UNDERSTANDING MCMC

Gareth O. Roberts and Jeffrey S. Rosenthal

Lancaster

July 2003

Contents

1 INTRODUCTION & MOTIVATION

About this course

A course on the theoretical underpinnings of Markov chain Monte Carlo designed for statisticians wishing to develop a deeper understanding of methods which they use (or expect to use in future).

What is this course not:

- a course in hands-on MCMC implementation;
- a course on statistical modelling with MCMC.

The problem

We're given a (possibly un-normalised) density function π_u. So there exists a probability density function π such that

$$\pi(x) = k^{-1}\pi_u(x)$$

for a constant $k = \int_{\mathcal{X}} \pi_u(x)dx$ which is unknown to us.

We're interested in the probability density π.
Why?

Perhaps we want to

- estimate expectations with respect to π:

$$\pi(f) = E_\pi[f(X)] = \int_{\mathcal{X}} f(x)\pi(x)dx;$$

- simulate a sample of points from distribution π.

Where does this problem arise?

Example 1: Bayesian Statistics

$L(\mathbf{y};\theta)$ is the likelihood of a statistical experiment with data $\mathbf{y}$ and unknown parameters, $\theta \in \Theta$. Let the prior on θ be $p(\theta)$.

From Bayes's Theorem, the posterior distribution of θ given $\mathbf{y}$ is

$$\pi(\theta|\mathbf{y}) \propto L(\mathbf{y};\theta)p(\theta) \equiv \pi_u(\theta|\mathbf{y}).$$

Example 2

Distributions defined in terms of their conditional distributions.

For example for $1 \leq i, j \leq n$, let $X_{i,j}$ take the value 0 or 1, with

$$P[X_{ij} = 1|\text{all other } X's] = \frac{\ell}{1+\ell}$$

where

$$\ell = \exp\{\beta(\#\text{neighbouring } 1's - \#\text{neighbouring } 0's)\}\ .$$

All these conditional distributions characterise the distribution (Clifford-Hammersley) and we can write

$$\pi_u = \exp\{\beta(\#\text{neighbouring similar values} - \#\text{neighbouring differing values})\}\ .$$

Example 3: Approximate counting

Let $A \subset B \subset \mathbb{Z}_+^d$. We wish to compute $|A|/|B|$. B is sufficiently complex that 'counting' its elements is impossible, but $|A|$ is known.

For $\mathbf{x} \in \mathbb{Z}_+^d$, take

$$\pi_u(\mathbf{x}) = \mathbb{1}_{\mathbf{x}\in B}.$$

Then

$$\frac{|A|}{|B|} = \frac{\sum_{\mathbf{x}\in\mathbb{Z}_+^d} \pi_u(\mathbf{x})\mathbb{1}_{\mathbf{x}\in A}}{\sum_{\mathbf{x}\in\mathbb{Z}_+^d} \pi_u(\mathbf{x})}\ .$$

So

$$|B| = \frac{|A|}{\pi(A)}\ ,$$

so that by estimating $\pi(A)$ in some way, we can approximate $|B|$.

Characteristics of problem

- 'Large' state spaces, often very high-dimensional
- Too complex for 'direct simulation' from π to be feasible.
- Identifying k is generally as hard as the whole simulation problem

Other solutions to the problem

Apart from MCMC, there are many 'solutions' to this simulation/integration/estimation problem. We'll look at a couple of the most versatile which have some connections to MCMC.

Rejection sampling

Suppose g is easy to sample from, and $\exists c$ such that $\pi_u(x) \leq cg(x)$ for all $x \in \mathcal{X}$. Then

1. 1. Draw $Z \sim g(\cdot)$.
2. Accept Z if $U \leq h(Z)/cg(Z)$.
3. Otherwise, repeat 1.

This algorithm outputs an observation from π under mild conditions on g and π_u. However

- Need to find g such that π/g is bounded.
- Even when we can find g, it needs to be a "good" enveloping function for the method to be reasonably efficient.

Importance sampling

Suppose that π/g is not bounded. Can we do anything now? Given a sample $X_1, X_2, \ldots X_n$ from g, we can estimate $E_\pi[f(X)]$ by

$$\frac{\sum_{i=1}^n f(X_i)w(X_i)}{\sum_{i=1}^n w(X_i)}$$

where $w(x) = \frac{\pi_u(x)}{g(x)}$.

This gets around the boundedness problem but

- still need a "good" approximating function g for this method to be efficient;
- this method doesn't actually produce a sample from π.

The problem of getting good approximating functions gets rapidly harder as dimension increases.

MCMC

Markov chain Monte Carlo is a method for drawing samples from π using Markov chains which have stationary distribution π.

They only need π_u for their implementation, that is the normalisation constant is not needed.

Metropolis Rosenbluth Rosenbluth Teller and Teller (J. Chemical Physics 1953).

Hastings (Biometrika, 1970)

Besag (JRSSB, 1974)

Suomela (PhD University of Helsinki, 1976)

Geman and Geman (IEEE Trans. Pattn. Anal. Mach. Intel., 1984)

Ripley (Stochastic Simulation, 1987)

Tanner and Wong (JASA, 1987)

Gelfand and Smith (JASA, 1990)

2 COMMON MCMC ALGORITHMS

The Challenge

Let $\pi(\cdot)$ be a target denstity, on some state space $\mathcal{X}$ (e.g. $\mathcal{X} = \mathbf{R}^d$), that we wish to sample from.

We wish to construct a *Markov chain* on $\mathcal{X}$ which has $\pi(\cdot)$ as its stationary distribution.

That is, we want to define Markov chain transition probabilities $P(x, dy)$ for $x, y \in \mathcal{X}$, with

$$\int_{x \in \mathcal{X}} \pi(dx)\, P(x, dy) \;=\; \pi(dy)\,.$$

[In words, if you begin in the distribution $\pi(\cdot)$, then one step later, you will still be in the distribution $\pi(\cdot)$.]

Then hopefully, if we run the Markov chain for a long time (started from anywhere), then for large n the distribution of X_n will be approximately $\pi(\cdot)$. [Good!]

But how can we construct $P(x, dy)$?

A Very Simple Example

Suppose $\mathcal{X} = \{1, 2, 3\}$, with:

$$\pi\{1\} = 1/6\,, \quad \pi\{2\} = 1/3\,, \quad \pi\{3\} = 1/2\,.$$

Let $P(1, \{2\}) = 1$, $P(2, \{1\}) = P(2, \{3\}) = 1/2$, and $P(3, \{2\}) = 1/3$, $P(3, \{3\}) = 2/3$.

Then

$$\int \pi\{x\}\, P(x, \{2\}) = \pi\{1\}\, P(1, \{2\}) + \pi\{3\}\, P(3, \{2\})$$
$$= (1/6)(1) + (1/2)(1/3) = 1/3 = \pi\{2\}\,.$$

Similarly $\int \pi\{x\}\, P(x, \{1\}) = \pi\{1\}$ and $\int \pi\{x\}\, P(x, \{3\}) = \pi\{3\}$. Success!

So, for large n, probably have $\mathcal{L}(X_n) \approx \pi(\cdot)$. [Good!]

But how did we know? And, how would we proceed for a more *complicated* example?

Reversibility

In above example, we have

$$\pi\{x\}\, P(x, \{y\}) = \pi\{y\}\, P(y, \{x\})$$

for every $x, y \in \mathcal{X}$.

DEFN: A Markov chain is *reversible* with respect to $\pi(\cdot)$ if

$$\pi(dx)\, P(x, dy) = \pi(dy)\, P(y, dx)\,, \quad x, y \in \mathcal{X}\,.$$

FACT: If Markov chain is reversible with respect to $\pi(\cdot)$, then $\pi(\cdot)$ is stationary.

