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, \(\psi(x) = s(x) + \int \psi(y)K(y \rightarrow x)dy\). \(G = \int \psi(x)g(x)dx = 4.6844 \times 10^{-2}\). (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, \(J(x) = g(x) + \int K(x \rightarrow y)J(y)dy\). \(G = \int J(x)s(x)dx = 4.6844 \times 10^{-2}\). (diffuse2d_opt_imp.m)

This lecture will be more theoretical than the first one.

### 1. Review

<table>
<thead>
<tr>
<th>Original Random Walk</th>
<th>Importance Sampled Random Walk</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source function</strong> $S(x)$</td>
<td>$\tilde{S}(x) = S(x) \frac{I(x)}{S_0}$</td>
</tr>
<tr>
<td>$\int S(x) , dx = 1$</td>
<td>$S_0 = \int S(x) , I(x) , dx$</td>
</tr>
<tr>
<td><strong>Gain function</strong> $g(x)$</td>
<td>$\tilde{g}(x) = \frac{S_0 , g(x)}{I(x)}$</td>
</tr>
<tr>
<td><strong>Transition probability</strong> $k(x \rightarrow y)$</td>
<td>$\tilde{k}(x \rightarrow y) = k(x \rightarrow y) \frac{I(y)}{I(x)}$</td>
</tr>
<tr>
<td>$\int k(x \rightarrow y) , dy \leq 1$</td>
<td>$\int \tilde{k}(x \rightarrow y) , dy = 1$</td>
</tr>
<tr>
<td>$1 - \int k(x \rightarrow y) , dy = g(x)$</td>
<td></td>
</tr>
<tr>
<td><strong>Integral equation</strong> $\psi(x) = S(x) + \int k(y \rightarrow x) \psi(y) , dy$</td>
<td>$\tilde{\psi}(x) = \tilde{S}(x) + \int \tilde{k}(y \rightarrow x) \tilde{\psi}(y) , dy$</td>
</tr>
<tr>
<td><strong>Gain</strong> $G = \int \psi(x) , g(x) , dx$</td>
<td>$\tilde{G} = \int \tilde{\psi}(x) , \tilde{g}(x) , dx$</td>
</tr>
</tbody>
</table>

### 2. The procedure of importance sampled Monte Carlo.

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

- $\tilde{\tilde{n}}(x) = 1 - \tilde{n}(x)$ termination probability
- $\tilde{k}'(x \rightarrow y) = \frac{1}{\tilde{n}(x)} \tilde{k}(x \rightarrow y)$ is a normalized probability

$\int \tilde{k}'(x \rightarrow y) \, dy = 1$.  

2
Rule #1:

1. Compute $\hat{\pi}(x)$, $\tilde{q}(x) = 1 - \hat{\pi}(x)$
2. Sample a random number $z$ uniformly distributed in $[0,1]$
3. If $z < \tilde{q}(x)$,
   - Yes: terminate walker at point $x$
4. Choose one of the neighboring point $y$ according to probability
   \[ \tilde{\kappa}(x \to y) = \frac{\hat{\pi}(x)}{\tilde{q}(x)} \tilde{\kappa}(x \to y) \]
   - $x := y$

But what if $\tilde{q}(x) > 1$?

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

What shall we do in this case?

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

Let $N_0$ be the total number of random walkers at the beginning of the simulation.
Let $N_c(x)$ be the number of random walkers at step $i$ of the simulation located at point $x$.
Let $N_v(x) = \sum_{i=1}^{\infty} N_i(x)$ 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,

\[ E \left[ N_0(x) \right] = N_0^{\text{tot}} \tilde{S}(x) \quad \text{........ (1)} \]

\[ E \left[ N_i(x) \right] = N_0^{\text{tot}} \tilde{\Psi}(x) \quad \text{........ (2)} \]

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 $\tilde{\Psi}(x)$ satisfy the integral equation

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

One possible way to solve this equation is by iteration:

\[ \tilde{\Psi}^{(0)}(x) = \tilde{S}(x) \]
\[ \tilde{\Psi}^{(1)}(x) = \tilde{S}(x) + \int \tilde{\Psi}^{(0)}(y) \tilde{K}(y-x) dy \]
\[ \tilde{\Psi}^{(2)}(x) = \tilde{S}(x) + \int \tilde{\Psi}^{(1)}(y) \tilde{K}(y-x) dy \]
\[ \vdots \]
\[ \tilde{\Psi}^{(n)}(x) = \tilde{S}(x) + \int \tilde{\Psi}^{(n-1)}(y) \tilde{K}(y-x) dy \]

If the iteration converges, i.e.

