The acceptance probability of the Hybrid Monte Carlo method in high-dimensional problems

A. Beskos[*], N. S. Pillai[†], G. O. Roberts[†], J. M. Sanz-Serna[**] and A. M. Stuart[‡]

[*]Department of Statistical Science, UCL, Gower Street, London, WC1E 6BT, UK
[†]Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK
[**]Departamento de Matemática Aplicada, Facultad de Ciencias, Universidad de Valladolid, Spain
[‡]Mathematics Institute, University of Warwick, Coventry, CV4 7AL, UK

Abstract. We investigate the properties of the Hybrid Monte-Carlo algorithm in high dimensions. In the simplified scenario of independent, identically distributed components, we prove that, to obtain an $\mathscr{O}(1)$ acceptance probability as the dimension d of the state space tends to ∞, the Verlet/leap-frog step-size h should be scaled as $h = \ell \times d^{-1/4}$. We also identify analytically the asymptotically optimal acceptance probability, which turns out to be 0.651 (with three decimal places); this is the choice that optimally balances the cost of generating a proposal, which *decreases* as ℓ increases, against the cost related to the average number of proposals required to obtain acceptance, which *increases* as ℓ increases.

Keywords: Hybrid Monte Carlo, Hamiltonian dynamics, Verlet method, Geometric integration
PACS: 02.50.Ng, 02.70.Bf, 02.70.Tt, 05.10.Ln

INTRODUCTION

The Hybrid Monte Carlo (HMC) algorithm originates from the physics literature [1] where it was introduced as a fast method for simulating molecular dynamics; it has since become popular in statistical physics, chemistry and neural networks. The algorithm has also been proposed for performing statistical inference in Bayesian contexts; nevertheless, HMC is still not a commonly used tool in the statistics literature.

HMC has been proposed as a method to improve on traditional Markov Chain Monte Carlo (MCMC) algorithms. There are heuristic arguments to suggest why HMC might perform better, for example based on the idea that it breaks the *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 and this lack of analytical guidance of choosing the free parameters for the algorithm partly accounts 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 and algotithms such as MALA [2] use the derivative of the target probability log-density to guide the Monte Carlo trajectory towards areas of high probability; this is to be compared with the standard Random-Walk Metropolis (RWM) algorithm which proposes symmetric moves around the current position. On the other hand, MALA takes local steps based on overdamped Langevin dynamics, while HMC takes *global* moves based on deterministic Hamiltonian dynamics.

In the simplified scenario where the target distribution $\Pi(Q)$ in $\mathscr{R}^N$ consists of $d \gg 1$ identically distributed vector components, we show analytically that in HMC (with the Verlet integrator) the time-step h has to be scaled as $h = \mathscr{O}(d^{-1/4})$ to ensure a nontrivial acceptance probability as $d \to \infty$. We also identify the associated optimal acceptance probability. Thus HMC requires $\mathscr{O}(d^{1/4})$ time-steps to traverse the state space, which compares favorably with the corresponding scalings $\mathscr{O}(d^1)$ and $\mathscr{O}(d^{1/3})$ that apply for the RWM and MALA algorithms respectively, see [3], [4] (the extension to non-product measures is considered in [5]).

Our analysis relies on ideas of geometric integration [6], [7], [8] and may be extended to time-integrators (other than the Verlet method) with favorable geometric properties: for a volume-preserving, time-reversible integrator of order ν (ν an even integer), the appropriate scaling is $h = \mathscr{O}(d^{-1/(2\nu)})$.

HYBRID MONTE CARLO

Assume that we wish to sample from a density Π on $\mathscr{R}^N$ with $\Pi(Q) = \exp\big(-\mathscr{V}(Q)\big)$, for $\mathscr{V} : \mathscr{R}^N \to \mathscr{R}$. HMC is based on the consideration of the Hamiltonian function

