{K-1}(K-j)\nabla f(F_j(G(x,q_0),q_0)), \quad \text{and}$$

$$x = \tilde{q}_0 + K\eta G(x,\tilde{q}_0) - K\frac{\eta^2}{2}\nabla f(\tilde{q}_0) - \eta^2\sum_{j=1}^{K-1}(K-j)\nabla f(F_j(G(x,\tilde{q}_0),\tilde{q}_0)).$$

Taking the difference between the two equations above, we obtain

$$\begin{aligned}G(x,q_0) - G(x,\tilde{q}_0) - \frac{q_0-\tilde{q}_0}{K\eta} - \frac{\eta}{2}\left(\nabla f(q_0)-\nabla f(\tilde{q}_0)\right) \\ = \frac{\eta^2}{K\eta}\sum_{k=1}^{K-1}(K-j)\left(\nabla f(F_k(G(x,q_0),q_0)) - \nabla f(F_k(G(x,\tilde{q}_0),\tilde{q}_0))\right).\end{aligned}$$

Applying L-smoothness of f along with the bound (32b) from Lemma 8, we find that

$$\begin{aligned}\|\nabla f(F_k(G(x,q_0),q_0)) - \nabla f(F_k(G(x,\tilde{q}_0),\tilde{q}_0))\|_2 &\leq L \left\| \frac{\partial F_k(G(x,y),y)}{\partial y} \right\|_2 \|q_0 - \tilde{q}_0\|_2 \\ &\leq 3L \|q_0 - \tilde{q}_0\|_2 .\end{aligned}$$

Putting the pieces together, we find that

$$\left\| G(x,q_0) - G(x,\tilde{q}_0) - \frac{q_0 - \tilde{q}_0}{K\eta} - \frac{1}{2}\left(\nabla f(q_0) - \nabla f(\tilde{q}_0)\right) \right\|_2 \leq \frac{3K\eta L}{2} \|q_0 - \tilde{q}_0\|_2 ,$$

which yields the claim (43).

5.3.3. Proof of claim (27b) in Lemma 6

We now bound the distance between the one-step proposal distribution $\mathcal{P}_x$ at point x and the one-step transition distribution $\mathcal{T}_x^{\text{before-lazy}}$ at x obtained after performing the accept-reject step (and no lazy step). Using equation (30a), we define the forward mapping E for the variable p_K as follows

$$p_K = E(p_0, q_0) := p_0 - \frac{\eta}{2}\nabla f(q_0) - \eta \sum_{j=1}^{K-1} \nabla f(q_j) - \frac{\eta}{2}\nabla f(q_K).$$

Consequently, the probability of staying at x is given by

$$\mathcal{T}_x^{\text{before-lazy}}(\{x\}) = 1 - \int_{\mathcal{X}} \min\left\{1, \frac{\exp(-\mathcal{H}(E(z,x),F(z,x)))}{\exp(-\mathcal{H}(z,x))}\right\} \varphi_x(z)dz,$$

where the Hamiltonian $\mathcal{H}(q,p) = f(q) + \frac{1}{2}\|p\|_2^2$ was defined in equation (7). As a result, the TV-distance between the proposal and transition distribution is given by

$$\begin{aligned} d_{\text{TV}}\left(\mathcal{P}_x, \mathcal{T}_x^{\text{before-lazy}}\right) &= 1 - \int_{\mathcal{X}} \min\left\{1, \frac{\exp(-\mathcal{H}(E(z,x),F(z,x)))}{\exp(-\mathcal{H}(z,x))}\right\} \varphi_x(z)dz \\ &= 1 - \mathbb{E}_{z\sim\mathcal{N}(0,\mathbb{I}_d)}\left[\min\left\{1, \frac{\exp(-\mathcal{H}(E(z,x),F(z,x)))}{\exp(-\mathcal{H}(z,x))}\right\}\right]. \end{aligned} \tag{46}$$

An application of Markov's inequality yields that

$$\begin{aligned} &\mathbb{E}_{z\sim\mathcal{N}(0,\mathbb{I}_d)}\left[\min\left\{1, \frac{\exp(-\mathcal{H}(E(z,x),F(z,x)))}{\exp(-\mathcal{H}(z,x))}\right\}\right] \\ &\qquad \geq \alpha \mathbb{P}_{z\sim\mathcal{N}(0,\mathbb{I}_d)}\left[\frac{\exp(-\mathcal{H}(E(z,x),F(z,x)))}{\exp(-\mathcal{H}(z,x))} \geq \alpha\right], \end{aligned} \tag{47}$$

for any $\alpha \in (0,1]$. Thus, to bound the distance $d_{\text{TV}}\left(\mathcal{P}_x, \mathcal{T}_x^{\text{before-lazy}}\right)$, it suffices to derive a high probability lower bound on the ratio $\exp(-\mathcal{H}(E(z,x),F(z,x)))/\exp(-\mathcal{H}(z,x))$ when $z \sim \mathcal{N}(0,\mathbb{I}_d)$.

We now derive a lower bound on the following quantity:

$$\exp\left(-f(F(p_0,q_0))+f(q_0)-\frac{1}{2}\left\|E(p_0,q_0)\right\|_2^2+\frac{1}{2}\left\|p_0\right\|_2^2\right), \quad \text{when } p_0\sim\mathcal{N}(0,\mathbb{I}_d).$$

We derive the bounds on the two terms $-f(F(p_0,q_0))+f(q_0)$ and $\|E(p_0,q_0)\|_2^2$ separately.

Observe that

$$f(F(p_0,q_0))-f(q_0)=\sum_{j=0}^{K-1}\left[f(F_{j+1}(p_0,q_0))-f(F_j(p_0,q_0))\right].$$

The intuition is that it is better to apply Taylor expansion on closer points. Applying the third order Taylor expansion and using the smoothness assumptions (10a) and (10c) for the function f, we obtain

$$f(x)-f(y)\le\frac{(x-y)^\top}{2}\left(\nabla f(x)+\nabla f(y)\right)+L_{\mathrm{H}}\left\|x-y\right\|_2^3.$$

For the indices $j\in\{0,\dots,K-1\}$, using F_j as the shorthand for $F_j(p_0,q_0)$, we find that

$$\begin{aligned}
f(F_{j+1})-f(F_j)&\le\frac{(F_{j+1}-F_j)^\top}{2}\left(\nabla f(F_{j+1})+\nabla f(F_j)\right)+L_{\mathrm{H}}\left\|F_{j+1}-F_j\right\|_2^3\\
&=\frac{1}{2}\eta p_0^\top\left(\nabla f(F_{j+1})+\nabla f(F_j)\right)\\
&\quad-\frac{\eta^2}{2}\left[\frac{1}{2}\nabla f(p_0)+\sum_{k=1}^{j}\nabla f(F_k)\right]^\top\left(\nabla f(F_{j+1})+\nabla f(F_j)\right)+L_{\mathrm{H}}\left\|F_{j+1}-F_j\right\|_2^3,
\end{aligned}\tag{48}$$

where the last equality follows by definition (30c) of the operator F_j.

Now to bound the term $E(p_0,q_0)$, we observe that

$$\begin{aligned}
\frac{\|E(p_0,q_0)\|_2^2}{2}&=\frac{\left\|p_0-\frac{\eta}{2}\nabla f(q_0)-\eta\sum_{j=1}^{K-1}\nabla f(F_j)-\frac{\eta}{2}\nabla f(F_K)\right\|_2^2}{2}\\
&=\frac{\|p_0\|_2^2}{2}-\eta p_0^\top\left(\frac{1}{2}\nabla f(q_0)+\sum_{j=1}^{K-1}\nabla f(F_j)+\frac{1}{2}\nabla f(F_K)\right)\\
&\qquad+\frac{\eta^2}{2}\left\|\frac{1}{2}\nabla f(q_0)+\sum_{j=1}^{K-1}\nabla f(F_j)+\frac{1}{2}\nabla f(F_K)\right\|_2^2.
\end{aligned}\tag{49}$$

Putting the equations (48) and (49) together leads to cancellation of many gradient terms and we obtain

$$\begin{aligned}
&- f(F(p_0, q_0)) + f(q_0) - \frac{1}{2}\|E(p_0, q_0)\|_2^2 + \frac{1}{2}\|p_0\|_2^2 \\
&\geq \frac{\eta^2}{8}\left(\nabla f(q_0) - \nabla f(F_K)\right)^\top \left(\nabla f(q_0) + \nabla f(F_K)\right) - L_{\mathrm{H}} \sum_{j=0}^{K-1} \|F_{j+1} - F_j\|_2^3 \\
&\geq -\frac{\eta^2 L}{4}\|q_0 - F(p_0, q_0)\|_2 \|\nabla f(q_0)\|_2 - \frac{\eta^2 L^2}{2}\|q_0 - F(p_0, q_0)\|_2^2 - L_{\mathrm{H}} \sum_{j=0}^{K-1} \|F_{j+1} - F_j\|_2^3
\end{aligned} \tag{50}$$

The last inequality uses the smoothness condition (10a) for the function f. Plugging the bounds (34a) and (34b) in equation (50), we obtain a lower bound that only depends on $\|p_0\|_2$ and $\|\nabla f(q_0)\|_2$:

$$\begin{aligned}
\text{RHS of (50)} \geq -2K^2\eta^4 L^2 \|p_0\|_2^2 - 2K\eta^3 L \|p_0\|_2 \|\nabla f(q_0)\|_2 - 2K^2\eta^4 L \|\nabla f(q_0)\|_2^2 \\
- L_{\mathrm{H}}\left(32K\eta^3 \|p_0\|_2^3 + 8K^4\eta^6 \|\nabla f(q_0)\|_2^3\right).
\end{aligned} \tag{51}$$

According to assumption (A), we have bounded gradient in the convex set Ω. For any $x \in \Omega$, we have $\|\nabla f(x)\|_2 \leq M$. Standard Chi-squared tail bounds imply that

$$\mathbb{P}\left[\|p_0\|_2^2 \leq d\alpha_1\right] \geq 1 - \frac{1}{16}, \quad \text{for } \alpha_1 = 1 + 2\sqrt{\log(16)} + 2\log(16). \tag{52}$$

Plugging the gradient bound and the bound (52) into equation (51), we conclude that there exists an absolute constant $c \leq 2000$ such that for η^2 satisfying equation (26b), namely

$$\eta^2 \leq \frac{1}{cL} \min\left\{ \frac{1}{K^2}, \frac{1}{Kd^{\frac{1}{2}}}, \frac{1}{K^{\frac{2}{3}} d^{\frac{1}{3}} \left(\frac{M^2}{L}\right)^{\frac{1}{3}}}, \frac{1}{K\frac{M}{L^{\frac{1}{2}}}}, \frac{1}{K^{\frac{2}{3}} d} \frac{L}{L_{\mathrm{H}}^{\frac{2}{3}}}, \frac{1}{K^{\frac{4}{3}}\frac{M}{L^{\frac{1}{2}}}} \left(\frac{L}{L_{\mathrm{H}}^{\frac{2}{3}}}\right)^{\frac{1}{2}} \right\},$$

we have

$$\mathbb{P}\left[-f(F(p_0, q_0)) + f(q_0) - \frac{1}{2}\|E(p_0, q_0)\|_2^2 + \frac{1}{2}\|p_0\|_2^2 \geq -1/16\right] \geq 1 - \frac{1}{16}.$$

Plugging this bound in the inequality (47) yields that

$$\mathbb{E}_{z \sim \mathcal{N}(0, \mathbb{I}_d)}\left[\min\left\{1, \frac{\exp(-\mathcal{H}(E(z,x), F(z,x)))}{\exp(-\mathcal{H}(z,x))}\right\}\right] \geq 1 - \frac{1}{8},$$

which when plugged in equation (46) implies that $d_{\mathrm{TV}}\left(\mathcal{P}_x, \mathcal{T}_x^{\text{before-lazy}}\right) \leq 1/8$ for any $x \in \mathcal{R}_s$, as claimed. The proof is now complete.

6. Discussion

In this paper, we derived non-asymptotic bounds on mixing time of Metropolized Hamiltonian Monte Carlo for log-concave distributions. Our results show that by choosing appropriate step-size and number of leapfrog steps, we obtain HMC convergence rate which is faster than the current best convergence rate of MALA. This improvement can be seen as the benefit of using multi-step gradients in HMC. An interesting open problem is to determine whether our HMC mixing rate is tight for log-concave sampling under the assumptions made in the paper.

Even though, we focused on the problem of sampling only from strongly and weakly log-concave distribution, our Theorem 1 applies to general distributions including nearly log-concave distributions as mentioned in Section C.2. It would be interesting to determine the explicit HMC mixing rate for these distributions. The other main contribution of our paper is to improve the warmness dependency in mixing rates of Metropolized algorithms that are proved previously such as MRW and MALA (Dwivedi et al., 2018). Our idea is inspired by the techniques used to improve warmness dependency in the literature of discrete-state Markov chains. It is interesting to ask if this warmness dependency can be further improved to prove a convergence sub-linear in d for HMC even for small condition number κ.

Acknowledgements

We would like to thank Wenlong Mou for his insights and helpful discussions. This work was supported by Office of Naval Research grant DOD ONR-N00014-18-1-2640, and National Science Foundation Grant NSF-DMS-1612948 to MJW and by ARO W911NF1710005, NSF-DMS-1613002, NSF-IIS-174134 and the Center for Science of Information (CSoI), US NSF Science and Technology Center, under grant agreement CCF-0939370 to BY.

Appendix

Appendix A. Proof of Lemmas 3, 4 and 6

In this appendix, we collect the proofs of Lemmas 3, and 4, as previously stated in Section 3.3, that are used in proving Theorem 1. Moreover, we provide the proof of auxiliary results related to HMC proposal that were used in the proof of Lemma 6.

A.1. Proof of Lemma 3

In order to prove Lemma 3, we begin by adapting the spectral profile technique (Goel et al., 2006) to the continuous state setting, and next we relate conductance profile with the spectral profile.

First, we briefly recall the notation from Section 2.2. Let $\Theta : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to \mathbb{R}_+$ denote the transition probability function for the Markov chain and let $\mathcal{T}$ be the corresponding transition operator, which maps a probability measure to another according to the transition probability Θ. Note that for a Markov chain satisfying the smooth chain assumption (16), if the distribution μ admits a density then the distribution $\mathcal{T}(\mu)$ would also admits a density. We use $\mathcal{T}_x$ as the shorthand for $\mathcal{T}(\delta_x)$, the transition distribution of the Markov chain at x.

Let $L_2(\pi^*)$ be the space of square integrable functions under function π^*. The *Dirichlet form* $\mathcal{E} : L_2(\pi^*) \times L_2(\pi^*) \to \mathbb{R}$ associated with the transition probability Θ is given by

$$\mathcal{E}(g,h) = \frac{1}{2}\int_{(x,y)\in\mathcal{X}^2} (g(x)-h(y))^2\,\Theta(x,dy)\pi^*(x)dx. \tag{53}$$

The expectation $\mathbb{E}_{\pi^*} : L_2(\pi^*) \to \mathbb{R}$ and the variance $\mathrm{Var}_{\pi^*} : L_2(\pi^*) \to \mathbb{R}$ with respect to the density π^* are given by

$$\mathbb{E}_{\pi^*}(g) = \int_{x\in\mathcal{X}} g(x)\pi^*(x)dx \quad \text{and} \quad \mathrm{Var}_{\pi^*}(g) = \int_{x\in\mathcal{X}} (g(x)-\mathbb{E}_{\pi^*}(g))^2\,\pi^*(x)dx. \tag{54a}$$

Furthermore, for a pair of measurable sets $(S,\Omega) \subset \mathcal{X}^2$, the Ω-*restricted spectral gap for the set* S is defined as

$$\lambda_\Omega(S) = \inf_{g\in c_0^+(S\cap\Omega)} \frac{\mathcal{E}(g,g)}{\mathrm{Var}_{\pi^*}(g)}, \tag{55a}$$

$$\text{where } c_0^+(S\cap\Omega) = \{g \in L_2(\pi^*) \mid \mathrm{supp}(g) \subset S\cap\Omega, g \geq 0, g \neq \text{constant}\}. \tag{55b}$$

Finally, the Ω-*restricted spectral profile* Λ_Ω is defined as

$$\Lambda_\Omega(v) = \inf_{\Pi^*(S\cap\Omega)\in[0,v]} \lambda_\Omega(S\cap\Omega), \quad \text{for all } v \in \big[0,\infty\big). \tag{56}$$

Note that we restrict the spectral profile to the set Ω. Taking Ω to be $\mathcal{X}$, our definition agrees with the standard definition definitions of the restricted spectral gap and spectral profile in the paper (Goel et al., 2006) for finite state space Markov chains to continuous state space Markov chains.
We are now ready to state a mixing time bound using spectral profile.

Lemma 11 *Consider a reversible irreducible ζ-lazy Markov chain with stationary distribution Π^* satifying the smooth chain assumption* (16)*. Given a β-warm start μ_0, an error tolerance $\epsilon \in (0,1)$ and a set $\Omega \subset \mathcal{X}$ with $\Pi^*(\Omega) \geq 1 - \frac{\epsilon^2}{2\beta^2}$, the L_2-mixing time is bounded as*

$$\tau_2(\epsilon; \mu_0) \leq \left\lceil \int_{4/\beta}^{8/\epsilon^2} \frac{dv}{\zeta \cdot v \Lambda_\Omega(v)} \right\rceil, \tag{57}$$

where Λ_Ω denotes the Ω-restricted spectral profile (56) *of the chain.*

See Appendix A.1.1 for the proof.

In the next lemma, we state the relationship between the Ω-restricted spectral profile (56) of the Markov chain to its Ω-restricted conductance profile (14).

Lemma 12 *For a Markov chain with state space $\mathcal{X}$ and stationary distribution Π^*, given any measurable set $\Omega \subset \mathcal{X}$, its Ω-restricted spectral profile* (56) *and Ω-restricted conductance profile* (14) *are related as*

$$\Lambda_\Omega(v) \geq \begin{cases} \frac{\Phi_\Omega^2(v)}{2} & \text{for all } v \in \left[0, \frac{\Pi^*(\Omega)}{2}\right] \\ \frac{\Phi_\Omega^2(\Pi^*(\Omega)/2)}{4} & \text{for all } v \in \left(\frac{\Pi^*(\Omega)}{2}, \infty\right). \end{cases} \tag{58}$$

See Appendix A.1.2 for the proof.

Lemma 3 now follows from Lemmas 11 and 12 as well as the definition (15) of $\tilde{\Phi}_\Omega$.

A.1.1. Proof of Lemma 11

We need the following lemma, proved in for the case of finite state Markov chains in Goel et al. (2006), which lower bounds the Dirichlet form in terms of the spectral profile.

