The Metropolis Sampling Algorithm

Lecture #11: Tuesday, 4 November 1997
Lecturer: Eric Veach
Scribe: Homan Igehy

1 Applications of Monte Carlo Methods

Monte Carlo methods were developed in order to solve difficult integration problems of high dimensionality. The techniques created came out of practical needs rather than theoretical exercises. Before going into the details of the Metropolis Sampling Algorithm, we shall first look at a few typical problems that require the use of Monte Carlo techniques. Monte Carlo is definitely not a solution in wait of a problem.

1.1 Computer Graphics

A large part of computer graphics deals with rendering and the transport of light. The light transport problem is a high dimensional problem: light leaving an emitter makes its way through several bounces on several surfaces (each with its own surface reflectance properties) before hitting the image plane. To make matters more complicated, discontinuities show up in a variety of ways. Within a BRDF, a singularity is present in the case of mirror reflections. A discontinuity in the surface normal of an object causes a discontinuous change in the amount of light reaching the object (e.g. along the edge of a cube). Even if all the BRDF's and surface normals are smooth, object boundaries form discontinuities. On the image plane, these discontinuities manifest themselves as silhouette edges. On a surface, these discontinuities form shadow boundaries. In the case of point light sources, the shadows are discontinuous in value, whereas in the case of area light sources, the shadows are discontinuous in their first or second derivatives. The presence of discontinuities in such a high dimensional problem accentuates the need for Monte Carlo techniques. Let us take a look at the dimensions involved in typical rendering methods.

1.1.1 Distribution Ray Tracing

Distribution ray tracing is similar to standard ray tracing in the sense that rays are traced from the eye towards the light source. The difference is that at each step of the light transport problem, an integral is evaluated. As illustrated in Figure 1, distribution ray tracing can be thought of as a simultaneous integration over:

Variable	Dimensions
time	1
pixel area	2
lens aperture	2
area light source	2
surface reflections	2 each

Thus, in distribution ray tracing we must evaluate a $(7 + 2r)$-dimensional integral, where r is the number of surface reflections considered.

1.1.2 Path Tracing

Path tracing is a generalization of distribution ray tracing in that it includes light transport paths of all lengths. A path can be started in any direction and can be represented as a sequence of points $x_0, x_1, \ldots, x_k$ on the scene surfaces. Each x_i can be either be chosen directly on a surface (e.g. sampling a point on a light source), or by choosing a random direction and casting a ray. In either case, each x_i has two degrees of freedom. If we include an integration over time as well, we get a problem of dimensionality $2(k + 1) + 1$ for a path of length k. In path tracing, k is not bounded, and we need to integrate over an infinite-dimensional space.

