CE 530 Molecular Simulation

Lecture 7
Monte Carlo Integration and Importance Sampling

David A. Kofke
Department of Chemical Engineering
SUNY Buffalo
kofke@eng.buffalo.edu

Monte Carlo Simulation

- Gives properties via ensemble averaging
  - *No time integration*
  - *Cannot measure dynamical properties*
- Employs stochastic methods to generate a (large) sample of members of an ensemble
  - *"random numbers" guide the selection of new samples*
- Permits great flexibility
  - *members of ensemble can be generated according to any convenient probability distribution…*
  - *…and any given probability distribution can be sampled in many ways*
  - *strategies developed to optimize quality of results*

    ergodicity — better sampling of all relevant regions of configuration space

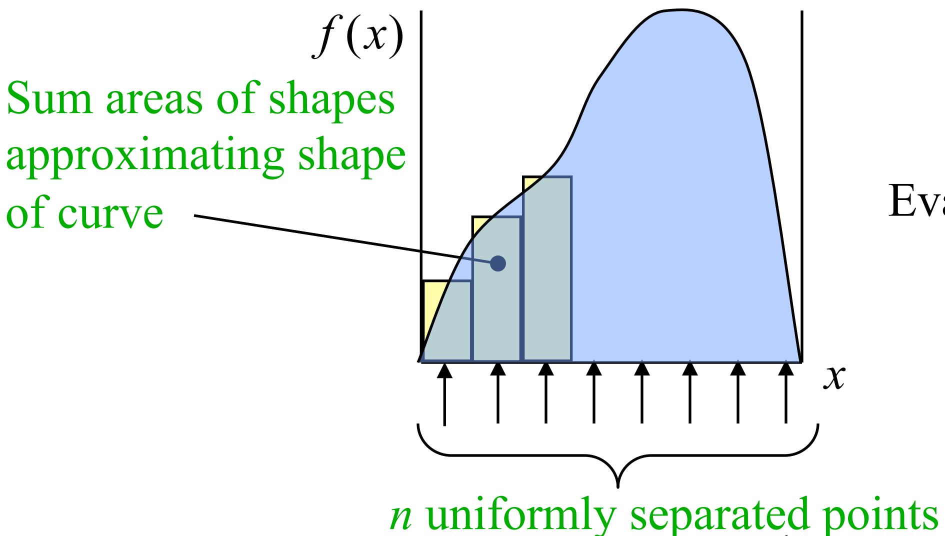
    variance minimization — better precision of results
- MC "simulation" is the evaluation of statistical-mechanics integrals

$$\langle U \rangle = \frac{1}{Z_N}\frac{1}{N!}\int dr^N U(r^N) e^{-\beta U(r^N)} \qquad Q = \frac{1}{h^{3N}N!}\int dp^N dr^N e^{-\beta E}$$

still too hard!

One-Dimensional Integrals

- Methodical approaches
  - *rectangle rule, trapezoid rule, Simpson's rule*



Evaluating the general integral $I = \int_a^b f(x)dx$

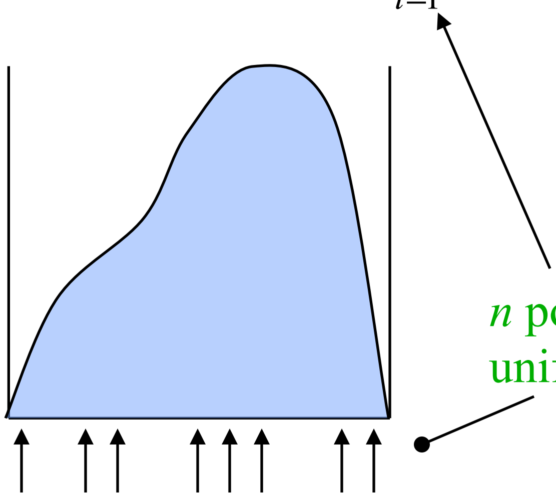
- Quadrature formula

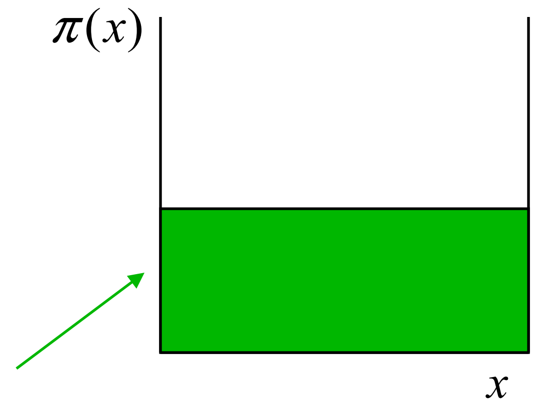
$$I \approx \Delta x \sum_{i=1}^{n} f(x_i) = \frac{b-a}{n} \sum_{i=1}^{n} f(x_i)$$