Lemma 13 *For any measurable set $\Omega \subset \mathcal{X}$, any non-constant function $g : \mathcal{X} \to \mathbb{R}_+$ such that $g \in L_2(\pi^*)$ and $supp(g) \subset \Omega$, we have*

$$\frac{\mathcal{E}(g,g)}{\mathrm{Var}_{\pi^*}(g)} \geq \frac{1}{2} \Lambda_\Omega \left(\frac{4 \left(\mathbb{E}_{\pi^*}(g)\right)^2}{\mathrm{Var}_{\pi^*}(g)} \right). \tag{59}$$

The proof of Lemma 13 is a straightforward extension of Lemma 2.1 from Goel et al. (2006), which deals with finite state spaces, to the continuous state Markov chain. See the end of Section A.1.1 for the proof.
We are now equipped to prove Lemma 11.

Proof of Lemma 11: We begin by introducing some notations. Recall that for any Markov chain satisfying the smooth chain assumption (16), given an initial distribution μ_0 that admits a density, the distribution of the chain at any step n also admits a density. As a result, we can define the ratio of the density of the Markov chain at the n-th iteration $h_{\mu_0,n} : \mathcal{X} \to \mathbb{R}$ with respect to the target density π^* via the following recursion

$$h_{\mu_0,0}(x) = \frac{\mu_0(x)}{\pi^*(x)} \quad \text{and} \quad h_{\mu_0,n+1}(x) = \frac{\mathcal{T}\left(\pi^* \cdot h_{\mu_0,n}\right)(x)}{\pi^*(x)},$$

where we have used the notation $\mathcal{T}(\mu)(x)$ to denote the density of the distribution $\mathcal{T}(\mu)$ at x. Note that

$$\mathbb{E}_{\pi^*}(h_{\mu_0,n}) = 1 \quad \text{and} \quad \mathbb{E}_{\pi^*}(h_{\mu_0,n} \cdot \mathbf{1}_\Omega) \leq 1 \quad \text{for all } n \geq 0, \tag{60}$$

where $\Omega \subset \mathcal{X}$ is a measurable set.

We also define the quantity $J(n) := \mathrm{Var}_{\pi^*}(h_{\mu_0,n})$ (we prove the existence of this variance below in Step (1)). Note that the L_2-distance between the distribution of the chain at step n and the target distribution is given by

$$d_{2,\pi^*}(\mathcal{T}^n(\mu_0), \Pi^*) = \left(\int_{x \in \mathbb{R}^d} \left(h_{\mu_0,n}(x) - 1\right)^2 \pi^*(x) dx\right)^{1/2} = \mathrm{Var}_{\pi^*}(h_{\mu_0,n}).$$

Consequently, to prove the ϵ-L_2 mixing time bound (57), it suffices to show that for any measurable set $\Omega \subset \mathcal{X}$, with $\Pi^*(\Omega) \geq 1 - \frac{\epsilon^2}{2\beta^2}$, we have

$$J(n) \leq \epsilon^2 \quad \text{for } n \geq \left\lceil \int_{4/\beta}^{8/\epsilon^2} \frac{dv}{\zeta \cdot v \Lambda_\Omega} \right\rceil \tag{61}$$

We now establish the claim (61) via a three step argument: (1) we prove the existence of the variance $J(n)$ for all $n \in \mathbb{N}$, (2) then we derive a recurrence relation for the difference $J(n+1) - J(n)$ in terms of Dirichlet forms that shows the J is a decreasing function, and (3) finally, using an extension of the variance J from natural indices to real numbers, we derive an explicit upper bound on the number of steps taken by the chain until J lies below the required threshold.

Step (1): Using the reversibility (2) of the chain, we find that

$$h_{\mu_0,n+1}(x)dx = \frac{\int_{y \in \mathcal{X}} \Theta(y, dx) h_{\mu_0,n}(y) \pi^*(y) dy}{\pi^*(x)} = \frac{\int_{y \in \mathcal{X}} \Theta(x, dy) h_{\mu_0,n}(y) \pi^*(x) dx}{\pi^*(x)}$$
$$= \int_{y \in \mathcal{X}} \Theta(x, dy) h_{\mu_0,n}(y) dx \tag{62}$$

Applying an induction argument along with the relationship (62) and the initial condition $h_{\mu_0,0}(x) \leq \beta$, we obtain that

$$h_{\mu_0,n}(x) \leq \beta, \quad \text{for all } n \geq 0. \tag{63}$$

As a result, the variances of the functions $h_{\mu_0,0}$ and $h_{\mu_0,n} \cdot \mathbf{1}_\Omega$ under the target density π^* are well-defined and

$$J(n) = \int_{\mathcal{X}} h^2_{\mu_0,n}(x) \pi^*(x) dx - 1 \tag{64}$$

Step (2): We now bound the difference between consecutive variance terms. We have

$$\begin{aligned} J(n) - \mathrm{Var}_{\pi^*}(h_{\mu_0,n} \cdot \mathbf{1}_\Omega) &= \mathrm{Var}_{\pi^*}(h_{\mu_0,n}) - \mathrm{Var}_{\pi^*}(h_{\mu_0,n} \cdot \mathbf{1}_\Omega) \\ &= \int_{x \in \mathcal{X} \setminus \Omega} h^2_{\mu_0,n}(x)\pi^*(x)dx - \left(\int_{x \in \mathcal{X}} h_{\mu_0,n}(x)\pi^*(x)dx \right)^2 \\ &\quad + \left(\int_{x \in \Omega} h_{\mu_0,n}(x)\pi^*(x)dx \right)^2 \\ &\leq \beta^2 \left(1 - \Pi^*(\Omega)\right) \leq \frac{\epsilon^2}{2} =: B, \end{aligned} \tag{65}$$

where the last inequality follows from the fact that Ω satisfies $\Pi^*(\Omega) \geq 1 - \epsilon^2/(2\beta^2)$. Also note the following bound on $J(0)$:

$$J(0) = \int_{x \in \mathcal{X}} \frac{\mu_0(x)^2}{\pi^*(x)} dx - 1 \leq \beta \int_{x \in \mathcal{X}} \mu_0(x)dx - 1 \leq \beta - 1. \tag{66}$$

Define the two step transition kernel $\Theta \circ \Theta$ as

$$\Theta \circ \Theta(y, dz) = \int_{x \in \mathcal{X}} \Theta(y, dx)\Theta(x, dz).$$

We have

$$\begin{aligned} J(n+1) := \mathrm{Var}_{\pi^*}(h_{\mu_0,n+1}) &= \int_{x \in \mathcal{X}} h^2_{\mu_0,n+1}(x)\pi^*(x)dx - 1 \\ &\overset{(i)}{=} \int_{x \in \mathcal{X}} \int_{y \in \mathcal{X}} \Theta(y, dx) h_{\mu_0,n}(y)\pi^*(y)dy \int_{z \in \mathcal{X}} \Theta(x, dz) h_{\mu_0,n}(z) - 1 \\ &= \int_{y,z \in \mathcal{X}^2} \Theta \circ \Theta(y, dz) h_{\mu_0,n}(y) h_{\mu_0,n}(z)\pi^*(y)dy - 1, \end{aligned}$$

where step (i) follows from the relation (62). Using the above expression for $J(n+1)$ and the expression from equation (64) for $J(n)$, we find that

$$\begin{aligned} J(n+1) - J(n) &= \int_{\mathcal{X}^2} \Theta \circ \Theta(y, dz) h_{\mu_0,n}(y) h_{\mu_0,n}(z)\pi^*(y)dy - \int_{\mathcal{X}} h^2_{\mu_0,n}(x)\pi^*(x)dx, \\ &\overset{(a)}{=} -\mathcal{E}_{\Theta \circ \Theta}(h_{\mu_0,n}, h_{\mu_0,n}), \end{aligned} \tag{67}$$

where $\mathcal{E}_{\Theta \circ \Theta}$ is the Dirichlet form (53) with transition probability Θ being replaced by $\Theta \circ \Theta$. We come back to the proof of equality (a) at the end of this paragraph. Assuming it as given at the moment, we proceed further. Since the Markov chain is ζ-lazy, we can relate the two Dirichlet forms $\mathcal{E}_{\Theta \circ \Theta}$ and $\mathcal{E}_\Theta$ as follows: For any $y, z \in \mathcal{X}$ such that $y \neq z$, we have

$$\begin{aligned} \Theta \circ \Theta(y, dz) = \int_{x \in \mathcal{X}} \Theta(y, dx)\Theta(x, dz) &\geq \Theta(y, dy)\Theta(y, dz) + \Theta(y, dz)\Theta(z, dz) \\ &\geq 2\zeta\Theta(y, dz). \end{aligned} \tag{68}$$

We have

$$\begin{aligned} J(n+1) - J(n) = -\mathcal{E}_{\Theta\circ\Theta}(h_{\mu_0,n}, h_{\mu_0,n}) &\overset{(i)}{\leq} -2\zeta\mathcal{E}_{\Theta}(h_{\mu_0,n}, h_{\mu_0,n}) \\ &\overset{(ii)}{\leq} -2\zeta\mathcal{E}_{\Theta}(h_{\mu_0,n}\cdot\mathbf{1}_\Omega, h_{\mu_0,n}\cdot\mathbf{1}_\Omega) \\ &\overset{(iii)}{\leq} -\zeta\,\mathrm{Var}_{\pi^*}(h_{\mu_0,n}\cdot\mathbf{1}_\Omega)\Lambda_\Omega\left(\frac{4\left[\mathbb{E}_{\pi^*}(h_{\mu_0,n}\cdot\mathbf{1}_\Omega)\right]^2}{\mathrm{Var}_{\pi^*}(h_{\mu_0,n}\cdot\mathbf{1}_\Omega)}\right) \\ &\overset{(iv)}{\leq} -\zeta\cdot(J(n)-B)\,\Lambda_\Omega\left(\frac{4}{J(n)-B}\right). \end{aligned} \tag{69}$$

where step (i) follows from inequality (68), step (ii) follows from the fact that Dirichlet forms satisfy $\mathcal{E}_\Theta(h_{\mu_0,n}, h_{\mu_0,n}) \geq \mathcal{E}_\Theta(h_{\mu_0,n}\cdot\mathbf{1}_\Omega, h_{\mu_0,n}\cdot\mathbf{1}_\Omega)$, step (iii) follows from Lemma 13, and finally step (iv) follows from inequality (65) which implies that $\mathrm{Var}_{\pi^*}(h_{\mu_0,n}\cdot\mathbf{1}_\Omega) \geq J(n)-B$, and the fact that the spectral profile Λ_Ω is a non-increasing function.

Proof of equality (a) in equation (67)**:** Since the distribution Π^* is stationary with respect to the kernel Θ, it is also stationary with respect to the two step kernel $\Theta\circ\Theta$. We now prove a more general claim: For any transition kernel K which has stationary distribution Π^* and any measurable function h, the Dirichlet form $\mathcal{E}_K$, defined by replacing Θ with K in equation (53), we have

$$\mathcal{E}_K(h,h) = \int_{\mathcal{X}} h^2(x)\pi^*(x)dx - \int_{\mathcal{X}}\int_{\mathcal{X}} h(x)h(y)K(x,dy)\pi^*(x)dx. \tag{70}$$

Note that invoking this claim with $K = \Theta\circ\Theta$ and $h = h_{\mu_0,n}$ implies step (a) in equation (67). We now establish the claim (70). Expanding the square in the definition (53), we obtain that

$$\begin{aligned} \mathcal{E}_K(h,h) &= \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}} h^2(x)K(x,dy)\pi^*(x)dx + \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}} h^2(y)K(x,dy)\pi^*(x)dx \\ &\qquad - \int_{\mathcal{X}}\int_{\mathcal{X}} h(x)h(y)K(x,dy)\pi^*(x)dx \\ &\overset{(i)}{=} \frac{1}{2}\int_{\mathcal{X}} h^2(x)\pi^*(x)dx + \frac{1}{2}\int_{\mathcal{X}} h^2(x)\pi^*(x)dx - \int_{\mathcal{X}}\int_{\mathcal{X}} h(x)h(y)K(x,dy)\pi^*(x)dx, \end{aligned}$$

where equality (i) follows from the following facts: For the first term, we use the fact that $\int_{\mathcal{X}} K(x,dy) = 1$ since K is a transition kernel, and, for the second term we use the fact that $\int_{\mathcal{X}} K(x,dy)\pi^*(x)dx = \pi^*(y)dy$, since Π^* is the stationary distribution for the kernel K. The claim now follows.

Step (3): Consider the domain extension of the function J from $\mathbb{N}$ to the set of non-negative real numbers $\mathbb{R}_+$ by piecewise linear interpolation. We abuse notation and denote this extension also by J. The extended function J is continuous and is differentiable on the set $\mathbb{R}_+\backslash\mathbb{N}$. Let $n^* \in \mathbb{R}_+ \cup \{\infty\}$ denote the index such that $J(n^*) < B$. Since Λ_Ω is non-increasing and J is non-increasing, we have

$$J'(t) \leq -\zeta\cdot(J(t)-B)\,\Lambda_\Omega\left(\frac{4}{J(t)-B}\right) \quad \text{for all } t \in \mathbb{R}_+\backslash\mathbb{N} \text{ such that } t \leq n^*. \tag{71}$$

Moving the J terms on one side and integrating for $t \leq n^*$, we obtain

$$\int_{J(0)}^{J(t)} \frac{dJ}{(J-B)\cdot \Lambda_\Omega\left(\frac{4}{J-B}\right)} \leq -\zeta t.$$

Using the change of variable $v = 4/(J-B)$, we obtain

$$\zeta t \leq \int_{4/(J(0)-B)}^{4/(J(t)-B)} \frac{dv}{v\Lambda_\Omega(v)} \tag{72}$$

Furthermore, equation (72) implies that for $T \geq \frac{1}{\zeta}\int_{4/\beta}^{8/\epsilon^2} \frac{dv}{v\Lambda_\Omega(v)}$, we have

$$\int_{4/\beta}^{8/\epsilon^2} \frac{dv}{v\Lambda_\Omega(v)} \leq \int_{4/(J(0)-B)}^{4/(J(T)-B)} \frac{dv}{v\Lambda_\Omega(v)}.$$

The bound (66) and the fact that $B = \epsilon^2/2$ imply that $4/(J(0)-B) > 4/\beta$. Using this observation and the fact that $0 \leq \Lambda_\Omega(v) < \infty$ for $v \geq 4/\beta$, we conclude that

$$J(T) \leq B = \frac{\epsilon^2}{2} \text{ or } \frac{4}{J(T)-B} \geq \frac{8}{\epsilon^2} \quad \text{for } T \geq \frac{1}{\zeta}\int_{4/\beta}^{8/\epsilon^2} \frac{dv}{v\Lambda(v)},$$

which implies the claimed bound (61).

Finally, we turn to the proof of Lemma 13.

Proof of Lemma 13: Fix a non-constant function $g : \mathcal{X} \to \mathbb{R}_+$ such that $g \in L_2(\pi^*)$ and $\text{supp}(g) \subset \Omega$. Note that for any constant $c \in \mathbb{R}$, we have

$$\begin{aligned}
\mathcal{E}(g,g) &= \frac{1}{2}\int_{(x,y)\in\mathcal{X}^2} (g(x)-g(y))^2\,\Theta(x,dy)\Pi^*(x)dx \\
&= \frac{1}{2}\int_{(x,y)\in\Omega^2} (g(x)-g(y))^2\,\Theta(x,dy)\Pi^*(x)dx \\
&= \frac{1}{2}\int_{(x,y)\in\Omega^2} ((g(x)-c)-(g(y)-c))^2\,\Theta(x,dy)\Pi^*(x)dx \\
&= \mathcal{E}\left((g-c)\cdot\mathbf{1}_\Omega, (g-c)\cdot\mathbf{1}_\Omega\right).
\end{aligned}$$

Consequently, we obtain that

$$\begin{aligned}
\mathcal{E}(g,g) = \mathcal{E}\left((g-c)\cdot\mathbf{1}_\Omega, (g-c)\cdot\mathbf{1}_\Omega\right) &\geq \mathcal{E}\left((g-c)_+\cdot\mathbf{1}_\Omega, (g-c)_+\cdot\mathbf{1}_\Omega\right) \\
&\overset{(i)}{\geq} \text{Var}_{\pi^*}\left((g-c)_+\cdot\mathbf{1}_\Omega\right) \inf_{f\in c_0^+(\{g>c\}\cap\Omega)} \frac{\mathcal{E}(f,f)}{\text{Var}_{\pi^*}(f)} \\
&\overset{(ii)}{\geq} \text{Var}_{\pi^*}\left((g-c)_+\cdot\mathbf{1}_\Omega\right)\cdot\Lambda_\Omega(\Pi^*(\{g>c\}\cap\Omega)).
\end{aligned} \tag{73}$$

Here $(x)_+ = \max\{0, x\}$ denotes the positive part of x. Inequality (i) follows from the infimum and inequality (ii) follows from the definition (56) of Ω-restricted spectral profile. Additionally, we have

$$\begin{aligned}\mathrm{Var}_{\pi^*}\left((g-c)_+ \cdot \mathbf{1}_\Omega\right) &= \mathbb{E}_{\pi^*}\left((g-c)_+ \cdot \mathbf{1}_\Omega\right)^2 - \left[\mathbb{E}_{\pi^*}\left((g-c)_+ \cdot \mathbf{1}_\Omega\right)\right]^2 \\ &\overset{(i)}{\geq} \mathbb{E}_{\pi^*}(g)^2 - 2(c\Pi^*(\Omega)) \cdot \mathbb{E}_{\pi^*}(g) - \left[\mathbb{E}_{\pi^*}(g)\right]^2 \\ &\geq \mathrm{Var}_{\pi^*}(g) - 2c\mathbb{E}_{\pi^*}(g), \end{aligned} \tag{74}$$

where inequality (i) follows from the fact that

$$(a-b)_+^2 \geq a^2 - 2ab \quad \text{and} \quad (a-b)_+ \leq a, \quad \text{for scalars } a, b \geq 0.$$

Setting $c = \mathrm{Var}_{\pi^*}(g)/4\mathbb{E}_{\pi^*}(g)$, we obtain from equation (74) that

$$\mathrm{Var}_{\pi^*}\left((g-c)_+ \mathbf{1}_\Omega\right) \geq \frac{1}{2}\mathrm{Var}_{\pi^*}(g) \tag{75}$$

Furthermore for any $c > 0$, applying Markov's inequality for the non-negative function $g \cdot \mathbf{1}_\Omega$, we also have $\Pi^*(\{g > c\} \cap \Omega) \leq \Pi^*(\{g > c\}) \leq \left[\mathbb{E}_{\pi^*}(g)\right]/c$. Combing equation (73) and (75), together with the fact that Λ_Ω is non-increasing, we obtain

$$\mathcal{E}(g,g) \geq \frac{1}{2}\mathrm{Var}_{\pi^*}(g) \cdot \Lambda_\Omega\left(\frac{4\left(\mathbb{E}_{\pi^*}(g)\right)^2}{\mathrm{Var}_{\pi^*}(g)}\right),$$

as claimed in the lemma.

