PHY256 Notes on Importance sampling and the Metropolis-Hastings method

A. C. Quillen

March 29, 2021

Contents

1 Monte Carlo methods

Monte Carlo numerical methods are a broad class of numerical methods that involves using numerically generated distributions of random events. They are particularly useful in optimization problems, and simulating systems that depend on a probability distribution. For example, Monte Carlo methods can be used interpretation of measurements. One can estimate the likelihood of experimental results. Similarly noise in experimental studies can be modeled with Monte Carlo methods. Statistical physics postulates that ensembles are well described with statistical ensembles. Processes that involve diffusion, such as diffusion limited aggregation can be modeled with Monte Carlo methods. Radiative transfer both in optically thin and optically thick limits can be modeled with Monte Carlo methods.

2 Importance sampling

To integrate a function $f(\mathbf{x})$ over a volume V

$$I = \int_V f(\mathbf{x}) dV$$

with a Monte Carlo method we can randomly sample $\mathbf{x}$ over the domain. Suppose $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 ...$ are a randomly and uniformly generated set of N samples. We can estimate the volume by computing

$$I \approx \frac{V}{N} \sum_{i=1}^{N} f(\mathbf{x}_i)$$

where N is the number of samples. The larger the number of samples, the closer our estimate is to the actual value. This kind of estimate could be done in any dimension. An example would be to integrate the volume of a unit sphere by uniform sampling of x, y, z in the domain $x, y, z \in [-1, 1]$ that looks like a cube.

Consider an integral

$$I = \int_a^b f(x) dx. \tag{1}$$

We can generate random variables uniformly $x_i \in [a, b]$ and estimate the integral as

$$I \approx \frac{(b-a)}{N} \sum_{i=1}^{N} f(x_i) \tag{2}$$

where N is the number of samples each with position x_i. The sum $\frac{1}{N} \sum_{i=1}^{N} f(x_i)$ is an estimate for the average value of $f()$ in the interval and the length of the interval is $(b-a)$.

The function $f(x)$ could be large in some regions of x. This would give large variations in the computed values for I (see Figure 1). If you happen to chose an x_i where $f(x_i)$ is really large you would get an inaccurate estimate for I. If the function $f()$ has peaks then you would need a very large number of samples N to get get an accurate measurement for the integral I.

We are generating x_i uniformly in $[a, b]$ which means x_i is described by the probability distribution

$$h_u(x) = \frac{1}{b-a}.$$

Equation 2 can be written as

$$I \approx \frac{1}{N} \sum_i \frac{f(x_i)}{h_u(x_i)}. \tag{3}$$

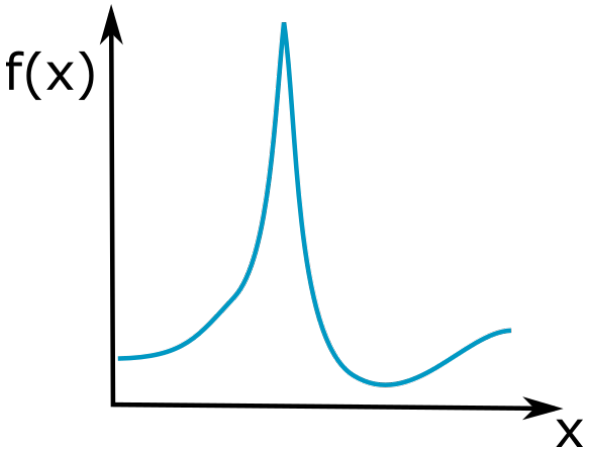

Figure 1: When estimating the integral of $f(x)$ from randomly chosen x values, the value you estimate for the integral depends on whether you happen to chose x values that lie near the peak.

