# Markov chain Monte Carlo methods for ensemble sampling

## Presentation for Metron Scientific Solutions

John Kerl

PhD Candidate, Department of Mathematics, University of Arizona

http://math.arizona.edu/~kerl

March 4, 2010

# Overview

In my research, I use Markov chain Monte Carlo methods to examine the relationship between interaction strength and critical temperature in a rather new model of random spatial permutations, which arises in statistical mechanics.

Today, I will leave the specialized topic of random spatial permutations aside, focusing instead on the widely applicable methods themselves. Content is taken from chapter 4 and appendix A of my dissertation.

Outline:

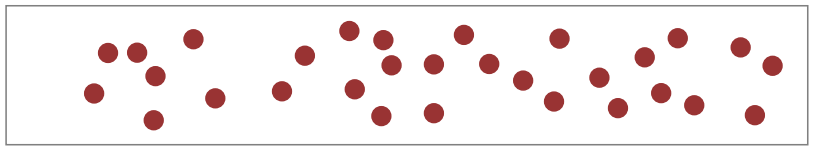
- Simplest motivating example
- Technical example
- Theoretical underpinnings
- Statistical analysis
- MCMC in broader contexts

Archetypal examples used throughout: mother-daughter sampling, die tipping, and lattice spin models.

# Sampling from a population

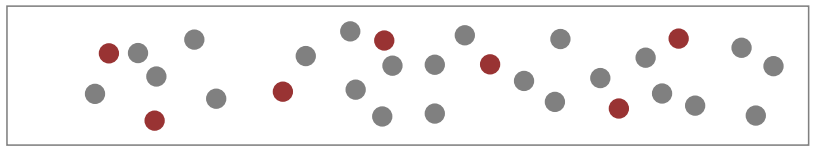
# Sampling from a population

If you could simultaneously measure the heights of all adults in the U.S., you would get an average: the population mean. It's a number with zero uncertainty (other than the uncertainty of the measurements themselves).



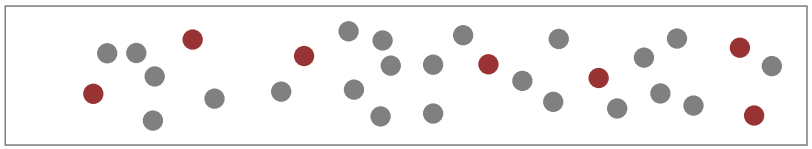
Full population

Since you can't do that, you might instead pick a few thousand people and hope it's a typical cross-section (e.g. you haven't gotten the entire NBA included in your sample). Now the sample mean is a random variable with its own uncertainty. The error bar (standard deviation of the sample mean) decreases in sample size $M$; the sample mean converges to the population mean as the sample size increases. We might call this random sampling a Monte Carlo method.



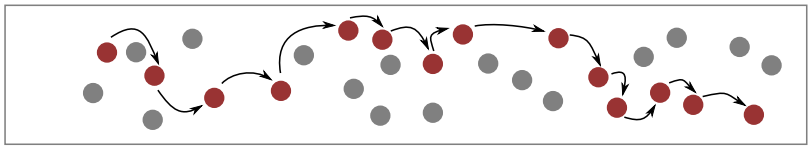
First random sample

Second random sample



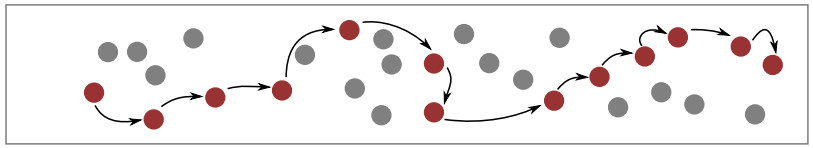
# Sampling from a population

Suppose instead that your sample consists solely of mother, daughter, granddaughter, and so on for $M$ generations. (Ignore generational drift in population height.) The sample mean still approaches the population mean, but more slowly: each successive data point tends to lie close to its predecessor. It takes time for the effect of a tall or short ancestor to dampen out.



First correlated sample

Second correlated sample



The error bar on the sample mean — the variation in the sample mean over many such experiments — is bigger due to these correlations between generations.

