Optimising and Adapting Metropolis Algorithms

Jeffrey S. Rosenthal
University of Toronto

jeff@math.toronto.edu
http://probability.ca/jeff/

(LMS/CRiSM Summer School, Warwick, July 2018)

(Brief) Background / Context / Motivation

Often have complicated, high-dimensional density functions $\pi : \mathcal{X} \to [0, \infty)$, for some $\mathcal{X} \subseteq \mathbf{R}^d$ with d large.

(e.g. Bayesian posterior distribution)

Want to compute probabilities like:

$$\Pi(A) \;:=\; \int_A \pi(x)\, dx\,,$$

and/or expected values of functionals like:

$$\mathbf{E}_\pi(h) \;:=\; \int_{\mathcal{X}} h(x)\, \pi(x)\, dx\,.$$

Or, if π is unnormalised:

$$\mathbf{E}_\pi(h) \;:=\; \int_{\mathcal{X}} h(x)\, \pi(x)\, dx \Big/ \int_{\mathcal{X}} \pi(x)\, dx\,.$$

Calculus? Numerical integration?

Impossible, if π is something like ...

Typical π: Variance Components Model

State space $\mathcal{X} = (0, \infty)^2 \times \mathbf{R}^{K+1}$, so $d = K + 3$, with

$$\begin{aligned}
&\pi(V, W, \mu, \theta_1, \ldots, \theta_K) \\
= \; & C\, e^{-b_1/V} V^{-a_1-1} e^{-b_2/W} W^{-a_2-1} \\
& \times e^{-(\mu-a_3)^2/2b_3} V^{-K/2} W^{-\frac{1}{2}\sum_{i=1}^K J_i} \\
\times \; & \exp\left[-\sum_{i=1}^K (\theta_i - \mu)^2/2V - \sum_{i=1}^K \sum_{j=1}^{J_i} (Y_{ij} - \theta_i)^2/2W \right],
\end{aligned}$$

where a_i and b_i are fixed constants (prior), and $\{Y_{ij}\}$ are the data.

In the application: $K = 19$, so $d = 22$.

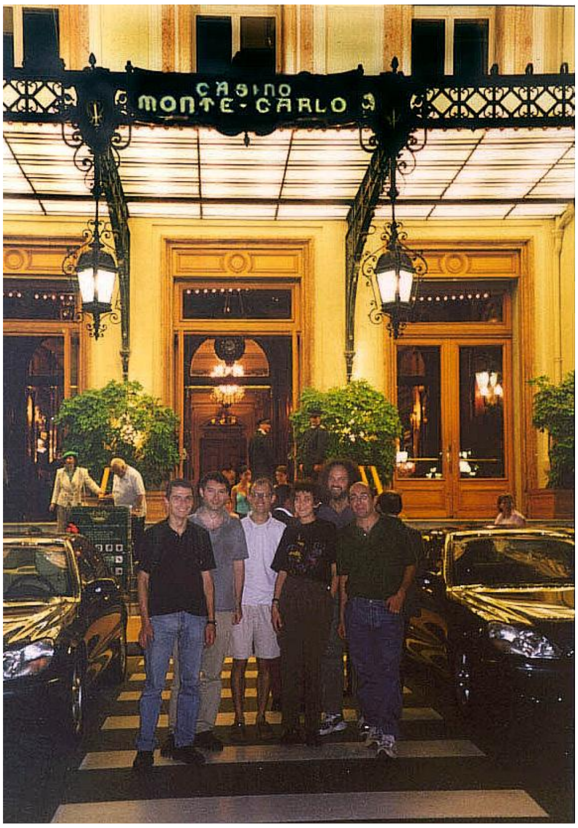
Integrate? Well, no problems *mathematically*, but ...

High-dimensional! Complicated! How to compute?

Try Monte Carlo!

Monte Carlo, Monaco

Nice Place for a Conference!



