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

Class notes: Monte Carlo methods
Week 5, Auto-correlation, Hamiltonian sampling
Jonathan Goodman
October 7, 2020

1 Auto-correlation, Kubo formula, linear algebra

To review, we're talking about MCMC. We have a sequence X_k and a time series $V_k = V(X_k)$. The quantity we want and the quantity we have are

$$B = \mathrm{E}_\rho[V(X)] \ , \ \ \widehat{B}_N = \frac{1}{N} \sum_{k=1}^{N} V_k \ .$$

To estimate the error bar, we want

$$\sigma_N^2 = \mathrm{var}\Big(\widehat{B}_N \Big) \ .$$

This is given by

$$\sigma_N^2 = \frac{1}{N^2} \sum_{k=1}^{N} \sum_{j=1}^{N} \mathrm{cov}(V_j, V_k) \ .$$

The steady state lag t auto-covariance function is

$$C_t = \mathrm{cov}_\rho(V(X_0), V(X_t)) = \lim_{k \to \infty} \mathrm{cov}_{\rho_0}(V(X_k), V(X_{k+t})) \ .$$

The *Kubo formula* is the approximate formula that holds when N is large

$$\sigma_N^2 \approx \frac{1}{N} \sum_{k=-\infty}^{\infty} C_t \ . \tag{1}$$

This formula holds for "good" MCMC algorithms. See below for more on what makes a good algorithm. This is the formula we will explain. It is equivalent to what we had last week.

We write it as

$$\sigma_N^2 \approx \frac{S}{N} \ .$$

Here, S is the sum

$$S = \sum_{t=-\infty}^{\infty} C_t \ . \tag{2}$$

You factor out the *static variance*, which is $C_0 = \mathrm{var}_\rho(V(X))$.

$$\sigma_N^2 \approx \frac{1}{N_{\text{eff}}} \mathrm{var}_\rho(V(X)) \tag{3}$$

$$N_{\text{eff}} = \frac{N}{\tau} \tag{4}$$

$$\tau = \sum_{-\infty}^{\infty} D_t \ , \quad D_t = \frac{C_t}{C_0} = \mathrm{corr}_\rho(V(X_0), V(X_t)) \ . \tag{5}$$

We explain the Kubo formula (1) for a "good" MCMC algorithm. We call an MCMC algorithm "good" if it has the following properties:

$$\mathrm{cov}(V(X_j), V(X_k)) \to \mathrm{cov}_\rho(V(X_0), V(X_{|j-k|})) \quad \text{as } j \to \infty \ , \ \ k \to \infty \tag{6}$$

$$\mathrm{cov}(V(X_j), V(X_k)) \to 0 \quad \text{as } |j-k| \to \infty \tag{7}$$

In that case, if N is large, then most of the terms of the sum have j and k large enough for (6), which gives

$$\frac{1}{N^2} \sum_{k=1}^{N} \sum_{j=1}^{N} \mathrm{cov}(V_j, V_k) \approx \frac{1}{N^2} \sum_{k=1}^{N} \sum_{j=1}^{N} C_{|k-j|} \ .$$

Most of the terms here have k large and $N-k$ large. Therefore, for most terms we have

$$\sum_{j=1}^{N} C_{|k-j|} \approx \sum_{j=-\infty}^{\infty} C_{|k-j|} = \sum_{t=-\infty}^{\infty} C_t = S \ . \tag{8}$$

To see this, consider the terms added for $j \leq 0$. If k is large, then these added terms are

$$\sum_{-\infty}^{0} C_{|k-j|} \ .$$

We use the approximation (8) and get

$$\begin{aligned} \frac{1}{N^2} \sum_{k=1}^{N} \sum_{j=1}^{N} \mathrm{cov}(V_j, V_k) &\approx \frac{1}{N} \left\{ \frac{1}{N} \sum_{k=1}^{N} \left[\sum_{t=-\infty}^{\infty} C_t \right] \right\} \\ &= \frac{1}{N} \left\{ \frac{1}{N} \sum_{k=1}^{N} S \right\} \\ &= \frac{S}{N} \ . \end{aligned}$$