Monte Carlo Integration

- Stochastic approach
- Same quadrature formula, different selection of points

$$I \approx \frac{b-a}{n}\sum_{i=1}^{n} f(x_i)$$





n points selected from uniform distribution p(x)

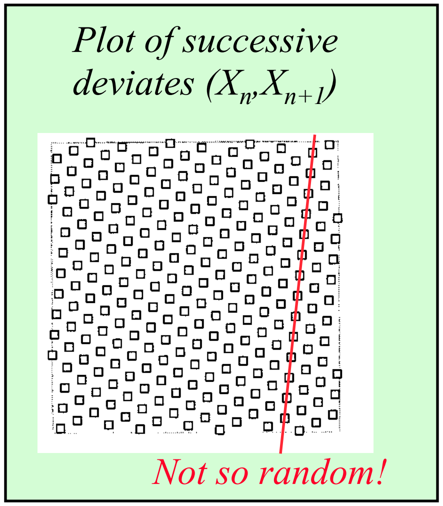
- Click here for an applet demonstrating MC integration

Random Number Generation

- Random number generators
  - *subroutines that provide a new random deviate with each call*
  - *basic generators give value on (0,1) with uniform probability*
  - *uses a deterministic algorithm (of course)*
    - usually involves multiplication and truncation of leading bits of a number

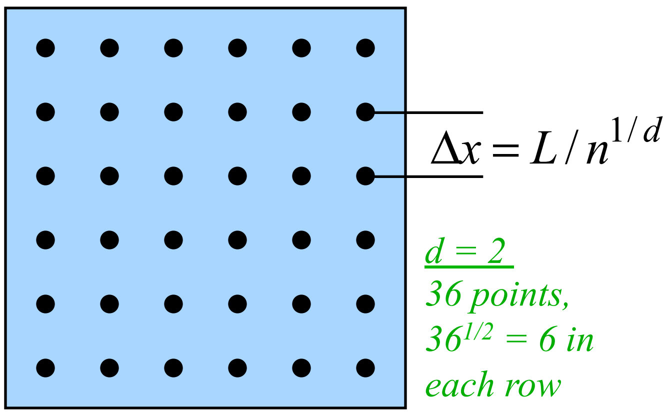
$$X_{n+1} = (aX_n + c) \bmod m$$ *linear congruential sequence*

- Returns set of numbers that meet many statistical measures of randomness
  - *histogram is uniform*
  - *no systematic correlation of deviates*
    - no idea what next value will be from knowledge of present value (without knowing generation algorithm)
    - but eventually, the series must end up repeating
- Some famous failures
  - *be careful to use a good quality generator*



Errors in Random vs. Methodical Sampling

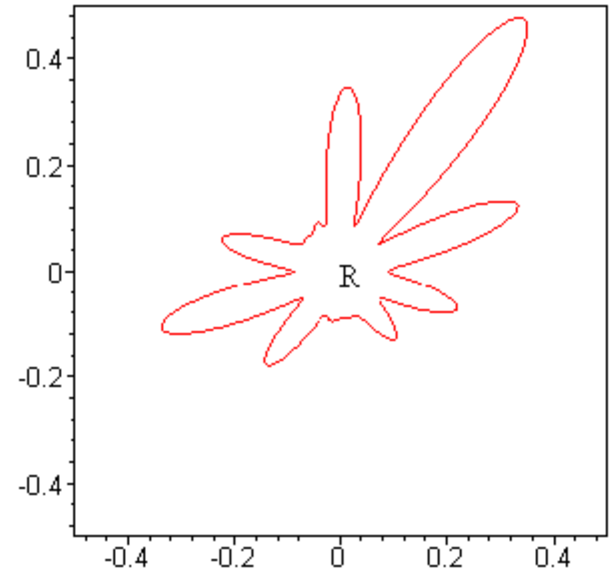
- Comparison of errors
  - *methodical approach* $\delta I \sim (\Delta x)^2 \sim n^{-2}$ ← *for example (Simpson's rule)*
  - *Monte Carlo integration* $\delta I \sim n^{-1/2}$
- MC error vanishes much more slowly for increasing *n*
- For one-dimensional integrals, MC offers no advantage
- This conclusion changes as the dimension *d* of the integral increases
  - *methodical approach* $\delta I \sim (\Delta x)^2 \sim n^{-2/d}$
  - *MC integration* $\delta I \sim n^{-1/2}$ ← *independent of dimension!*



- MC "wins" at about *d* = *4*

Shape of High-Dimensional Regions

- Two (and higher) dimensional shapes can be complex
- How to construct and weight points in a grid that covers the region *R*?



