Optimal Convergence Rate of Hamiltonian Monte Carlo for Strongly Logconcave Distributions

Zongchen Chen
School of Computer Science, Georgia Institute of Technology, USA
chenzongchen@gatech.edu

Santosh S. Vempala
School of Computer Science, Georgia Institute of Technology, USA
vempala@gatech.edu

Abstract

We study *Hamiltonian Monte Carlo* (HMC) for sampling from a strongly logconcave density proportional to e^{-f} where $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth (the condition number is $\kappa = L/\mu$). We show that the relaxation time (inverse of the spectral gap) of ideal HMC is $O(\kappa)$, improving on the previous best bound of $O(\kappa^{1.5})$; we complement this with an example where the relaxation time is $\Omega(\kappa)$. When implemented using a nearly optimal ODE solver, HMC returns an ε-approximate point in 2-Wasserstein distance using $\widetilde{O}((\kappa d)^{0.5}\varepsilon^{-1})$ gradient evaluations per step and $\widetilde{O}((\kappa d)^{1.5}\varepsilon^{-1})$ total time.

2012 ACM Subject Classification Theory of computation → Random walks and Markov chains; Theory of computation → Design and analysis of algorithms

Keywords and phrases logconcave distribution, sampling, Hamiltonian Monte Carlo, spectral gap, strong convexity

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2019.64

Category RANDOM

Funding *Zongchen Chen*: This work was supported in part by CCF-1563838 and CCF-1617306. *Santosh S. Vempala*: This work was supported in part by CCF-1563838, CCF-1717349 and DMS-1839323.

Acknowledgements We are grateful to Yin Tat Lee for helpful discussions.

1 Introduction

Sampling logconcave densities is a basic problem that arises in machine learning, statistics, optimization, computer science and other areas. The problem is described as follows. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Our goal is to sample from the density proportional to $e^{-f(x)}$. We study *Hamiltonian Monte Carlo* (HMC), one of the most widely-used *Markov chain Monte Carlo* (MCMC) algorithms for sampling from a probability distribution. In many settings, HMC is believed to outperform other MCMC algorithms such as the Metropolis-Hastings algorithm or Langevin dynamics. In terms of theory, rapid mixing has been established for HMC in recent papers [9, 10, 13, 14, 15] under various settings. However, in spite of much progress, there is a gap between known upper and lower bounds even in the basic setting when f is strongly convex (e^{-f} is strongly logconcave) and has a Lipschitz gradient.

Many sampling algorithms such as the Metropolis-Hastings algorithm or Langevin dynamics maintain a position $x = x(t)$ that changes with time, so that the distribution of x will eventually converge to the desired distribution, i.e., proportional to $e^{-f(x)}$. In HMC, besides the position $x = x(t)$, we also maintain a velocity $v = v(t)$. In the simplest Euclidean setting, the Hamiltonian $H(x, v)$ is defined as

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

Then in every step the pair (x, v) is updated using the following system of differential equations for a fixed time interval T:

$$\begin{cases} \dfrac{\mathrm{d}x(t)}{\mathrm{d}t} = \dfrac{\partial H(x,v)}{\partial v} = v(t), \\ \dfrac{\mathrm{d}v(t)}{\mathrm{d}t} = -\dfrac{\partial H(x,v)}{\partial x} = -\nabla f(x(t)). \end{cases} \tag{1}$$

The initial position $x(0) = x_0$ is the position from the last step, and the initial velocity $v(0) = v_0$ is chosen randomly from the standard Gaussian distribution $N(0, I)$. The updated position is $x(T)$ where T can be thought of as the step-size. It is well-known that the stationary distribution of HMC is the density proportional to e^{-f}. Observe that

$$\frac{\mathrm{d}H(x,v)}{\mathrm{d}t} = \frac{\partial H(x,v)}{\partial x}x'(t) + \frac{\partial H(x,v)}{\partial v}v'(t) = 0,$$

so the Hamiltonian $H(x, v)$ does not change with t. We can also write (1) as the following ordinary differential equation (ODE):

$$x''(t) = -\nabla f(x(t)), \quad x(0) = x_0, \quad x'(0) = v_0. \tag{2}$$

We state HMC explicitly as the following algorithm.

Algorithm 1 Hamiltonian Monte Carlo algorithm.

Input: $f : \mathbb{R}^d \to \mathbb{R}$ that is μ-strongly convex and L-smooth, ε the error parameter.

1. Set starting point $x^{(0)}$, step-size T, number of steps N, and ODE error tolerance δ.
2. For $k = 1, \ldots, N$:
   a. Let $v \sim N(0, I)$;
   b. Denote by $x(t)$ the solution to (1) with initial position $x(0) = x^{(k-1)}$ and initial velocity $v(0) = v$. Use the ODE solver to find a point $x^{(k)}$ such that
   $$\left\| x^{(k)} - x(T) \right\| \le \delta.$$
3. Output $x^{(N)}$.