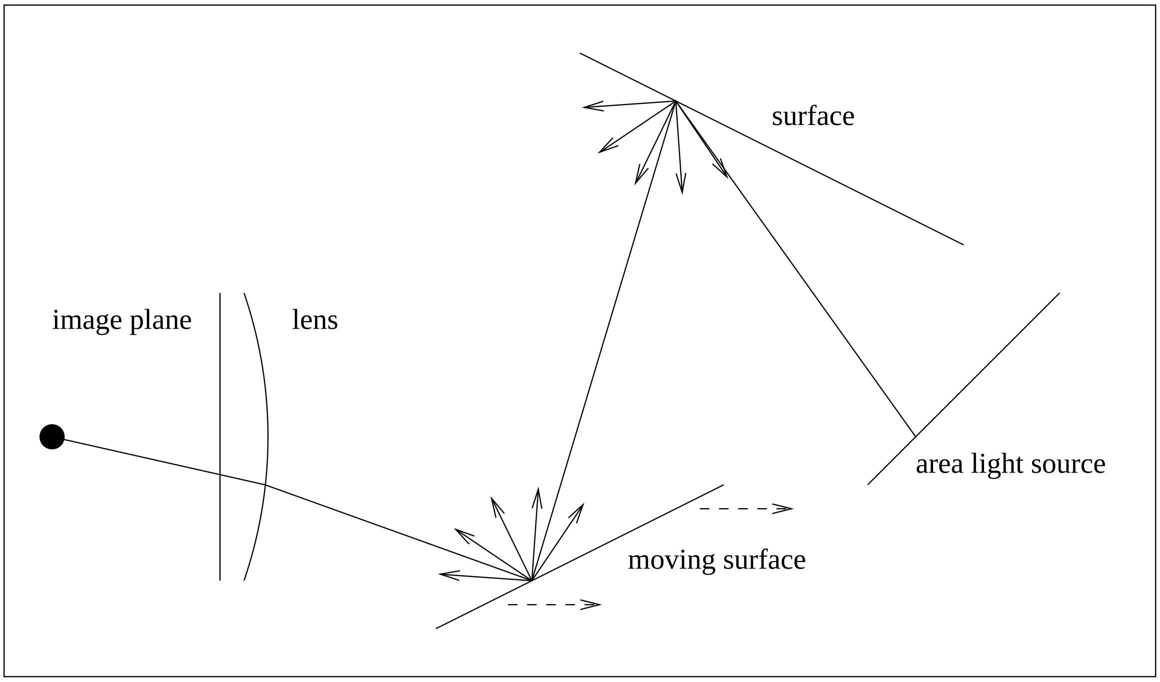

Figure 1: Distribution Ray Tracing.

1.1.3 Radiosity

The form factor computation of radiosity seeks to derive the fraction of light leaving patch A that falls onto patch B, assuming a Lambertian reflectance model. Although an analytic expression has been derived for two arbitrary polygonal patches, this model is slow to evaluate and does not take into account effects of visibility. Thus, this discontinuous 4-dimensional integration problem is amenable to Monte Carlo techniques. Similarly, finite element-based techniques for radiance (rather than radiosity) need to evaluate coefficients that represent the fraction of light that leaves patch A, falls onto patch B, and then gets reflected onto patch C. This is a 6-dimensional integral.

1.2 Finance

Monte Carlo techniques are widely used in the financial world for valuing securities. One class of securities traded is mortgage-backed securities. The mortgages backing these securities have a fixed or variable interest rate and typically have a maturity period of 30 years. The value of a mortgage-backed security depends on a complicated mix of the prime interest rate, the default rate, the remortgaging rate, and a variety of other interdependent factors that vary over time. By modeling these factors over time, a basis for valuation is provided. Solutions to these models involve integrating functions of 20 or 30 or even 360 dimensions. Since these functions are typically smooth, quasi-Monte Carlo techniques work well.

1.3 Physics and Chemistry

The proliferation of Monte Carlo techniques was due to the invention of electronic computers and their application to problems of physics and chemistry. Monte Carlo techniques have been used to solve problems in neutron transport (i.e., bombs), radiation shielding (i.e., reactors), and particle transport (e.g., electron beam lithography).

Another area where Monte Carlo methods are used is to predict the properties of materials. This is the application that originally led to the Metropolis Sampling Algorithm. Problems in this field include the calculation of the density of a liquid at a given temperature, the magnetic properties of alloys, and the modeling phenomena such as superfluidity.

2 A Typical Physics Problem

Before getting into the Metropolis Sampling Algorithm, let us motivate it with a typical physics problem. Imagine a bunch of particles inside a wrap-around box which is at thermal equilibrium, as illustrated in Figure 2. The state of this system can be described by a vector

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$$

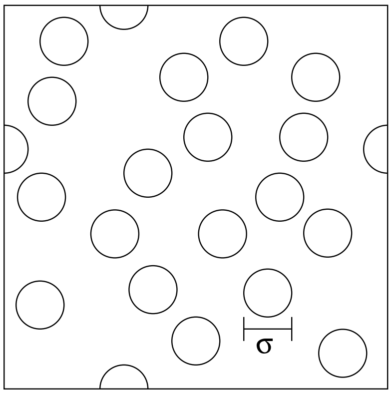

Figure 2: A Typical Physics Problem: Particles with extent σ.

