The Annals of Applied Probability
2021, Vol. 31, No. 5, 2019–2045
https://doi.org/10.1214/20-AAP1640

MIXING OF HAMILTONIAN MONTE CARLO ON STRONGLY LOG-CONCAVE DISTRIBUTIONS: CONTINUOUS DYNAMICS

BY OREN MANGOUBI[1] AND AARON SMITH[2]

[1] *Worcester Polytechnic Institute, omangoubi@gmail.com*
[2] *Department of Mathematics and Statistics, University of Ottawa, smith.aaron.matthew@gmail.com*

We obtain several quantitative bounds on the mixing properties of an "ideal" Hamiltonian Monte Carlo (HMC) Markov chain for a strongly log-concave target distribution π on $\mathbb{R}^d$. Our main result says that the HMC Markov chain generates a sample with Wasserstein error ϵ in roughly $O(\kappa^2 \log(\frac{1}{\epsilon}))$ steps, where the condition number $\kappa = \frac{M_2}{m_2}$ is the ratio of the maximum M_2 and minimum m_2 eigenvalues of the Hessian of $-\log(\pi)$. In particular, this mixing bound does not depend explicitly on the dimension d. These results significantly extend and improve previous quantitative bounds on the mixing of ideal HMC, and can be used to analyze more realistic HMC algorithms. The main ingredient of our argument is a proof that initially "parallel" Hamiltonian trajectories contract over much longer steps than would be predicted by previous heuristics based on the Jacobi manifold.

1. Introduction. Markov chain Monte Carlo (MCMC) methods are ubiquitous in Bayesian statistics and other areas, and Hamiltonian Monte Carlo (HMC) Markov chains are some the most widely-used MCMC methods [13, 23, 45]. In this paper, we improve upon the best earlier[1] analysis of the following "ideal" version of the HMC Markov chain for sampling from a distribution π on $\mathbb{R}^d$ with density $q(x) \propto e^{-U(x)}$:

Markov Chain 1 Ideal HMC with integration time T, initial point X_0

```
for $i = 1, 2, \ldots$ do
    Sample $\mathbf{p}_i$ according to the standard Gaussian.
    Set $X_{i+1} = Q_T^{X_i}(\mathbf{p}_i)$, as in equation (2.3) below.
end for
```

In this "ideal" version of HMC, the map Q_T^x defined in equation (2.3) below is the exact solution to a system of ordinary differential equation (ODE) called Hamilton's equations. In essentially all realistic situations, the solution Q_T^x must be approximated using for example, the leapfrog integrator or other numerical integration methods. It turns out to be the case that a careful analysis of Markov chain 1 can be used as a key ingredient in the analysis of numerical Markov chain algorithms. We use the present paper this way in our companion paper [42] and followup [43].

Received November 2019; revised July 2020.
MSC2020 subject classifications. Primary 60J05; secondary 65C40, 60J20, 68W20.
Key words and phrases. Markov chain Monte Carlo (MCMC), Hamiltonian Monte Carlo (HMC), momentum, strong convexity.

[1] Our initial arXiv paper [40] combined the results in this paper, our companion [42], and some other results. Since [40] was released, many authors have produced strong analyses of HMC. We defer discussion of these developments, and most related work, to Section 3.

Our main contribution is an analysis of this Markov chain under the following assumption:

ASSUMPTION 1.1. There exist $0 < m_2, M_2 < \infty$ and set $\mathcal{X} \subset \mathbb{R}^d$ so that

$$m_2\mathrm{Id} \preceq \nabla^2 U(x) \preceq M_2\mathrm{Id}$$

for all $x \in \mathcal{X}$, where $\preceq$ is the usual Loewner order on matrices and Id is the d-dimensional identity matrix.

A C^2-smooth log-density U satisfies Assumption 1.1 if and only if it is m_2-strongly convex and has M_2-Lipschitz gradient. Let $\kappa = \frac{M_2}{m_2}$, the *condition number* of U. Under Assumption 1.1 with $\mathcal{X} = \mathbb{R}^d$, our main result says that the Wasserstein mixing time of Markov chain 1 is $O(\kappa^2)$ for appropriate choice of $T = \frac{1}{2\sqrt{2}}\frac{\sqrt{m_2}}{M_2}$ (see Theorem 1 below for details).

Our result improves on the previous best bounds in [53] in two significant ways. They make the stronger assumption that π is *exactly* Gaussian (which we replace with the strictly-weaker Assumption 1.1), and they obtain the weaker conclusion that the Wasserstein mixing time is $O(d^2c(m_2, M_2))$ for some nonexplicit function c (while we remove entirely the dependence on dimension and give quantitative bounds on $c(m_2, M_2)$). We discuss the practical significance of these changes below.

Like [53], our proof strategy is to analyze an explicit coupling of two copies of the same Markov chain, showing that these two copies get closer (on average). Both our work and [53] rely on the observation that initially-parallel solutions to Hamilton's equations get closer for a short time under Assumption 1.1. More precisely, in our paper we show that

$$\|Q_T^x(p) - Q_T^y(p)\| \leq \left(1 - \frac{1}{4}m_2T^2\right)\|x - y\| \tag{1.1}$$

for all $x, y, p \in \mathbb{R}^d$ and all $T = T(x, y, p)$ sufficiently small. We obtain stronger results than [53] primarily because we give a more careful analysis of Hamilton's equations, allowing us to verify that something like inequality (1.1) holds for much larger values of T. See Section 3.1 for a more detailed discussion.

Another large difference between our paper and [53] is that our Assumption 1.1 is much weaker than the assumption made in [53], which says that the logdensity is a Gaussian distribution. As one example application with non-Gaussian log-density, we consider the problem of Bayesian logistic "ridge" regression (also discussed in Section 5 of our companion paper [42]). Here one wishes to sample from the log-density

$$U(x) = \frac{1}{2}x^\top \Sigma x - \sum_{i=1}^{r} \mathsf{Y}_i \log\bigl(\sigma(x^\top \mathsf{X}_i)\bigr) + (1 - \mathsf{Y}_i)\log\bigl(\sigma(-x^\top \mathsf{X}_i)\bigr), \tag{1.2}$$

where $\mathsf{X}_1, \dots, \mathsf{X}_r$ are the independent variable data vectors, $\mathsf{Y}_1, \dots, \mathsf{Y}_r \in \{0, 1\}$ are the binary dependent variable data labels, $\sigma(s) \equiv \frac{1}{e^{-s}+1}$ is the sigmoid function, and $\frac{1}{2}x^\top \Sigma x$ is the Gaussian log-prior with positive definite matrix Σ. This log-density satisfies our Assumption 1.1 with $m_2 = \lambda_{\min}(\Sigma^{-1})$ and $M_2 = \lambda_{\max}(\Sigma^{-1} + \sum_{i=1}^{r} \mathsf{X}_i\mathsf{X}_i^\top)$. We note that even though the ridge regression prior is Gaussian, the posterior density $\propto e^{-U(x)}$ we want to sample from is non-Gaussian because of the sigmoid function terms in equation (1.2).

Moreover, even if one replaces the Gaussian prior in equation (1.2) with a non-Gaussian prior, oftentimes we still have, with high probability, that the posterior log-density for logistic regression is still strongly convex in the "bulk" of the distribution containing most of its probability measure. Roughly speaking, this is true if the data comes from a distribution whose covariance matrix has all its eigenvalues bounded away from zero (see, e.g., Lemma 6.2 of

[30] for details). In such settings where our assumptions are satisfied in the bulk of the distribution containing most of the probability measure, our results can oftentimes be extended using "gluing" arguments (see the last paragraph of this section for more discussion on how one can extend our result to such settings).

Readers familiar with HMC will immediately notice that there are many realistic situations where our results do not directly apply, due to the following two problems:

1. It is almost never possible to compute the solution $Q_T^x(p)$ exactly, and
2. Assumption 1.1 is oftentimes not satisfied on all of $\mathbb{R}^d$.

At first glance, this might make an analysis of our "ideal" Markov chain seem pointless. In fact, a careful analysis of the "ideal" version of HMC can be a critical step in obtaining good estimates in more realistic settings:

We study the first problem in our own followup papers [42, 43]. The basic idea is to show that commonly-used approximations of Q_T^x are small perturbation of the ideal dynamics in some realistic settings. In this small-perturbation regime, the real-world HMC algorithm inherits the good mixing properties of our "ideal" HMC Markov chain. This approach allows us to apply our main result (Theorem 1) to show that real-world HMC algorithms have computational complexity that scale with dimension like $O(d^{0.25})$, providing the first rigorous results that match the famous heuristic in [1].

The second problem has been studied in several papers, including our followup [39]. We briefly sketch one approach to the issue. Consider the situation that Assumption 1.1 holds only on a compact subset $\mathcal{X}$ of $\mathbb{R}^d$. In this setting, we can "glue together" three estimates: (i) our bounds for the behaviour of the chain on $\mathcal{X}$, (ii) other bounds for the behaviour on $\mathcal{X}^c$, and (iii) very rough estimates on how much time is spent in $\mathcal{X}$. Useful and generic "gluing" arguments are given in [22]. This approach allows one to use our main bounds as a critical step in analyzing models that don't satisfy Assumption 1.1 on all of $\mathbb{R}^d$; this includes for example, mixtures of Gaussians (see our followup [39]) and also, as mentioned earlier, logistic regression with non-Gaussian priors. We call attention to the analysis of HMC [31], which has a particularly strong method for allowing one to ignore small-measure subsets on which assumptions may fail.

Section 3 gives further discussion on how our results can be applied, as well as the existing literature on analysis of HMC and related Markov chains.

1.1. *Paper overview.*

In Section 2, we give notation and a precise statement of our main results. In Section 3, we discuss related work (including work that has appeared since the first version of this note) and give more detailed discussion of our proof techniques and companion papers. In Section 4 we state the main lemmas required for our main theorems and give proofs of the simplest and most important bounds (the proofs of the remaining technical bounds are deferred to the Appendix). In particular, Section 4 contains a proof of our main contraction result (Theorem 3). Section 5 contains a proof of a related drift bound (the proof of this latter result is deferred to the Appendix). Finally, Section 6 contains some open problems.

2. Main notation and results.

2.1. *Preliminary notation.*

For any function $f : \mathbb{R}^a \to \mathbb{R}$, we use the shorthand $f' := \nabla f$, and denote by $D_v f := \langle v, \nabla f \rangle$ the directional derivative in the direction v. For a vector-valued function $g = (g_1, \dots, g_b)^\top$, we define the coordinate-wise directional derivative $D_v g := (D_v g_1, \dots, D_v g_b)$.

Throughout the paper, we will consider a function U that satisfies Assumption 1.1. Recall that any function satisfying Assumption 1.1 with $\mathcal{X} = \mathbb{R}^d$ has a unique minimizer; we assume without loss of generality that this minimum occurs at 0 in order to simplify notation.

Throughout the paper, we make a few small abuses of notation. For any function $f : X \to Y$ 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*

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

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. For $k = \infty$, define

$$W_\infty(\nu_1, \nu_2) = \inf_{(X,Y)\in\mathcal{C}(\nu_1,\nu_2)} \inf\{B \in \mathbb{R} : \mathbb{P}(\|X - Y\| \le B) = 1\}.$$

Denote by K a reversible transition kernel on $\mathbb{R}^d$ with unique invariant measure π. Recall that K acts as an operator from $L^2(\pi)$ to $L^2(\pi)$, and the *absolute spectral gap* of such an operator is given by

$$1 - \sup\{|\lambda| : \lambda \in (-1, 1), (K - \lambda \mathrm{Id}) \text{ is not invertible}\}.$$

We define the *relaxation time* $\tau_{\mathrm{rel}}(K)$ of a transition kernel K to be the reciprocal of its spectral gap.

For measures ν_1, ν_2 on a measurable space $(\Omega, \mathcal{F})$, the *total variation distance* between ν_1, ν_2 is given by

$$\|\nu_1 - \nu_2\|_{\mathrm{TV}} = \sup_{A\in\mathcal{F}} (\nu_1(A) - \nu_2(A)).$$

Finally, denote by $B(x, r) := \{y \in \mathbb{R}^d : \|x - y\| \le r\}$ the Euclidean ball with center $x \in \mathbb{R}^d$ and radius $r > 0$.

2.1.2. *Big-O notation.* For two nonnegative functions or sequences f, g, we write $f = O(g)$ as shorthand for the statement: there exists a constant $0 < C < \infty$ so that for all $x_1, \dots, x_n$, we have $f(x_1, \dots, x_n) \le Cg(x_1, \dots, x_n)$. We write $f = \Omega(g)$ for $g = O(f)$, and 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: $\lim_{x_1,\dots,x_n\to\infty} \frac{f(x_1,\dots,x_n)}{g(x_1,\dots,x_n)} = 0$.

