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

NUMERICAL INTEGRATORS FOR THE HAMILTONIAN MONTE CARLO METHOD, LECTURE II

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

8: AN IMPORTANT LEMMA

Beskos, Pillai, Roberts, SS & Stuart 2013
Bou-Rabee & SS 2018

• Denote the change in energy/energy error over an integration leg (i.e. when generating a single proposal) by

$$\Delta(q,p) = H(\Psi(q,p)) - H(q,p)$$

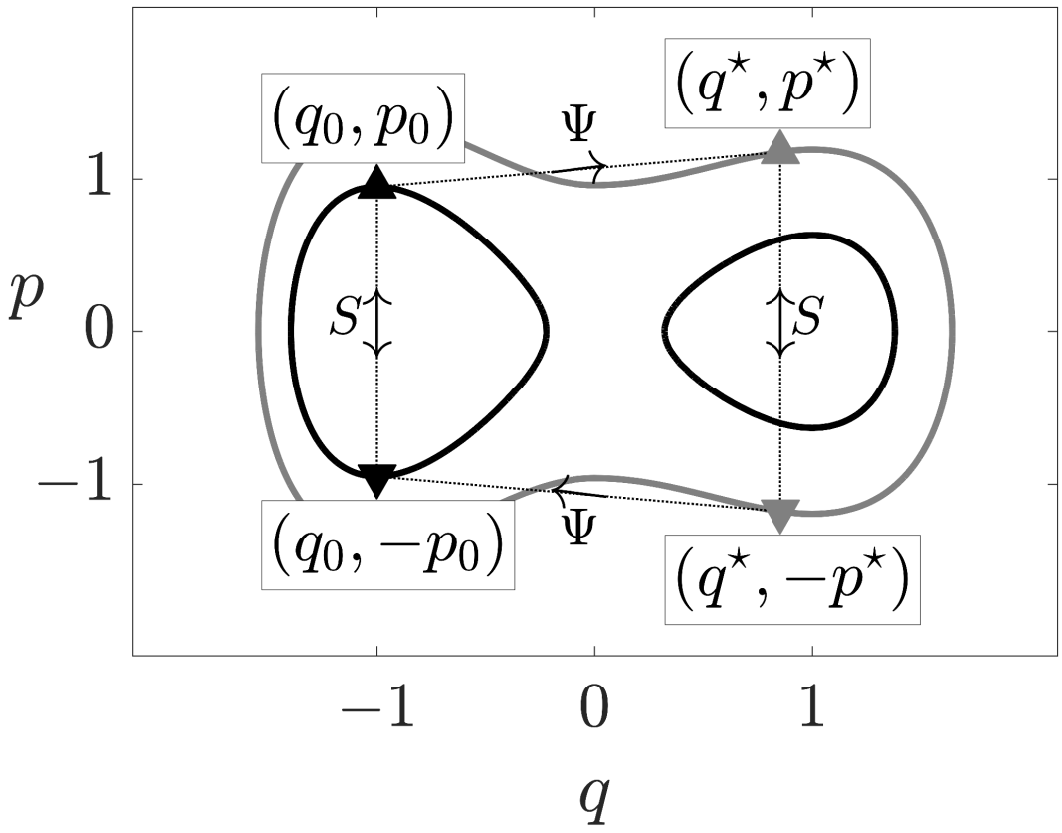
((q,p) is the initial point of the integration and $\Psi(q,p)$ the final point, so that the proposal is $S(\Psi(q,p))$).

• At each point in phase space $\Delta(q,p) = \mathcal{O}\big(h^r\big)$ as $h \downarrow 0$ with fixed T (r is the order of the integrator, 2 for leapfrog).

• The expected energy error at stationarity of chain is:

$$\mathbb{E}(\Delta) = \int_{\mathbb{R}^{2d}} \Delta(q,p) \exp\big(-H(q,p)\big)\, dq\, dp.$$

• We now take into account time reversibility. This implies that, in the next figure, the energy error at $(q^\star, -p^\star)$ is exactly the opposite of the energy error at (q_0, p_0):


(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

• Hence if we change variables in the integral and use the proposal as a new integration variable (renamed as (q, p))):

$$\mathbb{E}(\Delta) = -\int_{\mathbb{R}^{2d}} \Delta(q,p) \exp\big(-H(\Psi(q,p))\big)\, dq\, dp.$$

(the Jacobian of the transformation is 1 by conservation of volume).