where $\mathbf{x}_i$ is the position of particle i, and we let $\mathbf{\Omega}$ represent the space of all possible particle configurations $\mathbf{X}$. (The momenta of the particles are sometimes also required, but we will ignore them in this formulation.) If $\Phi(\mathbf{X})$ represents the energy of the system, then the probability $p(\mathbf{X})$ that the system will be in state $\mathbf{X}$ obeys the Boltzmann distribution:

$$\begin{aligned} p(\mathbf{X}) &= \frac{f(\mathbf{X})}{\int_{\mathbf{\Omega}} f(\mathbf{X})\, d\mathbf{X}} \\ \text{where} \quad f(\mathbf{X}) &= e^{-\Phi(\mathbf{X})/(kT)} \ . \end{aligned}$$

Here T is the temperature, and k is Boltzmann's constant. Note that $f(\mathbf{X})$ is just the unnormalized $p(\mathbf{X})$.

One model for the energy $\Phi(\mathbf{X})$ is to sum up all possible pair potentials:

$$\Phi(\mathbf{X}) = \sum_{1 \le i < j \le m} V(\mathbf{x}_i, \mathbf{x}_j)$$

The hard sphere potential model treats the particles as solid balls of diameter σ that cannot interpenetrate, and any configuration which satisfies this requirement is equally likely. This can be achieved by setting V as follows:

$$V(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 0 & r_{ij} > \sigma \\ \infty & \text{otherwise} \end{cases}$$

where r_{ij} is distance between particles i and j:

$$r_{ij} = |\mathbf{x}_i - \mathbf{x}_j|$$

Alternatively one could use the Lennard-Jones potential model (illustrated in Figure 3):

$$V(\mathbf{x}_i, \mathbf{x}_j) = 4\varepsilon \left\{ \left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6} \right\}$$

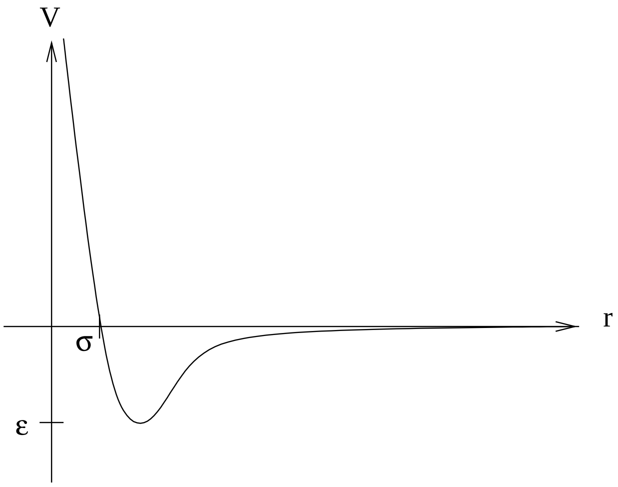

Figure 3: The Lennard-Jones Potential.

In any case, suppose there is some function $g(\mathbf{X})$ in which we are interested. We want to find the expected value of $g(\mathbf{X})$ where $\mathbf{X}$ is distributed proportionally to $f(\mathbf{X})$. In other words, the quantity of interest G is:

$$G = E[g(\mathbf{X})] = \int_{\mathbf{\Omega}} g(\mathbf{X})p(\mathbf{X})\, d\mathbf{X}$$

If we choose $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ according to the distribution specified by $p(\mathbf{X})$, an estimator for G is given by:

$$\hat{G} = \frac{1}{N}\sum_{t=1}^{N} g(\mathbf{X}_t)$$

Since $f(\mathbf{X})$ (and consequently $p(\mathbf{X})$) is a complicated function of a huge number of variables (in our case proportional to number of particles), finding an appropriate collection of samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ poses a formidable task. The Metropolis Sampling Algorithm solves precisely this problem.

