MCMC Theory: What is it Good For?

Jeffrey S. Rosenthal
University of Toronto

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

(MCMSki III, Utah, Jan. 5, 2011)

Introduction

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

So, what is MCMC theory good for?

Whitfield and Strong (Motown, 1969):

> War
> What is it good for?
> Absolutely nothin'!

Rosenthal (MCMSki, 2011):

> MCMC theory
> What is it good for?
> Perhaps a little somethin'!

Everyone uses SOME theory!

e.g. Metropolis-Hastings algorithm:

Given X_{n-1}, propose $Y_n \sim q(X_{n-1}, \cdot)$, accept with prob.

$$\min\left(1\,,\ \frac{\pi(Y_n)\ q(Y_n, X_{n-1})}{\pi(X_{n-1})\ q(X_{n-1}, Y_n)}\right)\,.$$

Why? To guarantee reversibility, i.e.

$$\pi(dx)\, P(x, dy) \ = \ \pi(dy)\, P(y, dx)\,,$$

This in turn guarantees that $\pi(\cdot)$ is a stationary distribution, i.e. $\int \pi(dx)\, P(x, A) = \pi(A)$.

Then, assuming irreducibility and aperiodicity, this guarantees ergodicity, i.e.

$$\lim_{n\to\infty} \mathbf{P}(X_n \in A) \;=\; \pi(A)\,.$$

And also laws of large numbers (LLN), e.g.

$$\lim_{M\to\infty} \frac{1}{M}\sum_{n=1}^{M} h(X_n) \;=\; \pi(h) \;:=\; \int h(y)\,\pi(dy)\,.$$

So, everybody uses this much theory!

(Tierney 1994, etc.)

But what about other theory?

Very Simple Running Example (Java Applet)

$\pi(\cdot)$ simple distribution on $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$:
$\pi(1) = \pi(3) = \pi(5) = 0.15$, $\pi(2) = 0.09$, $\pi(4) = \pi(6) = 0.23$

Do “random-walk Metropolis” (RWM):

For some fixed $\gamma \in \mathbf{N}$,

- Given X_n, first propose a state $Y_{n+1} \in \mathbf{Z}$, with $Y_{n+1} \sim \text{Uniform}\{X_n - \gamma, \ldots, X_n - 1, X_n + 1, \ldots, X_n + \gamma\}$.
- Then, with probability $\min[1,\ \pi(Y_{n+1})/\pi(X_n)]$, accept proposal and set $X_{n+1} = Y_{n+1}$.
- Otherwise, with probability $1 - \min[1,\ \pi(Y_{n+1})/\pi(X_n)]$, reject proposal and set $X_{n+1} = X_n$. [APPLET]

Tuning/Optimizing the Algorithm

This works, i.e. $\mathcal{L}(X_n) \to \pi(\cdot)$. (By reversibility!)

But should γ be 2, or 1, or 50, or … ?

- 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.

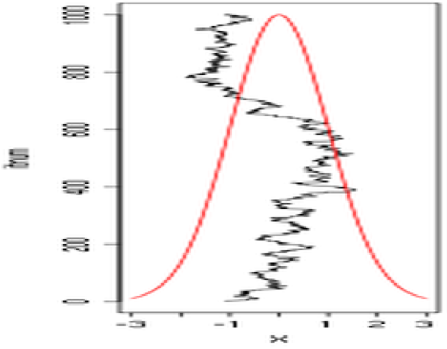
(“Goldilocks Principle”)

Obvious in this example.

Other examples??

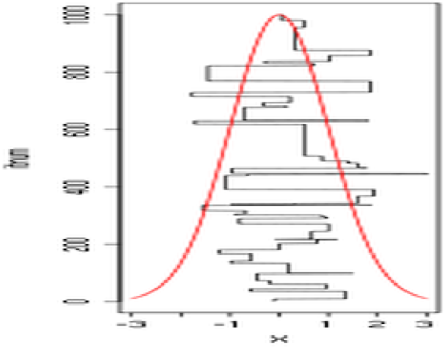
Example #2: N(0,1)

Target $\pi(\cdot) = N(0,1)$. Proposal $Q(x,\cdot) = N(x,\sigma^2)$.
How to choose σ?



