Definition of Simulation

- What is a simulation?
  - It has an internal state "*S*"
    - Classical mechanics: positions $\{q_i\}$ and momenta $\{p_i\}$ 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, new states *sample* a distribution $T(S_{n+1}|S_n)$.
  - From initial state S_0, we repeat the iteration many times: $n \Rightarrow \infty$

$$S_0 \Rightarrow S_1 \Rightarrow S_2 \Rightarrow S_3 \Rightarrow S_4 \Rightarrow S_5 \Rightarrow \cdots \Rightarrow S_n \Rightarrow S_{n+1} \Rightarrow \cdots$$

- Sometimes we call the iteration index n as "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 κ (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 κ, 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 egodicity 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* $F^*(S)$.

$$\lim_{t\to\infty} F_t(S|S_0) = F^*(S)$$

- In classical statistical systems, this is the canonical Boltzmann distribution:

$$F^*(S) = Z^{-1} \exp(-H(S)/k_B T)$$

  – for Hamiltonian $H(S)$ = kinetic + potential energy
- One goal is to compute averages to get properties in equilibrium. e.g. the internal energy:

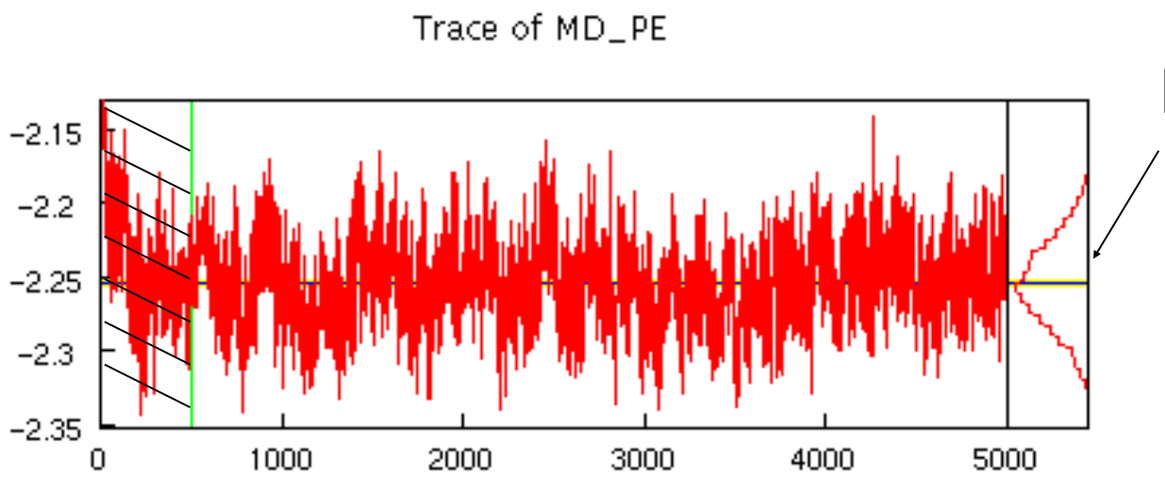
$$U = \int dS\, F^*(S) V(S) = \langle V(S) \rangle_{F^*}$$

- Another goal is to compute dynamics: for example the diffusion constant.

$$D = \lim_{t\to\infty} \frac{1}{2dt} \langle (r(t) - r(0))^2 \rangle$$

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. $(x_1, x_2, x_3, \ldots, x_N)$.

What is the pdf of mean $y = (1/N)\sum x_i$? Fourier transforms and cumulants.

Characteristic function:

$$c_x(k) = \langle e^{ikx} \rangle = \int_{-\infty}^{\infty} dx\, F^*(x) e^{ikx} = \exp\left[\sum_{n=1}^{\infty} \kappa_n \frac{(ik)^n}{n!} \right]$$