3 The Metropolis Sampling Algorithm

The Metropolis Sampling Algorithm was developed in 1953 by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (it is also called the $\mathrm{M(RT)}^2$ algorithm for obvious reasons). The algorithm takes a random walk to generate a set of samples that are distributed according to some arbitrary given function. Specifically, suppose want have a domain $\mathbf{\Omega}$ such that $\mathbf{X} \in \mathbf{\Omega}$. Furthermore, let $f(\mathbf{X})$ be some some scalar function. The goal of the $\mathrm{M(RT)}^2$ is to provide a series of samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ which have a probability density function proportional to $f(\mathbf{X})$. In other words:

Given: A function $f(\mathbf{X})$ where $\mathbf{X} \in \mathbf{\Omega}$.
Goal: Generate $\mathbf{X}_t \sim p(\mathbf{X})$ where $p(\mathbf{X}) \propto f(\mathbf{X})$ and $\int_{\mathbf{\Omega}} p(\mathbf{X})\, d\mathbf{X} = 1$.

3.1 Random Walks

M(RT)^2 generates a set of samples by taking a *random walk* across $\mathbf{\Omega}$. A random walk is simply a sequence of samples $\mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_N$ where each $\mathbf{X}_t$ is chosen with some probability distribution $p_t(\mathbf{X})$. Furthermore, the random walk generated by M(RT)^2 is a *Markov chain*, meaning that the choice of $\mathbf{X}_t$ depends only on $\mathbf{X}_{t-1}$. If the random walk is independent of the actual value of t, it is called a *time-homogeneous Markov chain*, which can be characterized by a single *transition function* for all t:

$$K(\mathbf{Y} \to \mathbf{X}) = \{\text{density that } \mathbf{X}_t = \mathbf{X} \text{ given that } \mathbf{X}_{t-1} = \mathbf{Y}\}$$

In other words, $K(\mathbf{Y} \to \mathbf{X})$ is the probability density of going from state $\mathbf{Y}$ to state $\mathbf{X}$. Since it is a pdf, it satisfies:

$$\int_{\mathbf{\Omega}} K(\mathbf{Y} \to \mathbf{X})\, d\mathbf{X} = 1 \qquad \text{for all } \mathbf{Y}$$

Thus, for a random walk, we choose an initial state $\mathbf{X}_1$ according to some initial distribution $p_1(\mathbf{X})$. Then $p_2(\mathbf{X} \mid \mathbf{X}_1)$ is set to $K(\mathbf{X}_1 \to \mathbf{X})$, and $\mathbf{X}_2$ is chosen according to that distribution. By repeating this process, $p_t(\mathbf{X} \mid \mathbf{X}_{t-1})$ is set to $K(\mathbf{X}_{t-1} \to \mathbf{X})$ and $\mathbf{X}_t$ is chosen according to that distribution. Now since each $\mathbf{X}_t$ is a random variable, we get the the following distributions for each step of the random walk:

$$\begin{aligned}
p_1(\mathbf{X}) &= \text{initial distribution} \\
p_2(\mathbf{X}) &= \int_{\mathbf{\Omega}} p_1(\mathbf{Y}) K(\mathbf{Y} \to \mathbf{X})\, d\mathbf{Y} \\
p_3(\mathbf{X}) &= \int_{\mathbf{\Omega}} p_2(\mathbf{Y}) K(\mathbf{Y} \to \mathbf{X})\, d\mathbf{Y} \\
&\vdots \\
p_t(\mathbf{X}) &= \int_{\mathbf{\Omega}} p_{t-1}(\mathbf{Y}) K(\mathbf{Y} \to \mathbf{X})\, d\mathbf{Y} \\
&\vdots
\end{aligned}$$

Now if our random walk, which is embodied by $K(\mathbf{Y} \to \mathbf{X})$, is *ergodic*, then there exists a *stationary distribution* $p(\mathbf{X})$ such that:

$$p(\mathbf{X}) = \int_{\mathbf{\Omega}} p(\mathbf{Y}) K(\mathbf{Y} \to \mathbf{X})\, d\mathbf{Y} \tag{1}$$

