Bernoulli **19**(5A), 2013, 1501–1534
DOI: 10.3150/12-BEJ414

Optimal tuning of the hybrid Monte Carlo algorithm

ALEXANDROS BESKOS[1], NATESH PILLAI[2], GARETH ROBERTS[3], JESUS-MARIA SANZ-SERNA[4] and ANDREW STUART[5]

[1]*Department of Statistical Science, UCL, Gower Street, London, WC1E 6BT, UK.*
E-mail: alex@stats.ucl.ac.uk
[2]*Department of Statistics, Harvard University, Cambridge, MA 02138-2901, USA.*
E-mail: pillai@fas.harvard.edu
[3]*Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK.*
E-mail: gareth.o.roberts@warwick.ac.uk
[4]*Departamento de Matematica Aplicada, Facultad de Ciencias, Universidad de Valladolid, Spain.*
E-mail: sanzsern@mac.uva.es
[5]*Mathematics Institute, University of Warwick, Coventry, CV4 7AL, UK.*
E-mail: a.m.stuart@warwick.ac.uk

We investigate the properties of the hybrid Monte Carlo algorithm (HMC) in high dimensions. HMC develops a Markov chain reversible with respect to a given target distribution Π using separable Hamiltonian dynamics with potential $-\log\Pi$. The additional momentum variables are chosen at random from the Boltzmann distribution, and the continuous-time Hamiltonian dynamics are then discretised using the leapfrog scheme. The induced bias is removed via a Metropolis–Hastings accept/reject rule. In the simplified scenario of independent, identically distributed components, we prove that, to obtain an $\mathcal{O}(1)$ acceptance probability as the dimension d of the state space tends to ∞, the leapfrog step size h should be scaled as $h = l \times d^{-1/4}$. Therefore, in high dimensions, HMC requires $\mathcal{O}(d^{1/4})$ steps to traverse the state space. We also identify analytically the asymptotically optimal acceptance probability, which turns out to be 0.651 (to three decimal places). This value optimally balances the cost of generating a proposal, which *decreases* as l increases (because fewer steps are required to reach the desired final integration time), against the cost related to the average number of proposals required to obtain acceptance, which *increases* as l increases.

Keywords: Hamiltonian dynamics; high dimensions; optimal acceptance probability; leapfrog scheme; squared jumping distance

1. Introduction

The hybrid Monte Carlo (HMC) algorithm originates from the physics literature [15], where it was introduced as a fast method for simulating molecular dynamics. It has since become popular in a number of application areas, including statistical physics [1,17,18,23,43], computational chemistry [22,24,30,42,45], data assimilation [2], geophysics [30] and neural networks [32,46]. The algorithm also has been proposed as a generic tool for Bayesian statistical inference [12,16, 31].

Many practitioners believe that HMC improves on traditional Markov chain Monte Carlo (MCMC) algorithms. There are heuristic arguments to suggest why HMC might perform bet-

ter, in particular based on the idea that it breaks down *random walk-like* behavior intrinsic to many MCMC algorithms, such as the random-walk Metropolis (RWM) algorithm. However, there is very little theoretical understanding of this phenomenon (although see [14]). This lack of theoretical guidance concerning the choice of the free parameters for the algorithm accounts in part for its relative obscurity in statistical applications. The aim of this paper is to provide insight into the behavior of HMC in high dimensions and develop theoretical tools for improving the efficiency of the algorithm.

HMC uses the derivative of the target probability log-density to guide the Monte Carlo trajectory toward areas of high probability. The standard RWM algorithm [29] proposes *local*, symmetric moves around the current position. In many cases (especially in high dimensions), the variance of the proposal must be small for the corresponding acceptance probability to be satisfactory. However smaller proposal variance leads to higher autocorrelations and increased computing time to explore the state space. In contrast, HMC exploits the information on the derivative of the log density to deliver guided, *global* moves, with higher acceptance probability. Thus HMC has the potential to effectively decorrelate by exploiting Hamiltonian evolution, conferring a potential advantage over random-walk-based methods, whose effective decorrelation time is determined by random-walk behavior.

HMC is closely related to the so-called Metropolis-adjusted Langevin algorithm (MALA) [39], which uses the derivative of the log density to propose steepest-ascent moves in the state space. MALA uses *Langevin* dynamics; the proposal is derived from an Euler discretisation of a Langevin stochastic differential equation that leaves the target density invariant. Note that statisticians' use of the term 'Langevin dynamics' refers to the dynamics of a first-order equation, which physicists normally term 'Brownian dynamics'; this model is derived from the second-order Langevin equation in the over-damped limit. Indeed the idea of using such dynamics as a proposal for Monte Carlo predates its appearance in the statistical literature [33,40]. In contrast, HMC uses *Hamiltonian* dynamics. The original variable q is seen as a "location" variable, and an auxiliary "momentum" variable p is introduced. Hamilton's ordinary differential equations are used to generate moves in the enlarged (q, p) phase space. These moves preserve the total energy, a fact that implies, in probability terms, that they preserve the target density Π of the original q variable, provided that the initial momentum is chosen randomly from an appropriate Gaussian distribution. Although seemingly of different origin, MALA can be viewed as a "localised" version of HMC in the case where Hamilton's equations are integrated for only one time step before the accept/reject mechanism is applied [28]. We return to this point later.

In practice, continuous-time Hamiltonian dynamics are discretised by means of a numerical scheme; the popular *Störmer–Verlet* or *leapfrog* scheme [19,25,41,44] is currently the scheme of choice. This integrator does not conserve energy exactly, and the induced bias is corrected via a Metropolis–Hastings accept/reject rule. In this way, HMC develops a Markov chain reversible with respect to Π, whose transitions incorporate information on Π in a natural way.

In this paper, we investigate the properties of HMC in high dimensions and this context offer some guidance regarding the optimal specification of the free parameters of the algorithm. We assume that we wish to sample from a density Π on $\mathbb{R}^N$ with

$$\Pi(Q) = \exp\bigl(-\mathcal{V}(Q)\bigr) \tag{1.1}$$

for $\mathcal{V}:\mathbb{R}^N \to \mathbb{R}$. We study the simplified scenario where $\Pi(Q)$ consists of $d \gg 1$ independent, identically distributed (i.i.d.) vector components,

$$\Pi(Q) = \exp\left(-\sum_{i=1}^{d} V(q_i)\right), \qquad V:\mathbb{R}^m \to \mathbb{R}; \qquad N = m \times d. \tag{1.2}$$

For the leapfrog integrator, we show analytically that under suitable hypotheses on V and as $d \to \infty$, HMC requires $\mathcal{O}(d^{1/4})$ steps to traverse the state space. We also identify the associated optimal acceptance probability.

To be more precise, if h is the step-size used in the leapfrog integrator, then we show that the choice

$$\text{HMC:} \quad h = l \cdot d^{-1/4} \tag{1.3}$$

leads to an average acceptance probability that is of $\mathcal{O}(1)$ as $d \to \infty$ (Theorem 3.6). This implies that $\mathcal{O}(d^{1/4})$ steps are required for HMC to make $\mathcal{O}(1)$ moves in state space. Furthermore we provide a result of perhaps greater practical relevance. We prove that for the leapfrog integrator and as $d \to \infty$, the asymptotically optimal algorithm corresponds to a well-defined value of the acceptance probability, *independent of the particular target* Π, in (1.2). This value is (to three decimal places) 0.651: Theorems 4.1 and 4.2. Thus, when applying HMC in high dimensions, one should attempt to tune the free algorithmic parameters to obtain an acceptance probability close to that value. We give the precise definition of optimality when stating the theorems, but roughly, it is determined by the choice of l that balances the cost of generating a proposal, which *decreases* as l increases (because fewer steps are required to reach the desired final integration time), against the cost related to the average number of proposals required to obtain acceptance, which *increases* as l increases.

The scaling $\mathcal{O}(d^{1/4})$ to make $\mathcal{O}(1)$ moves in state space contrasts favorably with the corresponding scalings $\mathcal{O}(d)$ and $\mathcal{O}(d^{1/3})$ required in a similar context by RWM and MALA, respectively (see the discussion that follows). Furthermore, the full analysis of the leapfrog scheme provided in this paper may be easily extended to high-order, volume-preserving, reversible integrators. For such an integrator, the corresponding scaling would be $\mathcal{O}(d^{1/(2\nu)})$, where ν (an integer) represents the order of the method. For the standard HMC algorithm, previous work has already established the relevance of the choice $h = \mathcal{O}(d^{-1/4})$ (by heuristic arguments; see [18]) and an optimal acceptance probability of around 0.7 (by numerical experiments; see [12]). Our analytic study of the scaling issues in HMC was prompted by those two papers.

We end this discussion with a transparent disclaimer about the range of validity of our optimal scaling results. Our work contains two central assumptions: (i) We work in the setting of an i.i.d. target measure; and (ii) this i.i.d. target is defined via a single potential V, which (see below) is assumed to grow no faster than quadratically at infinity, so that the tails of the distribution are no lighter than Gaussian. Regarding (i), we expect that our results will extend to some problems with a non-product structure, provided that the resulting measure is "close to i.i.d." Examples of such problems have been provided by [4–6] and [8,27,35]. The latter three papers reported that optimal scaling results for RWM and MALA type algorithms extend directly to target measures that have a density with respect to a Gaussian, uniformly as dimension $d \to \infty$. Because a Gaussian

measure is i.i.d. when represented in appropriate coordinates, such measures are indeed close to the i.i.d. case as almost sure properties of the Gaussian measure are inherited by the target measure. However, for all optimal scaling analyses of RWM, MALA and HMC, the extent and manner in which the "close to i.i.d." assumption can be violated, and yet the same optimality criteria apply, remains an open and interesting research question. Regarding (ii), we note that integration of Hamiltonian systems with superquadratic potentials (more precisely superlinear forces) typically requires adaptive time-step integration [41], and that an open and interesting research direction concerns the generalization of HMC algorithms to this situation. We discuss these issues related to possible relaxation of our key assumptions in the concluding section of this paper.

The paper is organized as follows. Section 2 presents the HMC method and reviews the literature concerning scaling issues for the RWM and MALA algorithms. Section 3 studies the asymptotic behavior of HMC as the dimensionality grows, $d \to \infty$, including the key Theorem 3.6. The optimal tuning of HMC is discussed in Section 4, including the key Theorems 4.1 and 4.2. Sections 5 and 6 are technical sections. The former presents the derivation of the required numerical analysis estimates on the leapfrog integrator, with careful attention to the dependence of constants in error estimates on the initial condition. Estimates of this kind are not available in the literature and may be of independent interest. Section 6 gathers the probabilistic proofs. The paper ends with some conclusions and discussion in Section 7.

2. Hybrid Monte Carlo (HMC)

The HMC method has been described from a statistician's perspective in [26]. Here we provide a precise definition of the algorithm, recalling several important concepts from the theory of Hamiltonian dynamics, such as volume preservation, Liouville equation. and reversible integration.[1] Rather than repeat these classical definitions here, we refer the reader to [41].

2.1. Hamiltonian dynamics

Consider the Hamiltonian function

$$\mathcal{H}(Q, P) = \tfrac{1}{2}\langle P, \mathcal{M}^{-1}P\rangle + \mathcal{V}(Q),$$

on $\mathbb{R}^{2N}$, where $\mathcal{M}$ is a symmetric positive definite matrix (the "mass" matrix), Q is the *location* argument, $\mathcal{V}(Q)$ is the potential energy of the system; P is the *momenta*, and $(1/2)\langle P, \mathcal{M}^{-1}P\rangle$ is the kinetic energy. Thus $\mathcal{H}(Q, P)$ gives the total *energy*: the sum of the potential and the kinetic energy. The Hamiltonian dynamics associated with $\mathcal{H}$ are governed by

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = \mathcal{M}^{-1}P, \qquad \frac{\mathrm{d}P}{\mathrm{d}t} = -\nabla\mathcal{V}(Q), \tag{2.1}$$

[1] Here "reversible" has a different meaning from that used in the study of Markov chains.

a system of ordinary differential equations whose solution flow, Φ_t, defined by

$$(Q(t), P(t)) = \Phi_t(Q(0), P(0)),$$

has some key properties relevant to HMC:

1. *Conservation of energy*. The change in the potential becomes kinetic energy; that is, $\mathcal{H} \circ \Phi_t = \mathcal{H}$, for all $t > 0$, that is $\mathcal{H}(\Phi_t(Q(0), P(0))) = \mathcal{H}(Q(0), P(0))$, for all $t > 0$ and all initial conditions $(Q(0), P(0))$.
2. *Conservation of volume*. The volume element, $\mathrm{d}P\,\mathrm{d}Q$, of the phase space is conserved under the mapping Φ_t.
3. *Time reversibility*. If $\mathcal{S}$ denotes the symmetry operator,

$$\mathcal{S}(Q, P) = (Q, -P),$$

then $\mathcal{H} \circ \mathcal{S} = \mathcal{H}$ and

$$\mathcal{S} \circ (\Phi_t)^{-1} \circ \mathcal{S} = \Phi_t. \tag{2.2}$$

Thus, changing the sign of the initial velocity, evolving backward in time, and changing the sign of the final velocity reproduces the forward evolution.