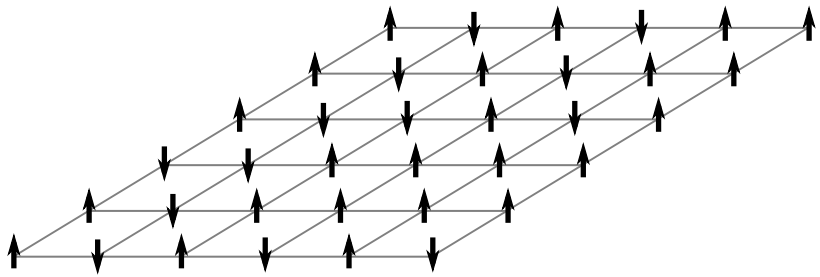
This is an idea of what the samples produced by a Markov chain Monte Carlo method look like.

# Lattice spin models

# Lattice spin models

A somewhat artificial example from statistical mechanics, which is easy to explain and visualize: lattice spin models are abstractions of real ferromagnetic materials. Picture an array of spins. A 2D checkerboard with spins either up or down is easy to think about (and suffices for this talk). There are 3D models with arbitrary-pointing spins.

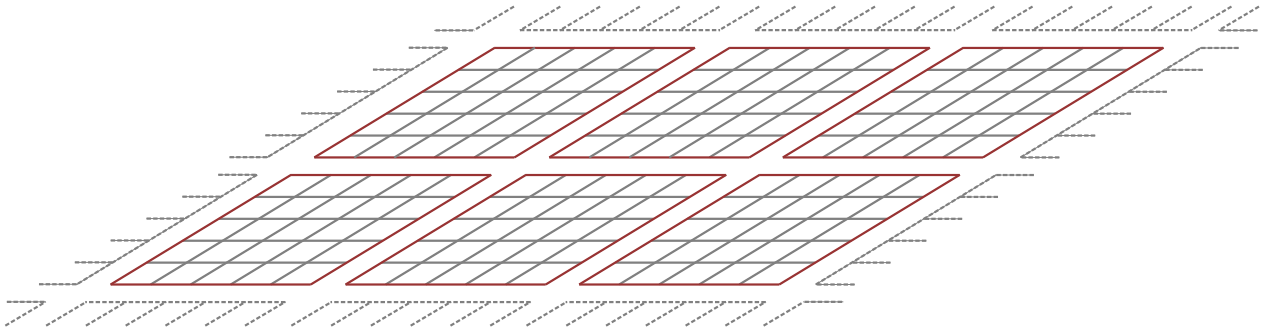
Spins at each site are induced to align with their neighbors. If a site's left-hand neighbor is up and the right-hand neighbor is down, what happens at the site? Worse, there is no leader — all spins simultaneously try to align with their neighbors.



- If the coupling is strong, all spins point in the same direction. The material is highly ordered.
- If the coupling is absent, spins can point in any direction, independently of one another. The material is disordered.
- In between: what happens? Are there perhaps islands of ups and downs? If so, with what average diameters? What do they look like?

# Lattice spin models: ergodic hypothesis

Bulk material has very many (on the order of Avogadro's number) spins. The bulk behavior is the average over many manageably small regions.



Outermost strategy when applying MCMC methods to statistical mechanics: examine $L \times L$ regions, applying statistical analysis (below). Then, use finite-size-scaling analysis on results obtained for larger and larger $L$. Most of this talk examines behavior with a fixed $L$.

# Lattice spin models: ergodic hypothesis

What makes Monte Carlo simulation of such systems work is the ergodic hypothesis: the spatial average (or time average, for time-evolving systems) is the same as the ensemble average. Meaning, weight each configuration $S$ of a region by its probability $P(S)$ of occurring in the bulk. E.g. on $2 \times 1$ lattice, there are 4 configurations $S$:

$$\uparrow\uparrow, \uparrow\downarrow, \downarrow\uparrow, \downarrow\downarrow .$$

They might occur with, say, respective probabilities $0.4, 0.1, 0.1, 0.4$.

We can measure a quantity of interest $Q(S)$ for each possible configuration. Then, given a probability distribution for all possible configurations $S$ of an $L \times L$ box,

$$\mathbb{E}[Q] = \sum_{\text{possible } S} P(S)Q(S).$$

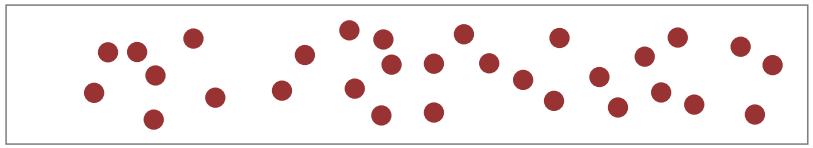
This is what would be measured in the bulk.