PROOF: If reversible, then

$$\int_{x \in \mathcal{X}} \pi(dx)\, P(x, dy) = \int_{x \in \mathcal{X}} \pi(dy)\, P(y, dx)$$
$$= \pi(dy) \int_{x \in \mathcal{X}} P(y, dx) = \pi(dy)\,.$$

So, suffices to make chain *reversible*.

The Metropolis-Hastings Algorithm

Suppose $\pi(\cdot)$ has a density:

$$\pi(dx) = \pi(x)\, dx\,.$$

Suppose $Q(x, \cdot)$ is some other (simple) Markov chain, also having a density:

$$Q(x, dy) = q(x, y)\, dy\,.$$

The Metropolis-Hastings algorithm proceeds as follows. Given X_n, generate Y_{n+1} from $Q(X_n, \cdot)$. Then, randomly either "accept" and set $X_{n+1} = Y_{n+1}$, or "reject" and set $X_{n+1} = X_n$.

Accept with probability $\alpha(X_n, Y_{n+1})$, or reject with probability $1 - \alpha(X_n, Y_{n+1})$, where

$$\alpha(x, y) = \min\left[1,\ \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right].$$

FACT: The formula for $\alpha(x, y)$ was chosen "just right", so that the resulting Markov chain $\{X_n\}$ is reversible with respect to $\pi(\cdot)$.

PROOF: Need to show

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

Suffices to assume $x \neq y$ (otherwise trivial).

But for $x \neq y$,

$$\begin{aligned}\pi(dx)\, P(x, dy) &= [\pi(x)\, dx]\, [q(x, y)\, \alpha(x, y)\, dy] \\ &= \pi(x)\, q(x, y)\ \min\left[1,\ \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right] dx\, dy \\ &= \min[\pi(x)\, q(x, y),\ \pi(y)q(y, x)]\, dx\, dy\end{aligned}$$

and similarly

$$\pi(dy)\, P(y, dx) = \min[\pi(x)\, q(x, y),\ \pi(y)q(y, x)]\, dx\, dy\,.$$

Summary So Far

- The Metropolis-Hastings algorithm proposes a new state according to the proposal kernel $Q(x, \cdot)$, and then either accepts or rejects it, with just the right probabilities to make $\pi(\cdot)$ be *reversible* (and hence *stationary*).
- To run this algorithm on a computer, we just need to be able to run the proposal chain $Q(x, \cdot)$ [easy, for appropriate choice of Q], and then do the accept/reject step [easy as long as we can compute the densities at individual points]. Good!
- Furthermore we need to compute only *ratios* of densities [e.g. $\pi(y)\,/\,\pi(x)$], so we don't require the *normalising constants*. Good!
- But, how to choose the proposal $Q(x, \cdot)$?
- And, will we really have $\mathcal{L}(X_n) \approx \pi(\cdot)$ for large enough n? (How large??)

A Metropolis-Hastings Example

Suppose $\mathcal{X} = \mathbf{R}^d$, with $\pi : \mathcal{X} \to (0, \infty)$ a complicated density function.

Let the proposal be $Q(x, \cdot) = N(x, 1)$, so that $q(x, y) = \frac{1}{\sqrt{2\pi}} e^{-(y-x)^2/2} = q(y, x)$.

Then

$$\alpha(x, y) = \min\left[1,\ \frac{\pi(y)\, q(y, x)}{\pi(x)\, q(x, y)}\right] = \min\left[1,\ \frac{\pi(y)}{\pi(x)}\right].$$

Take $X_0 = 0$ (say).

Then, given X_n, we choose $Y_{n+1} \sim N(X_n, 1)$, and then set

$$X_{n+1} = \begin{cases} Y_{n+1}\,, & \text{probability } \alpha(X_n, Y_{n+1}) \\ X_n\,, & \text{probability } 1 - \alpha(X_n, Y_{n+1}) \end{cases}$$

This creates a sequence $X_0, X_1, X_2, \ldots$

Hopefully, for large n, the density of X_n is approximately equal to π.

Metropolis-Hastings Variations

There are many different ways of choosing the proposal density, such as:

- **Symmetric Metropolis Algorithm.** Here

$$q(x, y) = q(y, x)$$

The acceptance probability simplifies to

$$\alpha(x, y) = \min\left[1,\ \frac{\pi(y)}{\pi(x)}\right]$$

- **Symmetric random walk Metropolis.**

$$q(x, y) = q(y - x)$$

[e.g. $Q(x, \cdot) = N(x, \sigma^2)$, or
$Q(x, \cdot) = \text{Uniform}(x - 1,\ x + 1)$, etc.]

- **Independence sampler.** Here

$$q(x, y) = q(y)\,,$$

i.e. $Q(x, \cdot)$ does not depend on x.

(Similar to rejection sampler ... but not identical.)

- **Langevin algorithm.**

Here the proposal is generated by

$$Y_{n+1} \sim N(X_n + (\delta/2)\,\nabla \log \pi(X_n),\ \delta)\,,$$

for some (small) $\delta > 0$.

(This is motivated by a discrete approximation to a "Langevin diffusion" processes.)

- **Multiplicative RWM**

This is just a symmetric random walk Metropolis algorithm 'on a log scale' and is sometimes useful for components which are strictly positive. For example

$$Q(x, \cdot) = x e^{N(0,\sigma^2)}\ .$$

$$\alpha(x, y) = \min\left[1, \frac{x\pi(y)}{y\pi(x)}\right]$$

Exercise check this!

The Gibbs Sampler

Suppose that $\pi(\cdot)$ is d-dimensional, i.e. that $\mathcal{X} \subseteq \mathbf{R}^d$.

Write $\mathbf{x} = (x_1, \ldots, x_d)$, and
$\mathbf{x}^{(-i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$.

Let $\pi_i(\mathbf{y} \,|\, \mathbf{x}^{(-i)})$ be the conditional density of $\pi(\cdot)$, conditional on knowing that $y_j = x_j$ for $j \neq i$:

$$\pi_i(\mathbf{y} \,|\, \mathbf{x}^{(-i)}) \ = \ \frac{g_{i,x}(\mathbf{y})}{\int_{\mathbf{z}\in\mathcal{X}} g_{i,x}(\mathbf{z})\, d\mathbf{z}}\,,$$

where $g_{i,x}(\mathbf{y}) = \pi(\mathbf{y})\mathbf{1}_{\{y_j = x_j \text{ for } j\neq i\}}$.

The i^{th} component Gibbs sampler is defined by

$$P_i(\mathbf{x}, d\mathbf{y}) = \pi_i(\mathbf{y} \,|\, \mathbf{x}^{(-i)})\, d\mathbf{y}\,.$$

That is, P_i leaves all components besides i unchanged, and replaces the i^{th} component by a draw from the full conditional distribution of $\pi(\cdot)$ given all the other components.

FACT: The i^{th} component Gibbs sampler, P_i, is reversible with respect to $\pi(\cdot)$.

(This follows from the definition of conditional density. In fact, P_i may be regarded as a special case of a Metropolis-Hastings algorithm.)

So, P_i leaves $\pi(\cdot)$ invariant. We then construct the Gibbs sampler out of P_i, as follows:

- The deterministic-scan Gibbs sampler is

$$P \;=\; P_1 P_2 \ldots P_d \,.$$

That is, it does the d different Gibbs sampler components, in order.

- The random-scan Gibbs sampler is

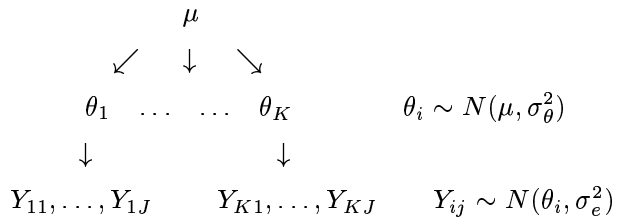
$$P \;=\; \frac{1}{d} \sum_{i=1}^{d} P_i \,.$$

That is, it does one of the d different Gibbs sampler components, chosen uniformly at random.