A.1.2. Proof of Lemma 12

The proof of the Lemma 12 follows along the lines of Lemma 2.4 in Goel et al. (2006), except that we have to deal with continuous-state transition probability. This technical challenge is the main reason for introducing the restricted conductance profile. At a high level, our argument is based on reducing the problem on general functions to a problem on indicator functions, and then using the definition of the conductance. Similar ideas have appeared in the proof of the Cheeger's inequality (Cheeger, 1969) and the modified log-Sobolev constants (Houdré, 2001).

We split the proof of Lemma 12 in two cases based on whether $v \in [\frac{4}{\beta}, \frac{\Pi^*(\Omega)}{2}]$, referred to as Case 1, or $v \geq \frac{\Pi^*(\Omega)}{2}$, referred to as Case 2.

Case 1: First we consider the case when $v \in [\frac{4}{\beta}, \frac{\Pi^*(\Omega)}{2}]$. First, we define $D^+ : L_2(\pi^*) \to L_2(\pi^*)$ as

$$D^+(g)(x) = \int_{y\in\mathcal{X}} \left(g(x) - g(y)\right)_+ \Theta(x, dy) \,\text{and}\, D^-(g)(x) = \int_{y\in\mathcal{X}} \left(g(x) - g(y)\right)_- \Theta(x, dy),$$

where $(x)_+ = \max\{0, x\}$ and (resp. $(\cdot)_-$) denote the positive and negative part of x respectively. We note that D^+ and D^- satisfy the following co-area formula:

$$\mathbb{E}_{\pi^*} D^+(g) = \int_{-\infty}^{+\infty} \mathbb{E}_{\pi^*} D^+ \mathbf{1}_{g>t} dt. \tag{76a}$$

See Lemma 1 in Houdré (2001) or Lemma 2.4 in Goel et al. (2006) for a proof of the equality (76a). Moreover, given any measurable set $A \subset \mathcal{X}$, scalar t, and function $g \in c_0^+(A \cap \Omega)$, we note that the term $\mathbb{E}_{\pi^*} D^+(\mathbf{1}_{g>t})(x)$ is equal to the flow ϕ (defined in equation (13)) of the level set $G_t = \{x \in \Omega \mid g(x) > t\}$:

$$\mathbb{E}_{\pi^*} D^+(\mathbf{1}_{g>t}) = \int_{x \in G_t} \Theta(x, G_t^c)\pi^*(x)dx = \phi(G_t). \tag{76b}$$

Since $G_t \subset \Omega$, we have

$$\phi(G_t) \geq \Pi^*(G_t) \cdot \inf_{0 \leq \Pi^*(S \cap \Omega) \leq \Pi^*(A \cap \Omega)} \frac{\phi(S)}{\Pi^*(S \cap \Omega)}. \tag{76c}$$

Combining the previous three equations, we find that[9]

$$\begin{aligned}\mathbb{E}_{\pi^*} D^+(g) = \int_{-\infty}^{+\infty} \mathbb{E}_{\pi^*} D^+ \mathbf{1}_{g>t} dt &\geq \int_{-\infty}^{+\infty} \Pi^*(G_t)dt \cdot \inf_{0 \leq \Pi^*(S \cap \Omega) \leq \Pi^*(A \cap \Omega)} \frac{\phi(S)}{\Pi^*(S \cap \Omega)} \\ &= \mathbb{E}_{\pi^*}(g) \cdot \Phi_\Omega(\Pi^*(A \cap \Omega)).\end{aligned}$$

In a similar fashion, we also obtain that

$$\mathbb{E}_{\pi^*} D^-(g) \geq \mathbb{E}_{\pi^*}(g) \cdot \Phi_\Omega(\Pi^*(A \cap \Omega)).$$

Combining these two bounds, we find that

$$\int_{\mathcal{X}} \int_{\mathcal{X}} |g(x) - g(y)| \, \Theta(x, dy)\pi^*(x)dx = \mathbb{E}_{\pi^*} D^+(g) + \mathbb{E}_{\pi^*} D^-(g) \geq 2\mathbb{E}_{\pi^*}(g) \cdot \Phi_\Omega(\Pi^*(A \cap \Omega)).$$

Applying this inequality with the function g^2, we have

$$\begin{aligned}
&2\mathbb{E}_{\pi^*}(g^2) \cdot \Phi_\Omega(\Pi^*(A \cap \Omega)) \\
&\leq \int_{\mathcal{X}} \int_{\mathcal{X}} \left|g^2(x) - g^2(y)\right| \Theta(x, dy)\pi^*(x)dx \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} |g(x) - g(y)| \, |g(x) + g(y)| \, \Theta(x, dy)\pi^*(x)dx \\
&\overset{(i)}{\leq} \left(\int_{\mathcal{X}} \int_{\mathcal{X}} |g(x) - g(y)|^2 \, \Theta(x, dy)\pi^*(x)dx \right)^{1/2} \cdot \left(\int_{\mathcal{X}} \int_{\mathcal{X}} |g(x) + g(y)|^2 \, \Theta(x, dy)\pi^*(x)dx \right)^{1/2} \\
&\overset{(ii)}{\leq} (2\mathcal{E}(g, g))^{1/2} \cdot \left(\int_{\mathcal{X}} \int_{\mathcal{X}} 2\left(g(x)^2 + g(y)^2\right) \Theta(x, dy)\pi^*(x)dx \right)^{1/2} \\
&= (2\mathcal{E}(g, g))^{1/2} \left(4\mathbb{E}_{\pi^*}(g^2)\right)^{1/2}.
\end{aligned}$$

Rearranging the last equation, we obtain that

$$\frac{\mathcal{E}(g, g)}{\mathbb{E}_{\pi^*}(g^2)} \geq \frac{\Phi_\Omega^2(\Pi^*(A \cap \Omega))}{2}. \tag{77}$$

9. Note that this step demonstrates that the continuous state-space treatment is different from the discrete state-space one in Lemma 2.4 of Goel et al. (2006).

In the above sequence of steps, inequality (i) follows from the Cauchy-Schwarz inequality, and inequality (ii) from the definition (53) and the fact that $(a+b)^2 \leq 2(a^2+b^2)$. Taking infimum over $g \in c_0^+(A \cap \Omega)$ in equation (77), we obtain

$$\lambda_\Omega(A) = \inf_{g \in c_0^+(A\cap\Omega)} \frac{\mathcal{E}(g,g)}{\mathrm{Var}_{\pi^*}(g)} \geq \inf_{g \in c_0^+(A\cap\Omega)} \frac{\mathcal{E}(g,g)}{\mathbb{E}_{\pi^*}(g^2)} \geq \frac{\Phi_\Omega^2(\Pi^*(A\cap\Omega))}{2},$$

where the first inequality follows from the fact that $\mathbb{E}_{\pi^*}(g^2) \geq \mathrm{Var}_{\pi^*}(g)$. Given $v \in [0, \frac{\Pi^*(\Omega)}{2}]$, taking infimum over $\Pi^*(A\cap\Omega) \leq v$ on both sides, we conclude the claimed bound for this case:

$$\Lambda_\Omega(v) = \inf_{\Pi^*(A\cap\Omega)\in[0,v]} \lambda_\Omega(A) \geq \inf_{\Pi^*(A\cap\Omega)\in[0,v]} \frac{\Phi_\Omega^2(\Pi^*(A\cap\Omega))}{2} = \frac{\Phi_\Omega^2(v)}{2},$$

where the last equality follows from the fact that the conductance profile Φ_Ω defined in equation (14) is non-increasing over its domain $[0, \frac{\Pi^*(\Omega)}{2}]$.

Case 2: Next, we consider the case when $v \geq \frac{\Pi^*(\Omega)}{2}$. We claim that

$$\Lambda_\Omega(v) \overset{(i)}{\geq} \Lambda_\Omega(\Pi^*(\Omega)) \overset{(ii)}{\geq} \frac{\Lambda_\Omega(\Pi^*(\Omega)/2)}{2} \overset{(iii)}{\geq} \frac{\Phi_\Omega(\Pi^*(\Omega)/2)^2}{4}, \tag{78}$$

where step (i) follows from the fact that the spectral profile Λ is a non-increasing function, and step (iii) from the result of Case 1. Note that the bound from Lemma 12 for this case follows from the bound above. It remains to establish inequality (ii), which we now prove.

Note that given the definition (56), it suffices to establish that

$$\frac{\mathcal{E}(g,g)}{\mathrm{Var}_{\pi^*}(g)} \geq \frac{\Lambda_\Omega(\Pi^*(\Omega)/2)}{2} \quad \text{for all functions } g \in c_0^+(\Omega). \tag{79}$$

Consider any fixed $g \in c_0^+(\Omega)$ and let $\nu \in \mathbb{R}$ be such that

$$\Pi^*(\{g > \nu\} \cap \Omega) = \Pi^*(\{g < \nu\} \cap \Omega) = \frac{\Pi^*(\Omega)}{2}.$$

Using the same argument as in the proof of Lemma 13, we have

$$\begin{aligned}\mathcal{E}(g,g) &= \mathcal{E}((g-\nu)\cdot\mathbf{1}_\Omega, (g-\nu)\cdot\mathbf{1}_\Omega) \\ &\geq \mathcal{E}((g-\nu)_+\cdot\mathbf{1}_\Omega, (g-\nu)_+\cdot\mathbf{1}_\Omega) + \mathcal{E}((g-\nu)_-\cdot\mathbf{1}_\Omega, (g-\nu)_-\cdot\mathbf{1}_\Omega).\end{aligned} \tag{80}$$

We have

$$\mathcal{E}((g-\nu)_+\cdot\mathbf{1}_\Omega, (g-\nu)_+\cdot\mathbf{1}_\Omega) \geq \mathbb{E}_{\pi^*}\left((g-\nu)_+^2\cdot\mathbf{1}_\Omega\right) \cdot \inf_{f\in c_0^+(\{g>\nu\}\cap\Omega)} \frac{\mathcal{E}(f,f)}{\mathbb{E}_{\pi^*} f^2}, \tag{81}$$

and similarly

$$\mathcal{E}((g-\nu)_-\cdot\mathbf{1}_\Omega, (g-\nu)_-\cdot\mathbf{1}_\Omega) \geq \mathbb{E}_{\pi^*}\left((g-\nu)_-^2\cdot\mathbf{1}_\Omega\right) \cdot \inf_{f\in c_0^+(\{g<\nu\}\cap\Omega)} \frac{\mathcal{E}(f,f)}{\mathbb{E}_{\pi^*} f^2}. \tag{82}$$

For $f \in c_0^+(\{g > \nu\} \cap \Omega)$, using Cauchy-Schwarz inequality, we have

$$\mathbb{E}_{\pi^*} f^2 = \int_{x \in \{g>\nu\} \cap \Omega} f(x)^2 \Pi^*(x) dx \geq \frac{\left(\int_{x \in \{g>\nu\} \cap \Omega} |f(x)| \, \Pi^*(x) dx\right)^2}{\Pi^*(\{g > \nu\} \cap \Omega)}$$

Using this bound and noting the ν is chosen such that $\Pi^*(\{g > \nu\} \cap \Omega) = \Pi^*(\Omega)/2$, for $f \in c_0^+(\{g > \nu\} \cap \Omega)$, we have

$$\mathrm{Var}_{\pi^*}(f) = \mathbb{E}_{\pi^*} f^2 - (\mathbb{E}_{\pi^*} f)^2 \geq \mathbb{E}_{\pi^*} f^2 \cdot \left(1 - \frac{\Pi^*(\Omega)}{2}\right). \tag{83}$$

Putting the equations (80), (81), (82) and (83) together, we obtain

$$\begin{aligned}
\mathcal{E}(g, g) &\geq \mathbb{E}_{\pi^*}\left((g-\nu)^2 \cdot \mathbf{1}_\Omega\right) \cdot \left(1 - \frac{\Pi^*(\Omega)}{2}\right) \cdot \inf_{\Pi^*(S) \in [0, \frac{\Pi^*(\Omega)}{2}]} \inf_{f \in c_0^+(S \cap \Omega)} \frac{\mathcal{E}(f, f)}{\mathrm{Var}_{\pi^*}(f)} \\
&= \mathrm{Var}_{\pi^*}(g) \cdot \frac{1}{2} \cdot \Lambda_\Omega(\Pi^*(\Omega)/2).
\end{aligned}$$

which implies the claim (79) and we are done.

A.2. Proof of Lemma 4

The proof of this lemma is similar to the conductance based proof for continuous Markov chains (see, e.g., Lemma 2 in our past work Dwivedi et al. (2018)). In addition to it, we have to deal with the case when target distribution satisfies the logarithmic isoperimetric inequality.

For any set A_1 such that $\Pi^*(A_1 \cap \Omega) \leq \frac{\Pi^*(\Omega)}{2}$, with its complement denoted by $A_2 = \mathcal{X} \backslash A_1$, we have $\Pi^*(A_2 \cap \Omega) \geq \frac{\Pi^*(\Omega)}{2} \geq \Pi^*(A_1 \cap \Omega)$, since $\Pi^*(A_1 \cap \Omega) + \Pi^*(A_2 \cap \Omega) = \Pi^*(\Omega)$. We claim that

$$\int_{x \in A_1} \Theta(x, A_2) \pi^*(x) dx \geq \Pi^*(A_1 \cap \Omega) \cdot \frac{\omega}{4} \cdot \min\left\{1, \frac{\Delta}{16 \psi_{\mathfrak{a}}} \cdot \log^{\mathfrak{a}}\left(1 + \frac{1}{\Pi^*(A_1 \cap \Omega)}\right)\right\}. \tag{84}$$

Note that the claim (20) of Lemma 4 can be directly obtained from the claim (84), by dividing both sides by $\Pi^*(A_1 \cap \Omega)$, taking infimum with respect to A_1 such $\Pi^*(A_1 \cap \Omega) \in (0, v]$ and noting that $\inf_{t \in (0,v]} \log^{\frac{1}{2}}(1 + 1/t) = \log^{\frac{1}{2}}(1 + 1/v)$.

We now prove the claim (84).

Define the following sets,

$$A_1' := \left\{x \in A_1 \cap \Omega \mid \Theta(x, A_2) < \frac{\omega}{2}\right\}, \quad A_2' := \left\{x \in A_2 \cap \Omega \mid \Theta(x, A_1) < \frac{\omega}{2}\right\}, \tag{85}$$

along with the complement $A_3' := \Omega \setminus (A_1' \cup A_2')$. Note that $A_i' \subset \Omega$ for $i = 1, 2, 3$. We split the proof into two distinct cases:

- Case 1: $\Pi^*(A_1') \leq \Pi^*(A_1 \cap \Omega)/2$ or $\Pi^*(A_2') \leq \Pi^*(A_2 \cap \Omega)/2$.
- Case 2: $\Pi^*(A_1') > \Pi^*(A_1 \cap \Omega)/2$ and $\Pi^*(A_2') > \Pi^*(A_2 \cap \Omega)/2$.

Note that these cases are mutually exclusive and exhaustive. We now consider these cases one by one.

Case 1: If we have $\Pi^*(A_1') \leq \Pi^*(A_1 \cap \Omega)/2$, then

$$\Pi^*(A_1 \cap \Omega \setminus A_1') \geq \Pi^*(A_1 \cap \Omega)/2. \tag{86}$$

We have

$$\begin{aligned}\int_{x \in A_1} \Theta(x, A_2)\pi^*(x)dx \geq \int_{x \in A_1 \cap \Omega \setminus A_1'} \Theta(x, A_2)\pi^*(x)dx &\overset{(i)}{\geq} \frac{\omega}{2} \int_{x \in A_1 \cap \Omega \setminus A_1'} \pi^*(x)dx \\ &\overset{(ii)}{\geq} \frac{\omega}{4}\Pi^*(A_1 \cap \Omega),\end{aligned}$$

where inequality (i) follows from the definition of the set A_1' in equation (85) and inequality (ii) follows from equation (86). For the case $\Pi^*(A_2') \leq \Pi^*(A_2 \cap \Omega)/2$, we use a similar argument with the role of A_1 and A_2 exchanged to obtain

$$\int_{x \in A_1} \Theta(x, A_2)\pi^*(x)dx = \int_{x \in A_2} \Theta(x, A_1)\pi^*(x)dx \geq \frac{\omega}{4}\Pi^*(A_2 \cap \Omega).$$

Putting the pieces together for this case, we have established that

$$\int_{x \in A_1} \Theta(x, A_2)\pi^*(x)dx \geq \frac{\omega}{4} \min\left\{\Pi^*(A_1 \cap \Omega), \Pi^*(A_2 \cap \Omega)\right\} = \frac{\omega}{4}\Pi^*(A_1 \cap \Omega). \tag{87}$$

Case 2: We have $\Pi^*(A_1') > \Pi^*(A_1 \cap \Omega)/2$ and $\Pi^*(A_2') > \Pi^*(A_2 \cap \Omega)/2$. We first show that in this case the sets A_1' and A_2' are far away, and then we invoke the logarithmic isoperimetry inequality from Lemma 16.

