Notes on the Metropolis et al. Algorithm

K.E. Schmidt
Department of Physics
Arizona State University
Tempe, AZ 85287

Abstract

The Metropolis et al. method is described and the conditions for convergence shown.

1 Metropolis et al.

The Metropolis et al. algorithm was invented by Metropolis et. al[1] and is often simply called the Metropolis algorithm. It is a method to sample a normalized probability distribution by a stochastic process. We define $\mathcal{P}_i^{(n)}$ to be the probability for finding the system in the state i at step n. The algorithm is then

- Sample a possible new state j with some probability $T_{i\to j}$.
- Accept the new state j with probability $A_{i\to j}$ and use it as the next sample. With probability $1 - A_{i\to j}$ the move is rejected and the original state i is used again as a sample.

We wish to derive the required properties of T and A such that $\mathcal{P}_i^{(n\to\infty)} \to p_i$ so that starting from any distribution, the method converges to the correct distribution. Note that the description here is for a discrete probability distribution. Replacing probabilities p_i with expressions like $p(x_i)dx_i$ will take all of these over to the corresponding continuum expressions.

The dynamical equation for $\mathcal{P}_i^{(n)}$ can be written directly from the description above. The probability of being in the state i at step n is given by the probability of being in any state j at the previous step, and making an accepted transition to i added to the probability of being in the state i, making a transition to any state j and rejecting the move:

$$\mathcal{P}_i^{(n)} = \sum_j \left[\mathcal{P}_j^{(n-1)} T_{j\to i} A_{j\to i} + \mathcal{P}_i^{(n-1)} T_{i\to j} \left(1 - A_{i\to j}\right)\right] . \tag{1}$$

Since the probability of making some transition must be 1, $\sum_j T_{i\to j} = 1$, and Eq. 1 becomes

$$\mathcal{P}_i^{(n)} = \mathcal{P}_i^{(n-1)} + \sum_j \left[\mathcal{P}_j^{(n-1)} T_{j\to i} A_{j\to i} - \mathcal{P}_i^{(n-1)} T_{i\to j} A_{i\to j}\right] . \tag{2}$$

For large n we require that $\mathcal{P}_i^{(n\to\infty)} = p_i$, the desired probability distribution. Taking this limit, gives the balance requirement

$$\sum_j \left[p_j T_{j\to i} A_{j\to i} - p_i T_{i\to j} A_{i\to j}\right] = 0 . \tag{3}$$

The balance requirement is very weak. Typically the much stronger detailed balance requirement is enforced, that is rather than the sum being set to zero, we set each term separately to zero and use this to determine the acceptance probabilities. Rearranging, the result is

$$\frac{A_{j\to i}}{A_{i\to j}} = \frac{p_i T_{i\to j}}{p_j T_{j\to i}} . \tag{4}$$

The Metropolis choice is to maximize the A values, that is

$$A_{j\to i} = \min\left(1, \frac{p_i T_{i\to j}}{p_j T_{j\to i}}\right). \quad (5)$$

Other choices are possible, but they all correspond to multilplying $A_{i\to j}$ and $A_{j\to i}$ by the same constant smaller than unity.[1]

Having chosen the acceptance probabilities, we have guaranteed that if the $\mathcal{P}_i^{(n)}$ has equilibrated, that is if it is equal to p_i, it will remain equilibrated. Next we need to find the circumstances for convergence to equilibrium.

The dynamical equation can be written as

$$\mathcal{P}_i^{(n)} = \sum_j M_{ij} \mathcal{P}_j^{(n-1)} \quad (6)$$

with the matrix M given by

$$M_{ij} = \delta_{ij}\left[1 - \sum_k T_{i\to k} A_{i\to k}\right] + T_{j\to i} A_{j\to i}\,. \quad (7)$$

Summing over i shows that $\sum_i M_{ij} = 1$, and since $\sum_k T_{i\to k} = 1$, and $A_{i\to k} \le 1$, the elements of the matrix satisfy $M_{ij} \ge 0$. The matrix M is therefore a stochastic matrix.

The Metropolis method is simply the power method for computing the right eigenvector of M with the largest magnitude eigenvalue. By construction, the correct probability distribution is a right eigenvector with eigenvalue 1. Therefore, for the Metropolis method to converge to this result, we must show that M has only one eigenvalue with this magnitude, and all other eigenvalues are smaller.