• . . . and, taking the mean of both expressions for the expected error:

$$\begin{aligned}
&\mathbb{E}(\Delta) \\
&= \frac{1}{2}\int_{\mathbb{R}^{2d}} \Delta(q,p)\Big[\exp\big(-H(q,p)\big) - \exp\big(-H(\Psi(q,p))\big)\Big]\, dq\, dp \\
&= \frac{1}{2}\int_{\mathbb{R}^{2d}} \Delta(q,p)\Big[1 - \exp\big(-\Delta(q,p)\big)\Big]\exp\big(-H(q,p)\big)\, dq\, dp.
\end{aligned}$$

• Therefore, by Taylor expanding $1 - \exp(-\zeta) \approx \zeta$:

$$\boxed{\mathbb{E}(\Delta) \approx \frac{1}{2} \int_{\mathbb{R}^{2d}} \Delta(q,p)^2 \exp\big(-H(q,p)\big)\, dq\, dp}.$$

• In fact by taking care of the Taylor remainder, we have the rigorous bound:

$$0 \leq \mathbb{E}(\Delta) \leq \int_{\mathbb{R}^{2d}} \Delta(q,p)^2 \exp\big(-H(q,p)\big)\, dq\, dp.$$

• Hence, on average over the phase space the energy error is much smaller than one may have expected. And it is > 0 (except trivial case integration is exact).

• Leapfrog has $\mathbb{E}(\Delta) = \mathcal{O}\big(h^4\big)$ —rather than $\mathcal{O}\big((h^2\big)$ as one may have anticipated.

• Also the box shows that for h small the average of Δ will approximately coincide with half its second moment and hence $\mathbb{E}(\Delta) \approx (1/2)\mathrm{Var}(\Delta)$.

9: SCALING THE TIME-STEP AS THE DIMENSIONALITY INCREASES

Beskos, Pillai, Roberts, SS & Stuart 2013

- Consider HMC in the following scenario:
  - H is a fixed Hamiltonian $H(q_1, p_1) = (1/2)p_1^T M p_1 + \mathcal{U}(q_1)$ in $\mathbb{R}^d \times \mathbb{R}^d$; q_1 has density $\exp(-\mathcal{U}(q_1))$.
  - For $m = 1, 2, \ldots$ consider Hamiltonian $H_m(q, p) = \sum_{j=1}^m H(q_j, p_j)$ in $\mathbb{R}^{md} \times \mathbb{R}^{md}$, with $q = (q_1, \ldots, q_m)$, $q_j \in \mathbb{R}^d$ (and similarly for the momentum).
  - Wish to sample the variable q with density $\exp(-\sum_j \mathcal{U}(q_j))$.
  - The integration time T is independent of m.
- The density $\exp(-H_m)$ is a product of m copies of $\exp(-H)$ and the target $\exp(-\sum_j \mathcal{U}(q_j))$ is the product of m copies of $\exp(-\mathcal{U})$. The q_j are iid.

• In the Hamiltonian system of ODEs for H_m, the dynamics of the different (q_j, p_j) are not coupled to each other. The same is true for the numerical approximations.

• The different components come together in the accept/reject mechanism, where the acceptance probility is

$$\min\left\{1, \quad \exp\left(-\sum_{j=1}^{m} \Delta(q_j^{(n)}, p_j^{(n)})\right)\right\}.$$

• At stationarity, $\Delta(q_j^{(n)}, p_j^{(n)})$, $j = 1, \ldots, m$, are iid and as we know have a positive expectation. If h is chosen independently of m the acceptance probability will vanish as $m \to \infty$.

• How to scale h (and therefore the computational work of the algorithm) as m increases?

Theorem: If integrator satisfies some mild technical requirements, there is a constant Σ such that the expectation and variance for a single component

$$\mu_h = \int_{\mathbb{R}^d \times \mathbb{R}^d} \Delta(q_1, p_1; h) \exp(-H(q_1, p_1))\, dq_1 dp_1,$$
$$\sigma_h^2 = \int_{\mathbb{R}^d \times \mathbb{R}^d} \Delta(q_1, p_1; h)^2 \exp(-H(q_1, p_1))\, dq_1 dp_1 - (\mu_h)^2$$

satisfy

$$\lim_h \mu_h / h^{2r} = \Sigma/2, \qquad \lim_h \sigma_h^2 / h^{2r} = \Sigma.$$

(This is a rigorous version of the "expectation $\approx$ half the variance" thing we found before.)

