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

Class notes: Monte Carlo methods
Week 6, Hamiltonian sampling, statistical physics
Jonathan Goodman
October 14, 2020

1 Background from statistical physics

Statistical mechanics is important to Monte Carlo. Many of the target applications for Monte Carlo come from statistical mechanics, sometimes with other subject names such as "physical chemistry" or "molecular dynamics". Moreover, ideas and intuition from statistical mechanics helps us understand high dimensional problems from statistics and other sources. The "physical" picture is helpful for Monte Carlo even if you're not interested in physics. Here we give some of the terminology, intuition, and elementary examples. It obviously is not a substitute for a real course.

A "statistical" description of a system might mean specifying a probability distribution of the possible states rather than specifying a specific state. A statistical description may be appropriate for a large complex system. An example would be the molecules that make up the air in a room. There are too many molecules to say where they all are. The precise locations are irrelevant for the properties we are interested in, such as pressure.

Notation. The notation this week will switch between two contradictory systems. Sometimes, $x \in \mathbb{R}^d$ will be the state and $\phi(x)$ will be the potential, with PDF $\rho(x) = \frac{1}{Z}e^{-\phi(x)}$. Sometimes, there will be momentum variables $p \in \mathbb{R}^d$. In that case, the state will be $y = (x,p) \in \mathbb{R}^{2d}$, The *Hamiltonian* (or small "h" *hamiltonian*) will be $H(x,p) = \phi(x) + \frac{1}{2}p^t M^{-1}p$, and the PDF will be $\rho(y) = \rho(x,p) = \frac{1}{Z}e^{-H(x,p)}$. If there is a temperature, it will be called T or possibly kT or k_BT. The parameter k or k_B is a physical constant called *Boltzmann's constant*. The PDF will be

$$\rho(x) = \frac{1}{Z(T)}e^{-\frac{\phi(x)}{kT}} \quad \text{or} \quad \rho(x,p) = \frac{1}{Z(T)}e^{-\frac{H(x,p)}{kT}} \ . \tag{1}$$

We may use the *inverse temperature*, which is $\beta = \frac{1}{k_BT}$. In that case the PDF is

$$\rho(x) = \frac{1}{Z(\beta)}e^{-\beta\phi(x)} \quad \text{or} \quad \rho(x,p) = \frac{1}{Z(\beta)}e^{-\beta H(x,p)} \ . \tag{2}$$

Either of these formulas is the *canonical ensemble. Ensemble* is a French word that means "set" or "collection". The canonical ensemble is a collection of states, the state space, together with a probability, or probability density, for each state.

The total energy of a moving system is the sum of its kinetic and potential energy. The potential energy is $\phi(x)$. The kinetic energy is $\frac{1}{2}p^tM^{-1}p$. The total energy is called the *hamiltonian.* There are situations where you neglect momentum and kinetic energy. If you do that, only the potential remains. In either case, the canonical ensemble PDF is proportional to $e^{-\beta \text{ energy}}$.

Here is a vague description of the picture that leads to the canonical distribution (1). You start with a system and some dynamics on the system that preserves energy. The *Boltzmann ergodic hypothesis* is that this dynamics visits every state on the *energy surface* equally often. This leads to a uniform probability distribution on the energy surface, which is called the *micro-canonical ensemble.* It is "micro" because it's smaller than the "full" canonical ensemble, being restricted to an energy surface. This is what a system would do if it no interaction with the outside world.

A canonical ensemble system is nearly but not completely isolated from the outside world. An example might the collection of air molecules in a room that interact mostly with each other but do sometimes bounce off the walls. The picture is that interacting with the outside world changes the energy a little bit but the internal dynamics still works to keep each state with a given energy equally likely. That makes the PDF of the system a function of energy alone. The canonical distribution (1) has this property.

What I called the "outside world" is sometimes called a *heat bath. Heat*, in this context, means thermal energy. Thermal energy means energy stored in small scale motion of parts (molecules) rather than large scale motion. If the air were all moving in one direction (wind), there would be kinetic energy that we don't count as heat. But the kinetic energy in the random motions of individual molecules is heat energy. The "bath" part of "heat bath" refers to our system being surrounded by a larger system, like someone in a bathtub surrounded by water (particularly if it's a small person in lots of water). The "total system" is the original system and the bath (outside world) together. The total system can conserve energy while energy flows back and forth (is "exchanged") between the original system and the heat bath (outside world). A completely isolated system would have a fixed energy, but a system in weak contact with a heat bath does not. It is contact with the heat bath that leads the system to the probability density (1). This is the origin of the term *heat bath algorithm.* You bring one variable X_j into its own probability density by "touching it to a heat bath".