Example quantity: $+1$ for up arrows and $-1$ for down arrows. Then $\mathbb{E}[Q]/N$ is mean magnetization per site: close to $\pm 1$ when long-range alignment is present; close to 0 when alignments are small relative to the bulk size. This quantity doesn't measure grain diameter.

Another example quantity: spin products $s_i s_j$ for two fixed sites $\mathbf{x}_i$ and $\mathbf{x}_j$. This is pair correlation. As a function of distance $\|\mathbf{x}_i - \mathbf{x}_j\|$, it helps in quantifying grain diameters.

# Lattice spin models: random sampling and sampling variability

Population sampling problem: Consider a $10 \times 10$ box, with up or down spins at each site. There are $N = 100$ sites, and $2^{100} \approx 10^{30}$ possible spin configurations. These can't all be summed over. As with heights of people, the population is too big.



Full population

Instead, try to select a sample of $M$ most-likely configurations — ones with high $P(S)$ which contribute significantly to the sum. (E.g. with moderately strong coupling, aligned configurations should happen more often; with external upward-pointing magnetic field, up-pointing configurations should happen more often.)

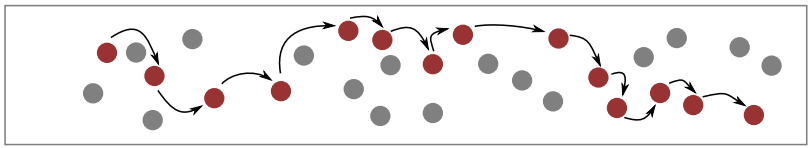
There is now sampling variability in the estimate $\overline{Q}_M$ of $\mathbb{E}[Q]$: it is now a random variable with its own error bar. This is not unlike the sampling error induced by polling 3,000 people to estimate their heights, or to gauge the opinions of millions of people[1]. This is the Monte Carlo — random sampling — part.

[1] Although people change their minds over time, adding another degree of complexity to political polling.

# Lattice spin models: invariant and limiting distributions

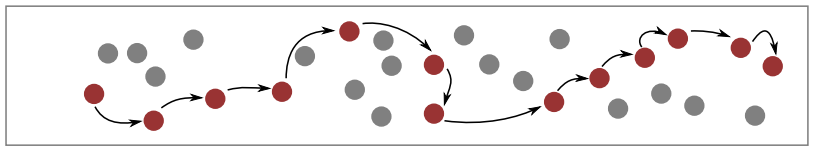
Solving one problem creates another: if there is a huge set of possible configurations (too many to even count), how do I pick out a few million most-likely ones — if I don't know what they look like in the first place?

This is where Markov chains come in. We pick *any* initial configuration. Then we propose, and maybe accept, a simple change (e.g. flipping one site's spin[2]). Keep doing that. A sequence of configurations results. As long as our change-proposal algorithm satisfies a few hypotheses, this sequence can be averaged over (with important caveats coming up).



First correlated sample

Second correlated sample



[2]More sophisticated cluster-update methods are needed in the critical parameter regime where the transition to long-range correlation begins to appear.

# Lattice spin models: invariant and limiting distributions

Sketch of an MCMC implementation:

