The Mathematics of MCMC Algorithms

Jeffrey S. Rosenthal
University of Toronto

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

(IWAP Workshop, Toronto, June 20, 2016)

Background / 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:

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

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

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

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

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

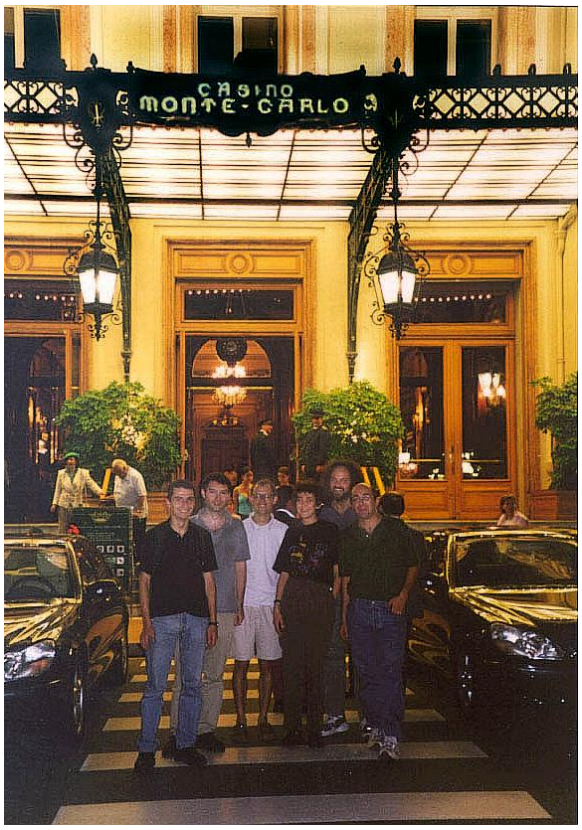
Integrate? Well, no problems *mathematically*, but ...

High-dimensional! Complicated! How to compute?

Try Monte Carlo!

Monte Carlo, Monaco

Nice Place for a Conference!



Estimation from sampling: Monte Carlo

Can try to sample from π, i.e. generate on a computer

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

(meaning that $\mathbf{P}(X_i \in A) = \int_A \pi(x)\, dx$).

Then can estimate by e.g.

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

(Like taking an opinion poll. As $M \to \infty$, the estimate gets more and more accurate. Just like how the gambling house always wins.)

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, ...

But how to create such a Markov chain?

Ex.: 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 $\alpha = \min\left[1, \frac{\pi(Y_n)}{\pi(X_{n-1})}\right]$.
- 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, it converges by general Markov chain theory.

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.
- Suggests new modifications of the algorithms.
- Determines which algorithm choices are best.
- Develops new MCMC directions (e.g. adaptive MCMC).

I'll discuss various Mathematical Research Questions (MRQ).

MRQ#1: How to Optimise MCMC Choices?

The theorem says that we can use 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")

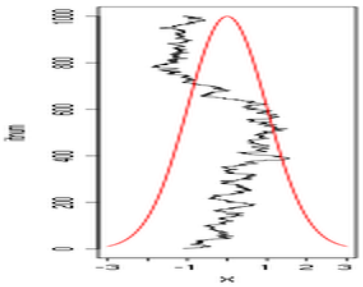
So what choice works best? e.g. What γ in [APPLET]?

- 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")

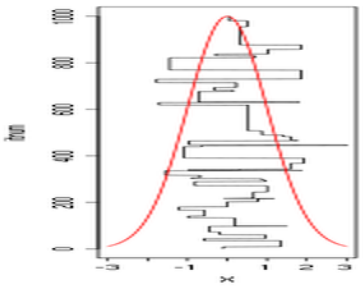
Example: Metropolis for N(0,1)

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

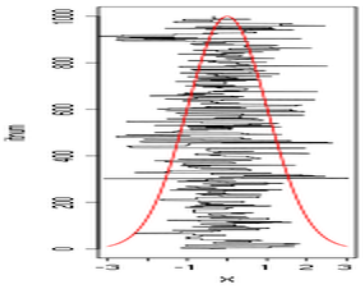
How to choose σ? Big? Small? What acceptance rate (A.R.)?