- Σ depends both on the target and on the integrator and is unknown in practice.

• Thus for the overall system with m components the expectation of the error and the variance are $m\mu_h \approx (\Sigma/2)mh^{2r}$ and $m\sigma_h^2 \approx \Sigma m h^{2r}$.

• For a nontrivial distributional limit, scale h as

$$h = \ell/m^{2r}, \qquad \ell > 0.$$

• Use CLT to show that, under this scaling, in the limit $\to \infty$, the distribution of $\Delta(q,p)$ is $\boxed{\mathcal{N}(\ell^{2r}\Sigma/2, \ell^{2r}\Sigma)}$.

• The acceptance probability then converges to $\mathbb{E}\big(\min\{1, \exp(-D)\}\big)$ with $D \sim \mathcal{N}(\ell^{2r}\Sigma/2, \ell^{2r}\Sigma)$.

• . . . which may be computed in closed form as $\boxed{A(\ell) = 2\Phi\big(-\ell^r\sqrt{\Sigma}/2\big)}$, with Φ the cdf of the standard normal variable. (As ℓ increases —longer time-steps, less cost— $A(\ell)$ decreases.)

• For the leapfrog integrator (and other integrators to be presented later), then h has to be scaled as $h = \ell/m^4$. This implies a cost of $\mathcal{O}(h^{5/4})$ to make $\mathcal{O}(1)$ moves to explore the target at stationarity.

• Random Walk Metropolis and MALA the cost is $\mathcal{O}(h^2)$ and $\mathcal{O}(h^{4/3})$ respectively (Roberts, Gelman & Gilks 1997; Roberts & Rosenthal 2001).

(For transient in Random Walk Metropolis, see Jourdain, Lelievre & Miasojedow 2015.)

10: OPTIMAL TUNING

Beskos, Pillai, Roberts, SS & Stuart 2013

• Rejections are unwelcome: they waste computational effort and increase the correlation of the Markov chain.

• For any integrator the number of rejections may be made arbitrarily low by decreasing h.

• However decreasing h may be counterproductive as it increases the computational effort.

• A very low empirical percentage of rejections may signal that h is too small: better sampling would be possible by increasing the number of elements in the Markov chain and generating cheaper proposals by using a larger h (even if more proposals would be rejected).

• In the scenario we just considered, for m large, the user should choose ℓ so as to maximize $\ell A(\ell)$, whose reciprocal measures the cost of generating an accepted proposal.

• Unfortunately

$$A(\ell) = 2\Phi\big(-\ell^r\sqrt{\Sigma}/2\big)$$

remains unknown, because Σ is not available.

• However one may use A as an independent variable and then the function to maximize is given by

$$\frac{2^{1/r}}{\Sigma^{1/2r}}\left[A\,\left(\Phi^{-1}(1-A/2)\right)^{1/r}\right].$$

• Hence it suffices to maximize the function of A in square brackets. For $r = 2$ the maximum occurs at $A \approx 0.651$. A user observing higher (lower) acceptance rates should decrease (increase) h.

11: THE UNIVARIATE GAUSSIAN MODEL

Blanes, Casas & SS 2014

Calvo, Sanz-Alonso & SS 2019

• The performance of a numerical integrator used within HMC depends on its global error, ie on the error at the final time T after L time-steps of length h have been taken.

• While, in the numerical analysis of ODEs, local errors (ie errors after a single time-step) are easy to analyze (just Taylor expand), global errors are notoriously difficult to pin down. For instance, most ODE codes do not attempt to control or even estimate global errors.

• This is particularly true in integrations that are not very accurate.

• For this reason the numerical analysis of ODEs strongly relies on investigating in detail the behaviour of the integrator in well-chosen model problems.

• For Hamiltonian problems, model problem is the harmonic oscillator:

$$H = \frac{1}{2}(p^2 + q^2), \qquad q, p \in \mathbb{R},$$

with equations of motion

$$\frac{d}{dt}q = -p, \qquad \frac{d}{dt}p = q.$$

• In the HMC setting this corresponds to q having density $\propto \exp(-q^2/2)$, ie $\mathcal{N}(0,1)$ and choosing the mass matrix to be 1.