From the Liouville equation for (2.1), it follows that if the initial conditions are distributed according to a probability measure with Lebesgue density depending only on $\mathcal{H}(Q, P)$, then this probability measure is preserved by the Hamiltonian flow Φ_t. In particular, if the initial conditions $(Q(0), P(0))$ of (2.1) are distributed with a density (proportional to, given that we omit the normalizing constant for the Gaussian part)

$$\exp\bigl(-\mathcal{H}(Q, P)\bigr) = \exp\bigl(-(1/2)\langle P, \mathcal{M}^{-1} P\rangle\bigr) \exp\bigl(-\mathcal{V}(Q)\bigr),$$

then for all $t > 0$, the marginal density of $Q(t)$ will also be $\exp(-\mathcal{V}(Q))$. This suggests that integration of equations (2.1) might form the basis for an exploration of the target density $\exp(-\mathcal{V}(Q))$.

2.2. The HMC algorithm

To formulate a practical algorithm, the continuous-time dynamics (2.1) must be discretised. The most popular *explicit* method for doing this is the Störmer–Verlet, or leapfrog, scheme (see [19, 25,41] and references therein), defined as follows. Assume a current state (Q_0, P_0). After one step of length $h > 0$, the system (2.1) will be at a state (Q_h, P_h) defined by the following three-stage procedure:

$$P_{h/2} = P_0 - \frac{h}{2}\nabla\mathcal{V}(Q_0); \tag{2.3a}$$

$$Q_h = Q_0 + h\mathcal{M}^{-1} P_{h/2}; \tag{2.3b}$$

$$P_h = P_{h/2} - \frac{h}{2}\nabla\mathcal{V}(Q_h). \tag{2.3c}$$

Table 1. Markov transition for the HMC algorithm. Iterative application for a given starting location Q^0, will yield a Markov chain $Q^0, Q^1, \ldots$

HMC(Q):

(i) Sample a momentum $P \sim N(0, \mathcal{M})$.

(ii) Accept the proposed update Q' defined via $(Q', P') = \Psi_h^{(T)}(Q, P)$ w.p.:

$$a((Q, P), (Q', P')) := 1 \wedge \exp\{\mathcal{H}(Q, P) - \mathcal{H}(Q', P')\}.$$

The scheme gives rise to a map,

$$\Psi_h : (Q_0, P_0) \mapsto (Q_h, P_h),$$

which approximates the flow Φ_h. The solution at time T is approximated by taking $\lfloor \frac{T}{h} \rfloor$ leapfrog steps,

$$(Q(T), P(T)) = \Phi_T(Q(0), P(0)) \approx \Psi_h^{\lfloor T/h \rfloor}(Q(0), P(0)).$$

Note that this is a *deterministic* computation. The map

$$\Psi_h^{(T)} := \Psi_h^{\lfloor T/h \rfloor}$$

may be shown to be volume-preserving and time-reversible (see [19,25,41]), but it does not exactly conserve energy. As a consequence, the leapfrog algorithm does not share the property of equations (2.1) following from the Liouville equation, namely that the probability density function proportional to $\exp(-\mathcal{H}(Q, P))$ is preserved. Restoring this property requires the addition of an accept/reject step must. The work of [31] provides a clear derivation of the required acceptance criterion.

We can now describe the complete HMC algorithm. Let the current state be Q. The next state for the HMC Markov chain is determined by the dynamics described in Table 1.

Given the time reversibility and volume conservation properties of the integrator map $\Psi_h^{(T)}$, the algorithm in Table 1 defines (see [15,31]) a Markov chain reversible with respect to $\Pi(Q)$. Sampling this chain up to equilibrium will provide correlated samples Q^n from $\Pi(Q)$. Note that the momentum P is merely an auxiliary variable, and that the user of the algorithm is free to choose h, T and the mass matrix $\mathcal{M}$. In this paper, we concentrate on the optimal choice of h for high-dimensional targets.

2.3. Connection with other Metropolis–Hastings algorithms

Previous research has studied the optimal tuning of other Metropolis–Hastings algorithms, namely RWM and MALA. In contrast with HMC, whose proposals involve a deterministic element, those algorithms use purely stochastic updates. For the target density $\Pi(Q)$ in (1.1), RWM is specified through the proposed update

$$Q' = Q + \sqrt{h}Z,$$

with $Z \sim N(0, I)$. (This simple case suffices for our exposition, but note that Z may be allowed to have an arbitrary mean zero distribution.) In contrast, MALA is determined through the proposal

$$Q' = Q + \frac{h}{2}\nabla \log \Pi(Q) + \sqrt{h}Z.$$

The density Π is invariant for both algorithms when the proposals are accepted with probability

$$a(Q, Q') = 1 \wedge \frac{\Pi(Q')T(Q', Q)}{\Pi(Q)T(Q, Q')},$$

where

$$T(x, y) = \mathrm{P}[Q' \in \mathrm{d}y | Q = x]/\mathrm{d}y$$

is the transition density of the proposed update. [Note that for RWM, the symmetry of the proposal implies that $T(Q, Q') = T(Q', Q)$.]

The proposed distribution for MALA corresponds to the Euler discretization of the stochastic differential equation (SDE)

$$\mathrm{d}Q = \tfrac{1}{2}\nabla \log \Pi(Q)\,\mathrm{d}t + \mathrm{d}W$$

for which Π is an invariant density. (Here W denotes a standard multivariate Brownian motion with covariance I.) Whether HMC and MALA are connected can be easily checked, because HMC reduces to MALA when $T \equiv h$, that is, when the algorithm makes only a single leapfrog step at each transition of the chain.

Assume now that RWM and MALA are applied with the scalings

$$\text{RWM:} \quad h = l \cdot d^{-1}, \qquad \text{MALA:} \quad h = l \cdot d^{-1/3} \tag{2.4}$$

for some constant $l > 0$, in the simplified scenario where the target Π has the i.i.d. structure (1.2) with $m = 1$. Previous authors ([36,37]) have shown that as $d \to \infty$ and under regularity conditions on V (i.e., the function V must be seven times differentiable,[2] with all derivatives having polynomial growth bounds, and all moments of $\exp(-V)$ finite, the acceptance probability approaches a nontrivial value,

$$\mathbb{E}[a(Q, Q')] \to a(l) \in (0, 1).$$

The limit $a(l)$ is different for the two algorithms. Furthermore, if $q_1^0, q_1^1, \ldots$ denotes the projection of the trajectory $Q^0, Q^1, \ldots$ onto its first coordinate, in the foregoing scenario it can be shown ([36,37]) that the continuous-time interpolation,

$$\text{RWM:} \quad t \mapsto q_1^{[t \cdot d]}, \qquad \text{MALA:} \quad t \mapsto q_1^{[t \cdot d^{1/3}]} \tag{2.5}$$

(with $[x]$ denoting the integer part of $x \in \mathbb{R}$) converges to the diffusion process governed by the SDE

$$\mathrm{d}q = -\tfrac{1}{2}la(l)V'(q)\,\mathrm{d}t + \sqrt{la(l)}\,\mathrm{d}w, \tag{2.6}$$

[2]This is mostly a technical requirement which may be relaxed.

where w represents a standard Brownian motion. In view of (2.4), (2.5) and (2.6), we deduce that the RWM and MALA algorithms cost $\mathcal{O}(d^2)$ and $\mathcal{O}(d^{4/3})$, respectively, to explore the invariant measure in stationarity for product measures where the cost of each step of the algorithm is $\mathcal{O}(d)$ (because m is fixed and $d \to \infty$). Furthermore, as the product $la(l)$ determines the *speed* of the limiting diffusion, the state space will be explored more rapidly for the choice l_{opt} of l that maximizes $la(l)$. Whereas l_{opt} depends on the target distribution, it turns out that the optimal acceptance probability $a(l_{\text{opt}})$ is independent of V. In fact, with three decimal places, we find that

$$\text{RWM:} \quad a(l_{\text{opt}}) = 0.234, \qquad \text{MALA:} \quad a(l_{\text{opt}}) = 0.574.$$

Asymptotically, as $d \to \infty$, this analysis identifies algorithms that may be considered *uniformly* optimal, given that (as discussed in [38]) ergodic averages of trajectories corresponding to $l = l_{\text{opt}}$ provide optimal estimation of expectations $\mathbb{E}[f(q)]$, $q \sim \exp(-V)$, irrespective of the choice of the (regular) function f. These investigations of the optimal tuning of RWM and MALA have been subsequently extended by [8] and [10] to non-product target distributions.

For HMC, we show that the scaling (1.3) leads to an average acceptance probability of $\mathcal{O}(1)$ and hence to a cost of $\mathcal{O}(d^{5/4})$ to make the $\mathcal{O}(1)$ moves necessary to explore the (product) invariant measure. However, in contrast to RWM and MALA, for HMC we are not able to provide a simple description of the limiting dynamics of a single coordinate of the Markov chain. Consequently, optimality is harder to define.

3. Hybrid Monte Carlo in the limit $d \to \infty$

Our primary aim in this section is to prove Theorem 3.6 concerning the scaling of the step size h in HMC. We also provide some insight into the limiting behavior of the resulting Markov chain under this scaling in Propositions 3.8 and 3.9.

3.1. HMC in the i.i.d. scenario

Here we study the asymptotic behavior of the HMC algorithm in the i.i.d. scenario (1.2), when the number d of 'particles' goes to infinity. Given such a scenario for our target, a global $m \times d$-dimensional implementation of the Hamiltonian dynamics (2.1), or indeed of the practical leapfrog scheme (2.3), can now be decomposed into d independent implementations along each of the identical m-dimensional constituent components (assuming that the auxiliary variable P is chosen to have a similar i.i.d. structure). We exploit this simplified structure in our analysis.

We write $Q = (q_i)_{i=1}^d$ and $P = (p_i)_{i=1}^d$ to distinguish the individual components, and use the following notation for the combination location/momentum:

$$X = (x_i)_{i=1}^d; \qquad x_i := (q_i, p_i) \in \mathbb{R}^{2m}.$$

We denote by $\mathcal{P}_q$ and $\mathcal{P}_p$ the projections onto the position and momentum components of x, that is, $\mathcal{P}_q(q, p) = q$, $\mathcal{P}_p(q, p) = p$. We then have

$$\mathcal{H}(Q, P) = \sum_{i=1}^{d} H(q_i, p_i); \qquad H(q, p) := \frac{1}{2}\langle p, M^{-1}p\rangle + V(q) - \log(c),$$

where M is a $m \times m$ symmetric, positive definite matrix and $c > 0$ is the normalizing constant for the Gaussian part, that is, $c^{-1} = \int \mathrm{e}^{-(1/2)\langle p, M^{-1}p\rangle}\,\mathrm{d}p$. We include this here only to avoid the repeated use of a normalizing constant in the mathematical expressions for expectations used below. Of course, HMC only uses differences in the energy $H(q, p)$ or its derivatives, and thus does not require normalizing constants under the distribution of p or q. The Hamiltonian differential equations for a single (m-dimensional) particle are

$$\frac{\mathrm{d}q}{\mathrm{d}t} = M^{-1}p, \qquad \frac{\mathrm{d}p}{\mathrm{d}t} = -\nabla V(q), \tag{3.1}$$

where $V : \mathbb{R}^m \to \mathbb{R}$. We denote the corresponding flow by φ_t and the leapfrog solution operator over one h step by ψ_h. Thus the acceptance probability for the evolution of the d particles is given by (see Table 1)

$$a(X, Y) = 1 \wedge \exp\left(\sum_{i=1}^{d}\left[H(x_i) - H\left(\psi_h^{(T)}(x_i)\right)\right]\right), \tag{3.2}$$

with $Y = (y_i)_{i=1}^d = \Psi_h^{(T)}(X)$ denoting the HMC proposal.

As mentioned earlier, in the i.i.d. scenario, the leapfrog scheme (2.3) disentangles into independent implementations for each of the d particles (q_i, p_i), with the different particles connected only through the accept/reject decision (3.2).

