Today: Fundamentals of Monte Carlo

What is Monte Carlo?

- Named at Los Alamos in 1940’s after the casino.

- Any method which uses \textit{(pseudo)random numbers} as an essential part of the algorithm.  
\textit{Stochastic - not deterministic!}

- A method for doing \textit{highly dimensional integrals} by sampling the integrand.

- Often a Markov chain, called Metropolis MC.

- Reading:  Lesar Chapter 7
Simple example: Buffon’s needle
Monte Carlo determination of $\pi$

Consider a square inscribed by circle. Consider one quadrant.

By geometry:
\[
\text{area of } \frac{1}{4} \text{ circle} = \frac{\pi r^2}{4} = \frac{\pi}{4}.
\]

\[
\frac{\text{area of square}}{r^2} = \frac{\pi}{4}.
\]

Simple MC like throwing darts at unit square target:
- Using RNG in (0,1), pick a pair of coordinates $(x,y)$.
- Count # of points (darts) in shaded section versus total.

\[
\text{Pts in shaded circle/pts in square} \sim \frac{\pi}{4}.
\]

\begin{verbatim}
hits = 0
DO  n=1,N{
   x = (random #)
   y = (random #)
   distance2 = (x^2 + y^2)
   If (distance2 <= 1) hits = hits + 1
}
pi= 4 * hits/N
\end{verbatim}
MC is advantageous for high dimensional integrals - the best general method

Consider an integral in the unit \textit{D-dimensional} hypercube:

$$I = \int dx_1 ... dx_D \ f(x_1,...,x_D)$$

\textbf{By conventional deterministic methods:}
- Lay out a grid with \(L\) points in each direction with \(h=1/L\)
- Number of points is \(N = L^D = h^{-D}\) proportional to CPU time.

\textbf{How does error scale with CPU time or Dimensionality?}
- Error in trapezoidal rule goes as \(\varepsilon = f''(x) h^2\) since
  \(f(x) = f(x_0) + f'(x_0) h + (1/2)f''(x_0)h^2 + ...\)
  - “direct” CPU time \(\sim \varepsilon^{-D/2}.\) \((\varepsilon \sim h^2 \text{ and CPU} \sim h^{-D})\)
  - But by sampling we find \(\varepsilon^{-2}.\) \((\varepsilon \sim M^{-1/2} \text{ and CPU} \sim M)\)
  \textbf{To get another decimal place takes 100 times longer!}
  \textbf{But MC is advantageous for \(D>4!\)
Improved Numerical Integration

Integration methods in D-dimensions:
• Trapezoidal Rule: \( \varepsilon \sim f^{(2)}(x) \ h^2 \)
• Simpson’s Rule: \( \varepsilon \sim f^{(4)}(x) \ h^4 \)
• ... generally: \( \varepsilon \sim f^{(\alpha)}(x) \ h^\alpha \)

And CPU time scales with sample points (grid size)
• CPU time \( \sim h^{-D} \) (e.g., 1-D like 1/L and 2-D like 1/L^2)
• Time to do integral: \( T_{\text{int}} \sim \varepsilon^{-D/\alpha} \)

By conventional deterministic methods:
• Lay out a grid with L points in each direction with \( h=1/L \)
• Number of points is \( N=h^{-D}=L^D \sim \) CPU time.

Stochastic Integration (Monte Carlo)
• Monte Carlo: \( \varepsilon \sim M^{-1/2} \) (sqrt of sampled points)
• CPU time: \( T \sim M \sim \varepsilon^{-2} \).

> In the limit of small \( \varepsilon \), MC wins if \( D > 2\alpha \)!
Other reasons to do Monte Carlo:
- Conceptually and practically simple.
- Comes with built in error bars.