Cumulants:

$$\kappa_1 = \mu = \langle x \rangle, \quad \kappa_2 = \sigma^2 = \langle (x-\mu)^2 \rangle, \quad \kappa_3 = \langle (x-\mu)^3 \rangle, \quad \kappa_4 = \langle (x-\mu)^4 \rangle - 3\kappa_2^2$$

$$c_{x_1+x_2}(k) = c_{x_1}(k) c_{x_2}(k) \text{ and } c_{\alpha x_1}(k) = c_{x_1}(\alpha k) \text{ so } c_y(k) = (c_x(k/N))^N$$

$$c_y(k) = \exp\left[\sum_{n=1}^{\infty} \frac{\kappa_n}{N^{n-1}} \frac{(ik)^n}{n!} \right] \qquad \kappa_1(y) = \mu \qquad \kappa_3(y) = \kappa_3/N^2$$

$$\kappa_2(y) = \sigma^2/N \qquad \ldots$$

Distribution with κ_2 non-zero and all higher cumulants zero? **Normal**.

With enough samples almost anything becomes a normal distribution.

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

Standard error in mean: $\sqrt{\frac{\kappa_2}{N}}$

Multidimensional central limit theorem

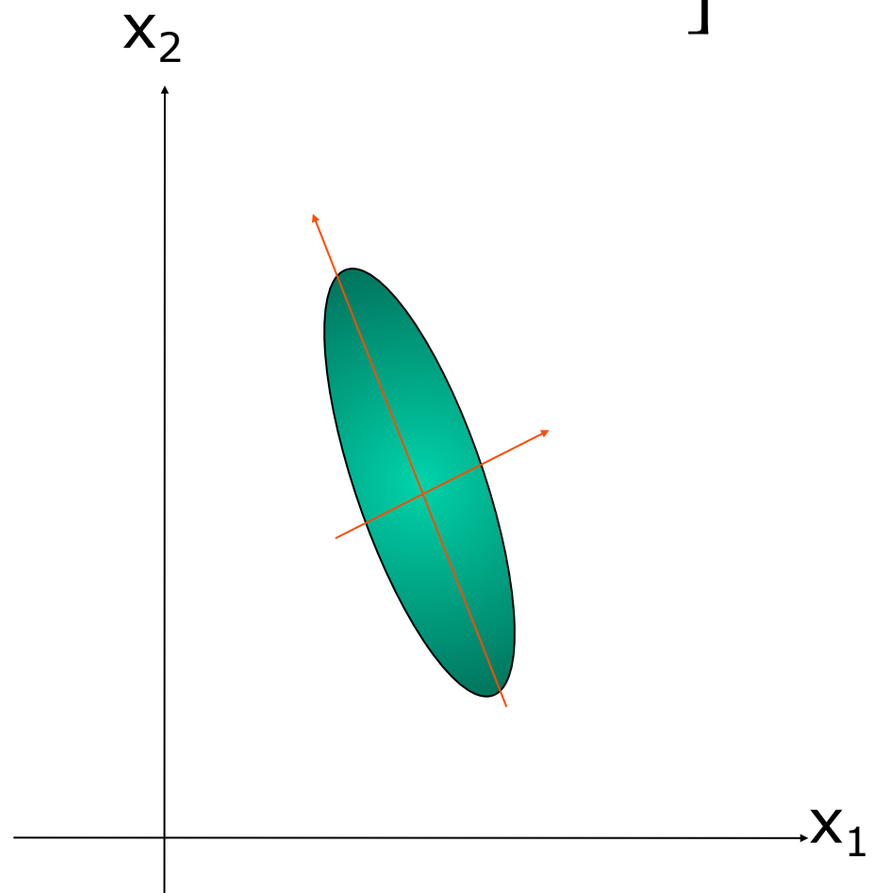
- Suppose the *x* to be sampled are *m* dimensional vectors from a multidimensional pdf: $P(x)d^m x$.
- The mean is defined as before.
- The *variance* becomes the *covariance*, a positive symmetric *m*×*m* matrix:

$$V_{ij} = \left\langle (x_i - \langle x_i \rangle)(x_i - \langle x_i \rangle) \right\rangle$$

- For sufficiently large *N*, the estimated mean (*y*) will approach the distribution:

$$P(y)d^m y = \sqrt{\frac{N}{2\pi \det V}} \exp\left[-\frac{N}{2} \sum_{ij} (y_i - \langle x_i \rangle)(V^{-1})_{ij}(y_j - \langle x_j \rangle) \right.$$

- 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.



