THEORY OF COMPUTING, Volume 18 (9), 2022, pp. 1–18
www.theoryofcomputing.org

SPECIAL ISSUE: APPROX-RANDOM 2019

Optimal Convergence Rate of Hamiltonian Monte Carlo for Strongly Logconcave Distributions

Zongchen Chen* Santosh S. Vempala†

Received February 29, 2020; Revised September 23, 2020; Published April 30, 2022

Abstract. We study the *Hamiltonian Monte Carlo* (HMC) algorithm 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 $O(\kappa^{1.5})$ (Lee et al., 2018); we complement this with an example where the relaxation time is $\Omega(\kappa)$, for any step-size. When implemented with an ODE solver, HMC returns an ε-approximate point in 2-Wasserstein distance using $\tilde{O}((\kappa d)^{0.5}\varepsilon^{-1})$ gradient evaluations per step and $\tilde{O}((\kappa d)^{1.5}\varepsilon^{-1})$ total time.

A conference version [5] of this paper appeared in the Proceedings of the 23rd International Conference on Randomization and Computation (RANDOM 2019).

*Supported by CCF-1563838 and CCF-1617306.

†Supported by CCF-1563838, CCF-1717349 and DMS-1839323.

ACM Classification: Theory of computation - Random walks and Markov chains, Theory of computation - Design and analysis of algorithms

AMS Classification: 60J05, 60J25, 68W20

Key words and phrases: logconcave distribution, sampling, Hamiltonian Monte Carlo, spectral gap, strong convexity

 DOI: 10.4086/toc.2022.v018a009

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 $\mathrm{e}^{-f(x)}$. 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 [14, 15, 18, 19, 20] in various settings. However, in spite of much progress, a gap remains between the 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 $\mathrm{e}^{-f(x)}$. 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

$$H(x, v) = f(x) + \frac{1}{2}\|v\|^2. \tag{1.1}$$

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

$$\begin{cases} \dfrac{\mathrm{d}x(t)}{\mathrm{d}t} = \nabla_v H(x, v) = v(t), \\ \dfrac{\mathrm{d}v(t)}{\mathrm{d}t} = -\nabla_x H(x, v) = -\nabla f(x(t)). \end{cases} \tag{1.2}$$

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

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

so the Hamiltonian $H(x, v)$ does not change with t. We can also write (1.2) as the following ordinary differential equation (ODE):

$$x''(t) = -\nabla f(x(t)), \quad x(0) = x_0, \quad x'(0) = v_0. \tag{1.3}$$

We state HMC explicitly as the following algorithm; also see Figure 1 for an illustration.

Hamiltonian Monte Carlo algorithm

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

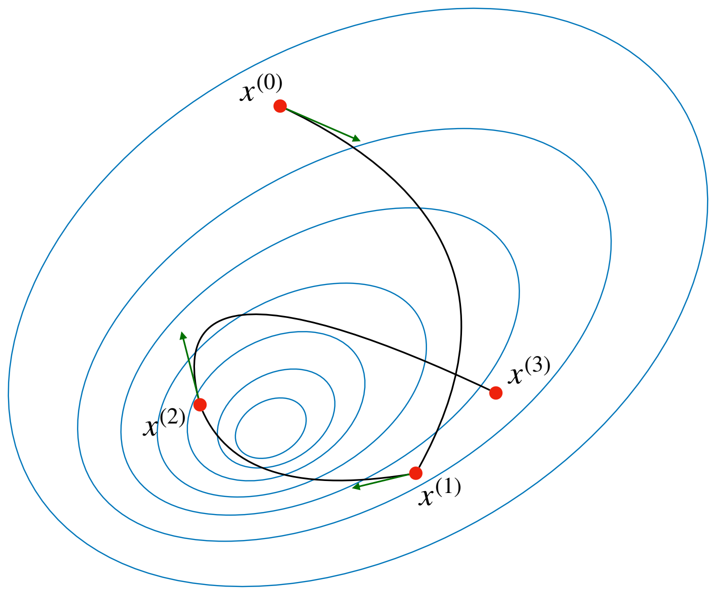

Figure 1: A trajectory of Hamiltonian Monte Carlo

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 (1.2) 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

$$\left\| x^{(k)} - x(T) \right\| \leq \delta.$$

3. Output $x^{(N)}$.

In our analysis, we first consider *ideal* HMC where in every step we have the exact solution to the ODE (1.2) and neglect the numerical error from solving the ODEs or integration ($\delta = 0$).