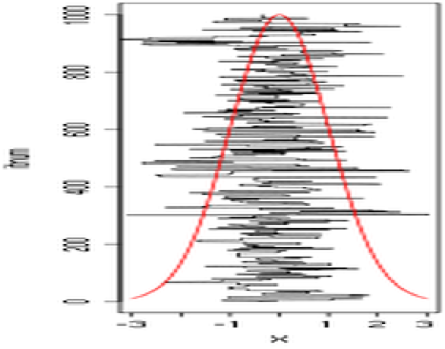
$\sigma = 0.1$?
too small!

A.R. $= 0.962$



$\sigma = 25$?
too big!

A.R. $= 0.052$



$\sigma = 2.38$?
(better!)

A.R. $= 0.441$

What about higher dimensions? (Need smaller σ?)

How to make theoretical progress?

Consider diffusion limits!

Analogy: if $\{X_n\}$ is simple random walk, and $Z_t = d^{-1/2} X_{dt}$ (i.e., we speed up time, and shrink space), then as $d \to \infty$, the process $\{Z_t\}$ converges to Brownian motion.

Theorem [Roberts, Gelman, Gilks, AAP 1997]:
If $\{X_n\}$ is a Metropolis algorithm in high dimension d, with $Q(x, \cdot) = N(x, \frac{\ell^2}{d} I_d)$, and $Z_t = d^{-1/2} X^{(1)}_{dt}$, then under "certain conditions" on $\pi(\cdot)$, the process $\{Z_t\}$ converges to a diffusion, whose speed $h(\ell)$ is explicitly related to its asymptotic acceptance rate $A(\ell)$.

Lots of information here!

- The speed $h(\ell)$ is related to the acceptance rate $A(\ell)$.
- To optimize the algorithm, we should maximize $h(\ell)$.
- The maximization is easy: $\ell_{opt} \doteq 2.38/C_\pi$.
- Then we can compute that: $A(\ell_{opt}) \doteq 0.234$.

So, for $Q(x,\cdot) = N(x, \sigma^2 I_d)$, it is <u>optimal</u> to choose

$$\sigma^2 \;=\; \frac{\ell_{opt}^2}{d} \;=\; \frac{(2.38)^2}{(C_\pi)^2 d},$$

which leads to an acceptance rate of 0.234.

Clear, simple rule. Good! Useful!

Generalizations to Langevin diffusions, other targets, etc.
(Roberts & R., JRSSB 1998; Bédard & R., CJS 2008)

How Quickly Does Java Applet Example Converge?

Use the "minorization condition" approach . . .
Take the case $\gamma = 3$ (say).

Then for all $x \in \mathcal{X}$ with $x \neq 3$,

$$P(x,3) = Q(x,3)\ \min(1, \pi(3)/\pi(x)) \geq (1/6)(0.15/0.23) > 0.1\,.$$

Also $P(3,3) \geq Q(3,0) = 1/6 > 0.1$.

Similarly $P(x,4) > 0.1$ for all $x \in \mathcal{X}$.

Conclusion: $P(x,y) \geq \epsilon\ \nu(y)$ for all $x, y \in \mathcal{X}$, where $\epsilon = 0.2$,
and $\nu(3) = \nu(4) = 1/2$ and $\nu(x) = 0$ otherwise.

"Minorization Condition"

How does this “Minorization Condition” help?

Theorem: if $P(x,y) \geq \epsilon\, \nu(y)$ for all $x, y \in \mathcal{X}$, then

$$\sup_A |P^n(x,A) - \pi(A)| \;\leq\; (1-\epsilon)^n\,.$$

In the above example,

$$\sup_A |P^n(x,A) - \pi(A)| \;\leq\; (1-\epsilon)^n = (1-0.2)^n = (0.8)^n\,,$$

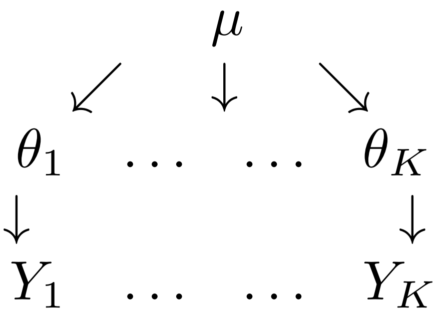
so e.g. $|P^n(x,A) - \pi(A)| < 0.01$ whenever $n \geq 21$.

“The chain converges in 21 iterations.”

What about a harder example??