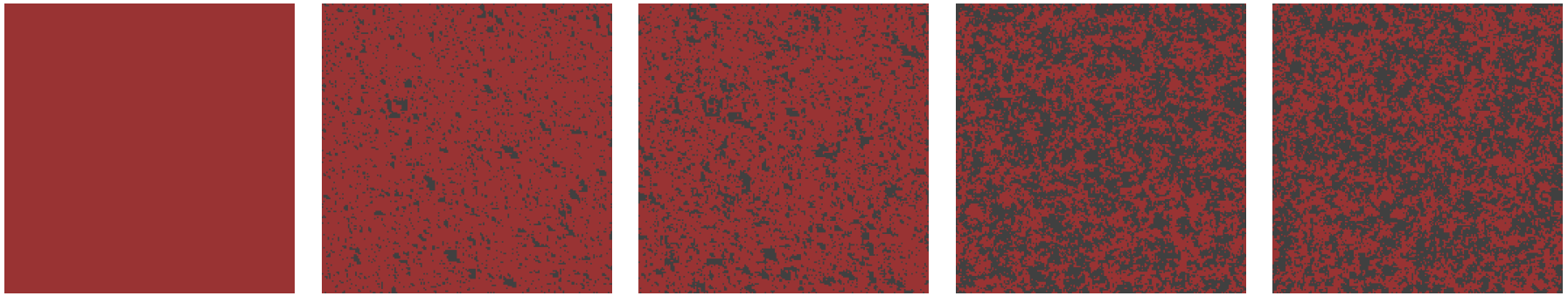
- Design a configuration-modification rule which satisfies the hypotheses below. Simple changes (e.g. flipping a single arrow) will turn out to be CPU-efficient (change of energy is easier to compute).
- Start with the system in a convenient configuration, even a highly unlikely one.
- Make a sequence of modifications until the configurations start to become "typical". (This isn't trivial but can be detected rather easily.) This is the burn-in or mixing or thermalization phase.
- Then, keeping making modifications, continuing the sequence of configurations. But now, remember quantities $Q(S_i)$ for each configuration $S_i$. This is the accumulation phase.
- Conduct necessary statistical analysis of the samples.

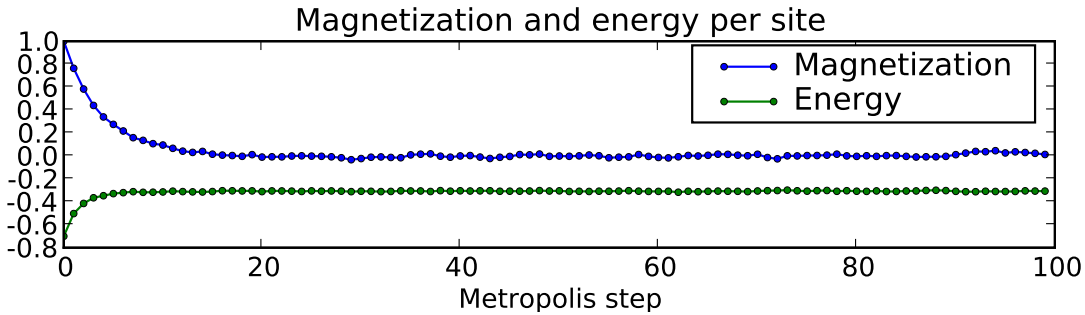
Note that adjacent configurations resemble one another — they are correlated — as in the mother-daughter sequence mentioned at the beginning.

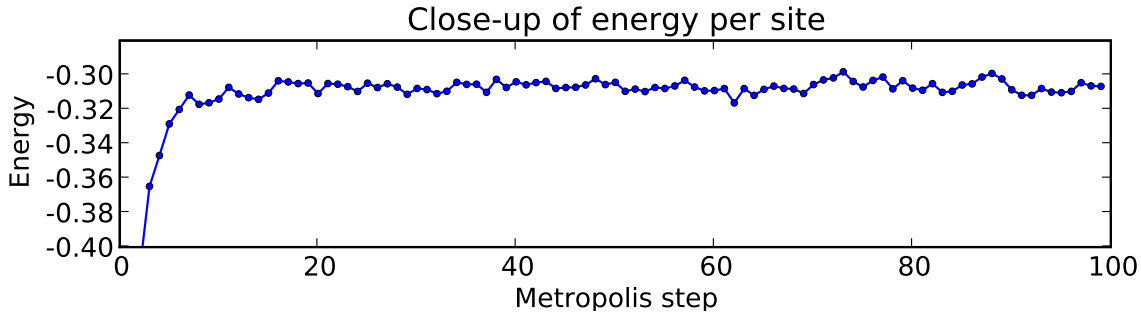
Terminology: a sweep involves proposing a change at each one of the $N$ lattice sites.

# Lattice spin models: pictures

Thermalization is rather quick. Here are configurations at sweep 0, 1, 2, 40, 100.





# Theoretical underpinnings

# Theoretical underpinnings

We have a finite probability space[3] $(\Omega, P_0)$. The probability space for sequences in $\Omega$ is $\Omega \times \Omega \times \cdots$, with the product measure $P$.