We look for a practical error bar estimator based on the Kubo formula (1). We start with the time series V_k, compute the sample mean $\overline{V} = \widehat{B}_N$, and the *empirical covariance function*

$$\widehat{C}_t = \frac{1}{N-t} \sum_{k=1}^{N-t} (V_k - \overline{V})(V_{k+t} - \overline{V}) \ . \tag{9}$$

This can be expensive to calculate directly when N is large because there are $O(N^2)$ terms. The same empirical auto-covariance function can be calculated using the FFT in $O(N \log(N))$ work. We approximate the infinite sum (2) by a finite sum

$$\widehat{S}_M = \widehat{C}_0 + 2 \sum_{t=1}^{M} \widehat{C}_t \; . \tag{10}$$

Note that the sum on the right is over positive t but C_t is a symmetric function of t, so the terms with $t > 0$ are counted twice. Now you just pick M and you're done.

It seems natural to a lot of terms to get an accurate approximation of the infinite sum (2). Don't do this because it is an *inconsistent* estimator. Suppose Q is some quantity we want to estimate and $\widehat{Q}_N$ is an estimator (a family of estimators depending on N). The estimator $\widehat{Q}_N$ is *weakly consistent* if $\widehat{Q}_N \to Q$ as $N \to \infty$ *in probability*. Convergence in probability means that for every $\epsilon > 0$, the probability of being wrong by ϵ goes to zero

$$\Pr\Big(\Big| \widehat{Q}_N - Q \Big| \geq \epsilon \Big) \to 0 \; , \quad \text{as } N \to \infty \; .$$

An estimator is *strongly consistent* if

$$\widehat{Q}_N \to Q \; , \quad \text{as } N \to \infty \text{ almost surely} \; .$$

I won't explain or use almost sure convergence. There is a mean-and-variance criterion for weak consistency that is simple to understand and use. The *bias* of an estimator is the error in its expected value

$$b_N = \mathrm{E}\Big[\widehat{Q}_N \Big] - Q \; . \tag{11}$$

You could put Q inside the expectation because Q is not random. An estimator is *unbiased* if $b_n = 0$. That's not common in MCMC. An estimator is *asymptotically unbiased* if

$$b_N \to 0 \; , \quad \text{as } N \to \infty \; . \tag{12}$$

An estimator that is not asymptotically unbiased can be consistent, but only (as far as I know) in very contrived examples. The variance of the estimator is

$$v_N = \mathrm{var}\Big(\widehat{Q}_N \Big) \tag{13}$$

It is an exercise (literally, exercise 1) to show that if the variance goes to zero and if it is asymptotically unbiased, then it is weakly consistent. I will say that an estimator is inconsistent in the mean-variance sense if it fails to satisfy one of these conditions. Strictly speaking, an estimator could be weakly consistent without being consistent in the mean-variance sense, but this never happens in practice, as far as I know.

The bias in our estimator $\widehat{S}$ comes from the terms C_t left out of the sum because $M < \infty$. This goes to zero asymptotically if $M \to \infty$ as $N \to \infty$. The estimator we ultimately propose does not have this property, so it is slightly biased. The variance of $\widehat{S}_N$ is sort of proportional to M/N. This means that if you take, say, $M = \frac{1}{2}N$, then the variance does not go to zero as $N \to \infty$. Roughly speaking, the terms $\widehat{C}_t$ for large t have mean more or less equal to zero, but variance on the order of $\frac{1}{N}$. If you add $O(N)$ of them together, and if they were independent, the variance would be $O(1)$. The fact that $\widehat{C}_t$ is not independent of $\widehat{C}_{t+1}$ should make this worse. If you want more convincing, exercise 2 gives an example where you can calculate everything, if you have the patience and interest.

