1 Motivation: time scale limit and rare events

Atomistic simulations such as Molecular Dynamics (MD) and Monte Carlo (MC) are playing an important role in our understanding of macroscopic behavior of materials in terms of the atomistic mechanisms. But the range of applicability of atomistic simulations is determined by their limits of time and length scales, see Fig. 1. While the length scale limit of atomistic simulations has been extended significantly through the use of massively parallel computing, the time scale limit is still a critical bottleneck to date.

Figure 1: Time and length scale limits of atomistic simulations.

*Lecture notes and Matlab codes available at [1].
To appreciate the time scale problem, consider the Molecular Dynamics simulation, which is essentially the numerical integration of the Newton’s equation of motion for a collection of interacting atoms. For the numerical integration to be stable, the time step should be much smaller than the inverse of the maximum vibration frequency of the atoms, which is the Debye frequency in a solid. Therefore, a time step on the order of 1fs ($10^{-15}$ second) is quite common. Thus even if the simulation is carried out for a million steps, we only cover a period of 1ns ($10^{-9}$ second). This is approximately the upper time scale limit of today’s atomistic simulations. On the other hand, there are many important processes for which the atomistic detail is important but which occurs at much longer time scales, ranging from 1ms ($10^{-3}$ second) to many years ($\sim 10^8$ second). The enormous gap between the limit of Molecular Dynamics simulations and these demands are too big to be closed by massively parallel computing alone.

![Figure 2: Typical energy landscape of many-body systems contains many local energy minima separated by energy barriers much higher than thermal energy ($kT$). Transitions between energy basins can be rare events at the atomistic time scale.](image)

A typical reason for the appearance of this gap in the time scale can be traced to the energy landscape of typical many-body systems, which usually contains many local energy minima separated by energy barriers that are much higher than the thermal energy $kT$, where $k$ is the Boltzmann’s constant. Every deep energy minima (or energy basin) like this corresponds to a meta-stable state, in which the system could spend a long time before making a transition to a neighboring meta-stable state. The long time scale evolution of the system can be characterized by the transition of the system from one meta-stable state to another. However, because such transitions can be very rare, a direct Molecular Dynamics simulation is very inefficient to capture the long time scale behavior – most of the computing cycle is “wasted” on “faithfully” tracing the quasi-equilibrium fluctuation of the system within a meta-stable state, which may not be of our interest at all.

Similar problems can occur in Monte Carlo simulations as well. For example, consider a Monte Carlo simulation of a particle moving on a 2-dimensional energy landscape shown in Fig. 3(a). When the thermal energy $kT$ is much lower than the energy barrier between the two metastable states ($A$ and $B$), the particle spends a long “time” (in terms of Monte
Figure 3: (a) Two-dimensional potential energy function $V(x, y)$. The two meta-stable states, $A$ for $x < 0$ and $B$ for $x > 0$, are separated by a ridge at $x = 0$. $S_1$ and $S_2$ mark locations of two saddles on the ridge. (b) Instantaneous values of coordinate $x$ in an MC simulation at $kT = 0.15$. The energy barrier at the saddle points are on the order of 1.

Carlo steps) in one state before jumping to the other one, as shown in Fig. 3(b). Again, when this is the case, Monte Carlo simulations are very inefficient in capturing the long time scale evolution of the system.

The purpose of these lectures is to introduce an approach to address this problem, based on the idea of importance sampling [2]. The hope is that the discussions here will eventually lead to a robust numerical method that can significantly extend the time scale limit of Monte Carlo (and possibly Molecular Dynamics) simulations of complex atomistic systems. In these lectures, we will describe the idea with examples on low dimensional model systems, on which the importance sampling method for rare event sampling has been successfully developed [3, 4].

The idea of importance sampling of rare events can be summarized as the following. We may regard the entire trajectory of a Monte Carlo simulation as consisting of many short trajectories, or paths. Most of the paths start from the bottom of an energy basin and returns to the bottom of the same basin, such as paths $0 \rightarrow 1, 1 \rightarrow 2$ and $2 \rightarrow 3$ in Fig. 4. These paths can be called “failed” paths. Very rarely a path may leave from the bottom of one basin and reach the bottom of another basin; such a path is called a “successful” path. The origin of the time scale problem is that the “successful” paths, while being very important to us, are sampled with very low efficiency. When the problem is formulated in this way, the importance sampling method seems to be a natural solution — it is developed to bias the sampling procedure in such a way that the more important events (here the successful paths) are sampled with higher efficiency. Compared with the many other approaches proposed to address the time scale problem, using the importance sampling method is a relatively “new” approach, even though the importance sampling method has a long history by itself and has been successfully applied to numerous other situations.

In the following, we will start our discussions with the original idea of importance sampling. The idea is best illustrated when Monte Carlo is used to compute the value of an integral.
Figure 4: Schematic representation of a reactive trajectory in an energy landscape with meta-stable states $A$ and $B$. Such a trajectory can be sliced into a sequence of “failed” paths followed by a single “successful” path. A failed path is defined as a sequence of microstates that initiates in region $A$, exits it at some instant, but returns to it before reaching $B$. In contrast, a successful segment is defined as a sequence of states that initiates in $A$ and succeeds in reaching $B$ before returning to state $A$. The shown reactive trajectory consists of three failed paths, namely, the sequences of states $0 \rightarrow 1$, $1 \rightarrow 2$, and $2 \rightarrow 3$, and the successful path $3 \rightarrow 4$.

2 Monte Carlo computation of integrals

Very often, the purpose of a Monte Carlo simulation can be reformulated as for computing the value of an integral. A typical example is to use Monte Carlo to evaluate the canonical ensemble average of a certain quantity, which may be expressed as a function of the positions of all atoms, e.g. $A(\{r_i\})$. The ensemble average of $A$ is,

$$\langle A \rangle = \frac{1}{Q} \int \prod_{i=1}^{N} dr_i A(\{r_i\}) \exp \left[ -\frac{V(\{r_i\})}{kT} \right]$$

(1)

where

$$Q = \int \prod_{i=1}^{N} dr_i \exp \left[ -\frac{V(\{r_i\})}{kT} \right]$$

(2)

and $N$ is the total number of atoms, $V$ is the potential energy function, $k$ is the Boltzmann’s constant and $T$ is the temperature. For large $N$, the dimension of the integral is too large to be carried out by numerical quadrature. Monte Carlo is a convenient method to evaluate this integral. By sampling a sequence of configurations, $\{r_i\}^{(m)}$, $m = 1, \cdots, M$, that satisfies the probability distribution,

$$f(\{r_i\}) = \frac{1}{Q} \exp \left[ -\frac{V(\{r_i\})}{kT} \right]$$

(3)
the ensemble average can be estimated from the numerical average [5],

\[
\langle A \rangle_{MC} = \frac{1}{M} \sum_{m=1}^{M} A(\{r_i\}^{(m)})
\] (4)

Contrary to the claims in [6], Eq. (4) is not importance sampling. As discussed below, we call something importance sampling if an importance function is introduced to further reduce the variance in the sampling.