$\sigma = 0.1$?
too small!
A.R. $= 0.962$



$\sigma = 25$?
too big!
A.R. $= 0.052$



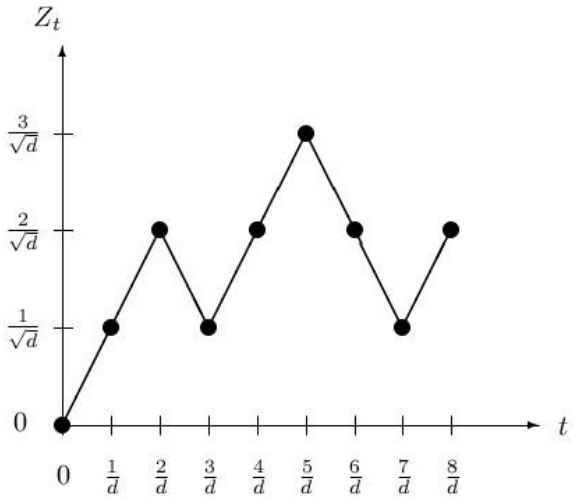
$\sigma = 2.38$?
just right!
A.R. $= 0.441$

The Goldilocks Principle in action!

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

Theoretical Progress: Diffusion Limits

Recall: 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 (i.e., a diffusion).



Do similar limits hold for a Metropolis algorithm, in dimension d, as $d \to \infty$? Yes!

Diffusion Limits for the Metropolis Algorithm

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

- So, to optimize the algorithm, we should maximize $h(\ell)$.
- The maximization gives: $\ell_{opt} \doteq 2.38/C_\pi$. (unknown)
- Then we compute that: $A(\ell_{opt}) \doteq 0.234$. (explicit!)

So, for $Q(x, \cdot) = N(x, \sigma^2 I_d)$, it is optimal to choose a scaling σ^2 which corresponds to an optimal acceptance rate of 0.234.

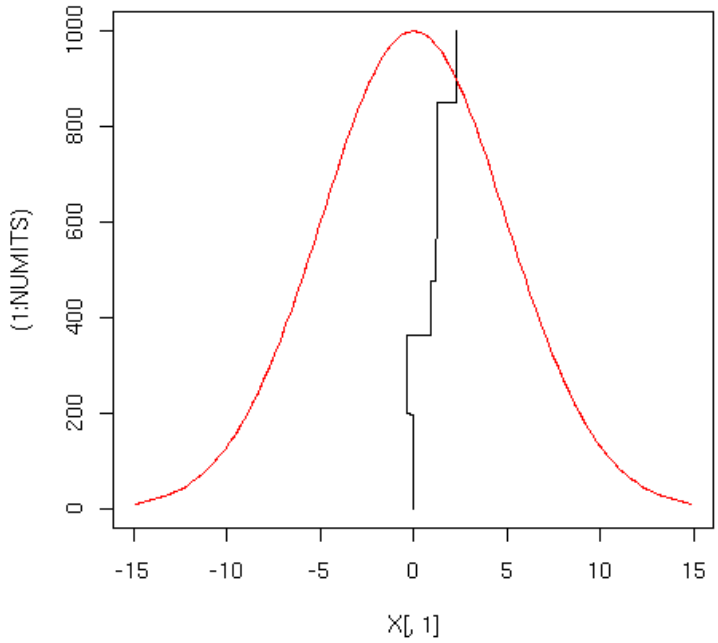
- Clear, simple "0.234" rule. Good! Useful! (Used in BUGS!)

Later generalizations to Langevin diffusions, other targets, etc. (Roberts & R., JRSSB 1998, Stat Sci 2001; Bédard, AAP 2007; Bédard & R., CJS 2008; Sherlock, JAP 2013; Stuart et al.; . . .)

What about further optimality, beyond "0.234"?