Conditions on Central Limit Theorem

raw moments: $I_n = \langle x^n \rangle = \int_{-\infty}^{\infty} dx\, F^*(x) x^n$

- We need the first three moments to exist.
  - If I_0 is not defined ⇒ not normalizable (not a probability distribution)
  - If I_1 does not exist ⇒ not mathematically well-posed (no finite mean)
  - If I_2 does not exist ⇒ infinite variance. Important to know if variance is finite for simulations.
- Divergence could happen because of tails of distribution
  - We need: $I_2 = \langle x^2 \rangle = \int_{-\infty}^{\infty} dx\, F^*(x) x^2$

$$\lim_{x \to \pm\infty} x^3 F^*(x) = 0$$

  - And non-singular behavior of F* at all x_0:

$$\lim_{x \to x_0} (x - x_0) F^*(x) = 0$$

- We need to establish **analytically** that variance exists!

Approach to normality

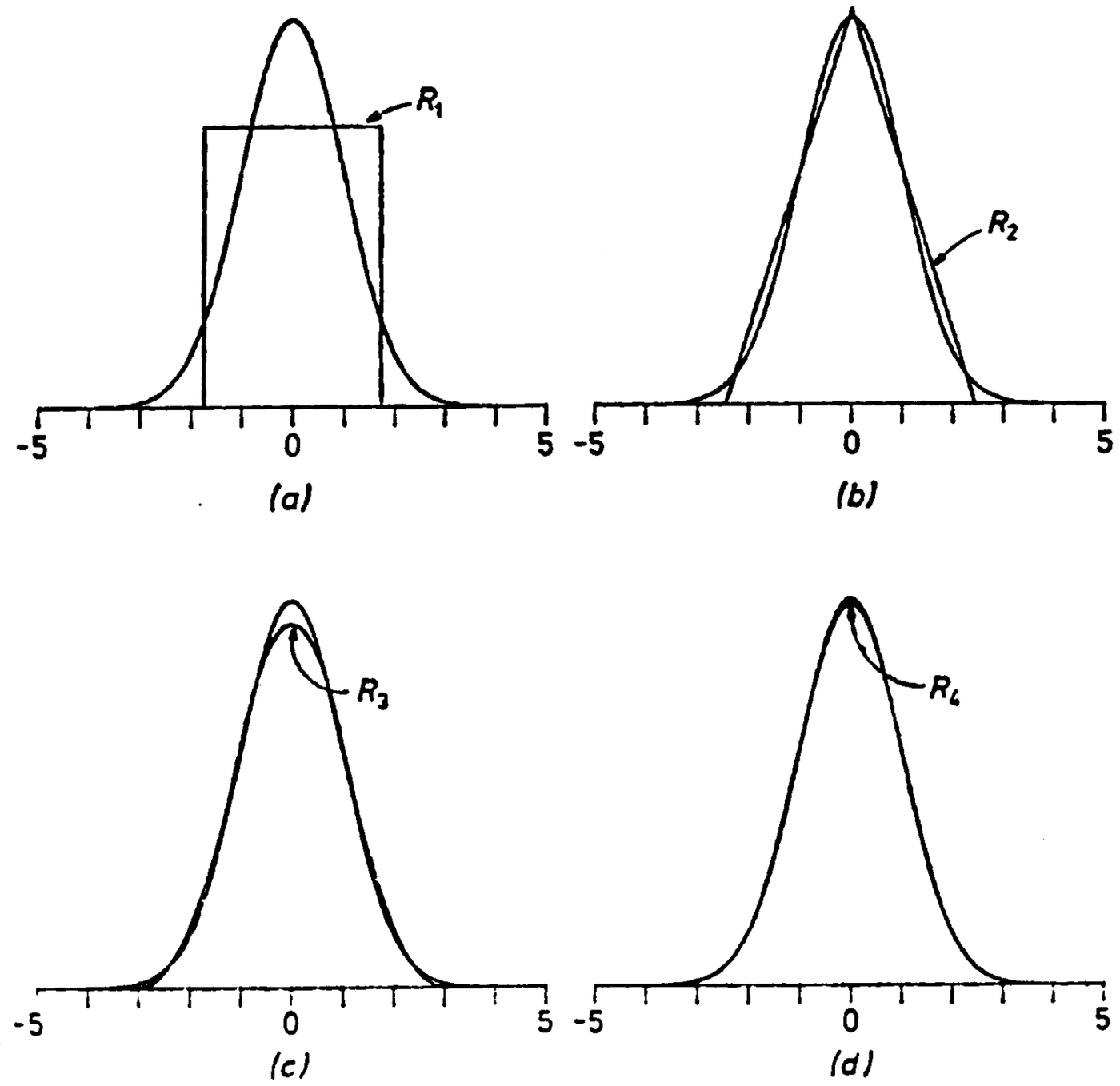
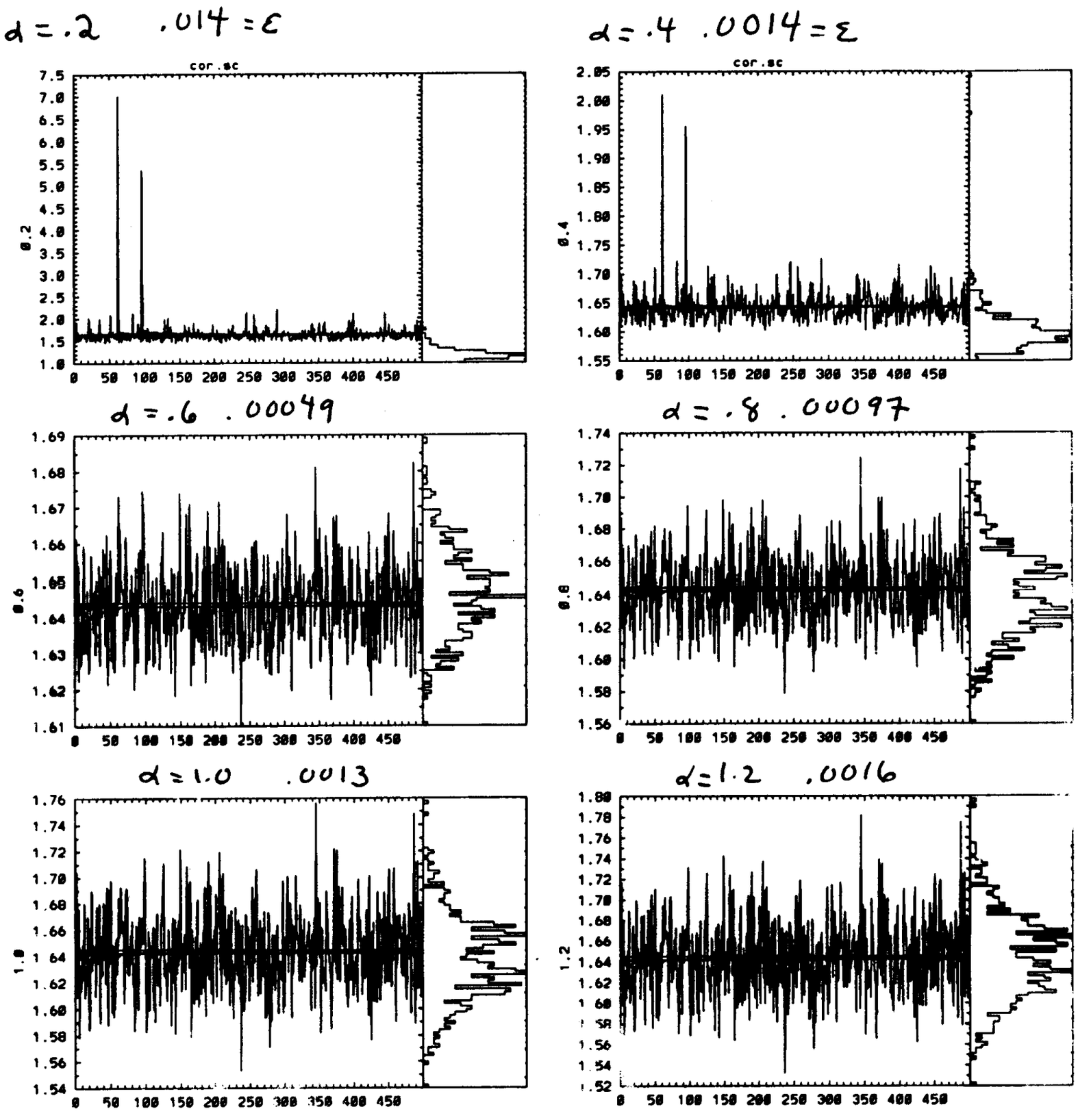