The probability distributions (1) and (2) have the property that low energy states have higher probability than high energy states. The temperature parameter determines how strong is the preference for low energy. For high temperature, or low β, the energy matters less. For low temperature, high β only states near the minimum energy states have much weight.

2 Simulated annealing

Simulated annealing is an optimization algorithm that is based on the Gibbs Boltzmann distribution (2). This is used to find points near the *ground state*, which minimized ϕ. In this algorithm, you have a *cooling sequence* or cooling *schedule*, which is a sequence of inverse temperatures $\beta_n \to \infty$. For each n, you use some MCMC algorithm to sample

$$\rho_n(x) = \frac{1}{Z_n} e^{-\beta_n \phi(x)} \; .$$

When you are done with ρ_n, you lower the temperature and do it again. Let $X_k^{(n)}$ be the MCMC chain for β_n. The starting point of an MCMC chain is not "typical", but the ending state should be if the chain length is long relative to the auto-correlation time. Therefore, even if $X_0^{(n)}$ is not a good sample of ρ_n, the ending state $X_N^{(n)}$ should be. We take the starting point of the next chain to be the ending point of the last one, which is $X_0^{(n+1)} = X_N^{(n)}$. You choose the cooling schedule so that β_{n+1} is not very much larger than β_n. In that way, a sample $X_N^{(n)}$ should be a reasonable starting point to sample ρ_{n+1}.

Simulated annealing is used to overcome the problem of a complex *energy landscape* with many local minima. *Energy landscape* refers to the shape of the graph of $\phi(x)$. A simple energy landscape would be something like a parabolic bowl $\phi(x) = |x|^2$. A *local minimum* is an x_* so that $\phi(x) > \phi(x_*)$ if x is close enough to x_*. Technically, x_* is a *strict* local minimum of ϕ if there is an $r > 0$ so that if $|x - x_*| \leq r$, then $\phi(x) > \phi(x_*)$. A complex energy landscape typically has lots of local minima. The *energy barrier* or *depth* of a local minimum is the amount of energy that it takes to "escape". This is the minimum extra energy it takes to go from x_* to a state with equal or lower energy. Technically, a *path* from x_* to x_1 is a continuous function of t so that $x(t = 0) = x_*$ and $x(t = 1) = x_1$. The maximum potential on a path is $\max \phi(x(t))$. A path *escapes* the local minimum if $\phi(x_1) \leq \phi(x_*)$. The barrier is

$$\Delta\phi = \min_{esc} \max_{t \in [0,1]} \phi(x) \; .$$

The subscript *esc* means "escaping paths", which means paths with $x(0) = x_*$ and $x(1) = x_1$, and $|x_1 - x_*| \geq r$.

Local minima are the bane of traditional gradient based optimization methods. These methods try to find the global minimum of ϕ by taking steps, called *iterates* with $\phi(x^{(n+1)}) < \phi(x^{(n)})$. An example is simple gradient descent

$$x^{(n+1)} = x^{(n)} - s_n \nabla\phi(x^{(n)}) \; . \tag{3}$$

The *step size* parameter s_n is also called the *learning rate*. If this is small, then a Taylor series calculation shows ϕ decreases:

$$\begin{aligned}
\phi(x^{(n+1)}) &= \phi(x^{(n)}) - \left[\nabla\phi(x^{(n)})\right] s_n \nabla\phi(x^{(n)}) + O(s_n^2) \\
&= \phi(x^{(n)}) - s_n \left\|\nabla\phi(x^{(n)})\right\|^2 + O(s_n^2) \\
&< \phi(x^{(n)}) \; .
\end{aligned}$$

Typical optimization packages choose s_n to insure that ϕ decreases. They take gradient step (3) only as a proposal. Then they evaluate $\phi(x^{(n+1)})$. If $\phi(x^{(n+1)}) > \phi(x^{(n)})$, they reject (in out terminology) s_n and try again, typically with $\frac{1}{2}s_n$. The Taylor series calculation shows that you get descent if s_n is small enough.

3 Hamiltonian sampler

The Hamiltonian sampler is a way to build *momentum* into an MCMC sampler. *Momentum* is what makes things moving in some direction want to keep moving in that direction. Our first MCMC sampler was Metropolis rejection built on a symmetric Gaussian proposal with some proposal step size r. If the proposals from X_k and X_{k+1} are independent with mean zero, which makes them steps from a random walk. This makes them likely to go in orthogonal directions and have less net motion than if they had gone in the same direction. Hamiltonian sampling *augments* the state space and doubles the dimension by adding a momentum variable that remembers direction.