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

($\mathscr{M}$ is a user-defined symmetric positive definite matrix) that leads to the canonical differential equations

$$\frac{dQ}{dt} = \mathscr{M}^{-1}P, \qquad \frac{dP}{dt} = -\nabla\mathscr{V}(Q), \tag{1}$$

with corresponding solution flow Φ_t defined by$(Q(t),P(t)) = \Phi_t(Q(0),P(0))$. For each fixed $T > 0$, the transformation Φ_T in the phase space $\mathscr{R}^{2N}$ conserves the Hamiltonian $\mathscr{H}$ and the volume element $dQdP$ and is time-reversible, i.e. $(Q^*,P^*) = \Phi_T(Q,P)$ is equivalent to $(Q,-P) = \Phi_T(Q^*,-P^*)$. As a consequence, if we assume that the initial conditions $(Q(0),P(0))$ of (1) are random and distributed with a density (proportional to) $\exp(-\mathscr{H}(Q,P)) = \exp(-(1/2)\langle P,\mathscr{M}^{-1}P\rangle)\exp(-\mathscr{V}(Q))$, then the initial density is preserved under the application of the Hamiltonian flow Φ_T. It follows that the marginal density of $Q(T)$ will be (proportional to) $\exp(-\mathscr{V}(Q))$, and the mapping $Q(0) \mapsto Q(T)$ may be used to define the transitions of a Markov chain in $\mathscr{R}^N$ invariant with respect to the target Π.

In practice, the analytic expression of Φ_T will not be available and it is necessary to resort to numerical approximations. The most popular *explicit* integrator is the second-order accurate Störmer-Verlet-leapfrog scheme[1] defined as follows: assuming a current state (Q_0,P_0), after one time-step of length $h > 0$ the system (1) will be at a state (Q_h,P_h) defined by the three-stage procedure

$$P_{h/2} = P_0 - \frac{h}{2}\nabla\mathscr{V}(Q_0), \quad Q_h = Q_0 + h\mathscr{M}^{-1}P_{h/2}, \quad P_h = P_{h/2} - \frac{h}{2}\nabla\mathscr{V}(Q_h).$$

The scheme gives rise to a map: $\Psi_h\colon (Q_0,P_0) \mapsto (Q_h,P_h)$ which approximates the flow Φ_h. The solution at time T is approximated by taking $\lfloor \frac{T}{h} \rfloor$ leapfrog steps:

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

The map $\Psi_h^{(T)} := \Psi_h^{\lfloor \frac{T}{h} \rfloor}$ is easily seen to be volume preserving and time-reversible (i.e. $(Q^*,P^*) = \Psi_h^{(T)}(Q,P) \Leftrightarrow (Q,-P) = \Psi_h^{(T)}(Q^*,-P^*)$), but it does not exactly conserve the energy $\mathscr{H}$ and therefore it does not preserve the measure $\exp(-\mathscr{H}(Q,P))dQdP$. In order to define a chain that preserves[2] Π, an accept/reject rule is introduced. More precisely, the Markov transitions $Q^{(n)} \mapsto Q^{(n+1)}$ in HMC are defined by:

(i) Given $Q^{(n)}$, sample a momentum $P^{(n)} \sim N(0,\mathscr{M})$.

(ii) Compute $(Q',P') = \Psi_h^{\lfloor \frac{T}{h} \rfloor}(Q^{(n)},P^{(n)})$, and calculate $a = a((Q^{(n)},P^{(n)}),(Q',P'))$, where a is the function

$$a((Q,P),(Q^*,P^*)) := 1 \wedge \exp\{\mathscr{H}(Q,P) - \mathscr{H}(Q^*,P^*)\}.$$

(iii) Set $Q^{(n+1)} = Q'$ with probability a; otherwise set $Q^{(n+1)} = Q^{(n)}$.

Here we consider the simplified scenario where $\Pi(Q)$ consists of $d \gg 1$ independent identically distributed vector components,

$$\Pi(Q) = \exp\big\{-\sum_{i=1}^{d} V(q_i)\big\}, \quad V : \mathscr{R}^m \to \mathscr{R}, \quad N = m \times d.$$

We investigate the asymptotic behaviour of the HMC algorithm when the number d of 'particles' goes to infinity. 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: $X = (x_i)_{i=1}^d$; $\quad x_i = (q_i,p_i) \in \mathscr{R}^{2m}$.

[1] An alternative integrator, suited for problems of large dimensionality, has been introduced in [9].