Example: Baseball Data Model

Model for baseball hitting percentages (J. Liu):



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

$Y_i \sim N(\theta_i, c)$

where $\{Y_i\}$ are observed hitting percentages, c is empirically estimated, and $\mu, A, \theta_1, \ldots, \theta_K$ are to be estimated.

Priors: $\mu \sim$ flat, $A \sim IG(a, b)$. $K = 18$, dim $= 20$.

Run a Gibbs sampler on this model.

Time to convergence??

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

- a minorization with $\epsilon = 0.0656$, at least for $x \in C \subseteq \mathcal{X}$;
- a corresponding "drift condition" back to C;

where $C = \left\{ \sum_i (\theta_i - \overline{Y})^2 \le 1 \right\}$.

Putting these two conditions together, can prove that

$$\sup_A |P^n(x, A) - \pi(A)| \;\le\; (0.967)^n + (1.17)(0.935)^n \,,$$

so e.g. $|P^n(x, A) - \pi(A)| < 0.01$ if $n \ge 140$.

"The chain converges in 140 iterations."

Realistic models/bounds! (Jones & Hobert, Stat Sci 2001)

But too tricky for everyday use . . . what else?

Geometric Ergodicity (qualitative convergence)

DEFN: Say the chain is geometrically ergodic if

$$\|P^n(x,\cdot) - \pi(\cdot)\| \;\leq\; C(x)\,\rho^n\,, \qquad n = 1, 2, 3, \ldots$$

for some $\rho < 1$, where $C(x) < \infty$ for π-a.e. $x \in \mathcal{X}$.
i.e., distance to stationarity decreases exponentially quickly (at some exponential rate).

Intuitively, if the chain is geometrically ergodic, then it "probably converges quickly" in some sense.

Always holds on finite state spaces (e.g. Java example).

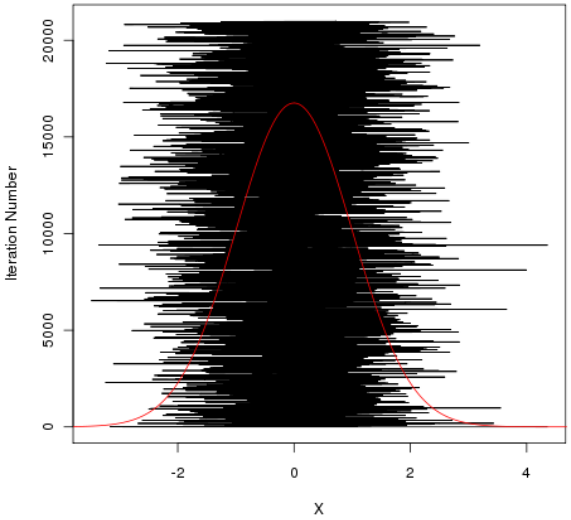
But on unbounded state spaces, may or may not hold.

Does this qualitative property actually matter??

Example #2 again: RWM for N(0,1)

RWM for $\pi(\cdot) = N(0,1)$, with $Q(x,\cdot) = N(x,\sigma^2)$, where σ is chosen to make A.R. $\doteq 0.234$.

Works well:

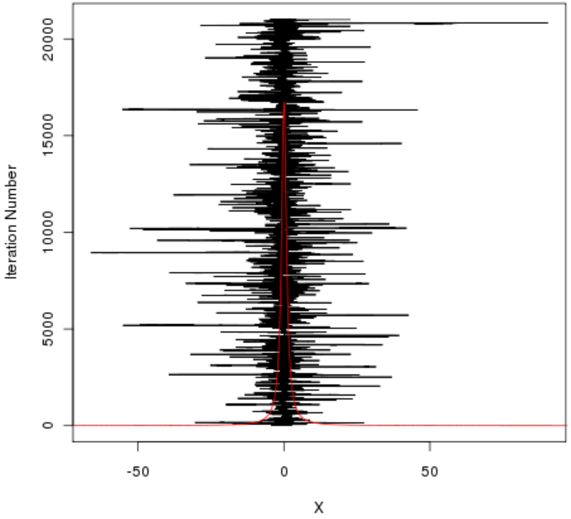


$\mathbf{P}(|X| > 2) \doteq 0.0455$; estimate $= 0.0453$. Good!