I have been using a *self consistent window* estimator proposed by Alan Sokal. This takes M to be a multiple of the estimated auto-correlation time. The multiplier is the *window size*, w

$$\widehat{S} = \widehat{C}_0 + 2 \sum_{t=1}^{w\widehat{\tau}} \widehat{C}_t \tag{14}$$

$$\widehat{\tau} = 1 + \frac{2}{\widehat{C}_0} \sum_{t=1}^{w\widehat{\tau}} \widehat{C}_t \; . \tag{15}$$

The estimator is *self consistent* because it must be consistent with itself. The estimate of τ used to cut off the sum is the same as the estimate of τ produced by the cut-off sum. I typically take $w = 5$ or $w = 10$. A larger w gives less bias but more variance. Here is a code sketch showing how keep adding to the sum until the estimate of τ becomes self consistent:

```
th = 1     # the t=0 contribution to tau hat
t  = 0
while (1):
    if ( t < w*th):   # end of the window?
        break         # the self consistent window
    t  = t + 1
    th = th + 2*C[t]/C[0]    # the next term in the sum
    if ( t == N):            # explained below
        complain             # the run was too short
```

There is a problem called *spectral density estimation* that is related to auto-correlation time. Te sum S in (2) is the spectral density at frequency zero (whatever that means). Spectral density estimation also has the issue that too much data (take too many terms in the sum), leads to large variance. They propose a similar solution, which they call a window estimator, or *smoothing* (for reasons I don't want to explain). This is like cutting the sum (10) at M and asking the "user" (the person doing the density estimation) to pick M. Some codes have a specific choice, such as $M = 100$ built in. I think this is dangerous for MCMC because you do not know in advance what order of magnitude τ will

have. If $\tau = 10$, then $M = 100$ is probably fine. If $\tau = 1000$, then $M = 100$ gives a serious under-estimate of τ and a serious under-estimate of σ_N^2.

In my opinion, error bar estimation still is a major open problem for MCMC practice. I don't know a better estimator than the self consistent window (15), but I know a few worse ones that are used in practice. One common way is to run L independent MCMC chains and getting that many independent samples $\widehat{B}_N^j$, for $j = 1, \cdots, L$. Because these are independent, you can use the sample variance of the separate estimators $\widehat{B}_N^j$

$$\sigma_N^2 = \mathrm{var}\widehat{B}_N) \approx \frac{1}{L} \sum_{j=1}^{L} \left(\widehat{B}_N^j - \overline{\widehat{B}_N} \right)^2$$

The problem with this is that it doesn't protect you from having a run that is too short: $N \sim \tau$. The total number of MCMC steps is NL. It would have been more accurate to take a single run of length NL and use the self consistent estimator. It makes sense to take random and independent starting configurations

$$X_0^j \sim \rho_0 \ .$$

But this is not much help if ρ_0 is not typical of the target distribution ρ. We saw in the one sided pinned walk that is it hard to pick a good starting point that is typical of the eventual distribution.

2 Finite state space and eigenvalues

As has been mentioned before, you can learn what to expect using the case of a finite state space, and eigenvalues and eigenvectors of the transition matrix, R. Here is a formula for the steady state auto-covariance function in terms of eigenvalues and eigenvectors. Recall that

$$R_{ij} = \Pr(X_{k+1} = j \mid X_k = i) = \Pr(i \to j \text{ in one hop}) \ .$$

Powers of the transition matrix give transition probabilities after more hops. Let R_{ij}^t be entry (i, j) of R^t. Then

$$R_{ij}^t = \Pr(X_{k+t} = j \mid X_k = i) = \Pr(i \to j \text{ in } t \text{ hops}) \ .$$

We let π_{ij} be the joint distribution of X_k and X_{k+t} in the steady state. Without loss of generality we take $k = 0$. The joint distribution is

$$\pi_{ij} = \Pr_\rho(X_t = j \text{ and } X_0 = i) \ .$$

This is found using Bayes' rule of conditional probability:

$$\Pr_\rho(X_t = j \text{ and } X_0 = i) = \Pr(X_t = j \mid X_0 = i) \Pr_\rho(X_0 = i)$$
$$\pi_{ij} = R_{ij}^t \rho_i \ .$$

