Definition of Simulation

- What is a simulation?
  - It has an internal state \( "S" \)
    - In classical mechanics, the state = positions \( \{q_i\} \) and velocities \( \{p_i\} \) of the particles.
    - In Ising model, they are the spins (up or down \( \{\sigma_i\} \)) of the particles.
  - A rule for changing the state \( S_{n+1} = T(S_n) \)
    - In a random case, the new state is sampled from a distribution \( T(S_{n+1}|S_n) \).
  - From initial state \( S_0 \), we repeat the iteration many times: \( n \to \infty \)
    \[
    S_0 \to S_1 \to S_2 \to S_3 \to S_4 \to \ldots \to S_n \to S_{n+1} \to \ldots
    \]

- Sometimes we call the iteration index "\( n \)" “time.”
  It could be either “real time” or an iteration count, a pseudo-time, sometimes called Monte Carlo time.

Simulations can be:
- Deterministic (e.g. Newton’s equations via Molecular Dynamics)
- Stochastic (Monte Carlo, Brownian motion,…)
- Combination of the two
  Nonetheless, you analyze the errors the same way.
  As with experiment: the rules of the simulation can be simple but output can be unpredictable.

Ergodicity

- Typically simulations are assumed to be ergodic:
  - after a certain time the system loses memory of its initial state, \( S_0 \)
    except possibly for certain conserved quantities such as the energy, momentum.
  - The correlation time \( \kappa \) (which we will define soon) is the number of iterations it takes to forget.
  - If you look at (non-conserved) properties for times much longer \( \kappa \), they are unpredictable as if randomly sampled from some distribution.
  - Ergodicity is often easy to prove for the random transition but usually difficult for the deterministic simulation. More later.
  - The assumption of ergodicity is used for:
    - Warm up period at the beginning (or equilibration)
    - To get independent samples for computing errors.
Equilibrium distribution

- Let $F_t(S|S_0)$ be the distribution of state after time $t$.
- If the system is ergodic, no matter what the initial state, one can characterize the state of the system for $t \gg \kappa$ by a unique probability distribution: the equilibrium state $F^*(S)$.
  \[ \lim_{t \to \infty} F(S|S_0) = F^*(S) \]
- In classical statistical systems, this is the canonical Boltzmann distribution: $F^*(S) = \exp(-V(S)/kT)/Z$
- One goal is to compute averages to get properties in equilibrium. e.g. the internal energy:
  \[ U = \int dS F^*(S) V(S) \equiv \langle V(S) \rangle^* \]
- Another goal is to compute dynamics: for example the diffusion constant.

Estimated Errors

- In what sense do we calculate exact properties? Answer: if we average long enough the error goes to zero. Hence the error is under control.
- Next, how accurate is the estimate of the exact value?
  - Simulation results without error bars are only suggestive.
    - All homework exercises must include errors estimates
    - Without error bars one has no idea of its significance.
    - You should understand formulas and be able to make an “eye-ball” estimate.
- Error bar: the estimated error in the estimated mean.
  - Error estimates based on Gauss’ Central Limit Theorem.
  - Average of statistical processes has normal (Gaussian) distribution.
  - Error bars: square root of the variance of the distribution divided by the number of uncorrelated steps.
Central Limit Theorem (Gauss)

Sample $N$ independent values from $F'(x)dx$, i.e. $(x_1, x_2, x_3, \ldots, x_N)$.
Calculate mean as $y = (1/N)\sum x_i$.

What is the pdf of mean? Solve by fourier transforms

Characteristic function: $c_y(k) = \langle e^{ikx} \rangle = \int_{-\infty}^{\infty} dx \ F'(x)e^{ikx}$  
$c_y(k) = c_y(k/N)^N$

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

Cumulants: Mean $= \kappa_1$  Variance $= \kappa_2$  Skewness $= \kappa_3$  Kurtosis $= \kappa_4$
The n=1 moment remains invariant but the rest get reduced by higher powers of $N$.

Given enough averaging almost anything becomes a Gaussian distribution.

$$P(y) = \left( N/2\pi \kappa_2 \right)^{1/2} \exp \left[ \frac{-N(y - \kappa_1)^2}{2\kappa_2} \right] \quad \text{standard error}(y) = \sigma = \sqrt{\frac{\kappa_2}{N}}$$

Conditions on Central Limit Theorem

$$I_n = \langle x^n \rangle = \int_{-\infty}^{\infty} dx \ F'(x)x^n$$

- 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 simulations.
- Divergence could happen because of tails of distribution
  $$I_2 = \langle x^2 \rangle = \int_{-\infty}^{\infty} dx \ F'(x)x^2$$
- We need:
  $$\lim_{x \to \pm \infty} x^3 F'(x) \to 0$$
- Or divergence because of singular behavior of $F'$ at finite $x$:
  $$\lim_{x \to 0} x F'(x) \to 0$$
- We need to establish analytically that variance exists!
Approach to normality

Figure 1. Distributions of sums of uniform random numbers, each compared with the normal distribution. (a) $R_1$, the uniform distribution. (b) $R_2$, the sum of two uniformly distributed numbers. (c) $R_3$, the sum of three uniformly distributed numbers. (d) $R_{12}$, the sum of twelve uniformly distributed numbers.

What does infinite variance look like?