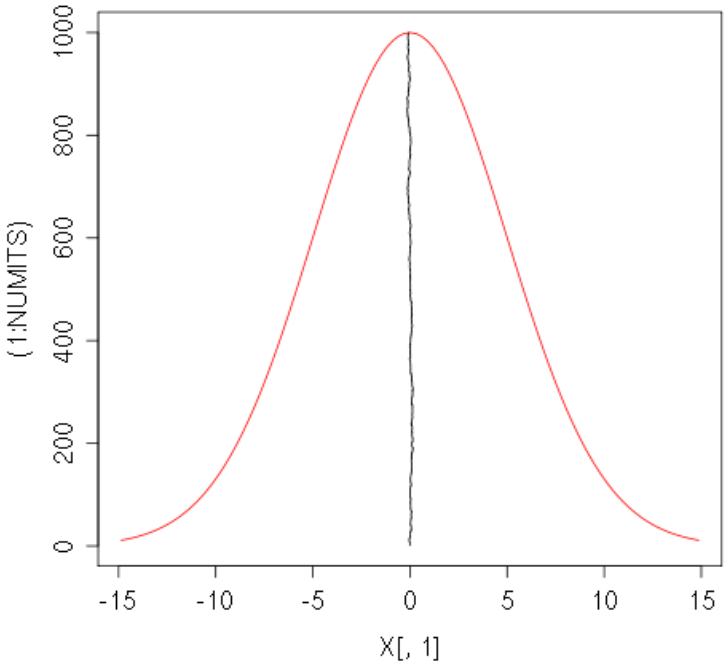
Example: $\pi = N(0, \Sigma)$ in dimension 20

First try: $Q(x, \cdot) = N(x, I_{20})$ (A.R. $= 0.006$)



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

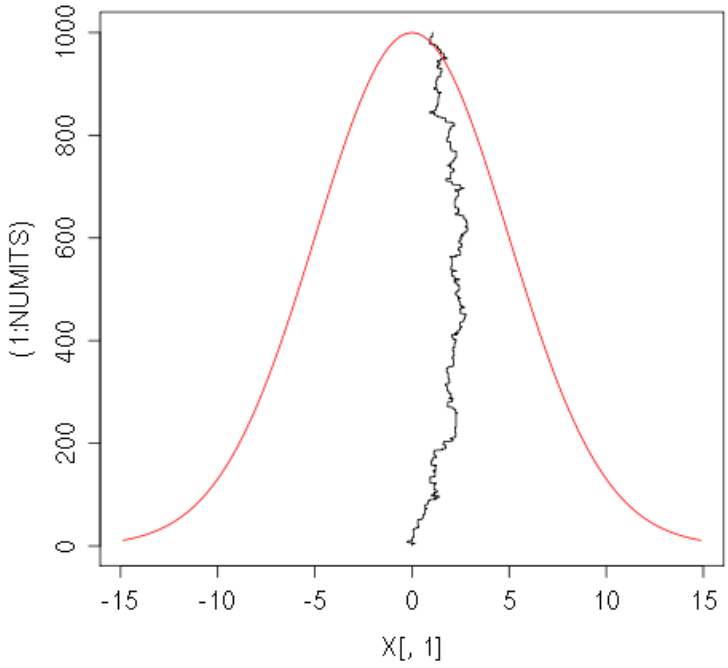
Second try: $Q(x, \cdot) = N\left(x, (0.0001)^2 I_{20}\right)$ (A.R.=0.9996)



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

Need bigger proposal!

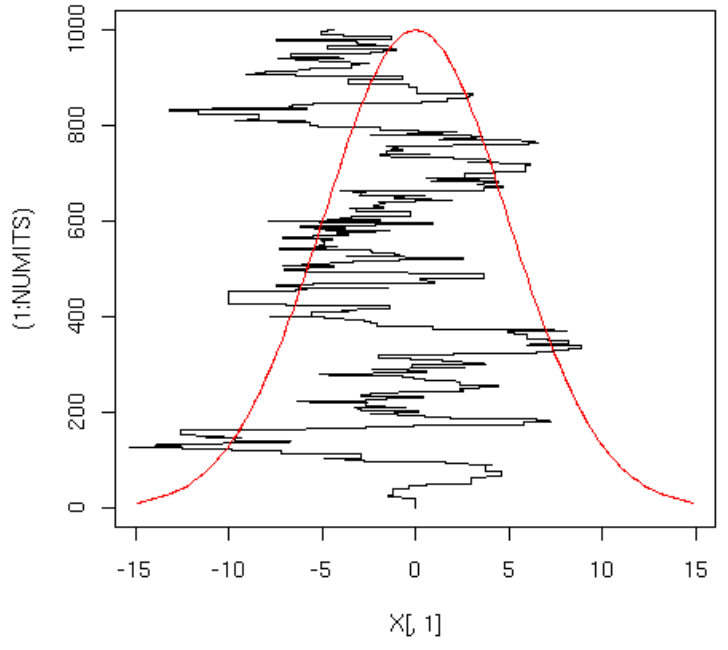
Third try: $Q(x,\cdot) = N\Big(x, (0.02)^2 I_{20}\Big)$ (A.R.=0.234)



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

But acceptance rate is "just right". What gives?