Estimation from sampling: Monte Carlo

Suppose we can sample from π, i.e. generate on a computer

$$X_1, X_2, \ldots, X_M \sim \pi \quad (i.i.d.)$$

(i.e., $\mathbf{P}(X_i \in A) = \int_A \pi(x)\, dx$ for each i, and independent).

Then can estimate by e.g.

$$\mathbf{E}_\pi(h) \;\approx\; \frac{1}{M} \sum_{i=1}^{M} h(X_i)\,.$$

As $M \to \infty$, the estimate converges to $\mathbf{E}_\pi(h)$ (by the Law of Large Numbers), which good error bounds / confidence intervals (by the Central Limit Theorem).

Good. But how to sample from π?

Often infeasible! (e.g. above example!)

Instead . . .

Markov Chain Monte Carlo (MCMC)

Given a complicated, high-dimensional target distribution $\pi(\cdot)$:

Find an ergodic Markov chain (random process) $X_0, X_1, X_2, \ldots$, which is easy to run on a computer, and which converges in distribution to π as $n \to \infty$.

Then for "large enough" B, $\mathcal{L}(X_B) \approx \pi$, so X_B, X_{B+1}, ... are approximate samples from π, and e.g.

$$\mathbf{E}_\pi(h) \approx \frac{1}{M} \sum_{i=B+1}^{B+M} h(X_i), \quad \text{etc.}$$

Extremely popular: Bayesian inference, computer science, statistical genetics, statistical physics, finance, insurance, ...

But how to create such a Markov chain?

Random-Walk Metropolis Algorithm (1953)

This algorithm defines the chain $X_0, X_1, X_2, \ldots$ as follows.

Given X_{n-1}:

- Propose a new state $Y_n \sim Q(X_{n-1}, \cdot)$, e.g. $Y_n \sim N(X_{n-1}, \Sigma_p)$.
- Let $\alpha = \min\left[1, \frac{\pi(Y_n)}{\pi(X_{n-1})}\right]$. (Assuming Q is symmetric.)
- With probability α, accept the proposal (set $X_n = Y_n$).
- Else, with prob. $1-\alpha$, reject the proposal (set $X_n = X_{n-1}$).

Try it: [APPLET] Converges to π!

Why? α is chosen just right so this Markov chain is reversible with respect to π, i.e. $\pi(dx)\, P(x, dy) = \pi(dy)\, P(y, dx)$. Hence, π is a stationary distribution. Also, chain will be aperiodic and (usually) irreducible. So, by general Markov chain theory, it converges to π in total variation distance: $\lim_{n\to\infty} \sup_A |\mathbf{P}(X_n \in A) - \pi(A)| = 0$.

More complicated example?

Example: Particle Systems

Suppose have n independent particles, each uniform on a region.

What is, say, the average "diameter" (maximal distance)?

Sample and see! [pointproc.java] Works! Monte Carlo!

Now suppose instead that the particles are not independent, but rather interact with each other, with the configuration probability proportional to e^{-H}, where H is an energy function, e.g.

$$H = \sum_{i<j} A\Big|(x_i, y_i) - (x_j, y_j)\Big| + \sum_{i<j} \frac{B}{\Big|(x_i, y_i) - (x_j, y_j)\Big|} + \sum_i C\, x_i$$

A large: particles like to be close together.
B large: particles like to be far apart.
C large: particles like to be towards the left.

Can't directly sample, but can use Metropolis! [pointproc.java]

Okay, but Where's the Math?

MCMC's greatest successes have been in ... applications!

- Medical Statistics / Statistical Genetics / Bayesian Inference / Chemical Physics / Computer Science / Mathematical Finance

So, what is MCMC mathematical theory good for?

- Informs and justifies the basic algorithms.
  (** Above Introduction)
- Quantifies how well the algorithms work.
  (** Quantitative Bounds)
- Suggests new modifications of the algorithms.
- Determines which algorithm choices are best.
  (** Optimal Scaling)
- Investigates high-dimensional behaviour. (** Complexity)
- Develops new MCMC directions. (** Adaptive MCMC)

First Topic: Quantitative Convergence Bounds

MCMC works eventually, i.e. $\mathcal{L}(X_n) \Rightarrow \pi$. Good!