2.1 One dimensional integral

To illustrate the idea of importance sampling, let us consider the integration of a simple one-dimensional function, \( g(x) = \sqrt{1 - x^2} \). Suppose we wish to calculate,

\[
G = \int_{0}^{1} \sqrt{1 - x^2} \, dx
\] (5)

Because \( G = \pi/4 \), we can regard this as a way to compute the numerical value of \( \pi \). This example is a favorite one in many text books and was developed in detail in [2]. Thus this section may be regarded as an illustration of chapter 4 of [2].

![Figure 5](image)

(a) \( g(x) = \sqrt{1 - x^2} \) and \( f(x) = 1 \) in domain \( x \in [0,1] \). (b) Dots: error of estimating \( \pi/4 \) by Monte Carlo integration of \( \int_{0}^{1} g(x) \, dx \) using \( N \) random numbers with distribution \( f(x) \). Solid line: \( \sigma_1/N^{1/2} \), where \( \sigma_1 \approx 0.223 \). (Matlab code `mcint1d.m` [1])

One way to compute this integral is to use numerical quadrature. For example, the trapezoidal rule gives,

\[
G \approx \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} [g(x_{i-1}) + g(x_i)]
\] (6)

where \( x_i = i/N \).

However, for illustration purposes, let us try to evaluate \( G \) by a Monte Carlo method. Instead of computing \( g(x) \) at regular intervals between 0 and 1, let us draw \( N \) random
numbers \( x_i, i = 1, \ldots, N \) that are independently and uniformly distributed in \([0, 1]\). Let \( G_N \)
be the average of these random numbers,

\[
G_N = \frac{1}{N} \sum_{i=1}^{N} g(x_i)
\]  

(7)

Notice that \( G_N \) is a random number itself. The expectation value of \( G_N \) is \( G \), i.e.,

\[
G = \langle G_N \rangle
\]  

(8)

Mathematically, this means we rewrite \( G \) as,

\[
G = \int g(x) f(x) \, dx = \langle g \rangle_f
\]  

(9)

where \( f(x) \) is the distribution density function of the random variables, \( f(x) = 1 \) if \( 0 \leq x \leq 1 \) and \( f(x) = 0 \) otherwise. In other words, \( G \) is the average of function \( g(x) \) under distribution \( f(x) \). The variance of \( G_N \) is

\[
\text{var}\{G_N\} = \frac{\sigma_1^2}{N}
\]  

(10)

where

\[
\text{var}\{g\} = \sigma_1^2 = \int g^2(x) f(x) \, dx - G^2
\]  

(11)

The error of \( G_N \) in estimating \( G \) is comparable to the standard deviation of \( G_N \), i.e.

\[
\text{error} = \varepsilon \approx \frac{\sigma_1}{N^{1/2}}
\]  

(12)

In this example, \( \sigma_1^2 = 2/3 - (\pi/4)^2 \approx 0.0498 \), \( \sigma_1 \approx 0.223 \). According to Eq. (12), we expect the error, \(|G_N - \pi/4|\), should decay as \( 0.223/N^{1/2} \) as the number \( N \) of random variables increase. This is confirmed by numerical results, as shown in Fig. 5(b).

### 2.2 Importance sampling in Monte Carlo integration

Using a uniform distribution function \( f(x) \) is not necessarily the optimal way of doing Monte Carlo integration, in the sense that it does not necessarily give the smallest statistical fluctuation, i.e. \( \text{var}\{G_N\} \).

Imagine that we use random variables that satisfy another density distribution function, \( \tilde{f}(x) \). Then the integral can be rewritten as,

\[
G = \int \left[ \frac{f(x)g(x)}{\tilde{f}(x)} \right] \tilde{f}(x) \, dx \equiv \int \tilde{g}(x) \tilde{f}(x) \, dx = \langle \tilde{g} \rangle_{\tilde{f}}
\]  

(13)

This means that \( G \) is also the expectation value of \( \tilde{g}(x) \equiv f(x)g(x)/\tilde{f}(x) \) under distribution function \( \tilde{f}(x) \). The variance of this estimator is,

\[
\text{var}\{\tilde{g}\} = \int \tilde{g}^2(x) \tilde{f}(x) \, dx - G^2 = \int \left[ \frac{f(x)g(x)}{\tilde{f}(x)} \right]^2 \tilde{f}(x) \, dx - G^2 = \int \frac{f^2(x)g^2(x)}{\tilde{f}(x)} \, dx - G^2
\]  

(14)
To illustrate that different distribution functions give rise to different variance, consider a family of \( \tilde{f}(x) \),

\[
\tilde{f}(x) = \begin{cases} 
(1 - \beta x^2)/(1 - \beta/3), & 0 \leq x \leq 1 \\
0, & \text{otherwise}
\end{cases}
\]  

(15)

Notice that when \( \beta = 0 \), \( \tilde{f}(x) \) corresponds to the uniform distribution considered above.

Fig. 6(a) shows the numerical error of \( G_N \) (compared with \( \pi/4 \)) as a function of \( N \) (number of random variables) for \( \beta = 0 \) and \( \beta = 0.74 \). The error is systematically lower in the \( \beta = 0.74 \) case, which scales as \( 0.054/N^{1/2} \). The variance of \( \tilde{g} \) under distribution function \( \tilde{f} \) can be obtained analytically [2],

\[
\text{var}\{\tilde{g}\} = \left(1 - \frac{\beta}{3}\right) \left[\frac{1}{\beta} - \frac{1 - \beta}{\beta^{3/2}} \tanh^{-1} \beta^{1/2}\right] - \left(\frac{\pi}{4}\right)^2
\]  

(16)

It is plotted in Fig. 6(b). The variance reaches a minimum at \( \beta \approx 0.74 \). This means that, among the family of importance functions given in Eq. (15), the one with Eq. (15) is the (local) optimal.

**Exercise:** for \( \tilde{f}(x) = (1 - \beta x^2)/(1 - \beta/2) \), derive \( \text{var}\{\tilde{g}\} \) and the optimal value of \( \beta \). Plot numerical error of \( \langle \tilde{g} \rangle \tilde{f} \) as a function of \( N \) (as in Fig. 6(a)) for \( \beta = 0 \) and the (local) optimal value of \( \beta \).