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



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

Not perfect, but fairly good. Why?

Theory about the Proposal Covariance (Shape)

Theorem [Roberts and R., Stat Sci 2001]:
Under "certain conditions" on π, 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.

The corresponding asymptotic acceptance rate is again 0.234.

- And, this turns out to be nearly optimal for many other high-dimensional densities, too.

This gives very useful advice ... if Σ is known!

But what if the target covariance Σ is unknown?

Can we make use of this optimality result anyway?

(Adaptive MCMC – later.) But first ...

MRQ#2: Quantitative Convergence Bounds?

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

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

One method: coupling. (Other methods: drifts, eigenvalues, ...)

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

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

If 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

Simplest version:

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 give the two copies probability ϵ of becoming equal. Hence, $\mathbf{P}(T > n) = (1 - \epsilon)^n$.

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

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

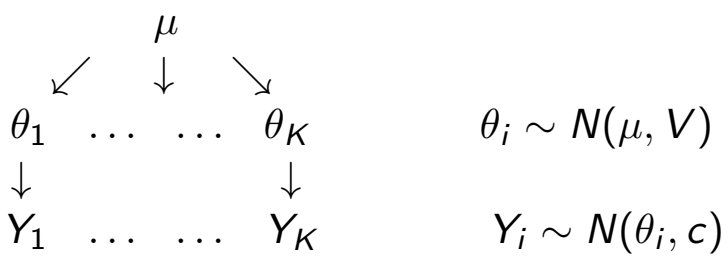
- So $|P^n(x, A) - \pi(A)| \leq (1 - \epsilon)^n = (1 - 0.2)^n = (0.8)^n$.
- Hence, $|P^n(x, A) - \pi(A)| < 0.01$ whenever $n \geq 21$.
- "The chain converges in 21 iterations." Good!

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 empirically estimated, with $\mu, V, \theta_1, \ldots, \theta_K$ to be estimated. Priors: $\mu \sim$ flat, $V \sim IG(a, b)$.

Diagram:



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

High dimensional! How to estimate?

Baseball Data Model (cont'd)

MCMC solution: Run a Gibbs sampler for π.

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

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

$$\mu^{(n)} \sim N(\overline{\theta}^{(n-1)},\ A^{(n)}/K);$$

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

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

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

Complicated! How to analyze convergence?

Example: Baseball Data Model (cont'd)

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

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

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

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

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

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

so e.g. $|\mathbf{P}(X_n \in A) - \pi(A)| < 0.01$ if $n \ge 140$.
"The chain converges in 140 iterations." Good!

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

MRQ#3: Qualitative Convergence Bounds

Quantitative bounds too tricky for everyday use . . . what else?

DEFN: Say the chain is geometrically ergodic if

$$\sup_A |\mathbf{P}(X_n \in A) - \pi(A)| \;\leq\; B_x\, \rho^n\,, \qquad n = 1, 2, 3, \ldots$$

for some $\rho < 1$, where $B_x < \infty$ for π-a.e. $x = X_0$.
i.e., convergence is exponentially quick (at some exponential rate).

This property always holds on finite state spaces.

- (e.g. must hold for [APPLET] example)

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

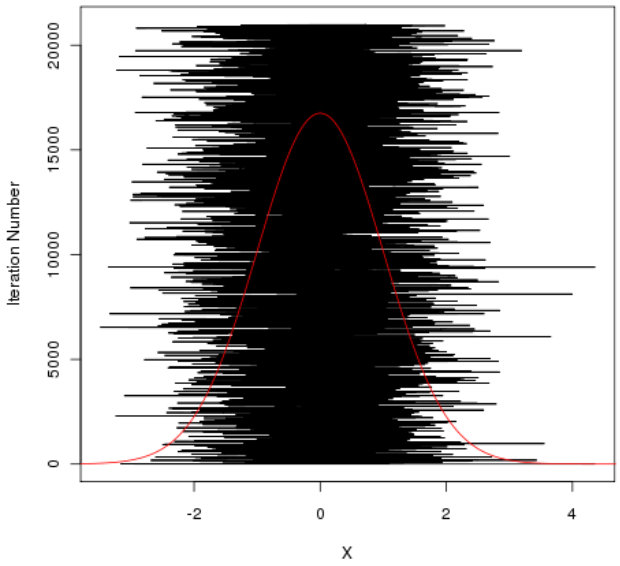
It says something about quick convergence (good).

