Reduction of Variance

As we discussed earlier, the statistical error goes as:

\[ \text{error} = \sqrt{\text{variance/computer time}}. \]

**Define:**

\[ \text{Efficiency} = \zeta = \frac{1}{vT} \]
\[ v = \text{error}^2 \text{ of mean} \quad \text{and} \quad T = \text{total CPU time} \]

How can you make simulation more efficient?

- Write a faster code,
- Get a faster computer
- Work on reducing the variance.
- Or all three

We will talk about the third option: *Importance sampling* and *correlated sampling*
Importance Sampling

Given the integral: \[ I = \int dxf(x) = \left\langle f(x) \right\rangle \]

How should we sample \( x \) to maximize the efficiency?

Transform the integral: \[ I = \int dx \, p(x) \left[ \frac{f(x)}{p(x)} \right] = \left\langle \frac{f(x)}{p(x)} \right\rangle_p \]

variance is: \[ \nu \approx \sigma^2 = \left\langle \left[ \frac{f(x)}{p(x)} - I \right]^2 \right\rangle_p = \int dx p(x) \left[ \frac{f(x)}{p(x)} \right]^2 - I^2 \]

Optimal sampling: \[ \frac{\delta \nu}{\delta p(x)} = 0 \quad \text{with} \quad \int dx \, p(x) = 1 \]

Mean value of estimator \( I \) is independent of \( p(x) \), but variance \( \nu \) is not!
Assume CPU-time/sample is independent of \( p(x) \), and vary \( p(x) \) to minimize \( \nu \).
Trick to parameterize as a positive definite PDF:

\[
p(x) = \frac{[q(x)]^2}{\int dx [q(x)]^2}
\]

Solution: 
\[
p^*(x) = \frac{|f(x)|}{\int dx |f(x)|}
\]

Estimator: 
\[
\frac{f(x)}{p^*(x)} = \frac{\text{sign}(f(x))}{\int dx |f(x)|}
\]

1. If \(f(x)\) is entirely positive or negative, estimator is constant. “zero variance principle.”
2. We can’t generally sample \(p^*(x)\), because, if we could, then we would have solved problem analytically! But the form of \(p^*(x)\) is a guide to lowering the variance.
3. Importance sampling is a general technique: it works in any dimension.
Example of importance sampling

Suppose $f(x)$ was given by

$$f(x) = \frac{e^{-x^2/2}}{1 + x^2}$$

Value is independent of $a$.

Optimize a Gaussian

$$p(x) = \frac{e^{-x^2/2a}}{(2\pi a)^{1/2}}$$

CPU time is not
What does infinite variance look like?

Spikes

Long tails on the distributions

Near optimal

Why (visually)?

2/20/2013  Atom Scale Simulation
Example of importance sampling

Suppose $f(x)$ was given by

\[ f(x) = \frac{e^{-x^2/2}}{1 + x^2} \]

Value is independent of $a$.

Optimize a Gaussian

\[ p(x) = \frac{e^{-x^2/2a}}{(2\pi a)^{1/2}} \]

CPU time is not
Importance sampling functions

\[ v \approx \sigma^2 = \int dx \ p(x) \left[ \frac{f(x)}{p(x)} - I \right]^2 = \int dx \left[ \frac{e^{-x^2(1-1/2a)}}{(1+x^2)^2} \right] - I^2 \]
What does infinite variance look like?

Spikes

Long tails on the distributions

Near optimal
Why (visually)?
General Approach to Importance Sampling

- Basic idea of importance sampling is to sample more in regions where function is large.
- Find a convenient approximation to $|f(x)|$.

- **Do not under-sample** -- could cause infinite variance.
- **Over-sampling** -- loss in efficiency but not infinite variance.
- **Always derive conditions for finite variance analytically.**
- To debug: *test that estimated value is independent of important sampling.*

- **Sign problem**: zero variance is not possible for oscillatory integral.
  
  "Monte Carlo can add but not subtract."
Correlated Sampling

Suppose we want to compute a function: \( G(F_1, F_2) \)
where F’s are integrals:
\[
F_k = \int dx \, f_k(x)
\]

Suppose we use same \( p(x) \) and same random numbers to do both integrals.