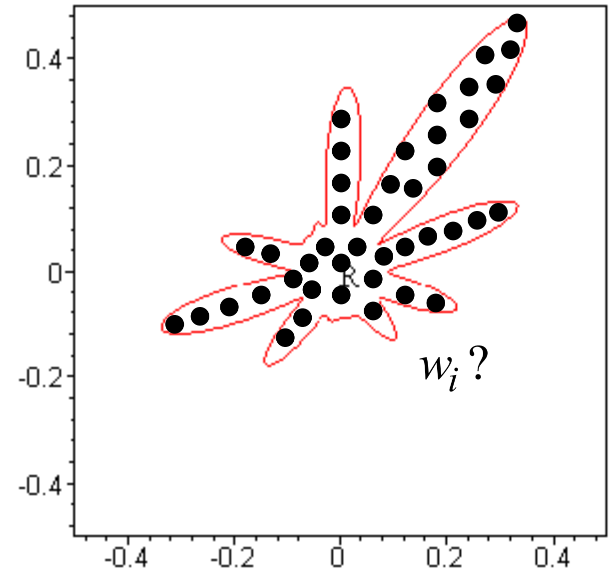
Example: mean-square distance from origin

$$\langle r^2 \rangle = \frac{\iint_R (x^2 + y^2)\,dxdy}{\iint_R dxdy}$$

Shape of High-Dimensional Regions

- Two (and higher) dimensional shapes can be complex
- How to construct and weight points in a grid that covers the region R?
  - *hard to formulate a methodical algorithm in a complex boundary*
  - *usually do not have analytic expression for position of boundary*
  - *complexity of shape can increase unimaginably as dimension of integral grows*
- We need to deal with 100+ dimensional integrals

$$\langle U \rangle = \frac{1}{Z_N}\frac{1}{N!}\int dr^N U(r^N) e^{-\beta U(r^N)}$$



Example: mean-square distance from origin

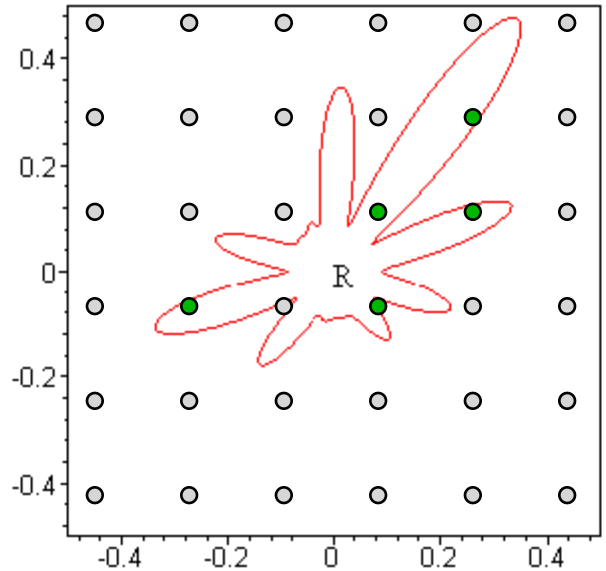
$$\langle r^2 \rangle = \frac{\iint_R (x^2+y^2)\,dxdy}{\iint_R dxdy}$$

Integrate Over a Simple Shape? 1.

- Modify integrand to cast integral into a simple shaped region
  - *define a function indicating if inside or outside* R

$$s = \begin{cases} 1 & \text{inside R} \\ 0 & \text{outside R} \end{cases}$$

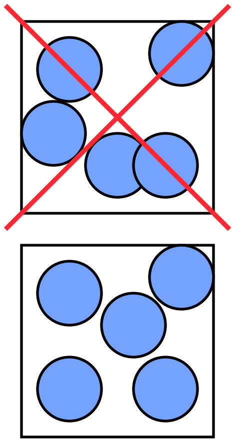
$$\left\langle r^2 \right\rangle = \frac{\int_{-0.5}^{+0.5} dx \int_{-0.5}^{+0.5} dy (x^2 + y^2) s(x,y)}{\int_{-0.5}^{+0.5} dx \int_{-0.5}^{+0.5} dy s(x,y)}$$



- Difficult problems remain
  - *grid must be fine enough to resolve shape*
  - *many points lie outside region of interest*
  - *too many quadrature points for our high-dimensional integrals (see applet again)*
- Click here for an applet demonstrating 2D quadrature

Integrate Over a Simple Shape? 2.

