Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions 2: Numerical integrators

Oren Mangoubi
EPFL

Aaron Smith
University of Ottawa

Abstract

We obtain quantitative bounds on the mixing properties of the Hamiltonian Monte Carlo (HMC) algorithm with target distribution in d-dimensional Euclidean space, showing that HMC mixes quickly whenever the target log-distribution is strongly concave and has Lipschitz gradients. We use a coupling argument to show that the popular leapfrog implementation of HMC can sample approximately from the target distribution in a number of gradient evaluations which grows like d^1/2 with the dimension and grows at most polynomially in the strong convexity and Lipschitz-gradient constants. Our results significantly extend and improve on the dimension dependence of previous quantitative bounds on the mixing of HMC and of the unadjusted Langevin algorithm in this setting.

1 Introduction

Markov chain Monte Carlo (MCMC) algorithms are ubiquitous in Bayesian statistics and other areas, and Hamiltonian Monte Carlo (HMC) algorithms are some the most widely-used MCMC algorithms [21, 10, 37] . Despite the popularity of HMC and the widespread belief that HMC outperforms other algorithms in high-dimensional statistical problems (see *e.g.* [2]), its theoretical properties are not as well-understood as some of its older cousins, such as the Random Walk Metropolis algorithm (RWM) or Metropolis-Adjusted Langevin Algorithm (MALA). This lack of theoretical results can make it harder to optimize HMC algorithms, and it means we do not have a good understanding of when HMC is better than other popular algorithms.

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

Several recent papers have begun to bridge this gap, most notably by proving geometric ergodicity of numerical implementations of HMC under general conditions [30], establishing some quantitative bounds on the rate of convergence for an "ideal" implementation of HMC with Gaussian target distributions [47], and relating HMC to Langevin dynamics [42]. In this paper and our companion paper [33], we extend this work by obtaining rapid mixing bounds for both "ideal" and "numerical" implementations of HMC in an important general class of target distributions π on $\mathbb{R}^d$: those which are strongly log-concave and also satisfy the Lipschitz-gradient condition. In this regime, we show upper bounds on the mixing rate of HMC that are better than those of many competitor algorithms, including the unadjusted Langevin algorithm [17, 16]. Our work is particularly close to that of [47], which is to our knowledge the only previous paper giving quantitative non-asymptotic bounds on the mixing of HMC. We improve on their conclusions by greatly improving the dependence of their bounds on the dimension of the target distribution, extending their analysis from Gaussian to general strongly log-concave targets, proving convergence in stronger norms, and providing rates for numerical implementations of HMC algorithms rather than merely "ideal" versions of HMC.

Although our assumptions are quite strong and so our results are far from providing a complete understanding of HMC, the strongly log-concave distributions are an important special case. Recall that a distribution π is strongly log-concave if the Hessian matrix of $-\log(\pi)$ has eigenvalues bounded below by a positive number m_2, and has log-density with Lipschitz gradient if the eigenvalues are bounded above by a positive number M_2. Many important posterior distributions in statistics, including the "ridge regression" posterior associated with Gaussian priors for logistic regression, are strongly log-concave (see Section 5 for explicit bounds). In addition to this, there is some historical interest in comparing MCMC algorithms based on their performance for log-concave targets. Most MCMC algorithms are expected to perform well in this situation, and so the performance of

many Monte Carlo algorithms has been studied extensively in the strongly log-concave setting (see *e.g.* [48, 11, 16, 18], and many others). Thus, extending the analysis of the strongly log-concave setting to HMC has the added advantage of giving a sensible comparison of the performance of HMC to its competitors, such as the Langevin algorithm and the ball walk.

Following the work of [47], our companion paper [33] studied an "idealized" HMC algorithm - one that has no numerical error. We showed that this ideal chain mixes in $\tilde{\mathcal{O}}((\frac{M_2}{m_2})^2)$ steps on strongly log-concave π (see Theorem 1 of [33]). Using this result, the present paper bounds the computational costs of various numerical implementations of HMC. This provides theoretical evidence that HMC can be much faster than competing algorithms in realistic situations (see Theorem 1). To our knowledge, these are the first quantitative mixing bounds for the popular leapfrog implementation of HMC.

More specifically, our main result shows that a simple numerical implementation of HMC can approximately sample from the stationary distribution with a number of gradient evaluations that grows at rate $\mathcal{O}_d(\sqrt{d})$ in the dimension (see Theorem 1). For comparison, the best available mixing time bound for the unadjusted Langevin algorithm on strongly log-concave π grows like $\mathcal{O}_d(d)$, a much larger dependence on dimension than our bounds for unadjusted leapfrog HMC [17, 16]. Our bounds also compare favorably to the ball walk, whose best available mixing time bound is roughly $\mathcal{O}(d^2 \frac{M_2}{m_2})$ [48] (note however that the particular assumptions made in different papers are slightly different).

Our main techniques in this paper are explicit comparisons of ODEs and probabilistic coupling bounds. Roughly speaking, we show that the popular leapfrog implementation of HMC is not too far from the "ideal" HMC algorithm near the mode of the target distribution, and that even poor numerical implementations exhibit drift towards the mode. These results allow us to show that numerical implementations are not too much slower than the ideal HMC algorithm analyzed in the companion paper [33]. There is a long history of obtaining contraction results to control the convergence of Markov chains, diffusions, and solutions to (S)PDEs. We mention especially [3] (which studies a process closely related to Langevin and HMC), [19] (which obtains coupling-based bounds even in the non-concave case), and [7] (which was an early paper on contraction for Markov chains).