In our analysis, we first consider *ideal* HMC where in every step we have the exact solution to the ODE (1) and neglect the numerical error from solving the ODEs or integration ($\delta = 0$).

1.1 Preliminaries

We recall standard definitions here. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function. We say f is μ-strongly convex if for all $x, y \in \mathbb{R}^d$,

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

We say f is L-smooth if ∇f is L-Lipschitz; i.e., for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f(x) - \nabla f(y)\| \le L \|x - y\| .$$

If f is μ-strongly convex and L-smooth, then the condition number of f is $\kappa = L/\mu$.

Consider a discrete-time *reversible* Markov chain $\mathcal{M}$ on $\mathbb{R}^d$ with stationary distribution π. Let

$$L_2(\pi) = \left\{ f : \mathbb{R}^d \to \mathbb{R} \,\middle|\, \int_{\mathbb{R}^d} f(x)^2 \pi(\mathrm{d}x) < \infty \right\}$$

be the Hilbert space with inner product

$$\langle f, g \rangle = \int_{\mathbb{R}^d} f(x) g(x) \pi(\mathrm{d}x)$$

for $f, g \in L_2(\pi)$. Denote by P the transition kernel of $\mathcal{M}$. We can view P as a self-adjoint operator from $L_2(\pi)$ to itself: for $f \in L_2(\pi)$,

$$(Pf)(x) = \int_{\mathbb{R}^d} f(y) P(x, \mathrm{d}y).$$

Let $L_2^0(\pi) = \{f \in L_2(\pi) : \int_{\mathbb{R}^d} f(x)\pi(\mathrm{d}x) = 0\}$ be a closed subspace of $L_2(\pi)$. The (absolute) *spectral gap* of P is defined to be

$$\gamma(P) = 1 - \sup_{f \in L_2^0(\pi)} \frac{\|Pf\|}{\|f\|} = 1 - \sup_{\substack{f \in L_2^0(\pi) \\ \|f\|=1}} |\langle Pf, f \rangle|.$$

The relaxation time of P is

$$\tau_{\mathrm{rel}}(P) = \frac{1}{\gamma(P)}.$$

Let ν_1, ν_2 be two distributions on $\mathbb{R}^d$. The 2-Wasserstein distance between ν_1 and ν_2 is defined as

$$W_2(\nu_1, \nu_2) = \left(\inf_{(X,Y) \in \mathcal{C}(\nu_1, \nu_2)} \mathbb{E}\left[\|X - Y\|^2 \right] \right)^{1/2},$$

where $\mathcal{C}(\nu_1, \nu_2)$ is the set of all couplings of ν_1 and ν_2.

1.2 Related work

Various versions of Langevin dynamics have been studied in many recent papers, see [5, 6, 21, 17, 7, 4, 3, 2, 8, 20, 19, 12]. The convergence rate of HMC is also studied recently in [9, 10, 13, 14, 15, 18]. The first bound for our setting was obtained by Mangoubi and Smith [13], who gave an $O(\kappa^2)$ bound on the convergence rate of ideal HMC.

▶ **Theorem 1** ([13, Theorem 1]). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function such that $\mu I \preceq \nabla^2 f(x) \preceq LI$ for all $x \in \mathbb{R}^d$. Then the relaxation time of ideal HMC for sampling from the density $\propto e^{-f}$ with step-size $T = \sqrt{\mu}/(2\sqrt{2}L)$ is $O(\kappa^2)$.*

This was improved by [9], which showed a bound of $O(\kappa^{1.5})$. They also gave a nearly optimal method for solving the ODE that arises in the implementation of HMC.

▶ **Theorem 2** ([9, Lemma 1.8]). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function such that $\mu I \preceq \nabla^2 f(x) \preceq LI$ for all $x \in \mathbb{R}^d$. Then the relaxation time of ideal HMC for sampling from the density $\propto e^{-f}$ with step-size $T = \mu^{1/4}/(2L^{3/4})$ is $O(\kappa^{1.5})$.*

Both papers suggest that the correct bound is linear in κ: [13] says linear is the best one can expect while [9] shows that there *exists* a choice of step-sizes (time for running the ODE) that might achieve a linear rate (Lemma 1.8, second part); however it was far from clear how to determine these step-sizes algorithmically.