Then $P_k(S_k = \omega) := P(\Omega^{k-1} \times \{\omega\} \times \Omega \times \cdots)$ is the marginal at the $k$th slot. A configuration sequence is a sequence of random variables, all on the same configuration space — but not necessarily either independent or identically distributed. This is a discrete-time stochastic process, indexed by the positive integers.

A stochastic process has the Markov property if, for all $k > 0$,

$$P(S_{k+1} = \omega_{k+1} \mid S_1 = \omega_1, S_2 = \omega_2, \ldots S_k = \omega_k) = P(S_{k+1} = \omega_{k+1} \mid S_k = \omega_k).$$

This is true whenever we choose the next configuration by looking only at the current configuration, without retaining memory of previous configurations. A discrete-time stochastic process with the Markov property, on a finite configuration space, is called a Markov chain.

[3] With $\Omega$ finite, the $\sigma$-field is $2^\Omega$. For infinite $\omega$, the $\sigma$-field must be specified.

# Theoretical underpinnings: homogeneous vs. stationary

A homogeneous Markov chain has the same transition probabilities for each $k$.

A stationary chain has the same probability distribution at each $k$.

Two examples illustrate the difference. Example 1:

- Take an ordinary die. $\Omega = 1, 2, 3, 4, 5, 6$.
- Place the die with 6 up. At this first step, $P_1(S_1 = 6)$ with probability 1 (viewed from the perspective of running many such experiments); the other five faces are up with probability 0.
- Picking one of the four sides at random, tip the die. Opposite faces sum to seven, so $P_2(S_2 = 2) = P_2(S_2 = 3) = P_2(S_2 = 4) = P_2(S_2 = 5) = 1/4$.
- Tip again. Distribution $P_3$: $(1/4, 1/8, 1/8, 1/8, 1/8, 1/4)$.
- After many tips, each face is up with probability approaching $1/6$. The memory of the initial configuration is forgotten.

The probability distributions aren't the same at each step, but the same die-tipping rule is applied at each step. This chain is homogeneous but not stationary.

# Theoretical underpinnings: homogeneous vs. stationary

Example 2:

- Just as before, but pick the initial configuration by rolling the die, i.e. $P_1$ is uniform.
- Enumerating cases, or computing with the Markov matrix, shows that $P_2$ is also uniform. Likewise for all subsequent steps.

The probability distributions are the same at each step, and the same die-tipping rule is applied at each step. This chain is homogeneous and stationary.

After many steps, the chains of examples 1 and 2 are indistinguishable — the former has converged to the latter.

To get a non-homogeneous chain, you'd have to change the rules along the way.

Summary for Markov chains:

- Choose an initial distribution $P_1$.
- Choose a transition rule $P(S_{k+1} = \omega_j \mid S_k = \omega_i)$. This is a $K \times K$ matrix $M$ if $\#\Omega = K$.
- This specifies probability distributions for all subsequent steps.

# Theoretical underpinnings: hypotheses

Definition: A Markov chain is irreducible if any configuration is reachable from any other, in one or more steps.

Definition: A configuration $S$ has period $p$ if any return to $S$ must occur at multiples of $p$ steps. A state is aperiodic if $p = 1$. The entire chain is said to be aperiodic if all states are aperiodic. An example of a periodic chain is die-inverting (or double-tipping): 1 goes to 6 goes to 1 goes to 6 goes to 1 ....

Definition: A Markov matrix $M$ on a configuration space $\Omega$ (with $\#\Omega = K$) and a distribution $P$ on $\Omega$ are reversible, or satisfy detailed balance, if for all $1 \leq i, j \leq K$, $M_{ij}P(S_i) = M_{ji}P(S_j)$.

Terminology: An irreducible, aperiodic chain on a finite configuration space is sometimes called ergodic.

# Theoretical underpinnings: the invariant-distribution and sampling theorem

Invariant distribution theorem: (1) If the chain with transition matrix $M$ and initial distribution $P$ is irreducible, aperiodic, and reversible, then $P$ is invariant for $M$. (2) If the chain with transition matrix and initial distribution $P_1$ is irreducible, aperiodic, and reversible, then for each configuration $S$ in $\Omega$, $P_n(S) \to P(S)$ as $n \to \infty$.