1.1 Updates on Recent Work

The papers [36, 19], which appeared after the first version of this note, also use coupling techniques to bound the running time of HMC when targeting high-dimensional targets, obtaining more refined results in various situations. They are almost certainly of interest to anyone reading the present paper.

1.2 Guide to the Paper

The rest of the paper is organized as follows:

- In Section 2, we set notation.
- In Section 3, we state our main result.
- In Section 4, we give a larger collection of references to related parts of the research literature. While the introduction focuses on related *results* this section includes references to many papers that use or originate the *techiques* used in this paper.
- In Section 5, we show that a popular statistical model satisfies the assumptions of Theorem 1.
- In Section 6 we discuss open problems related to the HMC algorithm, as well as how to precondition the target distribution to improve the running time of HMC.

The supplementary material contains most of the proofs, and is organized as follows:

- In Section 7 we recall several results about the idealized HMC dynamics from the companion paper [33].
- In Section 8 we bound the error of the leapfrog integrator that is used in Algorithm 2 to approximate the continuous Hamiltonian dynamics.
- In Section 9 we provide many generic MCMC bounds, which we then use to compare the "ideal" chain studied in [33] to the unadjusted leapfrog HMC chain studied in the present paper. This allows us to bound the running time of the unadjusted leapfrog HMC chain and prove Theorem 1.

2 Assumptions and Algorithms

2.1 Preliminary Notation

For any function $f : \mathbb{R}^a \to \mathbb{R}$, we use the shorthand $f' := \nabla f$, and for $v \in \mathbb{R}^a$ denote by $D_v f := \langle v, \nabla f \rangle$

the directional derivative in the direction v. For a vector-valued function $g = (g_1, \dots, g_b)^\top$ and $v \in \mathbb{R}^a$, we define the coordinate-wise directional derivative $D_v g := (D_v g_1, \dots, D_v g_b)$.

Throughout the paper, our goal is to sample from a stationary distribution $\pi(q)$ on $\mathbb{R}^d$, which we will write as $\pi(q) \propto e^{-U(q)}$ for some potential function $U : \mathbb{R}^d \mapsto \mathbb{R}^+$. We always assume that U is *strongly convex:*

Assumptions 2.1 (Strong Convexity)**.** *We assume that $U : \mathbb{R}^d \mapsto \mathbb{R}$ is differentiable and that there exists $m_2 > 0$ such that*

$$\langle U'(x) - U'(y), x - y \rangle \geq 2\, m_2 \|x - y\|^2 \qquad \forall x, y \in \mathbb{R}^d.$$

Recall that any strongly convex function has a unique minimizer. Throughout this paper, we assume without loss of generality that this minimizer is 0 in order to simplify notation. We will also assume that U is second-order differentiable, so that strong convexity implies that the eigenvalues of the Hessian of U are lower bounded by m_2. We prove a rapid mixing bound under the following additional assumption on the gradient:

Assumptions 2.2 (Lipschitz Gradient)**.** *We assume that there exists a constant $0 < M_2 < \infty$ so that $\|D_v U'(q)\| \leq M_2$ for all $q \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$.*

Throughout the paper, we make a few small abuses of notation. For any function $f : X \mapsto Y$ between two sets, and any $S \subset X$, we define

$$f(S) = \{f(x) : x \in S\}.$$

In addition, we will generally write x for the single-element set $\{x\}$ when this does not result in any ambiguity.

2.1.1 Distributions and Mixing

We denote the distribution of a random variable X by $\mathcal{L}(X)$ and write $X \sim \nu$ as a shorthand for $\mathcal{L}(X) = \nu$.

For two probability measures ν_1, ν_2 on $\mathbb{R}^d$, define the *Wasserstein-k distance*

$$W_k(\nu_1, \nu_2)^k = \inf_{(X,Y) \in \mathcal{C}(\nu_1, \nu_2)} \mathbb{E}[\|X - Y\|^k],$$

where $\mathcal{C}(\nu_1, \nu_2)$ is the set of all random variables on $\mathbb{R}^d \times \mathbb{R}^d$ with marginal distributions ν_1 and ν_2.

2.1.2 Big-O Notation

For two nonnegative functions $f, g : \mathbb{R}^n \mapsto [0, \infty)$, we write $f = O(g)$ as shorthand for the statement: there exist constants $0 < C_1, C_2 < \infty$ so that for all $x_1, \dots, x_n$ with $\|(x_1, \dots, x_n)\| > C_1$, we have $f(x_1, \dots, x_n) \leq C_2\, g(x_1, \dots, x_n)$. Similarly, we write $f = \Omega(g)$ as shorthand for the statement: there exist constants $0 < C_1, C_2 < \infty$ so that for all $x_1, \dots, x_n$ with $\|(x_1, \dots, x_n)\| > C_1$, we have $g(x_1, \dots, x_n) \leq C_2\, f(x_1, \dots, x_n)$. We write $f = \Theta(g)$ if both $f = O(g)$ and $g = O(f)$. Relatedly, we write $f = o(g)$ as shorthand for the statement: $\lim_{x_1, \dots, x_n \to \infty} \frac{f(x_1, \dots, x_n)}{g(x_1, \dots, x_n)} = 0$. Finally, we use "$f$ grows at most polynomially in x" as shorthand for "there exist $0 < C_1, C_2, C_3 < \infty$ such that $f(x) \leq C_1 x^{C_2}$ for all $x > C_3$."