It's easy to think you want to carry auxiliary variables that make the sampler more systematic, but it's hard to find methods that preserve the target density ρ. That's what makes Hamiltonian sampling important, both as a tool and as a way to think about augmenting the state space.

The Hamiltonian sampler grew out of the Hamiltonian formalism of *classical mechanics*, which is a mathematical formalism for "$F = ma$" Newtonian mechanics. I explain how the sampler was invented using the picture from mechanics. But the sampler itself applies to problems that do not come from physics. The components of x_j do not have to be coordinates and the "mass matrix" (see below) can be any symmetric positive definite matrix. I hope the physics picture helps you see "what's going on." If the Hamiltonian stuff is new to you, be aware that it was "old hat" to the people who invented the Hamiltonian sampler. They knew about phase apace volume conservation and the Gibbs-Boltzmann distribution. Their contribution was realizing that it led to a powerful MCMC algorithm, which is a brilliant contribution.

Newtonian mechanics describes the motion of "particles" (objects) depending on the forces between them. The *position variables* consist of all the coordinates of all the particles collected into a vector $x(t) \in \mathbb{R}^d$. It may not be as simple as $d = 3n$ for n particles in three dimensions, for example, because some of the variables might be angles. The components of x may be called *generalized coordinates*, but I just call them coordinates. *Hamiltonian mechanics* is for forces that come from a *potential energy* function $\phi(x)$. The force on coordinate j is

$$F_j(x) = -\partial_{x_j}\phi(x) \; .$$

If the "mass" (generalized mass?) of coordinate j is m_j, then $F = ma$ for that coordinate is

$$m_j \frac{d^2 x_j}{dt^2} = -\partial_{x_j}\phi(x) \; . \tag{4}$$

In vector form, $F = ma$ becomes

$$M \frac{d^2}{dt^2} x = -\nabla_x \phi(x) \; . \tag{5}$$

The ∇_x means "gradient with respect to x". This is redundant now, because there is no other variable the gradient could be with respect to. The M on the left is a diagonal matrix with masses m_j on the diagonal. For the Hamiltonian sampler, you can use any symmetric positive definite "mass matrix".

The *momentum* variable corresponding to x_j is

$$p_j = m_j \frac{dx_j}{dt} \; .$$

The dynamical equations (4) can be written using the momentum variables in the form

$$\begin{aligned} \frac{dp_j}{dt} &= -\partial_{x_j} \phi(x) \\ \frac{dx_j}{dt} &= \frac{1}{m_j} p_j \; . \end{aligned}$$

The vector form is

$$\frac{d}{dt} p = -\nabla_x \phi(x) \tag{6}$$

$$\frac{d}{dt} x = M^{-1} p \; . \tag{7}$$

The equations (5) are "second order" because they involve second derivatives. The equations (6) and (7) are an equivalent set of first order equations. The first one (6) says that the force pushes on coordinate x_j by changing its momentum. The second one (7) says that x goes in the direction of its momentum, redirected by the mass matrix M. It is typical to take $M = cI$ in MCMC applications that don't come from physical problems. In that case, x moves in the direction of p. If ϕ is constant, then $F = -\nabla_x \phi = 0$ so there is no force. In this case p does not change and x moves in a straight line.

The pair of dynamical equations (6) and (7) may be expressed in terms of derivatives of the *Hamiltonian function* (or just *Hamiltonian*

$$H(x,p) = \phi(x) + \frac{1}{2} p^t M^{-1} p \; . \tag{8}$$

The first term is potential energy $\phi(x)$. The second term is *kinetic energy*, which is the energy of motion. If the mass matrix is diagonal, then the kinetic energy is

$$\text{KE} = \frac{1}{2} p^t M^{-1} p = \frac{1}{2} \sum_{j=1}^{d} \frac{1}{m_j} p_j^2 = \frac{1}{2} \sum_{j=1}^{d} m_j v_j^2 \; , \;\; v_j = \frac{d}{dt} x_j \; .$$

The last form may be familiar to people who have learned Newtonian mechanics. You can check that

$$\begin{aligned}\nabla_x H(x,p) &= \nabla_x \phi(x) \\ \nabla_p H(x,p) &= \nabla_p \frac{1}{2} p^t M^{-1} p = M^{-1} p \; .\end{aligned}$$