Example #2b: RWM for Cauchy

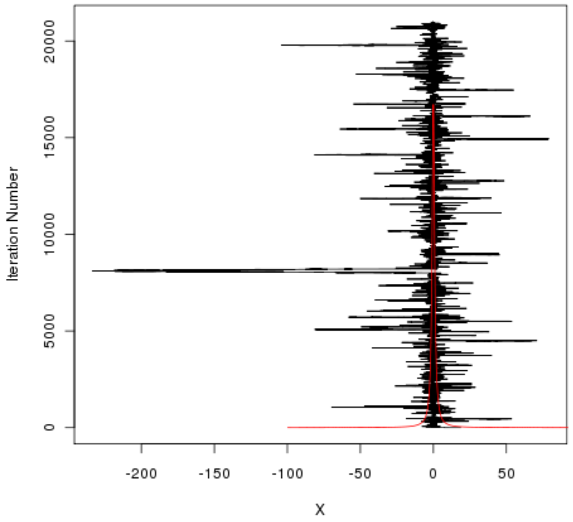
RWM for $\pi(x) = \frac{c}{1+x^2}$ (Cauchy), with $Q(x,\cdot) = N(x,\sigma^2)$, with σ again chosen to make A.R. $\doteq 0.234$.

Much worse!



$\mathbf{P}(|X| > 10) \doteq 0.0635$; estimate $= 0.0469$. Way too small!

Another try:



$\mathbf{P}(|X| > 10) \doteq 0.0635$; estimate $= 0.0746$. Way too big!

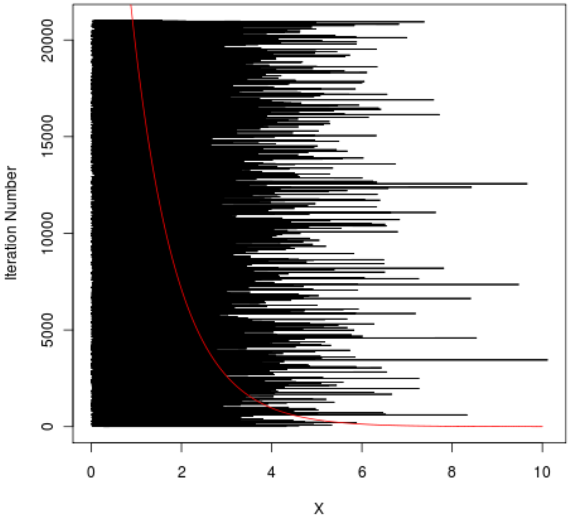
Theorem: RWM is geometrically ergodic if and only if $\pi(\cdot)$ has exponentially-small tails. [N(0,1): yes; Cauchy: no.] (Mengersen-Tweedie-Roberts, 1996)

It matters!

Example #3: Independence sampler

Independence sampler for $\pi(x) = e^{-x}$, with proposal $q(y) = ke^{-ky}$ for various possible choices of k:

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

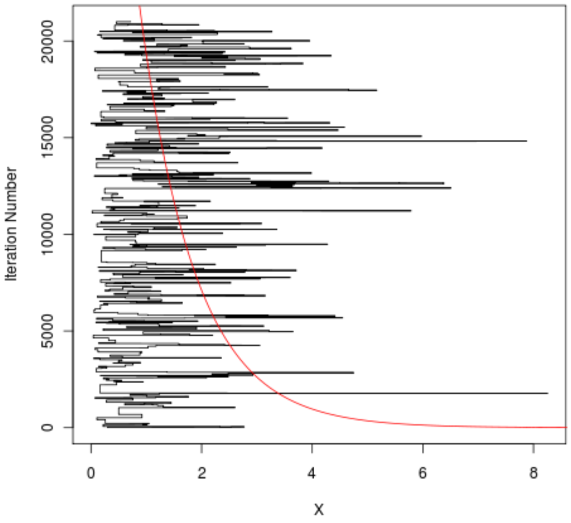


$\mathbf{E}(X) = 1$; estimate $= 0.9932$. Excellent!

Independence sampler (cont'd)

What about other values of k?

- $k = 0.01$

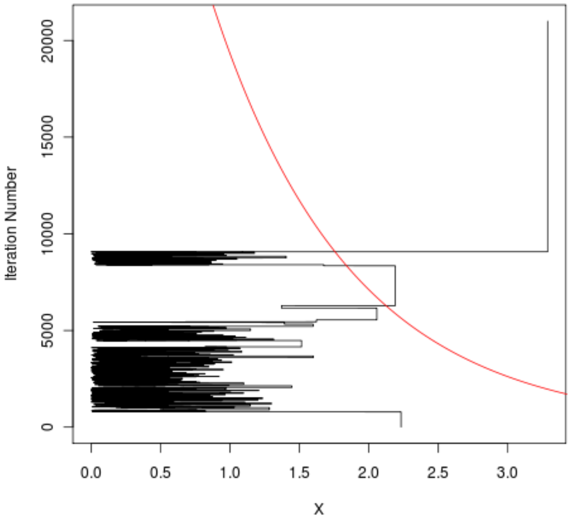


$\mathbf{E}(X) = 1$; estimate $= 1.0186$. Quite good.