Importance sampling is when the function is weighted so that we can achieve a better estimate with fewer randomly generated points. Let

$$x = g(y)$$

relating x to a variable y. The derivative

$$\frac{dx}{dy} = g'(y).$$

Our integral I can be written in terms of y

$$I = \int_a^b f(x)dx = \int_{g^{-1}(a)}^{g^{-1}(b)} f(g(y))\frac{dx}{dy}dy = \int_{g^{-1}(a)}^{g^{-1}(b)} f(g(y))g'(y)dy \tag{4}$$

We can now chose $y \in [g^{-1}(a), g^{-1}(b)]$ with a uniformly distribution and compute the integral I with (in analogy to equation 2)

$$I \approx \frac{g^{-1}(b) - g^{-1}(a)}{N} \sum_{i=1}^{N} f(g(y_i))g'(y_i). \tag{5}$$

Alternatively, rather than use a uniform probability distribution, we could chose x with any probability distribution $p(x)$. We use the probability distribution to weight the sum

$$I = \int_a^b f(x)dx \approx \frac{1}{N} \sum_{i=1}^{N} \frac{f(x_i)}{p(x_i)}$$

(compare this to equation 3). Here because $p(x)$ is a probability distribution $\int p(x)dx = 1$.

We can also use a function $w(x)$ that is not normalized. We call $w(x)$ a weighting function. We normalize a weighting function via defining

$$p(x) = \frac{w(x)}{\int w(x)dx}$$

and this gives an estimate for I

$$I \approx \frac{1}{N}\sum_{i=1}^{N} \frac{f(x_i)}{w(x_i)} \int w(x)dx. \tag{6}$$

Any weighting function $w(x)$ can be used weight generate randomly generated x values when computing and estimate for the integral I.

We can directly relate the probability function $p()$ with the weighting function $g()$ via comparing the two approaches (compare equation 4 to equation 6) giving

$$\frac{1}{p(x)} = \frac{dx}{dy} = g'(g^{-1}(x)).$$

2.1 The uncertainty or variance of the estimate

What is the uncertainty of our estimate for I? We want to compute the variance of our estimate for I.

Consider a series of values $x_1, x_2, ...$ with probabilities $p(x_i) = p_i$. The distribution is normalized; $\sum_i p(x_i) = 1$. The **expectation** value of x

$$E_p[x] = \sum_i x_i p(x_i).$$

This is also the mean value of x. I am subscripting with p because the probability distribution function is $p()$. Likewise the expectation value of a function of x

$$E_p[f(x)] = \sum_i f(x_i)p(x_i).$$

If the probability distribution is continuous

$$E_p[x] = \int xp(x)\ dx$$

and this is the mean value of x. Similarly

$$E_p[f(x)] = \int f(x)p(x)\ dx$$

If we call $\mu = E_p[x]$ then the variance σ^2 is

$$E_p[(x-\mu)^2] = \int (x-\mu)^2 p(x)\ dx.$$

Now let's go back to the integral that we want to estimate

$$I = \int_a^b\ dx f(x) = \int_a^b\ dx \frac{f(x)}{p(x)} p(x) \tag{7}$$

where $p(x)$ is a probability distribution. We can write this as

$$I = E_p\left[\frac{f(x)}{p(x)}\right]$$

This lets us estimate a variance for I as

$$E_p\left[\left(\frac{f(x)}{p(x)} - E_p\left[\frac{f(x)}{p(x)}\right]\right)^2\right] = E_p\left[\left(\frac{f(x)}{p(x)}\right)^2\right] - \left(E_p\left[\frac{f(x)}{p(x)}\right]\right)^2$$

How do we make a numerical estimate for I? Following equation 7 we generate x_i with probability distribution $p()$ and compute

$$I \approx \frac{1}{N}\sum_i \frac{f(x_i)}{p(x_i)}. \tag{8}$$

The variance in this estimate will depend on the expectation value (using probability distribution p) of $(f(x)/p(x))^2$.

We can minimize the variance by choosing $p(x)$ to have thicker tails than $f(x)$. One would chose the probability function $p(x)$ to depress $f's$ peaks and be larger than f when f is small.