Therefore, the equations of motion may be written in *Hamiltonian form* as

$$\frac{d}{dt}\, p = -\nabla_x H(x,p) \tag{9}$$

$$\frac{d}{dt}\, x = \quad \nabla_p H(x,p) \; . \tag{10}$$

These equations are nearly symmetric, with ∇_x in the p equation and ∇_p in the x equation, except that the p equation has a minus sign. Why is it useful to put the Newtonian dynamics in Hamiltonian form? You can get a sense of their convenience by verifying *conservation of energy*, which is the fact that if x and p evolve according to Hamiltonian dynamics, then the Hamiltonian is constant. In problems from mechanics, the Hamiltonian is the total energy (potential plus kinetic), but in other problems, it's just the Hamiltonian. The conservation law is

$$\frac{d}{dt}\, H(x(t),p(t)) = 0 \; . \tag{11}$$

The calculation behind this is just the chain rule, which uses the minus sign in (9). The arguments x and p are left out after the first line to un-clutter the formulas

$$\begin{aligned}\frac{d}{dt}\, H(x(t),p(t)) &= \left[\nabla_x H(x(t),p(t))\right]^t \frac{d}{dt}\, x + \left[\nabla_p H(x(t),p(t))\right]^t \frac{d}{dt}\, p \\ &= \left[\nabla_x H\right]^t \nabla_p H - \left[\nabla_p H\right]^t \nabla_x H \\ &= 0 \; .\end{aligned}$$

This calculation works even if the Hamiltonian is not a sum of the form (8). For example, it works if you add $x^t p$ to H. Hamiltonians like this have important use in mechanics, but not yet in MCMC.

In the rest of this section, we call x the *position* variable and p the *momentum.* We call the combined vector the *state* variable and denote it

$$y = \begin{pmatrix} x \\ p \end{pmatrix} \; .$$

The state variable has $2d$ components, which are d position components and d momentum components. If you know $x(0)$ and $p(0)$, then you can solve the dynamical equations (9) and (10) to calculate $x(t)$ and $p(t)$ for other t values. That justifies calling the combination the "state" of the mechanical system. The *state space* is the set of all possible states, which is R^{2d}. The Hamiltonian sampler samples a PDF in state space $f(x,p)$. This samples original

$$\rho(x) = \frac{1}{Z} e^{-\phi(x)} \; , \tag{12}$$

because ρ is the marginal density of f:

$$\rho(x) = \int_{\mathbb{R}^d} f(x,p)\, dp \; . \tag{13}$$

If you have a sequence of samples $Y_k = (X_k, P_k) \sim f$, then the position parts X_k are samples of ρ. This would be pointless if you had to use symmetric Gaussian proposal Metropolis to sample f. You would have a $2d$ dimensional sampler that most likely would be worse than the d dimensional sampler.

The Hamiltonian sampler samples the *Gibbs Boltzmann* distribution

$$f(x,p) = \frac{1}{Z} e^{-H(x,p)} \; . \tag{14}$$

[Gibbs is the American physicist and mathematician who discovered the "Gibbs phenomenon" in Fourier series. Boltzmann is the Austrian physicist who gave the modern definition of *entropy* and invented the term and the concept of *ergodic* dynamics.] This satisfies the marginal distribution formula (13) because of the sum structure (8) of the Hamiltonian, as we now verify. The value of the normalization constant Z is different in different places. The important thing is that each Z is a constant independent of x and p.

$$\begin{aligned}\int_{\mathbb{R}^d} f(x,p)\, dp &= \frac{1}{Z} \int_{\mathbb{R}^d} e^{-\phi(x) - \frac{1}{2}p^tMp}\, dp \\ &= \frac{1}{Z} e^{-\phi(x)} \int_{\mathbb{R}^d} e^{-\frac{1}{2}p^tMp}\, dp \\ &= \frac{1}{Z} e^{-\phi(x)}\end{aligned}$$

The crux is

$$\int_{\mathbb{R}^d} e^{-\frac{1}{2}p^tMp}\, dp = \;\; \text{something independent of } x.$$

I might call the function $e^{-\frac{1}{2}p^tMp}$ "Gaussian" or "quadratic exponential", but in this context, when it is a quadratic exponential of the momentum variable, it is called the *Maxwellian.* When the total energy (8) depends on the momentum through a position-independent quadratic, then the Gibbs Boltzmann distribution depends on the momentum as a Maxwellian. [Maxwell was a Scottish physicist who put the finishing touch on the equations that govern the interaction between electric and magnetic fields, now called *Maxwell's equations.* These explain the until then mysterious numerical relation $c = \frac{1}{\sqrt{\epsilon\mu}}$, where $c \approx 3 \cdot 10^8 \frac{\text{m}}{\text{sec}}$, ϵ is a constant related to electricity and μ is a constant related to magnetism. Maxwell suggested the Maxwellian distribution of momentum. He also wrote what is considered the first technical paper on control theory: *On Governors.* A *flyball governor* is a mechanical device that controls steam engines.]

4 Exercises