Independence sampler (cont'd)

And what if $k > 1$?

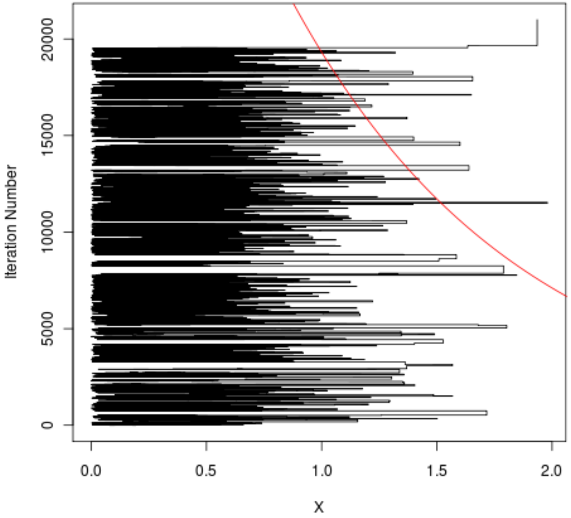
- $k = 5$



$\mathbf{E}(X) = 1$; estimate $= 2.4470$. Terrible: way too big!

What happened? Maybe we just got unlucky?

Another try with $k = 5$:



$\mathbf{E}(X) = 1$; estimate $= 0.7845$. Terrible: way too small!

In fact, we can prove (Roberts and R., MCAP, to appear) that with $k = 5$, the chain takes between 4,000,000 and 14,000,000 iterations to converge to within 0.01 of $\pi(\cdot)$!

So what's going on here?

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

Theorem: Independence samplers are geometrically ergodic if and only if there is $\delta > 0$ for which $q(x) \ge \delta\,\pi(x)$ for all $x \in \mathcal{X}$, in which case $|P^n(x, A) - \pi(A)| \le (1-\delta)^n$.

($k \le 1$: yes; $\quad k > 1$: no)

Again, it matters!

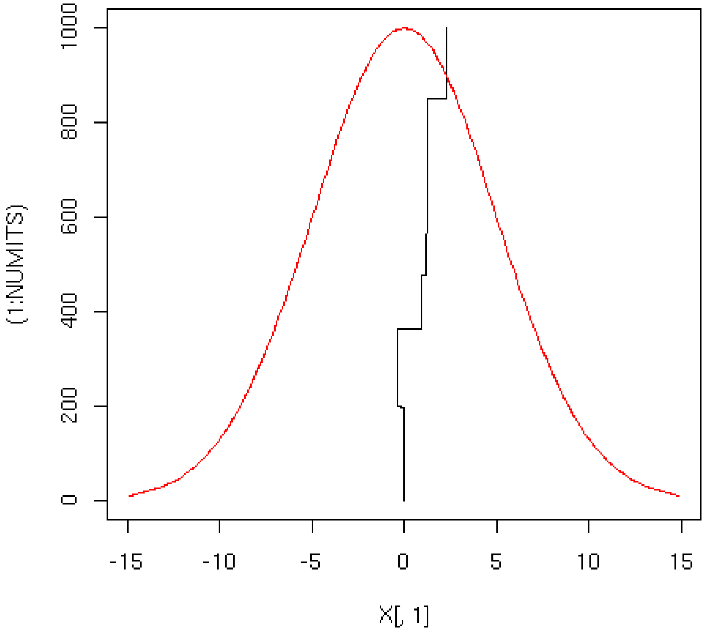
(Also important for ensuring CLTs of estimates: Jones 2004.)

Okay, so: geometric ergodicity is important, quantitative bounds are useful but difficult, and 0.234 is often an optimal acceptance rate.

What about further optimality, beyond "0.234"?