3.2. Energy increments

Our first aim is to estimate (in an analytical sense) the exponent on the right-hand side of (3.2). Because the d particles play the same role, studying a single term $H(x_i) - H(\psi_h^{(T)}(x_i))$ is sufficient. We set

$$\Delta(x, h) := H\left(\psi_h^{(T)}(x)\right) - H(\varphi_T(x)) = H\left(\psi_h^{(T)}(x)\right) - H(x). \tag{3.3}$$

This is the energy change due to the leapfrog scheme over $0 \le t \le T$, with step size h and initial condition x, which, by conservation of energy under the true dynamics, is simply the energy error at time T. We study the first and second moments,

$$\mu(h) := \mathbb{E}[\Delta(x, h)] = \int_{\mathbb{R}^{2m}} \Delta(x, h)\mathrm{e}^{-H(x)}\,\mathrm{d}x,$$

$$s^2(h) := \mathbb{E}[\Delta(x, h)^2],$$

and the corresponding variance,

$$\sigma^2(h) = s^2(h) - \mu^2(h).$$

If the integrator were exactly energy-preserving, then we would have $\Delta \equiv 0$, and all proposals would be accepted. However, as is well known, the size of $\Delta(x, h)$ generally is no better than the size of the integration error $\psi_h^{(T)}(x) - \varphi_T(x)$, that is, $\mathcal{O}(h^2)$. In fact, under natural smoothness assumptions on V, the following condition holds (see Section 5 for a proof):

Condition 3.1. *There exist functions $\alpha(x)$, $\rho(x, h)$ such that*

$$\Delta(x, h) = h^2\alpha(x) + h^2\rho(x, h) \tag{3.4}$$

with $\lim_{h\to 0} \rho(x, h) = 0$.

In the proofs of our theorems, we use an additional condition to control the variation of Δ as a function of x. In Section 5 we show that this condition holds under suitable assumptions on the growth of V and its derivatives.

Condition 3.2. *There exists a function $D : \mathbb{R}^{2m} \to \mathbb{R}$ such that*

$$\sup_{0\le h\le 1} \frac{|\Delta(x, h)|^2}{h^4} \le D(x),$$

with

$$\int_{\mathbb{R}^{2m}} D(x)\mathrm{e}^{-H(x)}\,\mathrm{d}x < \infty.$$

A key to the proof of Theorem 3.6 is the fact that the average energy increment scales as $\mathcal{O}(h^4)$. We show this in Proposition 3.4 using the following simple lemma, which holds for general volume-preserving, time-reversible integrators:

Lemma 3.3. *Let $\psi_h^{(T)}$ be any volume-preserving, time-reversible numerical integrator of the Hamiltonian equations (3.1), and let $\Delta(x, h) : \mathbb{R}^{2m} \times \mathbb{R}_+ \to \mathbb{R}$ be as in (3.3). If $g : \mathbb{R} \to \mathbb{R}$ is an odd function, then*

$$\int_{\mathbb{R}^{2m}} g(\Delta(x, h))\mathrm{e}^{-H(x)}\,\mathrm{d}x = -\int_{\mathbb{R}^{2m}} g(\Delta(x, h))\mathrm{e}^{-H(\psi_h^{(T)}(x))}\,\mathrm{d}x,$$

provided that at least one of the foregoing integrals exists. If g is an even function, then

$$\int_{\mathbb{R}^{2m}} g(\Delta(x, h))\mathrm{e}^{-H(x)}\,\mathrm{d}x = \int_{\mathbb{R}^{2m}} g(\Delta(x, h))\mathrm{e}^{-H(\psi_h^{(T)}(x))}\,\mathrm{d}x,$$

provided that at least one of the foregoing integrals exists.

Proof. See Section 6. □

Applying this lemma with $g(u) = u$, we obtain

$$\mu(h) = -\int_{\mathbb{R}^{2m}} \Delta(x,h)\mathrm{e}^{-H(\psi_h^{(T)}(x))}\,\mathrm{d}x,$$

which implies that

$$2\mu(h) = \int_{\mathbb{R}^{2m}} \Delta(x,h)\big[1-\exp\big(-\Delta(x,h)\big)\big]\mathrm{e}^{-H(x)}\,\mathrm{d}x. \tag{3.5}$$

We now use the inequality $|\mathrm{e}^u - 1| \le |u|(\mathrm{e}^u + 1)$ and then Lemma 3.3 with $g(u) = u^2$ to conclude that

$$\begin{aligned} |2\mu(h)| &\le \int_{\mathbb{R}^{2m}} |\Delta(x,h)|^2\mathrm{e}^{-H(\psi_h^{(T)}(x))}\,\mathrm{d}x + \int_{\mathbb{R}^{2m}} |\Delta(x,h)|^2\mathrm{e}^{-H(x)}\,\mathrm{d}x \\ &\le 2\int_{\mathbb{R}^{2m}} |\Delta(x,h)|^2\mathrm{e}^{-H(x)}\,\mathrm{d}x = 2s^2(h). \end{aligned} \tag{3.6}$$

The bound in (3.6) is important; it shows that the average of $\Delta(x,h)$ is actually of the order of (the average of) $\Delta(x,h)^2$. Given the second-order leapfrog scheme $\Delta(x,h) = \mathcal{O}(h^2)$, (3.6) shows that we may expect the average $\mu(h)$ to actually behave as $\mathcal{O}(h^4)$. This is made precise in the following theorem.

Proposition 3.4. *If the potential V is such that Conditions 3.1 and 3.2 hold for the leapfrog integrator $\psi_h^{(T)}$, then*

$$\lim_{h\to 0}\frac{\mu(h)}{h^4} = \mu, \qquad \lim_{h\to 0}\frac{\sigma^2(h)}{h^4} = \Sigma$$

for the constants

$$\Sigma = \int_{R^{2m}} \alpha^2(x)\mathrm{e}^{-H(x)}\,\mathrm{d}x; \qquad \mu = \Sigma/2.$$

Proof. See Section 6. □

We next perform explicit calculations for the example of the harmonic oscillator and verify (for this case) the conclusions of Proposition 3.4.

Example 3.5 (Harmonic oscillator). Consider the Hamiltonian

$$H(q,p) = \tfrac{1}{2}p^2 + \tfrac{1}{2}q^2$$

which gives rise to the system

$$\begin{pmatrix} \mathrm{d}q/\mathrm{d}t \\ \mathrm{d}p/\mathrm{d}t \end{pmatrix} = \begin{pmatrix} p \\ -q \end{pmatrix},$$

with solutions

$$\begin{pmatrix} q(t) \\ p(t) \end{pmatrix} = \begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix} \begin{pmatrix} q(0) \\ p(0) \end{pmatrix}.$$

In this case, the leapfrog integration can be written as

$$\psi_h = \psi_h(q, p) = \begin{pmatrix} 1 - h^2/2 & h \\ -h + h^3/4 & 1 - h^2/2 \end{pmatrix} \begin{pmatrix} q \\ p \end{pmatrix} = \Xi \begin{pmatrix} q \\ p \end{pmatrix}.$$

Accordingly, the numerical solution after $\lfloor \frac{1}{h} \rfloor$ steps is given by

$$\psi_h^{(1)}(q, p) = \Xi^{\lfloor 1/h \rfloor} \begin{pmatrix} q \\ p \end{pmatrix}.$$

Diagonalizing Ξ and exponentiating yields

$$\Xi^n = \begin{pmatrix} \cos(\theta n) & \dfrac{1}{\sqrt{1 - h^2/4}} \sin(\theta n) \\ -\sqrt{1 - h^2/4} \sin(\theta n) & \cos(\theta n) \end{pmatrix},$$

where $\theta = \cos^{-1}(1 - h^2/2)$. Using, for instance, MATHEMATICA, we can now obtain the Taylor expansion,

$$\Delta(x, h) = H\big(\psi_h^{(1)}(x)\big) - H(x) = h^2 \alpha(x) + h^4 \beta(x) + \mathcal{O}(h^6),$$

where

$$\alpha(q, p) = \big((p^2 - q^2)\sin^2(1) + pq \sin(2)\big)/8;$$

$$\beta(q, p) = \big(-q^2 \sin(2) + pq\big(2\cos(2) + 3\sin(2)\big) + p^2\big(3 - 3\cos(2) + \sin(2)\big)\big)/192.$$

Note that in the stationary regime, q and p are standard normal variables. Therefore, the expectation of $\alpha(x)$ is 0. Tedious calculations give

$$\mathrm{Var}[\alpha(x)] = \tfrac{1}{16} \sin^2(1), \qquad \mathbb{E}[\beta(x)] = \tfrac{1}{32} \sin^2(1),$$

in agreement with Proposition 3.4.

3.3. Expected acceptance probability

We are now in a position to identify the scaling for h that gives nontrivial acceptance probability as $d \to \infty$.

Theorem 3.6. *Assume that the potential V is such that the leapfrog integrator $\psi_h^{(T)}$ satisfies Conditions 3.1 and 3.2 and that*

$$h = l \cdot d^{-1/4} \tag{3.7}$$

for a constant $l > 0$. Then in stationarity, that is, for $X \sim \exp(-\mathcal{H})$ and $Y = \Psi_h^{(T)}(X)$,

$$\lim_{d\to\infty} \mathbb{E}[a(X,Y)] = 2\Phi\big(-l^2\sqrt{\Sigma}/2\big) =: a(l),$$

where the constant Σ is as defined in Proposition 3.4.

Proof. To grasp the main idea, note that the acceptance probability (3.2) is given by

$$a(X,Y) = 1 \wedge \mathrm{e}^{R_d}; \qquad R_d = -\sum_{i=1}^{d} \Delta(x_i, h). \tag{3.8}$$

Due to the simple structure of the target density and stationarity, the terms $\Delta(x_i, h)$ being added in (3.8) are i.i.d. random variables. Since the expectation and variance of $\Delta(x, h)$ are both $\mathcal{O}(h^4)$ and we have d terms, the natural scaling to obtain a distributional limit is given by (3.7). Then $R_d \approx N(-\frac{1}{2}l^4\Sigma, l^4\Sigma)$ and the desired result follows. See Section 6 for a detailed proof. □

In Theorem 3.6, the limit acceptance probability arises from the use of the central limit theorem. If Condition 3.2 is not satisfied and $\sigma^2(h) = \infty$, then a Gaussian limit is not guaranteed and it may be necessary to consider a different scaling to obtain a heavy-tailed limiting distribution such as, say, a stable law.

The scaling (3.7) is a direct consequence of the fact that the leapfrog integrator has second-order accuracy. Arguments similar to those used earlier prove that the use of a volume-preserving, symmetric νth-order integrator would result in a scaling $h = \mathcal{O}(d^{-1/(2\nu)})$ (ν is an even integer) to obtain an acceptance probability of $\mathcal{O}(1)$.

3.4. Displacement of one particle in a transition

We now turn our attention to the displacement $q_1^{n+1} - q_1^n$ of a single particle in a transition $n \to n+1$ of the chain. Note that, clearly,

$$q_1^{n+1} = I^n \cdot \mathcal{P}_q \psi_h^{(T)}(q_1^n, p_1^n) + (1 - I^n) q_1^n; \qquad I^n = \mathbb{I}_{U^n \le a(X^n, Y^n)}, \tag{3.9}$$

with U_n having a uniform distribution in $[0,1]$. Although Conditions 3.1 and 3.2 refer to the error in energy, the proof of the next results requires a condition on the leapfrog integration error in the dynamic variables q and p. In Section 5, we describe conditions on V that guarantee the fulfillment of this condition.

Condition 3.7. *There exists a function $E : \mathbb{R}^{2m} \to \mathbb{R}$ such that*

$$\sup_{0\le h\le 1} \frac{|\psi_h^{(T)}(x) - \varphi_T(x)|}{h^2} \le E(x),$$

with

$$\int_{\mathbb{R}^{2m}} E(x)^4 \mathrm{e}^{-H(x)}\,\mathrm{d}x < \infty.$$

Under the scaling (3.7) and at stationarity, the second moment $\mathbb{E}[(q_1^{n+1} - q_1^n)^2]$ will also approach a nontrivial limit.

Proposition 3.8. *Assume that the hypotheses of Theorem 3.6 and Condition 3.7 hold and also that the density* $\exp(-V(q))$ *has finite fourth moments. Then, in stationarity,*

$$\lim_{d\to\infty} \mathbb{E}[(q_1^{n+1} - q_1^n)^2] = C_J \cdot a(l),$$

where the value of the constant C_J *is given by*

$$C_J = \mathbb{E}\big[\big(\mathcal{P}_q \varphi_T(q,p) - q\big)^2\big]; \qquad (q,p) \sim \exp\big(-H(q,p)\big).$$

Proof. See Section 6. □

Note that the computational work required to integrate up to a fixed time T is inversely proportional to the parameter l. Thus Proposition 3.8 suggests that it is reasonable to choose a value for l that maximizes the quantity $a(l)l$. This choice of l is optimal in the sense that it seeks a middle ground between smaller values of l, which lead to a higher acceptance probability (and thus larger mean squared jumps) but need more computational work, and large values of l, which have a smaller acceptance probability (and hence smaller mean squared jumps) but require less computational resources. In Section 4, we expand this idea, define a precise notion of optimality that encodes this trade-off, and derive the resulting optimal acceptance probability.

3.5. The limit dynamics

We now discuss the limiting dynamics of the Markov chain, under the same assumptions as in Proposition 3.8. For HMC (as for RWM or MALA), the marginal process $\{q_1^n\}_{n\geq 0}$ is not Markovian with respect to its own filtration, because its dynamics depend on the current position of all d particles via the acceptance probability $a(X^n, Y^n)$ (see (3.9)). In the case of MALA and RWM, $\{q_1^n\}_{n\geq 0}$ is *asymptotically* Markovian: as $d \to \infty$ the effect of the rest of the particles are averaged to a constant via the strong law of large numbers. This allows for the interpolants of (2.5) to converge to solutions of the SDE (2.6), which defines a Markov process. We now argue that for HMC, $\{q_1^n\}_{n\geq 0}$ cannot be expected to be *asymptotically* Markovian. To simplify the exposition, we do not present all of the technicalities of the argument that follows.

