Haussdorf School on MCMC: Recent developments and new connections, September 21–25, 2020

NUMERICAL INTEGRATORS FOR THE HAMILTONIAN MONTE CARLO METHOD, LECTURE I

J. M. Sanz-Serna
Universidad Carlos III de Madrid

1: INTRODUCTION

• Wish to obtain (possibly correlated) samples $q^{(0)}$, $q^{(1)}$, ... from target pdf of the form $\propto \exp(-V(q))$, $q \in \mathbb{R}^d$. (This assumes density > 0 everywhere. This hypothesis is not essential.)

- Statistical mechanics/ Molecular dynamics: q configuration variables, V potential energy, target is Boltzmann distribution.

- Bayesian statistics: target pdf $\pi(\theta)$. Then $q = \theta$, $V(q) = -\mathcal{L}(\theta)$ is the negative log-likelihood of target.

• HMC (Hybrid/Hamiltonian Monte Carlo) (Duane et. al. 1987) is a very popular sampling method.

• HMC is an MCMC method: Generates trajectories $q^{(0)} \mapsto q^{(1)} \mapsto \cdots \mapsto q^{(n)} \mapsto \ldots$ of Markov chain with target as invariant distribution.

• HMC is based on ideas from Hamiltonian mechanics and statistical physics.

• I will start by providing some background.

2. STATISTICAL PHYSICS

• For a conservative mechanical system, Newton's second law reads

$$M\ddot{q} = -\nabla V(q),$$

($q \in \mathbb{R}^d$ collects the positions, d is the number of degrees of freedom, the symmetric, positive definite matrix M contains the masses and V is the potential energy).

• As t varies, the total energy $(1/2)\dot{q}(t)^T M\dot{q}(t) + V(q(t))$ is conserved.

• Now assume that the system, rather than being isolated from the environment, is inside a heat bath at constant (absolute) temperature $1/\beta$. (Think of a protein inside the human body.) Molecules of the heat bath hit the system and interchange energy with it.

• Keeping track of all interchanges is impossible and a statistical description is needed. (Maxwell, Boltzmann, Gibbs,. . .)

• Statistical mechanics uses the Hamiltonian formulation of mechanics. This introduces a new independent variable $p = M\dot{q}$ (momentum). The space $\mathbb{R}^d \times \mathbb{R}^d$ of pairs (q, p) is the phase space.

Newton's second law is rewritten as the first-order system

$$\dot{q} = M^{-1}p, \qquad \dot{p} = -\nabla V(q),$$

i.e. in the symmetric form, due to Hamilton:

$$\dot{q} = \partial H/\partial p, \qquad \dot{p} = -\partial H/\partial q,$$

where $H(q, p) = (1/2)p^T M^{-1}p + V(q)$ is the total energy of the system expressed as a function of q and p. (H defined up to an additive constant.)

• In a heat bath $q(t), p(t)$ evolve stochastically so as to preserve the canonical probability measure: $d\mu = (1/Z)\exp(-\beta H(q, p))\, dqdp$, where Z is the normalizing constant $\int_{\mathbb{R}^d \times \mathbb{R}^d} \exp(-\beta H) dqdp$ (partition function).

- In view of the product structure

$$\exp(-\beta H(q,p)) = \exp\Big(-\beta(1/2)p^T M^{-1} p\Big) \times \exp\Big(-\beta V(q)\Big),$$

q and p are stochastically independent.

- The momenta have a Gaussian density

$$\propto \exp(-\beta(1/2)p^T M^{-1} p)$$

(Maxwell's distribution). From here it follows that the average kinetic energy is $1/(2\beta) \times d$: the absolute temperature $1/\beta$ is twice the average kinetic energy per degree of freedom.

- The positions q have the Boltzmann density $\propto \exp(-\beta V(q))$: minima of the potential energy are modes of the probability. As the temperature diminishes those minima carry more and more probability.

