Journal of Machine Learning Research 25 (2024) 1-30 Submitted 11/23; Revised 8/24; Published 9/24

Hamiltonian Monte Carlo for efficient Gaussian sampling: long and random steps

Simon Apers APERS@IRIF.FR
IRIF
Université Paris Cité, CNRS
Paris, F-75013, France

Sander Gribling S.J.GRIBLING@TILBURGUNIVERSITY.EDU
Department of Econometrics and Operations Research
Tilburg University
Tilburg, 5000 LE, the Netherlands

Dániel Szilágyi SZILAGYI.D@GMAIL.COM
IRIF
Université Paris Cité
Paris, F-75013, France

Editor: Anthony Lee

Abstract

Hamiltonian Monte Carlo (HMC) is a Markov chain algorithm for sampling from a high-dimensional distribution with density $e^{-f(x)}$, given access to the gradient of f. A particular case of interest is that of a d-dimensional Gaussian distribution with covariance matrix Σ, in which case $f(x) = x^\top \Sigma^{-1} x$. We show that Metropolis-adjusted HMC can sample from a distribution that is ε-close to a Gaussian in total variation distance using $\widetilde{O}(\sqrt{\kappa} d^{1/4} \log(1/\varepsilon))$ gradient queries, where $\varepsilon > 0$ and κ is the condition number of Σ.

Our algorithm uses *long* and *random* integration times for the Hamiltonian dynamics, and it creates a warm start by first running HMC without a Metropolis adjustment. This contrasts with (and was motivated by) recent results that give an $\widetilde{\Omega}(\kappa d^{1/2})$ query lower bound for HMC with a fixed integration times or from a cold start, even for the Gaussian case.

Keywords: Markov chains, logconcave sampling, Metropolis-Hastings algorithm, numerical integration, Hamiltonian Monte Carlo

1. Introduction and main result

One of the most important tasks in statistics and machine learning is to sample from high-dimensional and potentially complicated distributions. Markov chains are an efficient means for sampling from such distributions, and there is a wide variety of Markov chain algorithms designed specifically for this purpose. Typically, the main difficulty in analyzing these algorithms is to bound the precise running time or *mixing time* of the Markov chain. While many algorithms have been in very broad (heuristic) usage for several decades, rigorous bounds on their performance are often missing. A key example is the *Hamiltonian Monte Carlo* (HMC) algorithm by Duane et al. (1987). This is an elegant Markov chain algorithm

that utilizes Hamiltonian dynamics to efficiently explore the state space, without straying too far away from the high probability region. One of its key features is that it overcomes the slow, diffusive behavior that is inherent to "small step" approaches such as the ball walk Vempala (2005) and Langevin algorithm Parisi (1981). While this is indeed observed in heuristic uses and simulation studies of the HMC algorithm (see e.g. Neal (2011); Bou-Rabee and Sanz-Serna (2017)), recent efforts in proving theoretical bounds are mostly restricted to step sizes much shorter than the heuristic choices (Chen et al. (2020); Chen and Vempala (2022)). In this work, we prove seemingly optimal[1] bounds on the Metropolis-adjusted HMC algorithm (with leapfrog integrator) for the special case of Gaussian distributions. This is the typical gateway to more complicated distributions such as logconcave or multimodal distributions. Our implementation of HMC exploits (i) long and randomized integration times, and (ii) an "algorithmic" warm start obtained by first running the *unadjusted* HMC chain. This surpasses recent roadblocks on sampling Gaussian distributions using HMC with either short (Chen and Vempala (2022)) integrations times, deterministic integration times, or Metropolis-adjustments applied to a cold start (Lee et al. (2021)).

Our bounds are stated most easily in the "black box model", where the goal is to sample from a density of the form $e^{-f(x)}$ for $x \in \mathbb{R}^d$, and we are given query access to both f and its gradient ∇f. The Gaussian case further restricts f to be a quadratic form $f(x) = \frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)$, where μ and Σ are the (unknown) mean and covariance matrix of the Gaussian, respectively. The *condition number* of the Gaussian distribution is simply the condition number of Σ^{-1}. Throughout we assume that we are given bounds $0 < \alpha \leq \beta$ such that $\alpha I \preceq \Sigma^{-1} \preceq \beta I$ and we use $\kappa = \beta/\alpha$ as an upper bound on the condition number. We prove the following theorem.

Theorem 1 (informal version of Theorem 20) *The Metropolis-adjusted HMC algorithm with leapfrog integrator*[2] *can sample from a distribution ε-close in total variation distance to a d-dimensional Gaussian distribution with condition number κ using a total number of gradient evaluations*[3]

$$\widetilde{O}(\sqrt{\kappa} d^{1/4} \log(1/\varepsilon)).$$

This theorem builds on an analysis of the *unadjusted* HMC algorithm, for which we get a bound of $\widetilde{O}(\sqrt{\kappa} d^{1/4}/\sqrt{\varepsilon})$ on the total number of gradient evaluations. Both bounds seem in line with expectation (see Duane et al. (1987); Neal (2011); Beskos et al. (2013)), and we expect they are tight when using the usual leapfrog integrator for simulating the Hamiltonian dynamics. Our algorithm surpasses the $\widetilde{\Omega}(\kappa\sqrt{d})$ lower bound on the complexity of HMC for Gaussian sampling from (Lee et al., 2021, Proposition 4 and Corollary 5) by using *randomized* integration times (which avoids aperiodicity issues associated to a deterministic integration time) and an algorithmic warm start (which avoids exponentially small acceptance probabilities).

Our work fits within the recent effort of proving non-asymptotic (and often tight) bounds on Markov chain algorithms for constrained distributions such as Gaussian distributions

1. See Section 5 for a discussion on why this bound seems optimal.
2. To be precise, we first run a number of HMC steps without Metropolis correction, providing a warm start for the Metropolis-adjusted HMC algorithm.
3. We use the $\widetilde{O}(\cdot)$- and $\widetilde{\Omega}(\cdot)$-notation to denote the usual big-O and big-Omega notation, $O(\cdot)$ and $\Omega(\cdot)$ respectively, where the tilde indicates that we hide polylogarithmic factors in the problem parameters d, α, β and $\log(1/\varepsilon)$.

and, more generally, logconcave distributions (where f is assumed to be convex). Most of these efforts have focused on short step dynamics such as the ball walk, the Langevin algorithm, and HMC with short integration times. The use of such "local steps" makes it easier to control the stability and acceptance probability of the algorithm. However, the restriction to short step dynamics is also what slows down these algorithms, and this is what we avoid in our HMC algorithm.

Another motivation for studying Gaussian sampling is that the restriction to sampling Gaussian and logconcave distributions precisely parallels the restriction to quadratic and convex functions in optimization. Nonetheless, a gap between the (first-order oracle) complexity for logconcave sampling and the $O(\min\{\sqrt{\kappa}, d\})$ complexity for convex optimization is apparently deemed plausible. More specifically, the authors in (Lee et al. (2020)) suggest an $\Omega(\kappa)$ lower bound for logconcave sampling. Our work shows that a sublinear κ-dependency is possible at least for the special case of Gaussian distributions, and we see it as evidence that a general $O(\sqrt{\kappa})$ bound for logconcave sampling is achievable.

On the same note of generalizing our results to logconcave distributions, we emphasize that our result strongly relies on the restriction to Gaussian sampling, for which explicit expressions can be derived on the Hamiltonian dynamics. While this makes progress towards the more general case of logconcave distributions, the lack of such explicit expressions in the general case is a clear obstacle when trying to generalize our techniques.

Finally, as a direct application of our work, we mention the use of Gaussian sampling in the contextual multi-armed bandit problem (see Agrawal and Goyal (2012)). A competitive exploration-exploitation strategy for this problem is called *Thompson sampling*, which is an efficient manner of maintaining a posterior on the set of arms. In the case of a linear payoff, as is considered by Agrawal and Goyal (2013), the prior and posterior distributions are Gaussian distributions. While recent works suggested the use of Langevin dynamics for Thompson sampling (Mazumdar et al. (2020); Xu et al. (2022)), our work shows that the use of Hamiltonian Monte Carlo should lead to faster algorithms, improving the mixing time from $\widetilde{O}(\kappa d^{1/3})$ for the Metropolis-adjusted Langevin algorithm (MALA) to $\widetilde{O}(\sqrt{\kappa} d^{1/4})$ for HMC.

1.1 Background and prior work

There is a vast body of work on the use of Markov chain algorithms for sampling from Gaussian and logconcave distributions. These works mostly consider the (Metropolized) random walk or ball walk (MRW), MALA,[4] and HMC. We discuss those works most directly related to ours.

The earliest works focus on *asymptotic* bounds or scaling limits on the performance as $d \to \infty$. A $d^{1/4}$-scaling was already suggested in Duane et al. (1987); Kennedy and Pendleton (1991); Beskos et al. (2013) for the complexity of HMC with leapfrog integrator for Gaussians and logconcave product distributions. This improves over the expected d- and $d^{1/3}$-scalings of MRW and MALA, respectively. Indeed, in the recent work by Chewi et al. (2021) it was proven that the complexity of MALA for standard Gaussian distributions (with $\kappa = 1$) from a warm start scales as $\widetilde{O}(d^{1/3})$. For leapfrog HMC, the first non-

4. MALA can be interpreted as HMC with a very short integration time (e.g., one leapfrog step), see for instance (Lee et al., 2020, Appendix A).

asymptotic bounds scaling with $d^{1/4}$ seem to have been proven recently in (Mangoubi and Vishnoi (2018); Mou et al. (2021)) for the *unadjusted* HMC chain, restricted to logconcave distributions that satisfy additional regularity assumptions. The final complexities in these works scale at least with κ^2 and $1/\sqrt{\varepsilon}$, and so scale worse in terms of both κ and ε compared to our bound.[5] An improved (linear) κ-dependency is obtained in recent works on MALA by Dwivedi et al. (2018); Lee et al. (2020); Altschuler and Chewi (2023); Wu et al. (2021) and HMC by Chen et al. (2020); Chen and Gatmiry (2023) (from a warm start).

Lower bounds. Many of the aforementioned works satisfy the $\widetilde{\Omega}(\kappa\sqrt{d})$ lower bounds on MALA and HMC from Wu et al. (2021); Lee et al. (2021), which even apply to the Gaussian case. Such lower bound typically follows from restricting to either of the following:

1. *short* integration times, which leads to diffusive behavior, c.f. Chen and Vempala (2022)

2. *fixed* integration times, which can lead to periodic behavior in the HMC algorithm, cf. (Lee et al., 2021, Proposition 4), or

3. when considering a Metropolis-adjusted chain starting from a cold start (leading to exponentially small acceptance probabilities), cf. (Lee et al., 2021, Corollary 5).

Any of these restrictions leads to the aforementioned lower bound, and indeed we are not aware of any former non-asymptotic bounds on the mixing time achieving a *sublinear* κ-dependency (while using a numerical integrator). We sidestep these bounds by using (1.) *long* and (2.) *random* integration times, and (3.) using an "algorithmic" warm start (as e.g. in Altschuler and Chewi (2023)) obtained by first running the *unadjusted* HMC chain. Although limited to Gaussians, our result is an important first step towards proving $\sqrt{\kappa}$-scalings for general logconcave distributions.

Variable integration times. The use of nonconstant integration times was also studied recently in the *randomized* HMC algorithm by Bou-Rabee and Sanz-Serna (2017). Similarly to our work, they motivate their algorithm by looking at the Gaussian case, and obtain similar scalings to our work for properties such as the autocorrelation time and mean displacement. In follow-up works by Deligiannidis et al. (2021); Lu and Wang (2022) bounds similar to ours are proven on the relaxation time. The related work Wang and Wibisono (2022) picks integration times by deterministically cycling through a set of integration times based on the roots of Chebyshev polynomials, and achieves a convergence time in Wasserstein-2 distance for Gaussians scaling with $\sqrt{\kappa}$. Similarly, Jiang (2023) achieves a $\sqrt{\kappa}$-scaling in Wasserstein-2 distance by randomly choosing integration times, similarly to our work. The main difference with these works is that they are proven only for the *idealized* case, and do not take into account the errors that arise from numerical integration, which is the technical bulk of this work.[6]

5. We note that such $1/\operatorname{poly}(\varepsilon)$-dependency is unavoidable for the unadjusted chain.
6. An exception is Jiang (2023), where a revised version contains a remark about using the leapfrog integrator. It states that the *unadjusted* HMC algorithm with randomized integration times achieve a scaling $\widetilde{O}(\sqrt{\kappa}d^{1/4}/\sqrt{\varepsilon})$ where ε is the error in Wasserstein-2 distance – this complements our $\widetilde{O}(\sqrt{\kappa}d^{1/4}/\sqrt{\varepsilon})$ bound on the unadjusted HMC chain in total variation distance (Proposition 7).

Gaussian sampling without Markov chains. For completeness we also mention that there are algorithms for Gaussian sampling that are *not* based on Markov chains. While these are generally incomparable (e.g., they require access to the precision or covariance matrix rather than gradient), we refer the interested reader to Vono et al. (2022). A very recent work that does combine matrix methods with the gradient query model is Chewi et al. (2023). They prove an $\Omega(\min\{\sqrt{\kappa}, d\})$ query lower bound for (centered) Gaussians, and an $O(\min\{\sqrt{\kappa}\log(d), d\})$ query upper bound based on a matrix Kyrylov method. The lower bound applies to our setting as well, and shows that the κ-dependency of our algorithm is optimal. The upper bound improves over our $\widetilde{O}(\sqrt{\kappa}d^{1/4})$ upper bound, but the matrix Krylov ideas from Chewi et al. (2023) are inherently restricted to Gaussians, contrasting with our HMC approach that in principle generalizes to arbitrary distribution.