We give two small extensions of this notation; all of these modifications apply in the obvious way to $\Omega(\cdot)$, $\Theta(\cdot)$ and $o(\cdot)$. First, when we wish to view a function as depending on only a subset of its arguments, we indicate the arguments of interest using a subscript. For example, if $f(x, y) = \frac{|x|}{1+y^2}$, we may write $f(x, y) = O(|x|)$ but may also write $f(x, y) = O_x(|x|)$ or $f(x, y) = O_y(1)$. Second, we use a "tilde" superscript $\sim$ to indicate that the relationship holds "up to logarithmic factors." For example, we write $f = \tilde{O}(g)$ if there exist constants $0 < C_1, C_2 < \infty$ so that, for all $x_1, \dots, x_n$, we have $f(x_1, \dots, x_n) \leq C_1\, g(x_1, \dots, x_n)\, \log(g(x_1, \dots, x_n))^{C_2}$. Finally, we note that we can always view a function as taking *additional* arguments, and do so without comment when needed. For example, we may write $(x + y)^2 = O(x^2 + y^2 + z^2)$, even though the variable z^2 does not appear explicitly in the expression on the left-hand side.

2.1.3 Ideal HMC Dynamics

A Hamiltonian of a simple system is written as

$$H(q, p) = U(q) + \frac{1}{2}\|p\|^2, \tag{2.1}$$

where q represents 'position', p represents 'momentum,' U represents 'potential energy,' and $\frac{1}{2}\|p\|^2$ represents 'kinetic energy.'

For fixed $\mathbf{q} \in \mathbb{R}^d$, $\mathbf{p} \in \mathbb{R}^d$, we denote by $\{q_t(\mathbf{q}, \mathbf{p})\}_{t \geq 0}$, $\{p_t(\mathbf{q}, \mathbf{p})\}_{t \geq 0}$ the solutions to Hamilton's equations:

$$\begin{aligned} \frac{\mathrm{d}q_t(\mathbf{q}, \mathbf{p})}{\mathrm{d}t} &= p_t(\mathbf{q}, \mathbf{p}), \\ \frac{\mathrm{d}p_t(\mathbf{q}, \mathbf{p})}{\mathrm{d}t} &= -U'(q_t(\mathbf{q}, \mathbf{p})), \end{aligned} \tag{2.2}$$

with initial conditions

$$q_0(\mathbf{q}, \mathbf{p}) = \mathbf{q}, \qquad p_0(\mathbf{q}, \mathbf{p}) = \mathbf{p}.$$

When the initial conditions $(\mathbf{q}, \mathbf{p})$ are clear from the context, we write q_t, p_t in place of $q_t(\mathbf{q}, \mathbf{p})$ and $p_t(\mathbf{q}, \mathbf{p})$, respectively. The dependence of these solutions on the Hamiltonian H is always suppressed in our notation, as it will always be clear from the context.

For a fixed integration time $T \in \mathbb{R}^+$ and starting point $\mathbf{q} \in \mathbb{R}^d$, we define the solution map $\mathcal{Q}_T^{\mathbf{q}} : \mathbb{R}^d \mapsto \mathbb{R}^d$ by

$$\mathcal{Q}_T^{\mathbf{q}}(\mathbf{p}) := q_T(\mathbf{q}, \mathbf{p}). \tag{2.3}$$

For $V > 0$, denote by $\Phi_V(\cdot)$ the normal distribution on $\mathbb{R}^d$ with mean 0 and variance V times the identity matrix. Algorithm 1 defines the simplest "ideal" HMC Markov chain (see Figure 1):

Algorithm 1 Idealized HMC

parameters: Potential U, trajectory time $T > 0$.
input: Initial point $X_0 \in \mathbb{R}^d$.
output: Markov chain $X_0, X_1, \ldots$.

1: **for** $i = 0, 1, \ldots$ **do**
2: Sample $\mathbf{p}_i \sim N(0, I_d)$.
3: Set $X_{i+1} = \mathcal{Q}_T^{X_i}(\mathbf{p}_i)$.
4: **end for**

In the context of HMC, we refer to q_t as the *position* variable and p_t as the *momentum* variable. In this context, we call $\mathbb{R}^d$ the *state space* of the algorithm and $\mathbb{R}^{2d}$ the *phase space* of the HMC algorithm.

Note that the sequence $\{X_i\}_{i\geq 0}$ is a deterministic function of its initial value X_0 and the i.i.d. sequence $\{\mathbf{p}_i\}_{i\geq 0}$ of momentum updates sampled during this algorithm. In the Markov chain literature, this fact is summarized by saying that this algorithm defines a *random mapping representation* of $\{X_i\}_{i\geq 0}$ with *update sequence* $\{\mathbf{p}_i\}_{i\geq 0}$ (see Chapter 1.2 of [29]). In particular, the fact that this algorithm gives a random mapping representation means that it is possible to define a coupling of two Markov chains evolving according to this algorithm by defining a coupling of the momentum updates. All of the HMC-based algorithms defined in this paper will also have this property, and we will use it throughout the paper to construct couplings of Markov chains. We will generally couple these Markov chains by setting their initial momenta to be equal at each step. Finally, note that this algorithm also naturally defines the nonreversible Markov chain $\{(X_i, \mathbf{p}_i)\}_{i\geq 0}$, which we call the *phase-space* chain on $\mathbb{R}^{2d}$.