It is well known (see, e.g., [44]) that, due to time-reversibility and under suitable smoothness assumptions on V, the energy increments of the leapfrog integrator may be expanded in even powers of h as follows (cf. (3.4)):

$$\Delta(x,h) = h^2\alpha(x) + h^4\beta(x) + \mathcal{O}(h^6).$$

Necessarily $\mathbb{E}[\alpha(x)] = 0$, because from Proposition 3.4 we know that $\mathbb{E}[\Delta(x,h)] = \mathcal{O}(h^4)$. Ignoring $\mathcal{O}(h^6)$-terms, we can write

$$a(X^n, Y^n) = 1 \wedge \mathrm{e}^{R^n_{1,d}+R^n_{2,d}}$$

with

$$R^n_{1,d} = -h^2 \sum_{i=1}^{d} \{\alpha(x_i^n) - \mathbb{E}[\alpha(x_i^n)|q_i^n]\} - h^4 \sum_{i=1}^{d} \beta(x_i^n),$$

$$R^n_{2,d} = -h^2 \sum_{i=1}^{d} \mathbb{E}[\alpha(x_i^n)|q_i^n].$$

Under appropriate conditions, $R^n_{1,d}$ converges, as $d \to \infty$, to a Gaussian limit independent of the σ-algebra $\sigma(q_1^n, q_2^n, \ldots)$. To see that, note that, due to the strong law of large numbers and because $h^4 = l^4/d$, the second sum in $R^n_{1,d}$ converges a.s. to a constant. *conditionally* on $\sigma(q_1^n, q_2^n, \ldots)$, the distributional limit of the first term in $R^n_{1,d}$ is Gaussian with mean 0 and variance determined by the a.s. constant limit of $h^4 \sum_{i=1}^{d} \{\alpha(x_i^n) - \mathbb{E}[\alpha(x_i^n)|q_i^n]\}^2$; this follows from the Martingale central limit theorem (see, e.g., Theorem 3.2 of [21]). On the other hand, the limit distribution of $R^n_{2,d}$ is Gaussian with mean 0 but in general cannot be asymptotically independent of $\sigma(q_1^n, q_2^n, \ldots)$. Instead, it seems that $R^n_{2,d}$ will result in a quantity appearing in the acceptance probability that is nontrivial as $d \to \infty$ and makes it impossible to have a Markovian limit for the trajectory of q_1. In the case of RWM or MALA, the conditional expectations that fulfill the role played here by $\mathbb{E}[\alpha(x_i^n)|q_i^n]$ are identically 0 (see the expansions for the acceptance probability in [36] and [37]), which implies that the corresponding acceptance probabilities are asymptotically independent from $\sigma(q_1^n, q_2^n, \ldots)$, and that the marginal processes $\{q_1^n\}_{n\geq 0}$ are asymptotically Markovian.

The final result in this section provides insight into the limit dynamics of $\{q_1^n\}_{n\geq 0}$.

Proposition 3.9. *Let* $Q^n \sim \Pi(Q)$. *Define*

$$\mathsf{q}_1^{n+1} = \mathsf{I}^n \cdot \mathcal{P}_q \varphi_T(q_1^n, p_1^n) + (1 - \mathsf{I}^n) q_1^n; \qquad \mathsf{I}^n = \mathbb{I}_{U^n \leq a(l)},$$

and consider q_1^{n+1} *in* (3.9). *Then, under the hypotheses of Proposition* 3.8, *as* $d \to \infty$,

$$(q_1^n, q_1^{n+1}) \xrightarrow{\mathcal{L}} (q_1^n, \mathsf{q}_1^{n+1}).$$

Proof. See Section 6. □

This proposition provides a simple description of the asymptotic behavior of the one-transition dynamics of the marginal trajectories of HMC. As $d \to \infty$, with probability $a(l)$, the HMC particle moves under the *correct* Hamiltonian dynamics. However, because of the energy errors accumulated from the leapfrog integration of all d particles, the deviation from the true Hamiltonian dynamics gives rise to the alternative event of staying at the current position q^n, with probability $1 - a(l)$.

4. Optimal tuning of HMC

In the previous section, we addressed the question of how to scale the step size in the leapfrog integration in terms of the dimension d, leading to Theorem 3.6. In this section we refine this analysis and study the choice of constant l in (3.7). Regardless of the metrics used to measure the efficiency of the algorithm, a good choice of l in (3.7) must balance the amount of work needed to simulate a full T-leg (interval of length T) of the Hamiltonian dynamics and the probability of accepting the resulting proposal. Increasing l decreases the acceptance probability but also decreases the computational cost of each T-leg integration; decreasing l will yield the opposite effects, suggesting an optimal value of l. In this section we present an analysis that avoids the complex calculations typically associated with the estimation of mixing times of Markov chains, but still provides useful guidance regarding the choice of l. We provide two alternative ways of doing this, summarized in Theorems 4.1 and 4.2.

4.1. Asymptotically optimal acceptance probability

Clearly, the number of leapfrog steps of length h needed to compute a proposal is $\lceil T/h \rceil$. Furthermore, at each step of the chain, it is necessary to evaluate $a(X, Y)$ and sample P. Thus the computing time for a single proposal will be

$$\mathsf{C}_{l,d} := \left\lceil \frac{T d^{1/4}}{l} \right\rceil \cdot d \cdot C_{\mathrm{LF}} + d \cdot C_{\mathrm{O}} \tag{4.1}$$

for some constants C_{LF}, C_{O} that measure for one particle the leapfrog costs and overheads. Let $\mathsf{E}_{l,d}$ denote the expected computing time until the first accepted T-leg in stationarity. Recall that Q denotes the vector of positions within the Hamiltonian model so that $X = (Q, P)$. If N denotes the number of proposals up to (and including) the first to be accepted, then

$$\mathsf{E}_{l,d} = \mathsf{C}_{l,d}\mathbb{E}[\mathsf{N}] = \mathsf{C}_{l,d}\mathbb{E}[\mathbb{E}[\mathsf{N}|Q]] = \mathsf{C}_{l,d}\mathbb{E}\left[\frac{1}{\mathbb{E}[a(X, Y)|Q]}\right].$$

Here we have used the fact that, given the locations Q, the number of proposed T-legs follows a geometric distribution with probability of success $\mathbb{E}[a(X, Y)|Q]$. Jensen's inequality yields

$$\mathsf{E}_{l,d} \geq \frac{\mathsf{C}_{l,d}}{\mathbb{E}[a(X, Y)]} =: \mathsf{E}^*_{l,d}, \tag{4.2}$$

and, from (4.1) and Theorem 3.6, we conclude that

$$\lim_{d\to\infty} d^{-5/4} \times \mathsf{E}^*_{l,d} = \frac{T C_{\mathrm{LF}}}{a(l)l}.$$

A sensible choice for l minimizes the asymptotic cost $\mathsf{E}^*_{l,d}$, that is,

$$l_{\mathrm{opt}} = \arg\max_{l>0} \mathrm{eff}(l); \qquad \mathrm{eff}(l) := a(l)l.$$

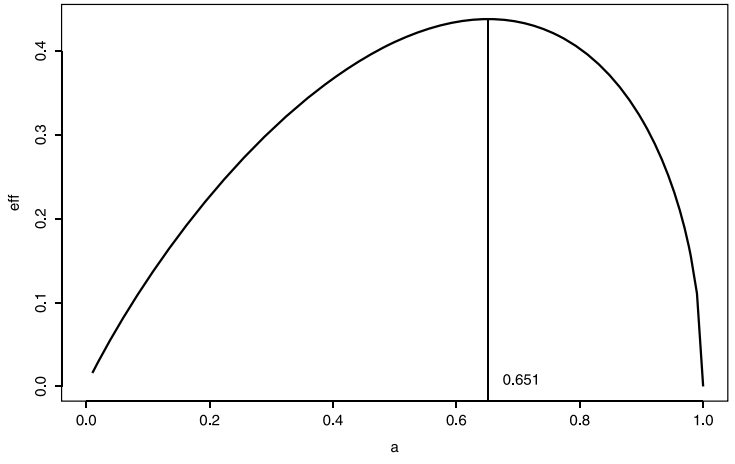

Figure 1. The efficiency function $\text{eff} = \text{eff}(a)$.

The value of l_{opt} generally will depend on the specific target distribution under consideration. However, by expressing eff as a function of $a = a(l)$, we may write

$$\text{eff} = \left(\frac{\sqrt{2}}{\Sigma^{1/4}}\right) \cdot a \cdot \left(\Phi^{-1}\left(1 - \frac{a}{2}\right)\right)^{1/2}, \tag{4.3}$$

and this equality makes it apparent that $a(l_{\text{opt}})$ *does not vary with the selected target.* Figure 1 illustrates the mapping $a \mapsto \text{eff}(a)$; different choices of target distribution change only the vertical scale. In summary, we have

Theorem 4.1. *Under the hypotheses of Theorem 3.6 and as $d \to \infty$, the measure of cost $\mathsf{E}^*_{l,d}$ defined in (4.2) is minimized for the choice l_{opt} of l that leads to the value of $a = a(l)$ that maximizes (4.3). Rounded to three decimal places, the (target independent) optimal value of the limit probability a is*

$$a(l_{\text{opt}}) = 0.651.$$

The optimal value identified in the preceding theorem is based on the quantity $\mathsf{E}^*_{l,d}$ that underestimates the expected number of proposals. It may be assumed that the practical optimal average acceptance probability is in fact *greater than* or equal to 0.651. In the next subsection we use an alternative measure of efficiency; the expected squared jumping distance. Consideration of this alternative metric will also lead to the same asymptotically optimal acceptance probability of precisely 0.651 as did the minimization of $\mathsf{E}^*_{l,d}$. This suggests that, as $d \to \infty$, the consequences of the fact that $\mathsf{E}^*_{l,d}$ underestimates $\mathsf{E}_{l,d}$ become negligible; proving analytically such a conjecture seems difficult given our current understanding of the limiting HMC dynamics.

4.2. Squared jumping distance

We now consider the chain $Q^0, Q^1, \ldots$ in stationarity (i.e., $Q^0 \sim \Pi(Q)$) and account for the computing cost $\mathsf{C}_{l,d}$ in (4.1) by introducing the continuous-time process $Q^{N(t)}$, where $\{N(t); t \geq 0\}$ denotes a Poisson process, independent of the HMC Markov chain, of intensity $\lambda_d = 1/\mathsf{C}_{l,d}$. If $q_d(t) := q_1^{N(t)}$ denotes the projection of $Q^{N(t)}$ onto the first particle and $\delta > 0$ is a time increment, we measure the efficiency of HMC algorithms using the expected squared jump distance,

$$\mathcal{SJD}_d(\delta) = \mathbb{E}\big[\big(q_d(t+\delta) - q_d(t)\big)^2\big].$$

This measure of efficiency is a fairly standard one; see [8,34], for example.

In the HMC algorithm the computational time (cost) expended between successive steps of the Markov chain is essentially fixed and equal to $\mathsf{C}_{l,d}$. Using an auxiliary Poisson process instead with mean interarrival time equal to $\mathsf{C}_{l,d}$ is merely a device that allows for the definition of processes (over the different choices of l) that take the computational time per step (that changes with l) under consideration in a reasonable manner and can be meaningfully compared via an easy to calculate measure such as $\mathcal{SJD}_d(\delta)$.

The following result shows that $\mathcal{SJD}_d(\delta)$ is indeed asymptotically maximized by maximizing $a(l)l$:

Theorem 4.2. *Under the hypotheses of Proposition 3.8,*

$$\lim_{d\to\infty} d^{5/4} \times \mathcal{SJD}_d = \frac{C_J \delta}{T C_{\mathrm{LF}}} \times a(l)l.$$

Proof. See Section 6. □

4.3. Optimal acceptance probability in practice

As $d \to \infty$, the computing time required for a proposal scales as $1/l$ (see (4.1)) and the number of proposals that may be performed in a given amount of time scales as l. Inspection of (4.1) reveals, however, that selecting a big value of l gives the full benefit of a proportional increase of the number of proposals only asymptotically, and at the slow rate of $\mathcal{O}(d^{-1/4})$. On the other hand, the average acceptance probability converges at the faster rate, $\mathcal{O}(d^{-1/2})$. (This is an application of Stein's method; see, e.g., [3].) These considerations suggest that unless $d^{-1/4}$ is very small, the algorithm will tend to benefit from average acceptance probabilities > 0.651.