Other papers focus on various aspects and use stronger assumptions (e.g., bounds on higher-order gradients) to get better bounds on the overall convergence time or the number of gradient evaluations in some ranges of parameters. For example, [15] shows that the dependence on dimension for the number of gradient evaluations can be as low as $d^{1/4}$ with suitable regularity assumptions (and higher dependence on the condition number). We note also that sampling logconcave functions is a polynomial-time solvable problem, without the assumptions of strong convexity or gradient Lipschitzness, and even when the function e^{-f} is given only by an oracle with no access to gradients [1, 11]. The Riemannian version of HMC provides a faster polynomial-time algorithm for uniformly sampling polytopes [10]. However, the dependence on the dimension is significantly higher for these algorithms, both for the contraction rate and the time per step.

1.3 Results

In this paper, we show that the relaxation time of ideal HMC is $\Theta(\kappa)$ for strongly logconcave functions with Lipschitz gradient.

▶ **Theorem 3.** *Suppose that f is μ-strongly convex and L-smooth. Then the relaxation time (inverse of spectral gap) of ideal HMC for sampling from the density $\propto e^{-f}$ with step-size $T = 1/(2\sqrt{L})$ is $O(\kappa)$, where $\kappa = L/\mu$ is the condition number.*

We remark that the only assumption we made about f is strongly convexity and smoothness (in particular, we do not require that f is twice differentiable, which is assumed in both [9] and [13]).

We also establish a matching lower bound on the relaxation time of ideal HMC, implying the tightness of Theorem 3.

▶ **Theorem 4.** *For any $0 < \mu \le L$, there exists a μ-strongly convex and L-smooth function f, such that the relaxation time of ideal HMC for sampling from the density $\propto e^{-f}$ with step-size $T = O(1/\sqrt{L})$ is $\Omega(\kappa)$, where $\kappa = L/\mu$ is the condition number.*

Using the nearly optimal ODE solver from [9], we obtain the following convergence rate in 2-Wasserstein distance for the HMC algorithm. We note that since our new convergence rate allows larger steps, the ODE solver is run for a longer time step.

▶ **Theorem 5.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function such that $\mu I \preceq \nabla^2 f(x) \preceq LI$ for all $x \in \mathbb{R}^d$. Let $\pi \propto e^{-f}$ be the target distribution, and let π_{HMC} be the distribution of the output of HMC with starting point $x^{(0)} = \arg\min_x f(x)$, step-size $T = 1/(16000\sqrt{L})$, and ODE error tolerance $\delta = \sqrt{\mu} T^2 \varepsilon/16$. For any $0 < \varepsilon < \sqrt{d}$, if we run HMC for $N = O\left(\kappa \log(d/\varepsilon)\right)$ steps where $\kappa = L/\mu$, then we have*

$$W_2(\pi_{\text{HMC}}, \pi) \le \frac{\varepsilon}{\sqrt{\mu}}.$$

Each step takes $O\left(\sqrt{\kappa} d^{3/2} \varepsilon^{-1} \log(\kappa d/\varepsilon)\right)$ time and $O\left(\sqrt{\kappa d} \varepsilon^{-1} \log(\kappa d/\varepsilon)\right)$ evaluations of ∇f, amortized over all steps.

The comparison of convergence rates, running times and numbers of gradient evaluations is summarized in the following table with polylog factors omitted.

reference	convergence rate	# gradients	total time
[13]	κ^2	$\kappa^{6.5}d^{0.5}$	$\kappa^{6.5}d^{1.5}$
[9]	$\kappa^{1.5}$	$\kappa^{1.75}d^{0.5}$	$\kappa^{1.75}d^{1.5}$
this paper	κ	$\kappa^{1.5}d^{0.5}$	$\kappa^{1.5}d^{1.5}$

2 Convergence of ideal HMC

In this section we show that the spectral gap of ideal HMC is $\Omega(1/\kappa)$, and thus prove Theorem 3. We first show a contraction bound for ideal HMC, which roughly says that the distance of two points is shrinking after one step of ideal HMC.

▶ **Lemma 6** (Contraction bound). *Suppose that f is μ-strongly convex and L-smooth. Let $x(t)$ and $y(t)$ be the solution to (1) with initial positions $x(0)$, $y(0)$ and initial velocities $x'(0) = y'(0)$. Then for $0 \le t \le 1/(2\sqrt{L})$ we have*

$$\|x(t) - y(t)\|^2 \le \left(1 - \frac{\mu}{4}t^2\right)\|x(0) - y(0)\|^2.$$

In particular, by setting $t = T = 1/(c\sqrt{L})$ for some constant $c \ge 2$ we get

$$\|x(T) - y(T)\|^2 \le \left(1 - \frac{1}{4c^2\kappa}\right)\|x(0) - y(0)\|^2$$

where $\kappa = L/\mu$.