2.2 Approximate HMC Dynamics and unadjusted leapfrog HMC Algorithm

It is difficult to solve Hamilton's equations (2.2) for most Hamiltonians of interest. In practice, one uses numerical integrators such as the Euler or leapfrog integrator in order to approximate solutions, with the leapfrog integrator being by far the most widely-used in practice (see [39]). In this paper we study the

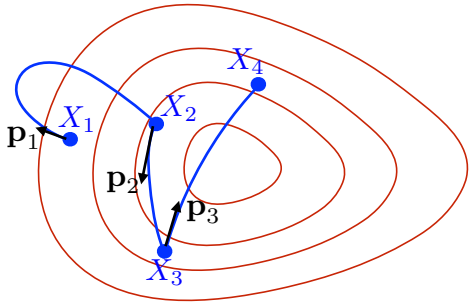

Figure 1: The Hamiltonian Monte Carlo Markov chain $X_1, X_2, \ldots$ with momentum $\mathbf{p}_1, \mathbf{p}_2, \ldots$.

(unadjusted) leapfrog method. In this case, one obtains the following *unadjusted* HMC (UHMC) Markov chain; the main result of this paper is a bound on the running time of this algorithm:

Algorithm 2 Unadjusted leapfrog HMC

parameters: Potential U, trajectory time $T > 0$, and integrator step size $\theta > 0$.
input: Initial point $X_0' \in \mathbb{R}^d$,
output: Markov chain $X_0', X_1', \ldots$.

1: **for** $i = 0, 1, \ldots$ **do**
2: Sample $\mathbf{p}_i \sim N(0, I_d)$
3: Set $\mathrm{q}_0 = X_i'$ and $\mathrm{p}_0 = \mathbf{p}_i$
4: **for** $j = 0$ to $\lfloor \frac{T}{\theta} \rfloor - 1$ **do**
5: Set

$$\mathrm{q}_{j+1} = \mathrm{q}_j + \theta \mathrm{p}_j - \frac{1}{2}\theta^2 U'(\mathrm{q}_j)$$
$$\mathrm{p}_{j+1} = \mathrm{p}_j - \frac{1}{2}\theta U'(\mathrm{q}_j) - \frac{1}{2}\theta U'(\mathrm{q}_{j+1}).$$

6: **end for**
7: Set $X_{i+1}' = \mathrm{q}_{\lfloor \frac{T}{\theta} \rfloor}$
8: **end for**

Algorithm 2 also naturally defines the phase-space unadjusted HMC Markov chain $\{(X_i', \mathbf{p}_i)\}_{i\geq 0}$. Note that this Markov chain will generally *not* have the desired stationary measure π.

We set some notation for the leapfrog integrator. For $\theta > 0$, we use the following shorthand for a single step of the leapfrog integrator:

$$q_\theta^\star(\mathbf{q}, \mathbf{p}) := \mathbf{q} + \theta\mathbf{p} - \frac{1}{2}\theta^2 U'(\mathbf{q}), \tag{2.4}$$
$$p_\theta^\star(\mathbf{q}, \mathbf{p}) := \mathbf{p} - \frac{1}{2}\theta U'(\mathbf{q}) - \frac{1}{2}\theta U'\left(q_\theta^\star(\mathbf{q}, \mathbf{p})\right), \forall \mathbf{q}, \mathbf{p} \in \mathbb{R}^d.$$

3 Main Result

Using the bounds on the mixing of ideal HMC dynamics in our companion paper [33], we can obtain similar results for the unadjusted numerical implementation of HMC with the leapfrog integrator. Note that we give a range on the value of the parameter θ for which the conclusion holds.

Theorem 1 (Mixing of first-order Unadjusted HMC)**.** *Fix $0 < \epsilon < e^{-1}$, let U satisfy Assumptions 2.1 and 2.2, and let $T = \frac{1}{2\sqrt{2}} \frac{\sqrt{m_2}}{M_2}$. For all $\theta > 0$, let Q_θ be the transition kernel defined in Algorithm 2 with these parameters.*

Then there exists some $\theta_0 = \theta_0(m_2, M_2, d, \epsilon)$ satisfying

$$\theta_0 = \tilde{\Omega}\left(d^{-\frac{1}{2}} \times \epsilon \times (\frac{M_2}{m_2})^{-4.5}\right)$$

and some universal constant $0 < c < \infty$ so that, for all $0 < \theta \leq \theta_0$, the following holds:

For all $\mathcal{I} \geq c \frac{M_2^2}{m_2^2} \log(\frac{M_2}{m_2 \epsilon})$ and all x satisfying $\|x\| \leq \frac{\sqrt{d}}{\sqrt{m_2}}$,

$$W_1(Q_\theta^{\mathcal{I}}(x, \cdot), \pi) \leq \epsilon. \tag{3.1}$$

Finally, if we choose $\theta = \tilde{\Omega}(\theta_0)$, then

$$\tilde{O}\left(d^{\frac{1}{2}} \times \epsilon^{-1} \times (\frac{M_2}{m_2})^{6.5}\right)$$

gradient evaluations are needed to compute the first $s = c \frac{M_2^2}{m_2^2} \log(\frac{M_2}{m_2 \epsilon})$ steps of the Markov chain.