Either version produces a "zig-zag pattern".

Example: Variance Components Model

MODEL:



PRIORS: $\sigma_\theta^2 \sim IG(a_1, b_1)$; $\sigma_e^2 \sim IG(a_2, b_2)$; $\mu \sim N(\mu_0,\, \sigma_0^2)$.

OBSERVED DATA: $Y_{ij} \quad (1 \le i \le K,\, 1 \le j \le J)$

TARGET DISTRIBUTION:
$\pi(\cdot) = \mathcal{L}(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K \,|\, \{Y_{ij}\})$.

Want to run a Gibbs sampler on $\pi(\cdot)$, i.e. on the $K+3$ variables $\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K$.

What are the conditional distributions?

Example (continued)

$$\mathcal{L}(\sigma_\theta^2 \mid \mu, \sigma_e^2, \theta_1, \ldots, \theta_K, Y_{ij}) \;=\; IG\left(a_1 + \frac{1}{2}K,\; b_1 + \frac{1}{2}\sum_i (\theta_i - \mu)^2\right) ;$$

$$\mathcal{L}(\sigma_e^2 \mid \mu, \sigma_\theta^2, \theta_1, \ldots, \theta_K, Y_{ij}) \;=\; IG\left(a_2 + \frac{1}{2}KJ,\; b_2 + \frac{1}{2}\sum_{i,j} (Y_{ij} - \theta_i)^2\right) ;$$

$$\mathcal{L}(\mu \mid \sigma_\theta^2, \sigma_e^2, \theta_1, \ldots, \theta_K, Y_{ij}) \;=\; N\left(\frac{\sigma_\theta^2 \mu_0 + \sigma_0^2 \sum_i \theta_i}{\sigma_\theta^2 + K\sigma_0^2},\; \frac{\sigma_\theta^2 \sigma_0^2}{\sigma_\theta^2 + K\sigma_0^2}\right) ;$$

$$\mathcal{L}(\theta_i \mid \mu, \sigma_\theta^2, \sigma_e^2, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_K, Y_{ij}) \;=\; N\left(\tfrac{J\sigma_\theta^2 \overline{Y}_i + \sigma_e^2 \mu}{J\sigma_\theta^2 + \sigma_e^2},\; \tfrac{\sigma_\theta^2 \sigma_e^2}{J\sigma_\theta^2 + \sigma_e^2}\right) .$$

The Gibbs sampler proceeds by updating the $K+3$ variables, in turn (either deterministic or random scan), according to the conditional distributions.

This is feasible since the conditional distributions are all easily simulated (IG and N).

In fact, it works well! [Gelfand and Smith, JASA, 1990]

All the algorithms constructed above can be combined in various different ways to produce more complex procedures tailored to different problems.

So, now we know how to construct (and run) lots of different MCMC algorithms. Good!

But do they converge to the distribution $\pi(\cdot)$? How quickly?

Stay tuned!

3 COMPUTER LABORATORY

4 ASYMPTOTIC CONVERGENCE OF MARKOV CHAINS

SUMMARY SO FAR:

We know how to construct Markov chain transition probabilities $P(x, \cdot)$ which have $\pi(\cdot)$ as a stationary distribution.

We then hope that, for large n, the distribution of X_n is close to $\pi(\cdot)$. But is it? Not necessarily!

EXAMPLE #1:

Suppose $\mathcal{X} = \{1, 2, 3\}$, with $\pi\{1\} = \pi\{2\} = \pi\{3\} = 1/3$. Let $P(1, \{2\}) = P(2, \{1\}) = 1/2$, and $P(3, \{3\}) = 1$. Let $X_0 = 1$.

Then $\pi(\cdot)$ is stationary. However, $X_n \in \{1, 2\}$ for all n, so $P(X_n = 3) = 0$ for all n, so $P(X_n = 3) \not\to \pi\{3\}$. No convergence! ("Reducible")

To avoid this problem, it suffices to have a single state x_*, which is "accessible" from all states x.

More generally, it suffices that the chain be "ϕ-irreducible" ...

ϕ-Irreducibility

Write $P^n(x, A)$ for the n-step transition law of the Markov chain:

$$P^n(x, A) = \mathbf{P}(X_n \in A \,|\, X_0 = x)\,.$$

DEFN: A chain is ϕ-irreducible if there exists a non-zero measure ϕ on $\mathcal{X}$ such that for all $A \subseteq \mathcal{X}$ with $\phi(A) > 0$, and for all $x \in \mathcal{X}$, there exists a positive integer $n = n(x)$ such that

$$P^n(x, A) > 0\,.$$

For example, if $\phi(A) = \delta_{x_*}(A)$, then this requires that x_* is accessible from any state x.

For a continuous Markov chain, $\phi(\cdot)$ might instead be e.g. Lebesgue measure.

Is ϕ-irreducibility the only property we require? No!

EXAMPLE #2:

Suppose again $\mathcal{X} = \{1, 2, 3\}$, with $\pi\{1\} = \pi\{2\} = \pi\{3\} = 1/3$. Let $P(1, \{2\}) = P(2, \{3\}) = P(3, \{1\}) = 1$. Let $X_0 = 1$.

Then $\pi(\cdot)$ is stationary, and the chain is ϕ-irreducible [e.g. take $\phi(\cdot) = \delta_1(\cdot)$]. However, $X_n = 1$ whenever n is a multiple of 3, so $P(X_n = 1)$ oscillates between 0 and 1, so $P(X_n = 1) \not\to \pi\{3\}$. Again no convergence! ("Periodic")

DEFN: The chain is <u>aperiodic</u> if there do not exist $d \geq 2$ and disjoint subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_d \subseteq \mathcal{X}$ with $\pi(\mathcal{X}_i) > 0$, such that $P(x, \mathcal{X}_{i+1}) = 1$ for all $x \in \mathcal{X}_i$ $(1 \leq i \leq d-1)$, and $P(x, \mathcal{X}_1) = 1$ for all $x \in \mathcal{X}_d$,

[In Example #2, have $d = 3$, so not aperiodic.]

Now we can state the main theorem!

DEFN: The <u>total variation distance</u> between two probability measures $\nu_1(\cdot)$ and $\nu_2(\cdot)$ is:

$$\|\nu_1(\cdot) - \nu_2(\cdot)\| = \sup_A |\nu_1(A) - \nu_2(A)| \,.$$

Theorem 4.1 *If a Markov chain is ϕ-irreducible and aperiodic, and has a stationary distribution $\pi(\cdot)$, then for π-a.e. $x = X_0 \in \mathcal{X}$,*

$$\lim_{n\to\infty} \|P^n(x, \cdot) - \pi(\cdot)\| \;=\; 0 \,.$$

In particular,

$$\lim_{n\to\infty} P^n(x, A) \;=\; \pi(A) \,, \quad A \subseteq \mathcal{X} \,.$$

Furthermore, for any $h : \mathcal{X} \to \mathbf{R}$,

$$\lim_{n\to\infty} (1/n) \sum_{i=1}^{n} h(X_i) \;=\; E_\pi[h(X)] \quad w.p.\ 1$$

Also, "usually" have a central limit theorem:

$$n^{1/2} \left(\frac{\sum_{i=1}^n h(\mathbf{X}_i)}{n} - \pi(h(X)) \right) \Rightarrow N(0,\, \sigma^2)$$

for some $\sigma^2 > 0$. (More later.)

In summary, the theorem says that if a chain is ϕ-irreducible and aperiodic, and has a stationary distribution $\pi(\cdot)$, then it will converge in distribution to $\pi(\cdot)$ from π-a.e. starting value. Good!

Now, in MCMC, always start with $\pi(\cdot)$ stationary – good.

Furthermore, usually easy to verify that chain is ϕ-irreducible, where e.g. ϕ is Lebesgue measure on appropriate region – good.