- Statistical-mechanics integrals typically have significant contributions from miniscule regions of the integration space
  - $\langle U \rangle = \frac{1}{Z_N} \frac{1}{N!} \int dr^N U(r^N) e^{-\beta U(r^N)}$
  - *contributions come only when no spheres overlap* $\left(e^{-\beta U} \neq 0\right)$
  - e.g., *100 spheres at freezing the fraction is* 10^{-260}



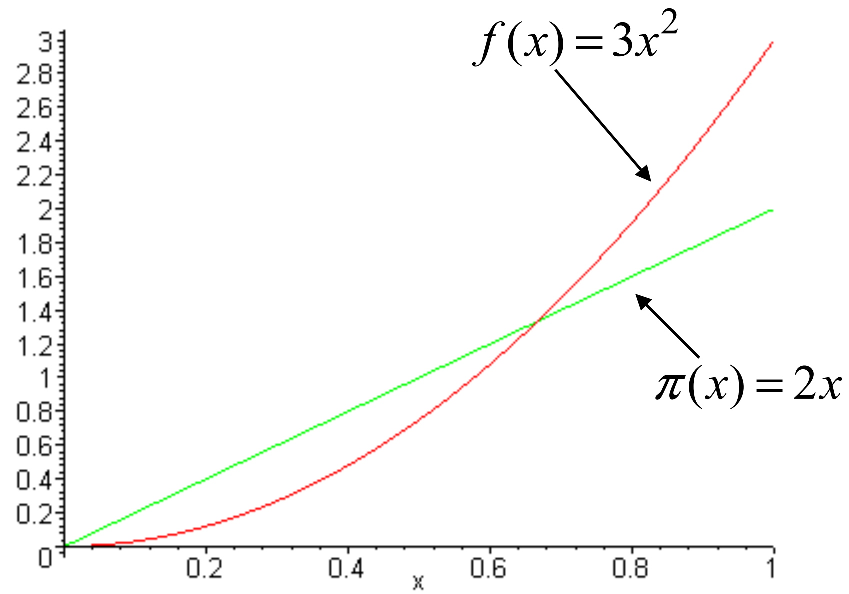
- Evaluation of integral is possible only if restricted to region important to integral
  - *must contend with complex shape of region*
  - *MC methods highly suited to "importance sampling"*

Importance Sampling

- Put more quadrature points in regions where integral receives its greatest contributions
- Return to 1-dimensional example

$$I = \int_0^1 3x^2 dx$$

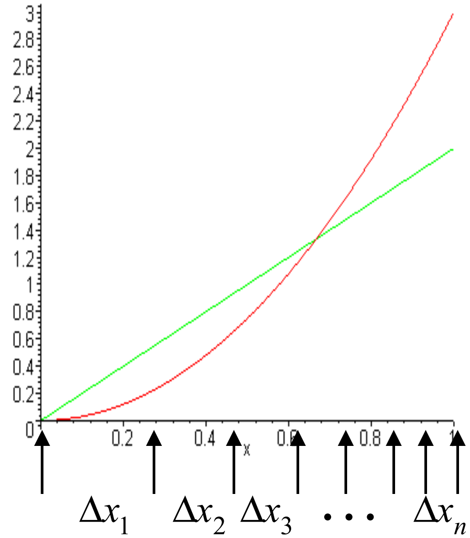
- Most contribution from region near x = 1
- Choose quadrature points not uniformly, but according to distribution p(x)
  - *linear form is one possibility*
- How to revise the integral to remove the bias?



The Importance-Sampled Integral

- Consider a rectangle-rule quadrature with unevenly spaced abscissas

$$I \approx \sum_{i=1}^{n} f(x_i)\Delta x_i$$



- Spacing between points
  - *reciprocal of local number of points per unit length*

$$\Delta x_i = \frac{b-a}{n}\frac{1}{\pi(x_i)}$$

Greater π ==> more points ➔ smaller spacing

- Importance-sampled rectangle rule
  - *Same formula for MC sampling*

$$I \approx \frac{b-a}{n}\sum_{i=1}^{n}\frac{f(x_i)}{\pi(x_i)}$$

$\pi(x)$ ← *choose x points according to p*

Generating Nonuniform Random Deviates

- Probability theory says...
  - *...given a probability distribution u(z)*
  - *if x is a function x(z),*
  - *then the distribution of π(x) obeys* $\pi(x) = u(z)\left|\frac{dz}{dx}\right|$
- Prescription for π(x)
  - *solve this equation for x(z)*
  - *generate z from the uniform random generator*
  - *compute x(z)*
- Example
  - *we want* $\pi(x) = ax$ *on* $x = (0,1)$
  - *then* $z = \frac{1}{2}ax^2 + c = x^2$ *a and c from "boundary conditions"*
  - *so* $x = z^{1/2}$
  - *taking square root of uniform deviate gives linearly distributed values*
- Generating π(x) requires knowledge of $\int \pi(x)dx$