1.2 Organization and proof overview

In Section 2 we formally introduce the problem and describe preliminaries related to Markov chains and Hamiltonian dynamics. In particular, for the Gaussian case, we discuss how the numerical leapfrog integrator exactly integrates the Hamiltonian of a closely related Gaussian. In Section 3 we bound the mixing time of the HMC algorithm with an idealized integrator. Using the observation about the leapfrog integrator, this mixing time extends to the "unadjusted" HMC algorithm, which is an exact HMC algorithm for a slightly perturbed Hamiltonian (and hence has a slightly perturbed stationary distribution). Finally, in Section 4, we consider the Metropolis-adjusted HMC algorithm with leapfrog integrator. This algorithm has the correct stationary distribution, but the mixing time might increase due to an additional accept-reject step. We use high-dimensional concentration bounds (in particular, the Hanson-Wright inequality) to show that the acceptance rate is usually large. This suffices to bound the mixing time through the use of s-conductance, which proves our main result.

2. Problem definition and preliminaries

2.1 Gaussian sampling

We consider a d-dimensional Gaussian distribution with unknown precision matrix B (equal to the inverse of the covariance matrix, $B = \Sigma^{-1}$) and mean $\mu = 0$.[7] In such case, the Gaussian distribution is $\pi(x) \propto \exp(-f(x))$ with $f(x) = \frac{1}{2}x^\top Bx$ for $x \in \mathbb{R}^d$ and B a positive definite matrix. The algorithms we use (Hamiltonian Monte Carlo with a leapfrog integrator) are basis invariant, and so for ease of notation we will assume throughout that B is diagonal with $B_{ii} = \omega_i^2$ for each $i \in [d]$. As input, we are given bounds $0 < \alpha \leq \beta$ such that $\alpha I \preceq B \preceq \beta I$, or, equivalently, $\alpha \leq \omega_i^2 \leq \beta$. The condition number of B is upper bounded by $\kappa = \beta/\alpha$ and we will also call this the condition number of π. We assume *first-order query access* to f, which means that a single query at a point $x \in \mathbb{R}^d$ provides both $f(x)$ and $\nabla f(x) = Bx$. The goal is to return a sample from a distribution that is ε-close to π in total variation distance, while making a minimal number of gradient queries to f.

7. This is without loss of generality. Using $\widetilde{O}(\sqrt{\kappa})$ gradient queries we can always determine the mean up to high precision and then translate the Gaussian to the origin.

2.2 Markov chains on $\mathbb{R}^d$

Throughout we work with Markov chains whose behaviour can be described as follows: when at $x \in \mathbb{R}^d$ move to $y \in \mathbb{R}^d$ with probability density $T(x,y) \geq 0$. We identify the Markov chain with the transition kernel (density) $T : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$. For a fixed $x \in \mathbb{R}^d$ we use T_x to denote the probability distribution on $\mathbb{R}^d$ with density $T(x,\cdot)$. Similarly (with some abuse of notation), we denote by T_μ the probability distribution on $\mathbb{R}^d$ with density $\int \mu(x)T(x,\cdot)\,\mathrm{d}x$. The K-step transition kernel T^K is defined recursively via $T^K(x,y) = \int_{\mathbb{R}^d} T^{K-1}(x,z)T(z,y)\,\mathrm{d}z$ for $K > 1$. We say that T satisfies the *detailed balance condition* with respect to the probability density $\pi : \mathbb{R}^d \to \mathbb{R}_+$ if

$$\pi(x)T(x,y) = \pi(y)T(y,x) \quad \text{for all } x, y \in \mathbb{R}^d.$$

The associated Markov chain is called *reversible*.

2.3 Hamiltonian dynamics, harmonic oscillator and leapfrog integrator

At its core, Hamiltonian Monte Carlo makes moves by integrating Hamiltonian dynamics. In general, these describe the evolution of a physical system parameterized by (generalized) *positions* and (generalized) *momenta*. For the purposes of this paper, we denote the former with $x \in \mathbb{R}^d$ and the latter with $v \in \mathbb{R}^d$. We sometimes refer to v as the *velocity*, which in classical physics is equal to the momentum of a unit mass. The Hamiltonian evolution of a d-dimensional system is governed by its Hamiltonian $\mathcal{H} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, which can be understood as the total energy of the system at position $x \in \mathbb{R}^d$ and with velocity $v \in \mathbb{R}^d$. The evolution of the system is described by the following equations:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial \mathcal{H}(x,v)}{\partial v}, \quad \frac{\mathrm{d}v}{\mathrm{d}t} = -\frac{\partial \mathcal{H}(x,v)}{\partial x}.$$

The simplest example is the (one-dimensional) harmonic oscillator with Hamiltonian $\mathcal{H}(x,v) = \frac{1}{2}\omega^2 x^2 + \frac{1}{2}v^2$ for some given $\omega > 0$. Its evolution is described by $\frac{\mathrm{d}x}{\mathrm{d}t} = v$ and $\frac{\mathrm{d}v}{\mathrm{d}t} = -\omega^2 x$, which can be solved analytically to yield

$$\begin{bmatrix} x(t) \\ v(t) \end{bmatrix} = \begin{bmatrix} \cos(\omega t) & \frac{1}{\omega}\sin(\omega t) \\ -\omega \sin(\omega t) & \cos(\omega t) \end{bmatrix} \begin{bmatrix} x(0) \\ v(0) \end{bmatrix} \tag{1}$$

A more interesting example is the d-dimensional harmonic oscillator. For a given positive (semi-)definite matrix $B \in \mathbb{R}^{d\times d}$, its Hamiltonian is $\mathcal{H}(x,v) = \frac{1}{2}x^\top Bx + \frac{1}{2}v^\top v$, and its evolution is described by

$$\frac{\mathrm{d}x}{\mathrm{d}t} = v, \quad \frac{\mathrm{d}v}{\mathrm{d}t} = -Bx. \tag{2}$$

If B has eigenvalues ω_i^2 then in the eigenbasis of B the system effectively decomposes into d independent harmonic oscillators with frequencies ω_i.

2.3.1 Leapfrog integrator

The leapfrog integrator, also known as the Störmer-Verlet method, is a well-known numerical integrator for Hamiltonian dynamics that uses two queries to $\frac{\partial \mathcal{H}(x,v)}{\partial x}$ in each integration

step. In the Gaussian case we have $\mathcal{H}(x,v) = \frac{1}{2}x^\top Bx + \frac{1}{2}v^\top v$ and a single step of the leapfrog integrator takes the following closed form:[8]

$$\begin{bmatrix} x^{(n+1)} \\ v^{(n+1)} \end{bmatrix} = \begin{bmatrix} I - \frac{\delta^2}{2}B & \delta I \\ -\delta B(I - \frac{\delta^2}{4}B) & I - \frac{\delta^2}{2}B \end{bmatrix} \begin{bmatrix} x^{(n)} \\ v^{(n)} \end{bmatrix}, \tag{3}$$

where $\delta > 0$ is a parameter used to describe the integration time. See for example (Leimkuhler and Reich, 2005, Sec. 2.6) for details. Following that reference we will exploit that, similarly as for the idealized Hamiltonian dynamics, the leapfrog dynamics decouple in the diagonal basis of B. Hence, for the analysis we can assume (without loss of generality) that B is diagonal with entries $0 < \alpha \leq \omega_i^2 \leq \beta$, and the leapfrog integrator can be interpreted as integrating d independent harmonic oscillators. Effectively, this corresponds to analyzing the leapfrog dynamics in the diagonalizing basis.[9] We can then understand the leapfrog integrator by restricting to a single harmonic oscillator with parameter ω.

The propagator from Eq. (3) has eigenvalues

$$\lambda^{\pm} = 1 - \frac{\delta^2\omega^2}{2} \pm \mathbf{i}\,\delta\omega\sqrt{1 - \frac{\delta^2\omega^2}{4}}.$$

If $\delta^2\omega^2 \leq 4$, we can set $\lambda^{\pm} = e^{\pm \mathbf{i}\,\varphi}$, where $\varphi \in [0,\pi]$ is uniquely defined by $\cos(\varphi) = 1 - \frac{\delta^2\omega^2}{2}$ and $\sin(\varphi) = \delta\omega\sqrt{1 - \frac{\delta^2\omega^2}{4}}$. We can use φ to rewrite the propagator as a rotation with angle φ

$$\begin{bmatrix} \cos(\varphi) & \frac{1}{\hat{\omega}}\sin(\varphi) \\ -\hat{\omega}\sin(\varphi) & \cos(\varphi) \end{bmatrix}, \qquad \text{where } \hat{\omega} = \omega\sqrt{1 - \frac{\delta^2\omega^2}{4}}.$$

Comparing this with (1), we see that the leapfrog trajectory *exactly* follows the Hamiltonian dynamics for the modified Hamiltonian $\hat{\mathcal{H}}$ given by

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

Indeed, if $(\hat{x}(t), \hat{v}(t))$ is the solution of Hamilton's equations with Hamiltonian $\hat{\mathcal{H}}(x,v)$ and initial conditions $(\hat{x}(0) = x_0, \hat{v}(0) = v_0)$, then the nth point on the leapfrog trajectory equals

$$\begin{bmatrix} \hat{x}^{(n)} \\ \hat{v}^{(n)} \end{bmatrix} = \begin{bmatrix} \cos(n\varphi) & \frac{1}{\hat{\omega}}\sin(n\varphi) \\ -\hat{\omega}\sin(n\varphi) & \cos(n\varphi) \end{bmatrix} \begin{bmatrix} \hat{x}_0 \\ \hat{v}_0 \end{bmatrix} = \begin{bmatrix} \cos(\hat{\omega}t_n) & \frac{1}{\hat{\omega}}\sin(\hat{\omega}t_n) \\ -\hat{\omega}\sin(\hat{\omega}t_n) & \cos(\hat{\omega}t_n) \end{bmatrix} \begin{bmatrix} \hat{x}_0 \\ \hat{v}_0 \end{bmatrix} = \begin{bmatrix} \hat{x}(t_n) \\ \hat{v}(t_n) \end{bmatrix},$$

where $t_n = n\varphi/\hat{\omega}$. We can now easily check that the difference between $\mathcal{H}$ and $\hat{\mathcal{H}}$ is

$$\mathcal{H}(x,v) - \hat{\mathcal{H}}(x,v) = \frac{\delta^2\omega^4x^2}{8}.$$

8. While the previous section gives an explicit form for the exact Hamiltonian dynamics as a function of the matrix B, in our algorithmic application we will not have direct access to B but only to gradient queries of the form Bx. This allows us to implement the leapfrog integrator without explicitly learning B.
9. We stress that this is only a (standard) trick for the analysis – it also appears in e.g. the reference work (Leimkuhler and Reich, 2005, Sec. 2.6)). The algorithms themselves work for general non-diagonal B, and they will only require the aforementioned bounds α and β on the eigenvalues ω_i^2.

By our former remark, this observation extends to general d-dimensional harmonic oscillators and the corresponding leapfrog integrator (3): we define $\hat{B}$ by replacing ω_i by $\hat{\omega}_i$ for each eigenvalue of B, where

$$\hat{\omega}_i := \omega_i \sqrt{1 - \frac{\delta^2 \omega_i^2}{4}}, \tag{4}$$

and we set $\hat{\mathcal{H}}(x,v) = \frac{1}{2} x^\top \hat{B} x + \frac{1}{2} v^\top v$. The leapfrog integrator is then an exact integrator for $\hat{\mathcal{H}}$ and we have that

$$\mathcal{H}(x,v) - \hat{\mathcal{H}}(x,v) = \frac{\delta^2}{8} \sum_{i \in [d]} \omega_i^4 x_i^2. \tag{5}$$

Finally we introduce the following notation: the tuple $(x', v') = \text{leapfrog}(x, v, t, \delta)$ is defined as the (position, momentum)-vector after taking t/δ leapfrog integration steps for Hamiltonian $\mathcal{H}$ with stepsize $0 \leq \delta \leq 1/\sqrt{\beta}$.[10]

3. Idealized and unadjusted HMC

We first analyze an idealized version of HMC, Algorithm 1, where we assume that we can exactly integrate the Hamiltonian dynamics. We use long and random integration times. In order to later apply the results from this section in the setting of a numerical integrator, we will use uniformly random integration times $t \sim U(\mathcal{T})$ from a *finite* set $\mathcal{T}$. We will require only that, for all $0 < \alpha \leq \omega^2 \leq \beta$, if t is chosen uniformly at random from $U(\mathcal{T})$ then with probability at least $1/2$ it holds that $|\cos(\omega t)| \leq 0.9$ (we denote this by $\mathbb{P}_{t \sim U(\mathcal{T})}\big[|\cos(\omega t)| \leq 0.9\big] \geq 1/2$). In the following lemma we show that this is satisfied for a simple choice of $\mathcal{T}$.

