http://www.math.nyu.edu/faculty/goodman/teaching/MonteCarlo20/

Class notes: Monte Carlo methods
Week 7, Hamiltonian sampling, Cramer's theorem

Jonathan Goodman

October 21, 2020

1 Symplectic integrators, Hamiltonian samplers

Hamiltonian MCMC relies on the fact that the *flow map* or *solution map* of a Hamiltonian system is volume preserving. To explain that, suppose $y \in \mathbb{R}^d$ and $z = F(y)$ with $z \in \mathbb{R}^d$. The jacobian matrix is $J(y)$ with entries

$$J_{ij}(x) = \partial_{y_j} F_i(y) \ .$$

The mapping F is one to one if $F(y) \neq F(y')$ whenever $y \neq y'$. If $A \subset \mathbb{R}^d$ is any region, we write $F(A) = \{F(y) | y \in A\}$. This is the *image* of A under the mapping F. The mapping is *volume preserving* if $\text{vol}(F(A)) = \text{vol}(A)$ for any region A where F is defined. If F is one to one and differentiable, then F is volume preserving if and only if

$$\det(J(y)) = 1 \ , \quad \text{for all } y \ .$$

A volume preserving transformation also preserves probability density. Suppose $Y \sim \rho_y(y)$ and $Z = F(Y)$, then $\Pr(Z \in F(A)) = \Pr(Y \in A)$ for any set A, whether or not F is volume preserving. The probability density transformation formula has a jacobian factor. If $Z \sim \rho_z(z)$ (OK, terrible notation!), then if $z = F(y)$, then

$$\rho_z(z) = \det(J(y))^{-1} \rho_y(y) \ .$$

In particular, if F is volume preserving, then the PDF doesn't change, in the sense that

$$\rho_z(F(y)) = \rho_y(y) \ .$$

This makes volume preserving maps useful for MCMC.

The differential equations (9) and (10) from Week 6 are a *Hamiltonian system*. Associated to any ODE system is a *flow map* that maps the initial data to the solution after a fixed amount of time. Consider solving (9) and (10) with initial data

$$\begin{pmatrix} x(0) \\ p(0) \end{pmatrix} = \begin{pmatrix} x_0 \\ p_0 \end{pmatrix}$$

The solution a time t later is

$$\begin{pmatrix} x_t \\ p_t \end{pmatrix} = \begin{pmatrix} x(t) \\ p(t) \end{pmatrix}$$

The flow map is

$$F(x_0, p_0, t) = (x_t, p_t) \ .$$

We don't have a formula, but you could compute F by solving the differential equation system numerically. The flow map for a Hamiltonian system is volume preserving.

The Verlet method for solving the Hamilton equations uses variables $x_k \approx x(t_k)$ and $p_{k+\frac{1}{2}} \approx p(t_k + \frac{1}{2}\Delta t)$. Here, Δt is the time step and $t_k = k\Delta t$. The notation $t_{k+\frac{1}{2}} = t_k + \frac{1}{2}\Delta t$ fits with this. The algorithm is

$$x_{k+1} = x_k \quad + \Delta t\, p_{k+\frac{1}{2}} \tag{1}$$

$$p_{k+\frac{1}{2}} = p_{k-\frac{1}{2}} - \Delta t\, \nabla\phi(x_k)\ . \tag{2}$$

The Verlet method is second order accurate because the variables are *staggered*. the momentum variables are defined only in the midpoints of the x intervals and vice versa. The special structure of this Hamiltonian system makes this possible, where $\dot{x}$ depends only on p and $\dot{p}$ depends only on x. This Verlet method (there are variants with similar properties) is also volume preserving. For this define

$$y = \begin{pmatrix} x_k \\ p_{k-\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} x \\ p \end{pmatrix}\ ,\quad z = \begin{pmatrix} x_{k+1} \\ p_{k+\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} x' \\ p' \end{pmatrix}\ . \tag{3}$$

The time step equations may be written

$$x' = x + \Delta t\, p'$$
$$p' = x - \Delta t\, \nabla\phi(x)\ .$$

The jacobian is the 2×2 block matrix

$$J = \begin{pmatrix} \partial_x x' & \partial_p x' \\ \partial_p x' & \partial_p p' \end{pmatrix}\ .$$

Each of the blocks is a $d \times d$ matrix. It is important that $\partial_x x' = I + \Delta t \partial_x p' \neq I$. This is a consequence of the staggered scheme (3) Exercise 1 asks you to see how this leads to $\det(J) = 1$.