[2] In fact the resulting chain will be reversible with respect to Π.

We have

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

where M is a $m \times m$ symmetric, positive definite matrix. The Hamiltonian differential equations for a single (m-dimensional) particle are then

$$\frac{dq}{dt} = M^{-1}p, \qquad \frac{dp}{dt} = -\nabla V(q)\ ,$$

where $V : \mathscr{R}^m \to \mathscr{R}$. We denote the corresponding flow by ϕ_t and the leapfrog solution operator over one h-step by ψ_h. Thus, the acceptance probability function a for the evolution of the d particles is given by:

$$a(X,X^*) = 1 \wedge \exp(R_d), \qquad R_d := \sum_{i=1}^{d} \big[H(x_i) - H(\psi_h^{(T)}(x_i))\big], \tag{2}$$

with $X^* = \Psi_h^{(T)}(X)$ denoting the HMC proposal. Note that the leapfrog scheme is applied independently for each of the d particles (q_i,p_i); the different co-ordinates are only connected through the accept/reject decision based on (2).

ANALYSIS

The analysis starts by estimating the sum in (2). Since the d particles play the same role, it is sufficient to study a single term $H(x_i) - H(\psi_h^{(T)}(x_i))$. We then set

$$\Delta(x,h) := H(\psi_h^{(T)}(x)) - H(\phi_T(x)) = H(\psi_h^{(T)}(x)) - H(x);$$

this is the energy change, due to the leapfrog scheme, over $0 \le t \le T$, with step-size h and initial condition x. We will study the first and second moments

$$\mu(h) := \mathscr{E}\,[\Delta(x,h)] = \int_{\mathscr{R}^{2m}} \Delta(x,h)\, e^{-H(x)} dx, \qquad s^2(h) := \mathscr{E}\,[|\Delta(x,h)|^2]$$

and the corresponding variance $\sigma^2(h) = s^2(h) - \mu^2(h)$. If the integrator were exactly energy-preserving, one would have $\Delta \equiv 0$ and all proposals would be accepted. However it is well known that the size of $\Delta(x,h)$ is in general no better than the size of the integration error $\psi_h^{(T)}(x) - \phi_T(x)$, i.e. $O(h^2)$. In fact under natural smoothness assumptions on V (see [10]) the following conditions hold:

Condition 1 *There exist functions* $\alpha(x)$, $\rho(x,h)$ *such that* $\Delta(x,h) = h^2\alpha(x) + h^2\rho(x,h)$ *with* $\lim_{h\to 0}\rho(x,h) = 0$.

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

$$\sup_{0\le h\le 1} \frac{|\Delta(x,h)|^2}{h^4} \le D(x), \qquad \int_{\mathscr{R}^{2m}} D(x)\, e^{-H(x)} dx < \infty\ .$$

The key point is now that, even though $\Delta(x,h) = \mathscr{O}(h^2)$ at each fixed x, the *expectation* $\mu(h)$ is much smaller, as a consequence of the geometric properties of the Verlet integrator. In fact, if we set $(q^*,p^*) = \psi_h^{(T)}(q,p)$ where $x = (q,p)$ is the generic point in $\mathscr{R}^{2m}$, then, by time-reversibility, $(q,-p) = \psi_h^{(T)}(q^*,-p^*)$ and therefore

$$\Delta\big((q,p),h\big) = H(q^*,p^*) - H(q,p) = -\big(H(q,-p) - H(q^*,-p^*)\big) = -\Delta\big((q^*,-p^*),h\big),$$

thus as (q,p) varies in $\mathscr{R}^{2m}$ the energy increments appear in pairs that cancel each other. Furthermore, by conservation of volume, $dq\,dp = dq^*\,d(-p^*)$, and therefore the integral of $\Delta(x,h)$ with respect to the standard Lebesgue measure $dq\,dp$ is exactly 0, the contribution to the integral of $\Delta((q,p),h)dq\,dp$ is cancelled by the contribution $\Delta((q^*,-p^*),h)dq^*\,d(-p^*)$. In μ, the integration of $\Delta(x,h)$ has to be taken with respect to the invariant measure $\exp(-H(x))\,dx$ and the cancellation in the integral is only partial because the contributions of the $\mathscr{O}(h^2)$ values