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 $\{q_t(\mathbf{q}, \mathbf{p})\}_{t\ge0}$, $\{p_t(\mathbf{q}, \mathbf{p})\}_{t\ge0}$ the solutions to Hamilton’s equations:

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

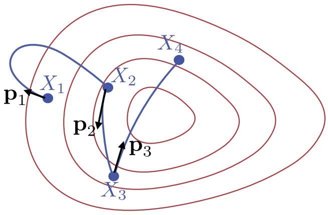

FIG. 1. *The Hamiltonian Monte Carlo Markov chain* $X_1, X_2, \dots$ *with momentum* $\mathbf{p}_1, \mathbf{p}_2, \dots$.

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 $\mathcal{Q}_T^{\mathbf{q}} : \mathbb{R}^d \to \mathbb{R}^d$ by

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

In the context of HMC, we refer to q_t as the *position* variable and p_t as the *momentum* variable.

In this paper we study Markov chain 1, the simplest Hamiltonian Monte Carlo Markov chain (see Figure 1).

Note that the sequence $\{X_i\}_{i \ge 0}$ appearing in the description of Markov chain 1 is a deterministic function of the initial value X_0 and the i.i.d. sequence $\{\mathbf{p}_i\}_{i \ge 0}$ of momentum updates. In the Markov chain literature, this fact is summarized by saying that this set of rules (given in the description of Markov chain 1) defines a *random mapping representation* of $\{X_i\}_{i \ge 0}$ with *update sequence* $\{\mathbf{p}_i\}_{i \ge 0}$ (see Chapter 1.2 of [35]). In particular, the fact that this gives a random mapping representation means that it is possible to define a coupling of two Markov chains evolving according to this set of rules by defining a coupling of the momentum updates. Finally, note that this set of rules also naturally defines the nonreversible Markov chains $\{(X_i, \mathbf{p}_i)\}_{i \ge 0}$ on the larger state space $\mathbb{R}^{2d}$.

2.2. *Main results.* For simplicity, all main results are stated for the largest integration time $T = \frac{1}{2\sqrt{2}} \frac{\sqrt{m_2}}{M_2}$ allowed by our proofs. We observe that these results can still be applied for any $0 < T \le \frac{1}{2\sqrt{2}} \frac{\sqrt{m_2}}{M_2}$, since a potential U that satisfies Assumption 1.1 for constants m_2, M_2 will also satisfy it for any pair m_2', M_2' satisfying $0 < m_2' < m_2 \le M_2 < M_2' < \infty$.

THEOREM 1 (Mixing for strongly log-concave targets). *Let K be the transition kernel defined by the ideal HMC Markov chain* (*Markov chain* 1) *with parameter $T = \frac{1}{2\sqrt{2}} \frac{\sqrt{m_2}}{M_2}$. Let Assumption* 1.1 *hold with $\mathcal{X} = \mathbb{R}^d$. Then K satisfies the contraction bound*

$$\sup_{x, y \in \mathbb{R}^d} \frac{W_k(K(x, \cdot), K(y, \cdot))}{\|x - y\|} \le 1 - \frac{1}{64}\left(\frac{m_2}{M_2}\right)^2 \quad \forall k \in \mathbb{N} \tag{2.4}$$

and the spectral bound

$$\tau_{\text{rel}}(K) \le 64\left(\frac{M_2}{m_2}\right)^2. \tag{2.5}$$

In particular, inequality (2.4) *implies that for all* $\epsilon > 0$ *there exists an* $\mathcal{I} = O((\frac{M_2}{m_2})^2 \log(\frac{D}{\epsilon}))$, *where* $D = W_k(\mathcal{L}(X_0), \pi)$, *such that*

$$W_k\big(\mathcal{L}(X_i), \pi\big) \le \epsilon \quad \forall i \ge \mathcal{I}, k \in \mathbb{N} \cup \{\infty\}. \tag{2.6}$$

PROOF. Inequality (2.4) follows immediately from Theorem 3, which we state and prove in Section 4. Inequality (2.5) follows immediately from inequality (2.4) and Proposition 30 of [47]. Inequality (2.6) follows from inequality (2.4) by a standard Markov chain coupling argument. □

We also give a bound for convergence of a slightly modified HMC Markov chain to the target distribution π in the total variation norm (Corollary 2). This bound is given for a slight modification of Markov chain 1, where one adds a very small random vector uniformly distributed on a very small ball to the final step of the Markov chain. [2]

COROLLARY 2. *Let* $Z_1, Z_2, \ldots$ *be independent random vectors uniformly distributed on the unit ball, and let* $\hat{X}_i = X_i + \frac{\epsilon\sqrt{m_2}}{144 M_2 \sqrt{d} \log^{\frac{1}{2}}(\frac{M_2}{(m_2\epsilon)})} Z_i$ *for every* i. *Then for all* $\epsilon > 0$ *there exists an* $\hat{\mathcal{I}} = O((\frac{M_2}{m_2})^2 \log(\frac{D \times d M_2}{\epsilon\sqrt{m_2}}))$, *where* $D = W_k(X_0, \pi)$, *such that*

$$\|\mathcal{L}(\hat{X}_i) - \pi\|_{\mathrm{TV}} \le \epsilon \quad \forall i \ge \hat{\mathcal{I}}.$$

The content of the following remarks (Remarks 2.1 and 2.2) are mentioned as some of the primary motivation for the papers [11, 32]. For this reason, we leave them here in their original forms, even though these later papers have made a great deal of progress on them (and [11] proves that our conjecture in the second remark is correct):

REMARK 2.1. 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. We give details in the arXiv version [40] and in the companion paper [42].

REMARK 2.2. We do not know if this dependence on the ratio $\frac{M_2}{m_2}$ is sharp, but we do know that the dependence cannot be removed entirely. In Section 5 of [39], we find a sequence $\{\pi_n\}_{n\in\mathbb{N}}$ of target distributions satisfying our assumptions for which the ratio $\{\frac{M_2(n)}{m_2(n)}\}_{n\in\mathbb{N}}$ goes to infinity and the associated relaxation times $\tau_{\mathrm{rel}}(n)$ satisfy $\tau_{\mathrm{rel}}(n) = \Omega(\frac{M_2}{m_2})$. We conjecture that this lower bound is sharp.

3. Related work and techniques.

3.1. *Discussion of coupling improvements.* Our main techniques in this paper are explicit comparisons of ODEs and probabilistic coupling bounds. To obtain these bounds, we use comparison theorems for ODEs to prove that initially "parallel" Hamiltonian trajectories contract for a relatively long time. It is well known and straightforward to check that, under a strong log-concavity assumption, all trajectories contract if they are *sufficiently short* (see inequality (4.3)). However, the final bounds on the mixing of the HMC Markov chain depend quite strongly on how long a trajectory can typically be before the contraction rate gets close to 0.

[2] In the arXiv version of our paper [40], we give a TV bound for the original Markov chain 1 with no modifications. However, that bound is weaker than the one in Corollary 2, and the proof is much longer.

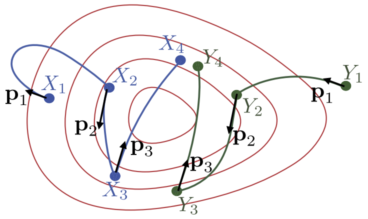

FIG. 2. *Coupling two copies $X_1, X_2, \ldots$ (blue) and $Y_1, Y_2, \ldots$ (green) of HMC by choosing the same momentum $\mathbf{p}_i$ at every step. Note that contraction is guaranteed at each step, despite the fact that the trajectory from X_1 to X_2 is not the shortest path between those two points. This is in contrast to typical comparison theorems from differential geometry, such as the Rauch comparison theorem, which require the assumption that paths are length-minimizing.*

By way of comparison, the previous work [53] was also based on an analysis of the contraction of HMC trajectories. The authors recall that Hamiltonian trajectories are exactly geodesics on an associated Jacobi manifold, then use the Rauch comparison theorem from differential geometry to show that the distance in the Jacobi metric between these geodesics contracts at least until one of the geodesics reaches a point that is "conjugate" to its initial point. In other words, [53] show that the trajectories contract until one of the trajectories no longer minimizes the distance between its initial point and the current point, if distance is measured in the Jacobi metric.

Since the Rauch comparison theorem is sharp, it is natural to guess that contraction only occurs until one of the trajectories reaches its conjugate point under the Jacobi metric. One of the main differences between the present paper and [53] is that we show that contraction persists for trajectories that are *vastly* longer than this heuristic would suggest in high dimensions. We show contraction up to a time $T = \Theta(\frac{\sqrt{m_2}}{M_2})$,,[3] which in high dimensions is much longer than the time for which contraction was proved to occur (with high probability) in [53]. [4]

This long-term contraction is illustrated in the first move shown in Figure 2, where the trajectory associated with X_1 does a "U-turn" (and in particular passes a conjugate point with respect to the Jacobi metric), but one can see clearly that contraction in the Euclidean distance between the two trajectories continues throughout the length of the trajectory.

3.2. *Relationship to companion paper and other works.* In our companion paper [42], we use the bounds on the "ideal" HMC Markov chain obtained in this paper to bound the computational costs of a numerical implementation of HMC. The contents of this paper and its companion paper are posted together on arXiv as a single long paper [40].

Our main result in the companion paper [42] uses the main result of this paper to show that a simple numerical implementation of HMC can approximately sample from the stationary distribution in a number of gradient evaluations that grows at rate $d^{\frac{1}{2}}$ in the dimension, if parameters M_2, m_2 do not grow with the dimension. For comparison, the best available mixing time bound for the unadjusted Langevin algorithm on strongly log-concave π grows roughly linearly in d, a much larger dependence on dimension than our bound for this numerical implementation of HMC [18, 19].

[3] Note that $T = \Theta(\frac{\sqrt{m_2}}{M_2})$ is a "long time" in the strongly log-concave setting when $\frac{\sqrt{m_2}}{\sqrt{M_2}}$ is small (the extra factor of $\frac{1}{\sqrt{M_2}}$ is a scaling factor which does not affect the mixing time bounds). However, if the target logdensity has a strong convexity constant m_2 that is very small in comparison to M_2, then T will be small as well.

[4] The contraction time in [53] is shorter by a factor of roughly d^{-1} than the contraction time we prove here, which results in a mixing time bound that is slower by a factor of roughly d^{-2}.

Under an additional separability assumption, the arXiv version [40] of our paper shows that an unadjusted numerical implementation of HMC based on a kth-order numerical integrator allows one to approximately sample the stationary distribution in a number of gradient evaluations that grows at rate of $d^{\frac{1}{2k}}$ in the dimension. Most implementations of HMC use second-order ($k = 2$) integrators, giving a dimension dependence of $d^{\frac{1}{4}}$, which matches the famous heuristic derived in [1].

Our bounds also compare favorably to the ball walk Markov chain, whose best available mixing time bound is roughly $O(d^2 \frac{M_2^2}{m_2^2} \log(\frac{1}{\epsilon}))$ [55] in the strongly log-concave setting where $\frac{M_2}{m_2}$ is small in comparison to the dimension d. On the other hand, in other settings which lack smoothness or strong convexity, or where $\frac{M_2}{m_2}$ is large in comparison to the dimension d, or where the support of the distribution is constrained to a convex body, the ball walk, and the closely related hit-and-run algorithm, oftentimes provide the best available bounds (see, e.g., [9, 34, 37, 38, 44] for bounds in various settings).

While discussing related literature in the remainder of this section, we include the main results of the companion paper [42] when referring to “the present work.” Since the present paper discusses only ideal HMC while the companion paper discusses only numerical implementation, we trust that this will not cause undue confusion. We also discuss related results for the Langevin and ball walk algorithms.

3.3. *Literature review.* 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 [16] 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., [17]) or on quantitative bounds on the running time of the algorithm as a function of the problem complexity (see, e.g., [3] or essentially any paper in the large Computer Science literature on the subject). Our work falls in the latter category.

Despite the popularity of HMC and the widespread belief that HMC outperforms other algorithms in high-dimensional statistical problems (see, e.g., [1]), its theoretical properties have not been as well-understood as some of its older cousins, especially the random walk Metropolis (RWM) algorithm with space-invariant proposal kernels (see [55] for a survey of results on the closely related ball walk algorithm) and to a lesser degree the Metropolis-adjusted Langevin algorithm (MALA) [5, 19, 50, 56]. 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.

Several recent papers have begun to bridge this gap, most notably by proving ergodicity [8] and geometric ergodicity [6, 36] of different versions of HMC under certain conditions, and also establishing some quantitative bounds on the rate of convergence for Gaussian target distributions [53]. A number of other papers, most prominently [1], have 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 [51] for a general discussion relating results similar to [1] 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 [36]).

Finally, in recent independent work [33], 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 [53] their bounds only apply to trajectories with very short step sizes and consequently do not imply the results in this paper.

Our work is most directly comparable to [53] (which studies the same Markov chain, but has quite different assumptions and conclusions) and [18] (which studies a different Markov chain, but has very similar assumptions and almost directly-comparable conclusions).

To our knowledge, [53] is the only previous paper giving quantitative nonasymptotic bounds on the mixing of HMC. In the present paper, 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, and proving convergence in stronger norms. In our companion paper [42], we apply these results to study numerical implementations of HMC, which are not studied in [53].