2.2 Importance sampling in statistical physics

In statistical physics one often wants to calculate an expectation value of a quantity for a system that is in thermal equilibrium at temperature T. States are weighted by a probability that depends on

$$e^{-\beta E_i}$$

(also known as the Boltzmann factor) with the state energy E_i and inverse temperature

$$\beta \equiv \frac{1}{kT}.$$

A normalized probability function for each energy state

$$P_B(E_i) = \frac{e^{-\beta E_i}}{Z} \tag{9}$$

where

$$Z \equiv \sum_i e^{-\beta E_i}$$

is known as the partition function. The expectation value of a physical quantity X that depends on the state is

$$\langle X \rangle = \sum_i P_B(E_i) X_i$$

A difficulty with computing this with a direct random sampling of energy states (using a uniform distribution in energy) is that there may be many states that are not very likely, or have exponentially small probability. We don't want to spend a lot of time computing the contribution of extremely unlikely energy states.

We could weight by $P_B(E_i)$, however this involves knowing the partition function. A work around is to generate a series of energy values with distribution consistent with a Boltzmann distribution without knowing or computing the partition function. This motivates our next topic, Markov chains.

3 Markov chains

A Markov chain is a sequence of randomly generated numbers $x_0 \to x_1 \to x_2 \to ...$ The numbers belong to a subset of possible *states*. Each number is generated from the previous one using a probability that only depends on the value you have at the previous step. We can define the Markov chain in terms of transition probabilities between possible states.

$$p_{ij} = P(x_{n+1} = j | x_n = i)$$

Here x_{n+1} is the $n+1$-th randomly generated variable and x_n is the n-th generated variable. The indices are the possible values. The probability p_{ij} is the probability that you would get j if you had i previously.

3.1 Markov chain model for a random walk

For example consider a random walk on a one-dimensional grid where a walker either goes to the right a step or to the left a step and each possibility has probability 1/2. The possible states are the grid points and so are described by integers. The transition probabilities from position 2 are $p_{23} = 1/2$ (to the right) , $p_{21} = 1/2$ (to the left), and $p_{22} = 0$ (not moving). The transition probabilities form a 2 dimensional square matrix

$$\mathbf{P} = \begin{pmatrix} .. & .. & .. & .. & .. & .. & .. & .. \\ .. & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & .. \\ .. & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & .. \\ .. & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & .. \\ .. & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & .. \\ .. & .. & .. & .. & .. & .. & .. & .. \end{pmatrix}$$

The probabilities in each row must sum to 1. If we have two endpoints where the walkers stay if they reach them, then

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & ... & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & ... & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & ... & 0 & 0 & 0 & 0 \\ .. & .. & .. & .. & ... & .. & .. & .. & .. \\ 0 & 0 & 0 & 0 & ... & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & ... & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & ... & 0 & 0 & 0 & 1 \end{pmatrix}$$

Suppose we start with a walker at a single position. This means the walker has a probability 1 of being at a particular location. Its initial distribution

$$\boldsymbol{\pi}_0 = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & ... \end{pmatrix}$$

if the particle is located at the 4th position initial. This vector is a list of the initial probabilities that the walkers are at each location. The sum of the values array should be 1, so $\sum_i \pi_i = 1$.

After one step of the random walk, the new distribution of probabilities (to be in each location) becomes a new vector of probabilities $\boldsymbol{\pi}_1$ with

$$\boldsymbol{\pi}_1 = \boldsymbol{\pi}_0 \mathbf{P}.$$

We multiply the horizontal vector $\boldsymbol{\pi}_0$ and the matrix $\mathbf{P}$.

$$\begin{pmatrix} . & . & . \end{pmatrix} \begin{pmatrix} . & . & . \\ . & . & . \\ . & . & . \end{pmatrix}$$

The probabilities at each position

$$\pi_{1,j} = \sum_i \pi_{0,i} P_{ij}.$$

The sum over i arises because you need to taken into account all possible previous values i for the random variable and the probability that the new state would transition from i to j.