But what about quantitative bounds, i.e. a specific number n_* such that, say, $|\mathbf{P}(X_{n_*} \in A) - \pi(A)| < 0.01 \quad \forall\, A$?

(Not just "as $n \to \infty$".)

One method: coupling. (Many other methods: spectral, ...)

Consider two copies of the chain, $\{X_n\}$ and $\{X'_n\}$.

Assume that $X'_0 \sim \pi$ (so $X'_n \sim \pi$ $\forall n$).

If we can "make" the two copies become equal for $n \geq T$, while respecting their marginal update probabilities, then $X_n \approx \pi$ too.

Specifically, the coupling inequality says:

$$|\mathbf{P}(X_n \in A) - \pi(A)| \;\equiv\; |\mathbf{P}(X_n \in A) - \mathbf{P}(X'_n \in A)| \;\leq\; \mathbf{P}(T > n)\,.$$

But how to apply this to a complicated MCMC algorithm?

Quantitative Bounds: Minorisation

Suppose there is $\epsilon > 0$, and a probability measure ν, such that $P(x,y) \geq \epsilon\, \nu(y)$ for all $x, y \in \mathcal{X}$.

This "minorisation condition" gives an ϵ-sized "overlap" between the transition distributions $P(x,\cdot)$ and $P(x',\cdot)$.

That means at each iteration, we can make the two copies become equal with probability at least ϵ. Hence, $\mathbf{P}(T > n) \leq (1-\epsilon)^n$.

Therefore, $|\mathbf{P}(X_n \in A) - \pi(A)| \leq \; (1-\epsilon)^n, \;\; \forall\, A$.

e.g. [APPLET], with that π, and $\gamma = 3$: find that $P(x,y) \geq \epsilon\; \nu(y)$ for all x, y, where $\epsilon = 0.2$, and $\nu(3) = \nu(4) = 1/2$.

- So $|P^n(x,A) - \pi(A)| \leq (1-\epsilon)^n = (1-0.2)^n = (0.8)^n$.
- Hence, $|P^n(x,A) - \pi(A)| < 0.01$ whenever $n \geq 21$.
- So $n_* = 21$. "The chain converges in 21 iterations." Good!

But what about a harder example??

Example: Baseball Data Model

Hierarchical model for baseball hitting percentages (J. Liu): observed hitting percentages satisfy $Y_i \sim N(\theta_i, c)$ for $1 \leq i \leq K$, where $\theta_1, \ldots, \theta_k \sim N(\mu, V)$, c is a given constant, with $V, \mu, \theta_1, \ldots, \theta_K$ to be estimated. Priors: $\mu \sim$ flat, $V \sim IG(a, b)$.

Diagram:

$$
\begin{array}{ccccccc}
 & & \mu & & & & \\
 & \swarrow & \downarrow & \searrow & & & \\
\theta_1 & \ldots & & \ldots & \theta_K & \qquad & \theta_i \sim N(\mu, V) \\
\downarrow & & & & \downarrow & & \\
Y_1 & \ldots & & \ldots & Y_K & \qquad & Y_i \sim N(\theta_i, c)
\end{array}
$$

For our data, $K = 18$, so dimension $= 20$.

High dimensional! How to estimate?

Baseball Data Model (cont'd)

MCMC solution: Run a Gibbs sampler for π.