The results in this paper and [42] most closely resemble those of [18], which studies the nonasymptotic mixing properties of the Langevin algorithm on strongly log-concave distributions. Their main results hold under essentially the same conditions as our Theorem 1. For closely related work, see [14], which studies very similar conditions to [18]. Recent independent work [12] also gives quantitative bounds for a second-order "underdamped" version of the Langevin algorithm that improve on those of [18] for some target distributions. See [28] and [46] for why "standard" HMC is believed to be more efficient than second-order Langevin Monte Carlo.

Although our results are far from providing a complete understanding of HMC, the strongly log-concave distributions are an important special case. Many important posterior distributions in statistics are strongly log-concave, such as the "ridge regression" posterior associated with Gaussian priors for logistic regression which was mentioned in the Introduction (see the examples in [18, 19] for calculations of the associated constants). Other distributions are strongly log-concave except on a set of small stationary measure. In addition to this, we expect most MCMC Markov chains to perform well for strongly log-concave targets. For these reasons, the performance of many Monte Carlo Markov chains has been studied extensively in the strongly log-concave setting [18]. This has the added advantage of allowing us to give a sensible comparison of the performance of HMC to its competitors, such as the Langevin algorithm and the ball walk.

REMARK 3.1. Roughly, geometric ergodicity results [6, 36] state that for each target distribution π and starting point x there exits a number $\lambda_\pi(x) > 0$ such that, after $i = O(\lambda_\pi(x)\log(\frac{1}{\epsilon}))$ steps,

$$\|\mathcal{L}(X_i) - \pi\|_{\mathrm{TV}} \le \epsilon$$

for the HMC Markov chain X_i started at $X_0 = x$. However, [6, 36] do not give a quantitative bound on λ_π. In contrast, we give a quantitative bound of $O((\frac{M_2}{m_2})^2 \log(\frac{D}{\epsilon}))$ on the number of steps i until the distance $W_k(\mathcal{L}(X_i), \pi)$ (in the k-Wasserstein metric for any $k \in \mathbb{N}$) between the distribution of the Markov chain and the stationary distribution satisfies $W_k(\mathcal{L}(X_i), \pi) \le \epsilon$, where $D = W_k(X_0, \pi)$.

While they do not provide quantitative bounds on the "mixing time," the geometric ergodicity results of [6, 36] apply under more general assumptions than our results, including for certain nonconvex log-densities. We suspect that the best analyses of HMC will combine results such as ours (for sharp analysis of the behaviour of HMC on the "bulk" of the target) with results such as theirs (for coarser analysis of the behaviour in the "tails" of the target), gluing them together as in [22].

3.3.1. *Subsequent developments.* Since the first arXiv version of our paper, there have been a number of improvements to the analysis of HMC in the literature, some of which make use of, or are motivated by, the results in our paper. In this section we provide a very brief collection of references to some work that has appeared since the first version of this paper (note that most of these papers cite the first arXiv version of our paper [40], which is substantially longer and has a slightly different title).

The paper [4] is probably the most similar to the present paper. Like our paper, it analyzes contractive couplings of HMC, though the details are different. Among other results, they extend our results to certain classes of nonconvex functions which are nevertheless strongly convex in a subset of $\mathbb{R}^d$.

The papers [11, 32] provide improvements to the polynomial dependence on $\frac{M_2}{m_2}$ in our Theorem 1; see the related work section of [11] for a summary of these subsequent developments. The paper [4] combines the contractive coupling introduced in our paper with a "synchronous" coupling to obtain fast mixing bounds for target distributions which are, roughly speaking, strongly log-concave in a sublevel set containing most of the target probability measure but need not be strongly log-concave in the tails. See also [2, 7, 10, 15, 26, 31, 43, 48].

It is not always easy to directly compare the results of the above papers to the present work. However, we find that it is particularly straightforward to see that [31] improves on our results in several important ways.

4. Technical results.

In this section, we give some useful bounds related to the solutions of Hamilton's equations (2.2). We begin with some notation and estimates that will be used for the remainder of the section, give a short proof sketch in Section 4.2, then prove the results.

4.1. *Definitions and notation.* Throughout this section, we assume that the potential U satisfies Assumption 1.1 with $\mathcal{X} = \mathbb{R}^d$, and we consider two solutions $(q_t^{(1)}, p_t^{(1)})$ and $(q_t^{(2)}, p_t^{(2)})$ of equation (2.2). Denote by $\tilde{q}_t := q_t^{(2)} - q_t^{(1)}$ and $\tilde{p}_t := p_t^{(2)} - p_t^{(1)}$ the differences between these solutions, and denote by $\hat{q}_t := \|\tilde{q}_t\|$ and $\hat{p}_t := \|\tilde{p}_t\|$ the magnitudes of these differences.

Hamilton's equations give

$$\frac{\mathrm{d}\tilde{q}_t}{\mathrm{d}t} = \tilde{p}_t, \qquad \frac{\mathrm{d}\tilde{p}_t}{\mathrm{d}t} = -(U'(q_t^{(2)}) - U'(q_t^{(1)})), \tag{4.1}$$

so we have

$$\frac{\mathrm{d}\hat{q}_t}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\|\tilde{q}_t\| = \frac{\langle \tilde{p}_t, \tilde{q}_t\rangle}{\|\tilde{q}_t\|} \leq \|\tilde{p}_t\| = \hat{p}_t$$

and

$$\frac{\mathrm{d}\hat{p}_t}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\|\tilde{p}_t\| = \frac{\langle \frac{\mathrm{d}}{\mathrm{d}t}\tilde{p}_t, \tilde{p}_t\rangle}{\|\tilde{p}_t\|} \leq \left\|\frac{\mathrm{d}}{\mathrm{d}t}\tilde{p}_t\right\| \overset{\text{Eq. (4.1)}}{=} \|U'(q_t^{(2)}) - U'(q_t^{(1)})\|.$$

This implies the following system of differential inequalities

$$\frac{\mathrm{d}\hat{q}_t}{\mathrm{d}t} \leq \hat{p}_t, \qquad \frac{\mathrm{d}\hat{p}_t}{\mathrm{d}t} \leq \|U'(q_t^{(2)}) - U'(q_t^{(1)})\|, \tag{4.2}$$

with initial conditions $\hat{q}_0$ and $\hat{p}_0$.

Finally, for two points $u, v \in \mathbb{R}^d$, we define the unit-speed parametrization of the line connecting u and v to be $\ell_s(u, v) := u + s\frac{v-u}{\|v-u\|}$ for all $0 \leq s \leq \|v - u\|$. We keep this last notation for the remainder of the paper.

4.2. *Proof sketch.* We will consider contraction of two Hamiltonian trajectories started in parallel, so that $\tilde{p}_0 = 0$. Using equation (4.1) and the strong convexity part of Assumption 1.1, it is straightforward to check that $\frac{\mathrm{d}\hat{q}_t}{\mathrm{d}t} = 0$ and $\frac{\mathrm{d}^2\hat{q}_t}{\mathrm{d}t^2} \leq -m_2 \hat{q}_t$ at $t = 0$. Taylor's theorem suggests that there is some small $T > 0$ for which

$$\frac{\mathrm{d}^2\hat{q}_t}{\mathrm{d}t^2} \leq -\frac{1}{2} m_2 \hat{q}_0 \tag{4.3}$$

over the interval $t \in [0, T]$. Standard ODE comparison results (see Lemma 4.2) can be used to "solve" inequalities of this form to give conclusions such as $\hat{q}_t \leq \hat{q}_0 - \frac{1}{4} m_2 \hat{q}_0 t^2$ over some time interval.

The bulk of our proof involves checking that a slightly more complicated inequality similar to (4.3) holds for a relatively large value of T, specifically, for $T = \frac{1}{2\sqrt{2}} \frac{\sqrt{m_2}}{M_2}$. Technically, the main ingredients are checking that $\hat{q}_t$, as well as certain related quantities, don't change too quickly over the time interval $[0, T]$ of interest (see Lemma 4.3), then using this to show that $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\hat{q}_t \ll 0$ for $t \in [0, T]$ (see inequality (4.23) of Lemma 4.4), and finally concluding that $\hat{q}_T$ is much smaller than $\hat{q}_0$ (see Lemma 4.4).

REMARK 4.1. The upper bounds in Lemmas 4.3 and 4.4 are only used in the regime where $t \leq T$, for $T = \frac{1}{2\sqrt{2}} \frac{\sqrt{m_2}}{M_2}$. The particular forms of the upper bounds come from Lemma 4.2, and we have made no effort to obtain good bounds for values of t much larger than T.

4.3. *ODE comparison theorem.* We make frequent use of the following comparison theorem for systems of ordinary differential equations, a generalization of Grönwall's inequality originally stated in Proposition 1.4 of [29]:

LEMMA 4.2 (ODE comparison theorem, Proposition 1.4 of [29], originally proved in Chapter 10 of [54]). *Let $U \subset \mathbb{R}^n$ and $I \subset \mathbb{R}$ be open, nonempty, and connected. Let $f, g : I \times U \to \mathbb{R}^n$ be continuous and locally Lipschitz maps. Then the following are equivalent*:

1. *For each pair (t_0, y), $(t_0, \overline{y})$ with $t_0 \in I$ and $y, \overline{y} \in U$, the inequality $y \leq \overline{y}$ implies $z(t) \leq \overline{z}(t)$ for all $t \geq t_0$, where*

$$\frac{d}{dt} z = f(t, z), \qquad z(t_0) = y,$$
$$\frac{d}{dt} \overline{z} = g(t, \overline{z}), \qquad \overline{z}(t_0) = \overline{y}.$$

2. *For all $i \in \{1, 2, \dots, n\}$ and all $t \geq t_0$, the inequality*

$$g\big(t, (\overline{x}[1], \dots, \overline{x}[i-1], x[i], \overline{x}[i+1], \dots, \overline{x}[n])\big)[i] \\ \geq f\big(t, (x[1], \dots, x[i-1], x[i], x[i+1], \dots, x[n])\big)[i]$$

holds whenever $\overline{x}[j] \geq x[j]$ for every $j \neq i$.

4.4. *Error bounds for HMC.* We give a simple estimate, showing that solutions to equation (2.2) don't diverge by very much on a small timescale:

LEMMA 4.3. *With notation as above*:

1. *For $t \geq 0$ we have*

$$\hat{q}_t \leq k_1 e^{t\sqrt{M_2}} + k_2 e^{-t\sqrt{M_2}},$$
$$\hat{p}_t \leq k_1 \sqrt{M_2} e^{t\sqrt{M_2}} - k_2 \sqrt{M_2} e^{-t\sqrt{M_2}}, \tag{4.4}$$

where $k_1 = \frac{1}{2}(\hat{q}_0 + \frac{\hat{p}_0}{\sqrt{M_2}})$, $k_2 = \frac{1}{2}(\hat{q}_0 - \frac{\hat{p}_0}{\sqrt{M_2}})$.

2. *Suppose that* $\hat{q}_0 = 0$, *and that* $\tau > 0$ *is any positive real number. Then*

$$\hat{q}_t \geq \hat{p}_0\big(t - \sqrt{M_2}\sinh(\tau\sqrt{M_2}) \times t^2\big) \quad \forall t \in [0, \tau]. \tag{4.5}$$

3. *Suppose instead that* $\hat{p}_0 = 0$ *and that* $\tau > 0$ *is such that* $\hat{q}_t \leq 2\hat{q}_0$ *on all* $t \in [0, \tau]$. *Then*

$$\hat{q}_t \geq \hat{q}_0\big(1 - 2M_2 t^2\big) \quad \forall t \in [0, \tau]. \tag{4.6}$$

PROOF. Recalling $\|D_v U'\| \leq M_2\|v\|$ for all $v \in \mathbb{R}^d$, we have the initial estimate:

$$\begin{aligned} \left\|\frac{\mathrm{d}\tilde{p}_t}{\mathrm{d}t}\right\| &= \|U'(q_t^{(2)}) - U'(q_t^{(1)})\| = \left\|\int_0^{\|\tilde{q}_t\|} D_{\frac{\tilde{q}_t}{\|\tilde{q}_t\|}} U'|_{\ell_s(q_t^{(1)}, q_t^{(2)})}\,\mathrm{d}s\right\| \\ &\leq \int_0^{\|\tilde{q}_t\|} \|D_{\frac{\tilde{q}_t}{\|\tilde{q}_t\|}} U'|_{\ell_s(q_t^{(1)}, q_t^{(2)})}\|\,\mathrm{d}s \leq \int_0^{\|\tilde{q}_t\|} M_2\,\mathrm{d}s = \|\tilde{q}_t\| M_2. \end{aligned} \tag{4.7}$$

We now prove our three conclusions in order.