Remark: The theorem does not address how many steps $n$ for $P_n$ to reach $P$ within some chosen tolerance.

Sampling theorem: Let $X$ be a random variable on the finite probability space $(\Omega, 2^\Omega, P)$. If the stationary Markov chain $(M, P)$ satisfies the hypotheses of the invariant distribution theorem, then

$$\frac{1}{M}\sum_{i=1}^{M} X(S_i) \to \mathbb{E}[X] \quad \text{as} \quad M \to \infty.$$

# Theoretical underpinnings: Metropolis methods

The preceding theory tells us what can happen if we have an ergodic reversible chain. But it doesn't tell us how. Nick Metropolis et al. have the following construction[4]:

- Each configuration $S$ in $\Omega$ has an energy $H(S)$.
- The probability distribution on $\Omega$ is $P = e^{-H(S)}/Z$, where the normalizing factor $Z$ is $\sum_{T \in \Omega} e^{-H(T)}$.
- Design an update rule so that in configuration $S$, a successor state $S'$ is chosen. One needs to check for aperiodicity and irreducibility.
- Accept the change with probability $\min\{1, e^{-\Delta H}\}$. This will give detailed balance.

For the figures produced above, the energy is the sum of nearest-neighbor spin products:

$$H(S) = -c \sum_{i \circ\!\!-\!\!\circ j} s_i s_j + h \sum_i s_i.$$

The constant $c$ determines the coupling strength and $h$ is the external field; above, they were $c = 0.35$ and $h = 0$. The only proposed updates I used were single spin flips, which is a naive algorithm suitable for a conceptual talk.

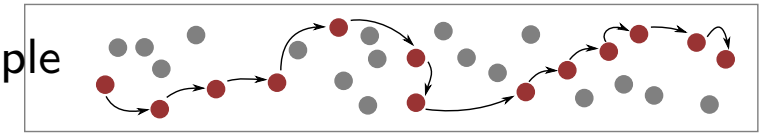
[4]Which can be presented in a more general setting, without Gibbs distributions. Moreover, there exist many other ways of constructing Markov chains for Monte Carlo methods.

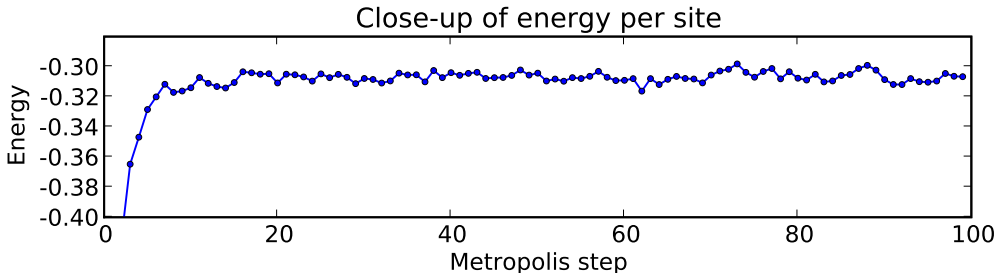
# Statistical analysis

# Statistical analysis: quantification of uncertainty

Recall what people's heights and lattice-spin energies, looked like:

Second correlated sample





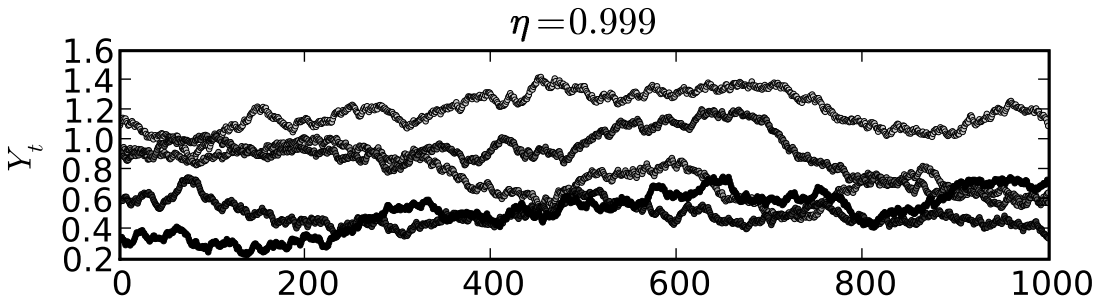

Given our time series of measurements from one experiment, we want to estimate the variation that would occur on many such experiments. Notation: $S$ is a sample (or outcome): person, spin configuration. $X(S)$ is a measurement (or random variable): height, magnetization per site. Also, abbreviate $X(S_i)$ as $X_i$.