Even a defective matrix has at least one left and right eigenvector for each eigenvalue.[2] The Gershgorin bounds for the eigenvalues can be derived by multiplying on the left with the eigenvector with the maximum and minimum eigenvalues,

$$\begin{aligned} \sum_i \psi_i^{\max} M_{ij} &= \lambda_{\max} \psi_j^{\max} \\ \sum_i \psi_i^{\min} M_{ij} &= \lambda_{\min} \psi_j^{\min} \end{aligned} \quad (8)$$

Next we choose the normalization of these eigenvectors so that the largest element (or one of the equally largest elements) has value 1. Let's call this element k, and we can therefore bound the magnitude of the other elements to be less than or equal to 1. This leads to the inequalities, using the property that $M_{ij} \ge 0$,

$$\begin{aligned} \sum_i M_{ik} &\le \lambda_{\max} \\ M_{kk} - \sum_{i\ne k} M_{ik} &\ge \lambda_{\min} \end{aligned} \quad (9)$$

where the equality from the maximum will occur only if the eigenvector takes the value 1 for all values of i where $M_{ik} \ne 0$, and the equality for the minimum will occur only if the eigenvector takes the value -1 for all values of $i \ne k$ where $M_{ik} \ne 0$.

That the maximum eigenvalue is 1 follows immediately from the property that $\sum_i M_{ik} = 1$. Similarly the minimum eigenvalue can be -1, but only if $M_{kk} = 0$ and the magnitude of all the other elements $\psi_i^{\min}$ of the eigenvector that multiply nonzero elements M_{ik} are -1.

Let's first see what the properties of M must be to eliminate any -1 eigenvalues. To have a -1 eigenvalue, the left eigenvector must contain only ± 1 and 0 values. Taking in turn each ± 1 value as the maximum, so that it corresponds to the index k, the nonzero M_{ik} values must correspond to i index values of the eigenvector which have opposite sign elements. That is, the M matrix must break up into sets of states that always make transitions from set A to set B ... back to set A. In particular, there can be no rejections of these moves in the cycle since the -1 eigenvalue requires $M_{kk} = 0$. To guarantee no eigenvalues with eigenvalue -1, we

[1] The penalty function method uses just such a factor to compensate for p_i that are evaluated stochastically and are therefore noisy.

[2] An example of a defective matrix is $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, with two zero eigenvalues, only one right eigenvector $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, and only one left eigenvector (0 1).

simply have to make sure that there are no cycles among states. Notice that this is generally trivial since such cycles cannot have any rejections at any stage. An example of such a cycle is sampling a noninteracting Ising spin. If the transition is taken to flip the spin, and the energy difference is zero, the Boltzmann factor will not change and the move will always be accepted. The system will simply flip from up to down to up to down ad infinitum. Including a rejection probability or using a heat bath algorithm immediately fixes the problem.

Next we need to make sure that there is only one left eigenvector with eigenvalue 1. To get an eigenvalue 1, the left eigenvector must be constructed from only ones and zeroes. It is straightforward to see that a vector made up of ones and zeroes can only be an eigenvector with eigenvalue 1 if the matrix element $M_{ij} = 0$ for all cases where $\psi_i \neq \psi_j$. That is we can choose an index i and take $\psi_i = 1$. We require all elements ψ_j where $M_{ij} \neq 0$ to also have the value 1. Continuing we then require all elements ψ_ℓ $M_{j\ell}$ to have value 1. Only if the matrix M can be put into block diagonal form can there be more than one choice for the left eigenvector with eigenvalue 1. We therefore require that the transition matrix not be in block diagonal form. This simply means that we must choose the transition probability so that we can get from any allowed state to any other in a series of transitions.

Finally, we note that for a defective matrix, with more eigenvalues than independent eigenvectors for eigenvalue 1, the left and right eigenvectors of eigenvalue 1 would be orthogonal. Here the left eigenvector is all 1 except for states that can never be reached, and the right eigenvector is $p_i > 0$ except for states that give zero probability. We already require that we can reach all states that contribute to p_i. Therefore the left and right eigenvectors with eigenvalue 1 do not correspond to a defective sector of the matrix and they are unique. The Metropolis algorithm therefore converges exponentially to the desired distribution.

The requirements for the transition $T_{i\to j}$ are

(a) A series of transitions must let us to get from any allowed state to any other by a finite series of transitions.

(b) The transitions cannot be grouped into sets of states, A, B, C ,... such that transitions from A go to B, B to C etc and finally back to A. With condition (a) satisfied, this condition will always be satisfied if either $T_{i\to i} \neq 0$ or there are some rejected moves.

References

[1] N. Metropolis, A.W. Rosenbluth, H. Rosenbluth, A. Teller, E. Teller, *Equation of state calculation by fast computing machines*, J. Chem. Phys. **21**, 1087 (1953).