Proving conclusion 1. Equations (4.2) and (4.7) together give the system of differential inequalities

$$\frac{\mathrm{d}\hat{q}_t}{\mathrm{d}t} \leq \hat{p}_t := f_1(\hat{q}_t, \hat{p}_t), \qquad \frac{\mathrm{d}\hat{p}_t}{\mathrm{d}t} \leq M_2\hat{q}_t := f_2(\hat{q}_t, \hat{p}_t).$$

Define $\hat{q}_t^\star$ and $\hat{p}_t^\star$ to be the solution to the system of differential equations

$$\frac{\mathrm{d}\hat{q}_t^\star}{\mathrm{d}t} = \hat{p}_t^\star = f_1(\hat{q}_t^\star, \hat{p}_t^\star), \qquad \frac{\mathrm{d}\hat{p}_t^\star}{\mathrm{d}t} = M_2\hat{q}_t^\star = f_2(\hat{q}_t^\star, \hat{p}_t^\star) \tag{4.8}$$

with initial conditions $\hat{q}_0^\star = \hat{q}_0$ and $\hat{p}_0^\star = \hat{p}_0$. We now compute $\hat{q}_t^\star$. Turning the system of equations (4.8) into a single second-order equation gives

$$\frac{\mathrm{d}^2\hat{q}_t^\star}{\mathrm{d}t^2} = M_2\hat{q}_t^\star$$

which has solution

$$\hat{q}_t^\star = k_1 e^{t\sqrt{M_2}} + k_2 e^{-t\sqrt{M_2}}$$

for some constants k_1, k_2. Using the initial conditions to solve for the constants

$$k_1 = \frac{1}{2}\left(\hat{q}_0 + \frac{\hat{p}_0}{\sqrt{M_2}}\right), \qquad k_2 = \frac{1}{2}\left(\hat{q}_0 - \frac{\hat{p}_0}{\sqrt{M_2}}\right).$$

Noting that

$$\hat{q}_t \leq \hat{q}_t^\star, \qquad \hat{p}_t \leq \hat{p}_t^\star, \quad t \in [0, \infty)$$

and then applying Lemma 4.2 separately to each n-dimensional vector $\hat{q}_t$ and $\hat{p}_t$ completes the proof of inequalites (4.4).

Proving conclusions 2 and 3. Define $z_t := \frac{\mathrm{d}}{\mathrm{d}t}\hat{q}_t$. Fix $\tau > 0$. Suppose that $\mathsf{C} > 0$ is a number such that $\hat{q}_t \le \mathsf{C}$ for all $t \in [0, \tau]$, to be fixed later in the proof. Now,

$$\left|\frac{\mathrm{d}}{\mathrm{d}t} z_t\right| = \left|\frac{\langle \frac{\mathrm{d}}{\mathrm{d}t}\tilde{p}_t, \tilde{q}_t\rangle}{\|\tilde{q}_t\|}\right| \le \left\|\frac{\mathrm{d}}{\mathrm{d}t}\tilde{p}_t\right\| \le M_2\hat{q}_t, \tag{4.9}$$

which implies

$$\frac{\mathrm{d}\hat{q}_t}{\mathrm{d}t} = z_t := g_1(\hat{q}_t, z_t), \qquad \frac{\mathrm{d}z_t}{\mathrm{d}t} \ge -M_2\hat{q}_t \ge -M_2\mathsf{C} := g_2(\hat{q}_t, z_t) \quad \forall t \in [0, \tau].$$

Define $\hat{q}_t^\dagger$ and $z_t^\dagger$ to be the solution to the system of differential equations

$$\frac{\mathrm{d}\hat{q}_t^\dagger}{\mathrm{d}t} = \hat{z}_t^\dagger = g_1(\hat{q}_t^\dagger, \hat{z}_t^\dagger), \qquad \frac{\mathrm{d}\hat{z}_t^\dagger}{\mathrm{d}t} = -M_2\mathsf{C} = g_2(\hat{q}_t^\dagger, \hat{z}_t^\dagger) \tag{4.10}$$

with initial conditions $\hat{q}_0^\dagger = \hat{q}_0$ and $z_0^\dagger = z_0$.

Since g_1 and g_2 are nondecreasing in each variable, Lemma 4.2, applied to the $2n$-dimensional vector $[\hat{q}_t; \hat{p}_t]$, implies

$$\hat{q}_t \ge \hat{q}_t^\dagger, \qquad \hat{p}_t \ge z_t^\dagger, \quad t \in [0, \tau].$$

Hence, all we need to do is solve for $z_t^\dagger$. We can turn the system of equations (4.10) into the following second-order equation:

$$\frac{\mathrm{d}^2\hat{q}_t^\dagger}{\mathrm{d}t^2} = -M_2\mathsf{C},$$

whose solutions are of the form

$$\hat{q}_t^\dagger = -M_2\mathsf{C}t^2 + z_0 t + \hat{q}_0.$$

Therefore, we have

$$\hat{q}_t \ge \hat{q}_t^\dagger = -M_2\mathsf{C}t^2 + z_0 t + \hat{q}_0 \quad \forall t \in [0, \tau].$$

In the special case that $\hat{q}_0 = 0$, we have that $z_0 = \hat{p}_0$. By inequality (4.4), for any $\tau > 0$ we also have in this special case that $\hat{q}_t \le \frac{\hat{p}_0}{\sqrt{M_2}}\sinh(\tau\sqrt{M_2})$ for all $t \in [0, \tau]$. So setting $\mathsf{C} = \frac{\hat{p}_0}{\sqrt{M_2}}\sinh(\tau\sqrt{M_2})$ completes the proof of inequality (4.5).

In the special case that $\hat{p}_0 = 0$, we have $|z_0| \le |\hat{p}_0| = 0$. Suppose that $\tau > 0$ is such that $\hat{q}_t \le 2\hat{q}_0$ holds for all $t \in [0, \tau]$. Then setting $\mathsf{C} = 2\hat{q}_0$ completes the proof of inequality (4.6). □

4.5. *Contraction for strongly log-concave targets.* In this section, we show that two solutions to Hamilton's equations with the same initial momenta will tend to move closer to each other over a moderate time interval (see again Figure 2).

For this section, we define the error function

$$\mathfrak{F}(t) := \frac{\sinh^2(t)}{1 - 2t^2} \tag{4.11}$$

for all $t \ge 0$ (note that $\mathfrak{F}(t) \approx t^2$ for $t > 0$ sufficiently small; see Figure 3 for a plot of $\mathfrak{F}(t)$ and the functions $\Psi_T(t)$ and $\psi(t)$ used in Lemma 4.4 for specific values of T, m_2, M_2). We can now prove the following contraction estimate.

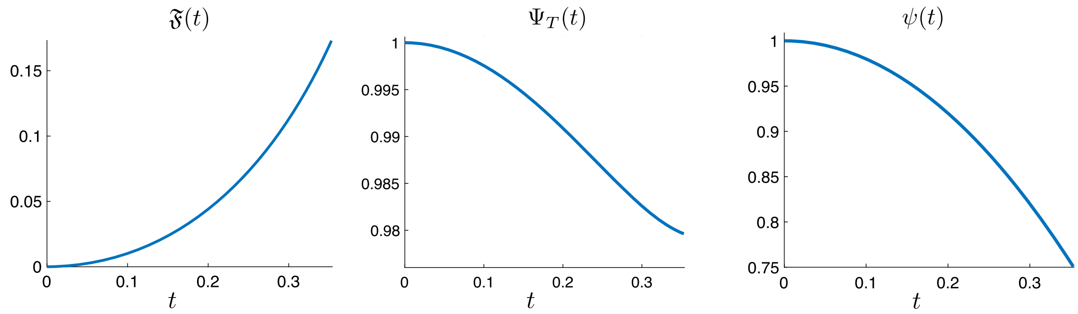

FIG. 3. *A plot of the functions* $\mathfrak{F}(t)$, $\Psi_T(t)$ *and* $\psi(t)$, *for* $t \in [0, T]$. *In these plots we take* $T = \frac{1}{2\sqrt{2}}\frac{\sqrt{m_2}}{M_2}$, $m_2 = 1$, *and* $M_2 = 1$.

LEMMA 4.4. *Define*

$$\Psi_T(t) := \left(-\frac{1}{2} + \frac{M_2}{m_2}\mathfrak{F}(T\sqrt{M_2})\right) \cdot \frac{1}{2}(\sqrt{m}_2 t)^2 + 1,$$
$$\psi(t) := -2M_2 t^2 + 1.$$

Suppose that $0 < T \le \frac{1}{2\sqrt{2}}\frac{\sqrt{m_2}}{M_2}$. *Then if* $\tilde{p}_0 = 0$,

$$\hat{q}_0\psi(t) \le \hat{q}_t \le \hat{q}_0\Psi_T(t) \quad \forall t \in [0, T]. \tag{4.12}$$

PROOF. We have

$$\langle -(U'(q_t^{(2)}) - U'(q_t^{(1)})), q_t^{(2)} - q_t^{(1)}\rangle \le -m_2\|q_t^{(2)} - q_t^{(1)}\|^2. \tag{4.13}$$

Note that

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\tilde{q}_t\| = \frac{\langle \tilde{p}_t, \tilde{q}_t\rangle}{\|\tilde{q}_t\|}, \tag{4.14}$$

so that

$$\|\tilde{q}_t\|\frac{\mathrm{d}}{\mathrm{d}t}\|\tilde{q}_t\| = \langle \tilde{p}_t, \tilde{q}_t\rangle. \tag{4.15}$$

Taking derivatives,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\|\tilde{q}_t\|\frac{\mathrm{d}}{\mathrm{d}t}\|\tilde{q}_t\|\right) &= \frac{\mathrm{d}}{\mathrm{d}t}\langle \tilde{p}_t, \tilde{q}_t\rangle = \left\langle \tilde{p}_t, \frac{\mathrm{d}\tilde{q}_t}{\mathrm{d}t}\right\rangle + \left\langle \frac{\mathrm{d}\tilde{p}_t}{\mathrm{d}t}, \tilde{q}_t\right\rangle \\ &= \|\tilde{p}_t\|^2 + \left\langle \frac{\mathrm{d}\tilde{p}_t}{\mathrm{d}t}, \tilde{q}_t\right\rangle \\ &\overset{\text{Eq. (4.1)}}{=} \|\tilde{p}_t\|^2 + \langle -(U'(q_t^{(2)}) - U'(q_t^{(1)})), q_t^{(2)} - q_t^{(1)}\rangle \\ &\overset{\text{Eq. (4.13)}}{\le} \|\tilde{p}_t\|^2 - m_2\|q_t^{(2)} - q_t^{(1)}\|^2.\end{aligned} \tag{4.16}$$

Applying the chain rule to the LHS of equation (4.15),

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\|\tilde{q}_t\|\frac{\mathrm{d}}{\mathrm{d}t}\|\tilde{q}_t\|\right) &= \|\tilde{q}_t\|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\|\tilde{q}_t\| + \left(\frac{\mathrm{d}}{\mathrm{d}t}\|\tilde{q}_t\|\right) \times \left(\frac{\mathrm{d}}{\mathrm{d}t}\|\tilde{q}_t\|\right) \\ &\ge \|\tilde{q}_t\|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\|\tilde{q}_t\|.\end{aligned} \tag{4.17}$$

Combining equality (4.17) with inequality (4.16), we get

$$\|\tilde{q}_t\|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\|\tilde{q}_t\| \le \|\tilde{p}_t\|^2 - m_2\|\tilde{q}_t\|^2,$$

and rearranging

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\|\tilde{q}_t\| \le \frac{\|\tilde{p}_t\|^2}{\|\tilde{q}_t\|} - m_2\|\tilde{q}_t\|. \tag{4.18}$$

Recall that we assume $\tilde{p}_0 = 0$ in the statement of this Lemma. Inequality (4.12) is obviously true if $\hat{q}_0 = 0$, so without loss of generality we may also assume that $\hat{q}_0 > 0$. By the fact that solutions to Hamilton's equations are continuous (first shown in [49]) and our assumptions that $\tilde{p}_0 = 0$ and $\|\tilde{q}_0\| = \hat{q}_0 \neq 0$, there must exist an $\epsilon > 0$ such that $\frac{\|\tilde{p}_t\|^2}{\|\tilde{q}_t\|} - m_2\|\tilde{q}_t\| < 0$ for every $t \in (0, \epsilon]$. Recall that $\hat{q}_t := \|\tilde{q}_t\|$ and define

$$\tau_1 := \max\{\tau \in [0, T] : \hat{q}_t \le 2\hat{q}_0\ \forall t \in [0, \tau]\},$$

where a maximum value exists by continuity of $\hat{q}_t$. By parts 3 and 1, respectively, of Lemma 4.3, we have

$$\|\tilde{q}_t\| \ge -2M_2\hat{q}_0 t^2 + \hat{q}_0 \quad \forall t \in [0, \tau_1] \tag{4.19}$$

and

$$\|\tilde{p}_t\| \le \frac{1}{2}\hat{q}_0\sqrt{M_2}\big(e^{t\sqrt{M_2}} - e^{-t\sqrt{M_2}}\big). \tag{4.20}$$

Hence, inequalities (4.18), (4.19), and (4.20) together imply that for all $t \in [0, \tau_1]$ we have