Lemma 2 *Let $0 < \sqrt{\alpha} \leq \sqrt{\beta}$. If $0 < \delta \leq \pi/(20\sqrt{\beta})$ and*

$$\mathcal{T} = \{k \cdot \delta \mid k \in \mathbb{N},\ k \cdot \delta < 10\pi/\sqrt{\alpha}\} \tag{6}$$

then we have for all $\omega \in [\sqrt{\alpha}, \sqrt{\beta}]$ that

$$\mathbb{P}_{t \sim U(\mathcal{T})}\big[|\cos(\omega t)| \leq 0.9\big] \geq 1/2. \tag{7}$$

Proof First, we prove that if $\zeta > \eta \geq 0$, $\omega > 0$, and $\widetilde{\mathcal{T}} = \{\eta + n\zeta : n \in \mathbb{N}, \eta + n\zeta \leq \frac{\pi}{2\omega}\}$ with $|\widetilde{\mathcal{T}}| \geq 10$, then for t chosen uniformly from $\widetilde{\mathcal{T}}$ we have

$$\mathbb{P}_{t \sim U(\widetilde{\mathcal{T}})}\left\{|\cos(\omega t)| \leq 0.9\right\} \geq 3/5. \tag{8}$$

To see this, note that $\zeta \leq \left\lfloor \frac{\pi}{2\omega(|\widetilde{\mathcal{T}}|-1)} \right\rfloor$ implies that

$$\begin{aligned}
\mathbb{P}_{t \sim U(\widetilde{\mathcal{T}})}\left\{|\cos(\omega t)| \leq 0.9\right\} &= \mathbb{P}_{t \sim U(\widetilde{\mathcal{T}})}\left\{t \geq \frac{1}{\omega}\arccos(0.9)\right\} \\
&\geq \frac{1}{|\widetilde{\mathcal{T}}|}\left\lfloor \frac{\pi/2 - \arccos(0.9)}{\omega\zeta} \right\rfloor \\
&\geq \frac{1}{|\widetilde{\mathcal{T}}|}\left\lfloor \big(1 - 2\arccos(0.9)/\pi\big)\,(|\widetilde{\mathcal{T}}| - 1) \right\rfloor.
\end{aligned}$$

10. We will always apply this with $t/\delta \in \mathbb{N}$.

The last quantity is at least $3/5$ for $|\widetilde{\mathcal{T}}| \geq 10$.

We now make use of the above to show the desired bound for the set $\mathcal{T}$ defined in Eq. (6). Let ω be such that $\sqrt{\alpha} \leq \omega \leq \sqrt{\beta}$. Note that $|\cos(\omega t)|$ is periodic with period $\frac{\pi}{2\omega}$. We write $\mathcal{T}$ as the disjoint union

$$\mathcal{T} = \bigcup_{n=1}^{N} \left(\mathcal{T} \cap [\frac{(n-1)\pi}{2\omega}, \frac{n\pi}{2\omega}] \right)$$

where N is the least integer such that $\frac{N\pi}{2\omega} > \frac{10\pi}{\sqrt{\alpha}}$, i.e., $N = \left\lfloor \frac{20\omega}{\sqrt{\alpha}} \right\rfloor$. Note that $N \geq 20$. Since $\delta \leq \frac{\pi}{20\sqrt{\beta}}$ and $\omega \leq \sqrt{\beta}$, the first $N-1$ such intervals contain at least

$$\left\lfloor \frac{\pi}{2\omega\delta} \right\rfloor \geq 10$$

equally spaced points. Now note that the subset $\mathcal{T} \cap [\frac{(n-1)\pi}{2\omega}, \frac{n\pi}{2\omega}])$ takes precisely the form as considered at the start of the proof, and we just proved that $|\mathcal{T} \cap [\frac{(n-1)\pi}{2\omega}, \frac{n\pi}{2\omega}])| \geq 10$. Hence, Eq. (8) shows that for each of these $N-1$ intervals we have

$$\mathbb{P}_{t \sim U(\mathcal{T} \cap [\frac{(n-1)\pi}{2\omega}, \frac{n\pi}{2\omega}])} \big[|\cos(\omega t)| \leq 0.9 \big] \geq \frac{3}{5}.$$

Given that there are $N \geq 20$ intervals in total, we get

$$\mathbb{P}_{t \sim U(\mathcal{T})} \big[|\cos(\omega t)| \leq 0.9 \big] \geq \frac{N-1}{N} \frac{3}{5} \geq \frac{19}{20} \frac{3}{5} \geq \frac{1}{2},$$

where the first inequality comes from the fact that with probability at least $(N-1)/N$ we pick a "good" interval, and conditioned on that we have that $|\cos(\omega t)| \leq 0.9$ with probability at least $3/5$. ■

We formulate the HMC algorithm using this definition of $\mathcal{T}$ as Algorithm 1.

Algorithm 1: Markov kernel P (idealized HMC with random integration time)

Input: $x \in \mathbb{R}^d$, stepsize $\delta \leq \frac{\pi}{20\sqrt{\beta}}$, $\mathcal{T} = \{k \cdot \delta \mid k \in \mathbb{N},\ k \cdot \delta < 10\pi/\sqrt{\alpha}\}$ as in Eq. (6)

Output: $x' \in \mathbb{R}^d$

1 Draw $v \sim \mathcal{N}(0, I_d)$ and $t \sim U(\mathcal{T})$;

2 Define x' by following Hamiltonian dynamics for $\mathcal{H}$ for time t, starting from (x, v);

It is well known that idealized HMC with a fixed integration time has the desired stationary distribution π whose density at (x, v) is related to the Hamiltonian $\mathcal{H}(x, v) = \frac{1}{2} x^\top \operatorname{diag}(\boldsymbol{\omega}^2) x + \frac{1}{2} v^\top v$, i.e., $\pi(x, v) \propto \exp(-\mathcal{H}(x, v))$ (cf. Duane et al. (1987); Neal (1996); Vishnoi (2021)). From this it follows that also P has stationary distribution π. In Section 3.1 we show that P has a small mixing time. We then extend this result to the setting where we use a numerical integrator (leapfrog) instead of the idealized time evolution according to Hamiltonian dynamics. For this we use the fact (cf. Section 2.3.1) that the leapfrog

integrator applied to $\mathcal{H}(x,v)$ can be viewed as an exact integrator for the Hamiltonian dynamics of a modified Hamiltonian $\hat{\mathcal{H}}(x,v)$. By bounding the distance between π and $\hat{\pi} \propto \exp(-\hat{\mathcal{H}}(x,v))$, we output a distribution that is ε-close to π in total variation distance using a number of gradient evaluations that scales as $\widetilde{O}(\sqrt{\kappa}d^{1/4}/\sqrt{\varepsilon})$, see Section 3.2.

3.1 Idealized HMC

Let P_x^t denote the density function of the proposal distribution from $x \in \mathbb{R}^d$, conditioned on having picked $t \in [0,T]$. Using the explicit expression Eq. (1), we can relate this to the density function of the standard Gaussian: the momentum term v needs to satisfy $\cos(\omega_i t)x_i + \frac{1}{\omega_i}\sin(\omega_i t)v_i = z_i \ \forall i \in [d]$. This gives us the density function P_x^t as

$$P_x^t(z) = (2\pi)^{-d/2} \prod_{i\in[d]} \frac{\omega_i}{|\sin(\omega_i t)|} \exp\left(-\frac{1}{2}\left(\frac{z_i - \cos(\omega_i t)x_i}{\frac{1}{\omega_i}\sin(\omega_i t)}\right)^2\right). \tag{9}$$

The probability density with which idealized HMC moves from x to z is then given by $P_x(z) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} P_x^t(z)$.

We analyze the convergence in total variation distance by explicitly writing out the distribution P^K obtained by taking K steps of the idealized HMC method. If we condition on the choice of random integration times in step 2 of Algorithm 1, then the resulting distribution is again a normal distribution. Indeed, let $(v^{(1)},\dots,v^{(K)})$, $(t_1,\dots,t_K)$ and $(x^{(1)},\dots,x^{(K)})$ denote the velocities, integration times and positions, respectively, encountered during the first K steps. By repeatedly applying (9), we can express

$$\begin{aligned} x_i^{(K)} &= x_i^{(K-1)}\cos(\omega_i t_K) + \frac{1}{\omega_i}\sin(\omega_i t_K)v_i^{(K)} \\ &= x_i^{(0)}\left(\prod_{k=1}^K \cos(\omega_i t_k)\right) + \frac{1}{\omega_i}\sum_{k=1}^K v_i^{(k)}\sin(\omega_i t_k)\left(\prod_{j=k+1}^K \cos(\omega_i t_j)\right). \end{aligned}$$

For a fixed tuple $\boldsymbol{t} = (t_1,\dots,t_K) \in \mathcal{T}^K$ of integration times, but *random* choices $(v^{(1)},\dots,v^{(K)}) \sim \mathcal{N}(0,I_d)^K$ of momenta, we can argue that this describes a Gaussian distribution, which we denote by $P_x^{\boldsymbol{t}}$. First, note that $P_x^{\boldsymbol{t}}$ is a product distribution: $P_x^{\boldsymbol{t}}(z) = \prod_{i\in[d]} P_x^{\boldsymbol{t},i}(z_i)$ where we use $P_x^{\boldsymbol{t},i}$ for the marginal distribution of $P_x^{\boldsymbol{t}}$ with respect to the i-th coordinate. Then, note that $P_x^{\boldsymbol{t},i}$ describes a sum of Gaussians with the same mean, and hence forms again a Gaussian. We formalize this in the next lemma.

Lemma 3 *Let $\boldsymbol{t} \in \mathcal{T}^K$, $\omega > 0$, $x \in \mathbb{R}$, and consider*

$$z = x\left(\prod_{k=1}^K \cos(\omega t_k)\right) + \frac{1}{\omega}\sum_{k=1}^K v^{(k)}\sin(\omega t_k)\left(\prod_{j=k+1}^K \cos(\omega t_j)\right)$$

where $v^{(k)} \sim \mathcal{N}(0,1)$ for each $k \in [K]$. Then $z \sim \mathcal{N}(x\prod_{k=1}^K \cos(\omega t_k), \frac{1}{\omega^2}(1-\prod_{k=1}^K \cos(\omega t_j)^2))$.

Proof It is clear that $\mathbb{E}[z] = x\prod_{k=1}^{K}\cos(\omega t_k)$. The sum of Gaussian random variables is again distributed according to a Gaussian whose variance is the sum of the individual variances. That is,

$$\begin{aligned}\mathbb{E}[(z-\mathbb{E}[z])^2] &= \frac{1}{\omega^2}\sum_{k=1}^{K}\sin(\omega t_k)^2\left(\prod_{j=k+1}^{K}\cos(\omega t_j)^2\right)\\ &= \frac{1}{\omega^2}\sum_{k=1}^{K}(1-\cos(\omega t_k)^2)\left(\prod_{j=k+1}^{K}\cos(\omega t_j)^2\right)\\ &= \frac{1-\prod_{j=1}^{K}\cos(\omega t_j)^2}{\omega^2}.\end{aligned}$$

■

If the term $\prod_{k=1}^{K}\cos(\omega t_k)$ is sufficiently small, then $P_x^{\boldsymbol{t}}$ is close to π. Lemma 2 and Hoeffding's inequality show that for a random tuple $\boldsymbol{t} = (t_1,\ldots,t_K) \sim U(\mathcal{T}^K)$ this term will indeed be small. Then we use this to prove convergence of the proposal distribution to π. We mention for completeness that we have made no effort to determine the optimal constants: the rate $0.9^{K/4}$ below suffices for our purposes, but can likely be improved.

Lemma 4 *Let $0 < \alpha \le \omega^2 \le \beta$ and $\mathcal{T}$ as in Lemma 2. Then*

$$\mathbb{P}_{\boldsymbol{t}\sim U(\mathcal{T}^K)}\left[\left|\prod_{k=1}^{K}\cos(\omega t_k)\right| \ge 0.9^{K/4}\right] \le \exp(-K/8).$$

Proof Let $\boldsymbol{t} = (t_1,\ldots,t_K)$ with $t_k \sim U(\mathcal{T})$, and define the i.i.d. Boolean variables X_k as indicating whether $|\cos(\omega t_k)| \le 0.9$. Define $\rho = \mathbb{P}[X_k = 1]$. By Lemma 2 we know that $\rho \ge 1/2$. By the multiplicative Chernoff bound this implies that

$$\mathbb{P}\left[\sum_{k=1}^{K}X_k \le \frac{K}{4}\right] \le \mathbb{P}\left[\sum_{k=1}^{K}X_k \le \frac{K\rho}{2}\right] \le \exp\left(-\frac{K}{8}\right).$$

It remains to note that if $\sum_{k=1}^{K}X_k > K/4$ then $\left|\prod_{k=1}^{K}\cos(\omega t_k)\right| < 0.9^{K/4}$, and this implies that

$$\mathbb{P}_{\boldsymbol{t}\sim U(\mathcal{T}^K)}\left[\left|\prod_{k=1}^{K}\cos(\omega t_k)\right| \ge 0.9^{K/4}\right] \le \mathbb{P}\left[\sum_{k=1}^{K}X_k \le \frac{K}{4}\right].$$

■

Using the above lemma, we show that the proposal distributions $P_x^K(z) := P^K(x,z)$ and $P_y^K := P^K(y,z)$ are close provided that x and y are close.