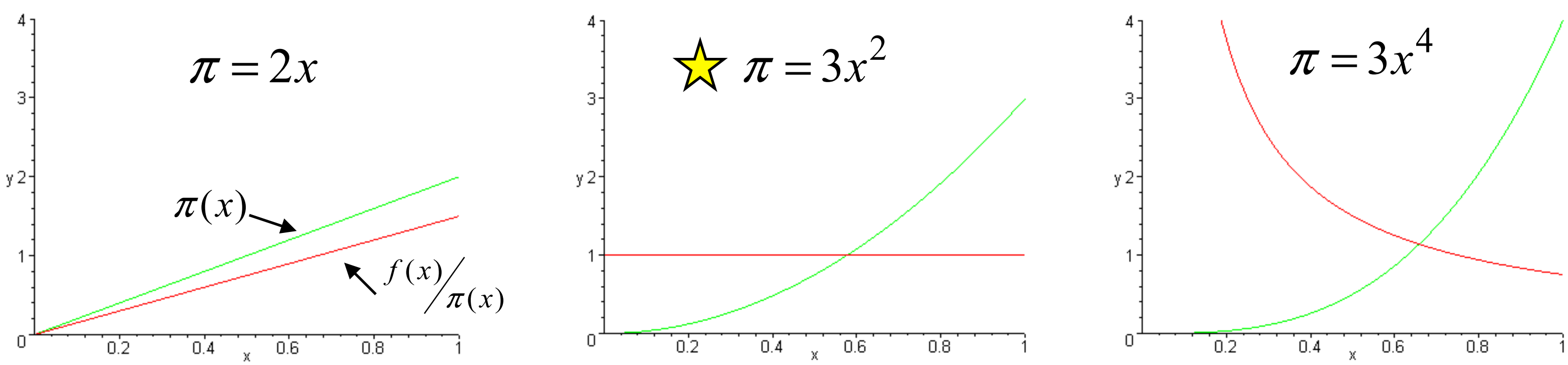
Choosing a Good Weighting Function

- MC importance-sampling quadrature formula

$$I \approx \frac{1}{n}\sum_{i=1}^{n}\frac{f(x_i)}{\pi(x_i)} = \left\langle \frac{f}{\pi} \right\rangle_{\pi}$$

$\pi(x)$

- Do not want π(x) to be too much smaller or too much larger than f(x)
  - *too small leads to significant contribution from poorly sampled region*
  - *too large means that too much sampling is done in region that is not (now) contributing much*



Variance in Importance Sampling Integration

- Choose π to minimize variance in average

$$\sigma_I^2 = \frac{1}{n}\left\{\int\left[\frac{f(x)}{\pi(x)}\right]^2 \pi(x)dx - \left[\int\left[\frac{f(x)}{\pi(x)}\right]\pi(x)dx\right]^2\right\}$$

$f(x) = 3x^2$

$\pi(x)$	σ_I	n = 100	n = 1000
1	$2/\sqrt{5n}$	0.09	0.03
$2x$	$1/\sqrt{8n}$	0.04	0.01
$3x^2$	0	0	0
$4x^3$	$1/\sqrt{8n}$	0.04	0.01

- Smallest variance in average corresponds to π(x) = c × f(x)
  - *not a viable choice*
  - *the constant here is selected to normalize π*
  - *if we can normalize π we can evaluate* $\int \pi(x)dx$
  - *this is equivalent to solving the desired integral of f(x)*
- Click here for an applet demonstrating importance sampling

Summary

- Monte Carlo methods use stochastic process to answer a non-stochastic question
  - *generate a random sample from an ensemble*
  - *compute properties as ensemble average*
  - *permits more flexibility to design sampling algorithm*
- Monte Carlo integration
  - *good for high-dimensional integrals*
    - better error properties
    - better suited for integrating in complex shape
- Importance Sampling
  - *focuses selection of points to region contributing most to integral*
  - *selecting of weighting function is important*
  - *choosing perfect weight function is same as solving integral*
- Next up:
  - *Markov processes: generating points in a complex region*