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\|\tilde{q}_t\| &\le \frac{\|\tilde{p}_t\|^2}{\|\tilde{q}_t\|} - m_2\|\tilde{q}_t\| \le \frac{[\frac{1}{2}\hat{q}_0\sqrt{M_2}(e^{t\sqrt{M_2}} - e^{-t\sqrt{M_2}})]^2}{-2M_2\hat{q}_0t^2 + \hat{q}_0} - m_2\|\tilde{q}_t\| \\ &= -m_2\|\tilde{q}_t\| + \frac{\frac{1}{4}\hat{q}_0M_2(e^{t\sqrt{M_2}} - e^{-t\sqrt{M_2}})^2}{-2(t\sqrt{M_2})^2 + 1}.\end{aligned} \tag{4.21}$$

But $T \le \frac{1}{2\sqrt{2}}\frac{\sqrt{m_2}}{\sqrt{M_2}}\frac{1}{\sqrt{M_2}}$ implies that $\mathfrak{F}(t\sqrt{M_2})$ is nondecreasing on $t \in [0, T]$, so by inequality (4.21) we have

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\|\tilde{q}_t\| \le -m_2\|\tilde{q}_t\| + \hat{q}_0M_2\mathfrak{F}(T\sqrt{M_2}) \quad \forall t \in [0, \tau_1]. \tag{4.22}$$

Define

$$\tau_2 := \max\left\{\tau \in [0, T] : \hat{q}_t \ge \frac{1}{2}\hat{q}_0\ \forall t \in [0, \tau]\right\},$$

where a maximum value exists by continuity of $\hat{q}_t$. Then by inequality (4.22),

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\|\tilde{q}_t\| \le -\frac{m_2}{2}\hat{q}_0 + \hat{q}_0M_2\mathfrak{F}(T\sqrt{M_2}) \quad \forall t \in [0, \tau_1] \cap [0, \tau_2]. \tag{4.23}$$

Let $\hat{q}_t^{\ddagger}$ be the solution to the differential equation

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\hat{q}_t^{\ddagger} = -\frac{m_2}{2}\hat{q}_0 + \hat{q}_0M_2\mathfrak{F}(T\sqrt{M_2}),$$

with initial conditions $\hat{q}_0^{\ddagger} = \hat{q}_0$ and $\frac{\mathrm{d}}{\mathrm{d}t}\hat{q}_0^{\ddagger} = \hat{p}_0 = 0$. Since the RHS of the differential inequality (4.23) is nondecreasing in both variables $\hat{q}_t$ and $\frac{\mathrm{d}}{\mathrm{d}t}\hat{q}_t$, we have, by Lemma 4.2 applied to the $2n$-dimensional vector $[\hat{q}_t; \frac{\mathrm{d}}{\mathrm{d}t}\hat{q}_t]$, that

$$\hat{q}_t \le \hat{q}_t^{\ddagger} \quad \forall t \in [0, \tau_1] \cap [0, \tau_2]. \tag{4.24}$$

Solving the differential equation for $\hat{q}_t^{\ddagger}$ gives

$$\begin{aligned} \hat{q}_t^{\ddagger} &= \left(-\frac{m_2}{2}\hat{q}_0 + \hat{q}_0 M_2 \mathfrak{F}(T\sqrt{M_2})\right) \cdot \frac{1}{2}t^2 + \hat{p}_0 t + \hat{q}_0 \\ &= \left(-\frac{m_2}{2}\hat{q}_0 + \hat{q}_0 M_2 \mathfrak{F}(T\sqrt{M_2})\right) \cdot \frac{1}{2}t^2 + \hat{q}_0 \\ &= \hat{q}_0\left[\left(-\frac{1}{2} + \frac{M_2}{m_2}\mathfrak{F}(T\sqrt{M_2})\right) \cdot \frac{1}{2}(\sqrt{m}_2 t)^2 + 1\right], \end{aligned}$$

where the second line uses the fact that $\hat{p}_0 = 0$. Therefore, by inequality (4.24), this implies

$$\hat{q}_t \le \hat{q}_0\left[\left(-\frac{1}{2} + \frac{M_2}{m_2}\mathfrak{F}(T\sqrt{M_2})\right) \cdot \frac{1}{2}(\sqrt{m}_2 t)^2 + 1\right] \tag{4.25}$$

for all $t \in [0, \tau_1] \cap [0, \tau_2]$.

Note that $\frac{\sqrt{m_2}}{\sqrt{M_2}} \le 1$, so

$$T \le \frac{1}{2\sqrt{2}}\frac{\sqrt{m_2}}{\sqrt{M_2}}\frac{1}{\sqrt{M_2}} \le \frac{1}{2\sqrt{2}}\frac{1}{\sqrt{M_2}}. \tag{4.26}$$

Hence,

$$\psi(t) = -2M_2 t^2 + 1 \ge \frac{3}{4} \quad \forall t \in [0, T]. \tag{4.27}$$

We calculate

$$\frac{M_2}{m_2}\mathfrak{F}(T\sqrt{M_2}) = \frac{M_2}{m_2}\frac{\sinh^2(\sqrt{M_2}T)}{1 - 2(\sqrt{M_2}T)^2} \le \frac{M_2}{m_2}\frac{\sinh^2(\sqrt{M_2}T)}{1 - 2(\sqrt{M_2}T)^2} \le \frac{1}{4}, \tag{4.28}$$

where both inequalities use the fact that $T \le \frac{1}{2\sqrt{2}}\frac{\sqrt{m_2}}{\sqrt{M_2}}\frac{1}{\sqrt{M_2}}$ and the first also uses the fact that $\sinh^2(t) \le 1.2t^2$ for all $t \in [0, \frac{1}{2}]$. This bound implies

$$\Psi_T(t) \le 1 - \frac{1}{4} \cdot \frac{1}{2}(\sqrt{m}_2 t)^2 \le 1 \quad \forall t \in [0, T]. \tag{4.29}$$

Therefore, inequalities (4.27) and (4.29), respectively, imply that

$$\hat{q}_0\psi(t) \ge \frac{3}{4}\hat{q}_0 \quad \text{and} \quad \hat{q}_0\Psi_T(t) \le \hat{q}_0 \quad \forall t \in [0, T]. \tag{4.30}$$

We now define τ_3 to be supremum of all values of $0 \le \tau \le T$ so that the following inequalities hold:

$$\frac{3}{4}\hat{q}_0 \le \hat{q}_0\psi(t) \le \hat{q}_t \le \hat{q}_0\Psi_T(t) \le \hat{q}_0 \quad \forall t \in [0, \tau]. \tag{4.31}$$

Observe that $\Psi_T(0) = \psi(0) = 1$, which implies $\tau_3 \ge 0$. Moreover, since $\hat{q}_t$, $\psi(t)$, and $\Psi_T(t)$ are continuous on $t \in [0, T]$, inequalities (4.31) are satisfied for $t = \tau_3$. We now prove by contradiction that in fact $\tau_3 = T$.

CLAIM: $\tau_3 = T$. Suppose (toward a contradiction) that $\tau_3 < T$. Since $\hat{q}_t$ is continuous on $t \in \mathbb{R}$, equation (4.31) (which is satisfied for $\tau = \tau_3$) implies that there exists a number δ with $0 < \delta \le T - \tau_3$ such that

$$\frac{1}{2}\hat{q}_0 \le \hat{q}_t \le 2\hat{q}_0 \quad \forall t \in [0, \tau_3 + \delta]. \tag{4.32}$$

Equation (4.32) implies that $\tau_3 + \delta \le \tau_1$ and $\tau_3 + \delta \le \tau_2$. Therefore, by equation (4.25), we have

$$\hat{q}_t \le \hat{q}_0 \Psi_T(t) \quad \forall t \in [0, \tau_3 + \delta] \tag{4.33}$$

and by part 3 of Lemma 4.3 we have

$$\hat{q}_0 \psi(t) \le \hat{q}_t \quad \forall t \in [0, \tau_3 + \delta]. \tag{4.34}$$

Hence, equations (4.30), (4.33), and (4.34) together imply that

$$\frac{3}{4}\hat{q}_0 \le \hat{q}_0 \psi(t) \le \hat{q}_t \le \hat{q}_0 \Psi_T(t) \le \hat{q}_0 \quad \forall t \in [0, \tau_3 + \delta]. \tag{4.35}$$

But equation (4.35) implies that equation (4.31) is satisfied for $\tau = \tau_3 + \delta$, which contradicts the fact that $\tau = \tau_3$ is the largest value of τ that satisfies equation (4.31). Therefore, by contradiction, our assumption that $\tau_3 < T$ must be false. □

Therefore, $\tau_3 = T$, and so equation (4.31) is satisfied for $\tau = T$:

$$\frac{3}{4}\hat{q}_0 \le \hat{q}_0 \psi(t) \le \hat{q}_t \le \hat{q}_0 \Psi_T(t) \le \hat{q}_0 \quad \forall t \in [0, T].$$

This completes the proof of the Lemma. □

This bound quickly implies the main result of this section:

THEOREM 3 (Contraction For Hamiltonian mechanics with convex potentials). *For* $0 \le T \le \frac{1}{2\sqrt{2}} \frac{\sqrt{m_2}}{M_2}$,

$$\hat{q}_T \le \left[1 - \frac{1}{8}(\sqrt{m}_2 T)^2\right] \times \hat{q}_0. \tag{4.36}$$

In particular, if $T = \frac{1}{2\sqrt{2}} \frac{\sqrt{m_2}}{M_2}$, *then*

$$\hat{q}_T \le \left[1 - \frac{1}{64}\left(\frac{m_2}{M_2}\right)^2\right] \times \hat{q}_0. \tag{4.37}$$

PROOF. By inequality (4.28),

$$\frac{M_2}{m_2} \mathfrak{F}(T\sqrt{M_2}) \le \frac{1}{4}. \tag{4.38}$$

This implies

$$\Psi_T(t) \le 1 - \frac{1}{4} \cdot \frac{1}{2}(\sqrt{m}_2 t)^2 \quad \forall t \in [0, T]. \tag{4.39}$$

Applying Lemma 4.4, equation (4.39) implies that

$$\hat{q}_T \le \left[1 - \frac{1}{8}(\sqrt{m}_2 T)^2\right] \times \hat{q}_0.$$

This completes the proof of inequality (4.36). Inequality (4.37) is an immediate consequence of inequality (4.36). □

5. Drift condition. Although this paper focuses on mixing bounds, we feel it is worth mentioning that the strong log-concavity assumption will also imply a quantitatively useful drift condition:

THEOREM 4 (Drift conditions for HMC: quadratic tails). *Fix* $1 < C < \infty$ *and define*

$$S = \{x \in \mathbb{R}^d : \|x\| \leq C\}.$$

Let Assumption 1.1 hold for $\mathcal{X} = S^c$. *Let* $\{X_t\}_{t \geq 0}$ *be the ideal HMC Markov chain* (*Markov chain 1*) *with parameter* $T = \frac{\sqrt{m_2}}{2\sqrt{2}M_2}$. *Then*

$$\mathbb{E}[e^{\|X_1\|} | X_0] \leq e^{-1} e^{\|X_0\|} + A, \tag{5.1}$$

where the constant $0 < A < \infty$ *satisfies*

$$\log(A) = O\left(\max\left(\frac{d}{m_2}, M_2, \left(\frac{M_2}{m_2}\right)^5 m_2^{-2}, \left(\frac{M_2}{m_2}\right)^{2.5} m_2^{-0.5}, C\frac{\sqrt{M_2}}{m_2}\right)\right).$$

PROOF. The proof is given in Appendix B. □

REMARK 5.1 (Other drift bounds in the HMC literature). Quantitative drift conditions are useful for extending quantiative mixing bounds in the present paper to more general distributions, which may have log-concave bulks but not tails. Finding such bounds is not generally easy, and is beyond the scope of this paper, but we mention some existing related work. Although we are not aware of published work obtaining general quantitative drift conditions in the HMC literature, there has been substantial work on finding drift conditions without explicit quantitative bounds, most noteably in [36] and [20]. Of course it is possible to obtain quantitative bounds by following the calculations in for example, Theorem 5.4 of [36]. However, in order to obtain useful results for the ideal HMC integrator, additional assumptions are required. We briefly sketch the difficulty.

In the last clause in the last sentence of the proof of Proposition 5.10, they assert that a certain sum is strictly positive because all of its terms are nonnegative and at least one is strictly positive. The continuous analogue to this argument fails: the analogous integral cannot be bounded away from 0 simply by noting that the integrand is strictly positive at at least one point. In order to avoid this problem, it is sufficient to add an assumption that $U'(q)$ changes slowly with q.

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 optimal dependence on dimension, but we leave many important questions open. In this section, we mention those that are most interesting to us.