What is the optimal importance function \( \tilde{f} \) among all possible functions? To answer this question, we should minimize \( \text{var}\{\tilde{g}\} \) with respect to \( \tilde{f} \) under the constraint that \( \int \tilde{f}(x) dx = 1, \tilde{f}(x) \geq 0 \) (because \( \tilde{f} \) is a probability distribution function). Introduce Lagrangian multiplier \( \lambda \), we wish to minimize a functional [2],

\[
L\{\tilde{f}\} = \int \frac{f^2(x)g^2(x)}{\tilde{f}(x)} + \lambda \tilde{f}(x) dx
\]  

(17)

At the minimum,

\[
\frac{\delta L}{\delta \tilde{f}} = -\frac{f^2(x)g^2(x)}{\tilde{f}^2(x)} + \lambda = 0
\]  

(18)
Hence,
\[ \tilde{f}(x) = \lambda |g(x)f(x)| \] (19)

The normalization condition of \( \tilde{f}(x) \) demands that \( \lambda = 1/G \), so that the (global) optimal importance function is,
\[ \tilde{f}(x) = \frac{1}{G} |g(x)f(x)| \] (20)

Notice that when the (global) optimal importance function is used,
\[ \tilde{g}(x) = G \frac{f(x)g(x)}{|f(x)g(x)|} \] (21)

If \( g(x) \) is non-negative, this means that \( \tilde{g}(x) \) is a constant, and equals \( G \). Therefore, when the (global) optimal importance function is used, the statistical variance is zero, because every random number gives identical contribution to the integral. This finding is not surprising because whenever \( \tilde{g}(x) \) is not a constant, there is always a finite variance when computing its average. Zero variance is reached if and only if \( \tilde{g}(x) \) is a constant.

Unfortunately, the (global) optimal importance function contains the information of the answer we were after (\( G \)) in the first place. This means finding the (global) optimal importance function is not easier than finding the integral \( G \) itself. Thus for practical purposes, we need to assume that we do not have the (global) optimal importance function when we compute \( G \). Fortunately, even if \( \tilde{f} \) is not the (global) optimum, it can still reduce the statistical variance of the estimate of \( G \). The closer \( \tilde{f} \) is to the (global) optimum, the smaller the variance becomes. Therefore, the importance sampling method is usually accompanied by the effort to improve the importance function.

3 Random walks, differential and integral equations

Usually, Monte Carlo simulations use a Markov process to generate a sequence of random variables to be used in the numerical integration. In this case, the system may be regarded as doing random walks in the phase space. We have already seen the connection between the Monte Carlo method and integrals. This is closely related to the (probably even deeper) connection between random walks and integral equations [2, 8, 9] and partial differential equations (PDEs) [7].

Consider the random walk problem (or game) shown in Fig. 7. The walkers move on discrete lattice points (with lattice spacing \( a \)) in domain \( \Omega \), with boundary \( \Gamma \). The walker start at point \( x \) with probability distribution \( s(x) \) (source function), with \( \int_{\Omega} s(x)dx = 1 \). At every step, the walker may be terminated with probability \( q(x) \). Otherwise, it moves to one of its adjacent lattice points with equal probabilities. The walker accumulates a score \( g(x) \geq 0 \) (gain function) whenever it visits point \( x \). Suppose we wish to compute \( G \), which is the average score a walker accumulates from its start until it is terminated. Let \( \psi(x) \) be the probability that a walker visits point \( x \) before its termination, then \( G \) can be expressed as an integral,
\[ G = \int_{\Omega} \psi(x)g(x)dx \] (22)
Figure 7: A random walk in a 2-dimensional domain $\Omega$ with boundary $\Gamma$. The walker start at point $x$ with probability distribution $s(x)$ (source function). At every step, the walker may be terminated with probability $q(x)$. Otherwise, it moves to one of its adjacent lattice points with equal probabilities. The walker accumulates a score $g(x)$ (gain function) whenever it visits point $x$.

Question: what is the expression for $G$ if the score is counted only when the path is terminated?

Note that we may demand $q(x) = 1$ at $x \in \Gamma$. This corresponds to the absorbing boundary condition. On the other hand, domain $\Omega$ can also be infinitely large for which $\Gamma$ does not exist. We may also let $g(x)$ to be zero everywhere in $\Omega$ except in $\Gamma$ and let $g(x) = 1$ on $\Gamma$. Then $G$ describes the possibility of the random walkers to escape from domain $\Omega$ without being absorbed (such as the escape of radiation from a container).

We can imagine that $s(x)$ describes a drug delivery device (implant), $\Omega$ is the tissue that the drug diffuses through (which may also absorb the drug molecules), and $g(x)$ describe the diseased tissue (target, e.g. cancer cells). $G$ the characterizes the effectiveness of the drug delivery device.

### 3.1 Partial differential equation

How do we compute $G$? One way to do so is to first compute $\psi(x)$ at all $x \in \Omega$ and perform the integral in Eq. (22) by numerical quadrature. In the continuum limit ($a \to 0$), $\psi(x)$ satisfies the following partial differential equation (PDE),

$$a^2 \nabla^2 \psi(x) - q(x)\psi(x) = s(x)$$  (23)
This is simply a absorption-diffusion equation. The problem with this approach is that, it can be very wasteful since we only want to know the value of $G$, which is a weighted integral of $\psi(x)$. Yet we obtained $\psi(x)$ at every point $x$ in $\Omega$. For example, $g(x)$ may be non-zero only in a small region of $\Omega$. Hence, in principle, we do not need to know $\psi(x)$ in the rest region of $\Omega$. $\Omega$ may also be infinite, in which case it may not be possible to store values of $\psi(x)$ everywhere in $\Omega$.

### 3.2 Monte Carlo simulation

Another approach is to perform Monte Carlo simulation of the random walkers according to the prescribed rules, and estimate $G$ by averaging the gain function $g(x)$ over the locations where the walkers have visited. Suppose we initiate $N$ walkers, then

$$
G_N = \frac{1}{N} \sum_{n=1}^{N} \sum_{i} g(x_i^{(n)})
$$

is an estimate of $G$, where $x_i^{(m)}$ is the location of walker $m$ at step $i$. The potential problem of this approach is that the efficiency may be very low, if the majority of the walkers do not visit regions where $g(x) > 0$.