Remark 3.1 (Starting Point)**.** Under the assumption that U is strongly log-concave, it is straightforward to find a starting point $x \in \mathbb{R}^d$ within distance $\frac{\sqrt{d}}{\sqrt{m_2}}$ of the unique minimum of U using algorithms from convex optimization.

Remark 3.2 (Trajectory Time)**.** Theorem 1 is stated for the largest-possible trajectory time $T = \frac{1}{2\sqrt{2}} \frac{\sqrt{m_2}}{M_2}$. Note, however, that the constants m_2 and M_2 in the statement of the theorem are merely *bounds* on the Hessian of U; they do not need to be the largest and smallest singular values. In particular, Theorem 1 as stated gives bounds on all trajectory times $0 \leq T \leq \frac{1}{2\sqrt{2}} \frac{\sqrt{m_2}}{M_2}$.

Remark 3.3 (Curvature Ratio)**.** The ratio $\frac{M_2}{m_2}$ that appears prominently in the conclusions of Theorem 1 can be made much smaller for realistic examples by the use of appropriate preconditioning steps. See Section 6.1 for details.

Remark 3.4 (Ergodicity, Volume-Preservation and Optimality)**.** Our proofs simply view Algorithm 2 as a first-order approximation of the "ideal" Algorithm 1, and do not greatly rely on the details of this approximation. As a consequence, we don't rely on, or capture, several features of HMC that are widely-believed to be important. Most blatantly, we do not prove that the chain Q_θ appearing in Theorem 1 is ergodic, and we don't need this for the result to hold[1]. Similarly, one oft-mentioned motivation for the leapfrog algorithm is that it preserves both volume and energy much more accurately than "generic" integrators, but we don't use this property in our proof.

This "blunt" viewpoint is sufficient to obtain the $\sqrt{d}$ dependence on dimension that we wished to find in this paper, but we suspect that a more detailed analysis of the leapfrog integrator could be useful for obtaining very precise bounds.

4 Further Related Literature

There is a large literature on obtaining quantitative bounds on the convergence rates of Markov chains (see [27] for an introduction to the statistical literature on the topic, and [12] for connections in other fields, including computer science). In general, it is difficult to obtain good quantitative bounds for large classes of chains. As such, the literature focuses on either finding very tight bounds for specific chains (see *e.g.* [13]) or on quantitative bounds on the running time of the algorithm as a function of the problem complexity (see *e.g.* [4] or essentially any paper in the large computer science literature on the subject). Our work falls in the latter category.

Although there are many papers that obtain quantitative bounds on the convergence of Markov chains, there are few quantitative bounds on the convergence of HMC. To our knowledge, [47] is the only previous work that focuses on obtaining quantitative bounds on the convergence rates of HMC, and it served as an inspiration for this work. A number of other papers, most prominently [2], had also worked on the problem of calculating the computational complexity of HMC algorithms by computing the rate at which certain proxies for the mixing or relaxation time of HMC increase with the dimension of the target distribution under reasonable conditions (see [44] for a general discussion relating results similar to [2] to the usual notions of complexity). Several other papers give calculations that imply or suggest quantitative bounds, though we are not aware of any that are close to tight (see *e.g.* the discussion in Section 7.5 of [30]).

[1] We do use the fact that the ideal chain is ergodic; this is an immediate consequence of the main contraction inequality in the companion paper [33].

In our opinion, the earlier work that most closely resembles our main result is [16], which studies the non-asymptotic mixing properties of the Langevin algorithm on strongly log-concave distributions. Their main results hold under essentially the same conditions as our Theorem 1, except that they require the additional assumption of a Lipschitz Hessian. See also [11], which studies very similar conditions to [16]. More recent independent work [9] also gives quantitative bounds for an "underdamped" version of the Langevin algorithm that matches our dimension dependence $d^{\frac{1}{2}}$; the authors explain that this underdamped version of the Langevin algorithm is closely related to HMC. The underdamped Langevin diffusion, the continuous-time process on which the underdamped Langevin algorithm is based, has been previously studied in [19, 6, 49, 25, 15, 50, 8, 14, 38].

In recent independent work [28], the authors have obtained quantitative bounds for a different version of HMC, called Riemannian HMC. Their bounds apply to a class of target distributions that include distributions that are not log-concave, although like [47] their bounds only apply to trajectories with very short step sizes and consequently do not imply the results in this paper. Since writing our first version of this paper, the paper [5] has used a refined coupling approach to obtain useful quantitative bounds on the convergence of the usual leapfrog HMC algorithm. [36] use many of the tools developed in our paper to obtain further improvements on the running time bounds for leapfrog HMC in special cases of strongly logconcave distributions with bounded higher-order derivatives. Some examples of these distributions include distributions used in logistic regression as well as other non-separable emperical distributions (see also the ArXiv version of our paper [34] for related improvements in the special case where these distributions are separable). Finally, our recent work [31, 32] quantitatively compares the convergence rates of HMC and RWM in a different, "highly-multimodal," regime.