6.1. *Relationship to the Jacobi metric.* The biggest difference between the approach of the previous paper that finds quantitative mixing bounds for HMC, [53], and our paper, is as follows. [53] uses concentration of measure to analyze contraction of "*typical*" Hamiltonian trajectories with parallel initial momenta on strongly convex potentials by expressing them as geodesic trajectories on a positively curved manifold under the Jacobi metric, applying the Rauch comparison theorem from differential geometry. Our paper instead proves contraction of *all* Hamiltonian trajectories with parallel initial momenta by applying comparison theorems for ordinary differential equations (ODEs) directly to the Hamilton's equations that define the Hamiltonian trajectories. Because the Jacobi manifold is never defined on the entire state space of an HMC Markov chain, it does not seem possible to extend an approach based on the Jacobi manifold to obtain uniform contraction estimates for all Hamiltonian trajectories. As a result, we are able to achieve bounds that do not grow explicitly with the dimension d, while the Jacobi metric approach in [53] yields bounds that grow like d^2.

We found this slightly surprising: the Jacobi metric is a natural tool for analyzing HMC, and it is far from clear to us if the technical difficulties that appear in [53] can be overcome. We leave as an open problem the question of whether the Jacobi metric approach of [53] can be refined to obtain bounds that do not grow explicitly with d, as well as to possibly further strengthen the relaxation time bound in our Theorem 1 from roughly $O(\frac{M_2^2}{m_2^2})$ to the conjectured value of roughly $O(\frac{M_2}{m_2})$. (We propose the conjectured dependences on M_2 and m_2 here based on the best available bounds for the Langevin diffusion [21] and the closely related geodesic walk [41], as well as exact solutions to Hamilton's equations that are available in the special case of a Gaussian target measure.)

6.2. *Riemannian HMC.* This paper analyzes one of the simplest possible HMC algorithms. However, many other variants exist. Riemannian HMC, introduced in [23], is one of the most popular. This approach seems to obviate the need for the preconditioning step discussed in Remark 2.1, but there are very few rigorous results on the performance of this algorithm. It would be interesting to check that Riemannian HMC does have this property, and that no additional problems arise.

6.3. *Quantitative drift conditions.* As discussed in [36], it is more difficult to obtain a Lyapunov condition for HMC than for RWM. Obtaining much more general quantitative drift conditions for HMC would allow us to check that bad behavior "in the tails" of the target distribution does not greatly influence mixing.

6.4. *De-biasing with coupling for parallel processing.* In [25], a coupling similar to the one described in Figure 2 is used to provide unbiased samples of the target density from HMC Markov chains that are numerically implemented in parallel. As the authors of [25] ask in their discussion section, it would be interesting to see if the contraction bounds obtained in our paper could be used to provide stronger convergence guarantees for their algorithm.

APPENDIX A: PROOF OF COROLLARY 2

PROOF. Fix $\epsilon > 0$ as in the statement of Corollary 2 and set $\hat{\mathcal{I}} = \mathcal{I} \times 10^3 \log(D \times \frac{144 M_2 d^2 \log^{\frac{1}{2}}(\frac{\kappa}{\epsilon})}{\epsilon\sqrt{m_2}}) = O(\kappa^2 \log(\frac{D \times d M_2}{\epsilon\sqrt{m_2}}))$, where $\mathcal{I}$ is the number in Theorem 1.

Next consider some $i \geq \hat{\mathcal{I}}$. By equation (2.6) of Theorem 1, $W_\infty(\mathcal{L}(X_i), \pi) \leq \frac{\epsilon\sqrt{m_2}}{144 M_2 d^2 \log^{\frac{1}{2}}(\frac{\kappa}{\epsilon})}$. Hence, there is a random variable $Y \sim \pi$ coupled to X_i such that $\|X_i - Y\| \leq \frac{\epsilon\sqrt{m_2}}{144 M_2 d^2 \log^{\frac{1}{2}}(\frac{\kappa}{\epsilon})}$ with probability 1.

By Assumption 1.1, and our assumption that $U(x)$ has a global minimum at $x = 0$, we have

$$
\begin{aligned}
\pi(x) &= \frac{1}{\int_{\mathbb{R}^d} e^{-U(x)}\,\mathrm{d}x} e^{-U(x)} \leq \frac{1}{\int_{\mathbb{R}^d} e^{-M_2\|x\|^2}\,\mathrm{d}x} e^{-m_2\|x\|^2} \\
&= \frac{\int_{\mathbb{R}^d} e^{-m_2\|x\|^2}\,\mathrm{d}x}{\int_{\mathbb{R}^d} e^{-M_2\|x\|^2}\,\mathrm{d}x} \times \frac{1}{\int_{\mathbb{R}^d} e^{-m_2\|x\|^2}\,\mathrm{d}x} e^{-m_2\|x\|^2} = \kappa^{\frac{d}{2}} \times \frac{1}{\int_{\mathbb{R}^d} e^{-m_2\|x\|^2}\,\mathrm{d}x} e^{-m_2\|x\|^2}.
\end{aligned}
\tag{A.1}
$$

Thus, letting $\xi \sim N(0, I_d)$ be a standard normal random variable, we have

$$\begin{aligned}\mathbb{P}\Big(\|Y\|^2 \geq s + \frac{d}{m_2}\Big) &\overset{\text{Eq. (A.1)}}{\leq} \kappa^{\frac{d}{2}}\mathbb{P}\Big(\frac{1}{m_2}\|\xi\|^2 \geq s + \frac{d}{m_2}\Big)\\ &= \kappa^{\frac{d}{2}}\mathbb{P}\big(\|\xi\|^2 \geq m_2 s + d\big)\\ &\leq \kappa^{\frac{d}{2}} e^{-\frac{m_2 s}{8}} \quad \forall s > \frac{d}{m_2},\end{aligned}$$

where the last inequality holds by the Hason–Wright inequality [52].

Hence, we have

$$\mathbb{P}\bigg(\|Y\| \geq 9\frac{\sqrt{d}}{\sqrt{m_2}}\log^{\frac{1}{2}}\Big(\frac{\kappa}{\epsilon}\Big)\bigg) \leq \frac{\epsilon}{8}.$$

Hence, since U has M_2-Lipschitz gradient, with probability at least $1 - \frac{\epsilon}{8}$, we have that

$$U(Y) - \frac{\epsilon}{8} \leq U(Y+z) \leq U(Y) + \frac{\epsilon}{8} \quad \forall z \in B\bigg(0, \frac{\epsilon\sqrt{m_2}}{72M_2\sqrt{d}}\log^{-\frac{1}{2}}\Big(\frac{\kappa}{\epsilon}\Big)\bigg),$$

and hence that

$$e^{-\frac{\epsilon}{8}} \leq \frac{\pi(Y+z)}{\pi(Y)} \leq e^{\frac{\epsilon}{8}} \quad \forall z \in B\bigg(0, \frac{\epsilon\sqrt{m_2}}{72M_2\sqrt{d}\log^{\frac{1}{2}}(\frac{\kappa}{\epsilon})}\bigg), \tag{A.2}$$

with probability at least $1 - \frac{\epsilon}{8}$.

Let $S \subset \mathbb{R}^d$ be the subset of points $Y \in \mathbb{R}^d$ for which inequality (A.2) holds. Then

$$\mathbb{P}(S) \geq 1 - \frac{\epsilon}{8}. \tag{A.3}$$

Let ζ be uniformly distributed on the unit ball $B(0,1)$ and let $\hat{Y} = Y + \frac{\epsilon\sqrt{m_2}}{144M_2\sqrt{d}\log^{\frac{1}{2}}(\frac{\kappa}{\epsilon})}\zeta$. Let V be the volume of the ball of radius R, where $\mathsf{R} := \frac{\epsilon\sqrt{m_2}}{72M_2\sqrt{d}\log^{\frac{1}{2}}(\frac{\kappa}{\epsilon})}$. Let ρ be the density of $\hat{Y}$. Then at any point $\hat{y} \in \mathbb{R}^d$, the density $\rho(\hat{y})$ is exactly the integral

$$\rho(\hat{y}) = \frac{1}{V}\int_{B(\hat{y},R)} \pi(z)\,dz.$$

Inequality (A.2) implies that for any point $\hat{y} \in S + B(0, \mathsf{R})$, where "+" denotes the Minkowski sum, we have

$$e^{-\frac{\epsilon}{8}} = \frac{1}{V}\int_{B(\hat{y},R)} e^{-\frac{\epsilon}{8}}\,\mathrm{d}z \leq \frac{1}{V}\int_{B(\hat{y},R)} \pi(z)\,\mathrm{d}z \leq \frac{1}{V}\int_{B(\hat{y},R)} e^{\frac{\epsilon}{8}}\,\mathrm{d}z = e^{\frac{\epsilon}{8}},$$

and hence that

$$e^{-\frac{\epsilon}{8}} \leq \rho(\hat{y}) \leq e^{\frac{\epsilon}{8}} \quad \forall \hat{y} \in S + B(0, \mathsf{R}). \tag{A.4}$$

Therefore, inequalities (A.3) and (A.4) together imply that

$$\begin{aligned}\|\mathcal{L}(\hat{Y}) - \mathcal{L}(Y)\|_{\mathrm{TV}} &\overset{\text{Eq. (A.4)}}{\leq} 1 - \mathbb{P}\big(\hat{Y} \in (S + B(0,\mathsf{R}))\big) + \big(e^{\frac{1}{8}} - 1\big)\\ &\leq 1 - \mathbb{P}(Y \in S) + \big(e^{\frac{1}{8}} - 1\big)\\ &\overset{\text{Eq. (A.3)}}{\leq} \frac{1}{8} + \big(e^{\frac{1}{8}} - 1\big)\\ &\leq \frac{\epsilon}{3}.\end{aligned} \tag{A.5}$$

But we also have that

$$\|X_i - Y\| \leq \frac{\epsilon\sqrt{m_2}}{144M_2 d^2 \log^{\frac{1}{2}}(\frac{\kappa}{\epsilon})}, \tag{A.6}$$

with probability 1. Moreover, inequality (A.6) implies that $B(X_i, \frac{\epsilon\sqrt{m_2}}{144M_2\sqrt{d}\log^{\frac{1}{2}}(\frac{\kappa}{\epsilon})}) \subseteq B(Y, \frac{\epsilon\sqrt{m_2}}{72M_2\sqrt{d}\log^{\frac{1}{2}}(\frac{\kappa}{\epsilon})})$.

Hence, since $\hat{X}_i = X_i + \frac{\epsilon\sqrt{m_2}}{144M_2\sqrt{d}\log^{\frac{1}{2}}(\frac{\kappa}{\epsilon})} Z_i$ and $\hat{Y} = Y + \frac{\epsilon\sqrt{m_2}}{144M_2\sqrt{d}\log^{\frac{1}{2}}(\frac{\kappa}{\epsilon})}\zeta$, standard concentration inequalities for the unit ball imply that

$$\|\mathcal{L}(\hat{X}_i) - \mathcal{L}(\hat{Y})\|_{\mathrm{TV}} \leq \frac{\epsilon}{8}. \tag{A.7}$$

Therefore, by inequalities (A.5) and (A.7), we have that

$$\|\mathcal{L}(\hat{X}_i) - \pi\|_{\mathrm{TV}} = \|\mathcal{L}(\hat{X}_i) - \mathcal{L}(Y)\|_{\mathrm{TV}} \leq \frac{\epsilon}{2}. \qquad \square$$

APPENDIX B: PROOF OF THEOREM 4

In order to prove Theorem 4, we need the following very rough bound on the distance that can be travelled by solutions to Hamilton's equations.

LEMMA B.1. *Assume that there exists some* $0 < C < \infty$ *so that*

$$\|U'(q)\| \leq C\|q\| \tag{B.1}$$

for all $q \in \mathbb{R}^d$. *Let* (q_t, p_t) *be solutions to Hamilton's equations* (2.2) *with initial conditions* $(q_0, p_0) = (\mathbf{q}, \mathbf{p}) \in \mathbb{R}^{2d}$. *Then for all* $t \geq 0$,

$$\|q_t - \mathbf{q}\| \leq \frac{1}{2C}\big(e^{-\sqrt{C}t}(e^{\sqrt{C}t} - 1)(\sqrt{C}\|\mathbf{p}\|(e^{\sqrt{C}t} + 1) + C\|\mathbf{q}\|(e^{\sqrt{C}t} - 1))\big).$$

PROOF. Note that

$$f(t) \equiv \frac{1}{2C}\big(e^{-\sqrt{C}t}(e^{\sqrt{C}t} - 1)(\sqrt{C}\|\mathbf{p}\|(e^{\sqrt{C}t} + 1) + C\|\mathbf{q}\|(e^{\sqrt{C}t} - 1))\big)$$

is a solution to the system of equations

$$\begin{aligned} &f(0) = 0, \\ &f'(0) = \|\mathbf{p}\|, \\ &f''(t) = Cf(t) + C\|\mathbf{q}\|, \quad t \geq 0. \end{aligned}$$

By Hamilton's equations (2.2) and inequality (B.1),

$$\begin{aligned} \frac{d^2}{dt^2}\|q_t - q_0\| &\leq \|U'(q_t)\| \\ &\leq C\|q_t\| \\ &\leq C\|q_t - q_0\| + C\|q_0\|. \end{aligned}$$

We also have

$$\frac{d}{dt}\Big|_{t=0}\|q_t - q_0\| \leq \|p_0\|.$$

By Lemma 4.2, this implies

$$\begin{aligned}\|q_t - q_0\| &\le f(t)\\ &= \frac{1}{2C}\big(e^{-\sqrt{C}t}\big(e^{\sqrt{C}t} - 1\big)\big(\sqrt{C}\|\mathbf{p}\|\big(e^{\sqrt{C}t} + 1\big) + C\|\mathbf{q}\|\big(e^{\sqrt{C}t} - 1\big)\big)\big)\end{aligned}$$

for all $t \ge 0$. This completes the proof. □

The following lemma gives us the main bounds in the proof of Theorem 4.