Markov chain is $X_k = (V^{(k)}, \mu^{(k)}, \theta_1^{(k)}, \ldots \theta_K^{(k)})$, updated by:

$$V^{(n)} \ \sim \ IG\left(a + \frac{K-1}{2},\ b + \frac{1}{2}\sum(\theta_i^{(n-1)} - \overline{\theta}^{(n-1)})^2\right) ;$$

$$\mu^{(n)} \ \sim \ N\left(\overline{\theta}^{(n-1)},\ \frac{V^{(n)}}{K}\right) ;$$

$$\theta_i^{(n)} \ \sim \ N\left(\frac{\mu^{(n)}c + Y_i V^{(n)}}{c + V^{(n)}},\ \frac{V^{(n)}c}{c + V^{(n)}}\right) \quad (1 \leq i \leq K) ;$$

where $\overline{\theta}^{(n)} = \frac{1}{K}\sum \theta_i^{(n)}$.

Recall that $K = 18$, so dimension $= 20$.

Complicated! How to analyze convergence?

Example: Baseball Data Model (cont'd)

Here we can find a minorisation $P(x,y) \geq \epsilon\,\nu(y)$, but only when $x \in C$ for a subset $C \subseteq \mathcal{X}$. ("small set")

But also find a "drift condition" $\mathbf{E}[f(X_1)\,|\,X_0 = x] \leq \lambda f(x) + \Lambda$, for some $\lambda < 1$ and $\Lambda < \infty$, where $f(x) = \sum_{i=1}^{K}(\theta_i - \overline{Y})^2$; this "forces" returns to $C \times C$.

Can compute (R., Stat & Comput. 1996):

- a drift condition towards $C = \left\{\sum_i(\theta_i - \overline{Y})^2 \leq 1\right\}$, with $\lambda = 0.000289$ and $\Lambda = 0.161$;
- a minorisation with $\epsilon = 0.0656$, at least for $x \in C \subseteq \mathcal{X}$.

Then can use coupling to prove (R., JASA 1995) that

$$|\mathbf{P}(X_n \in A) - \pi(A)| \;\leq\; (0.967)^n + (1.17)(0.935)^n\,, \quad n \in \mathbf{N}\,,$$

so e.g. $|\mathbf{P}(X_n \in A) - \pi(A)| < 0.01$ if $n \geq 140$.

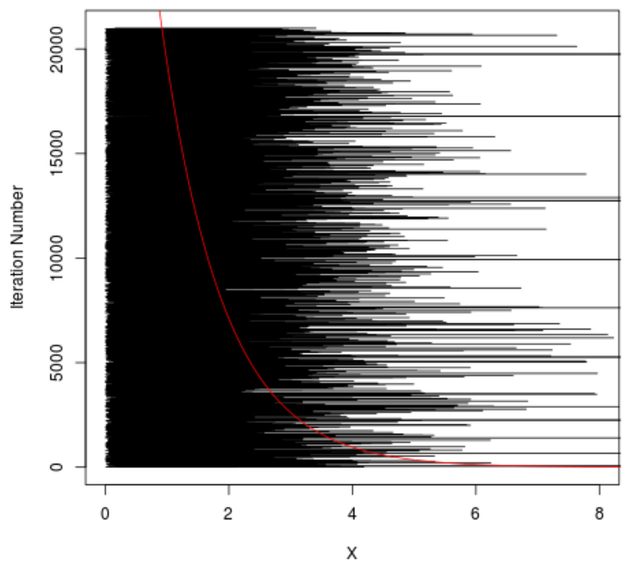
- So $n_* = 140$. "The chain converges in 140 iterations." Good!

Realistic bounds for complicated statistical models!
(See also Jones & Hobert, Stat Sci 2001, . . .)

Does it Matter? Case Study: Independence Sampler

Consider Metropolis-Hastings where $\pi(x) = e^{-x}$, and proposals are chosen i.i.d. $\sim \mathrm{Exp}(k)$ with density ke^{-ky}, for some $k > 0$.