Figure 2 shows the results of a numerical study on HMC. The target distribution is a product of $d = 10^5$ standard Gaussian densities $N(0,1)$. We have applied HMC with different choices of the step size h but the same length of time horizon $T = 1$ and in all cases allowed the algorithm to run during a computational time t_{comp} of 30 seconds. We used Monte Carlo averages of the output

$$\hat{f} = \frac{1}{N_{t_{\mathrm{comp}}}} \sum_{n=1}^{N_{t_{\mathrm{comp}}}} f(q_1^n)$$

to estimate, for different choices of f, the expectation $\mathbb{E}[f] = \mathbb{E}[f(q)]$, $q \sim N(0,1)$; here $N_{t_{\mathrm{comp}}}$ denotes the number of T-legs carried out within the allowed time t_{comp}. For each choice of h we ran the HMC algorithm 120 times.

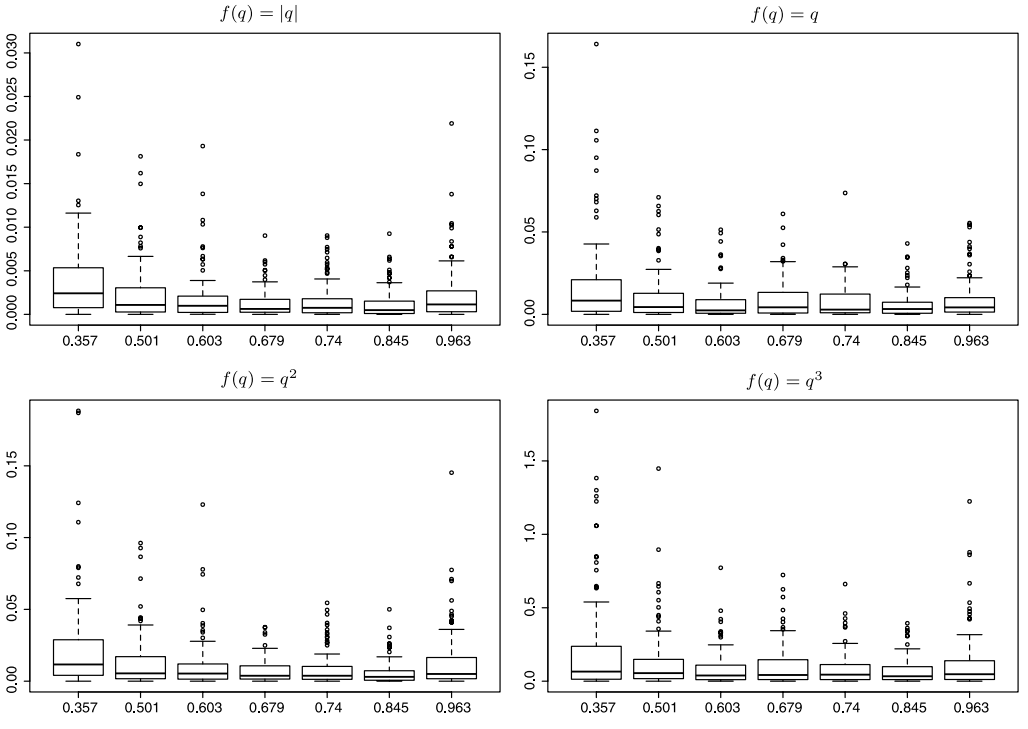

Figure 2. Boxplots of squared errors (SEs) from Monte Carlo averages of HMC with $T = 1$ for 7 different selections of number of leapfrog steps or step sizes h (corresponding to the different boxplots in each panel); the number of leapfrog steps used in the seven scenarios were $(6, 7, 8, 9, 10, 13, 27)$. We ran HMC 120 times, with each run allowed a computing time of 30 seconds. Each boxplot corresponds to the 120 SEs in estimating $\mathbb{E}[f(q)]$, for a particular h and $f(\cdot)$. Written at the bottom of each boxplots is the median of the 120 empirical average acceptance probabilities for the corresponding h. (Note that these medians change in a nonlinear fashion from one boxplot to the next.)

Each of the four panels in Figure 2 corresponds to a different choice of $f(\cdot)$. In each of the panels, the various boxplots correspond to choices of h; at the bottom of each boxplot we have written the median of the 120 empirical average acceptance probabilities. The boxplots themselves use the 120 realizations of the squared distances, $(\hat{f} - \mathbb{E}[f])^2$. The shape of the boxplots endorses the point made above, that the optimal acceptance probability for large (but finite) d is larger than the asymptotically optimal value of 0.651.

5. Estimates for the leapfrog algorithm

In this section we identify analytical hypotheses on V under which Conditions 3.1, 3.2 and 3.7 in Section 3 hold.

We set $f := -\nabla V$ (the 'force') and denote by $f'(q) := f^{(1)}(q), f^{(2)}(q), \ldots$ the successive Fréchet derivatives of f at q. Thus, at a fixed q, $f^{(k)}(q)$ is a multilinear operator from $(\mathbb{R}^m)^{k+1}$ to $\mathbb{R}$. For the rest of this section, we use the following assumptions on V:

Assumptions 5.1. *The function $V : \mathbb{R}^m \to \mathbb{R}$ satisfies*

(i) $V \in C^4(\mathbb{R}^m \to \mathbb{R}_+)$.
(ii) $f', f^{(2)}, f^{(3)}$ *are uniformly bounded by a constant* B.

These assumptions imply that the potential $V(q)$ can grow at most quadratically at infinity as $|q| \to \infty$. (If the growth of V is more than quadratic, then the leapfrog algorithm as applied with a constant value of h throughout the phase space is in fact unstable whenever the initial condition is large.) The case where V takes negative values but is bounded from below can be reduced to the case $V \geq 0$ by adding a suitable constant to V. In terms of the target measure, this simply involves changing the normalization constant and thus is irrelevant in the HMC algorithm.

5.1. Preliminaries

Differentiating (3.1) with respect to t, we find successively

$$\begin{aligned}
\ddot{p}(t) &= f'(q(t))M^{-1}p(t),\\
\ddot{q}(t) &= M^{-1}f(q(t)),\\
\dddot{p}(t) &= f^{(2)}(q(t))(M^{-1}p(t), M^{-1}p(t)) + f'(q(t))M^{-1}f(q(t)),\\
\dddot{q}(t) &= M^{-1}f'(q(t))M^{-1}p(t),\\
\ddddot{p}(t) &= f^{(3)}(q(t))(M^{-1}p(t), M^{-1}p(t), M^{-1}p(t))\\
&\quad + 3f^{(2)}(q(t))(M^{-1}f(q(t)), M^{-1}p(t)) + f'(q(t))M^{-1}f'(q(t))M^{-1}f(q(t)),\\
\ddddot{q}(t) &= M^{-1}f^{(2)}(q(t))(M^{-1}p(t), M^{-1}p(t)) + M^{-1}f'(q(t))M^{-1}f(q(t)).
\end{aligned}$$

In this section, K denotes a generic constant that may vary from one appearance to the next, but will depend only on B, T, $\|M\|$, $\|M^{-1}\|$. From the foregoing equations for the derivatives and using the assumptions on V, we obtain the following bounds:

$$\begin{aligned}
&|\dot{p}(t)| \leq |f(q(t))|, \qquad |\dot{q}(t)| \leq K|p(t)|,\\
&|\ddot{p}(t)| \leq K|p(t)|, \qquad |\ddot{q}(t)| \leq K|f(q(t))|,\\
&|\dddot{p}(t)| \leq K\big(|p(t)|^2 + |f(q(t))|\big), \qquad |\dddot{q}(t)| \leq K|p(t)|,\\
&|\ddddot{p}(t)| \leq K\big(|p(t)|^3 + |p(t)||f(q(t))| + |f(q(t))|\big), \qquad |\ddddot{q}(t)| \leq K\big(|p(t)|^2 + |f(q(t))|\big).
\end{aligned} \tag{5.1}$$

5.2. Asymptotic expansion for the leapfrog solution

In previous sections, we used a subscript to denote the different particles composing our state space. Here we consider leapfrog integration of a single particle and use the subscript to denote

the time level in this integration. The leapfrog scheme can then be compactly written as

$$q_{n+1} = q_n + hM^{-1}p_n + \frac{h^2}{2}M^{-1}f(q_n), \tag{5.2}$$

$$p_{n+1} = p_n + \frac{h}{2}f(q_n) + \frac{h}{2}f\left(q_n + hM^{-1}p_n + \frac{h^2}{2}M^{-1}f(q_n)\right). \tag{5.3}$$

We define the truncation error in the usual way:

$$-\tau_n^{(q)} := q(t_{n+1}) - \left(q(t_n) + hM^{-1}p(t_n) + \frac{h^2}{2}M^{-1}f(q(t_n))\right),$$

$$-\tau_n^{(p)} := p(t_{n+1}) - \left(p(t_n) + \frac{h}{2}f(q_n) + \frac{h}{2}f\left(q(t_n) + hM^{-1}p(t_n) + \frac{h^2}{2}M^{-1}f(q(t_n))\right)\right),$$

where we have set $t_n = nh \in [0, T]$. Expanding and using standard truncation error analysis (see, e.g., [19]), we obtain

$$\tau_n^{(q)} = \tfrac{1}{6}h^3\dddot{q}(t_n) + h^4\mathcal{O}(\|\ddddot{q}(\cdot)\|_\infty),$$
$$\tau_n^{(p)} = -\tfrac{1}{12}h^3\dddot{p}(t_n) + h^4\mathcal{O}(\|\ddddot{p}(\cdot)\|_\infty) + h\mathcal{O}\big(\tau_n^{(q)}\big),$$

where, for arbitrary function g,

$$\|g(\cdot)\|_\infty := \sup_{0\le t\le T}|g(t)|.$$

In view of these estimates, $\frac{1}{6}h^3\dddot{q}(t_n)$ and $-\frac{1}{12}h^3\dddot{p}(t_n)$ are the leading terms in the asymptotic expansion of the truncation error. Standard results (see, e.g., [20], Section II.8) show that the numerical solution has an asymptotic expansion,

$$\begin{aligned} q_n &= q(t_n) + h^2 v(t_n) + \mathcal{O}(h^3), \\ p_n &= p(t_n) + h^2 u(t_n) + \mathcal{O}(h^3), \end{aligned} \tag{5.4}$$

where functions $u(\cdot)$ and $v(\cdot)$ are the solutions, with initial condition $u(0) = v(0) = 0$, of the *variational* system

$$\begin{pmatrix} \dot{u}(t) \\ \dot{v}(t) \end{pmatrix} = \begin{pmatrix} 0 & M^{-1}f'(q(t)) \\ I & 0 \end{pmatrix}\begin{pmatrix} u(t) \\ v(t) \end{pmatrix} + \begin{pmatrix} \frac{1}{12}\dddot{p}(t) \\ -\frac{1}{6}\dddot{q}(t) \end{pmatrix}. \tag{5.5}$$

Remark 5.2. Note here that $u(\cdot), v(\cdot)$ depend on the initial conditions $(q(0), p(0))$ via $(q(\cdot), p(\cdot))$, but that this dependence is not reflected in the notation. One should keep in mind that most of the norms appearing in the sequel are functions of $(q(0), p(0))$.

Applying Gronwall's lemma and using the estimates (5.1), we obtain the bound

$$\|u(\cdot)\|_\infty + \|v(\cdot)\|_\infty \le K\big(\|p(\cdot)\|_\infty^2 + \|f(q(\cdot))\|_\infty\big) \tag{5.6}$$

and, by differentiating (5.5) with respect to t, expressing $\dot{u}, \dot{v}$ in terms of u, v, and using (5.1) again, in turn we obtain

$$\|\ddot{u}(\cdot)\|_\infty \le K\big(\|p(\cdot)\|_\infty^3 + \|p(\cdot)\|_\infty \|f(q(\cdot))\|_\infty + \|f(q(\cdot))\|_\infty\big), \tag{5.7}$$

$$\|\ddot{v}(\cdot)\|_\infty \le K\big(\|p(\cdot)\|_\infty^2 + \|f(q(\cdot))\|_\infty\big). \tag{5.8}$$

5.3. Estimates for the global error

With the leading coefficients u and v of the global errors $q_n - q(t_n)$, $p_n - p(t_n)$ estimated in (5.6), our task now is to obtain an explicit bound for the constants implied in the $\mathcal{O}(h^3)$ remainder in (5.4). Toward this end, we define the quantities

$$z_n := q(t_n) + h^2 v(t_n),$$
$$w_n := p(t_n) + h^2 u(t_n),$$

and denote by $\tau_n^{(q)*}$, $\tau_n^{(p)*}$ the residuals that they generate when substituted in (5.2) and (5.3), respectively, that is,

$$-\tau_n^{(q)*} = z_{n+1} - z_n - hM^{-1}w_n - \frac{h^2}{2}M^{-1}f(z_n),$$

$$-\tau_n^{(p)*} = w_{n+1} - w_n - \frac{h}{2}f(z_n) - \frac{h}{2}f\left(z_n + hM^{-1}w_n + \frac{h^2}{2}M^{-1}f(z_n)\right).$$

Because the leapfrog scheme is stable, standard numerical analysis techniques [20] show that it is possible to estimate the global errors in terms of the local residuals (truncation errors). This gives

$$\max_{0\le t_n\le T}(|q_n - z_n| + |p_n - w_n|) \le \frac{C}{h}\max_{0\le t_n\le T}\big(\big|\tau_n^{(q)*}\big| + \big|\tau_n^{(p)*}\big|\big), \tag{5.9}$$

with the constant C depending only on T and Lipschitz constant of the map $(q_n, p_n) \mapsto (q_{n+1}, p_{n+1})$, which in turn depends on $\|M^{-1}\|$ and the bound for f'. The stability bound (5.9) is the basis of the proof of the following estimation of the global error:

Proposition 5.3. *If the potential V satisfies Assumptions 5.1, then for $0 \le t_n \le T$,*

$$\big|p_n - \big(p(t_n) + h^2 u(t_n)\big)\big| \le Kh^3\big(\|p(\cdot)\|_\infty^4 + \|f(q(\cdot))\|_\infty^2 + 1\big),$$
$$\big|q_n - \big(q(t_n) + h^2 v(t_n)\big)\big| \le Kh^3\big(\|p(\cdot)\|_\infty^4 + \|f(q(\cdot))\|_\infty^2 + 1\big).$$

Proof. Our task is reduced to estimating $\tau_n^{(q)*}$, $\tau_n^{(p)*}$. We only present the estimation for $\tau_n^{(p)*}$, since the computations for $\tau_n^{(q)*}$ are similar but simpler.