$\Delta((q,p),h)$ and $-\Delta((q^*,-p^*),h)$ are weighted by the respective factors $\exp(-H(q,p))$ and $\exp(-H(q^*,-p^*))$ that are only equal up to an $\mathscr{O}(h^2)$ remainder. These ideas make it possible to prove the following result:

Theorem 1 *If the potential V is such that Conditions 1 and 2 hold for the leapfrog integrator $\psi_h^{(T)}$, then*

$$\lim_{h\to 0} \frac{\mu(h)}{h^4} = \mu\ , \qquad \lim_{h\to 0} \frac{\sigma^2(h)}{h^4} = \Sigma, \tag{3}$$

for the constants

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

The proof of the next result is based on the observation that in the sum R_d in (2) the random variables added possess expectations and variances that are both $\mathscr{O}(h^4)$ in view of (3); therefore the scaling $h = \ell \cdot d^{-1/4}$ is natural to have a distributional limit. Under such a scaling a Central Limit Theorem applies and R_d is asymptotically $N(-\ell^4\Sigma/2, \ell^4\Sigma)$. These considerations are the starting point to prove our main result:

Theorem 2 *Under the hypotheses of Theorem 1 assume that $h = l \cdot d^{-1/4}$, for a constant $l > 0$. Then at stationarity, i.e., for $X \sim \exp\{-\mathscr{H}\}$,*

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

(Here Φ denotes the standard $N(0,1)$ distribution function.)

Finally, under the scaling $h = \ell \cdot d^{-1/4}$, it is reasonable to accept[3] that eff $:= a(\ell) \times \ell$ is a sensible metric to assess the efficiency of the algorithm: the factor $a(\ell)$ reflects the fact that the efficiency increases as the acceptance probability increases, while the factor ℓ accounts for the fact that a larger value of ℓ implies fewer Verlet time-steps for given T.

Theorem 3 *Under the hypotheses of Theorem 2 and as $d \to \infty$, the measure of efficiency eff $:= a(\ell) \times \ell$ is maximized for the choice l_{opt} of l that leads to the value of $a = a(l)$ that maximises*

$$a \cdot \big(\Phi^{-1}\big(1 - \frac{a}{2}\big)\big)^{\frac{1}{2}}.$$

Rounded to 3 decimal places the, target independent, optimal value of the limit probability a is $a(l_{opt}) = 0.651$.

The conclusion is that when running the algorithm, the parameters should be set to obtain acceptance ratios close to 0.65.

REFERENCES

1. S. Duane, A. D. Kennedy, B. Pendleton, and D. Roweth, *Phys. Lett. B.* **195**, 216–222 (1987).
2. G. O. Roberts, and R. L. Tweedie, *Bernoulli* **2**, 341–363 (1996).
3. G. O. Roberts, A. Gelman, and W. R. Gilks, *Ann. Appl. Probab.* **7**, 110–120 (1997).
4. G. O. Roberts, and J. S. Rosenthal, *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60**, 255–268 (1998).
5. A. Beskos, G. O. Roberts, and A. M. Stuart, *Ann. Appl. Probab.* **19**, 863–898 (2009).
6. J. M. Sanz-Serna, and M. P. Calvo, *Numerical Hamiltonian Problems*, Chapman & Hall, London, 1994.
7. E. Hairer, Ch. Lubich, and G. Wanner, *Geometric Numerical Integration, 2nd ed.*, Springer, Berlin, 2006.
8. B. Leimkuhler, and S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge University Press, Cambridge, 2004.
9. A. Beskos, F. Pinski, J. M. Sanz-Serna, and A. M. Stuart, Hybrid Monte-Carlo on Hilbert Spaces, Technical Report (2010), `http://hermite.mac.cie.uva.es/sanzserna`.
10. A. Beskos, N. S. Pillai, G. O. Roberts, J. M. Sanz-Serna, and A. M. Stuart, Optimal Tuning of the Hybrid Monte Carlo Algorithm, Technical Report (2010), `http://hermite.mac.cie.uva.es/sanzserna`.

[3] See [10] for a precise analysis of this point.