Figure 1. Distributions of sums of uniform random numbers, each compared with the normal distribution. (*a*) R_1, the uniform distribution. (*b*) R_2, the sum of two uniformly distributed numbers. (*c*) R_3, the sum of three uniformly distributed numbers. (*d*) R_{12}, the sum of twelve uniformly distributed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

What does infinite variance look like?

Spikes

Long tails on the distributions



Estimating errors: role of correlation in samples

Sampling data: $\{a_t\} \quad 0 < t \leq N$ Expectation value: $\langle a \rangle \approx \bar{a} = \frac{1}{N}\sum_t a_t$

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

- Assumes a_{t1} and a_{t2} are **independent** random variables
  - As an extreme case, imagine repeating every sample to get 2*N* samples
  - Suggests that the mean is a good estimator, but our error estimate could have problems
- Requires an approach to deal with correlation *inside sampling set*
  - **Binning**: group data into "block averages" of increasing sizes N_b until error estimate converges:

$$b_{t'} = \frac{1}{N_b}\sum_{t=t'N_b+1}^{t'N_b+N_b} a_t \qquad \bar{b} = \bar{a} \qquad \text{error}(\langle a \rangle) \approx \sqrt{\frac{N_b \sigma^2(b_{t'})}{N}}$$

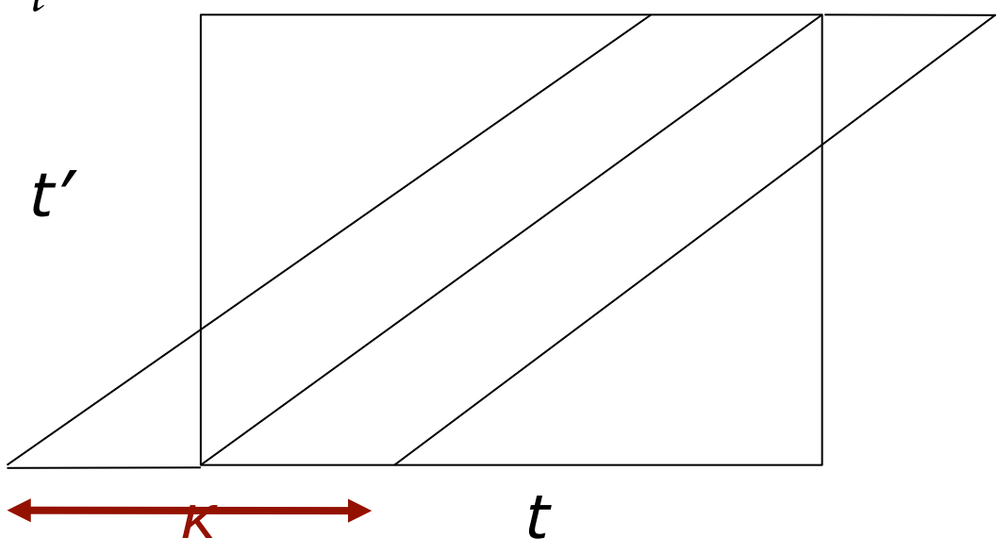
  - **Correlation time:** determine the number of steps required to make a_{t1} and a_{t2} uncorrelated. That time, κ, determines the number of truly independent samples present: N/κ

Estimating errors: correlation time

Auto time-correlation function: $C(t,t') \equiv \dfrac{\langle \delta a_t \delta a_{t'} \rangle}{\langle \delta a_t^2 \rangle} = C(|t-t'|)$

Correlation time:

$$\kappa = \int_{-\infty}^{\infty} dt\; C(t) \approx 1 + 2\sum_{t=1}^{\infty} C(t)$$