After regrouping the terms in $\tau_n^{(p)*}$, we find that

$$\begin{aligned}-\tau_n^{(p)*} = {} & \underbrace{p(t_{n+1}) - p(t_n) - \frac{h}{2} f(q(t_n)) - \frac{h}{2} f(q(t_{n+1})) + \frac{h^3}{12} \dddot{p}(t)}_{I_1} \\ & + \underbrace{h^2 \left(u(t_{n+1}) - u(t_n) - h f'(q(t_n)) v(t_n) - \frac{h}{12} \ddot{p}(t) \right)}_{I_2} \\ & + \underbrace{\frac{h}{2} \big(f(q(t_n)) - f(z_n) + h^2 f'(q(t_n)) v(t_n) \big)}_{I_3} \\ & - \underbrace{\frac{h}{2} \left(f\left(z_n + h M^{-1} w_n + \frac{h^2}{2} M^{-1} f(q(t_n)) \right) - f\big(q(t_{n+1}) - h^2 f'(q(t_n)) v(t_n) \big) \right)}_{I_4} \\ & + \underbrace{\frac{h}{2} \left(f\left(z_n + h M^{-1} w_n + \frac{h^2}{2} M^{-1} f(q(t_n)) \right) - f\left(z_n + h M^{-1} w_n + \frac{h^2}{2} M^{-1} f(z_n) \right) \right)}_{I_5}.\end{aligned}$$

Now we estimate the foregoing five terms separately.

I_1: We note that

$$p(t_{n+1}) - p(t_n) - \frac{h}{2} f(q(t_n)) - \frac{h}{2} f(q(t_{n+1})) = p(t_{n+1}) - p(t_n) - \frac{h}{2} \dot{p}(t_{n+1}) - \frac{h}{2} \dot{p}(t_n),$$

and by using the estimates in (5.1), it follows that

$$|I_1| \le K h^4 \big(\|p(\cdot)\|_\infty + \|p(\cdot)\|_\infty \|f(q(\cdot))\|_\infty + \|f(q(\cdot))\|_\infty \big).$$

I_2: Here we write $I_2 = h^2 (u(t_{n+1}) - u(t_n) - h \dot{u}(t_n))$ so that by (5.7)

$$|I_2| \le K h^4 \big(\|p(\cdot)\|_\infty^3 + \|p(\cdot)\|_\infty \|f(q(\cdot))\|_\infty + \|f(q(\cdot))\|_\infty \big).$$

I_3: This term is estimated, after Taylor expanding $f(z_n)$ near $f(q(t_n))$, by

$$|I_3| \le K h^5 \big(\|p(\cdot)\|_\infty + \|f(q(\cdot))\|_\infty \big)^2.$$

I_4: We rewrite this as

$$\frac{h}{2} \big(f\big(q(t_{n+1}) + \tau_n^{(q)} + h^2 v(t_n) + h^3 M^{-1} v(t_n) \big) - f(q(t_{n+1})) - h^2 f'(q(t_n)) v(t_n) \big)$$

and Taylor expand around $f(q(t_n))$ to derive the bound

$$|I_4| \le K h^4 \big(\|p(\cdot)\|_\infty^4 + \|f(q(\cdot))\|_\infty^2 \big).$$

I_5: This term is easily estimated as

$$|I_5| \le Kh^5 \|v(\cdot)\|_\infty \le Kh^5 \big(\|p(\cdot)\|_\infty^2 + \|f(q(\cdot))\|_\infty \big).$$

Combining all the foregoing estimates, we have the bound

$$\big|\tau_n^{(p)*}\big| \le Kh^4 \big(\|p(\cdot)\|_\infty^4 + \|f(q(\cdot))\|_\infty^2 \big).$$

A similar analysis for $\tau_n^{(q)*}$ yields the bound

$$\big|\tau_n^{(q)*}\big| \le Kh^4 \big(\|p(\cdot)\|_\infty^4 + \|f(q(\cdot))\|_\infty^2 \big).$$

The proof is completed by substituting the foregoing estimates in (5.9). □

We now use the estimates in Proposition 5.3 to derive the asymptotic expansion for the energy increment for the leapfrog scheme (cf. Condition 3.1).

Proposition 5.4. *Let potential V satisfy Assumptions 5.1. Then, for the leapfrog scheme, we get*

$$\Delta(x,h) = h^2 \alpha(x) + h^2 \rho(x,h),$$

with

$$\begin{aligned} \alpha(x) &= \langle M^{-1} p(T), u(T) \rangle - \langle f(q(T)), v(T) \rangle, \\ |\alpha(x)| &\le K \big(\|p(\cdot)\|_\infty^3 + \|f(q(\cdot))\|_\infty^2 + 1 \big), \\ |\rho(x,h)| &\le Kh \big(\|p(\cdot)\|_\infty^8 + \|f(q(\cdot))\|_\infty^2 + 1 \big), \qquad 0 < h \le 1, \end{aligned}$$

where $(q(\cdot), p(\cdot))$ denotes the solution of (3.1) with initial data $x \equiv (q(0), p(0))$ and $u(\cdot), v(\cdot)$ are the solutions of the corresponding variational system given in (5.5) with $u(0) = v(0) = 0$.

Proof. We only consider the case when T/h is an integer. The general case follows with minor adjustments. By Proposition 5.3,

$$\begin{aligned} \Delta(x,h) &= H\big(\psi_h^{(T)}(x)\big) - H(x) = H\big(\psi_h^{(T)}(x)\big) - H(\varphi_T(x)) \\ &= \langle M^{-1} p(T), h^2 u(T) + h^3 R_1 \rangle + \tfrac{1}{2} \big\langle M^{-1} \big(h^2 u(T) + h^3 R_1 \big), \big(h^2 u(T) + h^3 R_1 \big) \big\rangle \\ &\quad + V\big(q(T) + h^2 v(T) + h^3 R_2 \big) - V(q(T)), \end{aligned}$$

where R_1, R_2 are remainders with

$$|R_1| + |R_2| \le K \big(\|p(\cdot)\|_\infty^4 + \|f(q(\cdot))\|_\infty^2 + 1 \big).$$

By Taylor expanding $V(\cdot)$ around $q(T)$, we obtain

$$\Delta(x,h) = h^2 \big(\langle M^{-1} p(T), u(T) \rangle - \langle f(q(T)), v(T) \rangle \big) + \rho(x,h),$$

with

$$|\rho(x,h)| \le Kh^3\big(\|p(\cdot)\|_\infty^8 + \|f(q(\cdot))\|_\infty^2 + 1\big)$$

for $0 \le h \le 1$. From the bound (5.6), it follows that

$$\begin{aligned}|\alpha(x)| &\le K\big(\|p(\cdot)\|_\infty \|u(\cdot)\|_\infty + \|f(q(\cdot))\|_\infty \|v(\cdot)\|_\infty\big)\\ &\le K\big(\|p(\cdot)\|_\infty^3 + \|f(\cdot)\|_\infty^2 + 1\big)\end{aligned}$$

and the theorem is proved. □

Our analysis is completed by estimating the quantities $\|p(\cdot)\|_\infty$ and $\|q(\cdot)\|_\infty$ in the preceding theorems in terms of the initial data $(q(0), p(0))$. We obtain these estimates for two families of potentials that include most of the interesting/useful target distributions. The corresponding estimates for other potentials may be obtained using similar methods.

Proposition 5.5. *Let potential V satisfy Assumptions 5.1. If V also satisfies either of the conditions*

(1) *f is bounded and*

$$\int_{\mathbb{R}^m} |V(q)|^8 \mathrm{e}^{-V(q)}\,\mathrm{d}q < \infty \tag{5.10}$$

or,

(2) *there exist constants $C_1, C_2 > 0$, and $0 < \gamma \le 1$ such that for all $|q| \ge C_2$, we have $V(q) \ge C_1|q|^\gamma$*

then Conditions 3.1, 3.2 and 3.7 all hold.

Proof. Here we present only the treatment of Conditions 3.1 and 3.2. The derivation of Condition 3.7 is similar and simpler.

From Proposition 5.4, we observe that function $D(x)$ in Condition 3.2 may be taken to be

$$D(x) = K\big(\|p(\cdot)\|_\infty^{16} + \|f(q(\cdot))\|_\infty^4 + 1\big).$$

Thus, to prove the integrability of $D(\cdot)$, we need to estimate $\|p(\cdot)\|_\infty$ and $\|f(q(\cdot))\|_\infty$. Estimating $\|p(\cdot)\|_\infty$ is easier. Indeed, by conservation of energy,

$$\tfrac{1}{2}\langle p(t), M^{-1}p(t)\rangle \le \tfrac{1}{2}\langle p(0), M^{-1}p(0)\rangle + V(q(0)),$$

which implies

$$|p(t)|^{16} \le K\big(|p(0)|^{16} + |V(q(0))|^8\big). \tag{5.11}$$

We now prove the integrability of $D(\cdot)$ under each of the two previously stated hypotheses.

Under hypothesis (1), suppose that f is bounded. In this case, we obtain that $|D(x)| \le K(\|p(\cdot)\|_\infty^{16} + 1)$; therefore, it is sufficient to estimate $\|p(\cdot)\|_\infty$. Because the Gaussian distribution has all moments, integrability of D follows from (5.10) and (5.11).

Under hypothesis (2), using the stated hypothesis on $V(q)$, we obtain

$$C_1|q(t)|^\gamma \le V(q(t)) \le \tfrac{1}{2}\langle p(0), M^{-1}p(0)\rangle + V(q(0)),$$

which implies that

$$|q(t)| \le K\big(|p(0)|^{2/\gamma} + |V(q(0))|^{1/\gamma}\big).$$

By Assumptions 5.1(i), $|f(q(t))| \le K(1+|q(t)|)$, and arguing as before and using the bound (5.11), integrability of D follows if we show that

$$\int_{\mathbb{R}^m} |V(q)|^\delta \mathrm{e}^{-V(q)}\,\mathrm{d}q < \infty, \qquad \delta = \max\left(8, \frac{4}{\gamma}\right).$$

Because $|V(q)| \le K(1+|q|^2)$,

$$\int_{\mathbb{R}^m} |V(q)|^\delta \mathrm{e}^{-V(q)}\,\mathrm{d}q \le K\int_{\mathbb{R}^m} (1+|q|^{2\delta})\mathrm{e}^{-B|q|^\gamma}\,\mathrm{d}q < \infty$$

and we are done. □

6. Proofs of probabilistic results

Proof of Lemma 3.3. The volume preservation property of $\psi_h^{(T)}(\cdot)$ implies that the associated Jacobian is unit. Thus, setting $x = (\psi_h^{(T)})^{-1}(y)$, we get

$$\begin{aligned}\int_{\mathbb{R}^{2m}} g(\Delta(x,h))\mathrm{e}^{-H(x)}\,\mathrm{d}x &= \int_{\mathbb{R}^{2m}} g\big(H\big(\psi_h^{(T)}(x)\big) - H(x)\big)\mathrm{e}^{-H(x)}\,\mathrm{d}x \\ &= \int_{\mathbb{R}^{2m}} g\big[H(y) - H\big(\big(\psi_h^{(T)}\big)^{-1}(y)\big)\big]\mathrm{e}^{-H((\psi_h^{(T)})^{-1}(y))}\,\mathrm{d}y.\end{aligned}$$