Proposition 5 *For every $x, y \in \mathbb{R}^d$, if*

$$K \geq 38 \log\left(\frac{d(2+\sqrt{\beta}\|x-y\|_\infty)}{\varepsilon}\right),$$

then, with P the kernel of idealized HMC, we have

$$\|P_x^K - P_y^K\|_{\mathrm{TV}} \leq \varepsilon.$$

We remark that the lower bound on K is stated in a coordinate-dependent way through the use of $\|\cdot\|_\infty$ (which should be interpreted in the eigenbasis of B), however, it can be made coordinate-independent by using the 2-norm instead and the inequality $\|\cdot\|_\infty \leq \|\cdot\|_2$.

Proof Recall that $P_x^K = \frac{1}{|\mathcal{T}|^K}\sum_{\boldsymbol{t}\in\mathcal{T}^K} P_x^{\boldsymbol{t}}$ and $P_x^{\boldsymbol{t}} = \prod_{i\in[d]} P_x^{\boldsymbol{t},i}$ is a product distribution. Hence, we can twice apply a triangle inequality to obtain

$$\begin{aligned}\|P_x^K - P_y^K\|_{\mathrm{TV}} &\leq \frac{1}{|\mathcal{T}|^K}\sum_{\boldsymbol{t}\in\mathcal{T}^K}\|P_x^{\boldsymbol{t}} - P_y^{\boldsymbol{t}}\|_{\mathrm{TV}} \\ &\leq \sum_{i\in[d]}\frac{1}{|\mathcal{T}|^K}\sum_{\boldsymbol{t}\in\mathcal{T}^K}\|P_x^{\boldsymbol{t},i} - P_y^{\boldsymbol{t},i}\|_{\mathrm{TV}} \qquad (10)\end{aligned}$$

Now let $\delta = \frac{1}{\sqrt{2}}\min\left\{1, \frac{\varepsilon}{2d(2+\sqrt{\beta}\|x-y\|_\infty)}\right\}$ and $K \geq 38\log(1/\delta)$. We will invoke Lemma 4. By our choice of K we have that $0.9^{K/4} \leq 0.9^{38\log(1/\delta)} \leq \exp(-\log(1/\delta)) \leq \delta$, where the second inequality follows from $0.9^{38} < 1/e$, and $\exp(-K/8) \leq \varepsilon/(2d)$, and so the lemma ensures that

$$\mathbb{P}_{\boldsymbol{t}\sim U(\mathcal{T}^K)}\left[\left|\prod_{k=1}^K \cos(\omega_i t_k)\right| \geq \delta\right] \leq \frac{\varepsilon}{2d}$$

for each $i \in [d]$. Hence for each coordinate $i \in [d]$ we have

$$\begin{aligned}\frac{1}{|\mathcal{T}|^K}\sum_{\boldsymbol{t}\in\mathcal{T}^K}\|P_x^{\boldsymbol{t},i} - P_y^{\boldsymbol{t},i}\|_{\mathrm{TV}} &\leq \frac{\varepsilon}{2d} + \frac{1}{|\mathcal{T}|^K}\sum_{\boldsymbol{t}\in\mathcal{T}^K:|\prod_{k=1}^K\cos(\omega_i t_k)|\leq\delta}\|P_x^{\boldsymbol{t},i} - P_y^{\boldsymbol{t},i}\|_{\mathrm{TV}} \\ &\leq \frac{\varepsilon}{2d} + (1-\frac{\varepsilon}{2d})|x_i - y_i|\delta\sqrt{2}\omega_i \leq \varepsilon, \qquad (11)\end{aligned}$$

where we use that for $\boldsymbol{t} \in \mathcal{T}^K$ for which $\left|\prod_{k=1}^K \cos(\omega_i t_j)\right| \leq \delta \leq \frac{1}{\sqrt{2}}$, the proposal distributions $P_x^{\boldsymbol{t},i}$ and $P_y^{\boldsymbol{t},i}$ are univariate Gaussians with means μ_x, μ_y that satisfy $|\mu_x - \mu_y| \leq \delta|x_i - y_i|$, and both have variance $\sigma^2 \geq \frac{1-\delta^2}{\omega_i^2} \geq \frac{1}{2\omega_i^2}$. (For univariate Gaussians one has $\|\mathcal{N}(\mu,\sigma^2) - \mathcal{N}(\mu_2,\sigma^2)\|_{\mathrm{TV}} < |\mu_1 - \mu_2|/\sigma$.) Combining Eqs. (10) and (11) we obtain $\|P_x^K - P_y^K\|_{\mathrm{TV}} \leq \varepsilon$. ■

This bound then easily leads to a bound on the total variation distance between P_x^K and π for x that is sufficiently close to 0, and this is the main conclusion of this section.

Theorem 6 (Idealized HMC) *There exists a constant $C > 0$ such that for every $x \in \mathbb{R}^d$, if*

$$K \geq C \log\left(\frac{d\kappa(\sqrt{\alpha}\|x\|_\infty + 1)}{\varepsilon}\right),$$

then, with $\pi \propto \exp(-\frac{1}{2}x^\top Bx)$ and P the kernel of idealized HMC, we have

$$\|P_x^K - \pi\|_{\mathrm{TV}} \leq \varepsilon.$$

Proof We write $\pi = \int_{\mathbb{R}^d} \delta_y \,\mathrm{d}\pi(y)$. Using that π is stationary for P (and hence P^K), we also have that $\pi = \int_{\mathbb{R}^d} P_y^K \,\mathrm{d}\pi(y)$. Now we apply Jensen's inequality:

$$\begin{aligned}\|P_x^K - \pi\|_{\mathrm{TV}} &\leq \int_{y\in\mathbb{R}^d} \|P_x^K - P_y^K\|_{\mathrm{TV}} \,\mathrm{d}\pi(x) \\ &\leq \pi(\{y : \|y\| > \eta\}) + \int_{y\in\mathbb{R}^d:\|y\|\leq\eta} \|P_x^K - P_y^K\|_{\mathrm{TV}} \,\mathrm{d}\pi(x).\end{aligned}$$

We use Lemma 11 to choose an η that is sufficiently large to ensure that $\pi(\{y : \|y\| > \eta\}) \leq \varepsilon/2$. In particular, using the notation of that lemma, for $\gamma = \Theta(\log(1/\varepsilon))$ we know that $\pi(E_\gamma) \geq 1 - \varepsilon/2$, and we can bound the norm of each $y \in E_\gamma$ as

$$\alpha^2\|y\|^2 \leq y^\top \operatorname{diag}(\boldsymbol{\omega})^4 y \leq \sum_i \omega_i^2 + \gamma\sqrt{\sum_i \omega_i^4} \leq d\beta + \gamma\beta\sqrt{d},$$

which yields the bound $\|y\| \leq \sqrt{\frac{(\gamma+1)\kappa d}{\alpha}}$ for $y \in E_\gamma$. We use this to bound the quantity $\frac{d\sqrt{\beta}\|x-y\|_\infty}{\varepsilon}$ as follows

$$\begin{aligned}\frac{d\sqrt{\beta}\|x-y\|_\infty}{\varepsilon} &\leq \frac{d\sqrt{\beta}(\|x\|_\infty + \sqrt{(\gamma+1)\kappa d/\alpha})}{\varepsilon} \\ &\leq \frac{d\sqrt{\kappa}(\sqrt{\alpha}\|x\|_\infty + \sqrt{(\gamma+1)\kappa d})}{\varepsilon} \\ &\leq \frac{d^{3/2}\kappa\sqrt{(\gamma+1)}(\sqrt{\alpha}\|x\|_\infty + 1)}{\varepsilon}.\end{aligned}$$

Here the last inequality is a rather crude upper bound that only serves to show that there exists a $C > 0$ such that for $K \geq C\log(\frac{d\kappa(\sqrt{\alpha}\|x\|_\infty+1)}{\varepsilon})$ we have $K \geq 38\log\left(\frac{d(2+\sqrt{\beta}\|x-y\|_\infty)}{\varepsilon/2}\right)$. With such a bound on K, Proposition 5 implies that $\|P_x^K - P_y^K\|_{\mathrm{TV}} \leq \varepsilon/2$ for all $x, y \in \mathbb{R}^d$ with $\|x-y\|_\infty \leq \eta + \|x\|_\infty$. Combining these two bounds shows that $\|P_x^K - \pi\|_{\mathrm{TV}} \leq \varepsilon$. ■

3.2 Unadjusted HMC

The results from the previous section extend from the idealized setting where one can integrate exactly, to the setting where one uses the leapfrog integrator.

Algorithm 2: Markov kernel $\hat{Q}$ (leapfrog HMC with random integration time)

Input: $x \in \mathbb{R}^d$, stepsize $\delta \leq \frac{\pi}{20\sqrt{\beta}}$, $\mathcal{T} = \{k \cdot \delta \mid k \in \mathbb{N},\ k \cdot \delta < 10\pi/\sqrt{\alpha}\}$ as in Eq. (6)

Output: $x' \in \mathbb{R}^d$

1 Draw $v \sim \mathcal{N}(0, I_d)$ and move from x to (x, v) ;

2 Draw $t \sim U(\mathcal{T})$ and set $(x', v') = \text{leapfrog}(x, v, t, \delta)$;

As discussed in Section 2.3.1, the leapfrog dynamics correspond to Hamiltonian dynamics for a slightly modified Hamiltonian $\hat{\mathcal{H}}$. Bounding the distance between the stationary distribution $\hat{\pi}$ and π leads to the following poly$(1/\varepsilon)$-algorithm for sampling from a distribution ε-close to π.

Proposition 7 (Unadjusted HMC) *There exist constants $C, C' > 0$ such that for every $x \in \mathbb{R}^d$, if*

$$K \geq C \log\left(\frac{d\kappa(\sqrt{\alpha}\|x\|_\infty + 1)}{\varepsilon}\right) \quad \text{and} \quad \delta \leq C' \frac{\sqrt{\varepsilon}}{\sqrt{\beta} d^{1/4}},$$

then

$$\|\hat{Q}_x^K - \pi\|_{\text{TV}} \leq \varepsilon$$

where $\pi(x) \propto \exp(-\frac{1}{2}x^\top B x)$ and $\hat{Q}$ is the kernel of the unadjusted leapfrog HMC chain with step size δ. A sample from $\hat{Q}_x^K$ can be obtained using $O(\frac{\sqrt{\kappa} d^{1/4} K}{\sqrt{\varepsilon}})$ gradient evaluations.

Proof By our discussion of the leapfrog integrator in Section 2.3.1, we know that $\hat{Q}$ corresponds to the idealized HMC algorithm for the modified Hamiltonian $\hat{\mathcal{H}}$. Here we assume $\delta^2 \omega_i^2 \leq 4$ for all $i \in [d]$, i.e., $\delta \leq \frac{1}{\sqrt{\beta}}$. It thus follows from Theorem 6 that if we start from $x \in \mathbb{R}^d$ and take $K \geq C \log\left(\frac{d\kappa(\sqrt{\alpha}\|x\|_\infty + 1)}{\varepsilon}\right)$ steps of the chain $\hat{Q}$, for an appropriate constant $C > 0$, then it returns a distribution that is $\varepsilon/2$-close to the modified stationary $\hat{\pi}$ defined as

$$\hat{\pi}(x) \propto \exp(-\frac{1}{2}x^\top \hat{B} x).$$

Using that $\hat{\pi}$ and π are both multivariate Gaussians, one can show (see Lemma 8 below for completeness)

$$\|\pi - \hat{\pi}\|_{\text{TV}} \leq \frac{3}{8}\delta^2 \sqrt{\sum_i \omega_i^4} \leq \frac{3}{8}\delta^2 \beta \sqrt{d}.$$

Hence by choosing a sufficiently small stepsize $\delta \in O(\sqrt{\varepsilon}/(\sqrt{\beta} d^{1/4}))$, we have that $\|\hat{\pi} - \pi\|_{\text{TV}} \leq \varepsilon/2$. Together this shows that the resulting distribution after K steps will be ε-close to π.

It remains to bound the complexity of the algorithm. A single leapfrog step requires 2 gradient evaluations, and so a single step of the Markov chain $\hat{Q}$ requires $t/\delta \in O(\sqrt{\kappa} d^{1/4}/\sqrt{\varepsilon})$

gradient evaluations. Applying K steps of the Markov chain yields a total number of gradient evaluations

$$O\left(\frac{\sqrt{\kappa}d^{1/4}K}{\sqrt{\varepsilon}}\right).$$

■

Lemma 8 *Let* $\pi(x) \propto \exp(-x^\top \operatorname{diag}(\boldsymbol{\omega})x/2)$, $\hat{\omega}_i = \omega_i\sqrt{1-\frac{\delta^2\omega_i^2}{4}}$ *and* $\hat{\pi}(x) \propto \exp(-x^\top \operatorname{diag}(\hat{\boldsymbol{\omega}})x/2)$. *Then*

$$\|\pi - \hat{\pi}\|_{\mathrm{TV}} \le \frac{3}{8}\delta^2\sqrt{\sum_i \omega_i^4} \le \frac{3}{8}\delta^2\beta\sqrt{d}.$$

Proof For multivariate mean-zero Gaussians we have the following bound by (Devroye et al., 2022, Theorem 1.1):

$$\left\|\mathcal{N}(0,\Sigma_1) - \mathcal{N}(0,\Sigma_2)\right\|_{\mathrm{TV}} \le \frac{3}{2}\min\left\{1, \left\|\Sigma_1^{-1}\Sigma_2 - I\right\|_F\right\}. \tag{12}$$

Applying this bound for $\Sigma_1 = \operatorname{diag}(\hat{\boldsymbol{\omega}})$ and $\Sigma_2 = \operatorname{diag}(\boldsymbol{\omega})$ we get

$$\|\pi - \tilde{\pi}\|_{\mathrm{TV}} \le \frac{3}{2}\sqrt{\sum_i\left(\left(1-\frac{\delta^2\omega_i^2}{4}\right)-1\right)^2} = \frac{3}{8}\delta^2\sqrt{\sum_i \omega_i^4} \le \frac{3}{8}\delta^2\beta\sqrt{d}.$$

■

4. Metropolis-Adjusted HMC

Here we study the Metropolis-adjusted HMC algorithm. The algorithm applies a Metropolis filter to correct for the numerical errors of the integrator. This ensures that the algorithm has the correct stationary distribution, and leads to an overall improved error dependence.