What is optimal \( p(x) \)?
\[
p^*(x) \propto \left| f_1(x) \frac{dG}{dF_1} + f_2(x) \frac{dG}{dF_2} \right|
\]

It is a weighted average of the distributions for \( F_1 \) and \( F_2 \).
Sampling Boltzmann distribution

• Suppose we want to calculate a whole set of averages:

\[ \langle O_k \rangle = \frac{\int dR O_k(R)e^{-V(R)/kT}}{\int dR e^{-V(R)/kT}} \]

• Optimal sampling is:

\[ p_k^*(x) \propto e^{-V(R)/kT}(O_k(R) - \langle O_k \rangle) \]

• We need to sample the first factor because we want lots of properties

• Avoid undersampling.

• The Boltzmann distribution is very highly peaked.
Correlated Sampling

Suppose we want to compute a function: \( G(F_1, F_2) \)
where F’s are integrals:
\[
F_k = \int dx \ f_k(x)
\]

Suppose we use same \( p(x) \) and same random numbers to do both integrals.

What is optimal \( p(x) \)?
\[
p^*(x) \propto \left| f_1(x) \frac{dG}{dF_1} + f_2(x) \frac{dG}{dF_2} \right|
\]

It is a weighted average of the distributions for \( F_1 \) and \( F_2 \).
Sampling Boltzmann distribution

• Suppose we want to calculate a whole set of averages:

\[ \langle O_k \rangle = \frac{\int dR O_k(R)e^{-V(R)/kT}}{\int dR e^{-V(R)/kT}} \]

• Optimal sampling is:

\[ p_k^*(x) \propto e^{-V(R)/kT}(O_k(R) - \langle O_k \rangle) \]

• We need to sample the first factor because we want lots of properties

• **Avoid undersampling.**

• The Boltzmann distribution is *very highly peaked.*
Bias

- Bias is a *systematic error* caused by using a random number in another expression.
- You will get a result systematically too high or low.
- Suppose $Z' = \bar{Z} + \delta Z$ is the result of MC sampling.
- But we want $F(Z)$. *Example:* $F = -kT \ln(Z)$.
- What is the statistical error and bias of $F(Z')$?
- Expand $Z$ in power series about $<Z>$

$$F(Z') = F(\bar{Z}) + \frac{dF}{dZ} \bigg|_{\bar{Z}} \delta Z + \frac{1}{2} \frac{d^2 F}{dZ^2} \bigg|_{\bar{Z}} \delta Z^2 + L$$

$$\text{bias}(F) = <F(Z) - F(\bar{Z})> = \frac{1}{2} \frac{d^2 F}{dZ^2} \bigg|_{\bar{Z}} <\delta Z^2> + L = \frac{1}{2} \frac{d^2 F}{dZ^2} \bigg|_{\bar{Z}} \text{err}(Z)^2$$

$$\text{error}(F) = [<(F(Z')) - \langle F(Z) \rangle)^2>]^{1/2} = \left| \frac{dF}{dZ} \right|_{\bar{Z}} <\delta Z^2>^{1/2} + L = \left| \frac{dF}{dZ} \right|_{\bar{Z}} \text{err}(Z)$$

You may need to correct for the bias unless $N$ is very large.

You may need to correct for the bias unless $N$ is very large.
Independent Sampling for $\exp(-V/kT)$?

- Try hit or miss MC to get $Z = \exp(-V/kT)$.
- Sample $R$ uniformly in $(0,L)$: $P(R) = \Omega^{-N=1}$

What is the variance of the free energy and how does it depend on the number of particles?

$$Var(Z) = \int dR \left[ e^{-\beta V(R)} - Z(\beta) \right]^2 = Z(2\beta) - Z(\beta)^2$$

$$\text{var}(\beta F) = \frac{Var(Z)}{Z^2} = \frac{Z(2\beta)}{Z(\beta)^2} - 1 = e^{2\beta[F(\beta) - F(2\beta)]} - 1$$

$$\text{error}(\beta F) = e^{\beta[F(\beta) - F(2\beta)]}$$

- Blows up exponentially fast at large $N$ since $F$ is extensive!
- The number of sample points must grow exponentially in $N$, just like a grid based method.
Intuitive explanation

- Throw \( N \) points in a box, area \( A \).
- Say \textit{probability of no overlap is} \( q \).