For any two vectors $u \in A_1'$ and $v \in A_2'$, we have

$$d_{\mathrm{TV}}\left(\mathcal{T}_u, \mathcal{T}_v\right) \geq \Theta(u, A_1) - \Theta(v, A_1) = 1 - \Theta(u, A_2) - \Theta(v, A_1) > 1 - \omega.$$

Consequently, the assumption of the lemma implies that

$$d(A_1', A_2') \geq \Delta. \tag{88}$$

Using the fact that under the stationary distribution, the flow from A_1 to A_2 is equal to that from A_2 to A_1, we obtain

$$\begin{aligned}\int_{x \in A_1} \Theta(x, A_2)\pi^*(x)dx &= \frac{1}{2}\left(\int_{x \in A_1} \Theta(x, A_2)\pi^*(x)dx + \int_{x \in A_2} \Theta(x, A_1)\pi^*(x)dx\right) \\ &\geq \frac{1}{4}\left(\int_{x \in A_1 \cap \Omega \setminus A_1'} \Theta(x, A_2)\pi^*(x)dx + \int_{x \in x \in A_2 \cap \Omega \setminus A_2'} \Theta(x, A_1)\pi^*(x)dx\right) \\ &\geq \frac{\omega}{8}\Pi^*(\Omega \setminus (A_1' \cup A_2')),\end{aligned} \tag{89}$$

where the last inequality follows from the definition of the set A_1' in equation (85). Note that the sets A_1', A_2' and $\mathcal{X} \setminus (A_1' \cup A_2')$ partition $\mathcal{X}$. Using the condition (10d) with the Ω-restricted distribution Π_Ω^* with density π_Ω^* defined as

$$\pi_\Omega^*(x) = \frac{\pi^*(x)\mathbf{1}_\Omega(x)}{\Pi^*(\Omega)},$$

we obtain

$$\begin{aligned}
&\Pi^*(\Omega \setminus (A_1' \cap A_2')) \\
&= \Pi^*(\Omega) \cdot \Pi^*_\Omega(\mathcal{X} \setminus (A_1' \cap A_2')) \\
&\overset{(i)}{\geq} \Pi^*(\Omega) \cdot \frac{d(A_1', A_2')}{2\psi_{\mathfrak{a}}} \cdot \min\left\{\Pi^*_\Omega(A_1'), \Pi^*_\Omega(A_2')\right\} \cdot \log^{\mathfrak{a}}\left(1 + \frac{1}{\min\left\{\Pi^*_\Omega(A_1'), \Pi^*_\Omega(A_2')\right\}}\right) \\
&\overset{(ii)}{\geq} \Pi^*(\Omega) \cdot \frac{\Delta}{4\psi_{\mathfrak{a}}} \min\left\{\Pi^*(A_1 \cap \Omega), \Pi^*(A_2 \cap \Omega)\right\} \cdot \log^{\mathfrak{a}}\left(1 + \frac{2}{\min\left\{\Pi^*(A_1 \cap \Omega), \Pi^*(A_2 \cap \Omega)\right\}}\right) \\
&\geq \frac{1}{2} \cdot \frac{\Delta}{4\psi_{\mathfrak{a}}} \cdot \Pi^*(A_1 \cap \Omega) \cdot \log^{\mathfrak{a}}\left(1 + \frac{1}{\Pi^*(A_1 \cap \Omega)}\right), \qquad (90)
\end{aligned}$$

where step (i) follows from the assumption (10d), step (ii) from the bound (88) and the facts that $\Pi^*_\Omega(A_i') \geq \Pi^*(A_i') \geq \frac{1}{2}\Pi^*(A_i \cap \Omega)$ and that the map $x \mapsto x \log^{\mathfrak{a}}(1 + 1/x)$ is an increasing function for either $\mathfrak{a} = \frac{1}{2}$ or $\mathfrak{a} = 0$. Putting the pieces (89) and (90) together, we conclude that

$$\int_{x \in A_1} \Theta(x, A_2)\pi^*(x)dx \geq \frac{\omega}{16} \cdot \frac{\Delta}{4\psi_{\mathfrak{a}}} \cdot \Pi^*(A_1 \cap \Omega) \cdot \log^{\mathfrak{a}}\left(1 + \frac{1}{\Pi^*(A_1 \cap \Omega)}\right). \qquad (91)$$

Finally, the claim (84) follows from combining the two bounds (87) and (91) from the two separate cases.

A.3. Proofs related to Lemma 6

We now present the proof of the intermediate results related to the HMC chain that were used in the proof of Lemma 6, namely, Lemmas 7, 8, 9 and 10. For simplicity, we adopt following the tensor notation.

Notations for tensor: Let $\mathcal{T} \in \mathbb{R}^{d \times d \times d}$ be a third order tensor. Let $U \in \mathbb{R}^{d \times d_1}$, $V \in \mathbb{R}^{d \times d_2}$, and $W \in \mathbb{R}^{d \times d_3}$ be three matrices. Then the multi-linear form applied on (U, V, W) is a tensor in $\mathbb{R}^{d_1 \times d_2 \times d_3}$:

$$[\mathcal{T}(U, V, W)]_{p,q,r} = \sum_{i,j,k \in [d]} \mathcal{T}_{ijk} U_{ip} V_{jq} W_{kr}.$$

In particular, for the vectors $u, v, w \in \mathbb{R}^d$, the quantity $\mathcal{T}(u, v, w)$ is a real number that depends linearly on u, v, w (tensor analogue of the quantity $u^\top M v$ in the context of matrices and vector). Moreover, the term $\mathcal{T}(u, v, \mathbb{I}_d)$ denotes a vector in $\mathbb{R}^d$ (tensor analogue of the quantity Mv in the context of matrices and vector). Finally, the term $\mathcal{T}(u, \mathbb{I}_d, \mathbb{I}_d)$ represents a matrix in $\mathbb{R}^{d \times d}$.

A.3.1. Proof of Lemma 7

We will prove an equivalent statement: for $K^2\eta^2 \leq \frac{1}{4L}$, there is a matrix $Q(x, y) \in \mathbb{R}^{d \times d}$ with $|||Q|||_2 \leq \frac{1}{8}$ such that

$$\mathbf{J}_x F(x, y) = K\eta\left(\mathbb{I}_d - Q(x, y)\right), \quad \text{for all } x, y \in \mathcal{X}. \qquad (92)$$

Recall from equation (30b) that the intermediate iterate q_k is defined recursively as

$$q_k = F_k(p_0, q_0) = q_0 + k\eta p_0 - \frac{k\eta^2}{2}\nabla f(q_0) - \eta^2 \sum_{j=1}^{k-1} (k-j)\, \nabla f(q_j) \quad \text{for} \quad 1 \le k \le K.$$

Taking partial derivative with respective to the first variable, we obtain

$$\frac{\partial}{\partial p_0} q_k = \mathbf{J}_{p_0} F_k(p_0, q_0) = k\eta \mathbb{I}_d - \eta^2 \sum_{j=1}^{k-1} (k-j) \nabla^2 f_{q_j} \mathbf{J}_{p_0} F_j(p_0, q_0), \tag{93}$$

where $\nabla^2 f_{q_j}$ is the Hessian of f at q_j. We claim that for $1 \le k \le K$, there is a matrix $Q_k \in \mathbb{R}^{d \times d}$ with $\|\!|Q_k|\!\|_2 \le \frac{1}{8}$ such that

$$\mathbf{J}_{p_0} F_k(p_0, q_0) = k\eta \left(\mathbb{I}_d - Q_k\right). \tag{94}$$

Note that substituting $k = K$ in this claim yields the result of the lemma. We now prove the claim (94) using strong induction.

Base case ($k = 1, 2$): For the base case $k = 1, 2$, using equation (93), we have

$$\begin{aligned}
&\mathbf{J}_{p_0} F_1(p_0, q_0) = \eta \mathbb{I}_d, \quad \text{and} \\
&\mathbf{J}_{p_0} F_2(p_0, q_0) = 2\eta \mathbb{I}_d - \eta^2 \nabla^2 f_{q_1} \mathbf{J}_{p_0} F_1(p_0, q_0) = 2\eta \left(\mathbb{I}_d - \frac{\eta^2}{2} \nabla^2 f_{q_1}\right).
\end{aligned}$$

Combining the inequality $\left|\!\left|\!\left| \nabla^2 f_{q_1} \right|\!\right|\!\right|_2 \le L$ from smoothness assumption and the assumed stepsize bound $\eta^2 \le \frac{1}{4L}$ yields

$$\left|\!\left|\!\left| \frac{\eta^2}{2} \nabla^2 f_{q_1} \right|\!\right|\!\right|_2 \le \frac{1}{8}.$$

The statement in equation (94) is verified for $k = 1, 2$.

Inductive step: Assuming that the hypothesis holds for all iterations up to k, we now establish it for iteration $k+1$. We have

$$\begin{aligned}
\mathbf{J}_{p_0} F_{k+1}(p_0, q_0) &= (k+1)\eta \mathbb{I}_d - \eta^2 \sum_{j=1}^{k} (k+1-j) \nabla^2 f_{q_j} \mathbf{J}_{p_0} F_j(p_0, q_0) \\
&\overset{(i)}{=} (k+1)\eta \mathbb{I}_d - \eta^2 \sum_{j=1}^{k} (k+1-j) \nabla^2 f_{q_j} \cdot j\eta \left(\mathbb{I}_d - Q_j\right) \\
&= (k+1)\eta (\mathbb{I}_d - Q_{k+1}),
\end{aligned}$$

where $Q_{k+1} = \frac{\eta^2}{k+1}\sum_{j=1}^{k}(k+1-j)j\nabla^2 f_{q_j}(\mathbb{I}_d - Q_j)$. Equality (i) follows from the hypothesis of the induction. Finally, we verify that the spectral norm of Q_{k+1} is bounded by $\frac{1}{8}$,

$$\begin{aligned}
\|\!|Q_{k+1}|\!\|_2 &\leq \frac{1}{k+1}\sum_{j=1}^{k}\left\|\!\left|\eta^2(k+1-j)j\nabla^2 f_{q_j}\right|\!\right\|_2 \|\!|\mathbb{I}_d - Q_j|\!\|_2 \\
&\overset{(i)}{\leq} \frac{1}{k+1}\sum_{j=1}^{k}\left\|\!\left|\eta^2\frac{K^2}{4}\nabla^2 f_{q_j}\right|\!\right\|_2 \|\!|\mathbb{I}_d - Q_j|\!\|_2 \\
&\overset{(ii)}{\leq} \frac{1}{k+1}\sum_{j=1}^{k}\frac{1}{16}\left(1+\frac{1}{8}\right) \\
&\leq \frac{1}{8}.
\end{aligned}$$

Inequality (i) follows from the inequality $(k+1-j)j \leq \left(\frac{k+1-j+j}{2}\right)^2 \leq \frac{K^2}{4}$. Inequalilty (ii) follows from the assumption $K^2\eta^2 \leq \frac{1}{4L}$ and the hypothesis $\|\!|Q_j|\!\|_2 \leq \frac{1}{8}$. This completes the induction.

A.3.2. Proof of Lemma 8

Recall that the backward mapping G is defined implicitly as

$$x = y + K\eta G(x,y) - \frac{K\eta^2}{2}\nabla f(y) - \eta^2\sum_{k=1}^{K-1}(K-k)\nabla f\left(F_k(G(x,y),y)\right). \tag{95}$$

First we check the derivatives of $F_k(G(x,y),y)$. Since $F_k(G(x,y),y)$ satisfies

$$F_k(G(x,y),y) = y + k\eta G(x,y) - \frac{k\eta^2}{2}\nabla f(y) - \eta^2\sum_{j=1}^{k-1}(k-j)\nabla f(F_j(G(x,y),y)),$$

taking derivative with respect to y, we obtain

$$\begin{aligned}
\frac{\partial}{\partial y}F_k(G(x,y),y) = \mathbb{I}_d + k\eta\mathbf{J}_y G(x,y) - \frac{k\eta^2}{2}\nabla^2 f(y) \\
- \eta^2\sum_{j=1}^{k-1}(k-j)\nabla^2 f(F_j(G(x,y),y))\frac{\partial}{\partial y}F_j(G(x,y),y).
\end{aligned} \tag{96}$$

Using the same proof idea as in the previous lemma, we show by induction that for $1 \leq k \leq K$, there exists matrices $A_k, B_k \in \mathbb{R}^{d\times d}$ with $\|\!|A_k|\!\|_2 \leq \frac{1}{6}$ and $\|\!|B_k|\!\|_2 \leq \frac{1}{8}$ such that

$$\frac{\partial}{\partial y}F_k(G(x,y),y) = (\mathbb{I}_d - A_k) + k\eta\left(\mathbb{I}_d - B_k\right)\mathbf{J}_y G(x,y). \tag{97}$$

Case $k=1$: The case $k=1$ can be easily checked according to equation (96), we have

$$\frac{\partial}{\partial y}F_1\left(G(x,y),y\right) = \mathbb{I}_d - \frac{\eta^2}{2}\nabla^2 f(y) + \eta\mathbf{J}_y G(x,y)$$

It is sufficient to set $A_1 = \frac{\eta^2}{2}\nabla^2 f(y)$ and $B_1 = 0$.

Case k to $k+1$: Assume the statement is verified until $k \geq 1$. For $k+1 \leq K$, according to equation (96), we have

$$
\begin{aligned}
&\frac{\partial}{\partial y} F_{k+1}(G(x,y),y) \\
&= \mathbb{I}_d + (k+1)\eta \mathbf{J}_y G(x,y) - \frac{(k+1)\eta^2}{2}\nabla^2 f(y) - \eta^2 \sum_{j=1}^{k}(k+1-j)\nabla^2 f(F_j(G(x,y),y))\frac{\partial}{\partial y}F_j(G(x,y),y) \\
&= \mathbb{I}_d - \frac{(k+1)\eta^2}{2}\nabla^2 f(y) + (k+1)\eta \mathbf{J}_y G(x,y) \\
&\quad - \eta^2 \sum_{j=1}^{k}(k+1-j)\nabla^2 f(F_j(G(x,y),y))\left((\mathbb{I}_d - A_j) + j\eta\left(\mathbb{I}_d - B_j\right)\mathbf{J}_y G(x,y)\right) \\
&= \mathbb{I}_d - \frac{(k+1)\eta^2}{2}\nabla^2 f(y) - \eta^2 \sum_{j=1}^{k}(k+1-j)\nabla^2 f(F_j(G(x,y),y))(\mathbb{I}_d - A_j) \\
&\quad + (k+1)\eta \mathbf{J}_y G(x,y) - \eta^2 \sum_{j=1}^{k}(k+1-j)\nabla^2 f(F_j(G(x,y),y))\left(j\eta\left(\mathbb{I}_d - B_j\right)\mathbf{J}_y G(x,y)\right)
\end{aligned}
$$

To conclude, it suffices to note the following values of A_{k+1} and B_{k+1}:

$$
\begin{aligned}
A_{k+1} &= \frac{(k+1)\eta^2}{2}\nabla^2 f(y) + \eta^2 \sum_{j=1}^{k}(k+1-j)\nabla^2 f(F_j(G(x,y),y))(\mathbb{I}_d - A_j), \quad \text{and} \\
B_{k+1} &= \frac{1}{k+1}\eta^2 \sum_{j=1}^{k}(k+1-j)j\nabla^2 f(F_j(G(x,y),y))\left(\mathbb{I}_d - B_j\right).
\end{aligned}
$$

We now have the following operator norm bounds:

$$
\begin{aligned}
\|\!| A_{k+1} |\!\|_2 &\leq \frac{k+1}{2}\eta^2 L + \eta^2 \sum_{j=1}^{k}(k+1-j)L(1+\frac{1}{6}) \leq \frac{7}{12}(k+1)^2\eta^2 L \leq \frac{1}{6}, \quad \text{and} \\
\|\!| B_{k+1} |\!\|_2 &\leq \frac{1}{k+1}\eta^2(1+\frac{1}{8})L\sum_{j=1}^{k}(k+1-j)j = \frac{9}{8\cdot 6}k(k-1)\eta^2 L \leq \frac{1}{8}.
\end{aligned}
$$

This concludes the proof of equation (97). As a particular case, for $k = K$, we observe that

$$F_K\left(G(x,y),y\right) = x.$$

Plugging it into equation (97), we obtain that

$$\mathbf{J}_y G(x,y) = \frac{1}{K\eta}\left(\mathbb{I}_d - B_K\right)^{-1}\left(\mathbb{I}_d - A_K\right) \quad \Longrightarrow \quad \|\!| \mathbf{J}_y G(x,y) |\!\|_2 \leq \frac{4}{3K\eta}.$$

Plugging the bound on $\|\!| \mathbf{J}_y G(x,y) |\!\|_2$ back to equation (97) for other k, we obtain

$$\left\|\!\left| \frac{\partial}{\partial y} F_k(G(x,y),y) \right|\!\right\|_2 \leq 3.$$

This concludes the proof of Lemma 8.

A.3.3. Proof of Lemma 9

Recall that the backward mapping G is defined implicitly as

$$x = y + K\eta G(x,y) - \frac{K\eta^2}{2}\nabla f(y) - \eta^2 \sum_{k=1}^{K-1}(K-k)\nabla f\left(F_k(G(x,y),y)\right). \tag{98}$$

First we check the derivatives of $F_k(G(x,y),y)$. Since $F_k(G(x,y),y)$ satisfies

$$F_k(G(x,y),y) = y + k\eta G(x,y) - \frac{k\eta^2}{2}\nabla f(y) - \eta^2 \sum_{j=1}^{k-1}(k-j)\nabla f(F_j(G(x,y),y)),$$

we have

$$\frac{\partial}{\partial x}F_k(G(x,y),y) = k\eta \mathbf{J}_x G(x,y) - \eta^2 \sum_{j=1}^{k-1}(k-j)\nabla^2 f(F_j(G(x,y),y))\frac{\partial}{\partial x}F_j(G(x,y),y). \tag{99}$$

Similar to the proof of equation (94), we show by induction (proof omitted) that for $1 \le k \le K$, there exists matrices $\tilde{Q}_k \in \mathbb{R}^{d\times d}$ with $\left\|\tilde{Q}_k\right\|_2 \le \frac{1}{2}$ such that

$$\frac{\partial}{\partial x}F_k(G(x,y),y) = k\eta\left(\mathbb{I}_d - \tilde{Q}_k\right)\mathbf{J}_x G(x,y). \tag{100}$$

Then, by taking another derivative with respect to y_i in equation (99), we obtain

$$\begin{aligned}\frac{\partial\partial}{\partial x \partial y_i}F_k(G(x,y),y) = k\eta \mathbf{J}_{xy_i}G(x,y) \\ - \eta^2\sum_{j=1}^{k-1}(k-j)\Bigg\{ \quad \nabla^3 f_{F_j(G(x,y),y)}\left(\frac{\partial F_j(G(x,y),y)}{\partial y_i}, \mathbb{I}_d, \mathbb{I}_d\right)\frac{\partial}{\partial x}F_j(G(x,y),y) \\ + \nabla^2 f_{F_j(G(x,y),y)}\frac{\partial\partial}{\partial x\partial y_i}F_j(G(x,y),y)\Bigg\}\end{aligned} \tag{101}$$

Now we show by induction that for $1 \le k \le K$, for any $\alpha \in \mathbb{R}^d$, we have

$$\begin{aligned}\left\|\sum_{i=1}^d \alpha_i\left(\frac{\partial\partial}{\partial x\partial y_i}F_k(G(x,y),y)\mathbf{J}_x G(x,y)^{-1}\right)\right\|_2 \le 2k\eta\left\|\sum_{i=1}^d \alpha_i\left(\mathbf{J}_{xy_i}G(x,y)\mathbf{J}_x G(x,y)^{-1}\right)\right\|_2 \\ + 2\left\|\alpha\right\|_2 k^3\eta^3 L_{\mathrm{H}}.\end{aligned} \tag{102}$$

Case $k = 1$: We first examine the case $k = 1$. According to equation (101), we have

$$\sum_{i=1}^d \alpha_i\left(\frac{\partial\partial}{\partial x\partial y_i}F_1(G(x,y),y)\mathbf{J}_x G(x,y)^{-1}\right) = \eta\sum_{i=1}^d \alpha_i\left(\mathbf{J}_{xy_i}G(x,y)\mathbf{J}_x G(x,y)^{-1}\right).$$

The statement in equation (102) is easily verified for $k = 1$.