We also mention some papers that are on different topics, but have technical similarities. First, we remark that our arguments are based on viewing the leapfrog algorithm as a small perturbation of the "ideal" HMC algorithm. As such, they are closely related to bounds in the "approximate MCMC" literature, including *e.g.* [20, 46, 1, 26, 43, 40].

We also note that our arguments are based on computing the contraction rate of explicit couplings. Similar couplings have been used to study processes that are closely related to HMC (see *e.g.* [3, 19] and of course [5] for contractions of closely-related processes) and more generally for analyzing processes related to MCMC (see *e.g.* the analysis in [22], and *e.g.* [7, 41] for more generic discussion of contraction in the MCMC literature).

5 Application to Bayesian logistic regression

As an application, we consider Bayesian logistic regression with Gaussian priors, also called "ridge" regression. This example was previously considered in [16] as an application of the unadjusted Langevin algorithm, and we can bound the running time of HMC using the same strong convexity and Lipschitz gradient bounds that were used in [16] to bound the running time of ULA.

Recall that the "ridge" regression posterior is of the form

$$U(\theta) = \frac{1}{2}\theta^\top \Sigma^{-1}\theta - \sum_{i=1}^r \mathsf{Y}_i \log(F(\theta^\top \mathsf{X}_i)) + (1-\mathsf{Y}_i)\log(F(-\theta^\top \mathsf{X}_i)), \tag{5.1}$$

where the data vectors $\mathsf{X}_1, \dots \mathsf{X}_r \in \mathbb{R}^d$ are thought of as independent variables, the binary data $\mathsf{Y}_1, \dots, \mathsf{Y}_r \in \{0,1\}$ are the dependent variables, and $F(s) := (e^{-s}+1)^{-1}$ is the logistic function. The positive definite matrix Σ is the covariance matrix of the Gaussian prior.

The Hessian H_x of U is

$$H_x = \Sigma^{-1} + \sum_{k=1}^r F'(x^\top \mathsf{X}_k)\mathsf{X}_k\mathsf{X}_k^\top.$$

Therefore U satisfies Assumption 2.1 with strong convexity constant $m_2 = \lambda_{\min}(\Sigma^{-1})$, since this choice of m_2 gives a lower bound on the eigenvalues of H_x. Moreover, Assumption 2.2 is satisfied with Lipschitz gradient constant $M_2 = \lambda_{\max}(\Sigma^{-1} + \sum_{k=1}^r \mathsf{X}_k\mathsf{X}_k^\top)$, since this choice of M_2 gives an upper bound on the eigenvalues of H_x. Applying Theorem 1 with these bounds on m_2 and M_2 then gives running time bounds for generating samples from this class of targets using HMC.

6 Discussion

In this paper, we provide useful bounds on the convergence rate of HMC under rather strong assumptions of strong log-concavity. These bounds improve on several earlier results, and in particular give mixing bounds with near-optimal dependence on dimension for certain implementable variants of HMC, but we leave many important questions open. In this section, we mention some that seem most interesting.

6.1 Preconditioning and Optimization

Our main results are stated in terms of the ratio $\frac{M_2}{m_2}$ of upper and lower bounds m_2, M_2 on the Hessian of the potential U. This ratio can be quite large for many target distributions, such as the posterior distribution of a regression problem in which different coefficients have very different sizes. In this section, we show that simple preprocessing steps can make this ratio much smaller, thus making our bounds much better in practice than they might first appear. We note that this preprocessing is common in other "geometric" Markov chain applications (see *e.g.* the survey [48]). The basic idea is to find a linear transformation of the potential for which this ratio is small on the bulk of the target distribution. Fix a probability distribution π given by $\pi(x) = \frac{1}{\int_{\mathbb{R}^d} e^{-U(x)} \mathrm{d}x} e^{-U(x)}$ for some potential function U. We consider the assumption:

Definition 6.1 (Rounding Matrix)**.** Fix a constant $\epsilon > 0$. Define the *bulk level set* $\mathfrak{S}_\epsilon$ of π by the pair of equations

$$L_\epsilon = \inf\{C : \pi(\{x : \pi(x) \geq C\}) \leq 1 - \epsilon\},$$
$$\mathfrak{S}_\epsilon = \{x : \pi(x) \geq L_\epsilon\}.$$

Call a matrix A a *rounding matrix* for π with constants ϵ, m_2, M_2 if the eigenvalues of the Hessian of $\hat{U}(x) := U(A^{-1}x)$ are bounded below and above by m_2 and M_2, respectively, at every point $x \in A\mathfrak{S}_\epsilon$.