For example, consider the case when $g(x) = 1$ in a small region of $\Omega$ that is separated from the region where $s(x) = 1$ and that the termination rate $q(x)$ is high compared in the entire domain $\Omega$. Also assume that $q(x) = 1$ at locations where $g(x) = 1$, i.e. the walkers terminate immediately after it scores (but the majority of the walkers terminate at $g(x) = 0$, i.e. without score). Let $p$ be the probability that the walker makes a non-zero contribution ($g(x) = 1$) to the total score $G$, then,

$$
G = \langle G_N \rangle = p
$$

The variance of $G_N$ is,

$$
\text{var}\{G_N\} = \frac{p(1-p)}{N}
$$

so that the error in estimating $G$ is approximately,

$$
\text{error} = \epsilon \approx \frac{p^{1/2}(1-p)^{1/2}}{N^{1/2}}
$$

In the limit of $p \ll 1$,

$$
\epsilon \approx \frac{p^{1/2}}{N^{1/2}}
$$

and the relative error is,

$$
\frac{\epsilon}{G} \approx \frac{1}{(pN)^{1/2}}
$$

Thus we obtain meaningful results only if $N \gg 1/p$. For example, if $p \sim 10^{-6}$, the $N$ needs to be $10^8$ to achieve a relative error of 1%.
3.3 Integral equation and importance sampling

To achieve a lower variance in computing $G$ with smaller number of random variables, importance sampling is a promising method. Recall that importance sampling is useful to reduce variance when computing an integral. To apply importance sampling here, we first notice that the walker distribution function $\psi(x)$ satisfies the following integral equation (which is a counterpart of the PDE in Eq. (23)),

$$\psi(x) = s(x) + \int K(y \rightarrow x) \psi(y) \, dy$$

(30)

where $K(y \rightarrow x)$ is the transition probability of the random walkers. In this example, $K(y \rightarrow x) = (1 - q(y))/4$ if $x$ is one of the four nearest neighbors of $y$; otherwise, $K(y \rightarrow x) = 0$.

Let us now introduce an importance function $I(x) \geq 0$ for $x \in \Omega$. Define

$$S_0 = \int I(x) s(x) \, dx$$

(31)

Multiply both sides of Eq. (30) by $I(x)/S_0$, we obtain,

$$\tilde{\psi}(x) \equiv \frac{I(x) \psi(x)}{S_0} = \frac{s(x) I(x)}{S_0} + \int K(y \rightarrow x) \frac{I(x)}{I(y)} \frac{I(y)}{S_0} \psi(y) \, dy$$

(32)

$$\tilde{s}(x) \equiv \frac{s(x) I(x)}{S_0}$$

(34)

$$\tilde{K}(y \rightarrow x) \equiv K(y \rightarrow x) \frac{I(x)}{I(y)}$$

(35)

notice that $\int \tilde{s}(x) \, dx = 1$, we have,

$$\tilde{\psi}(x) = \tilde{s}(x) + \int \tilde{K}(y \rightarrow x) \tilde{\psi}(y) \, dy$$

(36)

which is a new integral equation relating $\tilde{\psi}$, $\tilde{s}$ and $\tilde{K}$. And the reward function can be rewritten as,

$$G = \int g(x) \psi(x) = S_0 \int \frac{g(x)}{I(x)} \tilde{\psi}(x) \, dx = \int \tilde{g}(x) \tilde{\psi}(x) \, dx$$

(37)

where $\tilde{g}(x) \equiv S_0 g(x)/I(x)$. This means that, we can estimate $G$ by simulating random walkers according to the transition probability matrix $\tilde{K}(Y \rightarrow X)$. Suppose one path is sampled following $\tilde{K}(Y \rightarrow X)$, which visits locations $\{x_1, x_2, \cdots, x_l\}$, then

$$G \approx S_0 \sum_{i=1}^l \frac{g(x_i)}{I(x_i)}$$

(38)
Figure 8: (a) Random walk is within the (red) rectangle on a square lattice with spacing $a = 1$. The source function $s(x)$ equals 1 at the circle and zero everywhere else. The contour line of $\psi(x)$ in (b) is also plotted. (b) The density distribution $\psi(x)$ from 1000 random walkers. (c) Estimates of $G$ using importance functions with different parameter $\mu$. (d) Standard deviation of estimates in (c) as a function of $\mu$. 
In the new random walk, the walker distribution density is $\tilde{\psi}(x)$. The variance is reduced if there is a significant overlap between regions where $\tilde{\psi}(x) > 0$ and $\tilde{g}(x) > 0$.

To give a specific example, consider a rectangular domain shown in Fig. 8(a). The source function $s(x)$ equals 1 at one point $x_0$ (marked by the circle) and zero everywhere else. The gain function $g(x) = 0$ and the termination probability $q(x) = r (r = 0.05)$ everywhere inside $\Omega$ except at boundary $\Gamma$. At boundary $\Gamma$, $g(x) = 1$ and $q(x) = 1$. Thus $G$ is the probability of the walkers escaping (reaching $\Gamma$) before being absorbed at the interior of $\Omega$.

The distribution $\psi(x)$ estimated from 1000 random walkers are plotted in Fig. 8(b). We notice that most of the walkers terminate at the interior of $\Omega$. The estimated value from $10^4$ walkers is $G \approx 4.84 \pm 0.28 \times 10^{-2}$.

Suppose we use the importance function of the following form,

$$I(x) = \exp \left[ -\mu \exp \left( -\alpha |x - x_0|^2 \right) \right] \quad (39)$$

Fig. 8(c) plots the estimate of $G$ for different values of $\mu$ when $\alpha = 0.03$ from $10^4$ walkers. The standard deviation of these estimates are plotted in Fig. 8(d). We observed that $\mu \approx 3$ is the optimal value giving rise to the lowest statistical error.

References


Importance Sampling in Monte Carlo Simulation of Rare Transition Events

Wei Cai

Lecture 2. August 2, 2005

Figure 1: (a) Contour line of the exact solution of the walker density distribution function \( \psi(x) \), which satisfies the integral equation, \( \psi(x) = s(x) + \int \psi(y)K(y \to x)\,dy \). \( G = \int \psi(x)g(x)\,dx = 4.6844 \times 10^{-2} \). (diffuse2d_exact_sol.m) (b) 3D plot of \( \psi(x) \). (c) Exact solution of the optimal importance function \( J(x) \), which satisfies the integral equation, \( J(x) = g(x) + \int K(x \to y)J(y)\,dy \). \( G = \int J(x)s(x)\,dx = 4.6844 \times 10^{-2} \). (diffuse2d_opt_imp.m)