Also, aperiodicity almost always holds, e.g. for virtually any Metropolis algorithm or Gibbs sampler – good.

But why just "from π-a.e. starting value"?

EXAMPLE:

Let P be any ϕ-irreducible, aperiodic Markov chain on $\mathcal{X} = \mathbf{R}$, with continuous stationary distribution $\pi(\cdot)$.

Let P' be defined as follows. Let $P'(x, \cdot) = P(x, \cdot)$ whenever x is <u>not</u> a positive integer. For x a positive integer, let

$$P'(x, \cdot) = (1/x^2)\pi(\cdot) + (1 - 1/x^2)\, \delta_{x+1}(\cdot) \,.$$

Then $\pi(\cdot)$ is stationary for P', and P' is still ϕ-irreducible and aperiodic.

But if $X_0 = 3$ (say), then could have $X_n = n + 3$ for all n, so that $\|\mathcal{L}(X_n) - \pi(\cdot)\| \not\to 0$.

[Not "Harris recurrent".]

Harris Recurrence

DEFN: Say a chain is Harris recurrent if for all $B \subseteq \mathcal{X}$ with $\pi(B) > 0$, and all $x \in \mathcal{X}$,

$$\mathbf{P}[\exists n;\ X_n \in B \mid X_0 = x] = 1\,.$$

(Stronger than π-irreducibility.)

Theorem 4.2 *If chain Harris recurrent, then convergence theorem holds from every starting point (not just π-a.e. starting point).*

For example, this always holds if $P(x, dy) = p(x, y)\,\pi(dy)$, or for any Metropolis algorithm with a π-irreducible proposal.

Summary

For an MCMC algorithm to be valid to sample from $\pi(\cdot)$, we need that:

- $\pi(\cdot)$ is a stationary distribution (of course);
- chain is ϕ-irreducible;
- chain is aperiodic.

If these conditions all hold, then:

- chain will converge to $\pi(\cdot)$ in total variation distance;
- $\mathbf{P}[X_n \in A] \to \pi(A)$ for all $A \subseteq \mathcal{X}$;
- $\mathbf{E}[h(X_n)] \to \pi(h)$ for all functionals h having finite expectation;
- usually, normalised errors of $\sum_{i=1}^{n} h(X_i)$ will be approximately normal (CLT).

If chain Harris recurrent, then this is true from all starting values X_0; otherwise just from π-a.a.

5 QUALITATIVE RATES OF CONVERGENCE

SO FAR we have looked at questions of whether an algorithm has converged or not. In practice we may wish to know alot more about how quickly our algorithm converges and how efficient our resulting algorithm is.

It is usually very difficult to give precise statements of this form in any level of generality, but it is often possible to make useful qualitiative statements as we shall see. This section gives a brief overview of the theory of geometric ergodicity applied to MCMC.

In discrete state spaces, we can characterise geometric ergodicity in terms of the chain's return times to any given state. For general state spaces, we don't necessarily return to any one state, so we need to define a collection of states which have similar properties as defined below.

Standard reference for the Markov chain theory in this chapter is Meyn and Tweedie, (MCs and Stochastic Stability, 1993, Springer).

Small sets

A set $C \in \mathcal{B}$ is **small** if there exists a positive integer n, $\delta > 0$, and a probability ν such that the following **minorisation condition** holds.

$$P^n(x, A) \geq \delta\nu(A)$$

for all $x \in C, A \in \mathcal{B}$. From now on, we assume that $\mathcal{B}$ is countably generated.

Theorem 5.1 *(Deep!) If X is ϕ-irreducible, then* **every** *set $A \in \mathcal{B}$ with $\pi(A) > 0$ contains a small set C with $\pi(C) > 0$.*

Theorem 5.2 *(Shallow!) If X is ϕ-irreducible and π is invariant, then X is π-irreducible.*

- C is small if $C = \{x\}$.
- C is small if C is finite and X is π-irreducible and aperiodic.
- C is small if C is **compact** under suitable topological conditions. For example: X is *open-set irreducible* ($P_x(\tau_A < \infty) > 0,\ \forall x, \forall$ open non-empty A) **AND** X satisfies the **weak Feller** property.

The weak Feller propertly of a Markov chain just says that for any continuous bounded function f, $\mathbf{E}_x(f(X_1))$ is a continuous function of x.

Continuity stops the transition probability measure from 'changing much' in a bounded region.

Stability of Markov chains

Stability can be formulated in terms of

$$\tau_C = \min\{n \geq 1 :\ X_n \in X\}\ .$$

Another way of definining **Harris recurrence** requires that

$$P(\tau_C < \infty | X_0 = x) = 1 \quad \forall x$$

for SOME small set C.

Theorem 5.3 *If X is Harris recurrent, then there exists a (necessarily unique) invariant measure π:*

$$\pi(A) = \int_{\mathcal{X}} \pi(dx)P(x, A), \quad \forall A.$$

Theorem 5.4 *(ergodicity theorem) Suppose X is ϕ-irreducible and aperiodic. The following are equivalent.*

- *There exists an invariant* **probability** *measure π.*
- *(Local convergence) There exists a small set C with $\phi(C) > 0$ and*

$$P^n(x, C) \to P^\infty(C) > 0 \quad \forall x \in C.$$

- *There exists a small set C such that*

$$\sup_{x \in C} E_x[\tau_C] < \infty\ .$$

- *(Foster's drift condition) There exists a small set C and a function $V(x) \geq 0$ with $V(x_0) < \infty$ for some $x_0 \in \mathcal{X}$ such that*

$$\int P(x, dy)V(y) \leq V(x) - 1 + b\mathbf{1}_C(x) \quad \forall x \in \mathcal{X}$$

When any of the above holds,

- $V_0 = \{x : V(x) < \infty\}$ is absorbing, and $\pi(V_0^c) = 0$.
- $\exists C_n \uparrow V_0$ with each C_n small, such that

$$\sup_{x \in C_n} E_x[\tau_{C_n}] < \infty \quad \forall n \ .$$

- $\forall x \in V_0$,

$$\|P^n(x, \cdot) - \pi\| \to 0 \ . \qquad (1)$$

- If X is Harris recurrent then (1) holds for all x.
- If $V < \infty$ everywhere, then (1) holds for all x.

Convergence of moments

For MCMC applications, we're typically interested in estimating $E_\pi[f(X)]$ for a collection of functions f. Let

$$S_n(f) = \sum_{j=1}^{n} f(X_j) \ .$$

Theorem 5.5 *If π is an invariant probability measure, the following are equivalent*

- $\forall f \in L_1(\pi)$,

$$\lim_{n \to \infty} \frac{1}{n} S_n(f) = \int f(y)\pi(dy) \quad a.s.[P_x], \quad \forall x;$$

- *X is Harris recurrent.*

A simple Cauchy example

$\pi(x) \propto (1 + x^2)^{-1}$ on the positive real axis. We use a random walk Metropolis algorithm with a Gaussian proposal distribution tuned to have around 30% of its moves accepted.

You may have looked at exactly this problem during the computer lab. Completely trivial problem, right?

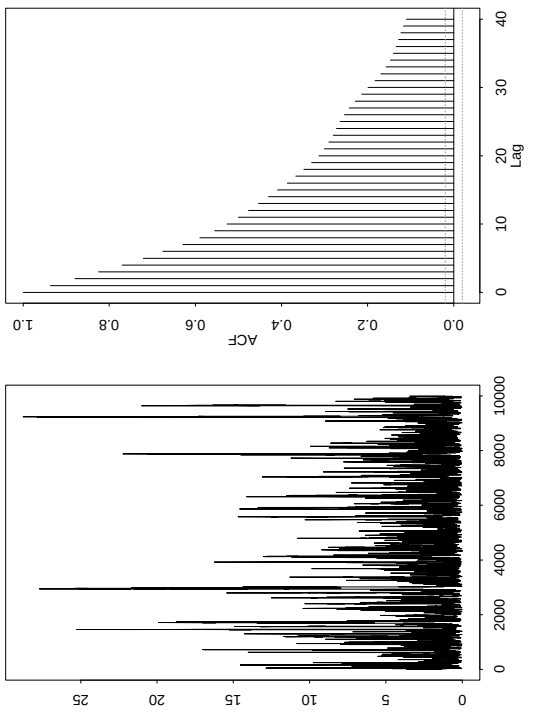


Figure 1: Does this look like reasonable output for a Cauchy distribution?