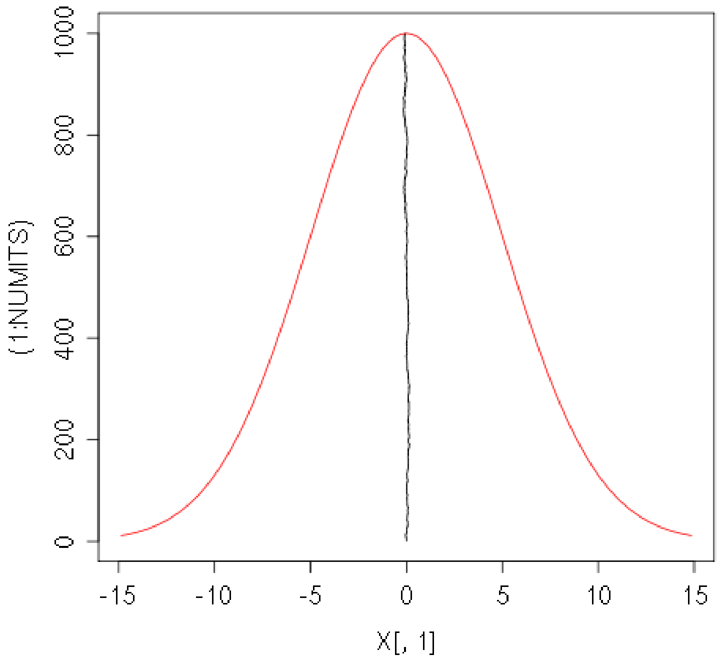
Example #4: $\pi(\cdot) = N(0, \Sigma)$ in dimension 20

First try: $Q(x, \cdot) = N(x, I_{20})$ (acc rate = 0.006)



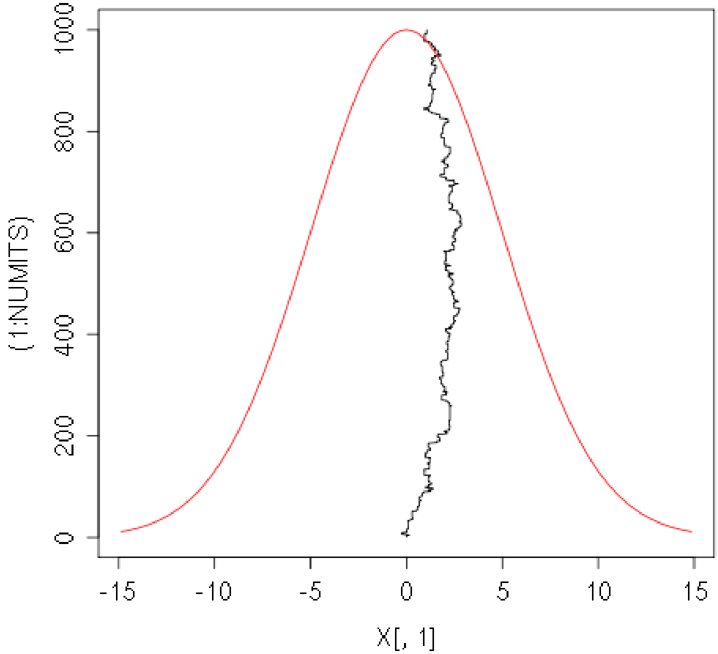
Horrible: $\Sigma_{11} = 24.54$, $E(X_1^2) = 1.50$.

Second try: $Q(x, \cdot) = N\left(x, (0.0001)^2 I_{20}\right)$ (acc=0.9996)



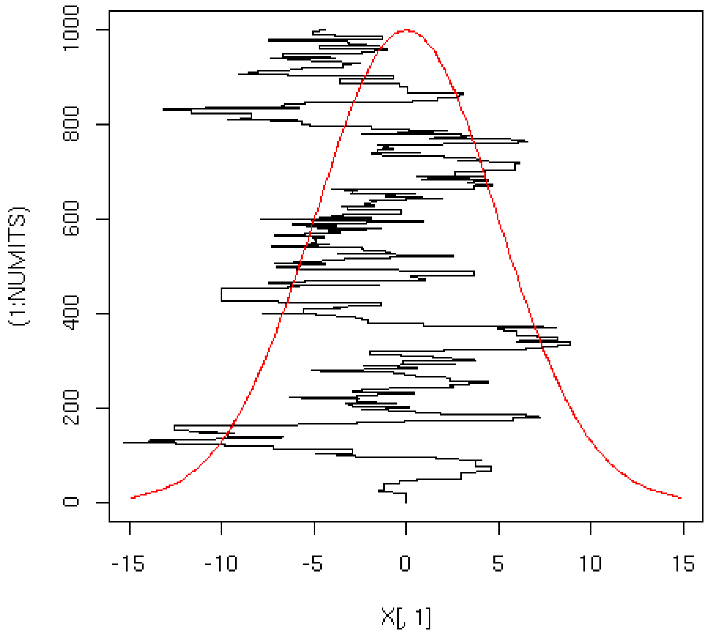
Also horrible: $\Sigma_{11} = 24.54$, $E(X_1^2) = 0.0053$.