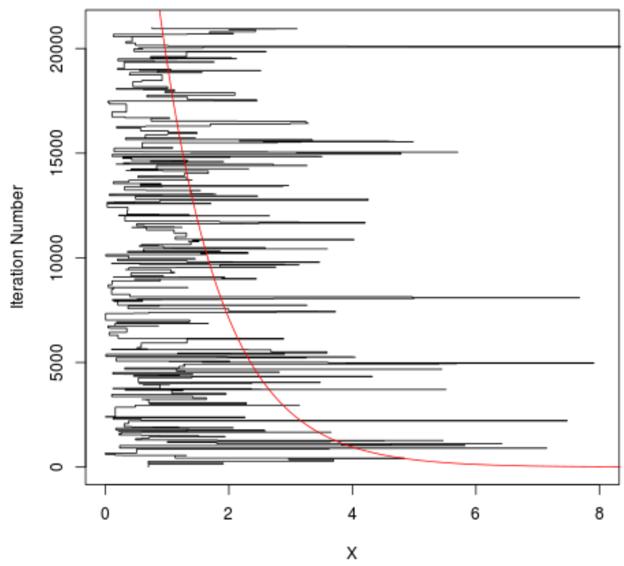
- $k = 1$ (i.i.d. sampling)



$\mathbf{E}(X) = 1$; estimate $= 1.001$. Excellent! Other k?

Independence Sampler (cont'd)

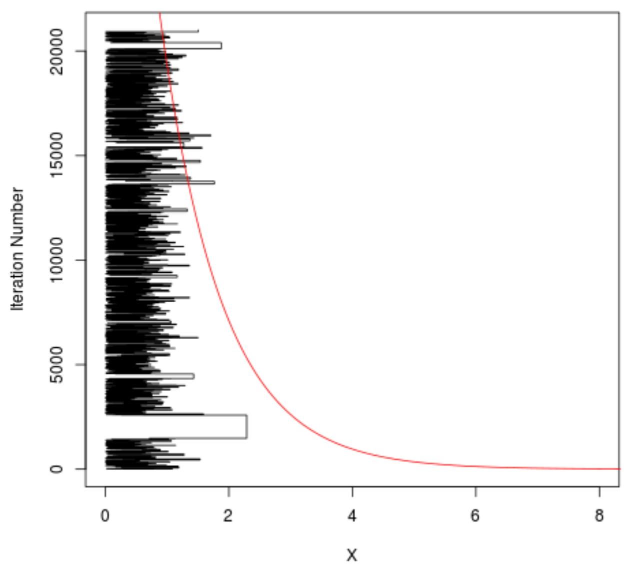
- $k = 0.01$



$\mathbf{E}(X) = 1$; estimate $= 0.993$. Quite good.

Independence Sampler (cont'd)

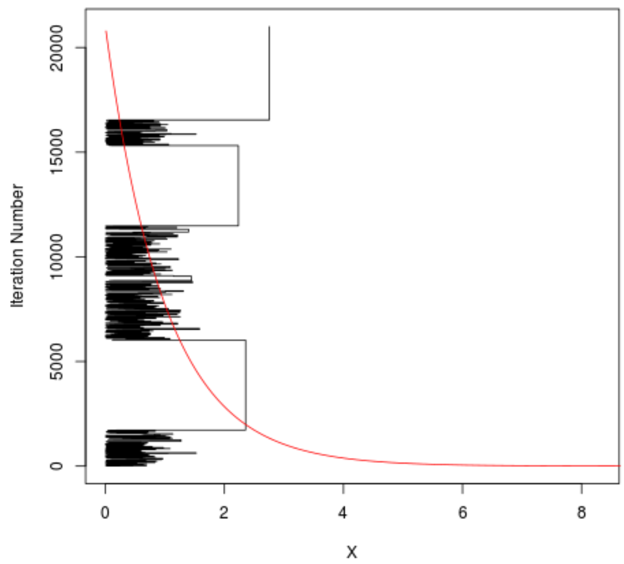
- $k = 5$



$\mathbf{E}(X) = 1$; estimate $= 0.687$. Terrible: way too small!

What happened? Maybe we just got unlucky? Try again!

- Another try with $k = 5$:



$\mathbf{E}(X) = 1$; estimate $= 1.696$. Terrible: way too big!