Without loss of generality, we may assume $\mathrm{E}_\rho[V(X)] = 0$. We write V_i for $V(i)$ Then

$$\begin{aligned}
\mathrm{cov}_\rho(V(X_0), V(X_t)) &= \mathrm{E}[\, V(X_0)V(X_t)] \\
&= \sum_{i=1}^n \sum_{j=1}^n \pi_{ij} V_i V_j \\
C_t &= \sum_{i=1}^n \sum_{j=1}^n R^t_{ij} \rho_i V_i V_j \qquad (16) \\
C_t &= \langle V, R^t V \rangle_\rho \qquad (17)
\end{aligned}$$

In the last line $\langle \cdot, \cdot \rangle_\rho$ is the $\rho-$inner product, which is defined by

$$\langle V, W \rangle_\rho = \sum_{i=1}^n V_i W_i \rho_i \ .$$

We think of V as a column vector with components V_i. Then $W = R^t V$ is the vector with components

$$W_i = \sum_{j=1}^n R^t_{ij} V_j \ .$$

Let r_j be the right eigenvectors of R. By convention

$$r_1 = \mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

This satisfies $Rr_1 = r_1$, so $\lambda_1 = 1$, because R is a stochastic matrix. The other eigenvectors and eigenvalues satisfy

$$Rr_j = \lambda_j r_j \ .$$

If the chain is non-degenerate (as we will show!), $|\lambda_j| < 1$ if $j \neq 1$. We suppose that there are n linearly independent eigenvectors and n corresponding eigenvalues, which do not have to be real if R does not satisfy detailed balance. The observable V may be expressed in terms of right eigenvalues and weights a_j

$$V = \sum_{j=2}^n a_j r_j \ .$$

The sum leaves out the term $j = 1$ because $\mathrm{E}_\rho[V] = 0$. This is justified in exercise 3 Therefore

$$R^t V = \sum_{j=2}^n \lambda_j^t a_j r_j \ . \qquad (18)$$

We put this into (17) and get the desired expression

$$C_t = \sum_{j \neq 1} w_j \lambda_j^t \ , \quad w_j = \langle V, a_j r_j \rangle_\rho \ . \tag{19}$$

The formula for τ follows from this, but you have to sum a geometric series. The terms with $t < 0$ involve $|t|$ because $C_{-t} = C_t$.

$$\begin{aligned} \sum_{-\infty}^{\infty} \lambda^{|t|} &= 1 + 2 \sum_{1}^{\infty} \lambda^t \\ &= 2 \left\{ \sum_{0}^{\infty} \lambda^t \right\} - 1 \\ &= \frac{2}{1-\lambda} - 1 \\ &= \frac{1+\lambda}{1-\lambda} \ . \end{aligned}$$

Therefore

$$S = \sum_{j \neq 1} w_j \frac{1+\lambda_j}{1-\lambda_j} \ .$$

The static variance is

$$C_0 = \mathrm{E}_\rho[V^2] = \langle V, V \rangle = \sum_{j \neq 1} w_j \ .$$

Combining these gives

$$\tau = \frac{\sum_{j \neq 1} w_j \frac{1+\lambda_j}{1-\lambda_j}}{\sum_{j \neq 1} w_j} \ . \tag{20}$$

The weights w_j are the same in the numerator and denominator. This is natural in view of the fact that auto-correlation time does not depend on the size of V. The observable $2V$ has the same auto-correlation time. The weights in the numerator are modified by the factor $\frac{1+\lambda_j}{1-\lambda_j}$. This is large if λ_j is close to 1. If $\lambda = 1 - g$ and g is small, then we have about $\frac{2}{g}$, which we saw on the homework for week 4.

The first formula (19) tells us what the auto-covariance function might look like. It decays to zero as $t \to \infty$ at a rate determined by the spectral gap

$$g = 1 - \max_{j \neq 1} |\lambda_j| \ .$$

A typical problem has a range of eigenvalues, some not close to 1 and others closer. For small values of t, the λ_j^t decreases rapidly if λ_j is not close to 1. This can make the overall C_t decrease rapidly at first. The rate of decrease can slow once the "energy" from the "rapidly decaying modes" is gone. A graph of C_t typically has a slope that is steep for small t and becomes less steep as t increases, leaving only the slowly decaying modes.