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

then \[ \lim_{n \to \infty} \tilde{\Psi}^{(n)}(x) = \tilde{\Psi}(x) \] is the solution.
\[ \tilde{Y}(x) = \tilde{S}(x) + \int dx' \tilde{S}(x') \tilde{K}(x' - x) \]
\[ + \int dx'' dx' \tilde{S}(x'') \tilde{K}(x'' - x') \tilde{K}(x' - x) \]
\[ + \int dx''' dx'' dx' \tilde{S}(x''') \tilde{K}(x'' - x'') \tilde{K}(x'' - x') \tilde{K}(x - x') \]
\[ + \cdots \]  

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

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

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

In matrix form, the equation \( \tilde{Y} = \tilde{S} + \tilde{Y} \cdot \tilde{K} \) can be formally solved as
\[ \tilde{Y} \cdot (I - \tilde{K}) = \tilde{S} \] where \( I \) is the identity matrix
\[ \tilde{Y} = \tilde{S} \cdot (I - \tilde{K})^{-1} = \tilde{S} \cdot (1 + \tilde{K} + \tilde{K} \cdot \tilde{K} + \tilde{K} \cdot \tilde{K} \cdot \tilde{K} + \cdots) \]
Every step of the Monte Carlo simulation can be regarded as an estimate of every term in Eq. (3).

Monte Carlo

0th step (initial condition)

\[ N_0(x) \] number of walkers at point \( x \).

\[ E[N_0(x)] = N_0^{\text{tot}} \tilde{S}(x) \]

1st step.

\[ N_1(x) \]

\[ E[N_1(x)] = E[\int dx' N_0(x) \tilde{K}(x' \to x)] \]
\[ = \int dx' \tilde{S}(x') \tilde{K}(x' \to x) \]

2nd step

\[ N_2(x) \]

\[ E[N_2(x)] = E[\int dx' N_1(x) \tilde{K}(x' \to x)] \]
\[ = \int dx'' \int dx' \tilde{S}(x') \tilde{K}(x' \to x) \tilde{K}(x'' \to x) \tilde{K}(x'' \to x) \]

\[ \vdots \]

\[ N_N(x) = N_0(x) + N_1(x) + N_2(x) + \ldots \]

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

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

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 $\tilde{R}(x) > 1$, we need to apply another "rule".

Notice that the goal is to satisfy

$$E[N(x)] = E[\int dx' N(x', \tilde{R}(x' \rightarrow x))] \quad \text{(4.1)}$$

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

Rule #2.

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

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

Such that

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

integer \quad \text{any real number.}$$

For example, suppose $\tilde{R}(x) = \lceil n(x') \rceil + b$

then we can let

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

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

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

7.
Notice that we compute contributions to $G$
AFTER the new walkers has move to the new sites.

If there are points $x$ in $\Omega$ such that $P(x) > 1$,
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 $G = \int_{\Omega} P(x) g(x) dx$.

Optimality is reached if every walker makes an identical contribution ($G$) 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

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

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

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

Recall that

\[
(7) \quad 4(x) = S(x) + \int 4(y) \cdot K(y \to x) \, dy
\]

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

\[
\int 4(x) J(x) \, dx = \int [4(x) \cdot g(x)] \, dx + \int \int 4(x) K(x \to y) \cdot J(y) \, dy \, dx
\]

\[
\int 4(x) J(x) \, dx = \int S(x) \cdot J(x) \, dx + \int \int 4(y) K(y \to x) \, J(x) \, dx \, dy
\]

This is another proof that

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

\[
G = 4 \cdot g = S \cdot J
\]

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

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

\[
(9) \quad G = 4 \cdot g = S \cdot J = S \cdot \left( I - K \right)^{-1} \cdot g
\]
In the 2D example, we can compute \( J(x) \) by Eq. (8) then result is plotted in Fig. 1(c).

From this result, \( G = \int S(x) J(x) dx = 4.6844 x^2 = \int h(x) g(x) dx \).

Eq. (8) may be rewritten as

\[
J(x) = g(x) + \int dx' k(x \to x') g(x') + \int dx' \int dx'' k(x \to x') k(x' \to x'') g(x'') + \int dx' \int dx'' \int dx''' k(x \to x') k(x' \to x'') k(x'' \to x''') g(x''') + \ldots.
\]

(10)

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

Why is that?

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

\[
\hat{K}(x \to y) = k(x \to y) \frac{J(y)}{J(x)}
\]

the normalization factor becomes

\[
\hat{N}(x) = \int \hat{K}(x \to y) dy = \int k(x \to y) \frac{J(y)}{J(x)} dy
\]

(but \( \int k(x \to y) J(y) dy = J(x) - g(x) \), from Eq. (6))

\[
\hat{N}(x) = \frac{J(x) - g(x)}{J(x)} = 1 - \frac{g(x)}{J(x)} \leq 1
\]

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

Recall that the gain of a walker at point \( x \) is
\[ \tilde{g}(x) = S_0 \frac{g(x)}{J(x)} \]

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 \( S_0 = \int g(x) \cdot J(x) \, dx = G_1 \).

In this case, every walker, upon termination, contribute a constant gain, which is equal to \( G_1 \).

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

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 \( \tilde{q}(x) \) or to accumulate gain \( \tilde{g}(x) \) whenever a walker visits point \( x \) ?