Opinionated Lessons in Statistics

by Bill Press

#39 MCMC and Gibbs Sampling

Markov Chain Monte Carlo (MCMC)

Data set $\mathbf{D}$

Parameters $\mathbf{x}$ (sorry, we've changed notation!)

We want to go beyond simply maximizing $P(\mathbf{D}|\mathbf{x})$
and get the whole Bayesian posterior distribution of $\mathbf{x}$

Bayes says this is proportional to $\pi(\mathbf{x}) \equiv P(\mathbf{D}|\mathbf{x})P(\mathbf{x})$
but with an unknown proportionality constant (the Bayes denominator). It seems as if we need this denominator to find confidence regions, e.g., containing 95% of the posterior probability.

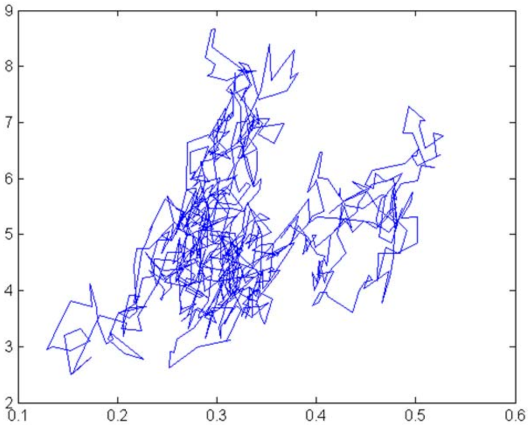
But no! MCMC is a way of drawing samples $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots$
from the distribution $\pi(\mathbf{x})$
without having to know its normalization!

With such a sample, we can compute any quantity of interest about the distribution of $\mathbf{x}$, e.g., confidence regions, means, standard deviations, covariances, etc.

Two ideas due to Metropolis and colleagues make this possible:

1. Instead of sampling unrelated points, sample a Markov chain $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots$ where each point is (stochastically) determined by the previous one by some chosen distribution $p(\mathbf{x}_i|\mathbf{x}_{i-1})$

Although locally correlated, it is possible to make this sequence *ergodic*, meaning that it visits every **x** in proportion to π(**x**).



2. Any distribution $p(\mathbf{x}_i|\mathbf{x}_{i-1})$ that satisfies

$$\pi(\mathbf{x}_1)\,p(\mathbf{x}_2|\mathbf{x}_1) = \pi(\mathbf{x}_2)\,p(\mathbf{x}_1|\mathbf{x}_2)$$

(“detailed balance”) will be such an ergodic sequence!

Deceptively simple proof: Compute distribution of x_1’s successor point

$$\int p(\mathbf{x}_2|\mathbf{x}_1)\pi(\mathbf{x}_1)\,d\mathbf{x}_1 = \pi(\mathbf{x}_2)\int p(\mathbf{x}_1|\mathbf{x}_2)\,d\mathbf{x}_1 = \pi(\mathbf{x}_2)$$

So how do we find such a $p(\mathbf{x}_i|\mathbf{x}_{i-1})$?

Metropolis-Hastings algorithm:

Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953), Hastings (1970)

1953

Pick more or less any "proposal distribution" $q(\mathbf{x}_2|\mathbf{x}_1)$
(A multivariate normal centered on $\mathbf{x}_1$ is a typical example.)

Then the algorithm is:

1. Generate a candidate point $\mathbf{x}_{2c}$ by drawing from the proposal distribution around $\mathbf{x}_1$

2. Calculate an "acceptance probability" by

$$\alpha(\mathbf{x}_1, \mathbf{x}_{2c}) = \min\left(1, \frac{\pi(\mathbf{x}_{2c})\, q(\mathbf{x}_1|\mathbf{x}_{2c})}{\pi(\mathbf{x}_1)\, q(\mathbf{x}_{2c}|\mathbf{x}_1)}\right)$$

Notice that the q's cancel out if symmetric on arguments, as is a multivariate Gaussian