But not too much, since ρ and B_x are unspecified (bad).

Easier. But does this qualitative property actually matter??

Example: Metropolis for N(0,1), again

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



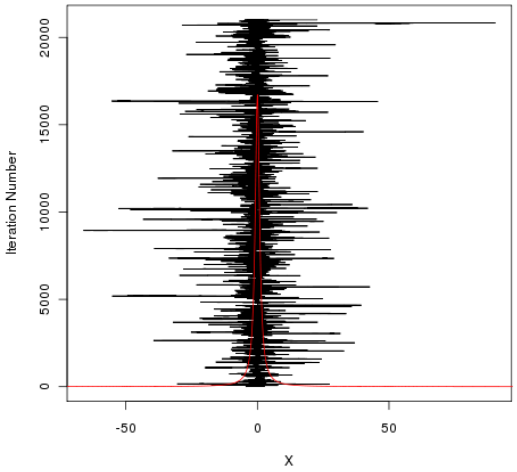
$\mathbf{P}(|X| > 2) \doteq 0.0455$; estimate $= 0.0453$. Great!

Does it always work so well?

Example: Metropolis for Cauchy

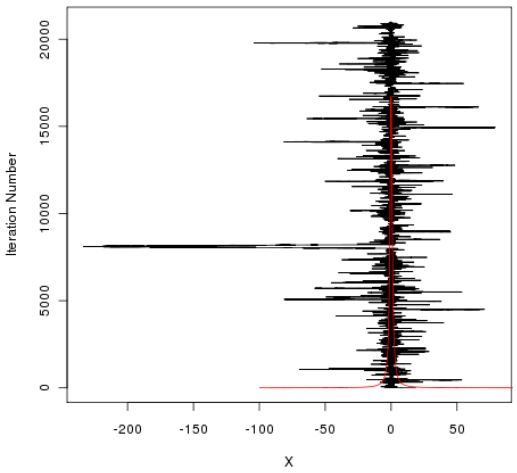
Random-walk Metropolis for $\pi(x) = \frac{c}{1+x^2}$ (Cauchy), with $Q(x, \cdot) = N(x, \sigma^2)$, with σ again chosen to make A.R. $\doteq 0.234$.

Much worse!



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

Example: Metropolis for Cauchy, second try



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

- So, MCMC is performing very badly here. Why??

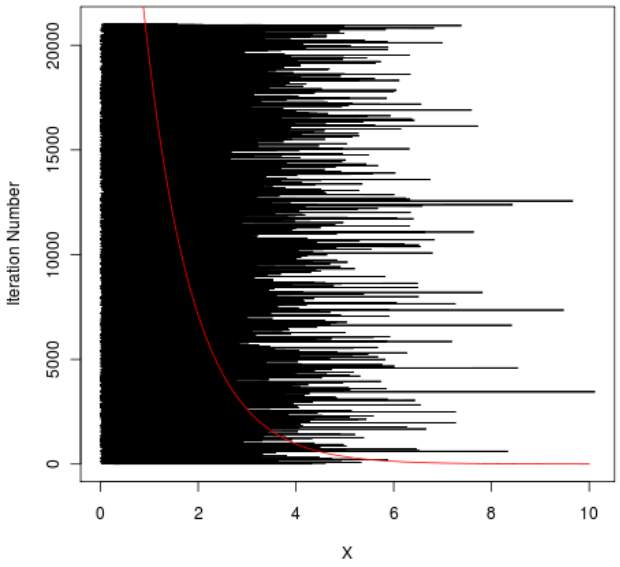
Theorem (Mengersen-Tweedie-Roberts, 1996): Metropolis is geometrically ergodic iff $\pi(\cdot)$ has exponentially-small tails.

N(0,1): yes. Cauchy: no. Makes a big difference!

MRQ#4: 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 $k > 0$.