The Verlet method (3) has a property stronger than being volume preserving. It is *symplectic*. The Hamiltonian flow also has this property, and it is important for some change of variable tricks of classical mechanics. Any symplectic map is volume preserving. It takes some time to explain *symplectic form*, but once you've figured that out, the calculations to show a map is symplectic are easier than the calculations to show it's volume preserving.

The actual Hamiltonian sampler with finite Δt is derived from an ideal sampler with $\Delta t = 0$. The ideal sampler uses partial resampling and a combination of two "moves". One of the moves it to evolve according to Hamilton's equations (9) and (10) of Week 6. The fact that Hamilton's equations preserve the Gibbs Boltzmann distribution is a big part of Boltzmann's reasoning behind the Gibbs Boltzmann distribution. We know that Hamilton's equations preserve $\rho(x,p) = \frac{1}{Z}e^{-\beta H(x,p)}$ because the flow is volume preserving and also preserves

H. The other move is to resample the momentum variables P. Exercise 2 explains how this may be done.

The actual Hamiltonian sampler has to deal with the fact that the Verlet dynamics does not preserve H exactly. One view (almost universally accepted in the molecular dynamics community) is to take a small $\Delta t > 0$ and hope that the error from not solving Hamilton's equations exactly is small. The other view (almost universally accepted outside the molecular dynamics community) is to *Metropolize* (apply a Metropolis rejection) to preserve the Gibbs Boltzmann distribution exactly even for finite Δt. To do this (details omitted), you do some number of Verlet steps and compute the Hamiltonian at the new point and compute the Metropolis Hastings ratio and accept with the Metropolis Hastings probability. If you reject (this is the crucial and brilliant observation), you have to reverse the momentum variable, replacing P with $-P$.

2 Exercises

1. Consider the Verlet method (3) in one dimension. Write ϕ' for $\nabla\phi$. Assume ϕ'' exists and show that

$$\det \begin{pmatrix} \frac{dx'}{dx} & \frac{dx'}{dp} \\ \frac{dp'}{dx} & \frac{dp'}{dp} \end{pmatrix} = 1 \; .$$

2. Suppose we want to sample $X \sim \mathcal{N}(0, C)$, where $C = H^{-1}$ is a $d \times d$ covariance matrix with information matrix H. [Everywhere else this week, H is the Hamiltonian.] The random number generator can make i.i.d. scalar standard normals, which fit together to form the components of the d component standard normal $Z \sim \mathcal{N}(0, I)$. Find conditions on the matrices A and B so that the following linear iteration preserves $\mathcal{N}(0, C)$:

$$X_{k+1} = AX_k + BZ_k \; .$$

   Show that this is equivalent to the relation between a and b we had before in the scalar case. Give an algorithm to resample the momentum variables without changing the position variables in the canonical ensemble with inverse temperature β. *Hint.* Try $A - aI$. You might think independent resampling (heat bath or Gibbs sampler) is best because it forgets the old P. But it turns out to be better in many cases to keep the old momentum so that the sample keeps moving in the same direction.

3. This exercise requires large amounts of numerical experimentation and curiosity. Please write some paragraphs describing your research findings and conclusions.

   Consider a probability density $\rho(x) = e^{-\beta\phi(x)}$. A *well* (or *potential well*) is a local minimum of ϕ. The *depth* of the well is the minimum height

above the minimum it takes to go to another well. Consider the famous model *double well* potential

$$\phi(x) = (x^2 - 1)^2 + \alpha x \ .$$

For $\alpha = 0$ the wells have equal "depth" and are equally likely. For $\alpha > 0$ but small, the well near $x = -1$ is deeper roughly by 2α. Run an MCMC sampler (Gaussian proposal Metropolis) for various values of the proposal step size and inverse temperature and α. Make plots of X_k as a function of k. At high temperature and for long runs, these just look like noise. At low temperature, you will see the algorithm "get stuck" in one well or the other. Make plots of the autocorrelation function of X_k and X_k^2 for various parameter values and comment on the differences. Comment on the the best value of the proposal size at high and low temperatures. What is the acceptance probability as a function of the temperature for the best proposal size. A good proposal size is one that gives low auto-correlation time. It is a common problem in MCMC that the algorithm seems to have converged from looking at plots of X_k or even the auto-correlation function but has not converged. That can happen in this problem. Suggest some practical steps you might take for this problem to detect that there might be a problem.

4. Describe in as much detail as you can your thoughts on a project for the class. Say who your partners are if you have partners. Otherwise, say you're looking for a partner or want to work alone. If you con't have a specific topic, give a topic area or a field. If you don't have a field, say enough about your background and goals that I can suggest somethong.