- Throw \( 2N \) points in a box, area \( 2A \).
- Probability of no overlap is \( q^2 \).

- Throw \( mN \) points in a box, area \( mA \).
- Probability of no overlap is \( q^m \).

\textit{Probability of getting a success is} \( p = \exp(m \ln(q)) \). Success defined as a reasonable sample of a configuration.

This is a general argument. We need to sample only near the peak of the distribution: to do so use \textit{random walks}.
Random Walks
A&T 110-123 and F&S 3.1.2

- As we explained last time, it is very difficult to sample directly a general probability distribution.
  - If we sample from another distribution, the overlap will be order \( \exp(-aN) \), where \( N \) is the number of variables. The error bars will get exponentially larger as \( N \) increases.

- Today we will discuss Markov chains (random walks), detailed balance and transition rules.
  - These methods were introduced by Metropolis et al. in 1953 who applied it to a hard sphere liquid.
  - It is one of the most powerful and used algorithms.
Equation of State Calculations by Fast Computing Machines

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, and Augusta H. Teller,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico

AND

Edward Teller,∗ Department of Physics, University of Chicago, Chicago, Illinois
(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of
state for substances consisting of interacting individual molecules is described. The method consists of a
modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere
system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared
to the free volume equation of state and to a four-term virial coefficient expansion.

I. INTRODUCTION

The purpose of this paper is to describe a general
method, suitable for fast electronic computing
machines, of calculating the properties of any substance
which may be considered as composed of interacting
individual molecules. Classical statistics is assumed,
only two-body forces are considered, and the potential
field of a molecule is assumed spherically symmetric.
These are the usual assumptions made in theories of
liquids. Subject to the above assumptions, the method
is not restricted to any range of temperature or density.
This paper will also present results of a preliminary two-
dimensional calculation for the rigid-sphere system.
Work on the two-dimensional case with a Lennard-
Jones potential is in progress and will be reported in a
later paper. Also, the problem in three dimensions is
being investigated.

∗ Now at the Radiation Laboratory of the University of Cali-
ifornia, Livermore, California.

II. THE GENERAL METHOD FOR AN ARBITRARY
POTENTIAL BETWEEN THE PARTICLES

In order to reduce the problem to a feasible size for
numerical work, we can, of course, consider only a finite
number of particles. This number N may be as high as
several hundred. Our system consists of a square containing N particles. In order to minimize the surface
effects we suppose the complete substance to be periodic,
consisting of many such squares, each square contain-
ing N particles in the same configuration. Thus we
define d_{AB}, the minimum distance between particles A
and B, as the shortest distance between A and any of
the particles B, of which there is one in each of the
squares which comprise the complete substance. If we
have a potential which falls off rapidly with distance,
there will be at most one of the distances AB which
can make a substantial contribution; hence we need
consider only the minimum distance d_{AB}.

† We will use the two-dimensional nomenclature here since it
is easier to visualize. The extension to three dimensions is obvious.

2/22/2013 18
Markov chain or Random Walk

• Markov chain is a random walk through phase space:
  \[ s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_4 \rightarrow \ldots \]
  Here “s” is the state of the system.

• The *transition probability* is \( P(s_n \rightarrow s_{n+1}) \) a *stochastic matrix*
  \[
  P(s \rightarrow s') \geq 0 \quad \sum_{s'} P(s \rightarrow s') = 1
  \]

• In a Markov chain, the distribution of \( s_{n+1} \) depends only on \( s_n \) (by definition). *A drunkard has no memory!*

• Let \( f_n(s) \) be the probability after “n” steps. It evolves according to a “master equation.”
  \[
  f_{n+1}(s') = \sum_s f_n(s) P(s \rightarrow s') \quad \text{or} \quad f_{n+1} = \mathbf{P} f_n
  \]

• The stationary states are eigenfunctions of \( \mathbf{P} \): \( \mathbf{P} \phi = \varepsilon \phi \)