Algorithm 3: Markov kernel Q (Adjusted leapfrog HMC with random integration time)

Input: $x \in \mathbb{R}^d$, stepsize $\delta \in O(1/(\sqrt{\beta}d^{1/4}))$, $\mathcal{T} := \{k\cdot\delta \mid k\in\mathbb{N},\ k\cdot\delta < 10\pi/\sqrt{\alpha}\}$
Output: $x' \in \mathbb{R}^d$

1 Draw $v \sim \mathcal{N}(0, I_d)$ and move from x to (x, v) ;
2 Draw $t \sim U(\mathcal{T})$ and set $(x', v') = \text{leapfrog}(x, v, t, \delta)$;
3 Accept with probability

$$\mathcal{A}(x,x') := \min\left\{1, \exp\big(-\mathcal{H}(x',-v') + \mathcal{H}(x,v)\big)\right\}$$

and return x'. Otherwise return $x' = x$;

We make a few (standard) observations about the adjusted HMC algorithm, whose proofs we defer to Appendix A.

Lemma 9 *The Markov kernel Q defined in Algorithm 3 has the following properties:*

1. *Kernel Q is reversible with respect to the stationary distribution $\pi(x) \propto \exp(-\frac{1}{2}x^\top Bx)$.*
2. *The acceptance probability is a function of only x and x':*

$$\mathcal{A}(x,x') = \min\left\{1, \exp\left(-\mathcal{H}(x',-v')+\mathcal{H}(x,v)\right)\right\} = \min\left\{1, \exp\left(\frac{\delta^2}{8}\sum_{i\in[d]}\omega_i^4(x_i^2-x_i'^2)\right)\right\},$$

and we can rewrite $Q_x(x') = \hat{Q}_x(x')\mathcal{A}(x,x')$ for $x \neq x'$.

4.1 Concentration bounds on high-dimensional Gaussian random variables

Here we use concentration bounds on high-dimensional Gaussians to show that if $x \sim \pi$ or $x \sim \hat{\pi}$ then with high probability the quantity $\sum_{i\in[d]}\omega_i^4 x_i^2$ is close to $\sum_{i\in[d]}\omega_i^2$. We moreover show that in that case $\pi(x)$ and $\hat{\pi}(x)$ differ by at most a small multiplicative factor.

We will use the following version of the Hanson-Wright inequality (Hanson and Wright (1971)) which gives a concentration inequality for quadratic forms of independent Gaussian random variables.

Theorem 10 (Hanson-Wright inequality (Vershynin, 2018, Thrm 6.2.1)) *Let $X = (X_1,\dots,X_d) \in \mathbb{R}^d$ be a random vector with independent $\mathcal{N}(0,1)$ coordinates. Let A be a $d\times d$ matrix. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\Big[|X^\top AX - \mathbb{E}[X^\top AX]| \geq t\Big] \leq 2\exp\left(-C_1\min\left\{\frac{t^2}{C_2^4\|A\|_F^2}, \frac{t}{C_2^2\|A\|}\right\}\right),$$

where $C_1, C_2 > 0$ are constants.[11]

Note that if $X \in \mathbb{R}^d$ is a random vector with independent $\mathcal{N}(0,1)$ coordinates, then so is $Y = UX$ for a rotation matrix U. This rotation-invariance allows us to again assume, for ease of notation, that the input precision matrix $B = \operatorname{diag}(\boldsymbol{\omega})$. For convenience, recall that $\pi(x) = \frac{\prod_i \omega_i}{(2\pi)^{d/2}}\exp\left(-\frac{1}{2}\sum_i x_i^2\omega_i^2\right)$, and (cf. Eq. (4)) that $\hat{\pi}$ is constructed similarly using $\hat{\boldsymbol{\omega}}$ which is defined, for each $i \in [d]$, as $\hat{\omega}_i = \omega_i\sqrt{1-\frac{\delta^2\omega_i^2}{4}}$. We have $\omega_i^2 - \hat{\omega}_i^2 = \frac{1}{4}\delta^2\omega_i^4$. For $\gamma \geq 1$, we define the measurable set

$$E_\gamma := \left\{x \in \mathbb{R}^d \mid \left|x^\top \operatorname{diag}(\boldsymbol{\omega})^4 x - \sum_i \omega_i^2\right| \leq \gamma\sqrt{\sum_{i\in[d]}\omega_i^4}\right\}. \tag{13}$$

The Hanson-Wright inequality gives us the following concentration of measure for π and $\hat{\pi}$.

Lemma 11 *Let $\gamma \geq 1$ and consider E_γ as in Eq. (13) then we have the following:*

11. The theorem holds more generally for independent mean zero *sub-gaussian* variables X_i. The constant K then upper bounds the *sub-gaussian norm* of all X_i.

1. *Let* $\pi(x) \propto \exp(-\frac{1}{2}x^\top \operatorname{diag}(\boldsymbol{\omega})^2 x)$, *then* $\pi(E_\gamma) \geq 1 - 2\exp\big(-C\gamma\big)$ *where* $C > 0$ *is a constant.*
2. *If* $0 < \delta \leq \beta^{-1/2} d^{-1/4}$, *then for* $\hat{\pi}(x) \propto \exp(-\frac{1}{2}x^\top \operatorname{diag}(\hat{\boldsymbol{\omega}})^2 x)$ *we have* $\hat{\pi}(E_\gamma) \geq 1 - 2\exp(-C'\gamma)$ *where* $C' > 0$ *is a constant.*

Proof We first prove the concentration of measure for π. We have

$$\pi(E_\gamma) = \mathbb{P}_{x\sim\pi}\left[\left|x^\top \operatorname{diag}(\boldsymbol{\omega})^4 x - \sum_i \omega_i^2\right| \leq \gamma\sqrt{\sum_{i\in[d]}\omega_i^4}\right]$$
$$= \mathbb{P}_{z\sim\mathcal{N}(0,I_d)}\left[\left|z^\top \operatorname{diag}(\boldsymbol{\omega})^2 z - \sum_i \omega_i^2\right| \leq \gamma\sqrt{\sum_{i\in[d]}\omega_i^4}\right]$$

where we set $z_i = \omega_i x_i$ for each $i \in [d]$ and observe that $z_i \sim \mathcal{N}(0,1)$. We apply Theorem 10 to the vector z, matrix $A = \operatorname{diag}(\boldsymbol{\omega})^2$, $t = \gamma\|A\|_F$, and note that $\|A\|_F \geq \|A\|$ implies the lower bound

$$C_1 \min\left\{\frac{(\gamma\|A\|_F)^2}{C_2^4\|A\|_F^2}, \frac{\gamma\|A\|_F}{C_2^2\|A\|}\right\} \geq C_1 \min\left\{\frac{\gamma^2}{C_2^4}, \frac{\gamma}{C_2^2}\right\} \geq \gamma C_1 \min\{C_2^{-2}, C_2^{-4}\}.$$

Therefore, for $C \leq C_1 \min\{C_2^{-2}, C_2^{-4}\}$ we obtain the desired bound for π.

We now use the same proof strategy to show concentration for $\hat{\pi}$. We have

$$\hat{\pi}(E_\gamma) = \mathbb{P}_{x\sim\hat{\pi}}\left[\left|x^\top \operatorname{diag}(\boldsymbol{\omega})^4 x - \sum_i \omega_i^2\right| \leq \gamma\sqrt{\sum_{i\in[d]}\omega_i^4}\right]$$
$$= \mathbb{P}_{z\sim\mathcal{N}(0,I_d)}\left[\left|z^\top \operatorname{diag}(\boldsymbol{\omega})^4 \operatorname{diag}(\hat{\boldsymbol{\omega}})^{-2} z - \sum_i \omega_i^2\right| \leq \gamma\sqrt{\sum_{i\in[d]}\omega_i^4}\right]$$
$$\geq \mathbb{P}_{z\sim\mathcal{N}(0,I_d)}\left[\left|z^\top \operatorname{diag}(\boldsymbol{\omega})^4 \operatorname{diag}(\hat{\boldsymbol{\omega}})^{-2} z - \sum_i \omega_i^4/\hat{\omega}_i^2\right| \leq \gamma\sqrt{\sum_{i\in[d]}\omega_i^4} - \left|\sum_i \omega_i^2 - \omega_i^4/\hat{\omega}_i^2\right|\right]$$

By definition $\omega_i^4/\hat{\omega}_i^2 = \omega_i^2/(1 - \delta^2\omega_i^2/4)$, and the upper bound on δ implies that $\delta^2\omega_i^2 \leq 2$. Using this bound, we get

$$\left|\sum_i \omega_i^2 - \omega_i^4/\hat{\omega}_i^2\right| = \sum_i \omega_i^2\left(1 - \frac{1}{1 - \delta^2\omega_i^2/4}\right)$$
$$\leq \sum_i \omega_i^2(1 - 1 + \delta^2\omega_i^2/2)$$
$$= \frac{1}{2}\sum_i \delta^2\omega_i^4 \leq \frac{1}{2\sqrt{d}}\sum_i \omega_i^2 \leq \frac{1}{2}\sqrt{\sum_i \omega_i^4}.$$

Again using the fact that $\omega_i^4/\hat{\omega}_i^2 \leq 2\omega_i^2$, we can further lower bound $\hat{\pi}(E_\gamma)$ as follows:

$$\hat{\pi}(E_\gamma) \geq \mathbb{P}_{z\sim\mathcal{N}(0,I_d)}\left[\left|z^\top \operatorname{diag}(\boldsymbol{\omega})^4 \operatorname{diag}(\hat{\boldsymbol{\omega}})^{-2} z - \sum_i \omega_i^4/\hat{\omega}_i^2\right| \leq \frac{\gamma}{4}\sqrt{\sum_{i\in[d]}(\omega_i^4/\hat{\omega}_i^2)^2}\right].$$

We can then again apply Theorem 10 to obtain $\hat{\pi}(E_\gamma) \geq 1 - 2\exp(-C'\gamma)$ for a suitable constant $C' > 0$. ■

Next we give a bound on $\hat{\pi}(x)/\pi(x)$ for all $x \in E_\gamma$, which we will use later to show that $\hat{\pi}$ can be used as a warm start for π.

Lemma 12 *Let $\pi(x) \propto \exp(-\frac{1}{2}x^\top \operatorname{diag}(\boldsymbol{\omega})^2 x)$, let $\gamma \geq 1$ and consider E_γ as defined in Eq. (13). Let $\delta = \frac{1}{10\sqrt{\gamma\beta}d^{1/4}}$, set $\hat{\omega}_i = \omega_i\sqrt{1-\frac{\delta^2\omega_i^2}{4}}$ for each $i \in [d]$, and let $\hat{\pi}(x) \propto \exp(-\frac{1}{2}x^\top \operatorname{diag}(\hat{\omega})^2 x)$. Then for all $x \in E_\gamma$ we have*

$$0.9 \leq \frac{\hat{\pi}(x)}{\pi(x)} \leq 1.1.$$

Proof For $x \in \mathbb{R}^d$ we have

$$\frac{\hat{\pi}(x)}{\pi(x)} = \left(\prod_i \left(1 - \frac{\delta^2\omega_i^2}{4}\right)\right)^{1/2} \exp\left(\frac{\delta^2}{8}\sum_i x_i^2\omega_i^4\right).$$

We first obtain an upper bound on $\frac{\hat{\pi}(x)}{\pi(x)}$ for $x \in E_\gamma$. Using the inequality $1 - z \leq \exp(-z)$ (which holds for all $z \in \mathbb{R}$), we obtain

$$\frac{\hat{\pi}(x)}{\pi(x)} \leq \exp\left(\frac{\delta^2}{8}\left(\sum_i x_i^2\omega_i^4 - \omega_i^2\right)\right) \leq \exp\left(\frac{1}{8}\delta^2\gamma\sqrt{\sum_{i\in[d]}\omega_i^4}\right) \leq \exp\left(\frac{1}{800}\right) \leq 1.1$$

where in the second inequality we use that $x \in E_\gamma$.