• The solution flow is a rotation by t radians in the phase plane. In matrix form:

$$\begin{bmatrix} q(t) \\ p(t) \end{bmatrix} = M_t \begin{bmatrix} q(0) \\ p(0) \end{bmatrix}, \qquad M_t = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}.$$

• For all numerical integrators of interest, the solution after a single time-step of length h is of the form

$$\begin{bmatrix} q_1 \\ p_1 \end{bmatrix} = \widetilde{M}_h \begin{bmatrix} q_0 \\ p_0 \end{bmatrix}, \qquad \widetilde{M}_h = \begin{bmatrix} A_h & B_h \\ C_h & D_h \end{bmatrix},$$

so that after L time-steps

$$\begin{bmatrix} q_L \\ p_L \end{bmatrix} = (\widetilde{M}_h)^L \begin{bmatrix} q_0 \\ p_0 \end{bmatrix}.$$

• Volume-preservation and reversibility read $A_h D_h - B_h C_h = 1$ and $A_h = D_h$. In particular the eigenvalues of M_h are a pair λ_h, $1/\lambda_h$.

• For instance, for leapfrog $A_h = D_h = 1 - h^2/2 \approx \cos h$, $B_h = h \approx \sin h$, $C_h = -h + h^3/4 \approx -\sin h$. For $h > 2$ the eigenvalues are real, one of them is > 1 and $(\widetilde{M}_h)^L$ grows unboundedly as L increases: instability. The case $h = 2$ is also unstable (double real eigenvalue and matrix does not diagonalize).

• An integrator is said to be stable for a given $h > 0$ if $(\widetilde{M}_h)^L$ remains bounded as L grows.

• If η is the largest number for which an integrator remains stable for $0 < h < \eta$ then $(0, \eta)$ is the stability interval of the integrator ($\eta = 2$ for leapfrog).

• Assuming the integrator is stable for the step-size h, $|A_h| \leq 1$ and there is a real θ_h with $A_h = \cos\theta_h$. Then

$$\widetilde{M}_h = \begin{bmatrix} \cos\theta_h & \chi_h \sin\theta_h \\ -\chi_h^{-1}\sin\theta_h & \cos\theta_h \end{bmatrix}$$

and, over L steps:

$$(\widetilde{M}_h)^L = \begin{bmatrix} \cos(L\theta_h) & \chi_h \sin(L\theta_h) \\ -\chi_h^{-1}\sin(L\theta_h) & \cos(L\theta_h) \end{bmatrix}.$$

• Numerical trajectory stays on an ellipse (true trajectory stays on level set $H = (1/2)(p^2 + q^2) = \text{constant}$: a circle).

• θ_h governs phase errors (in HMC may safely be ignored).

• χ_h governs shape of numerical orbits/energy errors. $\chi_h \equiv 1$ would be ideal (then numerical solution stay on circles, no energy error).

Theorem: The expectation at stationarity of the random variable $\Delta(q,p)$ is given by

$$\mathbb{E}(\Delta) = \sin^2(L\theta_h)\,\rho(h),$$

where

$$\rho(h) = \frac{1}{2}\left(\chi_h^2 + \frac{1}{\chi_h^2} - 2\right) = \frac{1}{2}\left(\chi_h - \frac{1}{\chi_h}\right)^2 \geq 0.$$

Accordingly

$$0 \leq \mathbb{E}(\Delta) \leq \rho(h).$$

• Note the bound is independent of the number L of time-steps in the integration leg.

- For leapfrog and $h < 2$, one finds

$$\mathbb{E}(\Delta) \leq \frac{h^4}{32(1 - \frac{h^2}{4})}.$$

- For $h \leq 1$ the expected energy error is $\leq 1/24$.

- Halving h to $h \leq 1/2$, leads to an expected energy error $\leq 1/480$.

Theorem: For the acceptance rate we have:

$$\mathbb{E}(a) = 1 - \frac{2}{\pi} \arctan \sqrt{\frac{\mathbb{E}(\Delta H)}{2}},$$

regardless of the (volume-preserving, time-reversible) integrator, the step-length h and the number L of time-steps in each integration leg.