3. SYMMETRIES OF THE HAMILTONIAN DYNAMICS

- For the Hamiltonian system

$$\dot{q} = \frac{\partial H}{\partial p}, \qquad \dot{p} = -\frac{\partial H}{\partial q},$$

with arbitrary H, denote by $\varphi_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$ the solution flow, i.e. $\varphi_t(q, p)$ is the value at time t of the solution with initial values (q, p) at the initial time $t = 0$.

- The flow has important geometric properties.

• For each t the flow preserves volume in phase space (Liouville): $\forall \Omega \subset \mathbb{R}^d \times \mathbb{R}^d$, $\varphi_t(\Omega)$ has the same $2d$-dimensional Lebesgue measure as Ω. [In fact, the flow has a stronger property: *symplecticness* (Poincaré).]

• The flow preserves energy: $H(\varphi_t(q,p)) = H(q,p)$.

• As a consequence, the flow preserves the canonical probability measure [$d\mu \propto \exp(-\beta H(q,p))\, dqdp$]: i.e. $\forall \Omega \subset \mathbb{R}^d \times \mathbb{R}^d$, $\varphi_t(\Omega)$ carries the same probability as Ω.

[But note that the —deterministic— Hamiltonian dynamics does not describe the random motions of the system in the heat bath.]

Time reversibility of classical mechanics.

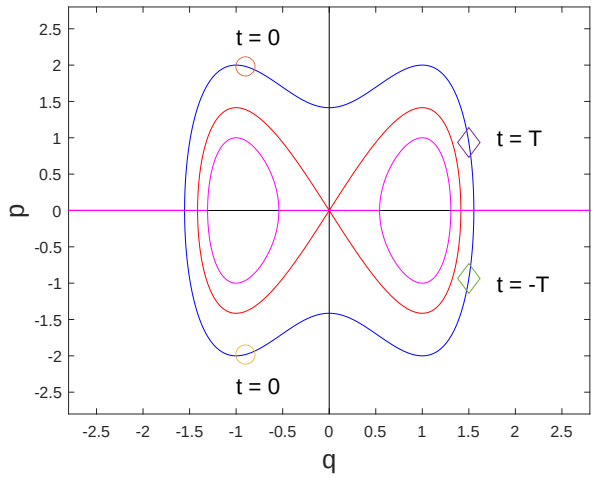
• For the special form $H(q,p) = (1/2)p^T M^{-1} p + V(q)$ we found above, the flow is reversible: if $\varphi_t(q,p) = (q^\star, p^\star)$, then $\varphi_t(q^\star, -p^\star) = (q, -p)$.

• Another formulation: if S denotes the momentum flip $(q,p) \mapsto (q,-p)$ ($S \circ S = Id$), then $S \circ \phi_t$ is an involution: $(S \circ \phi_t) \circ (S \circ \phi_t) = Id$.

• And yet another: $\phi_{-t} = (\phi_t)^{-1} = S \circ \phi_t \circ S$.

• Note S lets H invariant: $H \circ S = H$ and leaves invariant the Lebesgue measure in phase-space $dqdp$.

- Double-well potential $V(q) = (q^2 - 1)^2$ (bimodal distribution).



• For simplicity I'll set hereafter $\beta = 1$, but using different temperatures may of course be useful when sampling, because at higher temperatures moving between different probability modes becomes easier (tempering).

4. THE ALGORITHM

A Markov chain that preserves $\exp(-V(q))dq$:

In the phase space of the variable (q, p) consider the Hamiltonian system with $H = (1/2)p^T M^{-1} p + V(q)$ and its solution flow φ_T. [$T > 0$, M are parameters.]

- If $q^{(n)}$ is an element of the chain, then $q^{(n+1)}$ is defined as follows.
  - + Generate $p^{(n)}$ from pdf $\propto \exp(-(1/2)p^T M^{-1} p))$, independent from $q^{(n)}$ (and from past). (Momentum refreshment, needed for ergodicity.)
  - + Define $(q^{(n+1)}, \widetilde{p}^{(n+1)}) = (S \circ \varphi_T)(q^{(n)}, p^{(n)})$ [$\widetilde{p}^{(n+1)}$ will be discarded so S might have been omitted here].