Following the definition of time reversibility in (2.2), we have

$$S\circ\psi_h^{(T)} = \big(\psi_h^{(T)}\big)^{-1}\circ S$$

for the symmetry operator S such that $S(q,p) = (q,-p)$. Now using the volume-preserving transformation $y = Sz$ and continuing from the foregoing, we get

$$\begin{aligned}&\int_{\mathbb{R}^{2m}} g(\Delta(x,h))\mathrm{e}^{-H(x)}\,\mathrm{d}x \\ &\quad= \int_{\mathbb{R}^{2m}} g\big(H(Sz) - H\big(\big(\psi_h^{(T)}\big)^{-1}(Sz)\big)\big)\mathrm{e}^{-H((\psi_h^{(T)})^{-1}(Sz))}\,\mathrm{d}z \\ &\quad= \int_{\mathbb{R}^{2m}} g\big(H(Sz) - H\big(S\psi_h^{(T)}(z)\big)\big)\mathrm{e}^{-H(S(\psi_h^{(T)}(z)))}\,\mathrm{d}z \\ &\quad= \int_{\mathbb{R}^{2m}} g\big(H(z) - H\big(\psi_h^{(T)}(z)\big)\big)\mathrm{e}^{-H(\psi_h^{(T)}(z))}\,\mathrm{d}z,\end{aligned}$$

where in the last equation we used the identity $H(Sz) = H(z)$. □

Proof of Proposition 3.4. We first find the limit of $\sigma^2(h)/h^4$. Conditions 3.1 and 3.2 imply that

$$\frac{\Delta^2(x,h)}{h^4} = \alpha^2(x) + \rho^2(x,h) + 2\rho(x,h)\alpha(x) \le D(x).$$

Because for fixed x, $\Delta^2(x,h)/h^4 \to \alpha^2(x)$, the dominated convergence theorem shows that

$$\lim_{h\to 0} \frac{s^2(h)}{h^4} = \int_{\mathbb{R}^{2m}} \alpha^2(x)\mathrm{e}^{-H(x)}\,\mathrm{d}x = \Sigma.$$

Now (3.6) implies that

$$\lim_{h\to 0} \frac{\mu^2(h)}{h^4} = 0, \tag{6.1}$$

and the required limit for $\sigma^2(h)/h^4$ follows directly. Then, from (3.5) we obtain

$$\frac{2\mu(h) - \sigma^2(h)}{h^4} = -\int_{\mathbb{R}^{2m}} \frac{\Delta(x,h)}{h^2} \frac{[\exp(-\Delta(x,h)) - 1 + \Delta(x,h)]}{h^2} \mathrm{e}^{-H(x)}\,\mathrm{d}x + \frac{\mu^2(h)}{h^4}.$$

Because for any fixed x, Conditions 3.1 and 3.2 imply that $\Delta(x,h) \to 0$ as $h \to 0$ and $\Delta^2(x,h) = \mathcal{O}(h^4)$, we have the pointwise limit

$$\lim_{h\to 0} \frac{\exp(-\Delta(x,h)) - 1 + \Delta(x,h)}{h^2} = 0.$$

Using inequality $|u||\mathrm{e}^u - 1 - u| \le |u|^2(\mathrm{e}^u + 2)$, we deduce that for all sufficiently small h,

$$\begin{aligned}
&\int_{\mathbb{R}^{2m}} \frac{|\Delta(x,h)|}{h^2} \frac{|\exp(-\Delta(x,h)) - 1 + \Delta(x,h)|}{h^2} \mathrm{e}^{-H(x)}\,\mathrm{d}x \\
&\quad \le \int_{\mathbb{R}^{2m}} \frac{|\Delta^2(x,h)|}{h^4} \exp\bigl(-\Delta(x,h)\bigr)\mathrm{e}^{-H(x)}\,\mathrm{d}x + 2\int_{\mathbb{R}^{2m}} \frac{|\Delta^2(x,h)|}{h^4} \mathrm{e}^{-H(x)}\,\mathrm{d}x \\
&\quad \le 3\int_{\mathbb{R}^{2m}} D(x)\mathrm{e}^{-H(x)}\,\mathrm{d}x < \infty,
\end{aligned}$$

where the last line follows from applying Lemma 3.3 with $\varphi(x) = x^2$ and Condition 3.2. Thus the dominated convergence theorem yields

$$\lim_{h\to 0} \frac{2\mu(h) - \sigma^2(h)}{h^4} = 0.$$

This completes the proof of the proposition. □

Proof of Theorem 3.6. We continue from (3.8). In view of the scaling $h = l \cdot d^{-1/4}$ we obtain, using Proposition 3.4,

$$\mathbb{E}[R_d] = -d \cdot \mu(h) \to -\frac{l^4\sigma}{2}$$

and

$$\mathrm{Var}[R_d] = d \cdot \sigma^2(h) \to l^4\Sigma.$$

The Lindeberg condition is readily seen to hold, and thus

$$R_d \xrightarrow{\mathcal{L}} R_\infty := N\left(-\frac{l^4\Sigma}{2}, l^4\Sigma\right).$$

From the boundedness of $u \mapsto 1 \wedge \mathrm{e}^u$, we may write

$$\mathbb{E}[a(X, Y)] \to \mathbb{E}[1 \wedge \mathrm{e}^{R_\infty}],$$

where the last expectation can be found analytically (see, e.g., [36]) as

$$\mathbb{E}[1 \wedge \mathrm{e}^{R_\infty}] = 2\Phi\big(-l^2\sqrt{\Sigma}/2\big).$$

This completes the proof. □

Proof of Proposition 3.8. For simplicity, we write just q^n, q^{n+1} and p^n instead of q_1^n, q_1^{n+1} and p_1^n, respectively. Using (3.9), we get

$$(q^{n+1} - q^n)^2 = I^n\big(\mathcal{P}_q\psi_h^{(T)}(q^n, p^n) - q^n\big)^2.$$

We define

$$a^-(X^n, Y^n) := 1 \wedge \exp\left\{-\sum_{i=2}^{d}\Delta(x_i^n, h)\right\}; \qquad I^{n-} := \mathbb{I}_{U^n < a^-(X^n, Y^n)}, \tag{6.2}$$

and set

$$\xi^n = I^{n-}\big(\mathcal{P}_q\psi_h^{(T)}(q^n, p^n) - q^n\big)^2.$$

Using the Lipschitz continuity of $u \mapsto \mathbb{I}_{U \le 1 \wedge \mathrm{e}^u}$ and the Cauchy–Schwartz inequality, we get

$$\mathbb{E}|(q^{n+1} - q^n)^2 - \xi^n| \le |\Delta(x_1, h)|_{L_2}\big|\big(\mathcal{P}_q\psi_h^{(T)}(q^n, p^n) - q^n\big)^2\big|_{L_2}.$$

Now Conditions 3.1 and 3.2 imply that

$$|\Delta(x_1, h)|_{L_2} = \mathcal{O}(h^2).$$

In addition, from Condition 3.7 and the stated hypothesis on the density $\exp(-V)$, q^n and $\mathcal{P}_q\psi_h^{(T)}(q^n, p^n)$ have bounded fourth moments uniformly in h, and thus

$$\left|\left(\mathcal{P}_q\psi_h^{(T)}(q^n, p^n) - q^n\right)^2\right|_{L_2} \leq C$$

for some constant $C > 0$. The last two statements imply that

$$\mathbb{E}|(q^{n+1} - q^n)^2 - \xi^n| = \mathcal{O}(h^2). \tag{6.3}$$

Exploiting the independence between I^{n-} and the first particle,

$$\mathbb{E}[\xi_n] = \mathbb{E}[a^-(X, Y)] \times \mathbb{E}\left[\left(\mathcal{P}_q\psi_h^{(T)}(q^n, p^n) - q^n\right)^2\right] \to a(l) \cdot \mathbb{E}\left[\left(\mathcal{P}_q\varphi_T(q^n, p^n) - q^n\right)^2\right],$$

where for the first factor, we used its limit from Theorem 3.6, and for the second factor, the limit is a consequence of Condition 3.7 and the dominated convergence theorem. Equation (6.3) completes the proof. □

Proof of Proposition 3.9. Fix some $q_1^n \in \mathbb{R}^m$. We define $a^-(X^n, Y^n)$ and I^{n-} as in (6.2). For simplicity, we write just q^n, q^{n+1}, q^{n+1} and p^n instead of q_1^n, q_1^{n+1}, q_1^{n+1} and p_1^n, respectively.

We set

$$g^{n+1} = I^{n-} \cdot \mathcal{P}_q\varphi_T(q^n, p^n) + (1 - I^{n-})q^n.$$

Adding and subtracting $I^n \cdot \mathcal{P}_q(\varphi_T(q^n, p^n))$ yields

$$\begin{aligned} |q^{n+1} - g^{n+1}| \leq{}& \left|\mathcal{P}_q\left(\psi_h^{(T)}(q^n, p^n)\right) - \mathcal{P}_q(\varphi_T(q^n, p^n))\right| \\ &+ |I^{n-} - I^n|\left(|\mathcal{P}_q(\varphi_T(q^n, p^n))| + |q^n|\right). \end{aligned} \tag{6.4}$$

Using the Lipschitz continuity (with constant 1) of $u \mapsto \mathbb{I}_{U \leq 1 \wedge \exp(u)}$, we have

$$|I^{n-} - I^n| \leq |\Delta(x_1, h)|. \tag{6.5}$$

Now Condition 3.7 implies that the first term on the right-hand side of (6.4) vanishes with probability 1, and Condition 3.1 implies (via (6.5)) that the second term also vanishes with probability 1. Therefore, as $d \to \infty$,

$$q^{n+1} - g^{n+1} \to 0, \qquad \text{a.s.}$$

Theorem 3.6 immediately implies that $I^{n-} \xrightarrow{\mathcal{L}} \mathsf{I}^n$, and thus

$$g^{n+1} \xrightarrow{\mathcal{L}} \mathsf{q}^{n+1}.$$

From these two limits, we have $q^{n+1} \xrightarrow{\mathcal{L}} \mathsf{q}^{n+1}$, and this completes the proof. □

Proof of Theorem 4.2. To simplify the notation, we again drop the subscript 1. Conditionally on the trajectory $q^0, q^1, \ldots$, we get

$$\big(q(t+\delta)-q(t)\big)^2 = \begin{cases} 0, & \text{w.p. } 1-\lambda_d\delta+\mathcal{O}((\lambda_d\delta)^2), \\ \big(q^{N(t)+1}-q^{N(t)}\big)^2, & \text{w.p. } \lambda_d\delta+\mathcal{O}((\lambda_d\delta)^2), \\ \big(q^{N(t)+1+j}-q^{N(t)}\big)^2, & \text{w.p. } \mathcal{O}((\lambda_d\delta)^{j+1}),\ j\geq 1. \end{cases}$$

Therefore,

$$\begin{aligned} \mathcal{SJD}_d = {} & \mathbb{E}\big[\big(q^{N(t)+1}-q^{N(t)}\big)^2\big]\big(\lambda_d\delta+\mathcal{O}((\lambda_d\delta)^2)\big) \\ & +\sum_{j\geq 1}\mathbb{E}\big[\big(q^{N(t)+1+j}-q^{N(t)}\big)^2\big]\mathcal{O}((\lambda_d\delta)^{j+1}). \end{aligned} \tag{6.6}$$

Note now that

$$\begin{aligned} \mathbb{E}\big[\big(q^{N(t)+1+j}-q^{N(t)}\big)^2\big] &\leq \left(\sum_{k=1}^{j+1}\big|q^{N(t)+k}-q^{N(t)+k-1}\big|_{L_2}\right)^2 \\ &= (j+1)^2\mathbb{E}[(q^{n+1}-q^n)^2], \end{aligned}$$

because we have assumed stationarity. From (4.1),

$$\lambda_d = d^{-5/4}\frac{l}{TC_{\mathrm{LF}}}+\mathcal{O}(d^{-6/4})$$

and, from Proposition 3.8, $\mathbb{E}[(q^{n+1}-q^n)^2]=\mathcal{O}(1)$. Therefore,

$$d^{5/4}\times\sum_{j\geq 1}\mathbb{E}\big[\big(q^{N(t)+1+j}-q^{N(t)}\big)^2\big]\mathcal{O}((\lambda_d\delta)^{j+1})$$

is of the same order in d as

$$\lambda_d^2\cdot d^{5/4}\times\sum_{j\geq 1}(j+1)^2\mathcal{O}(\lambda_d^{j-1}),$$

and thus

$$d^{5/4}\times\sum_{j\geq 1}\mathbb{E}\big[\big(q^{N(t)+1+j}-q^{N(t)}\big)^2\big]\mathcal{O}((\lambda_d\delta)^{j+1})=\mathcal{O}(\lambda_d).$$

Using this result, and continuing from (6.6), Proposition 3.8 provides the required statement. □

7. Conclusion

The HMC methodology provides a promising framework for the study of sampling problems, especially in high dimensions. There are a number of directions in which the research direction

taken in this paper could be developed further, and a number of observations to be made concerning optimal tuning of MCMC methods in general. We conclude by listing some of these issues.