Proof. Consider the two ODEs for HMC:

$$\begin{cases} x'(t) = u(t); \\ u'(t) = -\nabla f(x(t)). \end{cases} \quad \text{and} \quad \begin{cases} y'(t) = v(t); \\ v'(t) = -\nabla f(y(t)). \end{cases}$$

with initial points $x(0), y(0)$ and initial velocities $u(0) = v(0)$. For the sake of brevity, we shall write $x = x(t)$, $y = y(t)$, $u = u(t)$, $v = v(t)$ and omit the variable t, as well as letting $x_0 = x(0)$, $y_0 = y(0)$. We are going to show that

$$\|x - y\|^2 \le \left(1 - \frac{\mu}{4}t^2\right)\|x_0 - y_0\|^2$$

for all $0 \le t \le 1/(2\sqrt{L})$.

Consider the derivative of $\frac{1}{2}\|x - y\|^2$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\|x - y\|^2\right) = \langle x' - y', x - y\rangle = \langle u - v, x - y\rangle. \tag{3}$$

Taking derivative on both sides, we get

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(\frac{1}{2}\|x - y\|^2\right) &= \langle u' - v', x - y\rangle + \langle u - v, x' - y'\rangle \\ &= -\langle \nabla f(x) - \nabla f(y), x - y\rangle + \|u - v\|^2 \\ &= -\rho\|x - y\|^2 + \|u - v\|^2, \end{aligned} \tag{4}$$

where we define

$$\rho = \rho(t) = \frac{\langle \nabla f(x) - \nabla f(y), x - y \rangle}{\|x - y\|^2}.$$

Since f is μ-strongly convex and L-smooth, we have $\mu \le \rho \le L$ for all $t \ge 0$.

We will upper bound the term $-\rho \|x - y\|^2 + \|u - v\|^2$, while keeping its dependency on ρ. To lower bound $\|x - y\|^2$, we use the following crude bound.

▷ Claim 7 (Crude bound). For all $0 \le t \le 1/(2\sqrt{L})$ we have

$$\frac{1}{2} \|x_0 - y_0\|^2 \le \|x - y\|^2 \le 2 \|x_0 - y_0\|^2 . \tag{5}$$

The proof of this claim is postponed to Section 2.1.

Next we derive an upper bound on $\|u - v\|^2$. The derivative of $\|u - v\|$ is given by

$$\begin{aligned} \|u - v\| \left(\frac{\mathrm{d}}{\mathrm{d}t} \|u - v\| \right) &= \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{1}{2} \|u - v\|^2 \right) \\ &= \langle u' - v', u - v \rangle \\ &= - \langle \nabla f(x) - \nabla f(y), u - v \rangle . \end{aligned}$$

Thus, its absolute value is upper bounded by

$$\left| \frac{\mathrm{d}}{\mathrm{d}t} \|u - v\| \right| = \frac{|- \langle \nabla f(x) - \nabla f(y), u - v \rangle|}{\|u - v\|} \le \|\nabla f(x) - \nabla f(y)\| .$$

Since f is L-smooth and convex, we have

$$\|\nabla f(x) - \nabla f(y)\|^2 \le L \langle \nabla f(x) - \nabla f(y), x - y \rangle = L\rho \|x - y\|^2 \le 2L\rho \|x_0 - y_0\|^2 ,$$

where the last inequality follows from the crude bound (5). Then, using the fact that $u_0 = v_0$ and the Cauchy-Schwarz inequality, we can upper bound $\|u - v\|^2$ by

$$\begin{aligned} \|u - v\|^2 &\le \left(\int_0^t \left| \frac{\mathrm{d}}{\mathrm{d}s} \|u - v\| \right| \mathrm{d}s \right)^2 \\ &\le \left(\int_0^t \sqrt{2L\rho} \, \|x_0 - y_0\| \, \mathrm{d}s \right)^2 \\ &\le 2Lt \left(\int_0^t \rho \, \mathrm{d}s \right) \|x_0 - y_0\|^2 . \end{aligned}$$

Define the function

$$P = P(t) = \int_0^t \rho \, \mathrm{d}s,$$

so $P(t)$ is nonnegative and monotone increasing, with $P(0) = 0$. Also we have $\mu t \le P(t) \le Lt$ for all $t \ge 0$. Then,

$$\|u - v\|^2 \le 2LtP \|x_0 - y_0\|^2 . \tag{6}$$

Plugging (5) and (6) into (4), we deduce that

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \left(\frac{1}{2} \|x - y\|^2 \right) \le -\rho \left(\frac{1}{2} \|x_0 - y_0\|^2 \right) + 2LtP \|x_0 - y_0\|^2 .$$