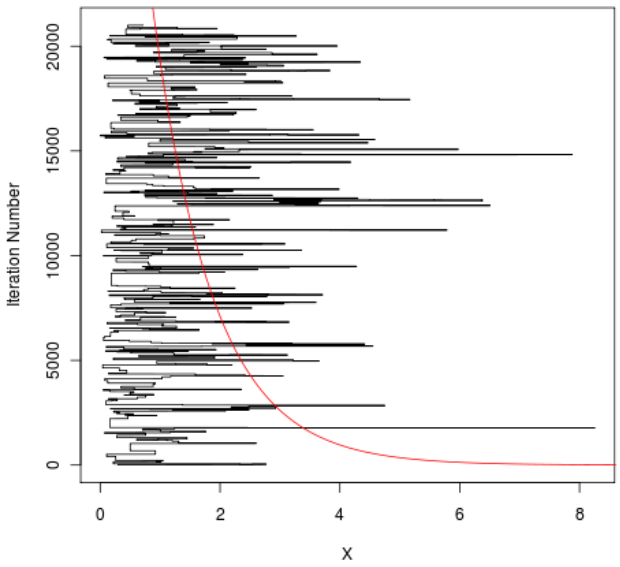
- $k = 1$ (i.i.d. sampling)



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

Independence sampler (cont'd)

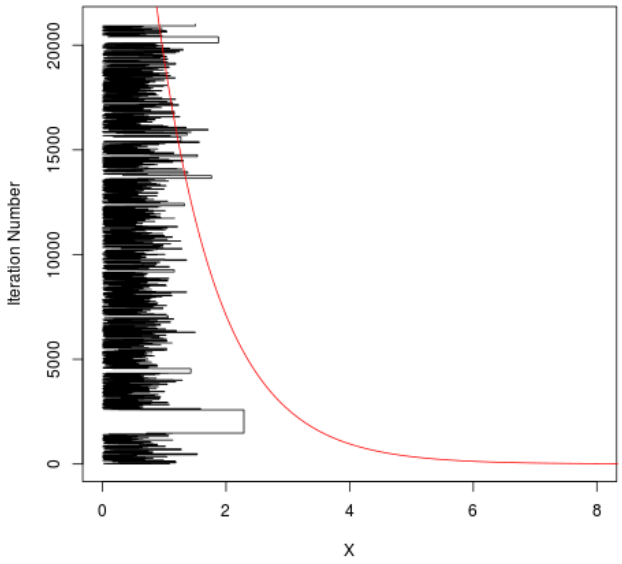
- $k = 0.01$



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

Independence sampler (cont'd)

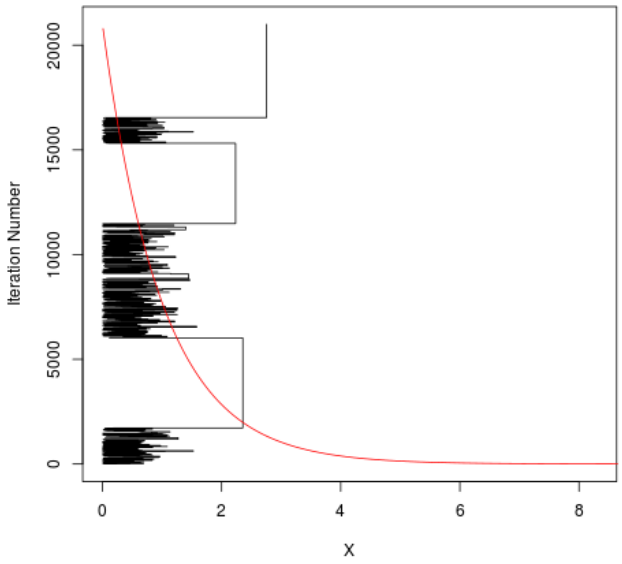
- $k = 5$



$\mathbf{E}(X) = 1$; 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!

In fact, we can prove (Roberts and R., MCAP, 2011) that with $k = 5$, the chain takes between 4,000,000 and 14,000,000 iterations to converge to within 0.01 of π! But why??

Independence Sampler: Theory

What's going on in this example?

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

Theorem [Mengersen & Tweedie, Ann Stat 1996]:
Independence samplers are geometrically ergodic if and only if there is $\delta > 0$ for which $Q(x) \geq \delta\, \pi(x)$ for all $x \in \mathcal{X}$.

If there is, then $|P^n(x, A) - \pi(A)| \leq (1 - \delta)^n$. (Quantitative!)