- The overall optimization involves tuning *three* free parameters (h, T, M). Because M is a symmetric matrix, the number of tuning parameters is $2 + m(m+1)/2$. In this paper, we have fixed M and T and illustrated that the choice $h = ld^{-1/4}$ provides nonvanishing $\mathcal{O}(1)$ acceptance probabilities as $d \to \infty$. We then focussed on optimizing the HMC algorithm over choice of l. The natural next step would be to study the algorithm for various choices of the mass matrix M and the integration time T.
- There is interesting recent computational work [16] concerning exploration of state space by means of nonseparable Hamiltonian dynamics. This work opens up several theoretical research directions.
- The issue of irreducibility for the transition kernel of HMC is subtle and requires further investigation, given that certain exceptional cases can lead to nonergodic behavior (see [11, 42] and references therein).
- Our analysis of the HMC algorithm is conducted in stationarity. It is possible that different scaling analyses will be needed to study the burn-in phase of the algorithm, as for the study of random-walk type algorithms in [13].
- There is evidence that the limiting properties of MALA for high-dimensional target densities do not appear to depend critically on the tail behavior of the target (see [37]). However, in the present work, for HMC, we have considered densities that are no lighter than Gaussian at infinity. It would be interesting to extend the work to light-tailed densities. This links naturally to the question of using variable step size integration [41] for HMC, because light-tailed densities will lead to superlinear vector fields at infinity in (2.1). This also links to the work of [16], in which nonseparable Hamiltonians arise via introduction of a nonstandard metric on phase space, related to the Fisher information. This metric introduces a rescaling of state space, and this rescaling induces similar algorithmic properties to those induced by variable time-stepping.
- We have shown how to scale the HMC method to obtain $\mathcal{O}(1)$ acceptance probabilities as the dimension of the target product measure grows. We have also shown how to minimize a reasonable measure of computational cost, defined as the work needed to make an $\mathcal{O}(1)$ move in state space. However, in contrast to similar work on RWM and MALA ([36,37]), in which a scalar SDE governs, for large d, the evolution of a single component of the Markov chain, we have not identified a limiting Markov process arising in the infinite-dimensional limit of HMC. This remains an interesting and technically demanding challenge.
- The work concerning optimal scaling of RWM and MALA in [36,37] and the identification of the optimal acceptance probabilities of 0.234 and 0.574, respectively, concerns target measures with an i.i.d. structure. However, recent work [27,35] shows that for measures that have density with respect to a Gaussian measure (in the limit $d \to \infty$) and hence are not necessarily i.i.d., the same optimal acceptance probabilities arise. A natural extension of the work on HMC presented in this paper is to non-i.i.d. target measures in a similar manner. Also note that for measures with this special structure, these results on optimal scaling are of mainly theoretical interest, because they extend known results out of the i.i.d. scenario. For the particular case of measures that have density with respect to a Gaussian, and from a

more practical perspective, the RWM, MALA and HMC algorithms should be modified to exploit this underlying Gaussian structure, as discussed next.

- We have concentrated on explicit integration by the leapfrog method. For measures that have density with respect to a Gaussian measure (in the limit $d \to \infty$) it is natural to use semi-implicit integrators that compute the linear dynamics implicitly, leading to exact statistics in the pure Gaussian case. This idea, first developed for the MALA algorithm [9] and for the RWM algorithm [10], leads to methods that explore state space in $\mathcal{O}(1)$ steps for measures with this special structure. The idea was recently developed for HMC methods by [7], and the resulting algorithm was shown to outperform the semi-implicit MALA algorithm for some problems arising in conditioned diffusions. Developing a theoretical understanding of this behavior would be of interest. Note that the optimal acceptance probabilities $0.234, 0.574$ and 0.651 will not necessarily apply for these semi-implicit proposals as the optimal proposal variance does not shrink to 0 as $d \to \infty$; as a result, different mechanisms may come into play when determining optimality.
- It would be interesting to conduct simulation studies which investigate the robustness of optimal scaling results for RWM, MALA and HMC in scenarios in which the target is not i.i.d. or change of measure from Gaussian. Such simulation studies could help guide future theoretical results on optimal scaling.

Acknowledgements

We thank Sebastian Reich for drawing our attention to the paper [18], which sparked our initial interest in the scaling issue for HMC. We also thank Gabriel Stoltz and Robert D. Skeel for stimulating discussions and useful comments. NP gratefully acknowledges the NSF Grant DMS 1107070; JMS gratefully acknowledges the Grant TM2010-18246-C03 by Ministerio de Ciencia e Innovacion, Spain; AS is grateful to EPSRC and ERC for financial support. Part of this work was done when NP was a postdoctoral member of CRiSM, University of Warwick, and visited JMS at University of Valladolid, so we thank both these institutions for their warm hospitality. Finally, we thank two referees for their comments that greatly improved the content and presentation of the paper.

References

[1] Akhmatskaya, E., Bou-Rabee, N. and Reich, S. (2009). A comparison of generalized hybrid Monte Carlo methods with and without momentum flip. *J. Comput. Phys.* **228** 2256–2265. MR2500680

[2] Alexander, F.J., Eyink, G.L. and Restrepo, J.M. (2005). Accelerated Monte Carlo for optimal estimation of time series. *J. Stat. Phys.* **119** 1331–1345.

[3] Barbour, A.D. and Chen, L.H.Y. (eds.) (2005). *An Introduction to Stein's Method. Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap.* **4**. Singapore: Singapore Univ. Press.

[4] Bédard, M. (2007). Weak convergence of Metropolis algorithms for non-i.i.d. target distributions. *Ann. Appl. Probab.* **17** 1222–1244. MR2344305

[5] Bédard, M. (2008). Efficient sampling using Metropolis algorithms: Applications of optimal scaling results. *J. Comput. Graph. Statist.* **17** 312–332. MR2439962

[6] Bédard, M. (2008). Optimal acceptance rates for Metropolis algorithms: Moving beyond 0.234. *Stochastic Process. Appl.* **118** 2198–2222. MR2474348
[7] Beskos, A., Pinski, F.J., Sanz-Serna, J.M. and Stuart, A.M. (2011). Hybrid Monte Carlo on Hilbert spaces. *Stochastic Process. Appl.* **121** 2201–2230. MR2822774
[8] Beskos, A., Roberts, G. and Stuart, A. (2009). Optimal scalings for local Metropolis–Hastings chains on nonproduct targets in high dimensions. *Ann. Appl. Probab.* **19** 863–898. MR2537193
[9] Beskos, A., Roberts, G., Stuart, A. and Voss, J. (2008). MCMC methods for diffusion bridges. *Stoch. Dyn.* **8** 319–350. MR2444507
[10] Beskos, A. and Stuart, A. (2009). MCMC methods for sampling function space. In *ICIAM* 07 – 6*th International Congress on Industrial and Applied Mathematics* 337–364. Zürich: Eur. Math. Soc. MR2588600
[11] Cancès, E., Legoll, F. and Stoltz, G. (2007). Theoretical and numerical comparison of some sampling methods for molecular dynamics. *M2AN Math. Model. Numer. Anal.* **41** 351–389. MR2339633
[12] Chen, L., Qin, Z. and Liu, J. (2000). Exploring hybrid Monte Carlo in Bayesian computation. In *ISBA Proceedings*, 2000.
[13] Christensen, O.F., Roberts, G.O. and Rosenthal, J.S. (2005). Scaling limits for the transient phase of local Metropolis–Hastings algorithms. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 253–268. MR2137324
[14] Diaconis, P., Holmes, S. and Neal, R.M. (2000). Analysis of a nonreversible Markov chain sampler. *Ann. Appl. Probab.* **10** 726–752. MR1789978
[15] Duane, S., Kennedy, A.D., Pendleton, B. and Roweth, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **195** 216–222.
[16] 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
[17] Gupta, R., Kilcup, G.W. and Sharpe, S.R. (1988). Tuning the Hybrid Monte Carlo algorithm. *Phys. Rev. D* **38** 1278–1287.
[18] Gupta, S., Irbäck, A., Karsch, F. and Petersson, B. (1990). The acceptance probability in the Hybrid Monte Carlo method. *Phys. Lett. B* **242** 437–443.
[19] Hairer, E., Lubich, C. and Wanner, G. (2006). *Geometric Numerical Integration*: *Structure-Preserving Algorithms for Ordinary Differential Equations*, 2nd ed. *Springer Series in Computational Mathematics* **31**. Berlin: Springer. MR2221614
[20] Hairer, E., Nørsett, S.P. and Wanner, G. (1987). *Solving Ordinary Differential Equations I*: *Nonstiff Problems*. *Springer Series in Computational Mathematics* **8**. Berlin: Springer. MR0868663
[21] Hall, P. and Heyde, C.C. (1980). *Martingale Limit Theory and Its Application*. *Probability and Mathematical Statistics*. New York: Academic Press. MR0624435
[22] Hansmann, U.H.E., Okamoto, Y. and Eisenmenger, F. (1996). Molecular dynamics, Langevin and Hydrid Monte Carlo simulations in a multicanonical ensemble. *Chem. Phys. Lett.* **259** 321–330.
[23] Hasenbusch, M. (2001). Speeding up the Hybrid Monte Carlo algorithm for dynamical fermions. *Phys. Lett. B* **519** 177–182.
[24] Izaguirre, J.A. and Hampton, S.S. (2004). Shadow hybrid Monte Carlo: An efficient propagator in phase space of macromolecules. *J. Comp. Phys.* **200** 581–604.
[25] Leimkuhler, B. and Reich, S. (2004). *Simulating Hamiltonian Dynamics*. *Cambridge Monographs on Applied and Computational Mathematics* **14**. Cambridge: Cambridge Univ. Press. MR2132573
[26] Liu, J.S. (2008). *Monte Carlo Strategies in Scientific Computing*. *Springer Series in Statistics*. New York: Springer. MR2401592
[27] Mattingly, J.C., Pillai, N. and Stuart, A.M. (2012). Diffusion limits of random walk Metropolis algorithms in high dimensions. *Ann. Appl. Probab.* **22** 881–930.

[28] Mehlig, B., Heermann, D.W. and Forrest, B.M. (1992). Exact Langevin algorithms. *Molecular Phys*. **76** 1347–1357. MR1181343
[29] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys*. **21** 1087–1092.
[30] Mohamed, L., Christie, M. and Demyanov, V. (2009). Comparison of stochastic sampling algorithms for uncertainty quantification. Technical report, Institute of Petroleum Engineering, Heriot-Watt Univ., Edinburgh. SPE Reservoir Simulation Symposium.
[31] Neal, R.M. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical report, Dept. Computer Science, Univ. Toronto.
[32] Neal, R.M. (1996). *Bayesian Learning for Neural Networks*. New York: Springer.
[33] Pangali, C.S., Rao, M. and Berne, B.J. (1978). On a novel Monte Carlo scheme for simulating water and aqueous solutions. *Chem. Phys. Lett*. **55** 413–417.
[34] Pasarica, C. and Gelman, A. (2010). Adaptively scaling the Metropolis algorithm using expected squared jumped distance. *Statist. Sinica* **20** 343–364. MR2640698
[35] Pillai, N.S., Stuart, A.M. and Thiery, A.H. (2012). Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. To appear.
[36] Roberts, G.O., Gelman, A. and Gilks, W.R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab*. **7** 110–120. MR1428751
[37] Roberts, G.O. and Rosenthal, J.S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol*. **60** 255–268. MR1625691
[38] Roberts, G.O. and Rosenthal, J.S. (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statist. Sci*. **16** 351–367. MR1888450
[39] Roberts, G.O. and Tweedie, R.L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2** 341–363. MR1440273
[40] Rossky, P.J., Doll, J.D. and Friedman, H.L. (1978). Brownian dynamics as smart Monte Carlo simulation. *J. Chem. Phys*. **69** 4628–4633.
[41] Sanz-Serna, J.M. and Calvo, M.P. (1994). *Numerical Hamiltonian Problems. Applied Mathematics and Mathematical Computation* **7**. London: Chapman & Hall. MR1270017
[42] Schütte, C. (1998). Conformational dynamics: Modelling, theory, algorithm, and application of biomolecules. Habilitation thesis, Dept. Mathematics and Computer Science, Free Univ. Berlin.
[43] Sexton, J.C. and Weingarten, D.H. (1992). Hamiltonian evolution for the Hybrid Monte Carlo algorithm. *Nuclear Phys. B* **380** 665–677.
[44] Skeel, R.D. (1999). Integration schemes for molecular dynamics and related applications. In *The Graduate Student's Guide to Numerical Analysis'*98 (*Leicester*). *Springer Ser. Comput. Math*. **26** 119–176. Berlin: Springer. MR1715033
[45] Tuckerman, M.E., Berne, B.J., Martyna, G.J. and Klein, M.L. (1993). Efficient molecular dynamics and hybrid Monte Carlo algorithms for path integrals. *J. Chem. Phys*. **99** 2796–2808.
[46] Zlochin, M. and Baram, Y. (2001). Manifold stochastic dynamics for Bayesian learning. *Neural Comput*. **13** 2549–2572.

Received February 2011 and revised October 2011