We can similarly bound $\frac{\hat{\pi}(x)}{\pi(x)}$ from below for $x \in E_\gamma$. For this we use the inequality $1 - z \geq \exp(-\eta z)$ which holds for $0 \leq z < 1$ and $\eta \geq \frac{1}{z}\ln(\frac{1}{1-z})$. For $z \leq 1/2$ one has $\frac{1}{z}\ln(\frac{1}{1-z}) \leq 1 + z$ and thus $\eta = 1 + z$ suffices. We apply this with $z = \frac{\delta^2\omega_i^2}{4} \leq \frac{1}{400\gamma\sqrt{d}} < 1/2$. This allows us to lower bound $\frac{\hat{\pi}(x)}{\pi(x)}$ as

$$\begin{aligned}
\frac{\hat{\pi}(x)}{\pi(x)} &\geq \exp\left(-\frac{1}{2}\left(1 + \frac{1}{400\gamma\sqrt{d}}\right)\frac{\delta^2}{4}\sum_i \omega_i^2\right)\exp\left(\frac{\delta^2}{8}\sum_i x_i^2\omega_i^4\right) \\
&\geq \exp\left(-\frac{1}{2}\frac{1}{400\gamma\sqrt{d}}\frac{\delta^2}{4}\sum_i \omega_i^2 - \frac{\delta^2}{8}\left|\sum_i x_i^2\omega_i^4 - \sum_i \omega_i^2\right|\right) \\
&\geq \exp\left(-\frac{1}{3200\cdot 100\gamma^2 d\beta}\sum_i \omega_i^2 - \frac{1}{800}\right) \geq \exp\left(-\frac{1}{400}\right) \geq 0.9
\end{aligned}$$

where in the third inequality we use that $\delta^2 = \frac{1}{100\gamma\beta\sqrt{d}}$ and $x \in E_\gamma$. ■

Finally, we note that the acceptance probability is large on E_γ.

Lemma 13 *Let $\mathcal{A}(x, x')$ be the acceptance probability of the adjusted leapfrog HMC with step size δ. If $x, x' \in E_\gamma$ then $\mathcal{A}(x, x') \geq \exp\left(-\frac{\delta^2\gamma}{4} d^{1/2}\beta\right)$.*

Proof If both $x, x' \in E_\gamma$ then we have that $\sum_{i\in[d]} \omega_i^4 (x_i^2 - x_i'^2) \leq 2\gamma\sqrt{\sum_i \omega_i^4} \leq 2\gamma d^{1/2}\beta$. ■

Lemma 12 and Lemma 13 tell us that the stepsize δ should scale with γ, d and β as

$$\delta = \frac{1}{10\sqrt{\gamma\beta} d^{1/4}}. \tag{14}$$

This choice of δ ensures a high acceptance probability whenever $x, x' \in E_\gamma$ and a pointwise bound on the ratio $\hat{\pi}(x)/\pi(x)$ for $x \in E_\gamma$. In the next section we tune the choice of $\gamma \geq 1$ to apply an argument based on the s-conductance.

4.2 s-conductance and warm start

We will bound the mixing time of the Metropolis-adjusted chain using the so-called *s-conductance*. This is a generalization of the conductance that allows one to ignore small subsets of measure $\pi(S) \leq s$. For context, the conductance of a set S with respect to a chain T roughly measures how quickly the chain T can escape the set S, which intuitively is a bottleneck on how fast a chain can mix. Jerrum and Sinclair (1989) made this intuition precise for discrete Markov chains, by showing that it is (quadratically) related to the spectral gap. In the continuous setting, Lovász and Simonovits (1993) showed that a related notion, s-conductance, implies a similar convergence rate. We state the necessary definitions and results below and we refer the interested reader to, e.g., Vempala (2005) for a survey on the analysis of geometric random walks.

Definition 14 (s-conductance) *Let $0 < s < 1/2$ and define the s-conductance C_s of a Markov chain with transition kernel T and stationary distribution π as*

$$C_s \coloneqq \inf\left\{ C_s(S) \mid S \subseteq \mathbb{R}^d \text{ measurable},\ s < \pi(S) \leq \frac{1}{2} \right\}, \quad \text{with } C_s(S) \coloneqq \frac{\int_S T(x, S^c)\pi(\mathrm{d}x)}{\pi(S) - s}.$$

The s-conductance leads to a mixing time bound through the following theorem from Lovász and Simonovits (1993) (the exact formulation below is from (Wu et al., 2021, Lem. 1)). It uses a *warmness* parameter $D_s^{\mu_0,\pi}$ between the initial distribution μ_0 and target distribution π, which for $0 < s < 1/2$ is defined by

$$D_s^{\mu_0,\pi} \coloneqq \sup\{|\mu_0(A) - \pi(A)| \,:\, A \subseteq \mathbb{R}^d \text{ measurable}, \pi(A) \leq s\}.$$

Lemma 15 (Lovász and Simonovits (1993)) *Consider a reversible, lazy*[12] *Markov chain with transition kernel R, stationary distribution π and initial distribution μ_0. Then for any $K \geq 0$ it holds that*

$$\|R^K_{\mu_0} - \pi\|_{\mathrm{TV}} \leq D^{\mu_0,\pi}_s + \frac{D^{\mu_0,\pi}_s}{s}\left(1 - \frac{C_s^2}{2}\right)^K.$$

Using Lemma 12 we can prove that the stationary distribution $\hat{\pi}$ of the *unadjusted* chain $\hat{Q}$ for sufficiently small step size forms a warm start, if we take $\gamma \in \Theta(\log(1/s))$.

Lemma 16 (unadjusted warm start) *Let $\pi(x) \propto e^{-\frac{1}{2}x^\top \operatorname{diag}(\boldsymbol{\omega})^2 x}$ and let $\hat{\pi}(x) \propto e^{-\frac{1}{2}x^\top \operatorname{diag}(\hat{\boldsymbol{\omega}})^2 x}$ with $\hat{\omega}_i = \omega_i\sqrt{1 - \frac{\delta^2\omega_i^2}{4}}$. For any $0 < s < 1/2$, if $\delta \leq \frac{C}{\sqrt{\beta\log(1/s)}d^{1/4}}$ for a sufficiently small constant $C > 0$, then*

$$D^{\hat{\pi},\pi}_s \leq 3s.$$

Proof Consider the set E_γ defined in (13) for a sufficiently large $\gamma \in O(\log(1/s))$. Then by Lemma 11 and Lemma 12 both $\pi(E_\gamma) \geq 1 - s$ and $\hat{\pi}(E_\gamma) \geq 1 - s$, and $\hat{\pi}(x)/\pi(x) \leq 1.1$ for all $x \in E_\gamma$. Now let $A \subseteq \mathbb{R}^d$ with $\pi(A) \leq s$. Then we have

$$\begin{aligned}
|\hat{\pi}(A) - \pi(A)| &= |\hat{\pi}(A \cap E_\gamma) + \hat{\pi}(A \cap E_\gamma^c) - \pi(A \cap E_\gamma) - \pi(A \cap E_\gamma^c)| \\
&\leq |\hat{\pi}(A \cap E_\gamma) - \pi(A \cap E_\gamma)| + \hat{\pi}(A \cap E_\gamma^c) + \pi(A \cap E_\gamma^c) \\
&\leq \pi(A \cap E_\gamma) + \hat{\pi}(A \cap E_\gamma^c) + \pi(A \cap E_\gamma^c) \\
&\leq \pi(A) + s + s \leq 3s.
\end{aligned}$$

Here in the second inequality we use that $|\hat{\pi}(x) - \pi(x)| \leq \pi(x)$ for all $x \in E_\gamma$. ■

4.3 Bounding the s-conductance of the adjusted HMC chain

To bound the s-conductance of the adjusted chain, we first bound the s-conductance of the *unadjusted* HMC chain $\hat{Q}$, and then relate both conductances. For the unadjusted chain, we can use our bounds on the mixing time of that chain to lower bound its conductance.

Lemma 17 (s-conductance unadjusted HMC) *Let $0 < s < 1/2$ and let $\hat{C}_s$ be the s-conductance of the unadjusted HMC chain $\hat{Q}$ with step size $\delta \leq \frac{C}{\sqrt{\beta\log(1/s)}d^{1/4}}$ for a sufficiently small constant $C > 0$. Then*

$$\hat{C}_s \in \Omega(1/\log(d\kappa\log(1/s))).$$

Proof First consider the s-conductance $\hat{C}_s^{(K)}$ of the *K-step* kernel $\hat{Q}^K$. From Lemma 7 we know that $\|\hat{Q}^K_x - \hat{\pi}\|_{\mathrm{TV}} \leq 1/10$ for $K \geq C\log(d\kappa(\sqrt{\alpha}\|x\|_\infty + 1))$ for an appropriate constant $C > 0$. In particular, if $x \in E_\gamma$ with $\gamma \geq 1$ then $\|x\|_\infty \leq \sqrt{\frac{(\gamma+1)\kappa d}{\alpha}}$ and hence

12. A lazy chain takes a step with probability 1/2, and otherwise does nothing.

$\|\hat{Q}_x^K - \hat{\pi}\|_{\mathrm{TV}} \leq 1/10$ for all $x \in E_\gamma$ and $K \geq C' \log(\gamma d\kappa)$ for an appropriate constant $C' > 0$. By Lemma 11 we can ensure $\hat{\pi}(E_\gamma) \geq 1 - s$ by picking $\gamma \in O(\log(1/s))$ (recall that $\delta = \frac{1}{10\sqrt{\gamma\beta}d^{1/4}}$). This choice of γ ensures there exists a $K \in O(\log(d\kappa \log(1/s)))$ with the above properties. Combining these properties, for any S for which $s < \hat{\pi}(S) \leq 1/2$ we have that

$$\begin{aligned}\hat{C}_s^{(K)}(S) = \frac{\int_S \hat{\pi}(x)\hat{Q}_x^K(S^c)}{\hat{\pi}(S) - s} &\geq \frac{\int_{S\cap E_\gamma} \hat{\pi}(x)\hat{Q}_x^K(S^c)}{\hat{\pi}(S) - s} \\ &\geq \frac{\hat{\pi}(S \cap E_\gamma)(\hat{\pi}(S^c) - 1/10)}{\hat{\pi}(S) - s} \geq \hat{\pi}(S^c) - \frac{1}{10} \geq \frac{2}{5},\end{aligned}$$

and hence $\hat{C}_s^{(K)} \geq 2/5$.

Now we can use the fact that $\hat{C}_s^{(K)} \leq K\hat{C}_s^{(1)} = K\hat{C}_s$ to conclude that $\hat{C}_s \geq 2/(5K)$, which is $\Omega(1/\log(d\kappa \log(1/s)))$ as claimed. To see that $\hat{C}_s^{(K)} \leq K\hat{C}_s^{(1)}$ (which is well-known, see e.g. (Levin et al., 2017, Eq. (7.10))), define $\hat{\pi}_S$ by $\hat{\pi}_S(x) = \hat{\pi}(x)$ for $x \in S$ and $\hat{\pi}_S(x) = 0$ elsewhere. Then note that $\hat{C}_s^{(K)}(S) = \|Q_{\hat{\pi}_S}^K - \hat{\pi}_S\|_{\mathrm{TV}}/(\hat{\pi}(S) - s)$. Using a telescoping sum and a triangle inequality we can bound

$$\begin{aligned}\|Q_{\hat{\pi}_S}^K - \hat{\pi}_S\|_{\mathrm{TV}} &\leq \|Q_{\hat{\pi}_S}^K - Q_{\hat{\pi}_S}^{K-1}\|_{\mathrm{TV}} + \|Q_{\hat{\pi}_S}^{K-1} - Q_{\hat{\pi}_S}^{K-2}\|_{\mathrm{TV}} + \cdots + \|Q_{\hat{\pi}_S} - \hat{\pi}_S\|_{\mathrm{TV}} \\ &\leq K\|Q_{\hat{\pi}_S} - \hat{\pi}_S\|_{\mathrm{TV}},\end{aligned}$$

where the second inequality follows from submultiplicativity of the total variation distance. Dividing both sides by $\hat{\pi}(S) - s$ and taking the infimum over S proves that $\hat{C}_s^{(K)} \leq K\hat{C}_s^{(1)}$. ■

To relate the s-conductance of the adjusted chain to the one of the unadjusted chain, we use the properties of π and $\hat{\pi}$ shown in Section 4.1: there is a set $E \subseteq \mathbb{R}^d$ of large measure on which π and $\hat{\pi}$ pointwise differ by at most a small multiplicative constant. Moreover, if both $x \in E$ and $x' \in E$, then the acceptance probability of the adjusted chain satisfies $\mathcal{A}(x, x') \geq 99/100$.

Lemma 18 (s-conductance adjusted HMC) *Let $0 < s < C/\log(d\kappa)$ for a sufficiently small constant $C > 0$, and let C_s and $\hat{C}_{s/2}$ be the s-conductance and the $s/2$-conductance of the adjusted and unadjusted chains Q and $\hat{Q}$ with step size $\delta \leq \frac{C'}{\sqrt{\beta \log(1/s)}d^{1/4}}$ for a sufficiently small constant $C' > 0$. Then*

$$C_s \geq \hat{C}_{s/2}/2.$$

Proof Our goal is to lower bound $\frac{1}{\pi(S)-s}\int_S \pi(x)Q(x, S^c)\,\mathrm{d}x$ for all sets S such that $s < \pi(S) \leq \frac{1}{2}$. To this end, we will use that by the Lemmas 11, 12 and 13 the set $E \coloneqq E_\gamma \subset \mathbb{R}^d$ (defined in Eq. (13)) for a suitable $\gamma \in \Theta(\log(1/s))$ and $\delta = \frac{1}{10\sqrt{\gamma\beta}d^{1/4}}$ (as in Eq. (14)) satisfies

1. $\pi(E^c) \leq s/10$,
2. $\hat{\pi}(E^c) \leq s^2/10$,

3. $0.9 \leq \frac{\hat{\pi}(x)}{\pi(x)} \leq 1.1$ for all $x \in E$,

4. the acceptance probability $\mathcal{A}(x, x') \geq 99/100$ for all $x, y \in E$.