For every $x \in \mathbb{R}^d$, define H_x to be the Hessian of U evaluated at x. The following theorem says that, if there *exists* a linear transformation with associated ratio $\frac{M_2}{m_2}$, then it is *easy to find* a linear transformation with associated ratio $\frac{M_2^2}{m_2^2}$:

Theorem 2. *Fix a probability distribution $\pi(x) = \frac{1}{\int_{\mathbb{R}^d} e^{-U(x)} \mathrm{d}x} e^{-U(x)}$. Suppose that there exists a rounding matrix A for π with constants $\epsilon, m_2, M_2 > 0$. Define $\tilde{U}(z) := U(\sqrt{H_x}^{-1} z)$ for $z \in \mathbb{R}^d$. For every $\zeta \in \mathbb{R}^d$, let $\tilde{H}_\zeta$ be the Hessian of $\tilde{U}$ evaluated at the point $z = \sqrt{H_x}\zeta$. Then for every $y \in \mathfrak{S}_\epsilon$, $\tilde{H}_y$ has all its eigenvalues bounded below and above by $\frac{m_2}{M_2}$ and $\frac{M_2}{m_2}$, respectively.*

Proof. As in Definition 6.1, let $\hat{U}(z) := U(A^{-1}z)$ for every $z \in \mathbb{R}^d$

For every $\zeta \in \mathbb{R}^d$, let $\hat{H}_\zeta$ be the Hessian of $\hat{U}(z)$ evaluated at the point $z = A\zeta$. By Definition 6.1,

$$m_2 u^\top u \leq u^\top \hat{H}_\zeta u \leq M_2 u^\top u, \tag{6.1}$$

for all $\zeta \in \mathfrak{S}_\epsilon$ and all $u \in \mathbb{R}^d$. Fixing $x \in \mathfrak{S}_\epsilon$ and applying Inequality (6.1) twice gives

$$\frac{m_2}{M_2} u^\top \hat{H}_x u \leq m_2 u^\top u \tag{6.2}$$

$$\begin{aligned} &\leq u^\top \hat{H}_y u \\ &\leq M_2 u^\top u \\ &\leq \frac{M_2}{m_2} u^\top \hat{H}_x u. \end{aligned}$$

Since $\sqrt{H}_z = A\sqrt{\hat{H}_z}$ for every $z \in \mathbb{R}^d$, applying Equation (6.2) with $v = Au$ gives

$$\frac{m_2}{M_2} v^\top H_x v \leq v^\top H_y v \leq \frac{M_2}{m_2} v^\top H_x v \tag{6.3}$$

for any $u \in \mathbb{R}^d$ (and hence for any $v \in \mathbb{R}^d$, since A is invertible).

Now $\sqrt{\tilde{H}_z} = \sqrt{H_x}^{-1}\sqrt{H_z}$ for every $z \in \mathbb{R}^d$. Therefore, Equation (6.3) implies that:

$$\frac{m_2}{M_2} v^\top v \leq v^\top \tilde{H}_y v \leq \frac{M_2}{m_2} v^\top v \tag{6.4}$$

for every $y \in \mathfrak{S}_\epsilon$ and every $v \in \mathbb{R}^d$. Therefore, by the minimax theorem for eigenvalues, $\tilde{H}_y$ has all its eigenvalues bounded between $\frac{m_2}{M_2}$ and $\frac{M_2}{m_2}$ for all $y \in \mathfrak{S}_\epsilon$. □

6.2 Riemannian HMC

This paper analyzes one of the simplest possible HMC algorithms. However, many other variants exist. Riemannian HMC, introduced in [21], is one of the most popular. This approach seems to obviate the need for the preconditioning step discussed in Section 6.1. Although the authors of [28] analyze the Riemannian HMC algorithm for very short trajectory times on the order of the step-size of the Langevin algorithm, it would be interesting to analyze Riemannian HMC for longer trajectory times. It is widely believed that longer trajectory times result in more efficient algorithms, but we are not aware of any work (besides the bounds of the present paper, for "standard" HMC) that provides quantitative bounds in support of this belief.

6.3 De-biasing with Coupling for Parallel Processing

In [24], a coupling similar to the one described in Section 2.1.3 is used to provide unbiased samples of the target density from HMC Markov chains that are numerically implemented in parallel. As the authors of [24] ask in their discussion section, it would be interesting to see if the quantitative bounds obtained in our paper with this coupling could be used to provide stronger convergence guarantees for their algorithm.

References

[1] Pierre Alquier, Nial Friel, Richard Everitt, and Aidan Boland. Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, 26(1-2):29–47, 2016.

[2] Alexandros Beskos, Natesh Pillai, Gareth Roberts, Jesus-Maria Sanz-Serna, and Andrew Stuart. Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 2013.

[3] François Bolley, Arnaud Guillin, and Florent Malrieu. Trend to equilibrium and particle approximation for a weakly selfconsistent Vlasov-Fokker-Planck equation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44(5):867–884, 2010.

[4] Christian Borgs, Jennifer T Chayes, Alan Frieze, Jeong Han Kim, Prasad Tetali, Eric Vigoda, and Van Ha Vu. Torpid mixing of some Monte Carlo Markov chain algorithms in statistical physics. In *40th Annual Symposium on Foundations of Computer Science*, pages 218–229. IEEE, 1999.

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

[6] Roger Brockett. Oscillatory descent for function minimization. In *Current and future directions in applied mathematics*, pages 65–82. Springer, 1997.

[7] Russ Bubley and Martin Dyer. Path coupling: A technique for proving rapid mixing in markov chains. In *Foundations of Computer Science, 1997. Proceedings., 38th Annual Symposium on*, pages 223–231. IEEE, 1997.

[8] Simone Calogero. Exponential convergence to equilibrium for kinetic fokker-planck equations. *Communications in Partial Differential Equations*, 37(8):1357–1390, 2012.

[9] Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323. PMLR, 06–09 Jul 2018.