3. Choose $\mathbf{x}_2 = \mathbf{x}_{2c}$ with probability α, $\mathbf{x}_2 = \mathbf{x}_1$ with probability $(1-\alpha)$

So, $$p(\mathbf{x}_2|\mathbf{x}_1) = q(\mathbf{x}_2|\mathbf{x}_1)\,\alpha(\mathbf{x}_1, \mathbf{x}_2), \qquad (\mathbf{x}_2 \neq \mathbf{x}_1)$$

It's something like: always accept a proposal that increases the probability, and sometimes accept one that doesn't. (Not exactly this because of ratio of q's.)

Proof:

$$\alpha(\mathbf{x}_1, \mathbf{x}_{2c}) = \min\left(1, \frac{\pi(\mathbf{x}_{2c})\, q(\mathbf{x}_1|\mathbf{x}_{2c})}{\pi(\mathbf{x}_1)\, q(\mathbf{x}_{2c}|\mathbf{x}_1)}\right)$$

So,

$$\begin{aligned}\pi(\mathbf{x}_1)\, q(\mathbf{x}_2|\mathbf{x}_1)\, \alpha(\mathbf{x}_1, \mathbf{x}_2) &= \min[\pi(\mathbf{x}_1)\, q(\mathbf{x}_2|\mathbf{x}_1),\ \pi(\mathbf{x}_2)\, q(\mathbf{x}_1|\mathbf{x}_2)]\\ &= \min[\pi(\mathbf{x}_2)\, q(\mathbf{x}_1|\mathbf{x}_2),\ \pi(\mathbf{x}_1)\, q(\mathbf{x}_2|\mathbf{x}_1)]\\ &= \pi(\mathbf{x}_2)\, q(\mathbf{x}_1|\mathbf{x}_2)\, \alpha(\mathbf{x}_2, \mathbf{x}_1)\end{aligned}$$

But

$$p(\mathbf{x}_2|\mathbf{x}_1) = q(\mathbf{x}_2|\mathbf{x}_1)\, \alpha(\mathbf{x}_1, \mathbf{x}_2), \qquad (\mathbf{x}_2 \neq \mathbf{x}_1)$$