1. Review

<table>
<thead>
<tr>
<th>Original random walk</th>
<th>Importance sampled random walk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source function $S(x)$</td>
<td>$S'(x) = S(x) I(x)/S_0$</td>
</tr>
<tr>
<td>$\int S(x) , dx = 1$</td>
<td>$S_0 = \int S(x) I(x) , dx$</td>
</tr>
<tr>
<td>$\int S'(x) , dx = 1$</td>
<td>$S_0 = \int S'(x) , dx$</td>
</tr>
<tr>
<td>Gain function $g(x)$</td>
<td>$g'(x) = \frac{S_0 g(x)}{I(x)}$</td>
</tr>
<tr>
<td>Transition probability $K(x \to y)$</td>
<td>$K'(x \to y) = K(x \to y) \frac{I(y)}{I(x)}$</td>
</tr>
<tr>
<td>$\int K(x \to y) , dy \leq 1$</td>
<td>$\int K'(x \to y) , dy \to 1$</td>
</tr>
<tr>
<td>$1 - \int K(x \to y) , dy = g(x)$</td>
<td>$1 - \int K'(x \to y) , dy = g'(x)$</td>
</tr>
<tr>
<td>Integral equation</td>
<td>$\tilde{g}(x) = \tilde{g}(x) \cdot I(x)/S_0$</td>
</tr>
<tr>
<td>Walker density: $\tilde{4}(x) = \tilde{S}(x) + \int \tilde{K}(y-x) \tilde{g}(y) , dy$</td>
<td>$\tilde{g}(x) = \tilde{S}(x) + \int \tilde{K}(y-x) \tilde{g}(y) , dy$</td>
</tr>
<tr>
<td>Gain</td>
<td>$G = \int \tilde{g}(x) g(x) , dx$</td>
</tr>
<tr>
<td>$G' = \int \tilde{g}'(x) \tilde{g}(x) , dx$</td>
<td>$G' = \int \tilde{g}'(x) \tilde{g}(x) , dx$</td>
</tr>
</tbody>
</table>

2. The procedure of importance sampled Monte Carlo.

Suppose $\int K(x \to y) \, dy = \tilde{K}(x) \leq 1$ as in the original MC.

$\tilde{K}(x) = 1 - \tilde{n}(x)$ termination probability

$\tilde{K}'(x \to y) = \frac{1}{\tilde{n}(x)} \tilde{K}(x \to y)$ is a normalized probability

$\int \tilde{K}'(x \to y) \, dy = 1.$
But what if $\tilde{\pi}(x) > 1$?

It does not make sense to have a negative termination probability $\tilde{\pi}(x) = 1 - \tilde{\pi}(x)$.

What shall we do in this case?

To answer this question, let's recall that we use the random walkers to estimate $\tilde{\pi}(x)$.

Let $N^0_\text{tot}$ be the total number of random walkers at the beginning of the simulation.

Let $N_i(x)$ be the number of random walkers at step $i$ of the simulation located at point $x$.

Let $N_i(x) = \sum_{i=1}^{\infty} N_i(x)$ be the total number of times a random walker visited point $x$ during the entire simulation.
We observe that even though \( N_i(x) \) are integers,

\[
E[N_0(x)] = N_0^* \tilde{S}(x) \quad \text{(1)}
\]

\[
E[N_0(x)] = N_0^* \tilde{\varphi}(x) \quad \text{(2)}
\]

Eq. (1) is satisfied by initializing the random walkers
according to distribution function \( \tilde{S}(x) \)

Eq. (2) should be satisfied by the "stochastic rules"
of the Monte Carlo simulation.

3. The relation between integral equation and Monte Carlo rules.

To see how the Monte Carlo rules make \( N_i(x) \) satisfy
Eq. (2), we recall that \( \tilde{\varphi}(x) \) satisfy the integral equation

\[
\tilde{\varphi}(x) = \tilde{S}(x) + \int \tilde{\varphi}(y) \tilde{R}(y-x) dy
\]

One way to solve this equation is
by iteration:

\[
\tilde{\varphi}^{(0)}(x) = \tilde{S}(x)
\]

\[
\tilde{\varphi}^{(1)}(x) = \tilde{S}(x) + \int \tilde{\varphi}^{(0)}(y) \tilde{R}(y-x) dy
\]

\[
\tilde{\varphi}^{(2)}(x) = \tilde{S}(x) + \int \tilde{\varphi}^{(1)}(y) \tilde{R}(y-x) dy
\]

\[
\vdots
\]

\[
\tilde{\varphi}^{(n)}(x) = \tilde{S}(x) + \int \tilde{\varphi}^{(n-1)}(y) \tilde{R}(y-x) dy
\]

If the iteration converges, i.e.

\[
\lim_{N \to \infty} \left| \tilde{\varphi}^{(n+1)}(x) - \tilde{\varphi}^{(n)}(x) \right| = 0
\]

then \( \lim_{N \to \infty} \tilde{\varphi}^{(n)}(x) = \tilde{\varphi}(x) \) is the solution.
\[ \Psi(x) = \tilde{S}(x) \]
\[ + \int dx' \tilde{S}(x') \tilde{K}(x'-x) \]
\[ + \int dx'' dx' \tilde{S}(x'') \tilde{K}(x''-x') \tilde{K}(x'-x) \]
\[ \vdots \]
\[ \Psi(\infty) = \tilde{S} + \sum_{i=1}^{n} \tilde{S} \cdot \tilde{K} \cdot \tilde{K} \cdots \tilde{K} \over \text{i terms} \]
\[ \Psi = \tilde{S} + \sum_{i=1}^{\infty} \tilde{S} \cdot \tilde{K} \cdot \tilde{K} \cdots \tilde{K} \over \text{i terms} \]

These expressions give rise to a numerical method to compute \( \Psi \) directly, which is applicable in low-dimensional systems.

The numerical result for \( \Psi \) is shown in Fig. 1.