- Proof: refreshment, Hamiltonian flow φ_T and momentum flip S preserve canonical probability measure $d\mu \propto \exp(-(1/2)p^T p - V(q))dqdp$ and hence the marginal on q (which is our target).

• Good news: by suitably choosing T, $q^{(n+1)}$ may be far away from $q^{(n)}$ (implications: low correlation, chain explores quickly $\mathbb{R}^d$) (cf. random walk Metropolis).

• Bad news: φ_T only known in trivial cases.

• Good idea: use a numerical approximation Ψ to φ_T, i.e. at each step of the Markov chain, integrate numerically the Hamiltonian dynamics with step-length h in the interval $0 \leq t \leq T$. If the integrator preserves exactly volume and energy then everything will work.

• Additional bad news: No numerical integrator preserves volume and energy (Ge and Mardsden 1988). Thus no Ψ preserves the canonical distribution μ. [An early result in Geometric Integration (SS 1995).]

• Additional good idea: Use an accept/reject mechanism to enforce conservation of μ.

ALGORITHM:

• Draw $p^{(n)} \sim \mathcal{N}(0, M)$. (Momentum refreshment.)

• From the initial condition $(q^{(n)}, p^{(n)})$ integrate numerically (see next slide) the Hamiltonian system of differential equations

$$\frac{d}{dt}q = M^{-1}p, \quad \frac{d}{dt}p = -\nabla V(q), \qquad 0 \le t \le T,$$

to get $(q^\star, p^\star)$. Proposal is $(q^\star, -p^\star) = S\big(\Psi(q^{(n)}, p^{(n)})\big)$.

• Calculate $a^{(n)} = \min\big(1, \exp(H(q^{(n)}, p^{(n)}) - H(q^*, (-)p^*))\big)$.

• Draw $u^{(n)} \sim U(0, 1)$. If $a^{(n)} \ge u^{(n)}$, set $q^{(n+1)} = q^*$ (acceptance); otherwise set $q^{(n+1)} = q^{(n)}$ (rejection).

• The numerical integration is performed by choosing and integer L, setting $h = T/L$ and performing L time-steps of length h:

$$(q_{j+1}, p_{j+1}) = \psi_h(q_j, p_j).$$

The approximation Ψ to φ_T is the L-fold composition $\psi_h \circ \cdots \circ \psi_h$.

• If this integrator is both volume-preserving and reversible, then, as I will prove later, the algorithm above is correct, in the sense that it preserves the target measure. [It would be possible to use more general integrators but then the accept/reject mechanism would have to be changed and become more complicated. See Fang, SS & Skeel, 2014.]

• I will not be concerned with the ergodicity/convergence to equilibrium of the algorithm (see Bou-Rabee, Eberle & Zimmer 2019 and its references). And I will not consider the many available variants.

- The acceptance rate

$$a^{(n)} = \min\left(1, \exp(H(q^{(n)}, p^{(n)}) - H(q^*, p^*))\right)$$

is a decreasing function of the change in energy over an integration leg

$$\Delta H(q^{(n)}, p^{(n)}) = H(q^*, p^*) - H(q^{(n)}, p^{(n)})$$

(recall that, conditional on $(q^{(n)}, p^{(n)})$, (q^*, p^*) is deterministic).

- Since, for the true solution $(q(t), p(t))$ starting at $(q^{(n)}, p^{(n)})$,

$$H(q(T), p(T)) = H(q^{(n)}, p^{(n)}),$$

we have (conservation of energy)

$$\Delta H(q^{(n)}, p^{(n)}) = H(q^*, p^*) - H(q(T), p(T))$$

and ΔH is also the integration error in H at the end of the integration leg.

• It follows that, for fixed T and $(q^{(n)}, p^{(n)})$, ΔH behaves as $\mathcal{O}(h^r)$ (r is the order of the integrator): by reducing h we may get acceptance rates arbitrarily close to 1 (but reducing the step-size implies more work to generate each proposal).

• By choosing T appropriately we expect to get proposals away from the current state $q^{(n)}$ of the Markov chain.