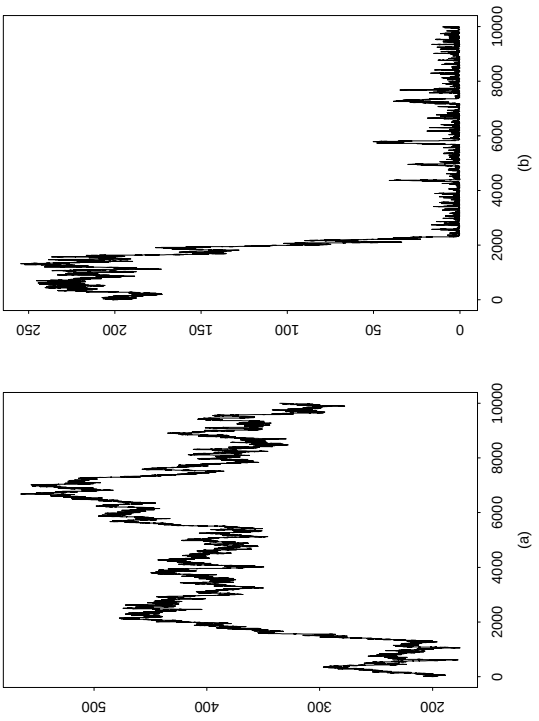


Figure 2: Random walk Metropolis on the Cauchy started out in the tail, showing the unstable heavy

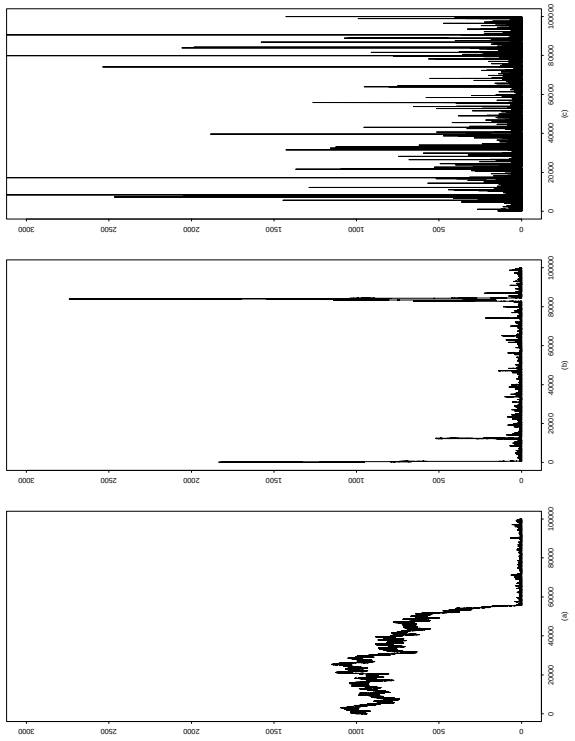


Figure 3: Three RWM algorithms on Cauchy example: (a) Gaussian proposal, (b) Cauchy proposal, (c) multiplicative RW Gaussian proposal

A minimal requirement for any sensible algorithm with transition probabilities P is that of ergodicity: that is for all $\mathbf{x} \in \mathcal{X}$,

$$\|P^n(\mathbf{x}, \cdot) - \pi(\cdot)\| = r(\mathbf{x}, n) \downarrow 0 \ . \qquad (2)$$

In fact it is easy to demonstrate that all three of the examples considered in the Cauchy example satisfy this requirement.

Why? Just show it is Lebesgue irreducible, aperiodic, and Harris recurrent in each of these three cases.

We need to consider more refined conditions on $r(\cdot, \cdot)$ in order to compare these methods.

Example: Independence sampler

$\pi(x) = e^{-x}$, $q(x) = ke^{-kx}$. We consider 2 possible algorithms:

1. $k = 0.01$
2. $k = 5$

Which will perform better?

Both algorithms were run for 1 million iterations started at the mean value of π, ie 1 in this case. The experiment was repeated 55 times for each case producing the following results.

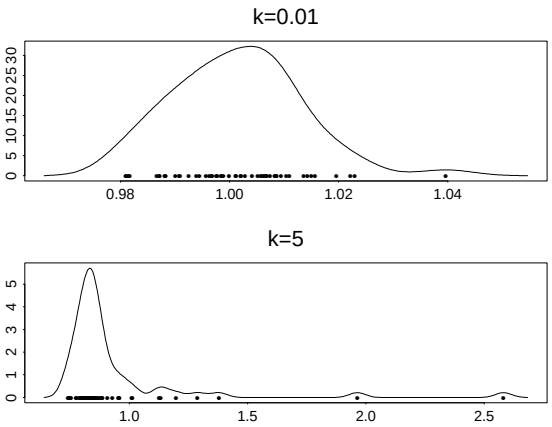


Figure 4: Sample means of each of the 110 runs. Monte carlo error in the case $k = 0.01$ is fairly small and symmetric about zero. However in the $k = 5$ case, the error is massively skewed, and most runs give very biased resuts.

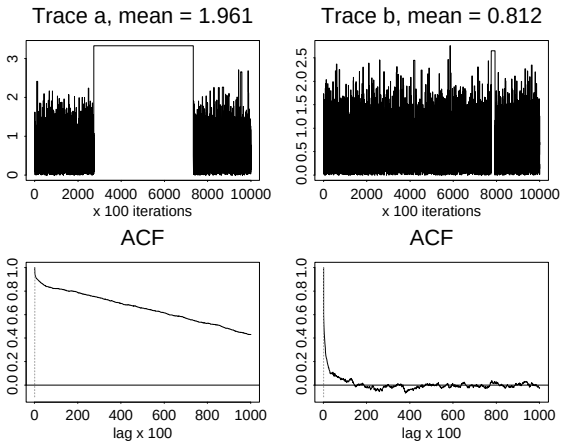


Figure 5: Two sample paths from the $k = 5$ simulation study.

Converging geometrically quickly

Recall, for Harris chains.

$$\|P^n(x,\cdot) - \pi\| \to 0, \quad \forall x$$

Natural questions are the following

- When is the chain **geometrically ergodic**?

  $$\|P^n(x,\cdot) - \pi\| \leq M(x)\rho^n,$$

  $M(x) < \infty$, $\rho < 1$.

- When is the chain **uniformly ergodic**?

  $$\|P^n(x,\cdot) - \pi\| \leq r(n) \to 0, \quad n \to \infty \qquad (3)$$

  Remark: In fact, that if (1) holds, then we can always take constants $\rho < 1$ and M such that $r(n) \leq M\rho^n$.

Therefore uniform ergodicity $\Rightarrow$ geometric ergodicity.