This means you have to be careful about estimating τ. There can be a large contribution from an eigenvalue close to 1 even if it has a small weight. If you stop the sum auto-covariance (10) too soon (M too small), you can miss this small slowly decaying "tail" that can have a large "mass" (sum). For that reason, I prefer to take w rather large (say, $w = 10$) in the self-consistent window algorithm. This almost certainly leads to a less accurate estimate of τ. But we should not "put error bars on error bars". We are looking for a rough estimate of error size, so knowing τ precisely is not important. What is more important is getting a warning when the run length N is too short. For $w = 10$, you need $N \geq 20\tau$ or the error bar code above will complain that N is too small. That means $N_{\text{eff}} \geq 20$.

In many applications you do a sequence of MCMC runs for different cases, such as models with different parameters. Even an MCMC expert does not want to examine the auto-covariance function for each case. Also common is someone using an MCMC code who doesn't know about auto-correlation, although he/she may be aware in qualitative terms that it is an issue. An MCMC code should warn such a user that her/his N is too small. For these reasons, it is good practice to give loud warnings if the software "thinks" that the run is too short.

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

$$m_j \frac{d^2 x_j}{dt^2} = -\partial_{x_j} \phi(x) \; . \tag{21}$$

In vector form, $F = ma$ becomes

$$M \frac{d^2}{dt^2} x = -\nabla_x \phi(x) \; . \tag{22}$$

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 (21) 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

$$\frac{d}{dt} \, p = -\nabla_x \phi(x) \tag{23}$$

$$\frac{d}{dt} \, x = M^{-1} p \; . \tag{24}$$

The equations (22) are "second order" because they involve second derivatives. The equations (23) and (24) are an equivalent set of first order equations. The first one (23) says that the force pushes on coordinate x_j by changing its momentum. The second one (24) 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 (23) and (24) may be expressed in terms of derivatives of the *Hamiltonian function* (or just *Hamiltonian*

$$H(x,p) = \phi(x) + \frac{1}{2} p^t M^{-1} p \; . \tag{25}$$

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

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

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

$$\frac{d}{dt} p = -\nabla_x H(x,p) \tag{26}$$

$$\frac{d}{dt} x = \quad \nabla_p H(x,p) \; . \tag{27}$$

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

$$\frac{d}{dt} H(x(t), p(t)) = 0 \; . \tag{28}$$

The calculation behind this is just the chain rule, which uses the minus sign in (26). 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 (25). 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 (26) and (27) 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

$$\rho(x) = \frac{1}{Z} e^{-\phi(x)} , \tag{29}$$

because ρ is the marginal density of f:

$$\rho(x) = \int_{\mathbb{R}^d} f(x, p) \, dp . \tag{30}$$

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

$$f(x, p) = \frac{1}{Z} e^{-H(x,p)} . \tag{31}$$

[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 (30) because of the sum structure (25) 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.

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

The crux is

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

I might call the function $e^{-\frac{1}{2}p^t M p}$ "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 (25) 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.]

Background on statistical mechanics

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. Here we give some of the terminology and intuition. It obviously is not a substitute for a real course.

One part of "equilibrium statistical mechanics" is guessing what the steady state PDF of of the state y might be. These are "physical" arguments related to the dynamical equations (26) and (27) but not on the specific form of the solution. They depend on three properties

1. Conservation of energy, (28).
2. Conservation of volume in phase space.
3. Dynamical chaos that forgets everything else.

Let $A(t)$ be an $n \times n$ matrix. We use a dot to represent derivative with respect to t, so, for any quantity $Q(t)$

$$\dot{Q} = \frac{dQ}{dt} \ .$$

$$\frac{d}{dt}\det(A) = \text{tr}\left(A^{-1}\dot{A}\right) \tag{32}$$

4 Exercises

1. Show that if an estimator is asymptotically unbiased and if the variance goes to zero as $N \to \infty$, then it is weakly consistent. You may use *Chebychev's inequality*, which says that if X is any random variable, then

$$\Pr\Big(|X - \text{E}[X]| \geq z\sqrt{\text{var}(X)}\Big] \leq \frac{1}{z^2} \ .$$

The variable z is the number of standard deviations X is away from its mean. Every math class should have at least one proof of this kind. If you're new to this kind of thing, start with the case of an unbiased estimator. Then use the fact that if $a_n \to a$ and $b_n \to b$, then $a_n + b_n \to a + b$. If b_n is the bias, then asymptotically unbiased means $b = 0$.