*Many methods of integration have been tried, and will be tried in this world of sin and woe. No one pretends that Monte Carlo is perfect or all-wise. Indeed, it has been said that Monte Carlo is the worst method except all those other methods that have been tried from time to time.*

Churchill 1947
Probability Distributions

- $P(x)dx =$ probability of observing a value in $(x, x+dx)$ is a \textit{probability distribution function} (p.d.f.)
  \[ \int dx \ P(x) = 1 \quad P(x) \geq 0 \]

- $x$ can be either a continuous or discrete variable.
- \textbf{Cumulative distribution}:
  \textit{Probability of $x<y$.}
  \textit{Useful for sampling}
  \[ c(y) = \int_{-\infty}^{y} dx \ P(x) \quad 0 \leq c(y) \leq 1 \]

- \textbf{Average or expectation}
  \[ \langle g(x) \rangle = \bar{g} = \int_{-\infty}^{\infty} dx \ P(x)g(x) \]

- \textbf{Moments}:
  - Zero\textsuperscript{th} moment $I_0 = 1$
  - Mean $\langle x \rangle = I_1$
  - Variance $\langle (x-\langle x \rangle)^2 \rangle = I_2 - (I_1)^2$
Mappings of random variables

Let $p_x(x)dx$ be a probability distribution
Let $y=g(x)$ be a new variable

-- e.g., $y=g(x) = -\ln(x)$ with $0 < x \leq 1$, so $y \geq 0$

What is the pdf of $y$?
With $p_y(y)dy = p_x(x)dx$

$$p_y(y) = p_x(x) \left| \frac{dx}{dy} \right| = p_x(x) \left| \frac{dx}{dg(x)} \right| = p_x(x) \left( \left| \frac{dg(x)}{dx} \right|_x \right)^{-1}$$

Example: $y=g(x) = -\ln(x)$

$$p_y(y)dy = p_x(x) \left| \frac{dg(x)}{dx} \right|^{-1} \ dy = e^{-y} dy$$

*Distributed exponentially, like in Poisson event, e.g. $y/\lambda$ has PDF of $\lambda e^{-\lambda y}$
What is Mapping Doing?

Generate random deviate

\[ p(x)dx = \begin{cases} 
  \frac{dx}{x} & 0 < x < 1 \\
  0 & \text{otherwise} 
\end{cases} \]

\[ \int_{-\infty}^{\infty} p(x)dx = 1 \]

PDF is normalized

Let \( p(y) = f(y) \), then \( y(x) = F^{-1}(x) \) (functional inverse), \( \frac{dx}{dy} = f(y) \) or \( x = F(y) \)

- Allows a random deviate \( y \) from a known probability distribution \( p(y) \).
- The indefinite integral of \( p(y) \) must be known and invertible.
- A uniform deviate \( x \) is chosen from \((0,1)\) such that its corresponding \( y \) on the definite-integral curve is desired deviate, i.e. \( x = F(y) \).
Example: Drawing from Normal Gaussian

\[ p(y_1, y_2, \ldots) dy_1 dy_2 \ldots = p(x_1, x_2, \ldots) \left| \frac{\partial(x_1, x_2, \ldots)}{\partial(y_1, y_2, \ldots)} \right| dy_1 dy_2 \ldots \]

**Box-Muller method:** get random deviates with normal distr.

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

**Consider:** uniform deviates (0,1), \( x_1 \) and \( x_2 \), and two quantities \( y_1 \) and \( y_2 \).

\[ y_1 = \sqrt{-\ln x_1} \cos 2\pi x_2 \quad y_2 = \sqrt{-\ln x_1} \sin 2\pi x_2 \]

Or, equivalently,

\[ x_1 = \exp(-[y_1^2 + y_2^2]/2) \quad x_2 = \frac{1}{2\pi} \arctan(y_2/y_1) \]

where

\[ \frac{\partial(x_1, x_2, \ldots)}{\partial(y_1, y_2, \ldots)} = -\begin{bmatrix} \frac{1}{\sqrt{2\pi}} e^{-y_1^2/2} \\ \frac{1}{\sqrt{2\pi}} e^{-y_2^2/2} \end{bmatrix} \]

So, each \( y \) is independently normal distributed.

**Better:** Pick \( R^2 = v_1^2 + v_2^2 \) so \( x_1 = \sqrt{R^2} \) and \( \angle(v_1, v_2) = 2\pi x_2 \)

Advantage: no sine and cosine by using \( v_1/\sqrt{R^2} \) and \( v_2/\sqrt{R^2} \) and get two RNG per calculation (1 for now, 2 for next time)
Reminder: Gauss’ Central Limit Theorem

Sample N values from $p(x)dx$, i.e. $(X_1, X_2, X_3, \ldots X_N)$.