If we define

$$\alpha(t) = \frac{1}{2}\|x-y\|^2,$$

then we have

$$\alpha''(t) \le -\alpha(0)\big(\rho(t) - 4LtP(t)\big).$$

Integrating both sides and using $\alpha'(0) = 0$, we obtain

$$\begin{aligned}\alpha'(t) &= \int_0^t \alpha''(s)\mathrm{d}s \\ &\le -\alpha(0)\left(\int_0^t \rho(s)\mathrm{d}s - 4L\int_0^t sP(s)\mathrm{d}s\right) \\ &\le -\alpha(0)\left(P(t) - 4LP(t)\int_0^t s\mathrm{d}s\right) \\ &= -\alpha(0)P(t)\left(1-2Lt^2\right),\end{aligned}$$

where the second inequality is due to the monotonicity of $P(s)$. Since for all $0 \le t \le 1/(2\sqrt{L})$ we have $P(t) \ge \mu t$ and $1 - 2Lt^2 \ge 1/2$, we deduce that

$$\alpha'(t) \le -\alpha(0)\frac{\mu}{2}t.$$

Finally, one more integration yields

$$\alpha(t) = \alpha(0) + \int_0^t \alpha'(s)\mathrm{d}s \le \alpha(0)\left(1 - \frac{\mu}{4}t^2\right),$$

and the theorem follows. ◀

Proof of Theorem 3. Lemma 6 implies that for any constant $c \ge 2$, the Ricci curvature of ideal HMC with step-size $T = 1/(c\sqrt{L})$ is at least $1/(8c^2\kappa)$. Then, it follows from [16, Proposition 29] that the spectral gap of ideal HMC is at least $1/(8c^2\kappa)$. Hence, the relaxation time is upper bounded by $8c^2\kappa = O(\kappa)$. ◀

2.1 Proof of Claim 7

We present the proof of Claim 7 in this section. We remark that a similar crude bound was established in [9] for general matrix ODEs. Here we prove the crude bound specifically for the Hamiltonian ODE, but without assuming that f is twice differentiable.

Proof of Claim 7. We first derive a crude upper bound on $\|u-v\|$. Since f is L-smooth, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|u-v\| &= \frac{-\langle \nabla f(x) - \nabla f(y), u-v\rangle}{\|u-v\|} \\ &\le \|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|.\end{aligned}$$

Then from $u_0 = v_0$ we get

$$\|u-v\| = \int_0^t \left(\frac{\mathrm{d}}{\mathrm{d}s}\|u-v\|\right)\mathrm{d}s \le L\int_0^t \|x-y\|\,\mathrm{d}s.$$

To obtain the upper bound for $\|x-y\|$, we first bound its derivative by

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\|x-y\|\right| = \frac{|\langle u-v, x-y\rangle|}{\|x-y\|} \le \|u-v\| \le L\int_0^t \|x-y\|\,\mathrm{d}s. \tag{7}$$

Therefore,

$$\begin{aligned}\|x-y\| &= \|x_0-y_0\| + \int_0^t \left(\frac{\mathrm{d}}{\mathrm{d}s}\|x-y\|\right)\mathrm{d}s\\ &\le \|x_0-y_0\| + L\int_0^t\int_0^s \|x-y\|\,\mathrm{d}r\mathrm{d}s\\ &= \|x_0-y_0\| + L\int_0^t (t-s)\|x-y\|\,\mathrm{d}s.\end{aligned}$$

We then deduce from [9, Lemma A.5] that

$$\|x-y\| \le \|x_0-y_0\|\cosh\left(\sqrt{L}t\right) \le \sqrt{2}\,\|x_0-y_0\|, \tag{8}$$

where we use the fact that $\cosh(\sqrt{L}t) \le \cosh(1/2) \le \sqrt{2}$.

Next, we deduce from (7) and (8) that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|x-y\| &\ge -L\int_0^t \|x-y\|\,\mathrm{d}s\\ &\ge -L\|x_0-y_0\|\int_0^t \cosh\left(\sqrt{L}s\right)\mathrm{d}s\\ &= -\sqrt{L}\|x_0-y_0\|\sinh\left(\sqrt{L}s\right).\end{aligned}$$

Thus, we obtain

$$\begin{aligned}\|x-y\| &= \|x_0-y_0\| + \int_0^t \left(\frac{\mathrm{d}}{\mathrm{d}s}\|x-y\|\right)\mathrm{d}s\\ &\ge \|x_0-y_0\| - \sqrt{L}\|x_0-y_0\|\int_0^t \sinh\left(\sqrt{L}s\right)\mathrm{d}s\\ &= \|x_0-y_0\|\left(2-\cosh\left(\sqrt{L}t\right)\right) \ge \frac{1}{\sqrt{2}}\|x_0-y_0\|,\end{aligned}$$

where we use $2-\cosh(\sqrt{L}t) \ge 2-\cosh(1/2) \ge 1/\sqrt{2}$. ◁

3 Lower bound for ideal HMC

In this section, we show that the relaxation time of ideal HMC can achieve $\Theta(\kappa)$ for some μ-strongly convex and L-smooth function, and thus prove Theorem 4.

