What is MCMC?

Jeffrey S. Rosenthal
University of Toronto
jeff@math.toronto.edu
http://probability.ca/jeff/

Some (Related!) Questions

- Medicine: How to make inferences from complicated medical studies involving many parameters (age, blood pressure, medical history, toxin levels, etc. for each patient, both before and after treatment)?

- Physics: How to understand models for physical systems involving thousands of interacting particles?

- Mathematics: How to numerically compute a very high-dimensional complicated integral?

- Why do casinos always make money?

Repeated gambling

Example: "craps". Probability of winning = 49.29%.

What happens in the long run? [APPLET]

Probability of doubling your fortune before going broke, with repeated \$10 bets at craps:

Start with \$100: 42.98%

Start with \$1,000: 5.58% (1 chance in 18)

Start with \$10,000: 1 chance in ten million billion

"Law of Large Numbers" – order from chaos.

Law of Large Numbers

Over time, slight edge leads to guaranteed victory.

Under repetition, averages converge to expected values.

Formally: if X_i is amount won/lost on i^{th} bet, then

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_i \;=\; \mathbf{E}(X_i)$$

so if lose money on average ($\mathbf{E}(X_i) < 0$), then will lose all in the long run ($\sum_{i=1}^{n} X_i \approx n\,\mathbf{E}(X_i) \to -\infty$).

Applies to gambling, investing, games, polls / surveys, "luck", traffic lights, ...

Markov chain Monte Carlo (MCMC)

Applied to complicated models / computations.

Analogy: Find average altitude of huge mountain range.

Systematic sampling of entire range too time-consuming.

Instead: explore randomly, to conduct a "mini-poll" of altitudes. Then take the sample average.

"Markov chain Monte Carlo"

Google hits: 775,000.

How does it work?

Have a target distribution $\pi(\cdot)$ that we want to sample from (e.g. uniform over a mountain range).

Starting from a state (position) X_n, we propose a new state Y_{n+1}, and then either accept it with probability α, or reject it with probability $1 - \alpha$, where

$$\alpha \;=\; \min[1,\; \pi(Y_{n+1}) \,/\, \pi(X_n)]\,.$$

Over time, it should converge . . . [APPLET]

Why Does it Work?

Key: $\pi(\cdot)$ is stationary distribution:

$$\sum_x \pi(x)\, P(x,y) \;=\; \pi(y)\,.$$

If start in $\pi(\cdot)$ then stay; otherwise converge to $\pi(\cdot)$.

Above equation holds since for $x \neq y$,

$$\begin{aligned}
\pi(x)\, P(x,y) &= \pi(x)\, Q(x,y)\, \alpha(x,y) \\
&= \pi(x)\, Q(x,y)\ \min[1,\ \pi(y)/\pi(x)] \\
&= Q(x,y)\ \min[\pi(x),\ \pi(y)] \\
&= \pi(y)\, P(y,x) \quad \text{(symmetric)}
\end{aligned}$$

and $\sum_x P(y,x) = 1$.

Example: Computing Integrals

Suppose want to compute $I \equiv \int_{\mathcal{X}} h(x)\, f(x)\, dx$, where $\mathcal{X}$ high-dimensional, and f is probability density (i.e., $f \geq 0$ and $\int_{\mathcal{X}} f(x)\, dx = 1$).

Run an MCMC algorithm having stationary distribution $\pi(dx) = f(x)\, dx$.

Then, for large B and M, $X_i \sim \pi(\cdot)$ for $i \geq B$, so

$$I \approx \frac{1}{M-B} \sum_{i=B+1}^{M} h(X_i)$$

by the Law of Large Numbers.

Example: Particle System

Suppose particle pairs contribute energy $h(dist)$.

System's overall energy is $H(\mathbf{x}) = \sum_{i<j} h(dist(x_i, x_j))$.

Probability of configuration $\mathbf{x}$ is proportional to $e^{-H(\mathbf{x})/\tau}$.

How to sample from this configuration?