So, not just bad luck: $k = 5$ is really bad. But why??

Independence Sampler: Theory

Why is $k = 0.01$ pretty good, and $k = 5$ so terrible?

Well, if $k \leq 1$, then $\forall x$, $q(x) = ke^{-kx} \geq ke^{-x} = k\pi(x)$. Then

$$\alpha(x,y) = \min(1, \frac{\pi(y)\,q(x)}{\pi(x)\,q(y)}) = \min(1, \frac{\pi(y)/q(y)}{\pi(x)/q(x)})$$

$$\geq \min(1, \frac{\pi(y)/q(y)}{(1/k)}) = k\,(\pi(y)/q(y)).$$

Then $P(x,y) \geq q(y)\,\alpha(x,y) \geq k\,\pi(y)$. Minorisation with $\epsilon = k$!

So, $|P^n(x,A) - \pi(A)| \leq (1-k)^n$.

- $k = 1$: yes, $\epsilon = 1$; converges <u>immediately</u> (of course). $n_* = 1$.
- $k = 0.01$: yes, $\epsilon = 0.01$; and $(1-0.01)^{459} < 0.01$, so $n_* = 459$; "chain converges within 459 iterations". (Pretty good.)
- $k = 5$: no such ϵ. <u>Not</u> geometrically ergodic. In fact, we can prove (Roberts and R., MCAP, 2011) that with $k = 5$, have $4,000,000 \leq n_* \leq 14,000,000$, i.e. takes <u>millions</u> of iterations!

Main Topic: How to Optimise MCMC Choices?

In theory, MCMC works with essentially any update rules, as long as they leave π stationary.

- Any symmetric proposal distribution Q. (Choices!)
- Non-symmetric proposals, with a suitably modified acceptance probability. ("Metropolis-Hastings") (e.g. Independent, Langevin)
- Update one coordinate at a time. ("Componentwise")
- Update from full conditional distributions. ("Gibbs Sampler")

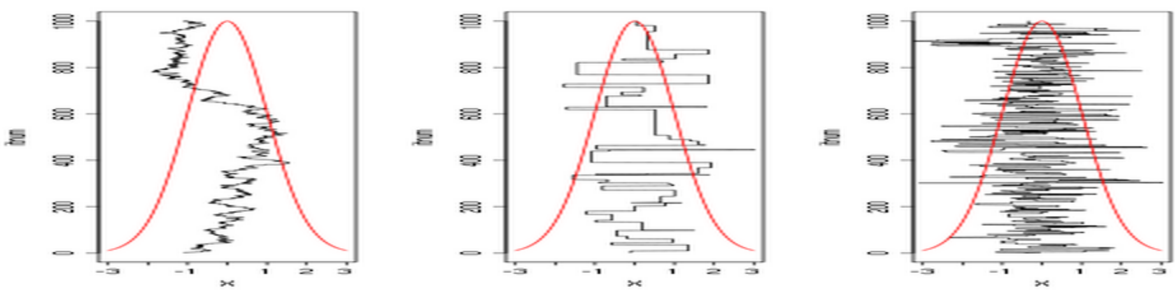
But what choice works best? e.g. What γ in [APPLET]?

- 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 γ is between the two extremes, i.e. acceptance rate should be far from 0 and far from 1. ("Goldilocks Principle")

Example: Metropolis for N(0,1)

Target $\pi = N(0,1)$. Proposal $Q(x,\cdot) = N(x,\sigma^2)$.

How to choose σ? Big? Small? What acceptance rate (A.R.)?



$\sigma = 0.1$?	$\sigma = 25$?	$\sigma = 2.38$?
too small!	too big!	just right!
A.R. $= 0.962$	A.R. $= 0.052$	A.R. $= 0.441$

The Goldilocks Principle in action!

What about higher-dimensional examples? If d increases, then σ should: decrease. But how quickly? On what scale? Theory?