1.1 Preliminaries

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

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

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

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

If f is μ-strongly convex and L-smooth, then the condition number of f is $\kappa = L/\mu$. When f is a twice differentiable function, f is μ-strongly convex if and only if $\nabla^2 f(x) \succeq \mu I$ for all $x \in \mathbb{R}^d$, where $\nabla^2 f(x)$ denotes the Hessian matrix of f at x and I is the identity matrix; similarly, f is L-smooth if and only if $\nabla^2 f(x) \preceq LI$ for all $x \in \mathbb{R}^d$.

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 $\mathbb{R}^d$ with finite p'th moments where $p \geq 1$. The p-Wasserstein distance between ν_1 and ν_2 is defined as

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

where $\mathcal{C}(\nu_1, \nu_2)$ is the set of all couplings of ν_1 and ν_2. The Ricci curvature of P is defined as

$$\mathrm{Ric}(P) = 1 - \sup_{x,y \in \mathbb{R}^d, x \neq y} \frac{W_1(P(x,\cdot), P(y,\cdot))}{\|x - y\|}.$$

It is known that $\gamma(P) \geq \mathrm{Ric}(P)$ (see [22, Proposition 29]), which provides an effective way to bound the spectral gap through coupling arguments. We refer to [22] for more backgrounds on the Ricci curvature.

1.2 Related work

Various versions of Langevin dynamics have been studied in many recent papers, see [8, 9, 29, 23, 10, 7, 6, 3, 12, 27, 26, 17]. The convergence rate of HMC has also been studied recently in [14, 15, 18, 19, 20, 24]. The first bound in our setting was obtained by Mangoubi and Smith [18], who gave an $O(\kappa^2)$ bound on the convergence rate of ideal HMC.

Theorem 1.1 ([18, 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 [14], which showed a bound of $O(\kappa^{1.5})$. They also gave an algorithm with nearly optimal running time for solving the ODE that arises in the implementation of HMC.

Theorem 1.2 ([14, 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. It is mentioned in [18] that linear dependency is the best one can expect. Meanwhile, [14] shows that there *exists* a choice of step-sizes (time for running the ODE), which depends on f, the current position, and the random initial velocity, that achieves a linear convergence 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, [20] 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, 16]. The Riemannian version of HMC provides a faster polynomial-time algorithm for uniformly sampling polytopes [15]. However, the dependence on the dimension is significantly higher for these algorithms, both for the contraction rate and the time per step.

After the conference version [5] of our paper, several articles appeared on sampling from logconcave densities. For example, [2, 11, 21, 25, 28] study the Langevin dynamics and its variants, and [4, 13] study Metropolized HMC. Shen and Lee [25] propose the Randomized Midpoint Method to discretize the underdamped Langevin diffusion; their algorithm achieves smaller number of total gradient calls $N = \tilde{O}(\kappa^{7/6} d^{1/6} \varepsilon^{-1/3} + \kappa d^{1/3} \varepsilon^{-2/3})$ and less running time compared to our HMC method.

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 1.3. *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 $\propto \mathrm{e}^{-f}$ 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 strong convexity and smoothness. (In particular, we do not require that f is twice differentiable, which is assumed in both [14] and [18].)

We also establish a matching lower bound on the relaxation time of ideal HMC, implying the tightness of Theorem 1.3.

Theorem 1.4. *For any $L \geq \mu > 0$ and $T > 0$, there exists a μ-strongly convex and L-smooth function f, such that the relaxation time of ideal HMC for sampling from the density $\propto \mathrm{e}^{-f}$ with step-size T is $\Omega(\kappa)$, where $\kappa = L/\mu$ is the condition number.*

Using the ODE solver from [14] with nearly optimal running time, we obtain the following convergence rate in 2-Wasserstein distance for the HMC algorithm.

