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:

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

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

Average gain:

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

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

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:

   $$K_{0}(x \to y) = \frac{1}{N_0} \min \left[ 1, e^{-\frac{E(y) - E(x)}{kT}} \right] \quad \text{for} \ y \neq x$$

   $N_0$: number of nearest neighbors

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

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

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

Motion to a lower energy point is more probable.

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

The transition 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

\[ 0 \leq K_{0} \leq 1, \quad \sum_{j} K_{0,ij} = 1. \text{ this means } \]

\( K_{0} \) is a stochastic matrix → the largest eigenvalue of \( K_{0} \) is 1.

The problem:

what is the rate of transition from \( A \) to \( B \) (\( V_{AB} \)) 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 \( B \)?

\[
V_{AB} = \frac{1}{N_{AB}}
\]

\( N_{AB} \) 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
absorption/gain random walk problem and then
apply importance sampling.

1. 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

$$k(x\rightarrow y) = (1 - q(x)) k_0(x\rightarrow y)$$

2. Select a region $\Omega_B$ inside state $B$, such that
a walker frequently visit $\Omega_B$ 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 $G = \int g(x) q(x) dx$ 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 $r_{AB}$?

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 $S_A$.

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

Then

$$ \frac{1}{r_{AB}} = N_{AB} = \frac{\langle L_{AA} \rangle}{p_s} + \langle L_{AB} \rangle $$

in the limit of $p_s \ll 1$, $p_s \ll 1$, $\frac{\langle L_{AA} \rangle}{p_s} \to \langle L_{AB} \rangle$.

$$ r_{AB} = \frac{p_s}{\langle L_{AA} \rangle} $$

$\langle L_{AA} \rangle$ 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.
Path distribution in importance sampled MC simulation.

For simplicity, let $\mathcal{S}_A = \{X_A\}$, i.e. $\mathcal{S}_A, X_0$ only contain one point.

Let $\mathcal{S}(X_A) = 1$ $\mathcal{S}(X_B) = 0$

$\mathcal{S}(X_B) = 1$ $\mathcal{S}(X_B) = 1$

$S(X) = K_{0}(X_A \rightarrow X)$

i.e. walkers always start from neighbors of $X_A$ and terminate at either $X_A$ or $X_B$.

What is the probability of sampling a path $(X, X_1, \ldots, X_{n-1}, X_B)$?

$p(X, X_1, \ldots, X_{n-1}, X_B) = S(X)K(X_1 \rightarrow X_2)K(X_2 \rightarrow X_3)\cdots K(X_{n-1} \rightarrow X_B)$

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

$= p_0(X_A, X_1, X_2, \ldots, X_{n-1}, X_B) (= p_0 \text{ in short})$

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
\[
\hat{P}(X_1, X_2, \ldots X_8) = \frac{S(X_1) \tilde{K}(X_1 \rightarrow X_2) \ldots \tilde{K}(X_{n-1} \rightarrow X_n)}{S_0}
\]
\[
= \frac{I(X_i) S(X_i)}{S_0} \cdot \frac{\tilde{K}(X_i \rightarrow X_j)}{I(X_i)} \ldots \frac{\tilde{K}(X_{n-1} \rightarrow X_n)}{I(X_{n-1})} \frac{I(X_n)}{I(X_n)}
\]
\[
= \frac{I(X_8)}{S_0} \cdot P_0(X_A, X_i, X_j, \ldots X_8)
\]

where\[
S_0 = \int I(X) S(X) \, dX = I(X_A) S(X_A)
\]
in this example.

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

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.

\[
N = \begin{cases} 1 & \text{with probability } P_0 \\ 0 & \text{with probability } 1 - P_0 \end{cases}
\]

\[
\langle N \rangle = P_0
\]

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 \(X_0\)) has normalization factor
\[
\tilde{W}(X_i) = \int \tilde{R}(x_i - y) dy > 1
\]
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 \((X_1, X_2, \ldots, X_{n-1}, X_n)\)
\[2, 3, 4, \ldots \text{ times.}\]
i.e. \(N = 0, 1, 2, 3, \ldots\) are all possible

\[
\langle N \rangle = \frac{I(X_n)}{S_0} \cdot P_0 \approx 1
\]

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

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

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

In this case, \(\langle N \rangle = \frac{I(X_n)}{S_0} \cdot P_0\)

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.

Several sets 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 $\frac{\partial^2 \text{var}[G]}{\partial a_i}$ and use iterative minimization algorithms.

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

9.
c. We may also use other quality forms.

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

We may construct another quality factor
\[ Q_c = \frac{1}{N} \sum_{x_i} \left[ \ln \hat{w}(x_i) \right]^2 \]
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
\[ \frac{2 \ln \hat{w}(x_j)}{\partial x_1} \quad \text{for every } x_j \]
and hence \[ \frac{\partial Q_c}{\partial x_1} \]

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