- As $\mathbb{E}(\Delta) \downarrow 0$, $\mathbb{E}(a) = 1 - (\sqrt{2}/\pi)\sqrt{\mathbb{E}(\Delta)} + \mathcal{O}\big((\mathbb{E}(\Delta))^{3/2}\big)$, and, as a consequence, as $h \downarrow 0$, $L \uparrow \infty$ with $L\epsilon = T$,

$$\mathbb{E}(a) = 1 - \mathcal{O}(h^r).$$

- As $\mathbb{E}(\Delta) \uparrow \infty$: $\mathbb{E}(a) \sim 2\sqrt{2}/\sqrt{\mathbb{E}(\Delta)}$. For $\mathbb{E}(\Delta) = 100$, $\mathbb{E}(a) \approx 0.089$. Even large energy errors provide nontrivial acceptance rates.

- All moments of Δ are also functions of $\mu = \mathbb{E}(\Delta)$:

$$\begin{aligned} \mathbb{E}(\Delta^2) &= 2\,\mu + 3\,\mu^2, \\ \mathbb{E}((\Delta^3) &= 18\,\mu^2 + 15\,\mu^3, \\ \mathbb{E}((\Delta^4) &= 36\,\mu^2 + 180\,\mu^3 + 105\,\mu^4, \\ \cdots &= \cdots \end{aligned}$$

(Used later to prove CLT.)

- For μ small, the first of these relations just restates the "expectation $\approx$ half the variance" principle we found before.

12: OTHER GAUSSIAN TARGETS

Blanes, Casas & SS 2014

• The results for the one-degree-of-freedom Hamiltonian $(1/2)(p^2 + q^2)$ give information sobre other Gaussian targets:

• For univariate target $\mathcal{N}(0, \sigma^2)$, we have $H = (1/2)p^2 + (1/2)(q^2/\sigma^2)$ and $dq/dt = p$, $dp/dt = -q/\sigma^2$ with angular frequency $\omega = 1/\sigma$.

• Reduce to the case $\mathcal{N}(0, 1)$, $M = 1$ by scaling variables. New variables are q/σ, p and t/σ. (Note integrators are equivariant with respect to such scalings.)

• In particular stability for time-step Δt demands $h = \Delta t/\sigma < \eta$.

• Note here Δt is a dimensional time-step; h and η are non-dimensional.

• For a d-variate Gaussian target distribution perform orthogonal change of variables so that new variables are not correlated—turn ODEs for a system of d coupled linear oscillators into d scalar, uncoupled oscillators.

• Stability demands that $\omega_j \Delta t < \eta$, where ω_j are the frequencies (which with a unit mass matrix coincide with the precisions $1/\sigma_j$).

• Assuming stability one then may prove:

$$\boxed{\mathbb{E}(\Delta) \leq \sum_{j=1}^{d} \rho(\omega_j \Delta t)},$$

where ρ is the function we encountered in the univariate case.

13: THE GAUSSIAN CASE AS THE DIMENSION GROWS

Calvo, Sanz-Alonso & SS 2019

- Scenario:
  - For each $d = 1, 2, \ldots$ choose a centered Gaussian target with non-singular covariance matrix.
  - Assume wlog covariance matrix diagonal and unit mass matrix. (If not change variables.)
  - For each d use HMC: the integrator, the time-step and the length L of the integration interval are allowed to change with d.
- Compare with (product structure) scenario in Beskos et al.

• For each d, $\mathbb{E}(\Delta_d)$ is a sum $\sum_{j=1}^{d} \mathbb{E}(\Delta_{j,d})$ where the $\Delta_{j,d}$ are the contributions corresponding to each of the d uncorrelated variables.

• Theorem (CLT): If

$$\max_{1 \le j \le d} \mathbb{E}(\Delta_{j,d}) \to 0$$

and, for some $\mu \in [0, \infty)$,

$$\mathbb{E}(\Delta_d) = \sum_{\ell=1}^{d} \mathbb{E}(\Delta_{d,\ell}) \to \mu,$$

Then, as $d \uparrow \infty$:

- At stationarity, the distributions of the random variables Δ_d converge to the distribution $\boxed{\mathcal{N}(\mu, 2\mu)}$.
- The acceptance rates a_d satisfy $\mathbb{E}(a_d) \to \boxed{2\Phi(-\sqrt{\mu/2})}$.

- Note that, in the limit, the expected acceptance rate is again a function of the expected energy error.

- This will be important in the next lecture.