LEMMA B.2 (Lyapunov function for Hamiltonian dynamics: Gaussian-like tails). *Let U satisfy Assumption 1.1 with $\mathcal{X} = \mathbb{R}^d$. Fix an initial position $(\mathbf{q}, \mathbf{p}) \in \mathbb{R}^{2d}$ that satisfies*

$$\frac{1}{\sqrt{2m_2}}\|\mathbf{p}\| - \frac{1}{512\kappa^2}\|\mathbf{q}\| \le -1 \tag{B.2}$$

and let $(q_t, p_t)_{t\ge 0}$ be a solution to equation (2.2) with initial conditions $q_0 = \mathbf{q}$, $p_0 = \mathbf{p}$. Then for $T = \frac{1}{2\sqrt{2}}\frac{\sqrt{m_2}}{M_2}$,

$$e^{\|q_T\|} \le e^{-1}e^{\|\mathbf{q}\|} \tag{B.3}$$

and also

$$\inf_{0\le s\le T}\|q_s\| \ge \frac{\|\mathbf{q}\|}{2} - \frac{\|\mathbf{p}\|}{2\sqrt{M_2}}. \tag{B.4}$$

PROOF. Let $(\alpha(t), \beta(t))_{t\ge 0}$ be a solution to equation (2.2) with initial conditions $\alpha(0) = 0$, $\beta(0) = \mathbf{p}$. By Theorem 3,

$$\begin{aligned}\|\alpha(T) - q_T\| &\le \left[1 - \frac{1}{64}(\sqrt{m}_2 T)^2\right] \times \|\alpha(0) - \mathbf{q}\|\\ &= \left(1 - \frac{1}{64}m_2 T^2\right)\|\mathbf{q}\|.\end{aligned} \tag{B.5}$$

By conservation of energy for Hamilton's equations,

$$\frac{1}{2}\|\beta(T)\|^2 + U\big(\alpha(T)\big) = \frac{1}{2}\|\beta(0)\|^2 + U\big(\alpha(0)\big) = \frac{1}{2}\|\beta(0)\|^2,$$

so

$$U\big(\alpha(T)\big) \le \frac{1}{2}\|\beta(0)\|^2. \tag{B.6}$$

By our assumptions, for $q \in \mathbb{R}^d$ we have

$$U(q) \ge m_2\|q\|^2.$$

Combining this with inequality (B.6), we have

$$m_2\|\alpha(T)\|^2 \le \frac{1}{2}\|\beta(0)\|^2,$$

so that

$$\|\alpha(T)\| \le \frac{1}{\sqrt{2m_2}}\|\beta(0)\|. \tag{B.7}$$

Combining this with inequality (B.5) and then the assumption (B.2),

$$\begin{aligned}\|q_T\| &\le \|\alpha(T)\| + \|\alpha(T) - q_T\| \\ &\le \frac{1}{\sqrt{2m_2}}\|\mathbf{p}\| + \left(1 - \frac{1}{64}m_2T^2\right)\|\mathbf{q}\| \\ &= \frac{1}{\sqrt{2m_2}}\|\mathbf{p}\| + \left(1 - \frac{1}{512\kappa^2}\right)\|\mathbf{q}\| \\ &\le \|\mathbf{q}\| - 1.\end{aligned}$$

We conclude that

$$e^{\|q_T\|} \le e^{-1}e^{\|q_0\|},$$

completing our proof of inequality (B.3).

We prove inequality (B.4) by an application of Lemma B.1, which gives

$$\begin{aligned}\|q_s\| &\ge \|q_0\| - \|q_0 - q_s\| \\ &\ge \|\mathbf{q}\| - \frac{1}{2M_2}\big(e^{-\sqrt{M_2}s}\big(e^{\sqrt{M_2}s} - 1\big)\big(\sqrt{M_2}\|\mathbf{p}\|\big(e^{\sqrt{M_2}s} + 1\big) + M_2\|\mathbf{q}\|\big(e^{\sqrt{M_2}s} - 1\big)\big)\big) \\ &\ge \|\mathbf{q}\|\left(1 - \frac{1}{2}\big(e^{\sqrt{M_2}s} - 1\big)\right) - \frac{\|\mathbf{p}\|}{2\sqrt{M_2}}\big(e^{\sqrt{M_2}s} - e^{-\sqrt{M_2}s}\big)\end{aligned}$$

for all $s \ge 0$. We note that the RHS is monotone nonincreasing in s. Thus, setting $s = T$ on the RHS and using the bounds $0 \le \frac{m_2}{M_2} \le 1$ and $\sqrt{M_2}T \le \frac{1}{2\sqrt{2}}$,

$$\begin{aligned}\inf_{0\le s\le T}\|q_s\| &\ge \|\mathbf{q}\|\left(1 - \frac{1}{2}\big(e^{\sqrt{M_2}T} - 1\big)\right) - \frac{\|\mathbf{p}\|}{2\sqrt{M_2}}\big(e^{\sqrt{M_2}T} - e^{-\sqrt{M_2}T}\big) \\ &\ge \|\mathbf{q}\|\left(1 - \frac{1}{2}\big(e^{\frac{1}{2\sqrt{2}}} - 1\big)\right) - \frac{\|\mathbf{p}\|}{2\sqrt{M_2}}\big(e^{\frac{1}{2\sqrt{2}}} - e^{-\frac{1}{2\sqrt{2}}}\big) \\ &\ge \frac{\|\mathbf{q}\|}{2} - \frac{\|\mathbf{p}\|}{2\sqrt{M_2}},\end{aligned}$$

completing the proof. □

We apply this lemma to prove Theorem 4:

PROOF OF THEOREM 4. For $\mathbf{q} \in \mathbb{R}^d$, define the associated "good" set to be

$$\mathcal{G}(\mathbf{q}) = \big\{\mathbf{p} \in \mathbb{R}^d : \|\mathbf{p}\| \le \min\big(c_1\|\mathbf{q}\| - c_3, c_2\|\mathbf{q}\| - c_4\big)\big\},$$

where

$$c_1 = \frac{\sqrt{2}m_2^{2.5}}{512M_2^2}, \qquad c_2 = \max(1, \sqrt{M_2}), \qquad c_3 = \sqrt{2m_2}, \qquad c_4 = 2C\sqrt{M_2}.$$

The condition $\mathbf{p} \in \mathcal{G}(\mathbf{q})$ guarantees that the solution to Hamilton's equations $(q_t, p_t)_{t=0}^T$ with initial conditions $(\mathbf{q}, \mathbf{p})$ will satisfy the following two properties:

- It will stay outside of the set S (by inequality (B.4) of Lemma B.2), *and*
- It will "drift" toward the origin (for a precise statement, see inequality (B.3) of Lemma B.2).

We now make this precise. Define Φ to be the standard Gaussian in d dimensions. Fix $X_0 = x \in \mathbb{R}^d$ as in the statement of the theorem, let $p \sim \Phi$, and set $X_1 = \mathcal{Q}_T^{(X_0)}(p)$; note that X_1 has the distribution required by the statement of the theorem. By Lemma B.2, we have

$$e^{\|X_1\|}\mathbb{1}_{p \in \mathcal{G}(X_0)} \le e^{-1}e^{\|X_0\|}. \tag{B.8}$$

Mimicking the calculation leading to inequality (B.7), we also have for *any* initial position and velocity the deterministic inequality

$$m_2\|X_1\|^2 \le M_2\|X_0\|^2 + \frac{1}{2}\|p\|^2,$$

so that

$$\|X_1\| \le \sqrt{\kappa}\|X_0\| + \frac{1}{\sqrt{2m_2}}\|p\|. \tag{B.9}$$

Combining inequalities (B.8) and (B.9),

$$\begin{aligned}
\mathbb{E}\big[e^{\|X_1\|}|X_0\big] &= \mathbb{E}\big[e^{\|X_1\|}\mathbb{1}_{p \in \mathcal{G}(X_0)}|X_0\big] + \mathbb{E}\big[e^{\|X_1\|}\mathbb{1}_{p \notin \mathcal{G}(X_0)}|X_0\big] \\
&\le e^{-1}e^{\|X_0\|} + \mathbb{E}\big[e^{\|X_1\|}\mathbb{1}_{p \notin \mathcal{G}(X_0)}|X_0\big] \\
&\le e^{-1}e^{\|X_0\|} + e^{\frac{\sqrt{M_2}}{m_2}\|X_0\|}\int_{p \notin \mathcal{G}(X_0)} e^{\frac{1}{\sqrt{2m_2}}\|p\|}\mathrm{d}\Phi(p).
\end{aligned}$$

Defining $A = \sup_{x \in \mathbb{R}^d} e^{\frac{\sqrt{M_2}}{m_2}\|x\|}\int_{p \notin \mathcal{G}(x)} e^{\frac{1}{\sqrt{2m_2}}\|p\|}\mathrm{d}\Phi(p) < \infty$, this implies

$$\mathbb{E}\big[e^{\|X_1\|}|X_0\big] \le e^{-1}e^{\|X_0\|} + A.$$

This completes the proof of inequality (5.1) with no bound on A. We will now bound A. For $R \ge 2\ell$, let $X \sim \Phi_1$. We then have the bound

$$\begin{aligned}
\int_{\|p\|>R} e^{\ell\|p\|}\mathrm{d}\Phi(p) &= \frac{1}{\sqrt{2\pi}^d}\int_{\|x\|>R} e^{\ell\|x\|}e^{-\frac{\|x\|^2}{2}}\mathrm{d}x \\
&\le e^{\frac{\ell^2}{2}}\frac{1}{\sqrt{2\pi}^d}\int_{\|x\|>R} e^{-\frac{(\|x\|-\ell)^2}{2}}\mathrm{d}x \\
&\le e^{\frac{\ell^2}{2}}\frac{1}{\sqrt{2\pi}^d}\int_{\|x\|>R} e^{-\frac{\|x\|^2}{8}}\mathrm{d}x \\
&\le 2^{\frac{d}{2}}e^{\frac{\ell^2}{2}}\mathbb{P}\big[2\|X\| > R\big] \\
&\le 2^{\frac{d}{2}}e^{\frac{\ell^2}{2}}e^{-\frac{R^2}{8}},
\end{aligned} \tag{B.10}$$

where the last inequality holds only for all $R > R_0$ larger than a single universal constant (see, e.g., inequalities (1.2) and (1.3) of [24]). For all ℓ, R (and in particular for $R < 2\ell$), we have the trivial bound

$$\begin{aligned}
\int_{\|p\|>R} e^{\ell\|p\|}\mathrm{d}\Phi(p) &\le \int e^{\ell\|p\|}\mathrm{d}\Phi(p) \\
&\le e^{d\frac{\ell^2}{2}}.
\end{aligned} \tag{B.11}$$

Thus, defining

$$\log\big(B_1(x)\big) = \frac{\sqrt{M_2}}{m_2}x + d + \frac{1}{8m_2} - \frac{1}{8}\left(c_1 x - c_3 - \sqrt{\frac{2}{m_2}}\right)^2,$$

$$\log\big(B_2(x)\big) = \frac{\sqrt{M_2}}{m_2}x + d + \frac{1}{8m_2} - \frac{1}{8}\left(c_2 x - c_4 - \sqrt{\frac{2}{m_2}}\right)^2,$$

$$\log(B_3) = \frac{\sqrt{M_2}}{m_2 c_1}\left(c_3 + \sqrt{\frac{2}{m_2}}\right) + \frac{d}{4m_2},$$

$$\log(B_4) = \frac{\sqrt{M_2}}{m_2 c_2}\left(c_4 + \sqrt{\frac{2}{m_2}}\right) + \frac{d}{4m_2},$$

for $x \in \mathbb{R}^+$, we have by inequalities (B.10) and (B.11)

$$\begin{aligned} A &= \sup_{x \in \mathbb{R}^d} e^{\frac{\sqrt{M_2}}{m_2}\|x\|} \int_{p \notin \mathcal{G}(x)} e^{\frac{1}{\sqrt{2m_2}}\|p\|} \, \mathrm{d}\Phi(p) \\ &\le \max\Big(\sup_{x \in \mathbb{R}^+} \max\big(B_1(x), B_2(x)\big), B_3, B_4\Big). \end{aligned} \tag{B.12}$$

Noting that $\log(B_1(x))$ and $\log(B_2(x))$ are quadratic, we can optimize these expressions by hand. In particular, for a quadratic of the form $f(x) = \alpha x + \beta - (\gamma x - \delta)^2$ for constants $\alpha, \beta, \gamma, \delta > 0$, we have

$$\sup_{x \in \mathbb{R}} f(x) = O\left(\max\left(\beta, \delta^2, \frac{\alpha^2}{\gamma^2}\right)\right).$$

Applying this with inequality (B.12) (and using the relationship $0 < m_2 \le M_2 < \infty$, $d \ge 1$ to remove some terms that cannot possibly be the largest) completes the proof of the bound on A, and thus the proof of the lemma. □

Acknowledgements. We are grateful to Natesh Pillai and Alain Durmus for helpful discussions. Oren Mangoubi was supported by a Canadian Statistical Sciences Institute (CANSSI) Postdoctoral Fellowship, and by an NSERC Discovery grant. Aaron Smith was supported by an NSERC Discovery grant. We are grateful to the anonymous reviewers of an earlier version of this paper for their helpful comments and suggestions.