When we learn statistics, we learn how to handle IID samples. We estimate the mean and variance of the sample mean (to get an error bar) by

$$\overline{X}_M = \frac{1}{M}\sum_{i=1}^{M} X_i \qquad \text{and} \qquad s^2_{\overline{X}_M} = \frac{1}{M(M-1)}\sum_{i=1}^{M}(X_i - \overline{X}_M)^2.$$

# Statistical analysis: quantification of uncertainty

When samples are identically distributed but *not* independent, the sample-mean calculation is still correct. But the naive error-bar estimate is far too small. Here's an example from another MCMC process, defined in appendix A of my dissertation. Here are five experiments, or time series, with pre-thermalization iterates discarded:




The sample mean computed using the topmost series would be too high, and so on. How can we quantify that without running the other four experiments?

A better estimator of the variance of the sample mean: multiply the naive variance estimator by the integrated autocorrelation time, $\hat{\tau}_{\mathrm{int}}$. This is simply the term that falls out when you compute the variance of an identically distributed but not independent time series, where correlation of $X_i$ and $X_j$ depends only on $|i - j|$.

# Statistical analysis: quantification of uncertainty

The exact integrated autocorrelation time is

$$\tau_{\text{int}} = \frac{\sigma_X^2}{M}\left[1 + 2\sum_{k=1}^{M}\left(1 - \frac{k}{M}\right)\text{Corr}(X_0, X_k)\right] \approx 1 + 2\sum_{k=1}^{\infty}\text{Corr}(X_0, X_k),$$

where the exact autocorrelation of the time series $X_1, \ldots, X_M$ is

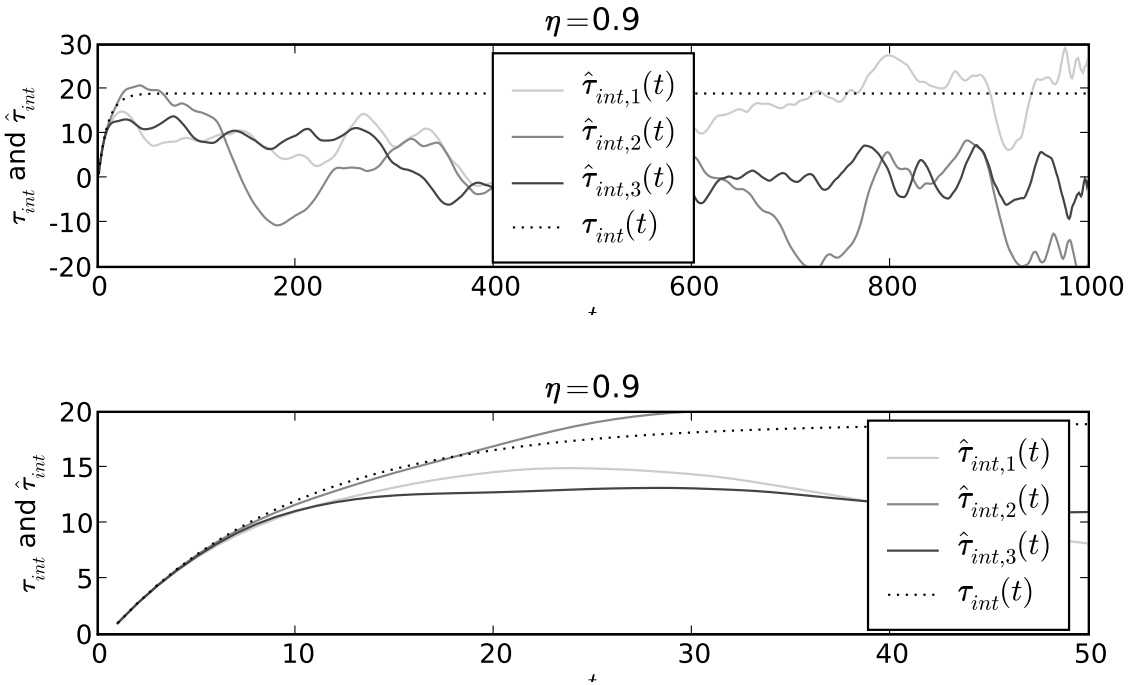
$$\text{Corr}(X_i, X_j) = \frac{\mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]}{\sigma_{X_i}\sigma_{X_j}}$$

and the expectation is over all possible experiments: $\sum_{S\in\Omega}(\ldots)$ which again are computationally intractable.