In above example, $\pi(x) = e^{-x}$ and $Q(x) = ke^{-kx}$, so:

- $k = 1$: yes, $\delta = 1$; converges immediately (of course).
- $k = 0.01$: yes, $\delta = 0.01$; and $(1 - 0.01)^{459} < 0.01$, so the chain "converges within 459 iterations". (Pretty good.)
- $k = 5$: no such δ. Not geometrically ergodic. (Bad.)

So, geometric ergodicity makes a big difference!

MRQ#5: Validity of Adaptive MCMC?

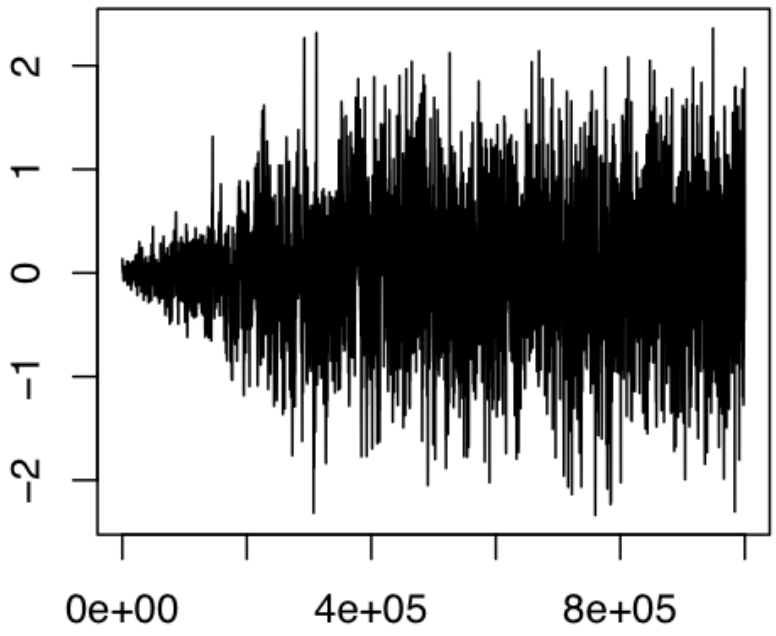
Recall:

- MCMC is really really really important.
- Some MCMC algorithms converge much faster than others.
- Can find optimality results from diffusion limits.
- e.g. Gaussian Random-Walk Metropolis: optimal choice has acceptance rate around 0.234 (how?), and proposal covariance $(2.38)^2\, d^{-1}\, \Sigma_t$ where Σ_t is the target covariance (unknown).
- So, we have guidance about optimising MCMC in terms of acceptance rate, target covariance matrix Σ_t, etc.
- But we don't know what proposal will lead to a desired acceptance rate, nor how to compute Σ_t.
- What to do? Trial and error? (difficult, especially in high dimension) Or ...

Adaptive MCMC

- Suppose have a family $\{P_\gamma\}_{\gamma \in \mathcal{Y}}$ of possible Markov chains, each with stationary distribution π.
- How to choose among them?
- Let the computer decide, on the fly!
- At iteration n, use Markov chain P_{Γ_n}, where $\Gamma_n \in \mathcal{Y}$ chosen according to some adaptive rules (depending on history, etc.).
- Simple example: [APPLET]
- e.g. Estimate true target covariance Σ_t by the empirical estimate, Σ_n, based on the observations so far $(X_1, X_2, \ldots, X_n)$.
- Can this help us to find better Markov chains? (Yes!)
- On the other hand, the Markov property, stationarity, etc. are all destroyed by using an adaptive scheme.
- Is the resulting algorithm still ergodic? (Sometimes!)

Example: 100-Dimensional Adaptive Metropolis



Plot of first coord. Takes about 300,000 iterations, then "finds" good proposal covariance and starts mixing well. Good!

- Similarly Adaptive Componentwise Metropolis, Gibbs, etc.

But What About the Theory?