Funding. Oren Mangoubi was supported by a Canadian Statistical Sciences Institute (CANSSI) Postdoctoral Fellowship, and by an NSERC Discovery grant. Aaron Smith was supported by an NSERC Discovery grant.

REFERENCES

[1] Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M. and Stuart, A. (2013). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli* **19** 1501–1534. MR3129023 https://doi.org/10.3150/12-BEJ414

[2] Biswas, N. and Jacob, P. E. (2019). Estimating convergence of Markov chains with L-lag couplings. Preprint. Available at arXiv:1905.09971.

[3] Borgs, C., Chayes, J. T., Frieze, A., Kim, J. H., Tetali, P., Vigoda, E. and Vu, V. H. (1999). Torpid mixing of some Monte Carlo Markov chain algorithms in statistical physics. In 40*th Annual Symposium on Foundations of Computer Science* (*New York*, 1999) 218–229. IEEE Comput. Soc., Los Alamitos, CA. MR1917562 https://doi.org/10.1109/SFFCS.1999.814594

[4] Bou-Rabee, N., Eberle, A. and Zimmer, R. (2020). Coupling and convergence for Hamiltonian Monte Carlo. *Ann. Appl. Probab.* **30** 1209–1250. MR4133372 https://doi.org/10.1214/19-AAP1528

[5] BOU-RABEE, N. and HAIRER, M. (2013). Nonasymptotic mixing of the MALA algorithm. *IMA J. Numer. Anal.* **33** 80–110. MR3020951 https://doi.org/10.1093/imanum/drs003

[6] BOU-RABEE, N. and SANZ-SERNA, J. M. (2017). Randomized Hamiltonian Monte Carlo. *Ann. Appl. Probab.* **27** 2159–2194. MR3693523 https://doi.org/10.1214/16-AAP1255

[7] BUCHHOLZ, A., CHOPIN, N. and JACOB, P. E. (2018). Adaptive tuning of Hamiltonian Monte Carlo within sequential Monte Carlo. Preprint. Available at arXiv:1808.07730.

[8] CANCÈS, E., LEGOLL, F. and STOLTZ, G. (2007). Theoretical and numerical comparison of some sampling methods for molecular dynamics. *ESAIM Math. Model. Numer. Anal.* **41** 351–389. MR2339633 https://doi.org/10.1051/m2an:2007014

[9] CHEN, Y., DWIVEDI, R., WAINWRIGHT, M. J. and YU, B. (2018). Fast MCMC sampling algorithms on polytopes. *J. Mach. Learn. Res.* **19** 2146–2231.

[10] CHEN, Y., DWIVEDI, R., WAINWRIGHT, M. J. and YU, B. (2020). Fast mixing of metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *J. Mach. Learn. Res.* **21** Paper No. 92. MR4119160

[11] CHEN, Z. and VEMPALA, S. S. (2019). Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions. Preprint. Available at arXiv:1905.02313.

[12] CHENG, X., CHATTERJI, N. S., BARTLETT, P. L. and JORDAN, M. I. (2018). Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on Learning Theory* 300–323.

[13] CHEUNG, S. H. and BECK, J. L. (2009). Bayesian model updating using hybrid Monte Carlo simulation with application to structural dynamic models with many uncertain parameters. *J. Eng. Mech.* **135** 243–255.

[14] DALALYAN, A. S. (2017). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 651–676. MR3641401 https://doi.org/10.1111/rssb.12183

[15] DELIGIANNIDIS, G., PAULIN, D. and DOUCET, A. (2018). Randomized Hamiltonian Monte Carlo as scaling limit of the bouncy particle sampler and dimension-free convergence rates. Preprint. Available at arXiv:1808.04299.

[16] DIACONIS, P. (2009). The Markov chain Monte Carlo revolution. *Bull. Amer. Math. Soc.* (*N.S.*) **46** 179–205. MR2476411 https://doi.org/10.1090/S0273-0979-08-01238-X

[17] DIACONIS, P., KHARE, K. and SALOFF-COSTE, L. (2008). Gibbs sampling, exponential families and orthogonal polynomials. *Statist. Sci.* **23** 151–178. With comments and a rejoinder by the authors. MR2446500 https://doi.org/10.1214/07-STS252

[18] DURMUS, A. and MOULINES, E. (2016). Sampling from strongly log-concave distributions with the unadjusted Langevin algorithm. Preprint. Available at arXiv:1605.01559.

[19] DURMUS, A. and MOULINES, É. (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.* **27** 1551–1587. MR3678479 https://doi.org/10.1214/16-AAP1238

[20] DURMUS, A., MOULINES, E. and SAKSMAN, E. (2017). On the convergence of Hamiltonian Monte Carlo. Preprint. Available at arXiv:1705.00166.

[21] EBERLE, A. (2016). Reflection couplings and contraction rates for diffusions. *Probab. Theory Related Fields* **166** 851–886. MR3568041 https://doi.org/10.1007/s00440-015-0673-1

[22] EBERLE, A. and MAJKA, M. B. (2019). Quantitative contraction rates for Markov chains on general state spaces. *Electron. J. Probab.* **24** 1–36. MR3933205 https://doi.org/10.1214/19-EJP287

[23] GIROLAMI, M. and CALDERHEAD, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 123–214. With discussion and a reply by the authors. MR2814492 https://doi.org/10.1111/j.1467-9868.2010.00765.x

[24] HASHORVA, E. and HÜSLER, J. (2003). On multivariate Gaussian tails. *Ann. Inst. Statist. Math.* **55** 507–522. MR2007795 https://doi.org/10.1007/BF02517804

[25] HENG, J. and JACOB, P. E. (2019). Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika* **106** 287–302. MR3949304 https://doi.org/10.1093/biomet/asy074

[26] HOFFMAN, M., SOUNTSOV, P., DILLON, J. V., LANGMORE, I., TRAN, D. and VASUDEVAN, S. (2019). Neutra-lizing bad geometry in Hamiltonian Monte Carlo using neural transport. Preprint. Available at arXiv:1903.03704.

[27] JONES, G. L. and HOBERT, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statist. Sci.* **16** 312–334. MR1888447 https://doi.org/10.1214/ss/1015346317

[28] KENNEDY, A. D. and PENDLETON, B. (2001). Cost of the generalised hybrid Monte Carlo algorithm for free field theory. *Nuclear Phys. B* **607** 456–510. MR1850796 https://doi.org/10.1016/S0550-3213(01)00129-8

[29] KIRKILIONIS, M. and WALCHER, S. (2004). On comparison systems for ordinary differential equations. *J. Math. Anal. Appl.* **299** 157–173. MR2091278 https://doi.org/10.1016/j.jmaa.2004.06.025

[30] LEE, H., MANGOUBI, O. and VISHNOI, N. K. (2019). Online sampling from log-concave distributions. Available at arXiv:1902.08179.

[31] LEE, Y. T., SHEN, R. and TIAN, K. (2020). Logsmooth gradient concentration and tighter runtimes for metropolized Hamiltonian Monte Carlo. In *Conference on Learning Theory* 2565–2597. PMLR.

[32] LEE, Y. T., SONG, Z. and VEMPALA, S. S. (2018). Algorithmic theory of odes and sampling from well-conditioned logconcave densities. Preprint. Available at arXiv:1812.06243.

[33] LEE, Y. T. and VEMPALA, S. S. (2018). Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. In *STOC'*18*—Proceedings of the* 50*th Annual ACM SIGACT Symposium on Theory of Computing* 1115–1121. ACM, New York. MR3826321 https://doi.org/10.1145/3188745.3188774

[34] LEE, Y. T. and VEMPALA, S. S. (2018). Stochastic localization + Stieltjes barrier = tight bound for log-Sobolev. In *STOC'*18*—Proceedings of the* 50*th Annual ACM SIGACT Symposium on Theory of Computing* 1122–1129. ACM, New York. MR3826322

[35] LEVIN, D. A., PERES, Y. and WILMER, E. L. (2009). *Markov Chains and Mixing Times*. Amer. Math. Soc., Providence, RI. MR2466937 https://doi.org/10.1090/mbk/058

[36] LIVINGSTONE, S., BETANCOURT, M., BYRNE, S. and GIROLAMI, M. (2016). On the geometric ergodicity of Hamiltonian Monte Carlo. Preprint. Available at arXiv:1601.08057.

[37] LOVÁSZ, L. and VEMPALA, S. (2006). Fast algorithms for logconcave functions: Sampling, rounding, integration and optimization. In 2006 47*th Annual IEEE Symposium on Foundations of Computer Science* (*FOCS'*06) 57–68. IEEE.

[38] LOVÁSZ, L. and VEMPALA, S. (2006). Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithm. *J. Comput. System Sci.* **72** 392–417. MR2205290 https://doi.org/10.1016/j.jcss.2005.08.004

[39] MANGOUBI, O., PILLAI, N. S. and SMITH, A. (2018). Does Hamiltonian Monte Carlo mix faster than a random walk on multimodal densities? Preprint. Available at arXiv:1808.03230.

[40] MANGOUBI, O. and SMITH, A. (2017). Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. Preprint. Available at arXiv:1708.07114.

[41] MANGOUBI, O. and SMITH, A. (2018). Rapid mixing of geodesic walks on manifolds with positive curvature. *Ann. Appl. Probab.* **28** 2501–2543. MR3843835 https://doi.org/10.1214/17-AAP1365

[42] MANGOUBI, O. and SMITH, A. (2019). Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions 2: Numerical integrators. In *The* 22*nd International Conference on Artificial Intelligence and Statistics* 586–595.

[43] MANGOUBI, O. and VISHNOI, N. (2018). Dimensionally tight bounds for second-order Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems* 6027–6037.

[44] MANGOUBI, O. and VISHNOI, N. K. (2019). Faster polytope rounding, sampling, and volume computation via a sub-linear ball walk. In 2019 *IEEE* 60*th Annual Symposium on Foundations of Computer Science* 1338–1357. IEEE Comput. Soc. Press, Los Alamitos, CA. MR4228229

[45] MEHLIG, B., HEERMANN, D. W. and FORREST, B. M. (1992). Hybrid Monte Carlo method for condensed-matter systems. *Phys. Rev. E* **45** 679.

[46] NEAL, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 113–162. CRC Press, Boca Raton, FL. MR2858447

[47] OLLIVIER, Y. (2009). Ricci curvature of Markov chains on metric spaces. *J. Funct. Anal.* **256** 810–864. MR2484937 https://doi.org/10.1016/j.jfa.2008.11.001

[48] PIPONI, D. and HOFFMAN, M. D. (2018). Antithetic sampling with Hamiltonian Monte Carlo.

[49] POINCARÉ, H. (1899). *Les Methods Nouvelles de la Mécanique Céleste*. Gauthier-Villars, Paris.

[50] RAGINSKY, M., RAKHLIN, A. and TELGARSKY, M. (2017). Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In *Conference on Learning Theory* 1674–1703.

[51] ROBERTS, G. O. and ROSENTHAL, J. S. (2016). Complexity bounds for Markov chain Monte Carlo algorithms via diffusion limits. *J. Appl. Probab.* **53** 410–420. MR3514287 https://doi.org/10.1017/jpr.2016.9

[52] RUDELSON, M. and VERSHYNIN, R. (2013). Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18** no. 82. MR3125258 https://doi.org/10.1214/ECP.v18-2865

[53] SEILER, C., RUBINSTEIN-SALZEDO, S. and HOLMES, S. (2014). Positive curvature and Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems* 586–594.

[54] SMOLLER, J. (1983). *Shock Waves and Reaction–Diffusion Equations*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Science*] **258**. Springer, New York. MR0688146

[55] VEMPALA, S. (2005). Geometric random walks: A survey. In *Combinatorial and Computational Geometry*. *Math. Sci. Res. Inst. Publ.* **52** 577–616. Cambridge Univ. Press, Cambridge. MR2178341

[56] ZHANG, Y., LIANG, P. and CHARIKAR, M. (2017). A hitting time analysis of stochastic gradient Langevin dynamics. In *Conference on Learning Theory* 1980–2022.