Consider an estimator that is consistent in the mean-variance sense and consider the following silly modification of it. Choose a sequence $p_N \to 0$ and $W_N \to \infty$ and define

$$R_N = \begin{cases} \widehat{Q}_N + W_N & \text{with probability } p_N \\ \widehat{Q}_N & \text{with probability } 1 - p_N \ . \end{cases}$$

This is a model of what would be a normal estimator but has a small chance of being very wrong. Show that if $\widehat{Q}_N$ is weakly consistent, then R_N is weakly consistent. Show that if you choose p_N and W_N correctly, you get a weakly consistent estimator that is not consistent in the mean-variance sense, and may not be asymptotically unbiased.

2. This exercise is quite time consuming. Read it but please do not do it unless you are finished everything else and are bored. Consider a linear scalar auto-regressive process with $X_0 = 0$ and

$$X_{k+1} = aX_k + Z_k \ , \quad Z_k \sim \mathcal{N}(0,1) \text{ i.i.d.}$$

We use a part of *Wick's theorem* that says if (Y_1, Y_2, Y_3, Y_4) is a four component multivariate normal with mean zero, then

$$\mathrm{E}[\, Y_1 Y_2 Y_3 Y_4] = \mathrm{E}[\, Y_1 Y_2]\, \mathrm{E}[\, Y_3 Y_4] + \mathrm{E}[\, Y_1 Y_3]\, \mathrm{E}[\, Y_2 Y_4] + \mathrm{E}[\, Y_1 Y_4]\, \mathrm{E}[\, Y_2 Y_3] \ .$$

There must be some formula like this because everything about a Gaussian is determined by second moments, so the fourth moment on the left must be some function of the second moments on the right. The Y_j do not have to be distinct. The case $Y_1 = Y_2 = Y_3 = Y_4 = Y$ gives the familiar formula $\mathrm{E}[Y^4] = 3\mathrm{E}[Y^2]^2$. Find an approximate formula for $\operatorname{var}\widehat{S}$ when N is large. This will show that $\widehat{S}$ is inconsistent in the mean variance sense if $M = CN$ for any C. With more calculations (please please don't do them), you would see that the estimator is also inconsistent in the weak sense, as the phenomenon of exercise 1 does not happen.

3. Here are some linear algebra facts to verify. Always assume that there are n linearly independent eigenvectors. The right eigenvector matrix is the matrix with right eigenvectors as columns. It is called Q instead of R because R is the transition matrix.

$$Q = \begin{pmatrix} | & | & \cdots & | \\ r_1 & r_2 & \cdots & r_n \\ | & | & \cdots & | \end{pmatrix}$$

The inverse of Q is L. It satisfies $QL = LQ = I$. The rows of L are row vectors l_j, so

$$L = \begin{pmatrix} -- & l_1 & -- \\ -- & l_2 & -- \\ \vdots & \vdots & \vdots \\ -- & l_n & -- \end{pmatrix}$$

The eigenvalues are the diagonal entries of the diagonal eigenvalue matrix

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

Assume that R represents a non-degenerate Markov chain.

(a) Show that $RQ = Q\Lambda$ and $LR = R\Lambda$ and $R = Q\Lambda L$.

(b) Show that l_j is the left eigenvector with $l_j R = \lambda_j L_j$.

(c) Show that we may take r_1 to be the vector of all ones, and if we do then $l_1 = \rho =$ the steady state probability distribution.

(d) Show that if V is a column vector with $\mathrm{E}_\rho[V] = 0$, then the eigenvector representation of V does not involve r_1 as in (18).

(e) Show that if R satisfies detailed balance with respect to ρ, then $C_t \geq 0$ for t even.