It's sometimes nice to write the new distribution in terms of the transpose of the transition matrix $\mathbf{P}^t$,

$$\boldsymbol{\pi}_1 = \mathbf{P}^t \boldsymbol{\pi}_0$$
$$\pi_{1,j} = \sum_i P^t_{ji} \pi_{0,i} = \sum_i \pi_{0,i} P_{ij}$$

Now we can think of $\boldsymbol{\pi}_1$ as a square transition probability matrix times a vertical probability distribution vector $\begin{pmatrix} . & . & . \\ . & . & . \\ . & . & . \end{pmatrix} \begin{pmatrix} . \\ . \\ . \end{pmatrix}$.

3.2 Detailed balance for MCMC models and stationary distributions

A **Markov chain Monte Carlo** or **MCMC** model is one where you simulate by generating a sequence of randomly generated variables. At each step you generate new variables based on the previous ones. Often your goal is to measure the distribution of the variables generated. In principle these could be calculated from the transition matrix and the initial probability distribution.

A distribution $\boldsymbol{\pi}$ of a Markov chain is **stationary** if

$$\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}.$$

That means the probability distribution is unchanged after you chose the next set of variables.

A distribution $\boldsymbol{\pi}$ satisfies **detailed balance** for a Markov chain if

$$\pi_i P_{ij} = P_{ji}\pi_j. \tag{10}$$

A distribution that satisfies detailed balance is a stationary distribution. Suppose we start with $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_0\mathbf{P}$. We compute the i-th element of $\boldsymbol{\pi}_1$

$$\pi_{1,j} = \sum_i \pi_{0,i} P_{ij}$$

With equation 10

$$\pi_{1,j} = \sum_i P_{ji}\pi_{0,j} = \pi_{0,j}\sum_i P_{ji}$$

However $\sum_i P_{ji} = 1$ because each row must sum to 1. Consequently $\pi_{1,j} = \pi_{0,j}$ and $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_0$. We have shown that a distribution that satisfies **detailed balance** (equation 10) is a **stationary** distribution.

A stationary distribution of a Markov chain is not necessarily an attracting or stable distribution.

The distribution of generated variables can become independent of the initial conditions and approach a *stable* or *limiting* stationary distribution $\boldsymbol{\pi}_l$. In this case

$$\mathbf{P}^n \to \begin{pmatrix} \boldsymbol{\pi}_l \\ \boldsymbol{\pi}_l \\ .. \\ \boldsymbol{\pi}_l \end{pmatrix}$$

where each row of the matrix is the limiting distribution $\boldsymbol{\pi}_l$. In this case $\lim_{n\to\infty} P^n_{ij} = \pi_{l,j}$ and is independent of i.

Consider an initial condition $x_i = i$. If the probability is non-zero that at some later iteration step that the random variable could be $x_j = j$ we say that i and j states **communicate**. There is some possible path between the states. If all states communicate with all other states the Markov chains is **irreducible**.

A state i is *recurrent* if the probability is 1 that at some later iteration step the state will again be i. Otherwise state i is *transient*.

An chain is *ergodic* if it is recurrent and aperiodic. (There is also a condition on the recurrence times).

An irreducible, ergodic Markov chain model has a unique limiting, attracting and stationary probability distribution. (This is a theorem!)

4 The Markov Chain methods for statistical physics

Start with a system in a particular energy state E_i. Each step involves a random choice to move the system into another energy state or not. We use a transition probability T_{ij} to move the system from state E_i to E_j. With indices reversed, T_{ji} moves the system from E_j back to E_i. As T_{ij} is a probability we require that $\sum_j T_{ij} = 1$.

We **choose** transition probabilities that satisfy

$$\frac{T_{ij}}{T_{ji}} = \frac{P_B(E_j)}{P_B(E_i)} = \frac{e^{-\beta E_j}/Z}{e^{-\beta E_i}/Z} = e^{-\beta(E_j - E_i)} \tag{11}$$

where P_B is defined in equation 9 via the Boltzmann factor. Notice the similarity between this condition and equation 10. We are choosing transition probabilities for a Markov chain model that will give us a stationary distribution consistent with Boltzmann probabilities.