Uniform ergodicity equivalences

Theorem 5.6 *For a ϕ-irreducible aperiodic chain, the following are all equivalent.*

- $\|P^n(x,\cdot) - \pi\| \leq M\rho^n$, $M < \infty$, $\rho < 1$, $\forall x \in \mathcal{X}$
- *There exists a small set C with*

  $$\sup_{x \in \mathcal{X}} E_x[\tau_C] < \infty$$

- *There exists a small set C and $\beta > 1$ with*

  $$\sup_{x \in \mathcal{X}} E_x[\beta^{\tau_C}] < \infty$$

- *$\mathcal{X}$ is small*

Theorem 5.7 *When $\mathcal{X}$ is small, there exists $m \geq 1$, $\delta > 0$, and a probability measure ν such that*

$$P^m(x, A) \geq \delta\nu(A), \quad \forall x, \forall A \ , \tag{4}$$

then for n such that $m|n$

$$\|P^n(x, \cdot) - \pi\| \leq 2(1-\delta)^{n/m} \ . \tag{5}$$

PROOF Use coupling of two 'copies' of the chain, one started from x, the other from the stationary measure π.

$$X_0 = x, \quad X_1, \ldots \quad : \text{so } X_n \sim P^n(x, \cdot)$$

$$X_0' \sim \pi, X_1', \ldots \quad : \text{so } X_n' \sim \pi \ .$$

Take $m = 1$ in (4). Then

$$P(y, A) \geq \delta\nu(A) \ \ \forall y.$$

Toss a 'δ-coin'. If 'heads', move from y with distribution ν; if 'tails', move from y with distribution

$$\frac{1}{1-\delta}\left[P(y, \cdot) - \delta\nu(\cdot)\right] \ .$$

This preserves the marginal distribution of the chains. However now we use the **same** coin for both chains X and X', so that as soon as a head is achieved the two chains remain equal forever more.

Now we use the **coupling inequality**:

$$T = \inf(n \geq 1 : \ \delta - \text{coin is head}).$$

$$\begin{aligned}\|P^n(x, \cdot) - \pi\| &= 2 \sup_{A \in \mathcal{B}} |P^n(x, A) - \pi(A)| \\ &\leq 2P(X_n = X_n') \\ &= 2P(T > n) \\ &= 2(1-\delta)^n\end{aligned}$$

■

This proof idea extends in a number of ways including

- geometrically (but not uniformly) ergodic chains;
- coupling from the past.

MCMC examples

Independence sampler with proposal density $q(x, y) = q(y)$.

Theorem 5.8 *(Mengersen-Tweedie) Suppose*

$$\frac{q(y)}{\pi(y)} \geq \beta > 0, \quad \forall y \in \mathcal{X} \tag{6}$$

then

$$\|P^n(x, \cdot) - \pi\| \leq 2(1-\beta)^n \ .$$

Conversely, if $\text{essinf}_\pi q(y)/\pi(y) = 0$, then P is not even geometrically ergodic.

PROOF Exercise! Just show that the minorisation condition holds.

So for the independence sampler to be uniformly ergodic, the proposal distribution needs to have tails that ar 'at least as heavy' as the target density. Note, if (6) holds, then (recall from Lecture 1) rejection sampling is possible.

If $\mathcal{X}$ is compact, and the chain is 'suitably continuous' (weak Feller plus a bit more), then $\mathcal{X}$ is small and so X is uniformly ergodic.

Example RWM on truncated normal.

$$\pi(x) \sim \frac{1}{(2\pi)^{1/2}} e^{-x^2/2} \mathbf{1}_{[-x_0, x_0]}(x)$$

$$\begin{aligned}Q(x, y) &= N(x, 1)\mathbf{1}_{[-x_0, x_0]}(x) \\ &= \frac{1}{(2\pi)^{1/2}} e^{(y-x)^2/2} \mathbf{1}_{[-x_0, x_0]}(x) \\ &\geq e^{-2x_0^2} \mathbf{1}_{[-x_0, x_0]}(x) \mathbf{1}_{-x_0, x_0}(y) \qquad (x \in [-x_0, x_0])\end{aligned}$$

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)} \geq e^{-x_0^2}$$

So we have directly verified the minorisation condition:

$$P(x, dy) \geq e^{-2x_0^2} \times e^{-x_0^2} \mathbf{1}_{[-x_0, x_0]}(y) dy \ .$$

Clearly these bounds can be improved on. In practice, even when chains are uniformly ergodic, δ can be extremely small, and better bounds on convergence time can sometimes be obtained by just using the weaker geometric ergodicity assumption.

Geometric ergodicity

Theorem 5.9 *Suppose C is ϕ-irreducible and aperiodic with invariant measure π. The following are equivalent*

- *There exists a small set C with $\phi(C) > 0$ such that for all $x \in C$:*

$$|P^n(x, C) - \pi(C)| \leq M_C \rho_C^n$$

  $M_C < \infty$, $\rho_C < 1$, (local geometric convergence).

- *There exists a small set C, a constant $\kappa > 1$ and a finite constant M_C such that*

$$\sup_{x \in C} E_x[\kappa^{\tau_C}] \leq M_C$$

  (geometric return times).

- There exists a small set C, and constants $0 < \lambda < 1$, $b < \infty$ and a function $V \geq 1$ which is finite at least for some x_0 say, then

$$\int_{\mathcal{X}} P(x, dy) V(y) \leq \lambda V(x) + b \mathbf{1}_C(x) \qquad (7)$$

  (geometric Foster-Lyapunov condition)

When these conditions hold,
$V_\lambda := \{x : V(x) < \infty\}$ is absorbing and $\pi(V_\lambda^c) = 0$ and for all $x \in V_\lambda$

$$\sup_{|g| \leq V} |\int P^n(x, dy) g(y) - \int \pi(dy) g(y)| \leq M V(x) \rho^n$$

for some $M < \infty$, $\rho < 1$.

So

- We get geometric convergence from all points in V_λ. So if V is finite everywhere, we get geometric convergence from everywhere.
- We can pick ρ in the above independently of x.
- This also identifies the dependence of convergence on the initial state x via the function V

Strategy for proving or disproving geometric ergodicity:

- To **prove** geometric ergodicity, find a drift function. For MCMC a generic choice turns out to be $\pi^{-d},\ 0 < d < 1$.
- To **disprove** geometric ergodicity, show that the return times to a small set of positive mass (according to π) is slower than exponential. Two strategies for doing this:
  - show that the Markov chain behaves 'like a random walk' out in the tails;
  - show that $P(x, \{x\})$ is NOT bounded away from zero.

MCMC and drift conditions

Why are drift function techniques appropriate for many statistical MCMC problems?

1. Because target densities frequently have a modal region towards which algorithms drift.
2. Drift conditions can often be calculated or approximated using only pocal properties of the target density.
3. $\pi(\mathbf{x})^{-d}$, for $0 < d$ proves to be a natural drift condition in many cases.

Example: Random walk Metropolis

Suppose π is a density on the real line with positive continuous density. Now if π has tails **heavier** that exponential (say $(\log(\pi(x))' \to 0$, for example the tails of π 'look like' x^{-r} for some $r > 1$). Fix $M > 0$. Then for $|y - x| < M$ say,

$$\log \pi(y) - \log \pi(x) = \int_x^y (\log(\pi(x))'(z) dz$$

which will be small for large positive or negative x. So $|y - x| < M$,

$$1 \wedge \frac{\pi(y)}{\pi(x)} \approx 1$$

and the algorithm behaves like a driftless random walk in the tails. Therefore geometric ergodicity fails.

Now suppose that there exists $\alpha > 0$ and x_1 such that for all $y \geq x \geq x_1$,

$$\log \pi(x) - \log \pi(y) \geq \alpha(y - x) \ ,$$

that is

$$\frac{\pi(x)}{\pi(y)} \geq \frac{e^{-\alpha x}}{e^{-\alpha y}} \ .$$

Theorem 5.10 *(Mengersen-Tweedie) Under these conditions, then for any mean zero proposal with continuous density $q(\cdot)$, the algorithm is geometrically ergodic, but NOT uniformly ergodic.*

PROOF Use the Lyapunov drift function $V(x) = e^{s|x|}$ for some $0 < s < \alpha$.

These results generalise to higher dimensions (Roberts-Tweedie, 1996, Biometrika) though many more things can (and do) go wrong so that the 'exponential tails' condition is no longer sufficient for geometric ergodicity.