That means that for some $p(\mathbf{X})$, the application of a step in our random walk will give back the exact same distribution. Furthermore, ergodicity guarantees that the random walk will converge to $p(\mathbf{X})$ for any initial distribution $p_1(\mathbf{X})$:

$$\lim_{t \to \infty} p_t(\mathbf{X}) = p(\mathbf{X})$$

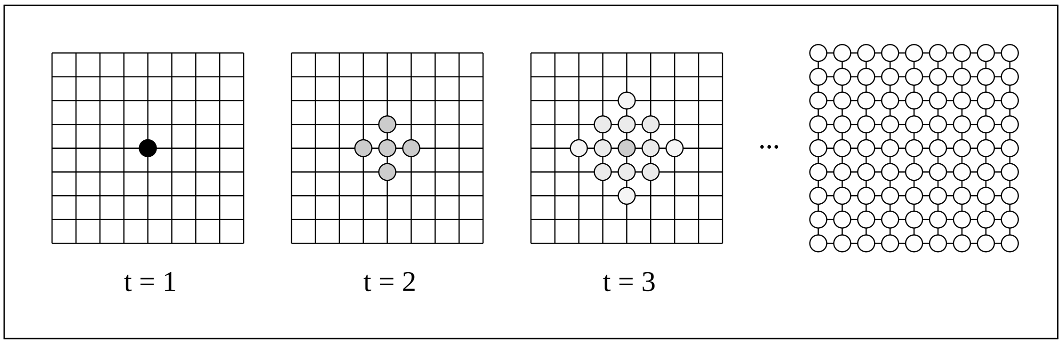

Figure 4: A Random Walk Across a Grid

$M(RT)^2$ constructs a $K(\mathbf{X} \to \mathbf{Y})$ so that its stationary distribution $p(\mathbf{X})$ is proportional to the desired $f(\mathbf{X})$. Thus, the samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ taken along a random walk will eventually obey the desired distribution $p(\mathbf{X})$ no matter what initial distribution $p_1(\mathbf{X})$ is used to pick $\mathbf{X}_1$. Although we will not get into the details of ergodicity, it is a relatively weak condition: there must exist a possibility of getting to any state from any other state in a finite number of steps in non-periodic fashion.

Before getting into the $M(RT)^2$ construction of $K(\mathbf{Y} \to \mathbf{X})$, we will first give an example of an ergodic random walk for clarity. Imagine a wrap-around grid of size $M \times M$, as illustrated in Figure 4. Each node on the grid represents a distinct point in the space $\mathbf{\Omega}$:

$$\begin{aligned} \mathbf{\Omega} &= \{\mathbf{X} = [i,j] \mid 1 \leq i,j \leq M\} = \{[1,1], \ldots, [M,M]\} \\ |\mathbf{\Omega}| &= M^2 \end{aligned}$$

We want to generate samples for $\mathbf{\Omega}$ with a uniform distribution:

$$\begin{aligned} f([i,j]) &= 1 \\ p([i,j]) &= \frac{1}{M^2} \end{aligned}$$

We propose a transition function $K(\mathbf{Y} \to \mathbf{X})$ that will keep us at the current node or move us to an adjacent node with equal probability:

$$K([i',j'] \to [i,j]) = \begin{cases} \frac{1}{5} & \text{if } |i'-i| + |j'-j| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Figure 4 shows what happens with our random walk after a few steps, assuming an initial distribution $p_1([i,j])$ which puts $\mathbf{X}_1$ at the center. The intensity of a location corresponds to the value of the density function $p_t([i,j])$ at step t of the random walk. By observation, we can see that $p_t([i,j])$ converges to $p([i,j])$, and that $p([i,j])$ is indeed a static distribution.