and also the other way around

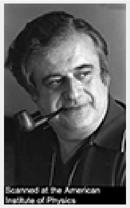
So,

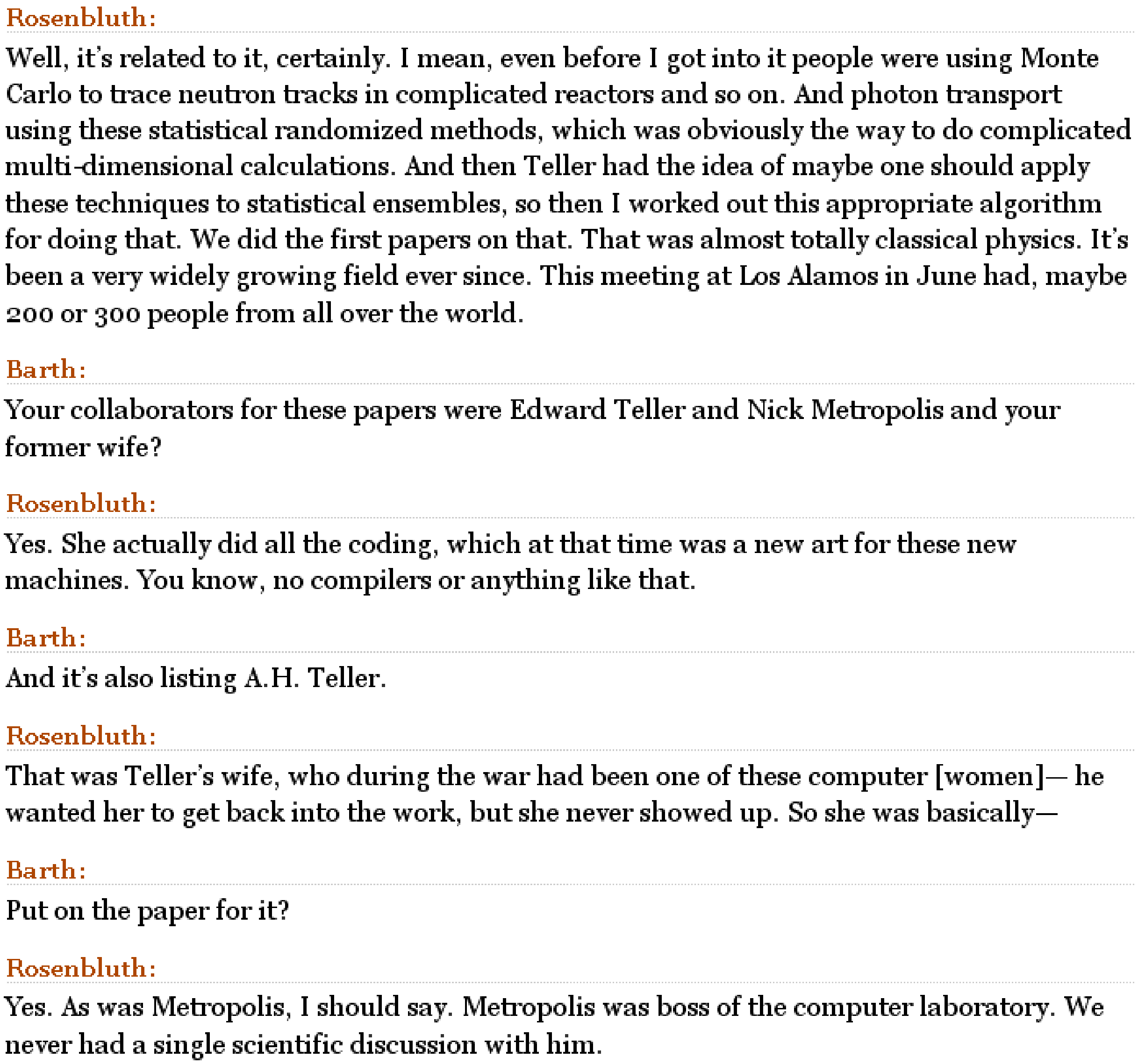
$$\pi(\mathbf{x}_1)\, p(\mathbf{x}_2|\mathbf{x}_1) = \pi(\mathbf{x}_2)\, p(\mathbf{x}_1|\mathbf{x}_2)$$

which is just detailed balance, q.e.d.

History has treated Metropolis perhaps more kindly than he deserves.

Interview with Dr. Marshall Rosenbluth
By Dr. Kai-Henrik Barth
At La Jolla, California
August 11, 2003





Rosenbluth:

Well, it's related to it, certainly. I mean, even before I got into it people were using Monte Carlo to trace neutron tracks in complicated reactors and so on. And photon transport using these statistical randomized methods, which was obviously the way to do complicated multi-dimensional calculations. And then Teller had the idea of maybe one should apply these techniques to statistical ensembles, so then I worked out this appropriate algorithm for doing that. We did the first papers on that. That was almost totally classical physics. It's been a very widely growing field ever since. This meeting at Los Alamos in June had, maybe 200 or 300 people from all over the world.

Barth:

Your collaborators for these papers were Edward Teller and Nick Metropolis and your former wife?

Rosenbluth:

Yes. She actually did all the coding, which at that time was a new art for these new machines. You know, no compilers or anything like that.

Barth:

And it's also listing A.H. Teller.

Rosenbluth:

That was Teller's wife, who during the war had been one of these computer [women]— he wanted her to get back into the work, but she never showed up. So she was basically—

Barth:

Put on the paper for it?

Rosenbluth:

Yes. As was Metropolis, I should say. Metropolis was boss of the computer laboratory. We never had a single scientific discussion with him.

