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) of the particles.
    • In any computer program, a finite number of bits
  – 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
    \[ S_0 \rightarrow S_1 \rightarrow S_2 \rightarrow S_3 \rightarrow S_4 \rightarrow S_5 \rightarrow \ldots \rightarrow S_n \rightarrow S_{n+1} \rightarrow \]

• Sometimes we call the iteration index “n” “time.”
  It could be either “real time,” an iteration count, or 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 and number of particles.
  – The correlation time $\tau$ (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 $\tau$,
    they are as unpredictable as if randomly sampled from some
    distribution.
  – Ergodicity can be proven within Monte Carlo but difficult for
    deterministic simulations. More about this later.
  – The assumption of egodicity is used for:
    • Warm up period (equilibration) at the beginning of a simulation
    • To get independent samples for computing errors.
Equilibrium distribution

• Let $F_t(S|S_0)$ be the distribution of the state after time $t$.
• If the system is ergodic, no matter what the initial state was, one can characterize the state of the system for $t >> \kappa$ 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:

$$F^*(S) = \frac{\exp(-V(S)/kT)}{Z}$$

• Typically, we want to compute properties in equilibrium. e.g. the internal energy, as an average over the simulation:

$$U = \int dS F^*(S)V(S) \equiv \langle V(S) \rangle_{F^*}$$

• 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 their significance.
    - You should understand formulas and be able to make an “eye-ball” estimate of errors
- 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 $F^*(x)dx$: $(x_1, x_2, x_3, \ldots, x_N)$.

Calculate mean as $y = (1/N) \sum x_i$.

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

**Characteristic function:**

$$c_x(k) = \langle e^{ikx} \rangle = \int_{-\infty}^{\infty} dx \ F^*(x) e^{ikx} \ c_y(k) = c_x(k/N)^N$$

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

**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.

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

Add n random numbers together.

From Kalos and Whitlock, “Monte Carlo Methods”
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 a mathematically well-posed integral.
  - 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 certain values of \( x \):

\[ \lim_{x \to 0} xF^*(x) \to 0 \]

- We need to establish \textit{analytically} that variance exists!
What does infinite variance look like?

Spikes

Long tails on the distributions

8/23/16

Atomic Scale Simulation
Interactive code to perform statistical analysis of data
Estimating Errors

- Uncorrelated data

\[
\{a_t\} \quad 0 < t \leq N
\]

\[
\langle a_t \rangle \approx \bar{a} = \frac{1}{N} \sum_t a_t
\]

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

\[
\delta a_t \equiv a_t - \bar{a}
\]

- Correlated data

\[
\text{error} (\bar{a}) = \left( \langle (\bar{a} - \langle a \rangle)^2 \rangle \right)^{1/2} \approx \left( \kappa \sum_t (a_t - \bar{a})^2 \right)^{1/2} \left[ \frac{\kappa \sum_t (a_t - \bar{a})^2}{N(N-1)} \right]
\]

\[
\kappa = 1 + 2 \sum_{t=1}^\infty \frac{\langle \delta a_t \delta a_0 \rangle}{\langle \delta a^2 \rangle} = \text{correlation time}
\]

- 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.
Correlated data

Uncorrelated data
Estimate of errors: how to deal with correlation

\[ \text{error}(\bar{a}) = \left\langle \left( \bar{a} - \langle a \rangle \right)^2 \right\rangle^{1/2} \approx \left\langle \frac{\kappa \sum_t (a_t - \bar{a})^2}{N(N-1)} \right\rangle^{1/2} \]

\[ \bar{a} = \frac{1}{N} \sum_t a_t \]

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

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

\[ \left\langle (\bar{a} - \langle a \rangle)^2 \right\rangle = \left\langle \frac{1}{N^2} \sum_{t, t'}^N \delta a_t \delta a_{t'} \right\rangle = \frac{\langle \delta a^2 \rangle}{N^2} \sum_{t, t'}^N C_{|t-t'|} \leq \frac{\langle \delta a^2 \rangle}{N^2} \sum_{t'=1}^N \sum_{t=-\infty}^\infty C_t = \langle \delta a^2 \rangle \frac{\kappa}{N} \]
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 $Z' = \bar{Z} + \delta Z$ 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 $<Z>$

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

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

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

**O(N^{-1})**

**O(N^{-1/2})**

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

\[
\text{standard error}(y) = \sigma = \sqrt{\frac{\kappa_2}{N}}
\]

• 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 refers to other types of errors, not sampling error. Even if you sample forever you do not get rid of systematic errors.
  – Systematic error can be 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.

- **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 $\tau$ to be “ergodic”
  - We need to correct for correlation-*this is a problem we can solve.*
  - Throw away the initial transient.

- We need about 20 *independent* data points to estimate errors. (so that the error of the error is only 20%)
Statistical Vocabulary

- Trace of A(t):
- Equilibration time.
- Histogram of values of A (P(A)).
- Mean of A (a).
- Variance of A (v).
- Bias of A
- estimate of the mean: \( \frac{\sum A(t)}{N} \)
- estimate of the variance
- Autocorrelation of A (C(t)).
- Correlation time k.
- The (estimated) error of the (estimated) mean (s).
- Efficiency \( \frac{1}{(CPU \ time \ * \ error^2)} \)
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
- Write Python script to compute errors.
- See the web site for the assignment.
- Homework due Thursday Sept 1.