Note that in Lemma 17 we have shown that $\hat{C}_{s/2} \in \Omega(1/\log(d\kappa \log(1/s)))$. Therefore, for $s < C/\log(d\kappa)$ for a small enough constant $C > 0$, we have $s \leq \hat{C}_{s/2}$ and thus $\hat{\pi}(E^c) \leq s\hat{C}_{s/2}/10$.

We can use this to lower bound the integral

$$\begin{aligned}
\int_S \pi(x)Q(x, S^c)\,\mathrm{d}x &\geq \int_{S\cap E} \pi(x)Q(x, S^c \cap E)\,\mathrm{d}x \\
&= \int_{S\cap E} \pi(x) \int_{S^c\cap E} Q(x, y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_{S\cap E} \pi(x) \int_{S^c\cap E} \hat{Q}(x, y)\mathcal{A}(x, y)\,\mathrm{d}y\,\mathrm{d}x \\
&\geq 0.85 \int_{S\cap E} \hat{\pi}(x) \int_{S^c\cap E} \hat{Q}(x, y)\,\mathrm{d}y\,\mathrm{d}x \\
&= 0.85 \int_{S\cap E} \hat{\pi}(x)\hat{Q}(x, S^c \cap E)\,\mathrm{d}x \\
&= 0.85 \left(\int_{S\cap E} \hat{\pi}(x)\hat{Q}(x, S^c \cup E^c)\,\mathrm{d}x - \int_{S\cap E} \hat{\pi}(x)\hat{Q}(x, E^c)\,\mathrm{d}x \right) \\
&\geq 0.85 \left(\int_{S\cap E} \hat{\pi}(x)\hat{Q}(x, S^c \cup E^c)\,\mathrm{d}x - \hat{\pi}(E^c) \right),
\end{aligned}$$

where the last inequality follows from detailed balance:

$$\int_{S\cap E} \hat{\pi}(x)\hat{Q}(x, E^c)\,\mathrm{d}x = \int_{E^c} \hat{\pi}(x)\hat{Q}(x, S \cap E)\,\mathrm{d}x \leq \hat{\pi}(E^c).$$

We recognize the last integral as the ergodic flow from the set $S' := S \cap E$ to its complement, and so we can lower bound it in terms of the conductance of $\hat{Q}$, provided that S' has an appropriate measure according to $\hat{\pi}$. We bound $\hat{\pi}(S')$ from below

$$\hat{\pi}(S') \geq 0.9\pi(S') = 0.9(\pi(S) - \pi(S \cap E^c)) \geq 0.9s - \pi(E^c) \geq 0.8s,$$

and from above:

$$\hat{\pi}(S') \leq 1.1\pi(S') \leq 1.1\pi(S) \leq 0.55.$$

We proceed in two different ways depending on the measure $\hat{\pi}(S')$.

1. If $0.8s \le \hat{\pi}(S') \le 1/2$, we have the lower bound

$$\begin{aligned} C_s = \frac{\int_S \pi(x) Q(x, S^c)\,\mathrm{d}x}{\pi(S) - s} &\ge 0.85 \frac{\hat{C}_{s/2}(\hat{\pi}(S') - s/2) - \hat{\pi}(E^c)}{\pi(S) - s} \\ &\ge 0.85 \frac{\hat{C}_{s/2}(\hat{\pi}(S') - 0.6s)}{\pi(S) - s} \\ &\ge 0.85 \frac{\hat{C}_{s/2}(0.9\pi(S') - 0.6s)}{\pi(S) - s} \\ &\ge 0.85 \frac{\hat{C}_{s/2}(0.9\pi(S) - \pi(E^c) - 0.6s)}{\pi(S) - s} \\ &\ge 0.85 \frac{\hat{C}_{s/2}(0.9\pi(S) - 0.7s)}{\pi(S) - s} \\ &\ge 0.85 \frac{\hat{C}_{s/2}(0.7\pi(S) - 0.7s)}{\pi(S) - s} \ge \frac{\hat{C}_{s/2}}{2}. \end{aligned}$$

2. If $1/2 \le \hat{\pi}(S') \le 0.55$, we have $s \le \hat{\pi}(S'^c) \le 1/2$. Additionally, we know that $\hat{Q}$ satisfies detailed balance:

$$\int_{S'} \hat{\pi}(x) \hat{Q}(x, S'^c)\,\mathrm{d}x = \int_{S'^c} \hat{\pi}(x) \hat{Q}(x, S')\,\mathrm{d}x.$$

   Therefore, we have the following lower bound

$$\begin{aligned} C_s = \frac{\int_S \pi(x) Q(x, S^c)\,\mathrm{d}x}{\pi(S) - s} &\ge 0.85 \frac{\hat{C}_{s/2}(\hat{\pi}(S'^c) - s/2) - \hat{\pi}(E^c)}{\pi(S^c) - s} \\ &\ge 0.85 \frac{\hat{C}_{s/2}(\hat{\pi}(S'^c) - 0.6s)}{\pi(S^c) - s} \\ &= 0.85 \frac{\hat{C}_{s/2}(1 - \hat{\pi}(S') - 0.6s)}{1 - \pi(S) - s} \\ &\ge 0.85 \frac{\hat{C}_{s/2}(1 - 1.1\pi(S) - 0.6s)}{1 - \pi(S) - s} \\ &\ge 0.85 \frac{\hat{C}_{s/2}(1 - 1.1\pi(S) - 0.6s)}{1 - \pi(S) - 0.6s} \ge \frac{\hat{C}_{s/2}}{2}. \end{aligned}$$

■

4.4 Mixing time of adjusted HMC

We can now plug our bounds on the s-conductance into Lemma 15 to get the following bound on the mixing time of the (lazy) Metropolis-adjusted HMC chain,[13] when starting from a warm start.

13. Making the chain lazy reduces the s-conductance only by a factor 2.

Theorem 19 (Metropolis-adjusted HMC with warm start) *Let $0 < \varepsilon < C/\log(d\kappa)$ for a sufficiently small constant $C > 0$, and let μ_0 be an initial distribution with warmness $D_s^{\mu_0,\pi} \leq \varepsilon/2$ for $s = \varepsilon/6$. There exist constants $C', C'' > 0$ such that for every $x \in \mathbb{R}^d$, if*

$$K \geq C' \log(d\kappa \log(1/\varepsilon)) \log(1/\varepsilon) \quad \text{and} \quad \delta \leq \frac{C''}{\sqrt{\beta \log(1/\varepsilon)} d^{1/4}},$$

then

$$\|Q_{\mu_0}^K - \pi\|_{\mathrm{TV}} \leq \varepsilon$$

where $\pi \propto \exp(-x^\top Bx/2)$ and Q is the kernel of the (lazy) Metropolis-adjusted leapfrog HMC chain with step size δ.

Proof For $s = \varepsilon/6$ and our choice of δ we know from Lemma 18 and Lemma 17 that Q has s-conductance $C_s \in \Omega(1/\log(d\kappa \log(1/s)))$. By invoking Lemma 15 we know that

$$\|Q_{\mu_0}^K - \pi\|_{\mathrm{TV}} \leq D_s + \frac{D_s}{s}\left(1 - \frac{C_s^2}{2}\right)^K \leq \frac{\varepsilon}{2} + 3\left(1 - \frac{C_s^2}{2}\right)^K \leq \varepsilon$$

for $K \in \Omega(\log(1/\varepsilon)/C_s)$ and hence $K \in \Omega(\log(d\kappa \log(1/\varepsilon)) \log(1/\varepsilon))$. ■

Hence, starting from a warm start μ_0 we can sample from a distribution ε-close to π in TV-distance using $\widetilde{O}(\sqrt{\kappa} d^{1/4} \log(1/\varepsilon))$ gradient evaluations. To get around this warm start, recall from Lemma 16 that the stationary distribution of the *unadjusted* chain (with sufficiently small step size δ) provides a warm start for the adjusted chain. This gives the following, main theorem.

Theorem 20 (Metropolis-adjusted HMC) *Let $0 < \varepsilon < C/\log(d\kappa)$ for a sufficiently small constant $C > 0$. There exists constants $C_0', C', C'' > 0$ such that for every $x \in \mathbb{R}^d$, if*

$$K \geq C' \log(d\kappa \log(1/\varepsilon)) \log(1/\varepsilon), \quad K_0 \geq C_0' \log\left(\frac{d\kappa(\sqrt{\alpha}\|x\|_\infty + 1)}{\varepsilon}\right), \quad \delta \leq \frac{C''}{\sqrt{\beta \log(1/s)} d^{1/4}},$$

then

$$\|(Q^K \circ \hat{Q}^{K_0})_x - \pi\|_{\mathrm{TV}} \leq \varepsilon$$

where $\pi \propto \exp(-x^\top Bx/2)$ and Q (resp. $\hat{Q}$) is the kernel of the (lazy) Metropolis-adjusted (resp. unadjusted) leapfrog HMC chain with step size δ. We can thus obtain a sample from a distribution that is ε-close to π in TV-distance using $\widetilde{O}(\sqrt{\kappa} d^{1/4} \log(1/\varepsilon))$ gradient evaluations.

Proof From Lemma 16 we know that there exists a constant $C'' > 0$ such that if $\delta \leq \frac{C''}{\sqrt{\beta \log(1/s)} d^{1/4}}$, then $\hat{\pi}$ is such that $D_s^{\hat{\pi},\pi} \leq \varepsilon/4$ for $s = \varepsilon/12$, i.e., $\hat{\pi}$ is warm for π. Theorem 19 shows that there exists a constant $C' > 0$ such that for all $K \geq C' \log(d\kappa \log(1/\varepsilon)) \log(1/\varepsilon)$ we have $\|Q_{\hat{\pi}}^K - \pi\|_{\mathrm{TV}} \leq \varepsilon/2$. On the other hand, for the unadjusted chain, by Corollary 6, there exists a constant $C_0' > 0$ such that for all $x \in \mathbb{R}^d$ and $K_0 \geq C_0' \log\left(\frac{d\kappa(\sqrt{\alpha}\|x\|_\infty + 1)}{\varepsilon}\right)$ we

have $\|\hat{Q}_x^{K_0} - \hat{\pi}\|_{\mathrm{TV}} \leq \varepsilon/2$. Combining these two estimates we obtain for such K and K_0 that

$$\begin{aligned}\|(Q^K \circ \hat{Q}^{K_0})_x - \pi\|_{\mathrm{TV}} &\leq \|(Q^K \circ \hat{Q}^{K_0})_x - Q_{\hat{\pi}}^K\|_{\mathrm{TV}} + \|Q_{\hat{\pi}}^K - \pi\|_{\mathrm{TV}} \\ &\leq \|\hat{Q}_x^{K_0} - \hat{\pi}\|_{\mathrm{TV}} + \|Q_{\hat{\pi}}^K - \pi\|_{\mathrm{TV}} \leq \varepsilon,\end{aligned}$$

where we used submultiplicativity ($\|Q_\mu^K - Q_\nu^K\|_{\mathrm{TV}} \leq \|\mu - \nu\|_{\mathrm{TV}}$) in the second inequality. ■

5. Conclusions and open questions

To conclude, we studied the Hamiltonian Monte Carlo algorithm for sampling from high-dimensional Gaussian distributions, focusing on the dependency on both condition number κ and dimension d of the Gaussian. We showed that a HMC algorithm with the leapfrog integrator and long, randomized integration times can be used to sample from a distribution ε-close in TV-distance to a Gaussian distribution by making only $\widetilde{O}(\sqrt{\kappa} d^{1/4} \log(1/\varepsilon))$ gradient queries. Such scaling for leapfrog HMC in both the dimension and the condition number matches with well-known scaling limits due to Duane et al. (1987); Neal (2011) and empirical observations (see e.g. the recent (Chada et al., 2023, Figure 4)). Rigorous optimality of the scaling with κ follows from the $\widetilde{\Omega}(\sqrt{\kappa})$ lower bound by Chewi et al. (2023).

The $\sqrt{\kappa}$-dependency also improves over similar, preceding work on leapfrog HMC that achieved at best a linear κ-dependency (Mangoubi and Vishnoi (2018); Chen et al. (2020)). While these works typically consider more general logconcave distributions, we feel that our work enhances the possibility of obtaining a similar $\sqrt{\kappa}$-dependency for such distributions as well. This would disprove the $\Omega(\kappa)$ versus $O(\sqrt{\kappa})$ gap that was suggested by Lee et al. (2020) between logconcave sampling and convex optimization, respectively.

Our analysis through s-conductance naturally leads to error bounds on the TV-distance. A natural open question is whether similar results can be obtained for different distance measures such as for example the KL-divergence or a Wasserstein distance. Understanding the KL-divergence would require pointwise bounds on the proposal distribution (which we do not obtain via our analysis). The Wasserstein distance depends on the geometry of the state space and is typically analysed via coupling methods. For the *unadjusted* HMC algorithm and the Wasserstein-2 distance, such a result has recently been obtained in Jiang (2023), see also Footnote 6.

Finally, another interesting question concerns the necessity of a warm start, which we obtain by first doing a number of *un-adjusted* HMC steps. A recent work Chada et al. (2023) provides a new perspective on this by combining Langevin dynamics with splitting methods, apparently avoiding the need for a warm start.

Acknowledgments

We thank Alain Durmus for useful discussions. We thank the anonymous reviewers for many suggestions that improved the manuscript.