Error estimation **including** correlations:

$$\langle (\bar{a} - \langle a \rangle)^2 \rangle = \left\langle \frac{1}{N^2} \sum_{t,t'} \delta a_t \delta a_{t'} \right\rangle$$

$$= \frac{\langle \delta a_t^2 \rangle}{N^2} \sum_{t,t'} C(t,t')$$

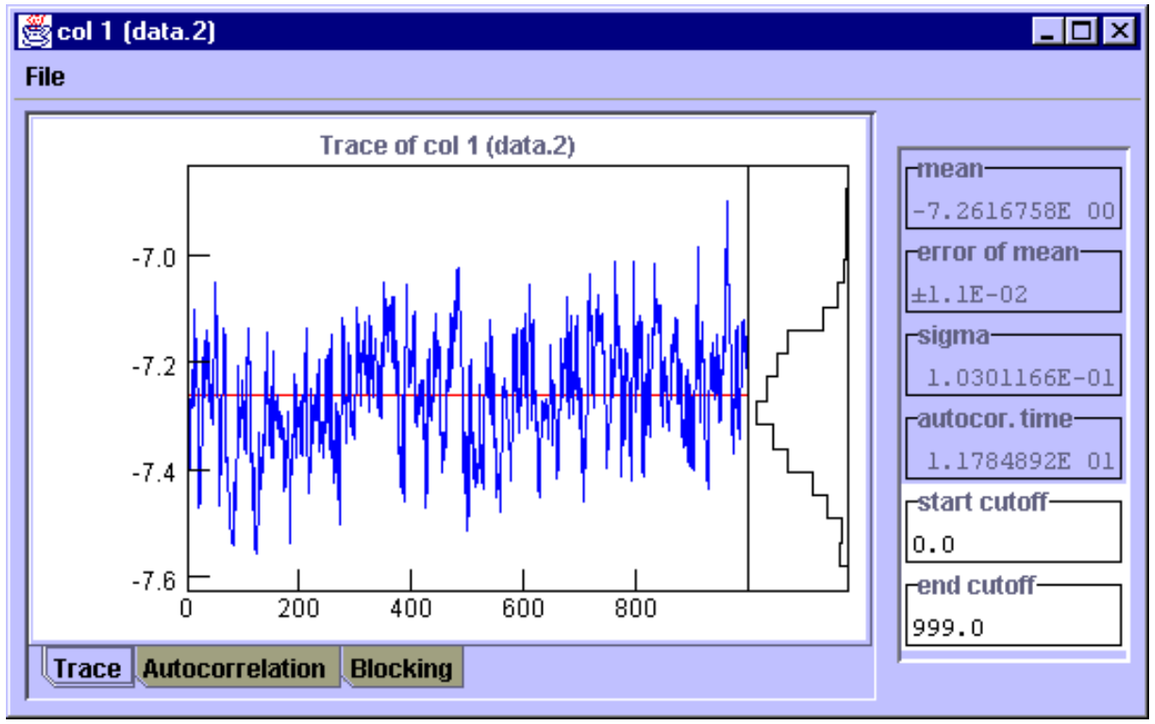
$$< \frac{\langle \delta a_t^2 \rangle}{N^2} \sum_{t=1}^{\infty} \sum_{\Delta t=-\infty}^{\infty} C(\Delta t)$$

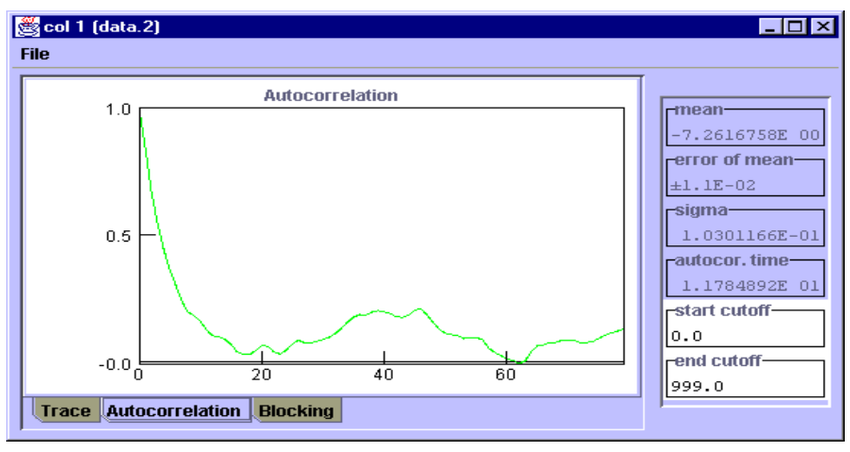
$$= \langle \delta a_t^2 \rangle \frac{\kappa}{N}$$