Behaviour of random walk Metropolis

(a) For target densities with tails uniformly bounded by exponentials in all directions are geometrically ergodic so long as their tails are sufficiently regular. One condition which is sufficient for this requires the contours of the target density to have curvature which converges to 0 as they move further away from the target density mode.

(b) If π has tails which are strictly heavier than exponential, then no random walk Metropolis algorithm can be geometrically ergodic.

(c) If the tails of π resemble $|\mathbf{x}|^{-(d+r)}$ in the tails, then light tailed Metropolis proposals lead to algorithms which converge at polynomial rate $r/2$ (that is $\|P^n(x,\cdot) - \pi\| \leq c/n^{r/2}$). Heavy tailed versions of the algorithms can increase the polynomial rate arbitrarily.

What can go wrong?

Geometric ergodicity is related to exponential moments of return times to a small set C.

There are two obvious ways in which exponential moments can fail to exist for a random walk Metropolis algorithm.

(1) The algorithm might have 'sticky' patches:

$$P(\mathbf{x}_i, \{\mathbf{x}_i\}) \to 1 \ , \tag{8}$$

along a suitable sequence of points $\{\mathbf{x}_i\}$. Then return times to small sets cannot have geometric moments.

Unfortunately, this phenomenon is very commmon. As an simple example consider the normal-Gamma density prevalent in Bayesian analysis of linear models:

$$\pi(\mu, \tau) \propto \tau^{n/2} \exp\{-\tau(S + n(\mu - \bar{x})^2)/2\} \ .$$

for suitable constants $n, S, \bar{x}$.

$P((\mu,\tau), \{(\mu,\tau)\}) \to 1$ along the ridges of high probability for large $|\mu|$, and so the algorithm fails to be geometrically ergodic.

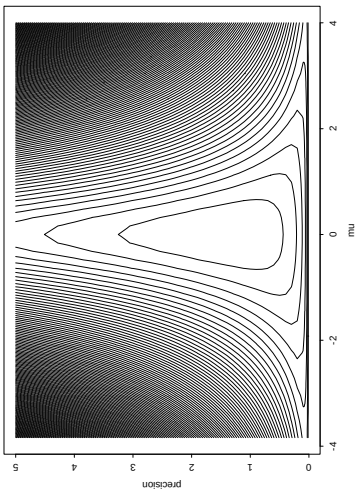


Figure 6: A contour plot of the contours of the posterior distribution in the normal-Gamma example.

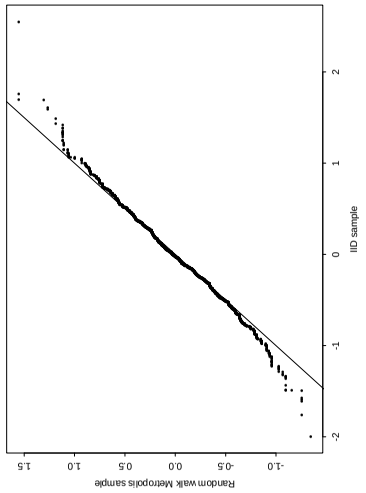


Figure 7: A q-q plot illustrating that the random walk Metropolis algorithm fails to adequately explore the tails of the normal-Gamma distribution.

What can go wrong?

(2) Another way in which random walk algorithms can fail to be geometrically ergodic is where the effect of the accept reject mechanism becomes negligible in the tails, so that the algorithm approximates a null-recurrent random walk. This is essentially what fails for heavy tailed target densities such as the Cauchy example.

Emphasis here on random walk Metropolis algorithm. However, this work easily extends to other algorithms such as Langevin, hybrid Metropolis methods, Gibbs, Independence sampler, etc.

Theorem 5.11 *Suppose that π is a $k-$dimensional density, and for each i, P_i is a Markov chain which updates just the ith coordinate. Consider running a random scan of the P_i's, that is a chain P with*

$$P = \frac{P_1 + P_2 + \ldots P_k}{k} \ .$$

Suppose that for some component i say, P_i is a random walk Metropolis algorithm with fixed increment proposal density q, and that

$$\lim_{K \to \infty} \frac{\log \pi(X_i \in (K, \infty))}{K} = 0 \ , \qquad (9)$$

then P fails to be geometrically ergodic.

The peculiar condition above can just be interpreted as the condition that the marginal distribution of the ith component has heavy tail.

Return to Cauchy example

$$\pi(x) \sim x^{-2} \ .$$

approximately in the tail.

Gaussian proposal random walk Metropolis The simple random walk Metropolis algorithm with light tailed proposal satisfies the drift condition for values of α less than $\alpha_0 = 1/3$. Don't get CLTs even for bounded functions from the CLT theorem.

Cauchy proposal random walk Metropolis illustrated in earlier figure satisfies the drift condition for all $\alpha < \alpha_0 = 1/2$. Nnot quite enough to ensure that CLTs hold for all bounded functions. (Any slightly heavier tailed proposal would achieve this.) However it is considerably more stable than the light tailed proposal case.

Multiplicative random walk Metropolis algorithm is geometrically ergodic, since the tails of the log of a Cauchy random variable are bounded by an exponential. Thus the algorithm is considerably more stable than either of the first two algorithms. It satisfies the drift condition for $\alpha = 1$ and therefore it is geometrically ergodic. CLTs hold for all square integrable functions.

Central limit theorems

The algorithm's convergence properties are closely linked to those of its excursions away from small sets and with the existence of CLT for the Markov chain, which are important for Monte Carlo implementation.

We say that a $\sqrt{n}$-CLT exists for a function f, if

$$n^{1/2}\left(\frac{\sum_{i=1}^{n} f(\mathbf{X}_i)}{n} - \mathsf{E}_\pi(f(X))\right) \Rightarrow \tag{10}$$

$$N(0, \tau_f \mathrm{Var}_\pi(f(X))) \tag{11}$$

where τ_f denotes the *integrated auto-correlation time* for estimating the function f using the Markov chain P.

Theorem 5.12 *CLT for Markov chains.*

If P is geometrically ergodic and reversible, then for all functions f for which $Var_\pi(f(X))$ is finite, a $\sqrt{n}$-CLT exists.

For polynomially ergodic MCs, CLTs can be shown to exist only for functions which do not grow too rapidly. For non-reversible chains, less clean but similar results are known.

6 COMPUTABLE QUANTITATIVE CONVERGENCE BOUNDS

The Set-Up

Have Markov chain $P(x, \cdot)$, with stationary distribution $\pi(\cdot)$.

Know that if ϕ-irreducible and aperiodic,

$$\lim_{n\to\infty} \|P^n(x,\cdot) - \pi(\cdot)\| \;=\; 0\,.$$

Good. But how large does n need to be to make $\|P^n(x,\cdot) - \pi(\cdot)\|$ small?

That is, how long do we need to run the algorithm to get approximate convergence to $\pi(\cdot)$?

Ideally, we would like to find explicit n for which we can prove that, say,

$$\|P^n(x,\cdot) - \pi(\cdot)\| < 0.01\,.$$

Can we do this?