Appendix A. Proof of Lemma 9

Proof [Proof of Lemma 9, part 1] This fact is well known for fixed integration times. Here we prove that it also holds for *randomized* integration times.

We prove first that Q leaves the distribution $\pi(x) \propto \exp(-x^\top Bx/2)$ invariant. To this end, we look at the larger *phase space*. Starting from $x \sim \pi$, the state (x, v) in step 2 is distributed according to the distribution

$$\tilde{\pi}(x, v) \propto \exp(-x^\top Bx/2 - v^\top v/2) = \exp(-\mathcal{H}(x, v)).$$

It remains to prove that steps 2. and 3. leave $\tilde{\pi}$ invariant. Let T denote the kernel of the proposal generated in step 2. (i.e., proposal $(x', -v')$ has density $T((x, v), (x', v'))$). First we note that T is *symmetric*, i.e., $T((x, v), (x', v')) = T((x', v'), (x, v))$. To see this, recall that leapfrog integration is reversible in the sense that $\text{leapfrog}(x, v, t/\delta, \delta) = (x', v')$ implies that $\text{leapfrog}(x', -v', t/\delta, \delta) = (x, -v)$, and hence

$$\begin{aligned} T((x, v), (x', v')) &= \frac{1}{|U(\mathcal{T})|} \sum_{t \in U(\mathcal{T})} \mathbb{1}\left\{\text{leapfrog}(x, v, t) = (x', -v')\right\} \\ &= \frac{1}{|U(\mathcal{T})|} \sum_{t \in U(\mathcal{T})} \mathbb{1}\left\{\text{leapfrog}(x', v', t) = (x, -v)\right\} = T((x', v'), (x, v)). \end{aligned}$$

Then, note that step 3. effectively implements a Metropolis filter w.r.t. distribution $\tilde{\pi}$, which has acceptance probability

$$\mathcal{A}((x, v), (x', v')) = \min\left\{1, \frac{\tilde{\pi}(x', v')}{\tilde{\pi}(x, v)}\right\} = \min\left\{1, \exp\big(-\mathcal{H}(x', -v') + \mathcal{H}(x, v)\big)\right\}.$$

It is then a direct consequence that steps 2. and 3. leave $\tilde{\pi}$ invariant as well.

Next, we show that Q is in fact *reversible* with respect to π, i.e.,

$$\pi(x)Q(x, x') = \pi(x')Q(x', x), \qquad \text{for all } x, x' \in \mathbb{R}^d.$$

To do this, we use the fact that for all $x, v \in \mathbb{R}^d$, the density $\tilde{\pi}(x, v)$ factorizes as $\tilde{\pi}(x, v) = \pi(x)\mu(v)$ with $\mu(v) \sim \exp(-v^\top v/2)$ a standard Gaussian. Using this, we get that

$$\begin{aligned} \pi(x)Q(x, x') &= \pi(x) \iint_{v, v' \in \mathbb{R}^d} T((x, v), (x', v'))\mathcal{A}((x, v), (x', v'))\, \mu(v)\, \mathrm{d}v\, \mathrm{d}v' \\ &= \pi(x) \iint_{v, v' \in \mathbb{R}^d} T((x, v), (x', v')) \min\left\{1, \frac{\pi(x')\mu(v')}{\pi(x)\mu(v)}\right\} \mu(v)\, \mathrm{d}v\, \mathrm{d}v' \\ &= \iint_{v, v' \in \mathbb{R}^d} T((x, v), (x', v')) \min\left\{\pi(x)\mu(v), \pi(x')\mu(v')\right\} \mathrm{d}v\, \mathrm{d}v'. \end{aligned}$$

Since each term in the last expression is symmetric under the exchange of (x, v) with (x', v'), we conclude that it is equal to $\pi(x')Q(x, x')$ for all x, x', and conclude that the chain is reversible. ∎

Proof [Proof of Lemma 9, part 2] First recall that $(x', v') = \text{leapfrog}(x, v, t/\delta, \delta)$. From Section 2.3.1 we know that the leapfrog integrator preserves the modified Hamiltonian and therefore we have

$$\hat{\mathcal{H}}(x, v) = \hat{\mathcal{H}}(x', v') = \hat{\mathcal{H}}(x', -v').$$

Moreover, by Eq. (5) we have

$$\mathcal{H}(x, v) - \hat{\mathcal{H}}(x, v) = \frac{\delta^2}{8} \sum_{i \in [d]} \omega_i^4 x_i^2$$

for all $x, v \in \mathbb{R}^d$. Combining these two identities we find that

$$\begin{aligned}\mathcal{H}(x, v) - \mathcal{H}(x', -v') &= \left(\hat{\mathcal{H}}(x, v) + \frac{\delta^2}{8} \sum_{i \in [d]} \omega_i^4 x_i^2\right) - \left(\hat{\mathcal{H}}(x', -v') + \frac{\delta^2}{8} \sum_{i \in [d]} \omega_i^4 {x'_i}^2\right) \\ &= \frac{\delta^2}{8} \sum_{i \in [d]} \omega_i^4 (x_i^2 - {x'_i}^2),\end{aligned}$$

and hence the acceptance probability takes the form $\mathcal{A}(x, x')$ as claimed.

From this, it easily follows that Q_x takes the form $Q_x(x') = \hat{Q}_x(x')\mathcal{A}(x, x')$ for $x \neq x'$:

$$\begin{aligned}Q_x(x') &= \iint_{v, v' \in \mathbb{R}^d} T((x, v), (x', v'))\mathcal{A}((x, v), (x', v'))\, \mu(v)\, \mathrm{d}v\, \mathrm{d}v' \\ &= \mathcal{A}(x, x') \iint_{v, v' \in \mathbb{R}^d} T((x, v), (x', v'))\, \mu(v)\, \mathrm{d}v\, \mathrm{d}v' = \mathcal{A}(x, x')\hat{Q}_x(x').\end{aligned}$$

■

References

Shipra Agrawal and Navin Goyal. Analysis of Thompson Sampling for the Multi-armed Bandit Problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, pages 39.1–39.26. JMLR Workshop and Conference Proceedings, 2012. URL `https://proceedings.mlr.press/v23/agrawal12.html`.

Shipra Agrawal and Navin Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, pages 127–135. PMLR, 2013. URL `https://proceedings.mlr.press/v28/agrawal13.html`.

Jason M Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2169–2176. IEEE, 2023.

Alexandros Beskos, Natesh Pillai, Gareth Roberts, Jesus-Maria Sanz-Serna, and Andrew Stuart. Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19:1501–1534, 2013. ISSN 1350-7265. doi: 10.3150/12-BEJ414.

Nawaf Bou-Rabee and Jesús María Sanz-Serna. Randomized Hamiltonian Monte Carlo. *The Annals of Applied Probability*, 27(4):2159–2194, 2017. ISSN 1050-5164, 2168-8737. doi: 10.1214/16-AAP1255.

Neil K Chada, Benedict Leimkuhler, Daniel Paulin, and Peter A Whalley. Unbiased kinetic langevin monte carlo with inexact gradients. *arXiv preprint arXiv:2311.05025*, 2023.

Yuansi Chen and Khashayar Gatmiry. When does Metropolized Hamiltonian Monte Carlo provably outperform Metropolis-adjusted Langevin algorithm? *arXiv:2304.04724*, 2023.

Yuansi Chen, Raaz Dwivedi, Martin J. Wainwright, and Bin Yu. Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *Journal of Machine Learning Research*, 21(92):1–72, 2020. ISSN 1533-7928. URL `http://jmlr.org/papers/v21/19-441.html`.

Zongchen Chen and Santosh S. Vempala. Optimal Convergence Rate of Hamiltonian Monte Carlo for Strongly Logconcave Distributions. *Theory of Computing*, 18(9):1–18, 2022. doi: 10.4086/toc.2022.v018a009.

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the Metropolis-Adjusted Langevin Algorithm. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pages 1260–1300. PMLR, 2021. URL `https://proceedings.mlr.press/v134/chewi21a.html`.

Sinho Chewi, Jaume de Dios Pont, Jerry Li, Chen Lu, and Shyam Narayanan. Query lower bounds for log-concave sampling. *arXiv:2304.02599*, 2023.

George Deligiannidis, Daniel Paulin, Alexandre Bouchard-Côté, and Arnaud Doucet. Randomized Hamiltonian Monte Carlo as scaling limit of the bouncy particle sampler and dimension-free convergence rates. *The Annals of Applied Probability*, 31(6):2612–2662, 2021. ISSN 1050-5164, 2168-8737. doi: 10.1214/20-AAP1659.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians with the same mean, 2022.

Simon Duane, A.D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987. ISSN 03702693. doi: 10.1016/0370-2693(87)91197-X.

Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Proceedings of the 31st Conference On Learning Theory*, pages 793–797. PMLR, 2018. URL `https://proceedings.mlr.press/v75/dwivedi18a.html`.

D. L. Hanson and F. T. Wright. A Bound on Tail Probabilities for Quadratic Forms in Independent Random Variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177693335.

Mark Jerrum and Alistair Sinclair. Approximating the permanent. *SIAM Journal on Computing*, 18(6):1149–1178, 1989. doi: 10.1137/0218077. URL `https://doi.org/10.1137/0218077`.

Qijia Jiang. On the dissipation of ideal Hamiltonian Monte Carlo sampler. *Stat*, 12(1):e629, 2023. doi: 10.1002/sta4.629.

A. D. Kennedy and Brian Pendleton. Acceptances and autocorrelations in hybrid Monte Carlo. *Nuclear Physics B - Proceedings Supplements*, 20:118–121, 1991. ISSN 0920-5632. doi: 10.1016/0920-5632(91)90893-J.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Logsmooth Gradient Concentration and Tighter Runtimes for Metropolized Hamiltonian Monte Carlo. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 2565–2597. PMLR, 2020. URL `https://proceedings.mlr.press/v125/lee20b.html`.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Lower Bounds on Metropolized Sampling Methods for Well-Conditioned Distributions. In *Advances in Neural Information Processing Systems*, volume 34, pages 18812–18824. Curran Associates, Inc., 2021. URL `https://papers.nips.cc/paper/2021/hash/9c4e6233c6d5ff637e7984152a3531d5-Abstract.html`.

Benedict Leimkuhler and Sebastian Reich. *Simulating Hamiltonian Dynamics*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2005. ISBN 978-0-521-77290-7. doi: 10.1017/CBO9780511614118.

David Asher Levin, Y. Peres, Elizabeth L. Wilmer, James Propp, and David B. Wilson. *Markov Chains and Mixing Times*. American Mathematical Society, second edition edition, 2017. ISBN 978-1-4704-2962-1.

László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random Structures & Algorithms*, 4(4):359–412, 1993. ISSN 1098-2418. doi: 10.1002/rsa.3240040402.

Jianfeng Lu and Lihan Wang. On explicit L2-convergence rate estimate for piecewise deterministic Markov processes in MCMC algorithms. *The Annals of Applied Probability*, 32(2):1333–1361, 2022. ISSN 1050-5164, 2168-8737. doi: 10.1214/21-AAP1710.

Oren Mangoubi and Nisheeth Vishnoi. Dimensionally Tight Bounds for Second-Order Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://papers.nips.cc/paper/2018/hash/e07bceab69529b0f0b43625953fbf2a0-Abstract.html`.

Eric Mazumdar, Aldo Pacchiano, Yian Ma, Michael Jordan, and Peter Bartlett. On Approximate Thompson Sampling with Langevin Algorithms. In *Proceedings of the 37th*

International Conference on Machine Learning, pages 6797–6807. PMLR, 2020. URL https://proceedings.mlr.press/v119/mazumdar20a.html.

Wenlong Mou, Yi-An Ma, Martin J. Wainwright, Peter L. Bartlett, and Michael I. Jordan. High-Order Langevin Diffusion Yields an Accelerated MCMC Algorithm. *Journal of Machine Learning Research*, 22(42):1–41, 2021. ISSN 1533-7928. URL http://jmlr.org/papers/v22/20-576.html.

Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer New York, 1996. ISBN 978-0-387-94724-2. doi: 10.1007/978-1-4612-0745-0.

Radford M. Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011. doi: 10.1201/b10905.

Giorgio Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180(3): 378–384, 1981.

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

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. ISBN 978-1-108-41519-4. doi: 10.1017/9781108231596.

Nisheeth K. Vishnoi. An Introduction to Hamiltonian Monte Carlo Method for Sampling, 2021. arXiv:2108.12107.

Maxime Vono, Nicolas Dobigeon, and Pierre Chainais. High-Dimensional Gaussian Sampling: A Review and a Unifying Approach Based on a Stochastic Proximal Point Algorithm. *SIAM Review*, 64(1):3–56, 2022. ISSN 0036-1445. doi: 10.1137/20M1371026.

Jun-Kun Wang and Andre Wibisono. Accelerating Hamiltonian Monte Carlo via Chebyshev Integration Time. arXiv:2207.02189, 2022.

Keru Wu, Scott Schmidler, and Yuansi Chen. Minimax Mixing Time of the Metropolis-Adjusted Langevin Algorithm for Log-Concave Sampling, 2021. arXiv:2109.13055.

Pan Xu, Hongkai Zheng, Eric V. Mazumdar, Kamyar Azizzadenesheli, and Animashree Anandkumar. Langevin Monte Carlo for Contextual Bandits. In *Proceedings of the 39th International Conference on Machine Learning*, pages 24830–24850. PMLR, 2022. URL https://proceedings.mlr.press/v162/xu22p.html.