Note that if $\mathbf{\Omega}$ is a discrete domain as in the above example, then $p(\mathbf{X})$ may be replaced by a column vector $\mathbf{p}$ with height equal to $|\mathbf{\Omega}|$. Each entry represents the discrete probability of being in a given state of $\mathbf{\Omega}$. Similarly, $p_t(\mathbf{X})$ may be replaced with a column vector $\mathbf{p_t}$. Furthermore, $K(\mathbf{Y} \to \mathbf{X})$ may be replaced by a matrix $\mathbf{K}$ of size $|\mathbf{\Omega}| \times |\mathbf{\Omega}|$. Each entry (l, k) represents the probability of the random walk moving from state l to state k, and the sum of any row is equal to 1. Given this matrix notation, the distribution for our random walk becomes:

$$\mathbf{p_t} = \mathbf{K}\mathbf{p_{t-1}} = \mathbf{K}^{t-1}\mathbf{p_1}$$

Now if $\mathbf{K}$ is ergodic, then for any $\mathbf{p_1}$ we get:

$$\lim_{t\to\infty} \mathbf{K}^t\mathbf{p_1} = \mathbf{p}$$

And the equation satisfied by the stationary distribution becomes:

$$\mathbf{p} = \mathbf{K}\mathbf{p}$$

Thus $\mathbf{p}$ is an eigenvector of $\mathbf{K}$ with eigenvalue 1. Thus in discrete domains, finding a transition function whose stationary distribution is proportional to an arbitrary function is thus equivalent to constructing a matrix with a specific eigenvector (modulo ergodicity).

3.2 The Tentative Transition Function

In order to construct $K(\mathbf{Y} \to \mathbf{X})$, we will break it up into two parts. A *tentative transition* is first used to propose a transition from the current state into some other state according to some chosen distribution. For example, in the aforementioned m-particle problem, a tentative transition could be to move a single particle by some small random distance. The pdf of this *tentative transition function* is denoted by $T(\mathbf{Y} \to \mathbf{X})$. Now, once this transition is proposed, we either choose to move to this new state for the next step of the random walk, or else we stay in the same state for the next step of the random walk. This *acceptance probability*, whose pdf is denoted by $A(\mathbf{Y} \to \mathbf{X})$, is chosen in a manner such that the combination of $T(\mathbf{Y} \to \mathbf{X})$ and $A(\mathbf{Y} \to \mathbf{X})$ form a $K(\mathbf{Y} \to \mathbf{X})$ whose stationary distribution is the desired function $p(\mathbf{X})$.

3.3 Detailed Balance

The question we ask ourselves at this point is as follows: what is the relationship between $T(\mathbf{Y} \to \mathbf{X})$, $A(\mathbf{Y} \to \mathbf{X})$, and $p(\mathbf{X})$? To answer this, we borrow an idea from physical equilibria. Imagine a box filled with a gas and another box that contains nothing. A clamped tube connects the two boxes. Once the tube is unclamped, gas begins to flow from the first box into the second box. However, once the second box starts holding some gas, some of that gas flows back into the first box. After some time, the two boxes reach a state of equilibrium. Even though gas flows back and forth between the two boxes and the state of the system is not static, the amount that goes in each direction is equal, and thus the system is at equilibrium.

Let us examine what happens when we apply the notion of *detailed balance* to our problem:

$$p(\mathbf{X})T(\mathbf{X} \to \mathbf{Y})A(\mathbf{X} \to \mathbf{Y}) = p(\mathbf{Y})T(\mathbf{Y} \to \mathbf{X})A(\mathbf{Y} \to \mathbf{X}) \tag{2}$$

This equation says that once a stationary distribution $p(\mathbf{X})$ is reached, then the chance of being in state $\mathbf{X}$ and then taking a proposed transition into state $\mathbf{Y}$ is equal to the chance of being in state $\mathbf{Y}$ and then taking a proposed transition into state $\mathbf{X}$. To see how this provides a solution to stationary distribution, we rewrite Equation (1) in terms of $T(\mathbf{Y} \to \mathbf{X})$ and $A(\mathbf{Y} \to \mathbf{X})$:

$$\begin{aligned} p'(\mathbf{X}) \;=\; & \left[\int_{\mathbf{\Omega}} p(\mathbf{Y})T(\mathbf{Y} \to \mathbf{X})A(\mathbf{Y} \to \mathbf{X})\, d\mathbf{Y}\right] + \\ & \left[p(\mathbf{X})\left(1 - \int_{\mathbf{\Omega}} T(\mathbf{X} \to \mathbf{Y})A(\mathbf{X} \to \mathbf{Y})\, d\mathbf{Y}\right)\right] \end{aligned}$$