From 4, \( G = \int \Psi(x) g(x) \, dx = 4.6844 \times 10^{-2} \) (convergence reached at \( n = 1000 \) terms)

In matrix form, the equation \( \tilde{\Psi} = \tilde{S} + \tilde{\Psi} \cdot \tilde{K} \) can be formally solved as

\[ \tilde{\Psi} \cdot (\mathbf{I} - \tilde{K}) = \tilde{S} \quad \text{where } \mathbf{I} \text{ is the identity matrix} \]

\[ \tilde{\Psi} = \tilde{S} \cdot (\mathbf{I} - \tilde{K})^{-1} = \tilde{S} \cdot (\mathbf{I} + \tilde{K} + \tilde{K} \cdot \tilde{K} + \tilde{K} \cdot \tilde{K} \cdot \tilde{K} + \ldots) \]
Every step of the Monte Carlo simulation can be regarded as an estimate of every term in Eq. (3).

\[ \Psi(x) = \tilde{S}(x) + \int dx' \tilde{S}(x') \tilde{K}(x' \rightarrow x) + \int dx'' dx' \tilde{S}(x') \tilde{K}(x' \rightarrow x') \tilde{K}(x'' \rightarrow x) \]

Monte Carlo

0th step (initial condition)
\[ N_0(x) \quad \text{number of walkers at point } x. \]
\[ E[N_0(x)] = N_0^{\text{tot}} \tilde{S}(x) \]

1st step
\[ N_1(x) \]
\[ E[N_1(x)] = E[\int dx' N_0(x') \tilde{K}(x' \rightarrow x)] = \int dx' \tilde{S}(x') \tilde{K}(x' \rightarrow x) \]

2nd step
\[ N_2(x) \]
\[ E[N_2(x)] = E[\int dx' N_1(x') \tilde{K}(x' \rightarrow x)] = \int dx''/dx' \tilde{S}(x') \tilde{K}(x' \rightarrow x) \tilde{K}(x'' \rightarrow x) \]

\[ \vdots \]

\[ N_N(x) = N_0(x) + N_1(x) + N_2(x) + \cdots \]
\[ E[N_N(x)] = N_N^{\text{tot}} \tilde{S}(x) \]

Eq. (4) and Eq. (5) are enforced by the stochastic rule of the MC simulation.

For example, Rule #1 guarantees Eqs (4) and (5), as long as \( \tilde{K}(x) \leq 1 \).
But if $\hat{N}(x) > 1$, we need to apply another "rule".

Notice that the goal is to satisfy

$$E[N_1(x)] = E[\int dx' N_0(x') \tilde{R}(x' \rightarrow x)] \quad \text{(4,}\n
\uparrow \quad \text{number of} \quad \text{number of}
\downarrow \quad \text{walkers in step 1} \quad \text{walkers in step 0}
\quad \text{at point } x \quad \text{at point } x'

There are many ways to satisfy this condition.
A possible way is given below.

**Rule #2:**

If $\hat{N}(x') = \int dx \tilde{R}(x' \rightarrow x) > 1$

then for every walker at point $x'$ in the previous step, we turn them into $N^*(x')$ number of walkers.

Such that

$$E[N^*(x')] = \tilde{R}(x')$$

$\uparrow$ integer $\quad \uparrow$ any real number.

For example, suppose $\hat{N}(x') = \lfloor n(x') \rfloor + b$

with $\lfloor n(x') \rfloor$ the largest integer that is not greater than $n(x')$

then we can let

$$N^*(x') = \begin{cases} 
\lfloor n(x') \rfloor & \text{with probability } 1-b \\
\lfloor n(x') \rfloor + 1 & \text{with probability } b
\end{cases}$$

Each new walker then moves to the neighboring points $x$

with probability

$$\tilde{R}'(x' \rightarrow x) = \frac{1}{\hat{N}(x)} \tilde{R}(x' \rightarrow x)$$

7.
Notice that we compute contributions to \( G \) AFTER the new walkers have moved to the new sites.

If there are points \( x \) in \( \Omega \) such that \( \Psi(x) > 1 \), then it is possible for the total number of walkers to “temporarily” increase from one MC step to the next.

However, on the average, the total number of walkers should decrease and the simulation will finish when the last walker is terminated.

Otherwise, it means that the summation series in Eq. (3) does not converge.

4. What is the (global) optimal importance function? How to achieve zero variance in computing \( G = \int_\Omega \Psi(x) g(x) dx \).

Optimality is reached if every walker makes an identical contribution \( G_i \) to the total integral.

Note: This condition is of theoretical value only, because finding the optimal importance function is most likely to be even more difficult than computing \( G \) itself. But it can teach us what the optimal importance function should look like and how can we approximate it.
Let \( J(x) \) be the average score a random walker accumulates if it starts at point \( x \).

By this definition, we have

\[
G = \int 4(x) \, g(x) \, dx = \int s(x) \, J(x) \, dx
\]

\( J(x) \) satisfies another integral equation

\[
(6) \quad J(x) = g(x) + \int k(x \to y) \, J(y) \, dy
\]

Recall that

\[
(7) \quad 4(x) = s(x) + \int 4(y) \, k(y \to x) \, dy
\]

We can regard \( J(x) \) as the "dual" of \( 4(x) \).

Multiply \( 4(x) \) to Eq. (6) and \( J(x) \) to Eq. (7) and integrate

\[
\begin{align*}
\int 4(x) \, J(x) \, dx &= \int 4(x) \, g(x) \, dx + \int \int 4(x) \, k(x \to y) \, J(y) \, dy \, dx \\
\int 4(x) \, J(x) \, dx &= \int s(x) \, J(x) \, dx + \int \int 4(y) \, k(y \to x) \, J(x) \, dx \, dy
\end{align*}
\]

This is another proof that

\[
G = \int 4(x) \, g(x) \, dx = \int s(x) \, J(x) \, dx
\]

Formally, we may solve the integral equation \( J = g + k \cdot J \) as

\[
(8) \quad J = (I - K)^{-1} \cdot g = g + k \cdot g + k \cdot k \cdot g + k \cdot k \cdot k \cdot g + \ldots
\]

\[
(9) \quad G = \frac{4 \cdot g}{J} = \frac{s \cdot J}{4} = \frac{J}{s \cdot (I - K)^{-1} \cdot g}
\]
In the 2D example, we can compute $J(x)$ by Eq. (8)
then result is plotted in Fig. 1 (c).
From this result, $G_1 = \int S(x) J(x) dx = 4.6844 \times 10^{-2} = \int g(x) dx$.

Eq. (8) may be rewritten as

$$J(x) = g(x) + \int dx' \ k(x \rightarrow x') \ g(x')$$
$$+ \int dx' \int dx'' \ k(x \rightarrow x') \ k(x' \rightarrow x'') \ g(x'')$$
$$+ \int dx' \int dx'' \int dx''' \ k(x \rightarrow x') \ k(x' \rightarrow x'') \ k(x'' \rightarrow x''') \ g(x''')$$
$$+ \ldots$$

5. $J(x)$ is the (global) optimal importance function.

Why is that?