• These facts are summarized in the slogan: HMC may give proposals that are both away from the current state and accepted with high probability.

• But in practice things are not so simple: how to choose T and h, Hamiltonian dynamics may backtrack its progress, artifacts arise due to special choices of T and/or h—various remedies available.

5. THE INTEGRATOR

• At present the (second-order) Störmer/Verlet/leapfrog scheme is the integrator of choice. In its velocity form one time-step consists of three sub-steps

$$\begin{aligned} p_{i+1/2} &= p_i - \frac{h}{2}\nabla V_q(q_i), \quad \text{(kick)} \\ q_{i+1} &= q_i + h\, M^{-1} p_{i+1/2}, \quad \text{(drift)} \\ p_{i+1} &= p_{i+1/2} - \frac{h}{2}\nabla_q V(q_{i+1}). \quad \text{(kick)} \end{aligned}$$

• Note that, over a single time-step, it is volume-preserving because each sub-step is volume preserving. It follows that over L time-steps is also volume-preserving. Hence the proposal map $S \circ \Psi$ also preserves volume.

• Time-reversibility of Ψ follows from the palindromic structure.

• Cost: A gradient evaluation per time-step. The first kick at the present time-step reuses the gradient at the second kick of the previous time-step.

• A position version of the integrator also exists, with a drift/kick/drift structure...

• ...but the velocity form has some advantages (Bou-Rabee & SS 2018).

6. CORRECTNESS OF THE ALGORITHM

• Momentum refreshment obviously leaves $\exp(-H)dqdp$ invariant, so we only have to deal with the Markov substep based on numerical integration.

• I will use a result, due to C. Andrieu and his coworkers (who claim that the result covers the correctness of 99% MCMC algorithms).

Theorem: Let μ be a probability distribution on $(\Xi, \mathcal{Q})$, $\sigma : \Xi \to \Xi$ an involution ($\sigma^2 = Id$) and consider for $\xi, \xi' \in \Xi$ the kernel

$$P(\xi, d\xi') = a(\xi)\, \delta_{\sigma(\xi)}(d\xi') + [1 - a(\xi)]\, \delta_{\xi}(d\xi'),$$

(i.e. σ provides proposals) with acceptance ratio

$$a(\xi) = \min\{1, \frac{\eta(\sigma(\xi))}{\eta(\xi)}\},$$

where η is the density of μ with respect to a σ-invariant measure ν.

Then for all measurable sets A and B:

$$\int_A \pi(d\xi) P(\xi, B) = \int_B \pi(d\xi) P(\xi, A)$$

(P is μ-reversible) and in particular has μ as an invariant distribution.

• Apply to HMC, with Ξ the phase space of the variable $\xi = (q, p)$, ν the Lebesgue measure and $\mu \propto \exp(-H(\xi))dqdp$ so that $\eta = \exp(-H(\xi))$.

• HMC kernel and acceptance probability formula are of the form considered in the theorem with $\sigma = S \circ \Psi$.

• Two things remain to be checked:

- σ has to be an involution—but this is just the demand that the integrator is reversible $S \circ \Psi \circ S \circ \Psi = Id$.

- $dqdp$ has to be left invariant by $\sigma = S \circ \Psi$—but this is just the demand that the integrator preserves volume .