- So, adaptive MCMC seems to work well in practice.
- But will it be ergodic, i.e. converge to π? (Converge at all ... never mind how quickly ...)
- Ordinary MCMC algorithms, with fixed choice γ, are automatically ergodic by standard Markov chain theory (since they're irreducible and aperiodic and leave π stationary). But adaptive algorithms are more subtle, since the Markov property and stationarity are destroyed by using an adaptive scheme.
- e.g. if the adaption of Γ_n is such that P_{Γ_n} usually moves slower when x is in a certain subset $\mathcal{X}_0 \subseteq \mathcal{X}$, then the algorithm will tend to spend much more than $\pi(\mathcal{X}_0)$ of the time inside $\mathcal{X}_0$, even if each update on its own preserves stationarity. [APPLET]
- Some previous results, but they require limiting / hard-to-verify conditions, like bounded state space, or existence of simultaneous geometric drift conditions, or Doeblin condition, or ...
- Need more general, easily-verified theorems ...

One Particular Convergence Theorem

- Theorem [Roberts and R., J.A.P. 2007]: Adaptive MCMC will converge, i.e. $\lim_{n\to\infty} \sup_{A\subseteq\mathcal{X}} \|\mathbf{P}(X_n \in A) - \pi(A)\| = 0$, if:

(a) [Diminishing Adaptation] Adapt less and less as the algorithm proceeds. Formally, $\sup_{x\in\mathcal{X}} \|P_{\Gamma_{n+1}}(x,\cdot) - P_{\Gamma_n}(x,\cdot)\| \to 0$ in prob. [Can always be made to hold, since adaption is user controlled.]

(b) [Containment] Times to stationary from X_n, if fix $\gamma = \Gamma_n$, remain bounded in probability as $n \to \infty$. [Technical condition, to avoid "escape to infinity". Holds if e.g. $\mathcal{X}$ and $\mathcal{Y}$ finite, or compact, or ... And always seems to hold in practice.]

(Also guarantees WLLN for bounded functionals. Various other results about LLN / CLT under stronger assumptions.)

Good, but ... Containment condition is a pain.

Can we eliminate it?

What about that "Containment" Condition?

- Recall: adaptive MCMC is ergodic if it satisfied Diminishing Adaptation (easy: user-controlled) and Containment (technical).
- Is Containment just an annoying artifact of the proof? No!
- Theorem (Latuszynski and R., 2014): If an adaptive algorithm does not satisfy Containment, then for all $\epsilon > 0$,

$$\lim_{K\to\infty} \limsup_{n\to\infty} \mathbf{P}(M_\epsilon(X_n, \gamma_n) > K) \;>\; 0\,,$$

where $M_\epsilon(x,\gamma) = \inf\{n \geq 1 : \|P_\gamma^n(x,\cdot) - \pi(\cdot)\| < \epsilon\}$ is the time to converge to within ϵ of stationarity.

That is, an adaptive algorithm without Containment will take arbitrarily large numbers of steps (K) to converge. Bad!

- Conclusion: Yay Containment!?!?
- But how to verify it??

Verifying Containment: "For Everyone"

- Proved general theorems about stability of "adversarial" Markov chains under various conditions (Craiu, Gray, Latuszynski, Madras, Roberts, and R., A.A.P. 2015).
- Then applied them to adaptive MCMC, to get a list of directly-verifiable conditions which guarantee Containment:
  - $\Rightarrow$ Never move more than some (big) distance D.
  - $\Rightarrow$ Outside (big) rectangle K, use fixed kernel (no adapting).
  - $\Rightarrow$ The transition or proposal kernels have continuous densities wrt Lebesgue measure. (or piecewise continuous: Yang & R. 2015)
  - $\Rightarrow$ The fixed kernel is bounded, above and below (on compact regions, for jumps $\leq \delta$), by constants times Lebesgue measure. (Easily verified under continuity assumptions.)
- Can directly verify these conditions in practice.
- So, this can be used by applied MCMC users.
- "Adaptive MCMC for everyone!"

Summary

- MCMC has tremendous application to many areas.
- MCMC mathematical theory plays a crucial supporting role.
- Theory can help verify and extend the algorithms, optimise proposal scaling / shape, bound convergence times with minorization conditions etc., show geometric ergodicity, and more.
- Theory also allows for adaption (if done carefully), to get the computer to "learn" good MCMC algorithms and run faster.
- Adaptive MCMC works very well, even in high-dimensional examples (good). But it must be done carefully, or it will destroy stationarity (bad). Suffices to have stationarity of each P_γ, plus Diminishing Adaptation (important), and Containment (technical condition, usually satisfied, necessary). "Adversarial Markov chain" theorems provide simple sufficient conditions.
- All my papers, applets, software: www.probability.ca