Number of "independent" samples in data

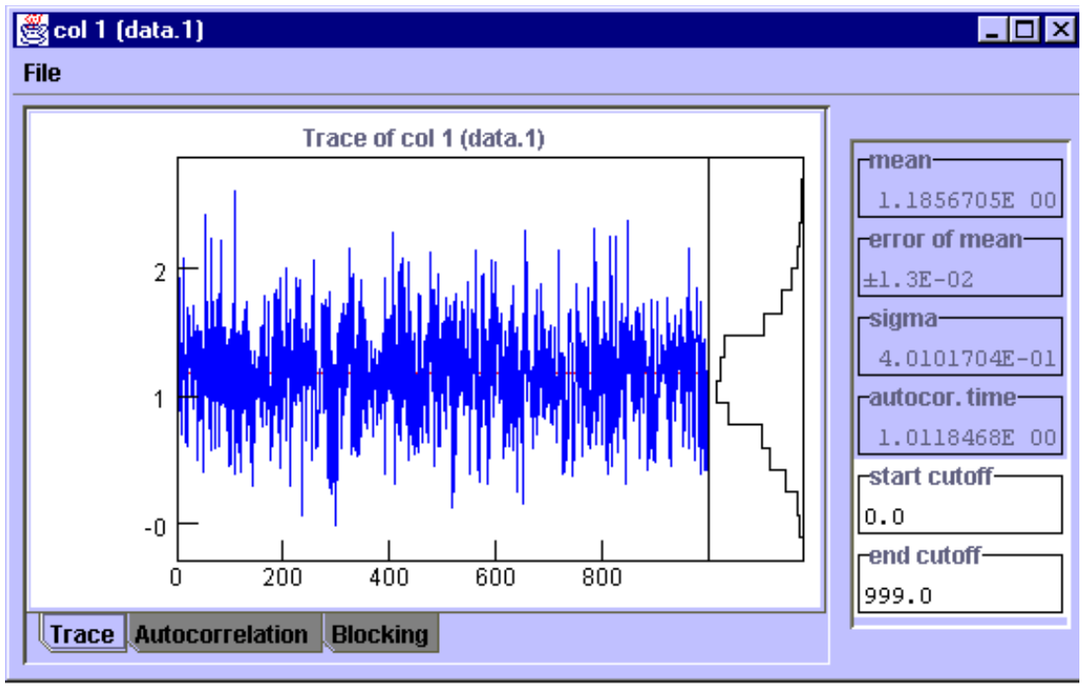
Note: correlation time evaluation is *cutoff* at finite length by finding C(t)=0

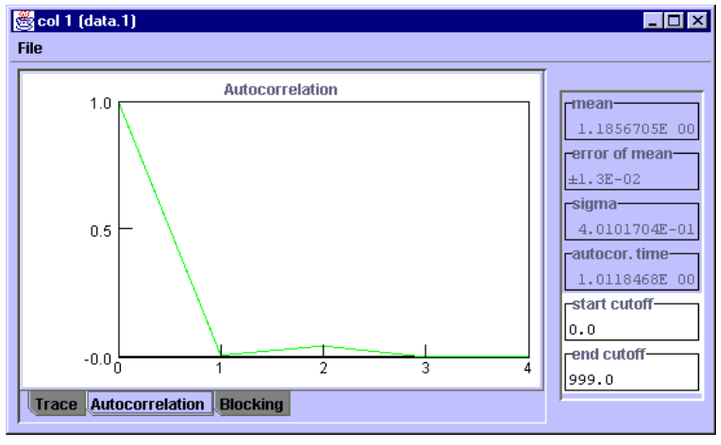
correlated data





uncorrelated data





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 $F(Z)$.
- What is the statistical error and bias of $F(Z)$?
- Expand Z in power series about mean value $\bar{Z}$*:*

$$F(Z) = F(\bar{Z}) + \left.\frac{dF}{dZ}\right|_{\bar{Z}} \delta Z + \frac{1}{2} \left.\frac{d^2F}{dZ^2}\right|_{\bar{Z}} \delta Z^2 + \cdots$$