Theorem 1.5. *Let $\pi \propto \mathrm{e}^{-f}$ 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*

$$W_2(\pi_{\text{HMC}}, \pi) \leq \frac{\varepsilon}{\sqrt{\mu}}. \tag{1.4}$$

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.

We note that since our new convergence rate allows larger steps, the ODE solver is run for a longer time step. One interesting open problem is to find a matching lower bound for HMC with respect to the 2-Wasserstein distance.

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
[18]	κ^2	$\kappa^{6.5} d^{0.5}$	$\kappa^{6.5} d^{1.5}$
[14]	$\kappa^{1.5}$	$\kappa^{1.75} d^{0.5}$	$\kappa^{1.75} d^{1.5}$
this paper	κ	$\kappa^{1.5} d^{0.5}$	$\kappa^{1.5} d^{1.5}$

Finally, we remark that Theorem 1.4 also implies an $\Omega(\kappa)$ computational lower bound for HMC in terms of the number of gradient queries. This differs from [25] by a factor of $\tilde{O}(\kappa^{1/6})$. As pointed out by one of the referees, a comparison with optimization literature seems to suggest that the optimal number of gradient queries that can be achieved by any gradient-based MCMC algorithm should scale as $\sqrt{\kappa}$. This corroborates with the fine analysis in [2] that the convergence rate of continuous underdamped Langevin dynamics is $O(\sqrt{m})$ where m is the Poincaré constant of the measure, suggesting that an ideal discretization of underdamped Langevin dynamics with step-size $O(\sqrt{1/L})$ might be able to achieve $O(\sqrt{\kappa})$ convergence. In this sense, HMC probably cannot achieve the optimal convergence.

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 1.3. 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 2.1 (Contraction bound)**.** *Suppose that f is μ-strongly convex and L-smooth. Let $x(t)$ and $y(t)$ be the solution to (1.2) 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*

$$\|x(t) - y(t)\|^2 \le \left(1 - \frac{\mu}{4}t^2\right) \|x(0) - y(0)\|^2 . \tag{2.1}$$

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

$$\|x(T) - y(T)\|^2 \le \left(1 - \frac{1}{4c^2\kappa}\right) \|x(0) - y(0)\|^2 \tag{2.2}$$

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 $u(0) = v(0)$. We will show that

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

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

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

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\|x(t) - y(t)\|^2\right) = \langle x'(t) - y'(t), x(t) - y(t)\rangle = \langle u(t) - v(t), x(t) - y(t)\rangle . \tag{2.3}$$

Taking derivative on both sides, we get

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(\frac{1}{2}\|x(t) - y(t)\|^2\right) &= \langle u'(t) - v'(t), x(t) - y(t)\rangle + \langle u(t) - v(t), x'(t) - y'(t)\rangle \\ &= -\langle \nabla f(x(t)) - \nabla f(y(t)), x(t) - y(t)\rangle + \|u(t) - v(t)\|^2 \\ &= -\rho(t)\|x(t) - y(t)\|^2 + \|u(t) - v(t)\|^2 , \end{aligned} \tag{2.4}$$

where we define

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

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

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

Claim 2.2 (Crude bound). *For all* $0 \le t \le 1/(2\sqrt{L})$ *we have*

$$\frac{1}{2}\|x(0) - y(0)\|^2 \le \|x(t) - y(t)\|^2 \le 2\|x(0) - y(0)\|^2. \tag{2.5}$$

The proof of this claim is postponed to Section 2.1.

Next we derive an upper bound on $\|u(t) - v(t)\|^2$. By taking the derivative of $\frac{1}{2}\|u(t) - v(t)\|^2$, we get

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

Thus, its absolute value is not greater than

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

Since f is L-smooth and convex, it holds that

$$\|\nabla f(x(t)) - \nabla f(y(t))\|^2 \le L \langle \nabla f(x(t)) - \nabla f(y(t)), x(t) - y(t) \rangle;$$

see, e. g., [30, Lemma 4] for a proof. Hence, we obtain

$$\|\nabla f(x(t)) - \nabla f(y(t))\|^2 \le L\rho(t)\|x(t) - y(t)\|^2 \le 2L\rho(t)\|x(0) - y(0)\|^2,$$