The first term of the above equation expresses the chance of starting in state $\mathbf{Y}$ and taking a transition into $\mathbf{X}$. The second term expresses the chance of starting in state $\mathbf{X}$ and rejecting a transition into another state $\mathbf{Y}$. These are all the ways that one can end up in state $\mathbf{X}$ after a step of the random walk. Rearranging the terms and moving $p(\mathbf{X})$ into the integral (since it is constant with respect to $\mathbf{Y}$), we get:

$$\begin{aligned} p'(\mathbf{X}) \;=\; & p(\mathbf{X}) + \\ & \int_{\mathbf{\Omega}} p(\mathbf{Y})T(\mathbf{Y} \to \mathbf{X})A(\mathbf{Y} \to \mathbf{X}) - p(\mathbf{X})T(\mathbf{X} \to \mathbf{Y})A(\mathbf{X} \to \mathbf{Y})\, d\mathbf{Y} \end{aligned}$$

Thus in order to get a stationary distribution, the integral in the above equation must evaluate to zero. One easy way to do this is to enforce the detailed balance of Equation (2) in our choice of $T(\mathbf{Y} \to \mathbf{X})$ and $A(\mathbf{Y} \to \mathbf{X})$. Thus, we are left with:

$$p'(\mathbf{X}) = p(\mathbf{X})$$

3.4 The Acceptance Probability

Since $p(\mathbf{X})$ is given to us (by way of $f(\mathbf{X})$), and $T(\mathbf{Y} \to \mathbf{X})$ is some guess for transitions, then our only choice is to set $A(\mathbf{Y} \to \mathbf{X})$ appropriately in order to meet the condition of detailed balance. From Equation (2), we know that we must have:

$$\frac{A(\mathbf{Y} \to \mathbf{X})}{A(\mathbf{X} \to \mathbf{Y})} = \frac{p(\mathbf{X})T(\mathbf{X} \to \mathbf{Y})}{p(\mathbf{Y})T(\mathbf{Y} \to \mathbf{X})} \tag{3}$$

One way to achieve this is:

$$A(\mathbf{Y} \to \mathbf{X}) = min\left(1, \frac{f(\mathbf{X})T(\mathbf{X} \to \mathbf{Y})}{f(\mathbf{Y})T(\mathbf{Y} \to \mathbf{X})}\right) \tag{4}$$

Noting that $f(\mathbf{X})/f(\mathbf{Y}) = p(\mathbf{X})/p(\mathbf{Y})$ since $p(\mathbf{X}) \propto f(\mathbf{X})$, it is clear from substitution that Equation (4) satisfies Equation (3). Note that all of the above scalar values can be readily evaluated for any pair of points. Thus we are now ready for the M(RT)^2 algorithm.

3.5 The M(RT)^2 Algorithm

Assuming we have a way of evaluating $f(\mathbf{X})$ at a point, and a way of evaluating $T(\mathbf{Y} \to \mathbf{X})$, then the algorithm for generating samples with a pdf proportional to $f(\mathbf{X})$ is as follows:

$\mathbf{X}_1 \sim p_1(\mathbf{X})$
for $t = 2$ to N
  $\mathbf{X}_{tent} \sim T(\mathbf{X}_{t-1} \to \mathbf{X})$
  if (random() $< A(\mathbf{X}_{t-1} \to \mathbf{X}_{tent})$)
    $\mathbf{X}_t = \mathbf{X}_{tent}$
  else
    $\mathbf{X}_t = \mathbf{X}_{t-1}$

Note that one may want to reject the first several samples since it takes a while for the random walk to reach the stationary distribution.