**Estimate mean** from $y = (1/N)\sum x_i$.

**What is the pdf of mean?** *Solve by fourier transforms.*

If you add together two random variables, you multiply together their characteristic functions:

$$c_x(k) = \langle e^{ikx} \rangle = \int_{-\infty}^{\infty} dx \, P(x) e^{ikx} \quad \text{so} \quad c_{x+y}(k) = c_x(k) c_y(k)$$

Then $c_{x_1 + \ldots + x_N}(k) = c_x(k)^N$ and $c_y(k) = c_x(k/N)^N$

Taylor expand $\ln[c_y(k)] = \sum_{n=1}^{\infty} \frac{(ik)^n}{n!} \kappa^n$ **cumulants**
Cumulants: \( \kappa_n \)

Mean \( \kappa_1 = \langle x \rangle = x \)

Variance \( \kappa_2 = \langle (x-\bar{x})^2 \rangle = s^2 \)

Skewness \( \kappa_3/s^3 = \langle ((x-\bar{x})/s)^3 \rangle \)

Kurtosis \( \kappa_4/s^4 = \langle ((x-\bar{x})/s)^4 \rangle > -3 \)

What happens to the reduced moments?
- The \( n=1 \) moment remains invariant.
- The rest get reduced by higher powers of \( N \).

\[
\hat{\kappa}_n = \kappa_n N^{1-n}
\]

\[
\lim_{N \to \infty} c_y(k) = e^{ik\kappa_1 - k^2\kappa_2 / 2N - ik^3\kappa_3 / 6N^2 ...}
\]

\[
P(y) = \frac{N}{2\pi \kappa_2}^{1/2} \exp \left[ -\frac{N(y-\kappa_1)^2}{2\kappa_2} \right]
\]

Given enough averaging almost anything becomes a Gaussian distribution.
Approach to normality

Gauss’ Central Limit Thm

For any population distribution, the *distribution of the mean* will approach Gaussian.
Conditions on Central Limit Theorem

\[ I_n = \langle x^n \rangle = \int_{-\infty}^{\infty} dx \, P(x)x^n \]

- We need the first three moments to exist.
  - If \( I_0 \) is not defined \( \Rightarrow \) not a pdf
  - If \( I_1 \) does not exist \( \Rightarrow \) not mathematically well-posed.
  - If \( I_2 \) does not exist \( \Rightarrow \) infinite variance. Important to know if variance is finite for Monte Carlo.

- Divergence could happen because of tails of distribution

\[ I_2 = \langle x^2 \rangle = \int_{-\infty}^{\infty} dx \, P(x)x^2 \]

- We need:

\[ \lim_{x \to \pm \infty} x^3 P(x) \to 0 \]

- Divergence because of singular behavior of \( P \) at finite \( x \):

\[ \lim_{x \to 0} xP(x) \to 0 \]
Multidimensional Generalization

• Suppose \( \mathbf{r} \) is an m dimensional vector from a multidimensional pdf: \( p(\mathbf{r})d^m\mathbf{r} \).
• The mean is defined as before.
• The variance becomes the covariance, a positive symmetric m x m matrix:
  \[
  \nu_{i,j} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle
  \]
• For sufficiently large N, the estimated mean (\( \mathbf{y} \)) will approach the distribution:

\[
P(\mathbf{y})d\mathbf{y} = \left[ \frac{2\pi}{N} \det(\nu) \right]^{-1/2} \exp[-(y_i - \langle y_i \rangle) \frac{N\nu^{-1}}{2} (y_j - \langle y_j \rangle)]
\]
2d histogram of occurrences of means

- Off-diagonal components of $\nu_{ij}$ are called the co-variance.
- Data can be uncorrelated, positively or negatively correlated depending on sign of $\nu_{ij}$
- Like a moment of inertia tensor
- 2 principal axes with variances
- Find axes with diagonalization or singular value decomposition
- Individual error bars on $x_1$ and $x_2$ can be misleading if correlated.