where the last inequality follows from the crude bound (2.5). Then, using the fact that $u(0) = v(0)$ and the Cauchy–Schwarz inequality, we can upper bound $\|u(t) - v(t)\|^2$ by

$$\begin{aligned} \|u(t) - v(t)\|^2 &\le \left(\int_0^t \left| \frac{\mathrm{d}}{\mathrm{d}s} \|u(s) - v(s)\| \right| \mathrm{d}s \right)^2 \\ &\le \left(\int_0^t \sqrt{2L\rho(s)} \|x(0) - y(0)\| \,\mathrm{d}s \right)^2 \\ &\le 2Lt \left(\int_0^t \rho(s)\,\mathrm{d}s \right) \|x(0) - y(0)\|^2. \end{aligned}$$

Define the function

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

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,

$$\|u(t) - v(t)\|^2 \le 2LtP(t)\, \|x(0) - y(0)\|^2 . \tag{2.6}$$

Plugging (2.5) and (2.6) into (2.4), we deduce that

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(\frac{1}{2}\|x(t) - y(t)\|^2\right) \le -\rho(t)\left(\frac{1}{2}\|x(0) - y(0)\|^2\right) + 2LtP(t)\,\|x(0) - y(0)\|^2 .$$

If we define

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

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 1.3. Recall from Section 1.1 that the Ricci curvature of a Markov chain P is defined to be $1 - \sup_{x \ne y} W_1(P(x,\cdot), P(y,\cdot))/\|x - y\|$ where W_1 is the 1-Wasserstein distance. Then, Lemma 2.1 immediately 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 $1/(8c^2\kappa)$. It follows from [22, Proposition 29] that the spectral gap of ideal HMC is at least $1/(8c^2\kappa)$. Hence, the relaxation time is at most $8c^2\kappa = O(\kappa)$. □

2.1 Proof of Claim 2.2

We present the proof of Claim 2.2 in this section. We remark that a similar crude bound was established in [14] 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 2.2. We first derive a crude upper bound on $\|u(t) - v(t)\|$. Since f is L-smooth, we have

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

Then from $u(0) = v(0)$ we get

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

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

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\|x(t) - y(t)\|\right| = \frac{|\langle u(t) - v(t), x(t) - y(t)\rangle|}{\|x(t) - y(t)\|} \le \|u(t) - v(t)\| \le L\int_0^t \|x(s) - y(s)\|\,\mathrm{d}s. \tag{2.7}$$

Therefore, we have

$$\begin{aligned}\|x(t) - y(t)\| &= \|x(0) - y(0)\| + \int_0^t \left(\frac{\mathrm{d}}{\mathrm{d}s}\|x(s) - y(s)\|\right)\mathrm{d}s \\ &\le \|x(0) - y(0)\| + L\int_0^t \int_0^s \|x(r) - y(r)\|\,\mathrm{d}r\mathrm{d}s \\ &= \|x(0) - y(0)\| + L\int_0^t (t-s)\|x(s) - y(s)\|\,\mathrm{d}s.\end{aligned}$$

Then, applying a Grönwall-type inequality from [14, Lemma A.5], we deduce that

$$\|x(t) - y(t)\| \le \|x(0) - y(0)\|\cosh\left(\sqrt{L}t\right) \le \sqrt{2}\,\|x(0) - y(0)\|, \tag{2.8}$$

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

Next, we deduce from (2.7) and (2.8) that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|x(t) - y(t)\| &\ge -L\int_0^t \|x(s) - y(s)\|\,\mathrm{d}s \\ &\ge -L\|x(0) - y(0)\|\int_0^t \cosh\left(\sqrt{L}s\right)\mathrm{d}s \\ &= -\sqrt{L}\,\|x(0) - y(0)\|\sinh\left(\sqrt{L}t\right).\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
\|x(t)-y(t)\| &= \|x(0)-y(0)\| + \int_0^t \left(\frac{\mathrm{d}}{\mathrm{d}s}\|x(s)-y(s)\|\right)\mathrm{d}s \\
&\geq \|x(0)-y(0)\| - \sqrt{L}\,\|x(0)-y(0)\| \int_0^t \sinh\left(\sqrt{L}s\right)\mathrm{d}s \\
&= \|x(0)-y(0)\|\left(2-\cosh\left(\sqrt{L}t\right)\right) \geq \frac{1}{\sqrt{2}}\|x(0)-y(0)\|,
\end{aligned}$$

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

3 Lower bound for ideal HMC

In this section, we show that the relaxation time of ideal HMC can be $\Theta(\kappa)$ for some μ-strongly convex and L-smooth function for any step-size T, and thus prove Theorem 1.4.