Third try: $Q(x, \cdot) = N\Big(x, (0.02)^2 I_{20}\Big)$ (acc=0.234)



Still poor: $\Sigma_{11} = 24.54$, $E(X_1^2) = 3.63$.

Fourth try: $Q(x,\cdot) = N\Big(x, [(2.38)^2/20]\,\Sigma\Big)$ (acc=0.263)



Much better: $\Sigma_{11} = 24.54$, $E(X_1^2) = 25.82$.

Optimizing the Proposal Covariance (Shape)

Theorem [Roberts and R., Stat Sci 2001]:

Under "certain conditions" on $\pi(\cdot)$, the optimal Metropolis algorithm Gaussian proposal distribution as $d \to \infty$ is

$$Q(x, \cdot) \;=\; N\Big(x,\; ((2.38)^2/d)\,\Sigma\Big)$$

[not $N(x, \sigma^2 I_d)$], where Σ is target covariance. And, the corresponding asymptotic acceptance rate is again 0.234.

Very useful, at least if Σ is known!

But what if it isn't??

Adaptive MCMC

What if Σ is unknown?
Can we still "approximately" optimize the RWM algorithm?

Could use "trial and error" (time-consuming, unreliable).
Or, could have computer adapt the algorithm . . .
That is, we design a rule for the computer to update its MCMC algorithm during the run, based on the history.

This destroys the Markov property, stationarity, etc.

But still valid (ergodic) under various conditions.
[Roberts and R., JAP 2007, JCGS 2009; Haario, Saksman, Tamminen, Vihola, Andrieu, Moulines, Robert, Fort, Atchadé, Craiu, Kohn, Giordani, Nott, . . . ; Adap'ski.]

Is it useful??

Example: High-Dimensional Adaptive Metropolis

$\pi(\cdot)$ is d-dimensional target distribution.

e.g. $d = 200$, so target covariance Σ is 200×200, with 20,100 distinct entries. (Can't possibly tune it by hand!)

Know that optimal Gaussian RWM proposal is:

$$N\Big(x,\ [(2.38)^2/d]\,\Sigma\Big).$$

But usually Σ unknown. Instead use empirical estimate, Σ_n.

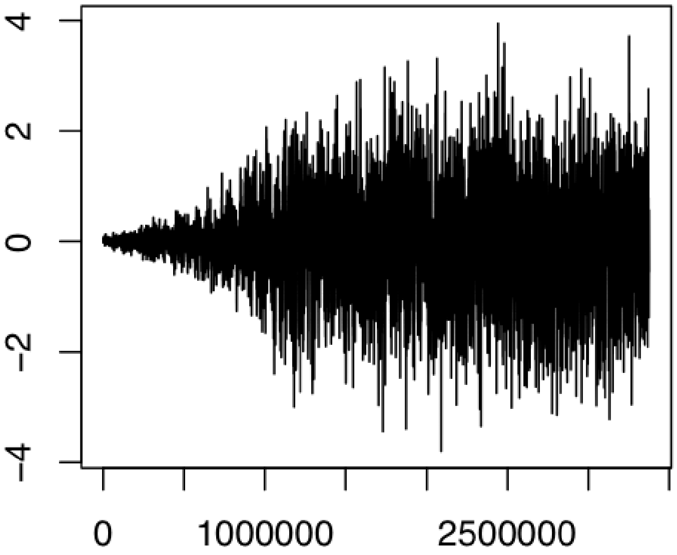
Specifically, let $0 < \beta < 1$, and use proposal distribution:

$$Q_n(x,\cdot)\ =\ (1-\beta)\,N\Big(x,\ [(2.38)^2/d]\,\Sigma_n\Big) + \beta\,N\Big(x,\ [(0.1)^2/d]\,I_d\Big)\,.$$

(Slight variant of algorithm of Haario et al., 2001.)