Spikes

Long tails on the distributions
Multidimensional Generalization of the central limit theorem

- 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 \( m \times m \) matrix:
  \[
  V_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle
  \]
- For sufficiently large \( N \), the estimated mean \( \mathbf{y} \) will approach the distribution:
  \[
  P(\mathbf{y})d\mathbf{y} = \left[ \frac{2\pi}{N} \det(V) \right]^{-1/2} \exp\left[-\frac{(\mathbf{y} - \langle \mathbf{y} \rangle)^T V^{-1} \frac{1}{2} (\mathbf{y} - \langle \mathbf{y} \rangle)}{2}\right]
  \]

2d histogram of occurrences of means

- Off-diagonal components of \( V_{ij} \) are called the co-variance.
- Data can be uncorrelated, positively or negatively correlated depending on sign of \( V_{ij} \).
- 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.
Estimate of errors

\[ \text{error}(\bar{a}) = \left( \frac{ \sum (a_i - \bar{a})^2 }{ N(N-1) } \right)^{1/2} \]

\[ \bar{a} = \frac{1}{N} \sum a_i \]

\[ \kappa = 1 + 2 \sum_{t=1}^{\infty} C(t) = \text{correlation time} \approx 2 \int_0^\infty \frac{dt}{\delta t} C(t) \]

\[ C(t,t') = \frac{\langle \delta a \delta a \rangle}{\langle \delta a \rangle^2} = C(|t-t'|) = \text{autocorrelation function} \]

\[ \left\langle \left( \bar{a} - \langle a \rangle \right)^2 \right\rangle = \left\langle \left( \frac{1}{N} \sum_{i \neq j} \delta a_i \delta a_j \right) \right\rangle \approx \frac{\langle \delta a \rangle^2}{N^2} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} C_i < \frac{\langle \delta a \rangle^2}{N^2} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} C_i = \frac{\langle \delta a \rangle^2 \kappa}{N} \]

\[ \text{Estimating Errors} \]

- Uncorrelated data
  \[ \langle a \rangle \approx \bar{a} = \frac{1}{N} \sum a_i \]
  \[ \text{error}(\bar{a}) = \left( \frac{ \sum (a_i - \bar{a})^2 }{ N(N-1) } \right)^{1/2} \]
  \[ \delta a_i = a_i - \bar{a} \]

- Correlated data
  \[ \text{error}(\bar{a}) = \left( \frac{ \sum (a_i - \bar{a})^2 }{ N(N-1) } \right)^{1/2} \]
  \[ \kappa = 1 + 2 \sum_{i=1}^{\infty} \frac{\langle \delta a_i \delta a_j \rangle}{\langle \delta a \rangle^2} = \text{correlation time} \]

- Problem: how to cut off the summation for \( \kappa \).
- Binning method: average together data in bins longer than the correlation time until it is uncorrelated.
Bias

- Bias is a systematic error caused by using a random number in a non-linear expression.
- You will get a result that is systematically too high or low.
- Suppose \( Z' = Z + \delta Z \) 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 \( <Z> \)

\[
F(Z) = F(Z) + \frac{dF}{dZ} \delta Z + \frac{1}{2} \frac{d^2F}{dZ^2} \delta Z^2 + L \quad \text{O}(N^{-1})
\]

\[
\text{bias}(F) = <F(Z) - F(Z)> = \frac{1}{2} \frac{d^2F}{dZ^2} \delta Z^2 + L \quad \text{error}(Z)^2 \quad \text{O}(N^{-1/2})
\]

\[
\text{error}(F) = [\langle (F(Z') - F(Z)) \rangle^2]^{1/2} = \frac{dF}{dZ} \langle \delta Z^2 \rangle^{1/2} + L \quad \text{err}(Z)
\]

\[
\text{bias}(F) = \frac{1}{2} \frac{d^2F}{dZ^2} \langle \delta Z^2 \rangle + L \quad \text{err}(Z)
\]

You may need to correct for the bias unless \( N \) is very large.

DataSpork

Interactive code to perform statistical analysis of data
Statistical vs. Systematic Errors

- What are statistical errors?
  - Statistical error measures the distribution of the averages about their avg.
  - *Statistical error can be reduced by extending or repeating runs*, increase N.
  
  \[
  \text{standard error}(y) = \sigma = \frac{\sqrt{\xi}}{N}
  \]

  - The efficiency is how we measure the rate of convergence of the statistical errors,
  
  \[
  \zeta = \frac{1}{T\sigma^2}
  \]

  - It depends on the computer, the algorithm, the property etc. But not on the length of the run.

- What are systematic errors?
  - Systematic error measures the error which is not sampling error. Even if you sample forever you do not get rid of systematic errors.
  - Systematic error is caused by round-off error, non-linearities, bugs, non-equilibrium, etc.
Recap: problems with estimating errors

- Any good simulation quotes systematic and statistical errors for anything important.

- The error and mean are simultaneously determined from the same data. HOW?

- Central limit theorem: the distribution of an average approaches a normal distribution (if the variance is finite).
  - One standard deviation means ~2/3 of the time the correct answer is within \( \sigma \) of the sample average.

- Problem in simulations is that data is correlated in time.
  - It takes a “correlation” time \( \kappa \) to be “ergodic”
  - We need to correct for correlation - this is a problem we can solve.
  - Also throw away the initial transient.

- We need about 20 independent data points to estimate errors. (so error of error is only 20%)
Statistical thinking is slippery: be careful

- “Shouldn’t the energy settle down to a constant”
  - NO. It fluctuates forever. It is the overall mean which converges.
- “The cumulative energy has converged”.
  - BEWARE. Even pathological cases have smooth cumulative energy curves.
- “Data set A differs from B by 2 error bars. Therefore it must be different”.
  - This is normal in 1 out of 10 cases. If things agree too well, something is wrong!
- “My procedure is too complicated to compute errors”
  - NO! NEVER! Run your whole code 10 times and compute the mean and variance from the different runs. If a quantity is important, you MUST estimate its errors.

Homework

- On computing error bars, using dataspork and writing Python script to compute errors.
- See the web site for the assignment.
- We can discuss python and errors next week in class.
- Homework due Monday, Jan 31st.