Consider the following one-dimensional quadratic function:

$$f(x) = \frac{x^2}{2\sigma^2}, \tag{3.1}$$

where $1/\sqrt{L} \leq \sigma \leq 1/\sqrt{\mu}$. Thus, f is μ-strongly convex and L-smooth. The probability density ν proportional to e^{-f} is the Gaussian distribution: for $x \in \mathbb{R}$,

$$\nu(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{3.2}$$

The following lemma shows that, for certain choice of σ, ideal HMC for the Gaussian distribution ν can have relaxation time $\Omega(\kappa)$. Theorem 1.4 then follows immediately.

Lemma 3.1. *For any $T > 0$ there exists $1/\sqrt{L} \leq \sigma \leq 1/\sqrt{\mu}$ such that the relaxation time of ideal HMC for sampling from ν with step-size T is at least $\kappa/(8\pi^2)$.*

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

$$x''(t) = -f'(x(t)) = -\frac{x(t)}{\sigma^2}$$

with initial position $x(0) \in \mathbb{R}$ and initial velocity $x'(0)$ drawn from the standard Gaussian $N(0,1)$. Solving the ODE and plugging in the step-size $t = T$, we get

$$x(T) = x(0)\cos(T/\sigma) + v(0)\sigma\sin(T/\sigma).$$

Consider first the case that $T > 4\pi/\sqrt{L}$. We may assume that $L \geq 4\mu$ (i. e., $\kappa \geq 4$), since otherwise the lemma trivially holds. Pick a positive integer k such that

$$\frac{T\sqrt{\mu}}{2\pi} \leq k \leq \frac{T\sqrt{L}}{2\pi}.$$

Note that such k exists as

$$\frac{T\sqrt{L}}{2\pi} - \frac{T\sqrt{\mu}}{2\pi} \geq \frac{T\sqrt{L}}{4\pi} > 1.$$

We then choose $\sigma = T/(2k\pi)$, which satisfies $1/\sqrt{L} \leq \sigma \leq 1/\sqrt{\mu}$. Observe that for such a choice HMC becomes degenerate: we have $x(T) = x(0)$, meaning that the chain does not move at all. Thus, in this extremal case the relaxation time becomes infinity and the lemma holds.

Next we consider the case that $T \leq 4\pi/\sqrt{L}$. We shall pick $\sigma = 1/\sqrt{\mu}$. Let P denote the transition kernel of ideal HMC. Then for $x, y \in \mathbb{R}$ we have

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

Namely, given the current position x, the next position y is from a normal distribution with mean $x\cos(T/\sigma)$ and variance $\sigma^2 \sin^2(T/\sigma)$. Denote the spectral gap of P by γ. Let $h(x) = x$ and note that $h \in L_2^0(\nu)$. Using the properties of spectral gaps, we deduce that

$$\gamma = 1 - \sup_{f \in L_2^0(\nu)} \frac{|\langle Pf, f\rangle|}{\|f\|^2} \leq 1 - \frac{|\langle Ph, h\rangle|}{\|h\|^2}.$$

Since we have

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

and

$$\langle Ph, h\rangle = \int_{-\infty}^{\infty} \nu(x)P(x,y)h(x)h(y)\mathrm{d}x\mathrm{d}y = \sigma^2 \cos(T/\sigma),$$

we then get

$$\gamma \leq 1 - |\cos(T/\sigma)| \leq \frac{T^2}{2\sigma^2} \leq 8\pi^2 \cdot \frac{\mu}{L}$$

where the last inequality is due to our assumption $T \leq 4\pi/\sqrt{L}$. Hence, $\tau_{\text{rel}} = 1/\gamma \geq \kappa/(8\pi^2)$. This completes the proof of the lemma. □

4 Convergence rate of discretized HMC

In this section, we show how our improved contraction bound (Lemma 2.1) implies that HMC returns a good enough sample after $\tilde{O}((\kappa d)^{1.5})$ steps. We will use the framework from [14] to establish Theorem 1.5.