The **Gibbs Sampler** is an interesting special case of Metropolis-Hastings

A “full conditional distribution” of $\pi(\mathbf{x})$ is the normalized distribution obtained by sampling along one coordinate direction (i.e. “drilling through” the full distribution. We write it as $\pi(x\,|\,\mathbf{x}^{-})$.

“given all coordinate values except one”

Theorem: A multivariate distribution is uniquely determined by all of its full conditional distributions.
Proof (sort-of): It’s a hugely overdetermined set of linear equations, so any degeneracy is infinitely unlikely!

Metropolis-Hastings along one direction looks like this:

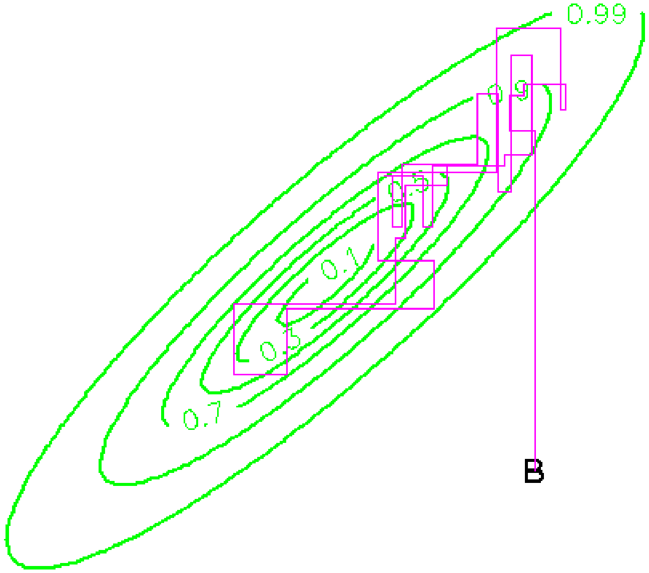
$$\alpha(x_1, x_{2c}|\mathbf{x}^{-}) = \min\left(1, \frac{\pi(x_{2c}|\mathbf{x}^{-})\,q(x_1|x_{2c}, \mathbf{x}^{-})}{\pi(x_1|\mathbf{x}^{-})\,q(x_{2c}|x_1, \mathbf{x}^{-})}\right)$$

Choose the proposal distribution $q(x_2|x_1, \mathbf{x}^{-}) = \pi(x_2|\mathbf{x}^{-})$
Then we always accept the step!

But a proposal distribution must be normalized, so we actually do need to be able to calculate $\int \pi(x|\mathbf{x}^{-})dx$ but only along one “drill hole” at a time!

So, Gibbs sampling looks like this:

- Cycle through the coordinate directions of **x**
- Hold the values of all the <u>other</u> coordinates fixed
- “Drill”, i.e., sample from the one dimensional distribution along the non-fixed coordinate.
  - this requires knowing the normalization, if necessary by doing an integral or sum along the line
- Now fix the coordinate at the sampled value and go on to the next coordinate direction.



source: Vincent Zoonekynd

Amazingly, this actually samples the full (joint) posterior distribution!

Gibbs sampling can really win if there are only a few discrete values along each drill hole.

Example: Assigning objects to clusters. Now **x** is the vector of assignments of each object. Each component of **x** has just (here) 3 possible values.