Case k to $k+1$: Assume the statement (102) is verified until k. For $k+1 \leq K$, according to equation (101), we have

$$\begin{aligned}
&\sum_{i=1}^{d} \alpha_i \left(\frac{\partial\partial}{\partial x \partial y_i} F_{k+1}(G(x,y),y) \mathbf{J}_x G(x,y)^{-1} \right) \\
&= (k+1)\eta \sum_{i=1}^{d} \alpha_i \left(\mathbf{J}_{xy_i} G(x,y) \mathbf{J}_x G(x,y)^{-1} \right) \\
&\quad - \eta^2 \sum_{j=1}^{k} (k+1-j) \left\{ \nabla^3 f_{F_j(G(x,y),y)} \left(\sum_{i=1}^{d} \alpha_i \frac{\partial F_j(G(x,y),y)}{\partial y_i}, \mathbb{I}_d, \mathbb{I}_d \right) \frac{\partial}{\partial x} F_j(G(x,y),y) \mathbf{J}_x G(x,y)^{-1} \right\} \\
&\quad - \eta^2 \sum_{j=1}^{k} (k+1-j) \nabla^2 f_{F_j(G(x,y),y)} \sum_{i=1}^{d} \alpha_i \left(\frac{\partial\partial}{\partial x \partial y_i} F_j(G(x,y),y) \mathbf{J}_x G(x,y)^{-1} \right).
\end{aligned}$$

In the last equality, we have used the fact that $\nabla^3 f_{F_j(G(x,y),y)}$ is a multilinear form to enter the coefficients α_i in the tensor. Let

$$M_\alpha = \left\| \sum_{i=1}^{d} \alpha_i \left(\mathbf{J}_{xy_i} G(x,y) \mathbf{J}_x G(x,y)^{-1} \right) \right\|_2.$$

Applying the hypothesis of the induction, we obtain

$$\begin{aligned}
&\left\| \sum_{i=1}^{d} \alpha_i \left(\frac{\partial\partial}{\partial x \partial y_i} F_{k+1}(G(x,y),y) \mathbf{J}_x G(x,y)^{-1} \right) \right\|_2 \\
&\overset{(i)}{\leq} (k+1)\eta M_\alpha + \eta^2 \sum_{j=1}^{k} 4(k+1-j) j L_{\mathrm{H}} \|\alpha\|_2 + \eta^2 \sum_{j=1}^{k} (k+1-j) L \left(2j\eta M + 2\|\alpha\|_2 j^3 \eta^3 L_{\mathrm{H}} \right) \\
&\leq 2(k+1)\eta M_\alpha + 2\|\alpha\|_2 (k+1)^3 \eta^3 L_{\mathrm{H}}.
\end{aligned}$$

The first inequality (i) used the second part of Lemma 8 to bound $\left\| \frac{\partial}{\partial} F_k(G(x,y),y) \right\|_2$. This completes the induction. As a particular case for $k = K$, we note that

$$F_K(G(x,y),y) = F(G(x,y),y) = x,$$

and equation (101) for $k = K$ gives

$$\begin{aligned}
0 &= K\eta \mathbf{J}_{xy_i} G(x,y) \\
&\quad - \eta^2 \sum_{j=1}^{K-1} (K-j) \left\{ \nabla^3 f_{F_j(G(x,y),y)} \left(\frac{\partial F_j(G(x,y),y)}{\partial y_i}, \mathbb{I}_d, \mathbb{I}_d \right) \frac{\partial}{\partial x} F_j(G(x,y),y) \right. \\
&\quad \left. + \nabla^2 f_{F_j(G(x,y),y)} \frac{\partial\partial}{\partial x \partial y_i} F_j(G(x,y),y) \right\}.
\end{aligned}$$

Using the bound in equation (102), we have

$$K\eta \left\| \sum_{i=1}^{d} \alpha_i \mathbf{J}_{xy_i} G(x,y) \mathbf{J}_x G(x,y)^{-1} \right\|_2 \leq \|\alpha\|_2 K^3 \eta^3 L_{\mathrm{H}} + \frac{1}{2} K\eta \left\| \sum_{i=1}^{d} \alpha_i \mathbf{J}_{xy_i} G(x,y) \mathbf{J}_x G(x,y)^{-1} \right\|_2.$$

Hence, we obtain

$$\operatorname{trace}\left(\sum_{i=1}^{d}\alpha_i \mathbf{J}_{xy_i}G(x,y)\mathbf{J}_x G(x,y)^{-1}\right) \le 2d\left\|\alpha\right\|_2 K^2\eta^2 L_{\mathrm{H}}.$$

This is valid for any $\alpha \in \mathbb{R}^d$, as a consequence, we have

$$\left\|\begin{bmatrix}\operatorname{trace}\left([\mathbf{J}_x G(x,q_0)]^{-1}\mathbf{J}_{xy_1}G(x,q_0)\right)\\ \vdots \\ \operatorname{trace}\left([\mathbf{J}_x G(x,q_0)]^{-1}\mathbf{J}_{xy_d}G(x,q_0)\right)\end{bmatrix}\right\|_2 \le 2dK^2\eta^2 L_{\mathrm{H}}.$$

This concludes the proof of Lemma 9.

A.3.4. Proof of Lemma 10

We first show equation (34b) by induction. Then equation (34a) is a direct consequence of equation (34b) by summing k terms together.

Case $k = 0$: We first examine the case $k = 0$. According to the definition of F_k in equation (30b), we have

$$F_1(p_0,q_0) = q_0 + \eta p_0 - \frac{\eta^2}{2}\nabla f(q_0).$$

Then the case $k = 0$ is verified automatically via triangle inequality,

$$\left\|F_1(p_0,q_0) - q_0\right\|_2 \le \eta\left\|p_0\right\|_2 + \frac{\eta^2}{2}\left\|\nabla f(q_0)\right\|_2.$$

Case k to $k+1$: Assume that the statement is verified until $k \ge 0$. For $k+1$, using F_j as the shorthand for $F_j(p_0,q_0)$, we obtain

$$\begin{aligned}&F_{k+2} - F_{k+1}\\ =&\eta p_0 - \frac{\eta^2}{2}\nabla f(q_0) - \eta^2\sum_{j=1}^{k+1}\nabla f(F_j).\end{aligned}$$

Taking the norm, we have

$$\begin{aligned}\left\|F_{k+2} - F_{k+1}\right\|_2 &\le \eta\left\|p_0\right\|_2 + \frac{(2k+3)\eta^2}{2}\left\|\nabla f(q_0)\right\|_2 + \eta^2\sum_{j=1}^{k+1}\left\|\nabla f(F_j) - \nabla f(q_0)\right\|_2\\ &\overset{(i)}{\le} \eta\left\|p_0\right\|_2 + \frac{(2k+3)\eta^2}{2}\left\|\nabla f(q_0)\right\|_2 + \eta^2\sum_{j=1}^{k+1}\sum_{l=0}^{j}\left\|\nabla f(F_{l+1}) - \nabla f(F_l)\right\|_2\\ &\overset{(ii)}{\le} \eta\left\|p_0\right\|_2 + \frac{(2k+3)\eta^2}{2}\left\|\nabla f(q_0)\right\|_2 + \eta^2 L\sum_{j=1}^{k+1}\sum_{l=0}^{j}\left\|F_{l+1} - F_l\right\|_2\\ &\overset{(iii)}{\le} \eta\left\|p_0\right\|_2 + \frac{(2k+3)\eta^2}{2}\left\|\nabla f(q_0)\right\|_2 + \eta^2 L\sum_{j=1}^{k+1}\sum_{l=0}^{j}\left(2\eta\left\|p\right\|_2 + 2(l+1)\eta^2\left\|\nabla f(q_0)\right\|_2\right)\\ &\overset{(iv)}{\le} 2\eta\left\|p_0\right\|_2 + (2k+2)\eta^2\left\|\nabla f(q_0)\right\|_2.\end{aligned}$$

Inequality (i) uses triangular inequality. Inequality (ii) uses L-smoothness. Inequality (iii) applies the hypothesis of the induction and inequalities relies on the condition $K^2\eta^2 \leq \frac{1}{4L}$. This completes the induction.

Appendix B. Proof of Corollary 2

In order to prove Corollary 2, we first state a more general corollary of Theorem 1 that does not specify the explicit choice of step size η and leapfrog steps K. Then we specify two choices of the initial distribution μ_0 and hyper-parameters (K, η) to obtain part (a) and part (b) of Corollary 2.

Corollary 14 *Consider an (L, L_H, m)-strongly log-concave target distribution Π^* (cf. Assumption (B)). Fix $s = \frac{\epsilon^2}{2\beta}$. Then the $\frac{1}{2}$-lazy HMC algorithm with initial distribution $\mu_\dagger = \mathcal{N}(x^*, \frac{1}{L}\mathbb{I}_d)$, step size η and leapfrog steps K chosen under the condition*

$$\eta^2 \leq \frac{1}{cL} \min\left\{ \frac{1}{K^2 d^{\frac{1}{2}}}, \frac{1}{K^2 d^{\frac{2}{3}}} \frac{L}{L_H^{\frac{2}{3}}}, \frac{1}{K d^{\frac{1}{2}}}, \frac{1}{K^{\frac{2}{3}} d^{\frac{2}{3}} \kappa^{\frac{1}{3}} r(s)^{\frac{2}{3}}}, \frac{1}{K d^{\frac{1}{2}} \kappa^{\frac{1}{2}} r(s)}, \frac{1}{K^{\frac{2}{3}} d} \frac{L}{L_H^{\frac{2}{3}}}, \frac{1}{K^{\frac{4}{3}} d^{\frac{1}{2}} \kappa^{\frac{1}{2}} r(s)} \left(\frac{L}{L_H^{\frac{2}{3}}}\right)^{\frac{1}{2}} \right\} \tag{103}$$

satisfies the mixing time bounds

$$\tau_2^{HMC}(\epsilon; \mu_0) \leq c \cdot \max\left\{ \log \beta, \frac{1}{K^2\eta^2 m} \log\left(\frac{d \log \kappa}{\epsilon}\right) \right\}.$$

Proof of part (a) in Corollary 2: Taking the hyper-parameters $K = d^{\frac{1}{4}}$ and $\eta = \eta_{\text{warm}}$ in equation (11b), we verify that η satisfies the condition (103). Given the warmness parameter $\beta = O\left(\exp\left(d^{\frac{2}{3}}\kappa\right)\right)$, we have

$$\frac{1}{K^2\eta^2 m} \geq \log(\beta).$$

Plugging in the choice of K and η into Corollary 14, we obtain the desired result.

Proof of part (b) in Corollary 2: We notice that the initial distribution $\mu_\dagger = \mathcal{N}(x^\star, \frac{1}{L}\mathbb{I}_d)$ is $\kappa^{d/2}$-warm (see Corollary 1 in Dwivedi et al. (2018)). It is sufficient to plug in the hyper-parameters $K = \kappa^{\frac{3}{4}}$ and $\eta = \eta_{\text{feasible}}$ into Corollary 14 to obtain the desired result.

Now we turn back to prove Corollary 14. In order to prove Corollary 14, we require the the following lemma, which relates a (L, L_{H}, m)-strongly-logconcave target distribution to a regular target distribution.

Lemma 15 *An (L, L_H, m)-strongly log-concave distribution is $(L, L_H, s, \psi_{\frac{1}{2}}, M)$-general with high mass set $\Omega = \mathcal{R}_s$, log-isoperimetric constant $\psi_{\frac{1}{2}} = m^{-\frac{1}{2}}$ and $M = L\left(\frac{d}{m}\right)^{\frac{1}{2}} r(s)$, where the radius is defined in equation* (11a) *and the convex measurable set $\mathcal{R}_s$ defined in equation* (29).

Taking Lemma 15 as given, Corollary 14 is a direct consequence of Theorem 1 by plugging the specific values of $(\Omega, \psi_{\frac{1}{2}}, M)$ as a function of strong convexity parameter m. The optimal choices of step-size η and leapfrog steps K in Corollary 14 are discussed in Appendix D.1.

We now proceed to prove Lemma 15.

B.1. Proof of Lemma 15

We now prove Lemma 15, which shows that any (L, L_{H}, m)-strongly-logconcave target distribution is in fact $(L, L_{\mathrm{H}}, s, \psi_{\frac{1}{2}}, M)$-regular.

First, we set Ω to $\mathcal{R}_s$ as defined in equation (29). It is known that this ball has probability under the target distribution lower bounded as $\Pi^*(\mathcal{R}_s) \geq 1 - s$ (e.g. Lemma 1 in the paper Dwivedi et al. (2018)). Second, the gradient bound is a consequence of the bounded domain. For any $x \in \mathcal{R}_s$, we have

$$\|\nabla f(x)\|_2 = \|\nabla f(x) - \nabla f(x^\star)\|_2 \leq L \|x - x^\star\|_2 \leq L \left(\frac{d}{m}\right)^{\frac{1}{2}} r(s). \tag{104}$$

Third, we make use of a logarithmic isoperimetric inequality for log-concave distribution. We note that the logarithmic isoperimetric inequality has been introduced in Kannan et al. (2006) for the uniform distribution on convex body and in Lee and Vempala (2018b) for log-concave distribution with a diameter. We extend this inequality to strongly log-concave distribution on $\mathbb{R}^d$ following a similar road-map and provide explicit constants.

Improved logarithmic isoperimetric inequality We now state the improved logarithmic isoperimetric inequality for strongly log-concave distributions.

Lemma 16 *Let γ denote the density of the standard Gaussian distribution $\mathcal{N}\left(0, \sigma^2 \mathbb{I}_d\right)$, and let Π^* be a distribution with density $\pi^* = q \cdot \gamma$, where q is a log-concave function. Then for any partition S_1, S_2, S_3 of $\mathbb{R}^d$, we have*

$$\Pi^*(S_3) \geq \frac{d(S_1, S_2)}{2\sigma} \min\left\{\Pi^*(S_1), \Pi^*(S_2)\right\} \log^{\frac{1}{2}} \left(1 + \frac{1}{\min\left\{\Pi^*(S_1), \Pi^*(S_2)\right\}}\right). \tag{105}$$

See Appendix B.2 for the proof.

Taking Lemma 16 as given for the moment, we turn to prove the logarithmic isoperimetric inequality for the Ω-restricted distribution Π^*_Ω with density

$$\pi^*_\Omega(x) = \frac{\pi^*(x) \mathbf{1}_\Omega(x)}{\Pi^*(\Omega)}.$$

Since f is m-strongly convex, the function $x \to f(x) - \frac{m}{2}\|x - x^\star\|_2^2$ is convex. Noting that the class of log-concave function is closed under multiplication and that the indicator function $\mathbf{1}_\Omega$ is log-concave, we conclude that the restricted density π^*_Ω can be expressed as a product of a log-concave density and the density of the Gaussian distribution $\mathcal{N}(x^\star, \frac{1}{m}\mathbb{I}_d)$. Applying Lemma 16 with $\sigma = \left(\frac{1}{m}\right)^{\frac{1}{2}}$, we obtain the desired logarithmic isoperimetric inequality with $\psi_{\frac{1}{2}} = \left(\frac{1}{m}\right)^{\frac{1}{2}}$, which concludes the proof of Lemma 15.

B.2. Proof of Lemma 16

The main tool for proving general isoperimetric inequalities is the localization lemma introduced by Lovász and Simonovits (1993). Similar result for the infinitesimal version of equation (105) have appeared as Theorem 1.1 in Ledoux (1999) and Theorem 30 in Lee and Vempala (2018b). Intuitively, the localization lemma reduces a high-dimensional isoperimetric inequality to a one-dimensional inequality which is much easier to verify directly. In a few key steps, the proof follows a similar road map as the proof of logarithmic Cheeger inequality (Kannan et al., 2006).

We first state an additional lemma that comes in handy for the proof.

Lemma 17 *Let γ be the density of the one-dimensional Gaussian distribution $\mathcal{N}\left(\nu, \sigma^2\right)$ with mean ν and variance σ^2. Let ρ be a one-dimensional distribution with density given by $\rho = q \cdot \gamma$, where q is a log-concave function supported on $[0,1]$. Let J_1, J_2, J_3 partition $[0,1]$, then*

$$\rho(J_3) \geq \frac{d(J_1, J_2)}{2\sigma} \min\left\{\rho(J_1), \rho(J_2)\right\} \log^{\frac{1}{2}}\left(1 + \frac{1}{\min\left\{\rho(J_1), \rho(J_2)\right\}}\right). \tag{106}$$

See Appendix B.3 for the proof.

We now turn to proving Lemma 16 via contradiction: We assume that the claim (105) is not true for some partition, and then using well known localization techniques, we construct a one-dimensional distribution that violates Lemma 17 resulting in a contradiction.

Suppose that there exists a partition S_1, S_2, S_3 of $\mathbb{R}^d$, such that

$$\Pi^*(S_3) < \frac{d(S_1, S_2)}{2\sigma} \min\left\{\Pi^*(S_1), \Pi^*(S_2)\right\} \log^{\frac{1}{2}}\left(1 + \frac{1}{\min\left\{\Pi^*(S_1), \Pi^*(S_2)\right\}}\right). \tag{107}$$

Let $\nu > 0$ denote a sufficiently small number (to be specified exactly later), such that $\nu < \min\left\{\Pi^*(S_1), \Pi^*(S_2)\right\}$.