In practice, replace the $\mathbb{E}$'s and the $\sigma$'s with computations using sliding-window sums over the elements in a single time series. This turns out to be a rather wild estimator, but it's what we have and it's what we use.

# Statistical analysis: quantification of uncertainty

Here are $\hat{\tau}_{\rm int}$ computed over five experiments, plotted versus a known true $\tau_{\rm int}$ for a process I constructed especially for this purpose. (The $\eta$ is a control parameter for that process.) The second figure is a zoom-in on the first.

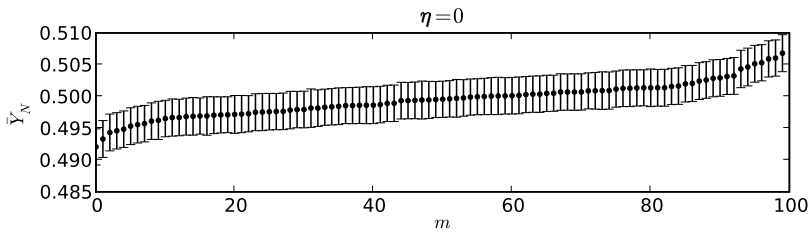


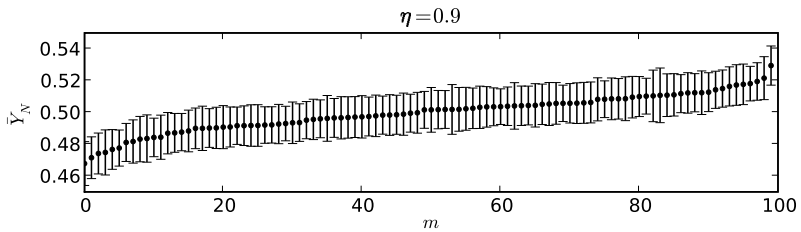

Since the $\hat{\tau}_{\rm int}$ estimator gets wilder further out, in practice one can simply look for its first flat spot.

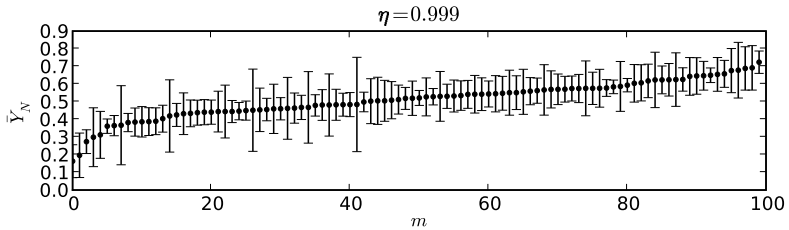
# Statistical analysis: bell-curve context

After all the computing and all the algebra, what does this finally mean in a bell-curve context? Here are the results of running 100 experiments, computing sample means and estimating error bars using the integrated-autocorrelation-time formula above, sorted by sample mean. The three plots show the results using an IID process, a weakly correlated process, and a strongly correlated process such as might occur in an MCMC simulation.

Summary: The magnitude and the variation of the error bars both increase with the amount of autocorrelation. And now you know what the error of the error bar is.



# MCMC in broader contexts

# MCMC in broader contexts

MCMC methods are also used in continuous probability distributions.

Example: Numerically integrate a function $f$ of one variable $x$ over $[a, b]$. Use Simpson's method, adaptive quadrature, etc.

Or, randomly sample points which walk around the interval $[a, b]$ with probability constrained by the height of $f$. Why bother with the latter when the former is simpler?

If you instead integrate $f(x_1, \ldots, x_{100})$ over the box $[a_1, b_1] \times \cdots \times [a_{100}, b_{100}]$, it takes $2^{100}$ function evaluations just to bracket the endpoints. Not even this can be done. Then, random sampling is necessary.

Such methods are a tool in the toolbox for many, many other contexts, including perhaps statistical inference in scientific consulting?

For more information, please visit `http://math.arizona.edu/~kerl`.

**Thank you for your time!**