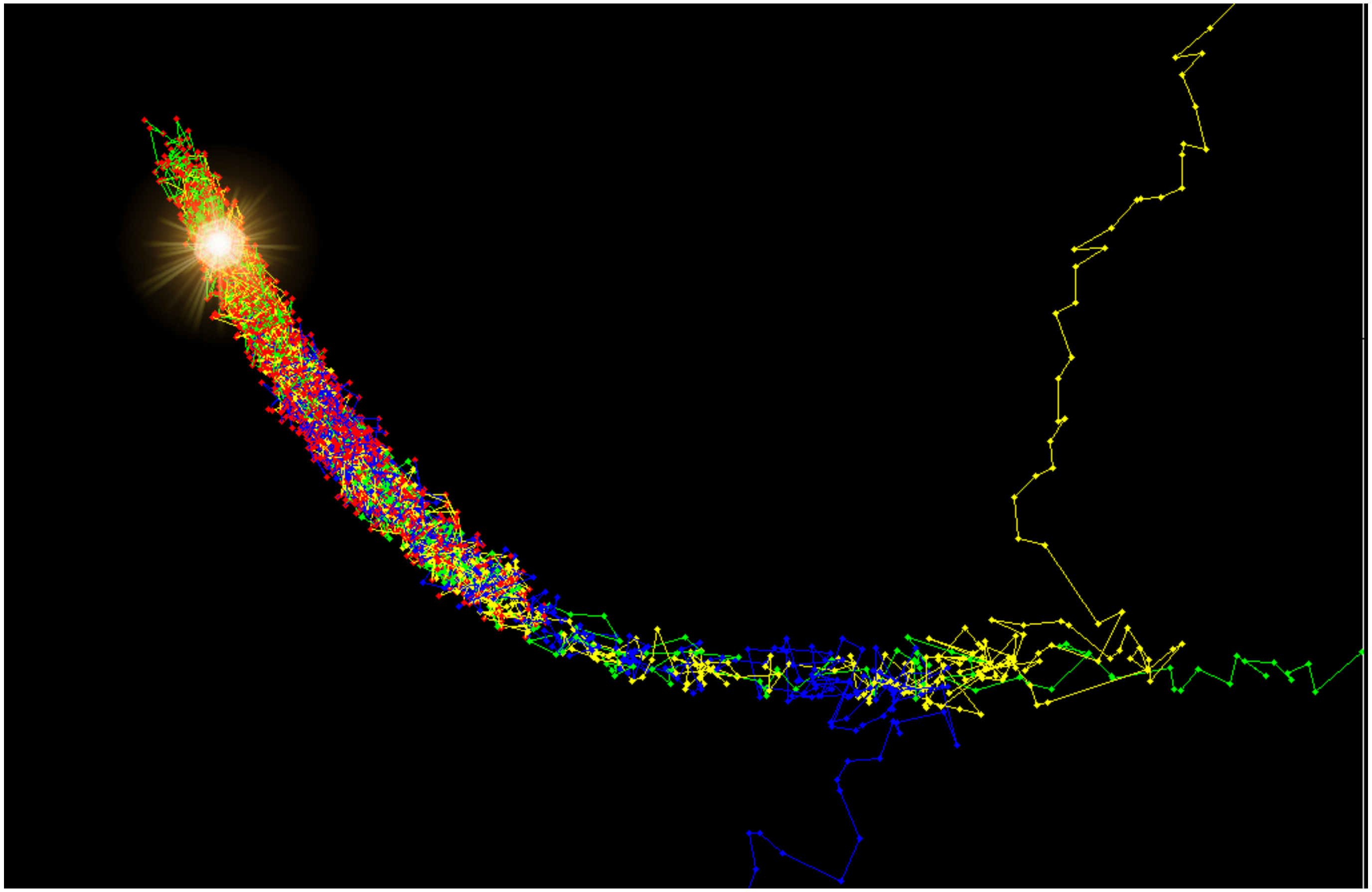
We assume of course some statistical model $\pi(\mathbf{x})$ that is the relative likelihood of a particular assignment (e.g., that all the objects in a given column are in fact drawn from that column's distribution).

Then Gibbs sampling says just cycle through the objects, reassigning each in turn by its conditional distribution.

Amazingly, this actually samples the full (joint) posterior distribution!

	Cl_1	Cl_2	Cl_3
A	o		
B			o
C			o
D	o		
E		o	
F			o
G	o		
H	o		
I			o
J		o	
K	o		

Burn-in can have multiple timescales (e.g., ascent to a ridge, travel along ridge)



Wikipedia