We now explain the construction of the one-dimensional density that is crucial for the rest of the argument. We define two functions $g : \mathcal{X} \to \mathbb{R}$ and $h : \mathcal{X} \to \mathbb{R}$ as follows

$$g(x) = \frac{\pi^*(x) \cdot \mathbf{1}_{S_1}(x)}{\Pi^*(S_1) - \nu} - \pi^*(x) \quad \text{and} \quad h(x) = \frac{\pi^*(x) \cdot \mathbf{1}_{S_2}(x)}{\Pi^*(S_2) - \nu} - \pi^*(x).$$

Clearly, we have

$$\int_{\mathcal{X}} g(x)dx > 0 \quad \text{and} \quad \int_{\mathcal{X}} h(x)dx > 0.$$

By the localization lemma (Lemma 2.5 in Lovász and Simonovits (1993); see the corrected form stated as Lemma 2.1 in Kannan et al. (1995)), there exist two points $a \in \mathbb{R}^d, b \in \mathbb{R}^d$ and a linear function $l : [0,1] \to \mathbb{R}_+$, such that

$$\int_0^1 l(t)^{d-1} g\left((1-t)a + tb\right) dt > 0 \quad \text{and} \quad \int_0^1 l(t)^{d-1} h\left((1-t)a + tb\right) dt > 0. \tag{108}$$

Define the one-dimensional density $\rho : [0,1] \to \mathbb{R}^+$ and the sets $J_i, i \in \{1,2,3\}$ as follows:

$$\rho(t) = \frac{l(t)^{d-1}\pi^*\left((1-t)a+tb\right)}{\int_0^1 l(u)^{d-1}\pi^*\left((1-u)a+ub\right)du}, \quad \text{and} \tag{109}$$

$$J_i = \{t \in [0,1] \mid (1-t)a+tb \in S_i\} \quad \text{for } i \in \{1,2,3\}. \tag{110}$$

We now show how the hypothesis (107) leads to a contradiction for the density ρ. Plugging in the definiton of g and h into equation (108), we find that

$$\rho(J_1) > \Pi^*(S_1) - \nu \quad \text{and} \quad \rho(J_2) > \Pi^*(S_2) - \nu.$$

Since J_1, J_2, J_3 partition $[0,1]$, it follows that

$$\rho(J_3) < \Pi^*(S_3) + 2\nu.$$

Since the function $x \mapsto x\log^{\frac{1}{2}}(1+1/x)$ is monotonically increasing on $[0,1]$, we have

$$\begin{aligned}
&\frac{d(S_1,S_2)}{2\sigma}\min\left\{\rho(J_1),\rho(J_2)\right\}\log^{\frac{1}{2}}\left(1+\frac{1}{\min\left\{\rho(J_1),\rho(J_2)\right\}}\right) - \rho(J_3) \\
&\quad \geq \frac{d(S_1,S_2)}{2\sigma}\min\left\{\left(\rho(S_1)-\nu\right),\left(\rho(S_2)-\nu\right)\right\}\cdot \\
&\qquad\qquad \log^{\frac{1}{2}}\left(1+\frac{1}{\min\left\{\left(\rho(S_1)-\nu\right),\left(\rho(S_2)-\nu\right)\right\}}\right) - \left(\rho(S_3)+2\nu\right)
\end{aligned}$$

The hypothesis (107) of the contradiction implies that we can find ν sufficiently small such that the RHS in the inequality above will be strictly positive. Consequently, we obtain

$$\frac{d(S_1,S_2)}{2\sigma}\min\left\{\rho(J_1),\rho(J_2)\right\}\log^{\frac{1}{2}}\left(1+\frac{1}{\min\left\{\rho(J_1),\rho(J_2)\right\}}\right) > \rho(J_3). \tag{111}$$

Additionally, for $t_1 \in J_1, t_2 \in J_2$, we have $(1-t_1)a+t_1b \in S_1$ and $(1-t_2)a+t_2b \in S_2$. As a result, we have

$$|t_1-t_2| = \frac{1}{\|b-a\|_2}\left\|[(1-t_1)a+t_1b]-[(1-t_2)a+t_2b]\right\|_2 \geq \frac{1}{\|b-a\|_2}d(S_1,S_2),$$

which implies that

$$d(J_1,J_2) \geq \frac{1}{\|b-a\|_2}d(S_1,S_2). \tag{112}$$

Combining equations (111) and (112), we obtain that

$$\frac{\|b-a\|_2\cdot d(J_1,J_2)}{2\sigma}\min\left\{\rho(J_1),\rho(J_2)\right\}\log^{\frac{1}{2}}\left(1+\frac{1}{\min\left\{\rho(J_1),\rho(J_2)\right\}}\right) > \rho(J_3), \tag{113}$$

which contradicts Lemma 17. Indeed, this contradiction is immediate once we note that the new density ρ can also be written as a product of log-concave density and a Gaussian density with variance $\frac{\sigma^2}{\|b-a\|_2^2}$.

B.3. Proof of Lemma 17

We split the proof into three cases. Each one is more general than the previous one. First, we consider the case when q is a constant function on $[0,1]$ and the sets J_1, J_2, J_3 are all intervals. In the second case, we consider a general log-concave q supported on $[0,1]$ while we still assume that the sets J_1, J_2, J_3 are all intervals. Finally, in the most general case, we consider a general log-concave q supported on $[0,1]$ and J_1, J_2, J_3 consist of an arbitrary partition of $[0,1]$. The proof idea follows roughly that of Theorem 4.6 in Kannan et al. (2006).

Our proof makes use of the Gaussian isoperimetric inequality which we now state (see e.g., equation (1.2) in Bobkov (1999)): Let Γ denote the standard univariate Gaussian distribution and let ϕ_Γ and Φ_Γ^{-1} denote its density and inverse cumulative distribution function respectively. Given a measurable set $A \subset \mathbb{R}$, define its Γ-perimeter $\Gamma^+(A)$ as

$$\Gamma^+(A) = \liminf_{h \to 0^+} \frac{\Gamma(A+h) - \Gamma(A)}{h},$$

where $A + h = \{t \in \mathbb{R} \mid \exists a \in A, |t-a| < h\}$ denotes an h-neighborhood of A. Then, we have

$$\Gamma^+(A) \geq \phi_\Gamma(\Phi_\Gamma^{-1}(\Gamma(A))), \tag{114}$$

Furthermore, standard Gaussian tail bounds[10] estimate imply that

$$\phi_\Gamma(\Phi_\Gamma^{-1}(t)) \geq \frac{1}{2} t \log^{\frac{1}{2}}\left(1 + \frac{1}{t}\right), \quad \text{for } t \in (0, \frac{1}{2}]. \tag{115}$$

Case 1: First, we consider the case when the function q is constant on $[0,1]$ and all of the sets J_1, J_2, J_3 are intervals. Without loss of generality, we can shift and scale the density function by changing the domain, and assume that the density ρ is of the form $\rho(t) \propto e^{-\frac{t^2}{2}} \mathbf{1}_{[a,d]}$. Additionally, we can assume that J_1, J_2, J_3 are of the form

$$J_1 = [a, b), \quad J_3 = [b, c], \quad \text{and} \quad J_2 = (c, d], \tag{116}$$

because the case when J_3 is not in the middle is a trivial case.

Applying the inequalities (114) and (115) with $A = J_2 = (c, d]$, we obtain that

$$\phi_\gamma(c) = \Gamma^+(J_2) \geq \phi_\gamma(\Phi_\gamma^{-1}(\Gamma(J_2))) \geq \frac{\Gamma(J_2)}{2} \log^{\frac{1}{2}}\left(1 + \frac{1}{\Gamma(J_2)}\right). \tag{117}$$

10. E.g., see the discussion before equation 1 in Barthe and Maurey (2000). The constant $1/2$ was estimated by plotting the continuous function on the left hand side via Mathematica.

Note that $\rho(t) = \frac{\phi_\gamma(t)}{\Phi_\gamma(d)-\Phi_\gamma(a)}\mathbf{1}_{[a,d]}(t)$ and $\rho(J_2) = \frac{\Gamma(J_2)}{\Phi_\gamma(d)-\Phi_\gamma(a)}$. We have

$$\begin{aligned}\rho(J_3) = \int_b^c \rho(t)dt \geq (c-b)\cdot\rho(c) &= (c-b)\frac{\phi_\gamma(c)}{\Phi_\gamma(d)-\Phi_\gamma(a)} \\ &\overset{(i)}{\geq} \frac{(c-b)}{2}\frac{\Gamma(J_2)}{\Phi_\gamma(d)-\Phi_\gamma(a)}\log^{\frac{1}{2}}\left(1+\frac{1}{\Gamma(J_2)}\right) \\ &\overset{(ii)}{\geq} \frac{c-b}{2}\rho(J_2)\log^{\frac{1}{2}}\left(1+\frac{\Phi_\gamma(d)-\Phi_\gamma(a)}{\Gamma(J_2)}\right) \\ &\overset{(iii)}{=} \frac{c-b}{2}\rho(J_2)\log^{\frac{1}{2}}\left(1+\frac{1}{\rho(J_2)}\right) \\ &\overset{(iv)}{\geq} \frac{c-b}{2}\min\{\rho(J_1),\rho(J_2)\}\log^{\frac{1}{2}}\left(1+\frac{1}{\min\{\rho(J_1),\rho(J_2)\}}\right),\end{aligned}$$

where step (i) follows from the bound (117) and step (ii) follows from the relationship between ρ and Γ and the facts that log is an increasing function and that $\Phi_\gamma(d)-\Phi_\gamma(a) \leq 1$. Step (iii) follows from the definition of ρ and finally step (iv) follows from the increasing nature of the map $t \mapsto t\log^{1/2}\left(1+\frac{1}{t}\right)$. This concludes the argument for Case 1.

Case 2: We now consider the case when q is a general log-concave function on $[0,1]$ and J_1, J_2, J_3 are all intervals. Again we can assume that J_1, J_2, J_3 are of the form (116), i.e., they are given by $J_1 = [a,b), J_3 = [b,c]$, and $J_2 = (c,d]$.

We consider a function $h(t) = \alpha e^{\beta t - \frac{t^2}{2\sigma^2}}$ such that $h(b) = q(b)$ and $h(c) = q(c)$.[11] Define $Q(t_1,t_2) = \int_{t_1}^{t_2} q(t)dt$ and $H(t_1,t_2) = \int_{t_1}^{t_2} h(t)dt$. Then since q has an extra log-concave component compared to h, we have

$$H(a,b) \geq Q(a,b), \quad H(c,d) \geq Q(c,d), \text{ but } H(b,c) \leq Q(b,c). \tag{118}$$

Using the individual bounds in equation (118), we have

$$\frac{H(a,b)}{H(b,c)} + \frac{H(c,d)}{H(b,c)} \geq \frac{Q(a,b)}{Q(b,c)} + \frac{Q(c,d)}{Q(b,c)}.$$

From the equation above and the fact that $H(a,b)+H(b,c)+H(c,d) = H(a,d)$, we obtain

$$\frac{H(b,c)}{H(a,d)} \leq \frac{Q(b,c)}{Q(a,d)}. \tag{119}$$

To prove the inequality in Case 2, here are two subcases depending on whether $H(a,d) \geq Q(a,d)$ or $H(a,d) < Q(a,d)$.

11. This idea of introducing exponential function appeared in Corollary 6.2 of Kannan et al. Kannan et al. (2006).

- If $H(a,d) \geq Q(a,d)$, then

$$\begin{aligned}\frac{Q(b,c)}{Q(a,d)} &\overset{(i)}{\geq} \frac{H(b,c)}{Q(a,d)} \\ &\overset{(ii)}{\geq} \frac{c-b}{2} \cdot \frac{H(a,d)}{Q(a,d)} \cdot \frac{\min(H(a,b),H(c,d))}{H(a,d)} \cdot \log^{\frac{1}{2}}\left(1 + \frac{H(a,d)}{\min(H(a,b),H(c,d))}\right) \\ &\overset{(iii)}{\geq} \frac{c-b}{2} \cdot \frac{H(a,d)}{Q(a,d)} \cdot \frac{\min(Q(a,b),Q(c,d))}{H(a,d)} \cdot \log^{\frac{1}{2}}\left(1 + \frac{H(a,d)}{\min(Q(a,b),Q(c,d))}\right) \\ &\overset{(iv)}{\geq} \frac{c-b}{2} \cdot \frac{\min(Q(a,b),Q(c,d))}{Q(a,d)} \cdot \log^{\frac{1}{2}}\left(1 + \frac{Q(a,d)}{\min(Q(a,b),Q(c,d))}\right).\end{aligned}$$

  Inequality (i) follows from equation (118); inequality (ii) follows from equation Case 1 because H is covered by Case 1; inequality (iii) uses the fact that the function $t \mapsto t\log^{\frac{1}{2}}\left(1+\frac{1}{t}\right)$ is increasing; inequality (iv) follows from the assumption in this subcase $H(a,d) \geq Q(a,d)$.

- Otherwise $H(a,d) < Q(a,d)$, then we have from equation (118)

$$\frac{H(a,b)}{H(a,d)} \geq \frac{Q(a,b)}{Q(a,d)}, \quad \frac{H(c,d)}{Q(a,d)} \geq \frac{Q(c,d)}{Q(a,d)}.$$

$$\begin{aligned}\frac{Q(b,c)}{Q(a,d)} &\overset{(i)}{\geq} \frac{H(b,c)}{H(a,d)} \\ &\overset{(ii)}{\geq} \frac{c-b}{2} \cdot \frac{\min(H(a,b),H(c,d))}{H(a,d)} \cdot \log^{\frac{1}{2}}\left(1 + \frac{H(a,d)}{\min(H(a,b),H(c,d))}\right) \\ &\overset{(iii)}{\geq} \frac{c-b}{2} \cdot \frac{\min(Q(a,b),Q(c,d))}{Q(a,d)} \cdot \log^{\frac{1}{2}}\left(1 + \frac{Q(a,d)}{\min(Q(a,b),Q(c,d))}\right).\end{aligned}$$

  Inequality (i) follows from equation (119); inequality (ii) follows from equation Case 1; inequality (iii) uses the fact that the function $t \mapsto t\log^{\frac{1}{2}}\left(1+\frac{1}{t}\right)$ is increasing.

In both subcases above, we conclude Case 2 using the results established in Case 1.

Case 3: Finally, we deal with the general case where J_1, J_2, J_3 each can be union of intervals and q is a general log-concave function on $[0,1]$. We show that this case can be reduced to the case of three intervals, namely, the previous case.

Let $\{(b_i,c_i)\}_{i\in\mathcal{I}}$ be all non-empty maximal intervals contained in J_3. Here the intervals can be either closed, open or half. That is, $(\cdot,\cdot)$ can be $[\cdot,\cdot]$, $]\cdot,\cdot[$, $[\cdot,\cdot[$ or $]\cdot,\cdot]$. For an interval (b_i,c_i), we define its left surround $LS((b_i,c_i))$ as

$$LS((b_i,c_i)) = \begin{cases} 2, \text{ if } \exists x_2 \in J_2, (x_2 \leq b_i) \text{ and } (\nexists x_1 \in J_1, x_2 < x_1 \leq b_i) \\ 1, \text{ if } \exists x_1 \in J_1, (x_1 \leq b_i) \text{ and } (\nexists x_2 \in J_2, x_1 < x_2 \leq b_i) \\ 0, \text{ otherwise .} \end{cases}$$

Similarly, we define $RS((b_i, c_i))$ as

$$RS((b_i,c_i)) = \begin{cases} 2, \text{ if } \exists x_2 \in J_2, (x_2 \geq c_i) \text{ and } (\nexists x_1 \in J_1, x_2 > x_1 \geq c_i) \\ 1, \text{ if } \exists x_1 \in J_1, (x_1 \geq c_i) \text{ and } (\nexists x_2 \in J_2, x_1 > x_2 \geq c_i) \\ 0, \text{ otherwise .} \end{cases}$$

We distinguish two types of intervals. Denote $G_2 \subset \mathcal{I}$ the set containing the indices of all intervals that are surrounded by either 1 or 2 but different.

$$G_2 := \{i \in \mathcal{I} \mid (LS((b_i,c_i)), RS((b_i,c_i))) = (1,2) \text{ or } (2,1)\}.$$

Denote $G_1 := \mathcal{I} \setminus G_2$ to be its complement. By the result settled in case 2, for $i \in G_2$, we have

$$\rho([b_i,c_i]) \geq \frac{d(J_1,J_2)}{2\sigma}\rho(I_i)\log^{\frac{1}{2}}\left(1+\frac{1}{\rho(I_i)}\right)$$

where I_i is either $[a,b_i]$ or $[c_i,d]$. Summing over all $i \in G_2$, we have

$$\begin{aligned}\rho(J_3) \geq \sum_{i\in G_2}\rho([b_i,c_i]) &\geq \frac{d(J_1,J_2)}{2\sigma}\sum_{i\in G_2}\rho(I_i)\log^{\frac{1}{2}}\left(1+\frac{1}{\rho(I_i)}\right) \\ &\geq \frac{d(J_1,J_2)}{2\sigma}\rho(\cup_{i\in G_2}I_i)\log^{\frac{1}{2}}\left(1+\frac{1}{\rho(\cup_{i\in G_2}I_i)}\right). \end{aligned} \tag{120}$$

The last inequality follows from the sub-additivity of the map: $x \mapsto x\log^{\frac{1}{2}}(1+x)$, i.e., for $x>0$ and $y>0$, we have

$$x\log^{\frac{1}{2}}\left(1+\frac{1}{x}\right)+y\log^{\frac{1}{2}}\left(1+\frac{1}{y}\right) \geq (x+y)\log^{\frac{1}{2}}\left(1+\frac{1}{x+y}\right).$$

Indeed the sub-additivity follows immediately from the following observation:

$$\begin{aligned} &x\log^{\frac{1}{2}}\left(1+\frac{1}{x}\right)+y\log^{\frac{1}{2}}\left(1+\frac{1}{y}\right)-(x+y)\log^{\frac{1}{2}}\left(1+\frac{1}{x+y}\right) \\ &= x\left[\log^{\frac{1}{2}}\left(1+\frac{1}{x}\right)-\log^{\frac{1}{2}}\left(1+\frac{1}{x+y}\right)\right]+y\left[\log^{\frac{1}{2}}\left(1+\frac{1}{y}\right)-\log^{\frac{1}{2}}\left(1+\frac{1}{x+y}\right)\right] \\ &\geq 0. \end{aligned}$$

Finally, we remark that either J_1 or J_2 is a subset of $\cup_{i\in G_2}I_i$. If not, there exists $u \in J_1 \setminus \cup_{i\in G_2}I_i$ and $v \in J_2 \setminus \cup_{i\in G_2}I_i$, such that u and v are separated by some inverval $(b_{i^*}, c_{i^*}) \subset J_3$ with $i^* \in G_2$. This is contradictory with the fact that either u or v must be included in I_{i^*}. Given equation (120), we use the fact that the function $x \mapsto x\log^{\frac{1}{2}}\left(1+\frac{1}{x}\right)$ is monotonically increasing:

$$\rho(J_3) \geq \frac{d(J_1,J_2)}{2\sigma}\min\{\rho(J_1),\rho(J_2)\}\log^{\frac{1}{2}}\left(1+\frac{1}{\min\{\rho(J_1),\rho(J_2)\}}\right)$$

to conclude the proof.