We first state the ODE solver from [14], 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 [14] for general kth order ODEs.

Theorem 4.1 ([14, Theorem 2.5]). *Let $x(t)$ be the solution to the ODE*

$$x''(t) = -\nabla f(x(t)), \quad x(0) = x_0, \quad x'(0) = v_0. \tag{4.1}$$

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*

$$\|q(t) - x''(t)\| \le \frac{\delta}{T^2}; \tag{4.2}$$

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 $x, y \in \mathbb{R}^d$,*

$$\|\nabla f(x) - \nabla f(y)\| \le L \|x - y\|. \tag{4.3}$$

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

$$\|\tilde{x}(t) - x(t)\| \le O(\delta). \tag{4.4}$$

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

$$C = O\left(\|v_0\| + T\|\nabla f(x_0)\|\right). \tag{4.5}$$

The following lemma, which combines Theorem 3.2, Lemma 4.1 and Lemma 4.2 from [14], establishes the conditions of Theorem 4.1 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 [14], holds in fact for the whole region $T \le 1/(2\sqrt{L})$ where the contraction bound (Lemma 2.1) is true. We omit these proofs here and refer the readers to [14] for more details.

Lemma 4.2. *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*

$$N = O\left(\frac{\log(d/\varepsilon)}{\mu T^2}\right) = O\left(\kappa \log(d/\varepsilon)\right) \tag{4.6}$$

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

1. ([14, Theorem 3.2]) *We have that*

$$W_2(\pi_{\text{HMC}}, \pi) \le \frac{\varepsilon}{\sqrt{\mu}}; \tag{4.7}$$

2. ([14, Lemma 4.1]) *For each k, let $x_k(t)$ be the solution to the ODE (1.3) in the kth step of HMC. Then there is a piece-wise constant function q_k of m_k pieces such that $\left\|q_k(t) - x_k''(t)\right\| \le \delta/T^2$ for all $0 \le t \le T$, where*

$$m_k = \frac{2LT^3}{\delta}\left(\left\|v_0^{(k-1)}\right\| + T\left\|\nabla f(x^{(k-1)})\right\|\right); \tag{4.8}$$

3. ([14, Lemma 4.2]) *We have that*

$$\frac{1}{N}\,\mathbb{E}\left[\sum_{k=1}^{N}\left\|\nabla f(x^{(k-1)})\right\|^2\right] \le O(Ld). \tag{4.9}$$

Proof of Theorem 1.5. The convergence of HMC is guaranteed by part 1 of Lemma 4.2. In the kth step, the number of evaluations of ∇f is $O(m_k \log(C_k\sqrt{\kappa}/\varepsilon))$ by Theorem 4.1 and part 2 of Lemma 4.2, 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 not greater than

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

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 4.2, 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

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

where the second inequality is due to Jensen's inequality and the fact that $\log^2(x)$ is concave when $x \geq 3$ (note that we can assume $M \geq 3$ by making $N \geq 3$). Therefore, the number of evaluations of ∇f per step, amortized over all steps, is

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

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

Acknowledgments. We are grateful to Yin Tat Lee and Andre Wibisono for helpful discussions. We also thank the anonymous referees for their valuable comments and suggestions.

References

[1] DAVID APPLEGATE AND RAVI KANNAN: Sampling and integration of near log-concave functions. In *Proc. 23rd STOC*, pp. 156–163. ACM Press, 1991. [doi:10.1145/103418.103439] 5

[2] YU CAO, JIANFENG LU, AND LIHAN WANG: On explicit L^2-convergence rate estimate for underdamped Langevin dynamics, 2019. [arXiv:1908.04746] 5, 6

[3] 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 *Proc. 35th Internat. Conf. Artificial Intelligence and Statistics (ICML'18)*, pp. 80:764–773. MLR Press, 2018. PMLR. [arXiv:1802.05431] 5

[4] YUANSI CHEN, RAAZ DWIVEDI, MARTIN J. WAINWRIGHT, AND BIN YU: Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *JMLR*, 21(92):1–72, 2019. JMLR. [arXiv:1905.12247] 5