Run MCMC for $\pi(d\mathbf{x}) = C\, e^{-H(\mathbf{x})/\tau}\, d\mathbf{x}$.

Works even with thousands of particles, provides <u>samples</u>.

Then can estimate mean inter-particle distance, etc.

Example: Medical Inference

Suppose have K patients, and J observations for each patient. Want to measure overall effect of treatment.

Use a complicated statistical model, e.g.

$$
\begin{array}{ccccccc}
 & & & \mu & & & \\
 & & \swarrow & \downarrow & \searrow & & \\
 & \theta_1 & & \ldots\ \ldots & & \theta_K & \\
 \swarrow & & \downarrow & & \downarrow & & \searrow \\
Y_{11}, \ldots, Y_{1J} & & & \ldots\ \ldots & & & Y_{K1}, \ldots, Y_{KJ}
\end{array}
$$

$$\theta_i \sim N(\mu, A)$$

$$Y_{ij} \sim N(\theta_i, V)$$

Run MCMC algorithm with $\pi(\cdot) =$ corresponding "posterior", then can estimate $\mathbf{E}(\mu)$, etc.

About MCMC

- It converges over the long run, just like for gambling, i.e. the Law of Large Numbers is crucial.

- MCMC only needed when more direct methods (e.g. numerical integration) infeasible due to complicated model / high dimension / limited computer speed.

- (historical) MCMC developed c. 1953, to study physical systems with many particles, using very slow computers. Then, became hugely more popular c. 1990 to study high-dimensional, complicated statistical models (esp. for medical studies).

How quickly does it converge?

To use MCMC, need time to convergence, i.e. how long to run it before black bars converge to blue bars?

Typically: Use "convergence diagnostics", to determine heuristically if MCMC has converged. For example, see if chain values appear "stationary", or if get same answer from different starting values. Problematic!

Better: Use mathematical theory to prove that, say, $|\mathbf{P}(X_B \in A) - \pi(A)| < 0.01$ for some explicit B.

But can be tricky. [Major research area . . .]

A Multitude of MCMC Algorithms

In applet example, with proposal distribution

$$\text{Uniform}\{X_n-\gamma,\ldots,X_n-1,X_n+1,\ldots,X_n+\gamma\},$$

which γ results in the "best" algorithm? [APPLET]

- If γ too small (say, $\gamma = 1$), then usually accept, but move very slowly – bad.

- If γ too large (say, $\gamma = 50$), then usually $\pi(Y_{n+1}) = 0$, i.e. hardly ever accept – bad.

- Best is a "moderate" value of γ, like 3 or 4. ["Goldilocks principle"]

Computer Learning: Adaptive MCMC

Suppose we don't know which γ is best. What to do?

Idea: Adapt, i.e. let the computer modify γ as it goes and "learn" the good values. In applet example:

Start with $\gamma = 2$ (say).

Each time proposal is accepted, increase γ.

Each time proposal is rejected, decrease γ (to min of 1).

Logical, natural adaptive scheme, which uses the computer to perform a "search" for a good γ, on the fly.

But does it work?? [APPLET]

About Adaptive MCMC

- We see that naive adaption can ruin the algorithm, and fail to converge to $\pi(\cdot)$.

- Hence, even obvious-seeming extensions of computer algorithms can go horribly wrong in the absence of theoretical justification; theory is important.

- Theorem: Adaptive MCMC will converge to $\pi(\cdot)$ provided it satisfies the Diminishing Adaptation property, e.g. only adapt with probability $p(n) \to 0$.

- So, can use adaption as long as your careful. Some successes on high-dimensional problems. Hopefully more in future as computers get faster (probability.ca/amcmc).

Summary

Law of Large Numbers creates order from chaos: averages converge to their expected values (e.g. gambling).

Can use this for scientific computation: MCMC.

MCMC runs a Markov chain (random process) which converges to the distribution of interest.

Can then use samples to draw inferences.

Time to convergence is a major research area.

Adaptive MCMC tries to get computer to help choose.

- Papers, applets, software: `probability.ca`