Appendix C. Beyond strongly log-concave target distributions

In this appendix, we continue the discussion of mixing time bounds of Metropolized HMC from Section 3.2. In the next two subsections, we discuss the case when the target is weakly log-concave distribution or a perturbation of log-concave distribution, respectively.

C.1. Weakly log-concave target

The mixing rate in the weakly log-concave case differs depends on further structural assumptions on the density. We now consider two different scenarios where either a bound on fourth moment is known or the covariance of the distribution is well-behaved:

(C) The negative log density of the target distribution is L-smooth (10a) and has L_{H}-Lipschitz Hessian (10c). Additionally for some point $x^\star$, its fourth moment satisfies the bound

$$\int_{\mathbb{R}^d} \|x - x^\star\|_2^4 \, \pi^*(x) dx \leq \frac{d^2 \nu^2}{L}. \tag{121}$$

(D) The negative log density of the target distribution is L-smooth (10a) and has L_{H}-Lipschitz Hessian (10c). Additionally, its covariance matrix satisfies

$$\left\| \int_{x \in \mathbb{R}^d} (x - \mathbb{E}[x])(x - \mathbb{E}[x])^\top \pi^*(x) dx \right\|_{\mathrm{op}} \leq 1, \tag{122}$$

and the norm of the gradient of the negative log density f is bounded by a constant in the ball $\mathbb{B}\left(\mathbb{E}\left[x\right], \log\left(\frac{1}{s}\right) d^{3/4}\right)$ for small enough $s \geq s_0$.

When the distribution satisfies assumption (C) we consider HMC chain with slightly modified target and assume that the μ_0 is β-warm with respect to this modified target distribution (see the discussion after Corollary 18 for details). Moreover, In order to simplify the bounds in the next result, we assume that $L_{\mathrm{H}}^{2/3} = O(L)$. A more general result with without this condition can be derived in a similar fashion.

Corollary 18 (HMC mixing for weakly-log-concave) *Let μ_0 be a β-warm start, $\epsilon \in (0,1)$ be fixed and consider $\frac{1}{2}$-lazy HMC chain with leapfrog steps $K = d^{\frac{1}{2}}$ and step size $\eta^2 = \frac{1}{cLd^{\frac{4}{3}}}$.*

(a) If the distribution satisfies assumption (C), then we have

$$\tau_{\mathrm{TV}}^{HMC}(\epsilon; \mu_0) \leq c \cdot \max\left\{ \log \beta, \frac{d^{\frac{4}{3}} \nu}{\epsilon} \log\left(\frac{\log \beta}{\epsilon}\right) \right\}. \tag{123}$$

(b) If the distribution satisfies assumption (D) such that $s_0 \leq \frac{\epsilon^2}{2\beta}$, then we have

$$\tau_2^{HMC}(\epsilon; \mu_0) \leq c \cdot d^{\frac{5}{6}} \log\left(\frac{\log \beta}{\epsilon}\right). \tag{124}$$

As an immediate consequence, we obtain that the number of gradient evaluations in the two cases is bounded as

$$\mathcal{B}_1 = \max\left\{ d^{\frac{1}{2}} \log\beta, \frac{d^{\frac{11}{6}}\nu}{\epsilon} \log\left(\frac{\log\beta}{\epsilon}\right)\right\} \quad \text{and} \quad \mathcal{B}_2 = d^{\frac{4}{3}} \log\left(\frac{\log\beta}{\epsilon}\right).$$

We remark that the bound $\mathcal{B}_1$ for HMC chain improves upon the bound for number of gradient evaluations required by MALA to mix in a similar set-up. Dwivedi et al. (2018) showed that under assumption (C) (without the Lipschitz-Hessian condition), MALA takes $O(\frac{d^2}{\nu\epsilon}\log\frac{\beta}{\epsilon})$ steps to mix. Since each step of MALA uses one gradient evaluation, our result shows that HMC takes $O(d^{\frac{1}{6}})$ fewer gradient evaluations. On the other hand, when the target satisfies assumption (D), Mangoubi and Vishnoi (2019) showed that MALA takes $O(d^{\frac{3}{2}}\log\frac{\beta}{\epsilon})$ steps.[12] Thus even for this case, our result shows that HMC takes $O(d^{\frac{1}{6}})$ fewer gradient evaluations when compared to MALA.

Proof sketch: When the target distribution has a bounded fourth moment (assumption (C)), proceeding as in Dalalyan (2016), we can approximate the target distribution Π^* by a strongly log-concave distribution $\tilde{\Pi}$ with density given by

$$\tilde{\pi}(x) = \frac{1}{\int_{\mathbb{R}^d} e^{-\tilde{f}(y)}dy} e^{-\tilde{f}(x)} \quad \text{where } \tilde{f}(x) = f(x) + \frac{\lambda}{2}\left\|x - x^\star\right\|_2^2.$$

Setting $\lambda := \frac{2L\epsilon}{d\nu}$ yields that $\tilde{f}$ is $\lambda/2$-strongly convex, $L + \lambda/2$ smooth and L_{H}-Hessian Lipschitz and that the TV distance $d_{\mathrm{TV}}\left(\Pi^*, \tilde{\Pi}\right) \le \epsilon/2$ is small. The new condition number becomes $\widetilde{\kappa} := 1 + d\nu/\epsilon$. The new logarithmic-isoperimetric constant is $\widetilde{\psi}_{1/2} = (d\nu/(L\epsilon))^{1/2}$. Thus, in order to obtain an ϵ-accurate sample with respect to Π^*, it is sufficient to run HMC chain on the new strongly log-concave distribution $\widetilde{\Pi}$ upto $\epsilon/2$-accuracy. Invoking Corollary 2 for $\widetilde{\Pi}$ and doing some algebra yields the bound (123).

For the second case (assumption (D)), Lee and Vempala (2017) showed that when the covariance of Π^* has a bounded operator norm, it satisfies isoperimetry inequality (10d) with $\psi_0 \le O(d^{\frac{1}{4}})$. Moreover, using the Lipschitz concentration (Gromov and Milman, 1983), we have

$$\mathbb{P}_{x\sim\Pi^*}\left(\left\|x - \mathbb{E}_{\Pi^*}[x]\right\|_2 \ge t\psi_0 \cdot \sqrt{d}\right) \le e^{-ct},$$

which implies that for $\Omega_s = \mathbb{B}\left(\mathbb{E}_{\Pi^*}[x], \frac{1}{c}\log\left(\frac{1}{s}\right)\psi_0 \cdot \sqrt{d}\right)$, we have $\Pi^*(\Omega_s) \ge 1 - s$. In addition, assuming that the gradient is bounded in this ball Ω_s for $s = \frac{\epsilon^2}{2\beta}$ enables us to invoke Theorem 1 and obtain the bound (124) after plugging in the values of ψ_0, K and η.

12. Note that Mangoubi and Vishnoi (2019) assume an infinity-norm third order smoothness which is a stronger assumption than the L_{H}-Lipschitz Hessian assumption that we made here. Under our setting, the infinity norm third order smoothness is upper bounded by $\sqrt{d}L_{\mathrm{H}}$ and plugging in this bound changes their rate of MALA from $d^{7/6}$ to $d^{3/2}$.

C.2. Non-log-concave target

We now briefly discuss how our mixing time bounds in Theorem 1 can be applied for distributions whose negative log density may be non-convex. Let Π be a log-concave distribution with negative log density as f and isoperimetric constant ψ_0. Suppose that the target distribution $\widetilde{\Pi}$ is a perturbation of Π with target density $\widetilde{\pi}(x)$ such that $\widetilde{\pi}(x) \propto e^{-f(x)-\xi(x)}$, where the perturbation $\xi : \mathbb{R}^d \to \mathbb{R}$ is uniformly lower bounded by some constant $-b$ with $b \geq 0$. Then it can be shown that the distribution $\widetilde{\Pi}$ satisfies isoperimetric inequality (10d) with a constant $\widetilde{\psi}_0 \geq e^{-2b}\psi_0$. For example, such type of a non-log-concave distribution distribution arises when the target distribution is that of a Gaussian mixture model with several components where all the means of different components are close to each other (see e.g. Ma et al. (2018)). If a bound on the gradient is also known, Theorem 1 can be applied to obtain a suitable mixing time bound. However deriving explicit bounds in such settings is not the focus of the paper and thereby we omit the details here.

Appendix D. Optimal choice for HMC hyper-parameters

In this section, we provide a detailed discussion about the optimal leapfrog steps choice for Metropolized HMC with strongly log-concave target distribution (Corollary 2). We also discuss a few improved convergence rates for Metropolized HMC under additional assumptions on the target distribution. Finally, we compare our results for Metropolized HMC with other versions of HMC namely unadjusted HMC and ODE-solved based HMC in Subsection D.2.

D.1. Optimal choices for Corollary 14

Corollary 14 provides an implicit condition that the step size η and leapfrog steps K should satisfy and provides a generic mixing time upper bound that depends on the choices made. We claim that the optimal choices of η and K according to Table 4 lead to the following upper bound on number of gradient evaluations required by HMC to mix to ϵ-tolerance:

$$K \cdot \tau_{\mathrm{TV}}^{\mathrm{HMC}}(\epsilon; \mu_0) \leq O\left(\max\left\{d\kappa^{\frac{3}{4}}, d^{\frac{11}{12}}\kappa, d^{\frac{3}{4}}\kappa^{\frac{5}{4}}, d^{\frac{1}{2}}\kappa^{\frac{3}{2}}\right\} \cdot \log\frac{1}{\epsilon}\right). \tag{125}$$

This (upper) bound shows that HMC always requires fewer gradient evaluations when compared to MALA for mixing in total variation distance. However, such a bound requires a delicate choice of the leap frog steps K and η depending on the condition number κ and the dimension d, which might be difficult to implement in practice. We summarize these optimal choices in Table 4.

Proof of claim (125)**:** Recall that under the condition (103) (restated for reader's convenience)

$$\eta^2 \leq \frac{1}{cL}\min\left\{\frac{1}{K^2 d^{\frac{1}{2}}}, \frac{1}{K^2 d^{\frac{2}{3}}}\frac{L}{L_{\mathrm{H}}^{\frac{2}{3}}}, \frac{1}{K d^{\frac{1}{2}}}, \frac{1}{K^{\frac{2}{3}} d^{\frac{2}{3}} \kappa^{\frac{1}{3}} r(s)^{\frac{2}{3}}}, \frac{1}{K d^{\frac{1}{2}} \kappa^{\frac{1}{2}} r(s)}, \frac{1}{K^{\frac{2}{3}} d}\frac{L}{L_{\mathrm{H}}^{\frac{2}{3}}}, \frac{1}{K^{\frac{4}{3}} d^{\frac{1}{2}} \kappa^{\frac{1}{2}} r(s)}\left(\frac{L}{L_{\mathrm{H}}^{\frac{2}{3}}}\right)^{\frac{1}{2}}\right\},$$

Case	K	η^2
$\kappa \in (0, d^{\frac{1}{3}})$	$\kappa^{\frac{3}{4}}$	$\frac{1}{cL} \cdot d^{-1}\kappa^{-\frac{1}{2}}$
$\kappa \in [d^{\frac{1}{3}}, d^{\frac{2}{3}}]$	$d^{\frac{1}{4}}$	$\frac{1}{cL} \cdot d^{-\frac{7}{6}}$
$\kappa \in (d^{\frac{2}{3}}, d]$	$d^{\frac{3}{4}}\kappa^{-\frac{3}{4}}$	$\frac{1}{cL} \cdot d^{-\frac{3}{2}}\kappa^{\frac{1}{2}}$
$\kappa \in (d, \infty)$	1	$\frac{1}{cL} \cdot d^{-\frac{1}{2}}\kappa^{-\frac{1}{2}}$

Table 4: Optimal choices of leapfrog steps K and the step size η for the HMC algorithm for an (m, L, L_{H})-regular target distribution such that $L_{\mathrm{H}} = O(L^{\frac{3}{2}})$ used for the mixing time bounds in Corollary 14. Here c denotes a universal constant.

Corollary 2 guarantees that the HMC mixing time for the $\kappa^{\frac{d}{2}}$-warm initialization $\mu_\dagger = \mathcal{N}(x^\star, L^{-1}\mathbb{I}_d)$, is

$$\tau_2^{\mathrm{HMC}}(\epsilon; \mu_0) = O\left(d + \frac{\kappa}{K^2\eta^2 L}\right),$$

where we have ignored logarithmic factors. In order to compare with MALA and other sampling methods, our goal is to optimize the number of gradient evaluations $\mathcal{G}_{\mathrm{eval}}$ taken by HMC to mix:

$$\mathcal{G}_{\mathrm{eval}} := K \cdot \tau_{\mathrm{TV}}^{\mathrm{HMC}}(\epsilon; \mu_0) = O\left(Kd + \frac{\kappa}{K\eta^2 L}\right). \tag{126}$$

Plugging in the condition on η stated above, we obtain

$$\mathcal{G}_{\mathrm{eval}} \leq \max\left\{ \underbrace{Kd}_{=:T_1}, \quad \underbrace{K\max\left(d^{\frac{1}{2}}\kappa, d^{\frac{2}{3}}\kappa\vartheta\right)}_{=:T_2}, \quad \underbrace{d^{\frac{1}{2}}\kappa^{\frac{3}{2}}}_{=:T_3}, \quad \underbrace{K^{-\frac{1}{3}}d^{\frac{2}{3}}\kappa^{\frac{4}{3}}}_{=:T_4}, \quad \underbrace{K^{-\frac{1}{3}}d\kappa\cdot\vartheta}_{=:T_5}, \quad \underbrace{K^{\frac{1}{3}}d^{\frac{1}{2}}\kappa^{\frac{3}{2}}\cdot\vartheta^{\frac{1}{2}}}_{=:T_6} \right\} \tag{127}$$

where $\vartheta = L_{\mathrm{H}}^{\frac{2}{3}}/L$. Note that this bound depends only on the relation between d, κ and the choice of K. We now summarize the source of all of these terms in our proofs:

- T_1: This term is attributed to the warmness of the initial distribution. The distribution $\mu_\dagger$ is $O(\kappa^d)$-warm. This term could be improved if we have a warmer initial distribution.
- T_2: This term appears in the proposal overlap bound from equation (27a) of Lemma 6 and more precisely, it comes from equation (35).
- T_3, T_4, T_5 and T_6: These terms pop-out from the accept-reject bound from equation (27b) of Lemma 6. More precisely, T_3 and T_4 are a consequence of the first three terms in equation (51), and T_5 and T_6 arise the last two terms in equation (51).

In Table 5, we summarize how these six terms can be traded-off to derive the optimal parameter choices for Corollary 14. The effective bound on $\mathcal{G}_{\text{eval}}$-the number of gradient evaluations required by HMC to mix, is given by the largest of the six terms.

κ versus d	optimal choice K	T_1	T_2	T_3	T_4	T_5	T_6
		Kd	$Kd^{\frac{2}{3}}\kappa$	$d^{\frac{1}{2}}\kappa^{\frac{3}{2}}$	$K^{-\frac{1}{3}}d^{\frac{2}{3}}\kappa^{\frac{4}{3}}$	$K^{-\frac{1}{3}}d\kappa$	$K^{\frac{1}{3}}d^{\frac{1}{2}}\kappa^{\frac{3}{2}}$
$\kappa \in [1, d^{\frac{1}{3}})$	$K = \kappa^{\frac{3}{4}}$	$\mathbf{d\kappa^{\frac{3}{4}}}$	$d^{\frac{2}{3}}\kappa^{\frac{7}{4}}$	$d^{\frac{1}{2}}\kappa^{\frac{3}{2}}$	$d^{\frac{2}{3}}\kappa^{\frac{13}{12}}$	$\mathbf{d\kappa^{\frac{3}{4}}}$	$d^{\frac{1}{2}}\kappa^{\frac{7}{4}}$
$\kappa \in [d^{\frac{1}{3}}, d^{\frac{2}{3}}]$	$K = d^{\frac{1}{4}}$	$d^{\frac{5}{4}}$	$\mathbf{d^{\frac{11}{12}}\kappa}$	$d^{\frac{1}{2}}\kappa^{\frac{3}{2}}$	$d^{\frac{7}{12}}\kappa^{\frac{4}{3}}$	$\mathbf{d^{\frac{11}{12}}\kappa}$	$d^{\frac{7}{12}}\kappa^{\frac{3}{2}}$
$\kappa \in (d^{\frac{2}{3}}, d]$	$K = d^{\frac{3}{4}}\kappa^{-\frac{3}{4}}$	$d^{\frac{7}{4}}\kappa^{-\frac{3}{4}}$	$d^{\frac{19}{12}}\kappa^{\frac{1}{4}}$	$d^{\frac{1}{2}}\kappa^{\frac{3}{2}}$	$d^{\frac{5}{12}}\kappa^{\frac{19}{12}}$	$\mathbf{d^{\frac{3}{4}}\kappa^{\frac{5}{4}}}$	$\mathbf{d^{\frac{3}{4}}\kappa^{\frac{5}{4}}}$
$\kappa \in (d, \infty]$	$K = 1$	d	$d^{\frac{2}{3}}\kappa$	$\mathbf{d^{\frac{1}{2}}\kappa^{\frac{3}{2}}}$	$d^{\frac{2}{3}}\kappa^{\frac{4}{3}}$	$d\kappa$	$\mathbf{d^{\frac{1}{2}}\kappa^{\frac{3}{2}}}$

Table 5: Trade-off between the the six terms $T_i, i = 1, \ldots 6$, from the bound (127) under the assumption $\vartheta = L_{\text{H}}^{2/3}/L \leq 1$. In the second column, we provide the optimal choice of K for the condition on κ stated in first column such that the maximum of the T_i's is smallest. For each row the dominant (maximum) term, and equivalently the effective bound on $\mathcal{G}_{\text{eval}}$ is displayed in bold (red).

D.1.1. Faster mixing time bounds

We now derive several mixing time bounds under additional assumptions: (a) when a warm start is available, and (b) the Hessian-Lipschitz constant is small.

Faster mixing time with warm start: When a better initialization with warmness $\beta \leq O(e^{d^{\frac{2}{3}}\kappa})$ is available, and suppose that κ is much smaller than d. In such a case, the optimal choice turns out to be $K = d^{\frac{1}{4}}$ (instead of $\kappa^{\frac{3}{4}}$) which implies a bound of $O\left(d^{\frac{11}{12}}\kappa \log\left(\frac{1}{\epsilon}\right)\right)$ on $\mathcal{G}_{\text{eval}}$ (this bound was also stated in Table 1).