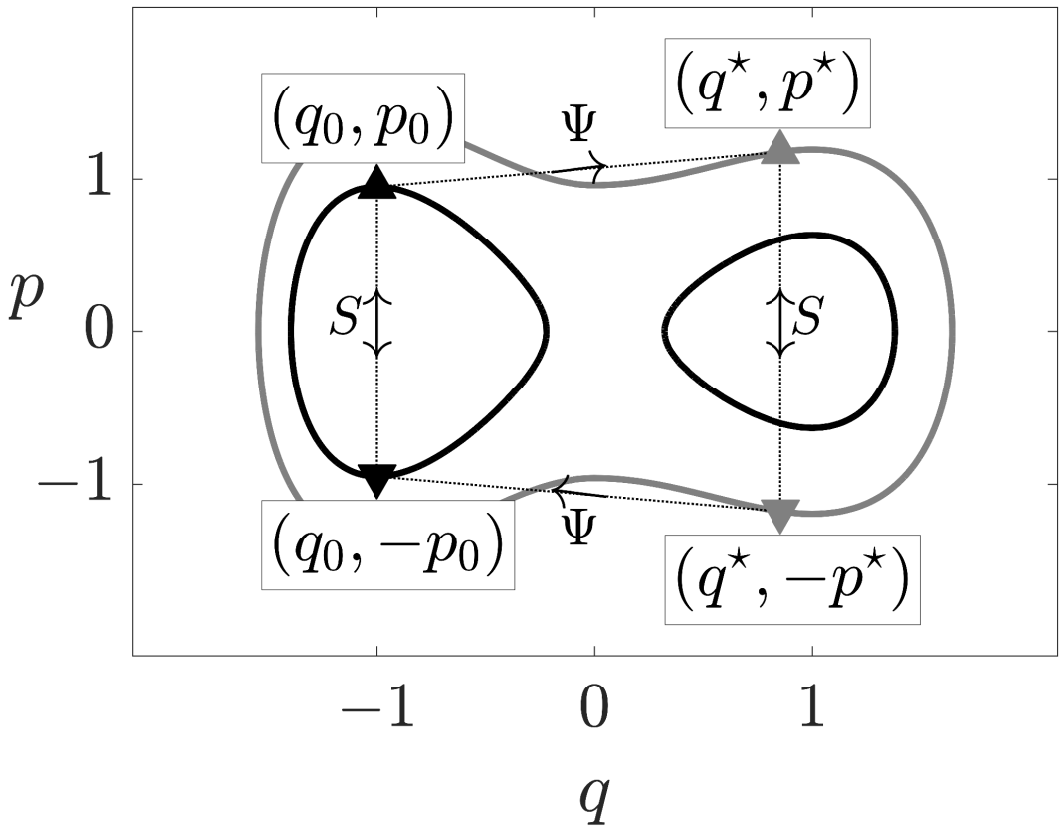
• The proof of Andrieu's result (inspired by a well-known 1998 paper by Tierney) hinges on the fact that

$$\min\left\{1, \exp\big(-H\big(S\circ\Psi(\xi)\big)\big)/\exp\big(-H(\xi)\big)\right\}$$

is the factor that when multiplying $\exp(-H(\xi))d\xi$ turns it into a measure

$$\min\left\{\exp\big(-H\big(S\circ\Psi(\xi)\big)\big), \exp\big(-H(\xi)\big)\right\} d\xi$$

that is invariant by the proposal map $\xi \mapsto S\circ\Psi(\xi)$. (The map switches the terms within the curly brackets —reversibility— and preserves $d\xi$ —volume preservation.)


(q_0, p_0)
Ψ
$(q^\star, p^\star)$
S
S
$(q_0, -p_0)$
Ψ
$(q^\star, -p^\star)$
p
1
0
-1
-1
0
1
q

• Note that all proposals with $\Delta H < 0$ are accepted.

• In addition, by using the arguments used to prove the reversibility of the chain, we may prove that if $\mathbb{P}(\Delta H) = 0$, then, at stationarity:

$$\mathbb{E}(a) = 2\mathbb{P}(\Delta H > 0) = \mathbb{P}(\Delta H > 0) + \mathbb{P}(\Delta H < 0).$$

More generally, even if $\mathbb{P}(\Delta H) \neq 0$,

$$\mathbb{P}(\Delta H > 0) = \mathbb{P}(\Delta H < 0)$$

7. OVERVIEW

In the remaining two lectures I will study the interplay between the numerical integrator and the sampling properties of HMC. More precisely:

• I will show that the properties of volume-preservation and reversibility have a big impact on the behaviour of the errors in the numerical integration. This has important implications on the acceptance rate and on the sampling properties of HMC.

• I will study whether the leapfrog algorithms is the best one may use within HMC.

I have aimed at a self-contained presentation that does not assume much background.