So, how well does it work?

Adaptive Metropolis in dimension 200



In dimension 200, takes over 1,000,000 iterations, then finds good proposal covariance and starts mixing well.

Good!

Adaptive Metropolis-within-Gibbs Example

Update i^{th} coordinate using proposal distribution $N(x_i,\, e^{2\, ls_i})$, while leaving the other $d-1$ coordinates fixed.

How to choose good values for the proposal scalings ls_i??

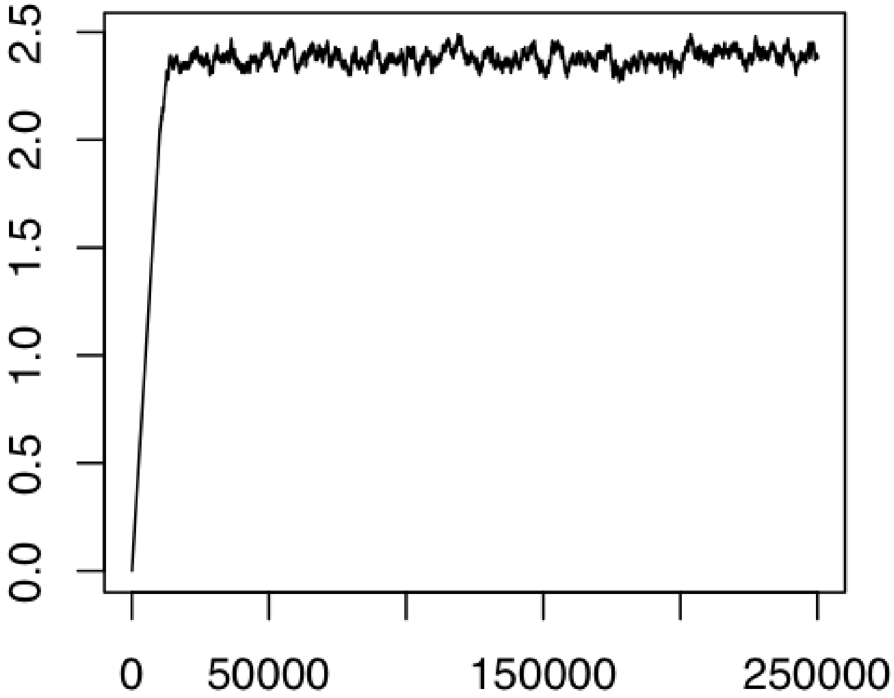
Adaptive algorithm:

- Start with $ls_i \equiv 0$ (say).
- Adapt each ls_i, in batches, to seek 0.44 acceptance rate.

(Approximately optimal for one-dim proposals.)

Test on Variance Components Model, with $K = 500$ (dim=503), J_i chosen with $5 \leq J_i \leq 500$, data $Y_{ij} \sim N(i-1,\, 10^2)$.

How well does it work?



Adaption finds "good" values for the ls_i (R&R, JCGS 2009).

Algorithm applied to statistical genetics models by Turro, Bochkina, Hein, and Richardson (BMC Bioinformatics, 2007).

Adapting Random-Scan Gibbs Sampler Weights

If some coordinates less “significant” than others, may want to update them less often.

Hence, adapt the random-scan coordinate weights, $\{\alpha_{n,i}\}$.

If done carefully, then can prove its ergodic, and it can significantly speed up convergence time in e.g. statistical genetics models (Richardson, Bottolo, R., Valencia 9).

Many other adaptions possible too (e.g. R & R, JCGS 2009).

General-purpose software: probability.ca/amcmc

Lots of recent activity (Adap’ski, . . .); requires theory!

Summary

- MCMC theory is good for a little somethin'!
- Need some theory to define the basic algorithms.
- Theory can help optimize scaling, acceptance rate (e.g. 0.234, etc.), tuning parameters, . . .
- Theory can help improve proposal shape (e.g. $\propto \Sigma$).
- Can compute time to convergence with minorization conditions etc., to ensure correct sampling distributions.
- Geometric ergodicity is an important property that greatly affects performance (convergence, accuracy, CLTs, etc.).
- Theory allows for adaption (if done carefully), to get the computer to help us find good MCMC algorithms.

- The next time you see an MCMC theorist . . . smile.

All my papers, applets, software: probability.ca/jeff