Faster mixing time with small L_{H}: Suppose in addition to warmness being not too large, $\beta \leq O(e^{d^{\frac{2}{3}}\kappa})$, the Hessian-Lipschitz constant L_{H} is small enough $L_{\text{H}}^{\frac{2}{3}} \ll L$. In such a scenario, the terms T_5 and T_6 become negligible because of small L_{H} and T_1 is negligible because of small β. The terms T_3 and T_4 remain unchanged, and the term T_2 changes slightly. More precisely, for the case $L_{\text{H}}^{\frac{2}{3}} \leq \frac{L}{d^{\frac{1}{2}}\kappa^{\frac{1}{2}}}$ we obtain a slightly modified trade-off for the terms in the (127) for $\mathcal{G}_{\text{eval}}$ (summarized in Table 6). If κ is small too, then we obtain a mixing time bound of order $d^{\frac{5}{8}}$. Via this artificially constructed example, we wanted to demonstrate two things. First, faster convergence rates are possible to derive under additional assumptions directly from our results. Suitable adaptation of our proof techniques might provide a faster rate of mixing for Metropolized HMC under additional assumptions like infinity semi-norm regularity condition made in other works (Mangoubi and Vishnoi, 2018) (but we leave a detailed derivation for future work). Second, it also

demonstrates the looseness of our proof techniques since we were unable to recover an $O(1)$ mixing time bound for sampling from a Gaussian target.

κ versus d	K optimal choice	T_1	T_2	T_3	T_4	T_5	T_6
		-	$Kd^{\frac{1}{2}}\kappa$	$d^{\frac{1}{2}}\kappa^{\frac{3}{2}}$	$K^{-\frac{1}{3}}d^{\frac{2}{3}}\kappa^{\frac{4}{3}}$	-	-
$\kappa \in (0, d^{\frac{1}{2}})$	$K = d^{\frac{1}{8}}\kappa^{\frac{1}{4}}$	-	$\mathbf{d^{\frac{5}{8}}\kappa^{\frac{5}{4}}}$	$d^{\frac{1}{2}}\kappa^{\frac{3}{2}}$	$\mathbf{d^{\frac{5}{8}}\kappa^{\frac{5}{4}}}$	-	-

Table 6: Six terms in the HMC number of gradient evaluations bound under small hessian-Lipschitz constant and very warm start. The dominant term is highlighted in red.

Linearly transformed HMC (effect of mass function): In practice, it is often beneficial to apply linear transformations in HMC (cf. Section 4 in Neal (2011)). At a high level, such a transformation can improve the conditioning of the problem and help HMC mix faster. For the target distribution Π^* with density proportional to e^{-f}, we can define a new distribution Π_h with density e^{-h} (up to normalization) such that $h(x) = f(M^{-\frac{1}{2}}x)$ where $M \in \mathbb{R}^{d\times d}$ is an invertible matrix. Then for a random sample $\tilde{q} \sim \Pi_h$, the distribution of $M^{\frac{1}{2}}\tilde{q}$ is Π^*. When the new distribution h has a better condition number κ_h than the condition number κ of f, we can use HMC to draw approximate sample from Π_h and then transform the samples using the matrix M. Clearly the bound from Corollary 14 guarantees that when κ_h is much smaller than κ, HMC on the new target Π_h would mix much faster than the HMC chain on Π^*. This transformation is equivalent to the HMC algorithm with modified kinetic energy

$$\frac{dq_t}{dt} = M^{-1}p_t \quad \text{and} \quad \frac{dp_t}{dt} = -\nabla f(q_t),$$

which is easier to implement in practice. For a detailed discussion of this implementation, we refer the readers to the paper by Neal (2011).

D.2. Comparison with guarantees for unadjusted versions of HMC

In this appendix, we compare our results with mixing time guarantees results on unadjusted and ODE solver based HMC chains. We summarize the number of gradient evaluations needed for Metropolized HMC to mix and those for other existing sampling results in Table 7. Note that all the results summarized here are the best upper bounds in the literature for log-concave sampling. We present the results for a (L, L_{H}, m)-regular target distribution. We remark that all methods presented in Table 7 requires the regularity assumptions (10a) and (10b), even though some do not require assumption (10c).

Two remarks are in order. First, the error metric for the guarantees in the works (Mangoubi and Vishnoi, 2018; Cheng et al., 2017; Lee et al., 2018) is 1-Wasserstein distance, while our results make use of $\mathcal{L}_2$ or TV distance. As a result, a direct comparison between these results is not possible although we provide an indirect comparison below. Second, the

Sampling algorithm	#Grad. evals
$^{\ddagger,\diamond}$Unadjusted HMC with leapfrog integrator (Mangoubi and Vishnoi, 2018)	$d^{\frac{1}{4}}\kappa^{\frac{11}{4}} \cdot \frac{1}{\epsilon^{1/2}}$
‡Underdamped Langevin (Cheng et al., 2017)	$d^{\frac{1}{2}}\kappa^{2} \cdot \frac{1}{\epsilon}$
‡HMC with ODE solver, Thm 1.6 in (Lee et al., 2018)	$d^{\frac{1}{2}}\kappa^{\frac{7}{4}} \cdot \frac{1}{\epsilon}$
*MALA (Dwivedi et al., 2018)[this paper]	$\max\left\{d\kappa, d^{\frac{1}{2}}\kappa^{\frac{3}{2}}\right\} \cdot \log\frac{1}{\epsilon}$
*Metropolized HMC with leapfrog integrator [this paper]	$\max\left\{d\kappa^{\frac{3}{4}}, d^{\frac{11}{12}}\kappa, d^{\frac{3}{4}}\kappa^{\frac{5}{4}}, d^{\frac{1}{2}}\kappa^{\frac{3}{2}}\right\} \cdot \log\frac{1}{\epsilon}$

Table 7: Summary of the number of gradient evaluations needed for the sampling algorithms to converge to a (m, L, L_{H})-regular target distribution with $L_{\mathrm{H}} = O(L^{\frac{3}{2}})$ within ϵ error from the target distribution (in total-variation distance* or 1-Wasserstein distance‡) (and $\diamond$ certain additional regularity conditions for the result by Mangoubi and Vishnoi (2018)). Note that the unadjusted algorithms suffer from an exponentially worse dependency on ϵ when compared to the Metropolis adjusted chains. For MALA, results by Dwivedi et al. (2018) had an extra d factor which is sharpened in Theorem 5 of this paper.

previous guarantees have a polynomial dependence on the inverse of error-tolerance $1/\epsilon$. In contrast, our results for MALA and Metropolized HMC have a logarithmic dependence $\log(1/\epsilon)$. For a well-conditioned target, i.e., when κ is a constant, all prior results have a better dependence on d when compared to our bounds.

Logarithmic vs polynomial dependence on $1/\epsilon$: We now provide an indirect comparison, between prior guarantees based on Wasserstein distance and our results based on TV-distance, for estimating expectations of Lipschitz-functions on bounded domains. MCMC algorithms are used to estimate expectations of certain functions of interest. Given an arbitrary function g and an MCMC algorithm, one of the ways to estimate $\Pi^*(g) := \mathbb{E}_{X\sim\Pi^*}[g(X)]$ is to use the k-th iterate from N independent runs of the chain. Let $X_i^{(k)}$ for $i = 1, \ldots, N$ denote the N i.i.d. samples at the k-th iteration of the chain and let μ_k denote the distribution of $X_i^{(k)}$, namely the distribution of the chain after k iterations. Then for the estimate $\widehat{\Pi}_k(g) := \frac{1}{N}\sum_{i=1}^{N} g(X_i^{(k)})$, the estimation error can be decomposed as

$$\begin{aligned}\Pi^*(g) - \widehat{\Pi}_k(g) &= \int_{\mathbb{R}^d} g(x)\pi^*(x)dx - \frac{1}{N}\sum_{i=1}^{N} g(X_i^{(k)}) \\ &= \underbrace{\int_{\mathbb{R}^d} g(x)\left[\pi^*(x) - \mu_k(x)\right]dx}_{=:J_1 \text{ (Approximation bias)}} + \underbrace{\mathbb{E}_{\mu_k}\left[g(X)\right] - \frac{1}{N}\sum_{i=1}^{N} g(X_i^{(k)})}_{=:J_2 \text{ (Finite sample error)}}. \end{aligned} \tag{128}$$

To compare different prior works, we assume that $\mathrm{Var}_{\mu_k}[g(X_1)]$ is bounded and thereby that the finite sample error J_2 is negligible for large enough N.[13] It remains to bound the error J_1 which can be done in two different ways depending on the error-metric used to provide mixing time guarantees for the Markov chain.

If the function g is ω-Lipschitz and k is chosen such that $\mathcal{W}_1(\Pi^*, \mu_k) \leq \epsilon$, then we have $J_1 \leq \omega\epsilon =: J_{\mathrm{Wass}}$. On the other hand, if the function g is bounded by B, and k is chosen such that $d_{\mathrm{TV}}(\Pi^*, \mu_k) \leq \epsilon$, then we obtain the bound $J_1 \leq B\epsilon =: J_{\mathrm{TV}}$. We make use of these two facts to compare the number of gradient evaluations needed by unadjusted HMC or ODE solved based HMC and Metropolized HMC. Consider an ω-Lipschitz function g with support on a ball of radius R. Note that this function is uniformly bounded by $B = \omega R$. Now in order to to ensure that $J_1 \leq \delta$ (some user-specified small threshold), the choice of ϵ in the two cases (Wasserstein and TV distance) would be different leading to different number of gradient evaluations required by the two chains. More precisely, we have

$$J_1 \leq J_{\mathrm{Wass}} = \omega\epsilon \leq \delta \quad \Longrightarrow \quad \epsilon_{\mathrm{wass}} = \frac{\delta}{\omega} \quad \text{and}$$
$$J_1 \leq J_{\mathrm{TV}} = B\epsilon = \omega R\epsilon \leq \delta \quad \Longrightarrow \quad \epsilon_{\mathrm{TV}} = \frac{\delta}{\omega R}.$$

To simplify the discussion, we consider well-conditioned (constant κ) strongly log-concave distributions such that most of the mass is concentrated on a ball of radius $O(\sqrt{d})$ (cf. Appendix B.1) and consider $R = \sqrt{d}$. Then plugging the error-tolerances from the display above in Table 7, we obtain that the number of gradient evaluations $\mathcal{G}_{\mathrm{MC}}$ for different chains[14] would scale as

$$\mathcal{G}_{\text{unadj.-HMC}} \leq O(\sqrt{\frac{d\omega}{\delta}}), \quad \mathcal{G}_{\text{ODE-HMC}} \leq O(\frac{\omega\sqrt{d}}{\delta}), \text{and} \quad \mathcal{G}_{\text{Metro.-HMC}} \leq O(d \log \frac{\omega\sqrt{d}}{\delta})$$

Clearly, depending on ω and the threshold δ, different chains would have better guarantees. When ω is large or δ is small, our results ensure the superiority of Metropolized-HMC over other versions. For example, higher-order moments can be functions of interest, i.e., $g(x) = \|x\|^{1+\nu}$ for which the Lipschitz-constant $\omega = O(d^\nu)$ scales with d. For this function, we obtain the bounds:

$$\mathcal{G}_{\text{unadj.-HMC}} \leq O(\frac{d^{\frac{1+\nu}{2}}}{\sqrt{\delta}}), \quad \mathcal{G}_{\text{ODE-HMC}} \leq O(\frac{d^{\frac{1}{2}+\nu}}{\delta}), \quad \text{and} \quad \mathcal{G}_{\text{Metro.-HMC}} \leq O(d(1+\nu) \log \frac{d}{\delta})$$

and thus Metropolized HMC takes fewer gradient evaluations than ODE-based HMC for $\nu > 1/2$ and unadjusted HMC for $\nu > 1$ (to ensure $J_1 \leq \delta$ (128)). We remark that the bounds for unadjusted-HMC require additional regularity conditions. From this informal comparison, we demonstrate that both the dimension dependency d and error dependency ϵ should be accounted for comparing unadjusted algorithms and Metropolized algorithms. Especially for estimating high-order moments, Metropolized algorithms with $\log(\frac{1}{\epsilon})$ dependency will be advantageous.

13. Moreover, this error should be usually similar across different sampling algorithms since several algorithms are designed in a manner agnostic to a particular function g.
14. The results for other HMCs often assume (different) additional conditions so that a direct comparison should be taken with a fine grain of salt.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `http://tensorflow.org/`. Software available from tensorflow.org.

Berni J Alder and T E Wainwright. Studies in molecular dynamics. I. General method. *The Journal of Chemical Physics*, 31(2):459–466, 1959.

David Aldous and James Allen Fill. Reversible Markov chains and random walks on graphs, 2002. Unfinished monograph, recompiled 2014, available at `http://www.stat.berkeley.edu/~aldous/RWG/book.html`.

Franck Barthe and Bernard Maurey. Some remarks on isoperimetry of Gaussian type. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 36, pages 419–434. Elsevier, 2000.

Alexandros Beskos, Natesh Pillai, Gareth Roberts, Jesus-Maria Sanz-Serna, Andrew Stuart, et al. Optimal tuning of the hybrid monte carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 2013.

MJ Betancourt, Simon Byrne, and Mark Girolami. Optimizing the integrator step size for Hamiltonian Monte Carlo. *arXiv preprint arXiv:1411.6669*, 2014.

Sergey G Bobkov. Isoperimetric and analytic inequalities for log-concave probability measures. *The Annals of Probability*, 27(4):1903–1921, 1999.

Nawaf Bou-Rabee, Andreas Eberle, and Raphael Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *arXiv preprint arXiv:1805.00452*, 2018.

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

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.

Jeff Cheeger. A lower bound for the smallest eigenvalue of the Laplacian. In *Proceedings of the Princeton Conference in honor of Professor S. Bochner*, 1969.

Lingyu Chen, Zhaohui Qin, and Jun S Liu. Exploring hybrid monte carlo in bayesian computation. *sigma*, 2:2–5, 2001.

Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Fast MCMC sampling algorithms on polytopes. *The Journal of Machine Learning Research*, 19(1):2146–2231, 2018.

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 and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. *arXiv preprint arXiv:1705.09048*, 2017.

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.

T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

Michael Creutz. Global Monte Carlo algorithms for many-fermion systems. *Physical Review D*, 38(4):1228, 1988.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 2019.

Persi Diaconis, Laurent Saloff-Coste, et al. Logarithmic Sobolev inequalities for finite Markov chains. *The Annals of Applied Probability*, 6(3):695–750, 1996.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

Alain Durmus, Eric Moulines, and Eero Saksman. On the convergence of Hamiltonian Monte Carlo. *arXiv preprint arXiv:1705.00166*, 2017.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *arXiv preprint arXiv:1801.02309*, 2018.

Andreas Eberle. Error bounds for Metropolis-Hastings algorithms applied to perturbations of Gaussian measures in high dimensions. *The Annals of Applied Probability*, 24(1):337–377, 2014.

Walter R Gilks and Pascal Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, pages 337–348, 1992.

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

Mikhail Gromov and Vitali D Milman. A topological application of the isoperimetric inequality. *American Journal of Mathematics*, 105(4):843–854, 1983.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623, 2014.

Christian Houdré. Mixed and isoperimetric estimates on the log-Sobolev constants of graphs and Markov chains. *Combinatorica*, 21(4):489–513, 2001.

Mark Jerrum and Alistair Sinclair. Conductance and the rapid mixing property for Markov chains: the approximation of permanent resolved. In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, pages 235–244. ACM, 1988.

Ravi Kannan, László Lovász, and Miklós Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13(3-4):541–559, 1995.

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

Michel Ledoux. Concentration of measure and logarithmic Sobolev inequalities. In *Seminaire de Probabilites XXXIII*, pages 120–216. Springer, 1999.

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, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 1115–1121, 2018a.

Yin Tat Lee and Santosh S Vempala. Stochastic localization+ Stieltjes barrier= tight bound for log-Sobolev. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1122–1129. ACM, 2018b.

Yin Tat Lee and Santosh Srinivas Vempala. Eldan's stochastic localization and the KLS hyperplane conjecture: An improved lower bound for expansion. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 998–1007. IEEE, 2017.

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.

Samuel Livingstone, Michael Betancourt, Simon Byrne, and Mark Girolami. On the geometric ergodicity of Hamiltonian Monte Carlo. *arXiv preprint arXiv:1601.08057*, 2016.

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 31st annual ACM Symposium on Theory of Computing*, pages 282–287. ACM, 1999.

László Lovász and Miklós Simonovits. The mixing rate of Markov chains, an isoperimetric inequality, and computing the volume. In *Proceedings of 31st Annual Symposium on Foundations of Computer Science, 1990*, 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 et al. Random walks on graphs: A survey. *Combinatorics, Paul erdos is eighty*, 2(1):1–46, 1993.

Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I Jordan. Sampling can be faster than optimization. *arXiv preprint arXiv: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.

Oren Mangoubi and Nisheeth K Vishnoi. Dimensionally tight running time bounds for second-order Hamiltonian Monte Carlo. *arXiv preprint arXiv:1802.08898*, 2018.

Oren Mangoubi and Nisheeth K Vishnoi. Nonconvex sampling with the Metropolis-adjusted Langevin algorithm. *arXiv preprint arXiv:1902.08452*, 2019.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Ben Morris and Yuval Peres. Evolving sets, mixing and heat kernel bounds. *Probability Theory and Related Fields*, 133(2):245–266, 2005.

Radford M Neal. An improved acceptance procedure for the hybrid Monte Carlo algorithm. *Journal of Computational Physics*, 111:194–203, 1994.

Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer texts in statistics. Springer-Verlag, New York, NY, 1999.

Gareth O Roberts and Osnat Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002.

Gareth O Roberts, Jeffrey S Rosenthal, et al. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

Adrian FM Smith and Gareth O Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(1):3–23, 1993.

BJ Smith. Mamba: Markov chain Monte Carlo (MCMC) for Bayesian analysis in julia, 2014. URL `https://mambajl.readthedocs.io/en/latest/`. Software available at mambajl.readthedocs.io.

Luke Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, pages 1701–1728, 1994.

Santosh Vempala. Geometric random walks: a survey. *Combinatorial and Computational Geometry*, 52(573-612):2, 2005.

Ziyu Wang, Shakir Mohamed, and Nando Freitas. Adaptive Hamiltonian and Riemann manifold Monte Carlo. In *International Conference on Machine Learning*, pages 1462–1470, 2013.

Changye Wu, Julien Stoehr, and Christian P Robert. Faster Hamiltonian Monte Carlo by learning leapfrog scale. *arXiv preprint arXiv:1810.04449*, 2018.