We start with E_i. What is the distribution of new energy states E_j after we make transitions using the transition probability T_{ij}? The distribution of E_j is given by T_{ij}.

If E_i has a distribution $p(E_i)$ what is the probability distribution of E_j after using the transition probabilities? It is

$$p(E_j) = \sum_i T_{ij} p(E_i).$$

What if the probability that the state is in E_i is equal to the Boltzmann distribution $p(E_i) = P_B(E_i)$? Then using equation 11

$$\begin{aligned} p(E_j) &= \sum_i T_{ij} p(E_i) \\ &= \sum_i T_{ij} P_B(E_i) \\ &= \sum_i T_{ji} P_B(E_j) \\ &= P_B(E_j) \sum_i T_{ji} \\ &= P_B(E_j). \end{aligned}$$

If the probability of being in state i satisfies the Boltzmann distribution then so will the probability of being in state j. We have just shown that the Boltzmann distribution is a **stationary** distribution of the iterative MCMC method as long as the transition probabilities satisfy equation 11.

Now we need a way to generate a transition probabilities T_{ij} for the Markov chain in such a way as to satisfy equation 11.

We would like that choice, when iterated, would give us a system that not only has the Boltzmann distribution as a limiting distribution but also converges onto set of states consistent with a Boltzmann distribution (and so in thermal equilibrium) even if our initial state is highly unlikely at our assumed temperature.

4.1 Metropolis-Hastings method

The Metropolis algorithm applies to a system that has a variety of energy states. The goal is to find states that are likely if the system is in thermal equilibrium. To apply the algorithm you need a way to compute differences in energy of the system as it changes state. You need to know the temperature T.

Start the system in some energy state.

- Randomly choose some part of the system to vary and some way to vary it. This would give a new state with a new energy. Compute the difference between this energy and the energy of the original state, ΔE.
- Accept the new state with a probability of acceptance P_a.

Repeat numerous times.

After running the algorithm for a while the distribution of states reached should be given by the Boltzmann distribution.

Let's describe the acceptance probability. Starting in state E_i the Metropolis method accepts or rejects a change of state to E_j based on an acceptance probability

$$P_a = \begin{cases} 1 & \text{if} \quad E_j \leq E_i \\ e^{-\beta(E_j - E_i)} & \text{if} \quad E_j > E_i \end{cases} \tag{12}$$

Accept the change if the new state lowers the energy, otherwise only accept it based on the Boltzmann distribution of the energy change.

We will now show that this acceptance probability is consistent with the requirement for the transition probabilities in equation 11.

Consider transition probability T_{ij}. If $E_j \leq E_i$ we use the first case in equation 11. If the number of states with $E_j \leq E_i$ is M then

$$T_{ij} = \frac{1}{M}.$$

The transition in the opposite direction T_{ji} involves going from higher to lower energy so we use the second case in equation 12,

$$T_{ji} = \frac{1}{M} e^{-\beta(E_i - E_j)}.$$

(We flip i, j in the exponential because we are going in the opposite direction). The ratio is

$$\frac{T_{ij}}{T_{ij}} = e^{-\beta(E_j - E_i)}, \tag{13}$$

and this is consistent with our choice in equation 11.

Our recipe for the acceptance probability P_a (in equation 12) satisfies equation 11. And that means that the recipe has as stationary distribution of states that is consistent with a Boltzmann distribution. It does not mean necessarily that if you start out of equilibrium that you would necessarily converge onto statistical equilibrium. However, all states are reachable in principle so the Markov chain is irreducible. If it is also ergodic then the stationary distribution is also attracting and stable.

The Metropolis algorithm has been used to explore phase transitions in a variety of systems.

The Metropolis-Hastings algorithm works by generating a sequence of sample values, so that as more and more sample values are produced, the distribution of values more closely approximates the desired distribution which in the case described here is a Boltzmann distribution. These sample values are produced iteratively, with the probability distribution of the next sample being dependent only on the current sample value. The procedure for generating the sequence of samples is a Markov chain Monte Carlo or MCMC model.