$O(N^{-1})$

$$\text{bias}(F) = \langle F(Z) - F(\bar{Z}) \rangle = \frac{1}{2} \left.\frac{d^2F}{dZ^2}\right|_{\bar{Z}} \langle \delta Z^2 \rangle + \cdots \approx \frac{1}{2} \left.\frac{d^2F}{dZ^2}\right|_{\bar{Z}} \text{error}(Z)^2$$

$$\text{error}(F) = \left\langle \left(F(Z) - \langle F(Z) \right\rangle\right)^2 \rangle^{1/2} = \left| \left.\frac{dF}{dZ}\right|_{\bar{Z}} \right| \langle \delta Z^2 \rangle^{1/2} + \cdots \approx \left| \left.\frac{dF}{dZ}\right|_{\bar{Z}} \right| \text{error}(Z)$$

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

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

Statistical vs. Systematic Errors

- What are **statistical errors**?
  - Statistical error measures distribution of the averages about their avg.
  - Reduce statistical error with extended or repeating runs, increase N.

$$\text{Standard error in mean: } \sqrt{\frac{\kappa_2}{N}}$$

- Efficiency is a measure of 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 σ of the sample average.
- Problem in simulations is that data is correlated in time.
  - It takes a "correlation" time κ to be independent (and approach 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 Vocabulary

Name	Formula	Estimator
Data ("Trace")	$A_k;\ -\infty < k < +\infty$	$A_k;\ 1 \leq k \leq N$
Equilibriated Data Range	$-\infty < k < +\infty$	$k_1 \leq k \leq k_2;\ N_{eq} = k_2 - k_1 + 1$
Mean of A	$\langle A \rangle = \lim_{N\to\infty} \frac{1}{N} \sum_k A_k$	$a = \frac{1}{N_{eq}} \sum_{k=k_1}^{k_2} A_k \approx \langle A \rangle$
Variance of A	$V = \langle (A - \langle A \rangle)^2 \rangle$	$v = \frac{1}{N_{eq}-1} \sum_{k=k_1}^{k_2} (A_k - a)^2$
Standard Deviation of Data	$\sigma = \sqrt{V}$	
Autocorrelation of A	$C_A(i) = \frac{1}{V} \sum_{k=1}^{N-i} \langle (A_k - \langle A \rangle)(A_{k+i} - \langle A \rangle) \rangle$	$c_A(i) = \frac{1}{v} \frac{1}{N_{eq}-i} \sum_{k=k_1}^{k_2-i} (A_k - a)(A_{k+i} - a)$
Correlation Time of A	$\kappa = 1 + 2 \sum_{i=1}^{\infty} C_A(i)$	$\kappa = 1 + 2 \sum_{i=1}^{i_{cutoff}} c_A(i)$
Effective Number of Points	$N_{eff} = \lim_{N\to\infty} \frac{N}{\kappa} \to \infty$	$N_{eff} = \frac{N_{eq}}{\kappa}$
Error of Mean	$\sigma = \sqrt{\frac{V_A}{N_{eff}}} \to 0$	$\sigma = \sqrt{\frac{v_A}{N_{eff}}} = \sqrt{\frac{v_A \kappa}{N_{eq}}} \propto \frac{1}{\sqrt{N_{eq}}}$

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.

Python Intro - Links

Useful Links:

- Python related:
  - https://www.python.org/about/gettingstarted/
  - http://tdc-www.harvard.edu/Python.pdf (slightly outdated)

- Libraries and Math:
  - http://matplotlib.org/
  - http://www.numpy.org/
  - http://www.scipy.org/
  - http://numba.pydata.org/

- iPython/Jupyter notebook:
  - http://ipython.org/notebook.html
  - https://www.codecogs.com/latex/eqneditor.php