[5] ZONGCHEN CHEN AND SANTOSH S. VEMPALA: Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions. In *Proc. 23rd Internat. Conf. on Randomization and Computation (RANDOM'19)*, pp. 64:1–12. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019. [doi:10.4230/LIPIcs.APPROX-RANDOM.2019.64, arXiv:1905.02313] 1, 5

[6] 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, 2018. [arXiv:1805.01648] 5

[7] XIANG CHENG, NILADRI S. CHATTERJI, PETER L. BARTLETT, AND MICHAEL I. JORDAN: Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proc. 31st Ann. Conf. on Learning Theory (COLT'18)*, pp. 75:1–24. MLR Press, 2018. PMLR. [arXiv:1707.03663] 5

[8] ARNAK S. DALALYAN: Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. Royal Statistical Soc. Ser. B*, 79(3):651–676, 2017. [doi:10.1111/rssb.12183, arXiv:1412.7392] 5

[9] ARNAK S. DALALYAN AND AVETIK KARAGULYAN: User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stoch. Proc. Appl.*, 129(12):5278–5311, 2019. [doi:10.1016/j.spa.2019.02.016, arXiv:1710.00095] 5

[10] ARNAK S. DALALYAN AND LIONEL RIOU-DURAND: On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020. [doi:10.3150/19-BEJ1178, arXiv:1807.09382] 5

[11] ARNAK S. DALALYAN, LIONEL RIOU-DURAND, AND AVETIK KARAGULYAN: Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets, 2019. [arXiv:1906.08530] 5

[12] RAAZ DWIVEDI, YUANSI CHEN, MARTIN J. WAINWRIGHT, AND BIN YU: Log-concave sampling: Metropolis-Hastings algorithms are fast! *JMLR*, 20(183):1–42, 2019. JMLR, Preliminary version in COLT'18. [arXiv:1801.02309] 5

[13] YIN TAT LEE, RUOQI SHEN, AND KEVIN TIAN: Logsmooth gradient concentration and tighter runtimes for Metropolized Hamiltonian Monte Carlo. In *Proc. 33rd Ann. Conf. on Learning Theory (COLT'20)*, pp. 125:2565–2597. MLR Press, 2020. PMLR. [arXiv:2002.04121] 5

[14] YIN TAT LEE, ZHAO SONG, AND SANTOSH S. VEMPALA: Algorithmic theory of ODEs and sampling from well-conditioned logconcave densities, 2018. [arXiv:1812.06243] 2, 5, 6, 10, 12, 13, 14

[15] YIN TAT LEE AND SANTOSH S. VEMPALA: Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. In *Proc. 50th STOC*, pp. 1115–1121. ACM Press, 2018. [doi:10.1145/3188745.3188774, arXiv:1710.06261] 2, 5

[16] LÁSZLÓ LOVÁSZ AND SANTOSH S. VEMPALA: Fast algorithms for logconcave functions: Sampling, rounding, integration and optimization. In *Proc. 47th FOCS*, pp. 57–68. IEEE Comp. Soc., 2006. [doi:10.1109/FOCS.2006.28] 5

[17] YI-AN MA, NILADRI CHATTERJI, XIANG CHENG, NICOLAS FLAMMARION, PETER BARTLETT, AND MICHAEL I. JORDAN: Is there an analog of Nesterov acceleration for MCMC?, 2019. [arXiv:1902.00996] 5

[18] OREN MANGOUBI AND AARON SMITH: Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions 1: Continuous dynamics, 2017. [arXiv:1708.07114] 2, 5, 6

[19] OREN MANGOUBI AND AARON SMITH: Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions 2: Numerical integrators. In *Proc. 22nd Internat. Conf. Artificial Intelligence and Statistics (AISTATS'19)*, pp. 89:586–595. MLR Press, 2019. PMLR. 2, 5

[20] OREN MANGOUBI AND NISHEETH VISHNOI: Dimensionally tight bounds for second-order Hamiltonian Monte Carlo. In *Proc. 31st Adv. Neural Info. Proc. Sys. (NeurIPS'18)*, pp. 6027–6037. Curran Assoc., 2018. NeurIPS. 2, 5