If we use $J(x)$ as importance function, then

$$\hat{k}(x \rightarrow y) = k(x \rightarrow y) \frac{J(y)}{J(x)}$$

the normalization factor becomes

$$\hat{n}(x) = \int \hat{k}(x \rightarrow y) \ dy = \int k(x \rightarrow y) \frac{J(y)}{J(x)} \ dy$$

$$\left( \text{but } \int k(x \rightarrow y) J(y) \ dy = J(x) - g(x), \ \text{from Eq. (6)} \right)$$

$$\hat{n}(x) = \frac{J(x) - g(x)}{J(x)} = 1 - \frac{g(x)}{J(x)} \leq 1$$

This means that there is no need to duplicate existing workers.
Termination probability
\[ \tilde{q}(x) = 1 - \hat{q}(x) = \frac{q(x)}{J(x)} \]

Recall that the gain of a walker at point \( x \) is
\[ \tilde{g}(x) = \frac{s_0 \cdot g(x)}{J(x)} \]

thus
\[ \tilde{q}(x) = \frac{1}{s_0} \tilde{g}(x) \]

The probability of terminating a walker is proportional to its gain — optimal condition.

To achieve zero variance, we may account for the gain of a random walker only when it is terminated. Because the probability of terminating a walker at point \( x \) is \( \tilde{q}(x) \), upon termination, the walker should contribute a gain of \( \frac{\tilde{g}(x)}{\tilde{q}(x)} = s_0 \).

Notice that \( s_0 = \int g(x) \cdot J(x) \, dx = G_i \).

In this case, every walker, upon termination, contributes a constant gain, which is equal to \( G_i \).

Hence, we have a zero-variance estimate of \( G_i \).

**Question:** When the importance function \( I(x) \) is not exactly the same as the optimal one, \( J(x) \),
is it better to accumulate gain \( \frac{\tilde{g}(x)}{\tilde{q}(x)} \) when a walker terminates, \( x \), or to accumulate gain \( \tilde{g}(x) \) whenever a walker visits point \( x \)?
Importance Sampling in Monte Carlo Simulation of Rare Transition Events*

Wei Cai

Lecture 3. August 3, 2005

Figure 1: (a) 2D potential landscape. (b) Contour plot of (a). Circle: $q(x) = 1$, $g(x) = 0$. Cross: $q(x) = 1$, $g(x) = 1$. (c) Exact solution of walker density $\psi(x)$. Average gain $G = \int \psi(x)g(x)dx = 2.7598 \times 10^{-12}$. (d) Exact solution of optimal importance function $J(x,y)$. $G = \int s(x)J(x)dx = 2.7598 \times 10^{-12}$. (pot2d_exact_sol.m)

1. random walk and rare event problem

In the last 2 lectures, we talked about a random walk problem with absorption and gain, specified by:

- $K(x \rightarrow y)$: transition probability matrix
- $S(x)$: source function
- $g(x)$: gain function

Every walker eventually gets terminated $\int K(x \rightarrow y) \, dy \leq 1$ during its lifetime, it visited point $x$ $\gamma(x)$ times on the average.

Average gain

$$G = \int \gamma(x) \, g(x) \, dx$$

How can we apply the importance sampling method developed here to the rare event problem in Monte Carlo Simulations?

![Diagram](image1)

A MC simulation with rare events has many similarities with the random walk problem above, but there are also some major differences.

1. The transition probability matrix is anisotropic and depends on energy landscape.

E.g. in the Metropolis algorithm

$$K_0(x \rightarrow y) = \frac{1}{N_n} \min\left[1, e^{-\frac{E(y) - E(x)}{kT}}\right] \quad \text{for } y \neq x$$

$\uparrow$ Number of nearest neighbors

$$K_0(x \rightarrow x) = 1 - \int K(x \rightarrow y) \, dy$$
In our example (Fig. 1) we use a different Ko matrix—there is no probability for staying, $K_{0}(x \rightarrow x) = 0$.

$$K_{0}(x \rightarrow y) = \frac{1}{\eta_{0}(x)} \min \left[ 1, e^{-\frac{E(y) - E(x)}{kT}} \right] \quad \text{for} \quad y \neq x$$

$$\eta_{0}(x) = \int \min \left[ 1, e^{-\frac{E(y) - E(x)}{kT}} \right] dy$$

motion to a lower energy point is more probable.

2. $\eta(x) = \int K_{0}(x \rightarrow y) dy = 1$

transition probability matrix is normalized at every $x$.
the random walker never “dies.” It goes back and forth between A and B.

We have a rare event problem when the walker stays in A for a very “long time” (i.e. many MC steps) before jumping into B.

The transition probability matrix satisfies

$$0 \leq K_{0,ij} < 1 \quad \sum_{j} K_{0,ij} = 1.$$

This means $K_{0}$ is a stochastic matrix $\rightarrow$ the largest eigenvalue of $K_{0}$ is 1.

The problem:

what is the rate of transition from A to B ($R_{AB}$) if the walker follows $K_{0}$?

in other words, if we put a walker in A, what is the average number ($N_{AB}$) of MC steps before we find it in B?

$$R_{AB} = \frac{1}{N_{AB}}$$

$N_{AB}$ can be extremely large. and it may not be possible to run a MC simulation for that many steps.
We will reformulate the rare event problem to a \textit{absorption/gain} random walk problem and then apply importance sampling.

We select a region $\Omega_A$ inside state $A$ such that the walker visits $\Omega_A$ very many times before it visits $\Omega_B$. In this region, let $S(x) > 0$, $q(x) > 0$, i.e., let the random walk have a finite probability to terminate in this region. After termination, a new walker starts in $\Omega_A$ according to $S(x)$.

In this way, we break a long trajectory of a random walker (following $K(x \rightarrow y)$) into many short trajectories of random walkers following

\[ K(x \rightarrow y) = (1 - q(x)) K_0(x \rightarrow y). \]

Select a region $\Omega_B$ inside state $B$, such that a walker frequently visits $\Omega_B$ once it enters $B$. Let $g(x) = 1$, $q(x) = 1$ in $\Omega_B$, i.e., once the walker enters $\Omega_B$, it is terminated and accumulates a gain.

The average gain $G = \int 4(x) q(x) dx$ is the probability of a walker, once started in $\Omega_A$, terminates in $\Omega_B$ instead of in $\Omega_A$.

Let us call this probability the "success probability" $P_s$, i.e.,

\[ G = P_s \]

4.
How is $p_s$ related to transition rate $Y_{AB}$?

Once a walker starts in $S_A$ (in the original problem, $K$) it visits $S_A$ many times before it visits $S_B$.

Let $\langle L_{AA} \rangle$ be the average length of a path that starts from $S_A$ and ends in $S_A$.

$\langle L_{AB} \rangle$ be the average length of a path that starts from $S_A$ and ends in $S_B$.