Consider a two-dimensional quadratic function:

$$f(x_1,x_2) = \frac{x_1^2}{2\sigma_1^2} + \frac{x_2^2}{2\sigma_2^2},$$

where $\sigma_1 = 1/\sqrt{\mu}$ and $\sigma_2 = 1/\sqrt{L}$. Thus, f is μ-strongly convex and L-smooth. The probability density ν proportional to e^{-f} is essentially the bivariate Gaussian distribution: for $(x_1,x_2)\in\mathbb{R}^2$,

$$\nu(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{x_1^2}{2\sigma_1^2} - \frac{x_2^2}{2\sigma_2^2}\right).$$

The following lemma shows that ideal HMC for the bivariate Gaussian distribution ν has relaxation time $\Omega(\kappa)$, and then Theorem 4 follows immediately.

▶ **Lemma 8.** *For any constant $c > 0$, the relaxation time of ideal HMC for sampling from ν with step-size $T = 1/(c\sqrt{L})$ is at least $2c^2\kappa$.*

Proof. The Hamiltonian curve for f is given by the ODE

$$(x_1'', x_2'') = -\nabla f(x_1, x_2) = \left(-\frac{x_1}{\sigma_1^2}, -\frac{x_2}{\sigma_2^2}\right)$$

with initial position $(x_1(0), x_2(0))$ and initial velocity $(x_1'(0), x_2'(0))$ from the bivariate standard Gaussian $N(0, I)$. Observe that $\nu = \nu_1 \otimes \nu_2$ is a product distribution of the two coordinates and HMC for f is a product chain. Thus, we can consider the dynamics for each coordinate separately. The Hamiltonian ODE for one coordinate becomes

$$x_i'' = -\frac{x_i}{\sigma_i^2}, \quad x_i(0), \quad x_i'(0) = v_i(0) \sim N(0, 1)$$

where $i = 1, 2$. Solving the ODE above and plugging in the step-size $t = T$, we get

$$x_i(T) = x_i(0)\cos(T/\sigma_i) + v_i(0)\sigma_i \sin(T/\sigma_i).$$

Let P_i be the transition kernel of ideal HMC for the ith coordinate (considered as a Markov chain on $\mathbb{R}$). Then for $x, y \in \mathbb{R}$ we have

$$P_i(x, y) = \frac{1}{\sqrt{2\pi}\sigma_i \sin(T/\sigma_i)} \exp\left(-\frac{(y - x\cos(T/\sigma_i))^2}{2\sigma_i^2 \sin^2(T/\sigma_i)}\right).$$

Namely, given the current position x, the next position y is from a normal distribution with mean $x\cos(T/\sigma_i)$ and variance $\sigma_i^2 \sin^2(T/\sigma_i)$. Denote the spectral gap of P_i by γ_i for $i = 1, 2$ and that of ideal HMC by γ. Let $h(x) = x$ and note that $h \in L_2^0(\nu_i)$. Using the properties of product chains and spectral gaps, we deduce that

$$\gamma \le \min\{\gamma_1, \gamma_2\} \le \gamma_1 = 1 - \sup_{f \in L_2^0(\nu_1)} \frac{|\langle P_1 f, f\rangle|}{\|f\|^2} \le 1 - \frac{|\langle P_1 h, h\rangle|}{\|h\|^2}.$$

Since we have

$$\|h\|^2 = \int_{-\infty}^{\infty} \nu_1(x) h(x)^2 \mathrm{d}x = \sigma_1^2$$

and

$$\langle P_1 h, h\rangle = \int_{-\infty}^{\infty} \nu_1(x) P_1(x, y) h(x) h(y) \mathrm{d}x\mathrm{d}y = \sigma_1^2 \cos(T/\sigma_1),$$

it follows that $\gamma \le 1 - |\cos(T/\sigma_1)|$. Suppose that $T = 1/(c\sqrt{L})$ for some $c > 0$. Then we get

$$\gamma \le \frac{T^2}{2\sigma_1^2} = \frac{1}{2c^2}\frac{\mu}{L},$$

and consequently $\tau_{\mathrm{rel}} = 1/\gamma \ge 2c^2\kappa$. ◀

4 Convergence rate of discretized HMC

In this section, we show how our improved contraction bound (Lemma 6) implies that HMC returns a good enough sample after $\widetilde{O}((\kappa d)^{1.5})$ steps. We will use the framework from [9] to establish Theorem 5.

We first state the ODE solver from [9], which solves an ODE in nearly optimal time when the solution to the ODE can be approximated by a piece-wise polynomial. We state here only for the special case of second order ODEs for the Hamiltonian system. We refer to [9] for general kth order ODEs.