[21] WENLONG MOU, YI-AN MA, MARTIN J. WAINWRIGHT, PETER L. BARTLETT, AND MICHAEL I. JORDAN: High-order Langevin diffusion yields an accelerated MCMC algorithm. *JMLR*, 22(42):1–41, 2021. JMLR. [arXiv:1908.10859] 5

[22] YANN OLLIVIER: Ricci curvature of Markov chains on metric spaces. *J. Functional Anal.*, 256(3):810–864, 2009. [doi:10.1016/j.jfa.2008.11.001, arXiv:math/0701886] 4, 9

[23] MAXIM RAGINSKY, ALEXANDER RAKHLIN, AND MATUS TELGARSKY: Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In *Proc. 30th Ann. Conf. on Learning Theory (COLT'17)*, pp. 65:1–30. MLR Press, 2017. PMLR. [arXiv:1702.03849] 5

[24] CHRISTOF SEILER, SIMON RUBINSTEIN-SALZEDO, AND SUSAN HOLMES: Positive curvature and Hamiltonian Monte Carlo. In *Proc. 27th Adv. Neural Info. Proc. Sys. (NIPS'14)*, pp. 586–594. Curran Assoc., 2014. NIPS. 5

[25] RUOQI SHEN AND YIN TAT LEE: The Randomized Midpoint Method for log-concave sampling. In *Proc. 32nd Adv. Neural Info. Proc. Sys. (NeurIPS'19)*, pp. 2100–2111. Curran Assoc., 2019. NeurIPS. [arXiv:1909.05503] 5, 6

[26] SANTOSH S. VEMPALA AND ANDRE WIBISONO: Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Proc. 32nd Adv. Neural Info. Proc. Sys. (NeurIPS'19)*, pp. 8094–8106. Curran Assoc., 2019. NeurIPS. [arXiv:1903.08568] 5

[27] ANDRE WIBISONO: Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Proc. 31st Ann. Conf. on Learning Theory (COLT'18)*, pp. 75:2093–3027. MLR Press, 2018. PMLR. [arXiv:1802.08089] 5

[28] ANDRE WIBISONO: Proximal Langevin algorithm: Rapid convergence under isoperimetry, 2019. [arXiv:1911.01469] 5

[29] YUCHEN ZHANG, PERCY S. LIANG, AND MOSES CHARIKAR: A hitting time analysis of stochastic gradient Langevin dynamics. In *Proc. 30th Ann. Conf. on Learning Theory (COLT'17)*, pp. 65:1980–2022. MLR Press, 2017. PMLR. [arXiv:1702.05575] 5

[30] XINGYU ZHOU: On the Fenchel Duality between strong convexity and Lipschitz Continuous Gradient, 2018. SLIDES. [arXiv:1803.06573] 8

AUTHORS

Zongchen Chen
Instructor
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
zongchen@mit.edu
https://sites.google.com/view/zongchenchen

Santosh S. Vempala
Professor
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
vempala@gatech.edu
https://www.cc.gatech.edu/~vempala/

ABOUT THE AUTHORS

ZONGCHEN CHEN is an instructor (postdoc) in Mathematics at the Massachusetts Institute of Technology. He obtained his Ph. D. in Algorithms, Combinatorics and Optimization at the Georgia Institute of Technology in 2021, advised by Eric Vigoda. Before that, he received his B. S. in Mathematics and Applied Mathematics from the Zhiyuan College at the Shanghai Jiao Tong University in 2016. He has broad interests in randomized algorithms, discrete probability, and combinatorics. Currently, his research focuses on Markov chain Monte Carlo (MCMC) methods for sampling from Gibbs distributions, and machine learning problems related to undirected graphical models.

SANTOSH VEMPALA is the Frederick Storey professor in the College of Computing and former director of the Algorithms and Randomness Center at Georgia Tech. His research interests, ironically, are in algorithms, randomness, and geometry. He graduated from CMU in 1997 following the advice of Avrim Blum and was at MIT till 2006 except for a year as a Miller fellow at UC Berkeley. He gets unreasonably excited when a phenomenon that appears complex from one perspective turns out to be simple from another. At Georgia Tech, he started the Computing-for-Good (C4G) program.