Then
\[
\frac{1}{Y_{AB}} = N_{AB} = \frac{\langle L_{AA} \rangle}{p_s} + \langle L_{AB} \rangle
\]

in the limit of $P_i \gg 0$, $p_s \ll 1$, \(\frac{\langle L_{AA} \rangle}{p_s} \approx \langle L_{AB} \rangle\),

\[
Y_{AB} = \frac{p_s}{\langle L_{AA} \rangle}
\]

$\langle L_{AA} \rangle$ can be easily computed by Monte Carlo simulation (no importance sampling is necessary).

The difficulty now lies in computing $p_s = G_1$ for which importance sampling can help.

The procedure of importance sampling is the same as described in earlier lectures. An example is given in Fig. 1.
2. Path distribution in importance sampled MC simulation.

For simplicity, let \( S_A = \{ X_A \} \), i.e. \( S_A, X_0 \) only contain one point.

Let \( q(X_A) = 1 \) \( g(X_A) = 0 \)
\( q(X_B) = 1 \) \( g(X_B) = 1 \)

\( S(X) = K_0( X_A \rightarrow X) \)

i.e. walkers always start from neighbors of \( X_A \) and terminates at either \( X_A \) or \( X_B \).

What is the probability of sampling a path \( (X, X_1, \ldots, X_n, X_B) \)

\[
P(X, X_1, \ldots, X_n, X_B) = S(X) K(X_1 \rightarrow X) K(X_2 \rightarrow X_1) \cdots K(X_n \rightarrow X_{n-1}) K(X_{n+1} \rightarrow X_B)
\]

\[
= K_0(X_A \rightarrow X_1) K_0(X_1 \rightarrow X_2) \cdots K_0(X_{n-1} \rightarrow X_B)
\]

\[
= p_0(X_A, X_1, X_2, \ldots, X_{n+1}, X_B) \quad (= p_o \text{ in short})
\]

It is the same as sampling a path \( (X_A, X_1, X_2, \ldots, X_B) \) in the original simulation (as specified by \( K_0 \)).

Suppose now we introduce an importance function \( I(X) \)

\[
K(X \rightarrow Y) = K(X \rightarrow Y) \frac{I(Y)}{I(X)}
\]

Question:
what is the probability of sampling the same path \( (X_1, X_2, \ldots, X_{n+1}, X_B) \) in the importance sampled MC simulation?
The "probability" is
\[ p(X_1, X_2, \ldots X_8) = \frac{S(X_1) \tilde{K}(X_1 \rightarrow X_2) \cdots \tilde{K}(X_{n-1} \rightarrow X_8)}{S_0} \]
\[ = \frac{I(X_1) S(X_1)}{S_0} \tilde{K}(X_1 \rightarrow X_2) \cdots \tilde{K}(X_{n-1} \rightarrow X_8) \]
\[ = \frac{I(X_8)}{S_0} p_0(X_1, X_2, X_3, \ldots X_8) \]

where \[ S_0 = \int I(x) S(x) \, dx = I(X_1) S(X_1) \quad \text{in this example.} \]

Thus the "probability" of sampling all paths going from \( X_1 \) to \( X_8 \)
are charged by a uniform amount \[ \frac{I(X_8)}{I(X_1) S(X_1)} \to 1 \]

Here I put "probability" under quotation mark, because what we have here is not a true probability. Instead, it is the average number of times this specific path is sampled when we run the simulation once.

How can these two concepts be different from each other? When no importance function is used.

Let \( N \) be the number of times this path is sampled in one MC Simulation.

\[ N = \begin{cases} 0 & \text{with probability } p_0 \\ 1 & \text{with probability } 1 - p_0 \end{cases} \]

\[ \therefore \langle N \rangle = p_0 \]

Average number of path samples and the probability is the same thing.
When an importance function is used, if any point along the path \((X_1, X_2, \ldots, X_{n-1}, X_n)\) (not including \(X_n\)) has normalization factor
\[
\tilde{N}(X_i) = \int \tilde{K}(x_i, y) \, dy > 1
\]
then for every MC simulation, it is possible to generate more than one paths.

Even if we started with one random walker, in principle it is possible that we obtain the same path \((X_1, X_2, \ldots, X_{n-1}, X_n)\) \(2, 3, 4, \ldots\) times.

i.e. \(N = 0, 1, 2, 3, \ldots\) are all possible

\[
\langle N \rangle = \frac{I(X_0)}{S_0} \cdot p \approx 1
\]

Our goal is to let
\[
\langle N \rangle = \frac{I(X_0)}{S_0} \cdot p \approx 1
\]

but
\[
\text{Prob} \{ N = 1 \} < \langle N \rangle
\]
\[
\text{var} \{ N \} > \langle N \rangle
\]

Thus allowing \(\tilde{N}(X) > 1\) (i.e. walker shall multiply) increases variance.

Notice that when the optimal importance function is used, \(\tilde{N}(X) \leq 1\) everywhere.

In this case, \(\langle N \rangle = \frac{I(X_0)}{S_0} \cdot p\)

because \(N\) can only be 0 or 1.
3. How to improve importance function during MC Simulation?

a. First, we propose a functional form for $I(x)$ based on intuition and leave some adjustable free parameters.


Intuition: in the 2D example, we find that the optimal importance function $I(x)$ corresponds to a bias potential $\Delta E(x) = -2kT \ln I(x)$ which, when added to the original energy landscape, inverts the basin into a hill along the reaction coordinate.

In the direction perpendicular to the reaction coordinate, $\Delta E$ stays roughly constant.

b. The simplest way to optimize $I(x)$ is to do it by hand. Not necessarily efficient.

For a set of parameters $\{ a_i \}$, compute $G$ and $\text{var}[G]$ choose the parameter set $\{ a_i \}$ that gives the lowest variance $\text{var}[G]$ and repeat this process.

It would be useful if we can compute $2[\text{var}[G]]^{-1}$ and use iterative minimization algorithms.

i.e. we can define $\bar{Q} = \text{var}[G]$ as a quality factor for the parameter set $\{ a_i \}$ and try to minimize $\bar{Q}$. 

q.
c. We may also use other quality forms.

Notice that for the optimal importance function
\[ \tilde{h}(x_i) = 1 \quad \text{(in this example)} \]
for all \( x_i \).

We may construct another quality factor
\[ Q_c = \frac{1}{N} \sum_{x_i} \left[ \ln \tilde{h}(x_i) \right]^2 \]
where \( x_i \) are the locations the walkers visited.

\( Q_c = 0 \) for optimal importance function and
\( Q_c > 0 \) for any other importance function.

It is not difficult to compute
\[ \frac{2 \ln \tilde{h}(x_j)}{\partial x_i} \quad \text{for every } x_j \]
and hence \( \frac{\partial Q_c}{\partial x_i} \).

With this information, we may minimize \( Q_c \) by
steepest descent iterations.