▶ **Theorem 9** ([9, Theorem 2.5]). *Let $x(t)$ be the solution to the ODE*

$$x''(t) = -\nabla f(x(t)), \quad x(0) = x_0, \quad x'(0) = v_0.$$

where $x_0, v_0 \in \mathbb{R}^d$ and $0 \le t \le T$. Suppose that the following conditions hold:

1. *There exists a piece-wise polynomial $q(t)$ such that $q(t)$ is a polynomial of degree D on each interval $[T_{j-1}, T_j]$ where $0 = T_0 < T_1 < \cdots < T_m = T$, and for all $0 \le t \le T$ we have*

$$\|q(t) - x''(t)\| \le \frac{\delta}{T^2};$$

2. *$\{T_j\}_{j=1}^m$ and D are given as input to the ODE solver;*
3. *The function f has a L-Lipschitz gradient; i.e., for all $x, y \in \mathbb{R}^d$,*

$$\|\nabla f(x) - \nabla f(y)\| \le L \|x - y\|.$$

If $\sqrt{L}T \le 1/16000$, then the ODE solver can find a piece-wise polynomial $\tilde{x}(t)$ such that for all $0 \le t \le T$,

$$\|\tilde{x}(t) - x(t)\| \le O(\delta).$$

The ODE solver uses $O(m(D+1)\log(CT/\delta))$ evaluations of ∇f and $O(dm(D+1)^2 \log(CT/\delta))$ time where

$$C = O\left(\|v_0\| + T\|\nabla f(x_0)\|\right).$$

The following lemma, which combines Theorem 3.2, Lemma 4.1 and Lemma 4.2 from [9], establishes the conditions of Theorem 9 in our setting. We remark that Lemmas 4.1 and 4.2 hold for all $T \le 1/(8\sqrt{L})$, and Theorem 3.2, though stated only for $T \le O(\mu^{1/4}/L^{3/4})$ in [9], holds in fact for the whole region $T \le 1/(2\sqrt{L})$ where the contraction bound (Lemma 6) is true. We omit these proofs here and refer the readers to [9] for more details.

▶ **Lemma 10.** *Let f be a twice differentiable function such that $\mu I \preceq \nabla^2 f(x) \preceq LI$ for all $x \in \mathbb{R}^d$. Choose the starting point $x^{(0)} = \arg\min_x f(x)$, step-size $T = 1/(16000\sqrt{L})$, and ODE error tolerance $\delta = \sqrt{\mu}T^2\varepsilon/16$ in the HMC algorithm. Let $\{x^{(k)}\}_{k=1}^N$ be the sequence of points we get from the HMC algorithm and $\{v_0^{(k)}\}_{k=1}^N$ be the sequence of random Gaussian vector we choose in each step. Let $\pi \propto e^{-f}$ be the target distribution and let π_{HMC} be the distribution of $x^{(N)}$, i.e., the output of HMC. For any $0 < \varepsilon < \sqrt{d}$, if we run HMC for*

$$N = O\left(\frac{\log(d/\varepsilon)}{\mu T^2}\right) = O\left(\kappa \log(d/\varepsilon)\right)$$

steps where $\kappa = L/\mu$, then:

1. ([9, Theorem 3.2]) *We have that*

$$W_2(\pi_{\text{HMC}}, \pi) \le \frac{\varepsilon}{\sqrt{\mu}};$$

2. ([9, Lemma 4.1]) *For each k, let $x_k(t)$ be the solution to the ODE (2) in the kth step of HMC. Then there is a piece-wise constant function q_k of m_k pieces such that $\|q_k(t) - x_k''(t)\| \le \delta/T^2$ for all $0 \le t \le T$, where*

$$m_k = \frac{2LT^3}{\delta}\left(\left\|v_0^{(k-1)}\right\| + T\left\|\nabla f(x^{(k-1)})\right\|\right);$$

3. ([9, Lemma 4.2]) *We have that*

$$\frac{1}{N}\,\mathbb{E}\left[\sum_{k=1}^{N}\left\|\nabla f(x^{(k-1)})\right\|^2\right] \le O(Ld).$$

Proof of Theorem 5. The convergence of HMC is guaranteed by part 1 of Lemma 10. In the kth step, the number of evaluations of ∇f is $O(m_k \log(C_k\sqrt{\kappa}/\varepsilon))$ by Theorem 9 and part 2 of Lemma 10, where

$$m_k = O\left(\frac{\sqrt{\kappa}}{\varepsilon}\right)\left(\left\|v_0^{(k-1)}\right\| + T\left\|\nabla f(x^{(k-1)})\right\|\right)$$

and

$$C_k = O\left(\left\|v_0^{(k-1)}\right\| + T\left\|\nabla f(x^{(k-1)})\right\|\right).$$