[10] Sai Hung Cheung and James L Beck. Bayesian model updating using hybrid Monte Carlo simulation with application to structural dynamic models with many uncertain parameters. *Journal of engineering mechanics*, 135(4):243–255, 2009.

[11] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

[12] Persi Diaconis. The Markov chain Monte Carlo revolution. *Bulletin of the American Mathematical Society*, 46(2):179–205, 2009.

[13] Persi Diaconis, Kshitij Khare, and Laurent Saloff-Coste. Gibbs sampling, exponential families and orthogonal polynomials. *Statistical Science*, 23(2):151–178, 2008.

[14] Jean Dolbeault, Clément Mouhot, and Christian Schmeiser. Hypocoercivity for linear kinetic equations conserving mass. *Transactions of the American Mathematical Society*, 367(6):3807–3828, 2015.

[15] C dric Villani. *Hypocoercivity*. Number 949-951. American Mathematical Soc., 2009.

[16] Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

[17] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the Unadjusted Langevin Algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

[18] Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! *To appear in Proceedings of COLT 2018, arXiv preprint arXiv:1801.02309*, 2018.

[19] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *arXiv preprint arXiv:1703.01617*, 2017.

[20] D. Ferré, L. Hervé, and J. Ledoux. Regular perturbation of V-geometrically ergodic Markov chains. *Journal of Applied Probability*, 50(1):184–194, 2013.

[21] Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[22] Martin Hairer, Andrew M Stuart, and Sebastian J Vollmer. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *The Annals of Applied Probability*, 24(6):2455–2490, 2014.

[23] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms

in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

[24] Jeremy Heng and Pierre E Jacob. Unbiased Hamiltonian Monte Carlo with couplings. *arXiv preprint arXiv:1709.00404*, 2017.

[25] Frédéric Hérau and Francis Nier. Isotropic hypoellipticity and trend to equilibrium for the Fokker-Planck equation with a high-degree potential. *Archive for Rational Mechanics and Analysis*, 171(2):151–218, 2004.

[26] James E Johndrow, Jonathan C Mattingly, Sayan Mukherjee, and David Dunson. Approximations of Markov chains and high-dimensional Bayesian inference. *arXiv preprint arXiv:1508.03387*, 2015.

[27] Galin L Jones and James P Hobert. Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, 16(4):312–334, 2001.

[28] Yin Tat Lee and Santosh S Vempala. Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1115–1121. ACM, 2018.

[29] David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. *Markov chains and mixing times*. American Mathematical Soc., 2009.

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

[31] Oren Mangoubi, Natesh Pillai, and Aaron Smith. Does Hamiltonian Monte Carlo mix faster than a random walk on multimodal densities? *arXiv preprint arXiv:1808.03230*, 2018.

[32] Oren Mangoubi, Natesh Pillai, and Aaron Smith. Simple conditions for metastability of continuous Markov chains. *arXiv preprint arXiv:1808.03239*, 2018.

[33] Oren Mangoubi and Aaron Smith. Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions 1: Continuous dynamics. *preprint*, 2017.

[34] Oren Mangoubi and Aaron Smith. Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. *arXiv preprint*, 2017.

[35] Oren Mangoubi and Aaron Smith. Rapid mixing of geodesic walks on manifolds with positive curvature. *The Annals of Applied Probability*, 28(4):2501–2543, 2018.

[36] Oren Mangoubi and Nisheeth K Vishnoi. Dimensionally tight running time bounds for second-order Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems*, 2018.

[37] B Mehlig, DW Heermann, and BM Forrest. Hybrid Monte Carlo method for condensed-matter systems. *Physical Review B*, 45(2):679, 1992.

[38] Stéphane Mischler and Clément Mouhot. Exponential stability of slowly decaying solutions to the kinetic Fokker-Planck equation. *Archive for rational mechanics and analysis*, 221(2):677–723, 2016.

[39] Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2:113–162, 2011.

[40] Jeffrey Negrea and Jeffrey S Rosenthal. Error bounds for approximations of geometrically ergodic Markov chains. *arXiv preprint arXiv:1702.07441*, 2017.

[41] Yann Ollivier. Ricci curvature of Markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009.

[42] Michela Ottobre, Natesh S Pillai, Frank J Pinski, and Andrew M Stuart. A function space HMC algorithm with second order Langevin diffusion limit. *Bernoulli*, 22(1):60–106, 2016.

[43] Natesh S Pillai and Aaron Smith. Ergodicity of approximate MCMC chains with applications to large data sets. *arXiv preprint arXiv:1405.0182*, 2014.

[44] Gareth O Roberts and Jeffrey S Rosenthal. Complexity bounds for Markov chain Monte Carlo algorithms via diffusion limits. *J. Appl. Prob*, 53(2):1–11, 2016.

[45] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab*, 18(82):1–9, 2013.

[46] Daniel Rudolf and Nikolaus Schweizer. Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli*, 24(4A):2610–2639, 2018.

[47] Christof Seiler, Simon Rubinstein-Salzedo, and Susan Holmes. Positive curvature and Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 586–594, 2014.

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

[49] Cédric Villani. Hypocoercive diffusion operators. In *International Congress of Mathematicians*, volume 3, pages 473–498, 2006.

[50] Raphael Zimmer. Explicit contraction rates for a class of degenerate and infinite-dimensional diffusions. *Stochastics and Partial Differential Equations: Analysis and Computations*, 5(3):368–399, 2017.