Thus, the average number of evaluations of ∇f per step is upper bounded by

$$\frac{1}{N}\,\mathbb{E}\left[\sum_{k=1}^{N} O(m_k \log(C_k\sqrt{\kappa}/\varepsilon))\right] \le \frac{1}{N}\,\mathbb{E}\left[\sum_{k=1}^{N} O(m_k \log m_k)\right] \le \frac{1}{N}\,O\left(\mathbb{E}\left[M \log M\right]\right),$$

where $M = \sum_{k=1}^{N} m_k$. Since each $v_0^{(k-1)}$ is sampled from the standard Gaussian distribution, we have $\mathbb{E}\left[\left\|v_0^{(k-1)}\right\|^2\right] = d$. Thus, by the Cauchy-Schwarz inequality and part 3 of Lemma 10, we get

$$\begin{aligned}\mathbb{E}\left[M^2\right] \le N\sum_{k=1}^{N}\mathbb{E}\left[m_k^2\right] &\le O\left(\frac{N\kappa}{\varepsilon^2}\right)\sum_{k=1}^{N}\mathbb{E}\left[\left\|v_0^{(k-1)}\right\|^2\right] + T^2\mathbb{E}\left[\left\|\nabla f(x^{(k-1)})\right\|^2\right] \\ &\le O\left(\frac{N^2\kappa d}{\varepsilon^2}\right).\end{aligned}$$

We then deduce again from the Cauchy-Schwarz inequality that

$$(\mathbb{E}[M\log M])^2 \le \mathbb{E}\left[M^2\right]\cdot\mathbb{E}\left[\log^2 M\right] \le \mathbb{E}\left[M^2\right]\cdot\log^2\left(\mathbb{E}M\right) \le \mathbb{E}[M^2]\cdot\log^2\left(\sqrt{\mathbb{E}\left[M^2\right]}\right),$$

where the second inequality is due to that $h(x) = \log^2(x)$ is concave when $x \ge 3$. Therefore, the number of evaluations of ∇f per step, amortized over all steps, is

$$\frac{1}{N}\,O\left(\sqrt{\mathbb{E}[M^2]}\log\left(\sqrt{\mathbb{E}\left[M^2\right]}\right)\right) \le O\left(\frac{\sqrt{\kappa d}}{\varepsilon}\log\left(\frac{\kappa d}{\varepsilon}\right)\right).$$

Using a similar argument we have the bound for the expected running time per step. This completes the proof. ◀

References

1. David Applegate and Ravi Kannan. Sampling and integration of near log-concave functions. In *Proceedings of the 23rd Annual ACM Symposium on Theory of Computing (STOC)*, pages 156–163. ACM, 1991.
2. Niladri S. Chatterji, Nicolas Flammarion, Yi-An Ma, Peter L. Bartlett, and Michael I. Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

3 Xiang Cheng, Niladri S. Chatterji, Yasin Abbasi-Yadkori, Peter L. Bartlett, and Michael I. Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *ArXiv preprint*, 2018. arXiv:1805.01648.

4 Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 2018 Conference on Learning Theory (COLT)*, 2018.

5 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.

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

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

8 Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Proceedings of the 2018 Conference on Learning Theory (COLT)*, 2018.

9 Yin Tat Lee, Zhao Song, and Santosh S. Vempala. Algorithmic Theory of ODEs and Sampling from Well-conditioned Logconcave Densities. *ArXiv preprint*, 2018. arXiv:1812.06243.

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

11 László Lovász and Santosh S. Vempala. Fast Algorithms for Logconcave Functions: Sampling, Rounding, Integration and Optimization. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 57–68. IEEE, 2006.

12 Yi-An Ma, Niladri Chatterji, Xiang Cheng, Nicolas Flammarion, Peter Bartlett, and Michael I. Jordan. Is There an Analog of Nesterov Acceleration for MCMC? *ArXiv preprint*, 2019. arXiv:1902.00996.

13 Oren Mangoubi and Aaron Smith. Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions 1: continuous dynamics. *ArXiv preprint*, 2017. arXiv:1708.07114.

14 Oren Mangoubi and Aaron Smith. Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions 2: Numerical integrators. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 586–595, 2019.

15 Oren Mangoubi and Nisheeth Vishnoi. Dimensionally tight bounds for second-order Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6027–6037, 2018.

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

17 Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory (COLT)*, 2017.

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

19 Santosh S. Vempala and Andre Wibisono. Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices. *ArXiv preprint*, 2019. arXiv:1903.08568.

20 Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Proceedings of the 2018 Conference on Learning Theory (COLT)*, 2018.

21 Yuchen Zhang, Percy S. Liang, and Moses Charikar. A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics. In *Proceedings of the 2017 Conference on Learning Theory (COLT)*, 2017.