Recall minorisation condition:

$$P(x,\cdot) \geq \epsilon\nu(\cdot) \qquad x \in C\,.$$

Also need "bivariate drift condition":

$$\overline{P}h(x,y) \;\leq\; h(x,y)\,/\,\alpha\,, \qquad (x,y) \notin C\times C$$

for some $h : \mathcal{X}\times\mathcal{X} \to [1,\infty)$, and $\alpha > 1$, where

$$\overline{P}h(x,y) \;\equiv\; \int_{\mathcal{X}}\int_{\mathcal{X}} h(z,w)\,P(x,dz)\,P(y,dw)\,.$$

Finally, define

$$B = \max[1,\, \alpha(1-\epsilon)\sup_{C\times C}\overline{R}h]\,,$$

where for $(x,y) \in C\times C$,

$$\begin{aligned}\overline{R}h(x,y) \;&= \textstyle\int_{z\in\mathcal{X}}\int_{w\in\mathcal{X}}(1-\epsilon)^{-2}h(z,w)\\ &\quad\times(P(x,dz)-\epsilon\nu(dz))\,(P(y,dw)-\epsilon\nu(dw))\,.\end{aligned}$$

Main Result

Theorem 6.1 *Then for any integers* $1 \leq j \leq n$,

$$\|\mathcal{L}(X_n) - \mathcal{L}(X_n')\|_{TV} \;\leq$$
$$(1-\epsilon)^j + \alpha^{-n}B^{j-1}\,E_{\mu\times\pi}[h(X_0,X_0')]\,,$$

where $\mu = \mathcal{L}(X_0)$.

[Rosenthal, 1995; Roberts and Tweedie, 1999; Rosenthal, 2002; other variations available.]

Thus, if we can very explicit minorisation and drift conditions, then we can get explicit bounds on how large n has to be to make, say,

$$\|P^n(x,\cdot) - \pi(\cdot)\| \;<\; 0.01\,.$$

But can we apply this to real examples?

SOME FURTHER SIMPLIFICATIONS:

Have that

$$B \leq \max[1,\, \alpha(B_0 - \epsilon)]\,,$$

where

$$B_0 = \sup_{(x,y)\in C\times C}\hat{P}h(x,y)\,;$$

here

$$\hat{P} = \epsilon(\nu\times\nu) + (1-\epsilon)\overline{R}\,.$$

Also, if $PV(x) \leq \lambda V(x) + b\mathbf{1}_C(x)$ for all $x \in \mathcal{X}$, where $V : \mathcal{X} \to [1,\infty)$, and $C = \{x \in \mathcal{X};\; V(x) \leq d\}$, and $PV(x) = \mathbf{E}[V(X_{n+1}) \,|\, X_n = x]$, then can take

$$h(x,y) = (V(x) + V(y)/2\,,$$

$$\alpha^{-1} = \lambda + \frac{b}{d+1}\,,$$

$$B_0 = \lambda d^* + b\,.$$

Very explicit!

EXAMPLE: Hierarchical Poisson Model ("pump failures"). [Gelfand-Smith, 1990; Tierney, 1994]

$$\begin{array}{ccccc} & & \beta & & \beta \sim G(\gamma, \delta) \\ & \swarrow & \downarrow & \searrow & \\ \theta_1 & \ldots & \ldots & \theta_n & \theta_j \sim G(\alpha_0, \beta) \\ \downarrow & \ldots & \ldots & \downarrow & \\ Y_1 & \ldots & \ldots & Y_n & Y_j \sim Pois(\theta_j T_j) \end{array}$$

Gibbs Sampler run on $\beta, \theta_1, \ldots, \theta_n$, conditional on observed Y_j, T_j. (Updating distributions: Gamma)

Theorem 6.2 (Rosenthal, JASA, 1995) *: For data and priors as in Gelfand and Smith (1990), ($n = 10$), initial distribution μ_0, we have*

$$\|\mathcal{L}(X^{(k)}) - \pi(\cdot)\|_{\text{var}} \;\leq\; (0.976)^k + (0.951)^k (6.2 + E)\,,$$

where $E = E_{\mu_0}\left(\sum_j \theta_j^{(0)} \;- 6.5\right)^2$.

(e.g. $E = 2$, $k = 200$, bound is 0.008. Numerical work suggests convergence actually occurs earlier, perhaps around $k = 10$ or $k = 20$.)

EXAMPLE: related to James-Stein estimators [suggested by Jun Liu]

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

Gibbs sampler run on $\mu, A, \theta_1, \ldots, \theta_K$. (Conditionals: IG and N.)

Theorem 6.3 *[Rosenthal, Stat. and Comp., to appear]:*

For priors $\mu \sim$ flat, $A \sim IG(a, b)$, for any $0 < r < 1$, we have

$$\begin{aligned} \|\mathcal{L}(X^{(k)}) \;-\; \pi(\cdot)\| \;&\leq\; (1-\epsilon)^{rk} \\ &+\; \left(\alpha^{-(1-r)}\gamma^r\right)^k \;\left(1 + \tfrac{\Lambda}{1-\lambda} + \mathbf{E}\left(f(X^{(0)})\right)\right), \end{aligned}$$

where

$$f(x) \;=\; \sum_{i=1}^{K} (\theta_i - \overline{Y})^2\,;$$

$$\lambda \;=\; \mathbf{E}\left(1 + \frac{W}{V}\right)^{-2} \quad \text{with } W \sim IG\left(a + \frac{K-1}{2}, b\right)$$

$$\Lambda = \Delta + (K + \tfrac{2}{9})V, \quad \Delta = \textstyle\sum (Y_i - \overline{Y})^2,$$

$$\begin{aligned} \epsilon \;=\; \; & 2\textstyle\int_0^\infty dA \;\min\Big[IG\left(a + \frac{K-1}{2}, b; A\right), \\ & IG\left(a + \frac{K-1}{2}, b + \frac{d}{2}; A\right)\Big] \;\times \\ & \times\; \textstyle\int_0^\infty d\mu \; N\left(\sqrt{\frac{d}{K}}, \frac{A}{K}; \mu\right), \end{aligned}$$

$$\alpha^{-1} \;=\; \frac{1 + 2\Lambda + \lambda d}{1 + d} \;<\; 1\,; \quad \gamma \;=\; 1 + 2(\lambda d + \Lambda)\,.$$

Gives general formula bounding distance to stationarity in terms of prior values, data, initial distribution.

[e.g. baseball data of Efron-Morris, with appropriate priors, bound equals 0.009 for $k = 140$ iterations.]

Diagnosing Convergence

Theoretical computable bounds sometimes work well.

However, for complicated models, difficult to verify drift and minorisation conditions.

Can sometimes "approximately" verify drift and minorisation conditions using auxiliary simulation [Cowles and R., Stat and Comp 1998].

Otherwise, need to use "convergence diagnostics": do statistical analysis (or just informal observation) on Markov chain's output to hopefully conclude that convergence has occurred.

e.g. Gelman and Rubin (1987, Stat Sci): Run the Markov chain from many different starting values X_0 ("overdispersed starting distribution"). Hopefully converged when "inter-chain variances" comparable to "intra-chain variances".

No guarantees, though!

Example: "Witch's Hat"

$\mathcal{X} = [0,1]^d \quad (d \text{ large})$

$\pi_u(\mathbf{x}) = 1 + \delta^{-d+1}\mathbf{1}_S(\mathbf{x})$, where $\delta > 0$ very small, and

$$S = \{\mathbf{x} \in \mathcal{X} : \ x_i < \delta \ \forall i\}\,.$$

Then $\pi(S) \approx 1$.

However, unless $X_0 \in S$, or "get lucky" and find $X_n \in S$, then Gibbs Sampler or Metropolis algorithm may well miss S entirely.

Convergence diagnostics would suggest $\pi(\cdot) \approx \text{Uniform}(\mathcal{X})$. Wrong!!

Chain converges extremely slowly, but is still geometrically ergodic. Misleading!!

Overall, the "convergence time problem" remains largely unresolved ... but usually okay in practice.

7 OPTIMAL SCALING OF METROPOLIS-HASTINGS ALGORITHMS

Let $\pi : \mathbf{R}^d \to [0, \infty)$ be a d-dimensional density (d large).

Consider running a Metropolis-Hastings algorithm for π. How should we choose the proposal?

- RWM: Consider proposal $N(\mathbf{X}_n, \sigma^2 I_d)$.
- Langevin: Consider proposal $N(\mathbf{X}_n + \frac{\sigma^2}{2}\nabla \log \pi(\mathbf{X}_n),\ \sigma^2 I_d)$.

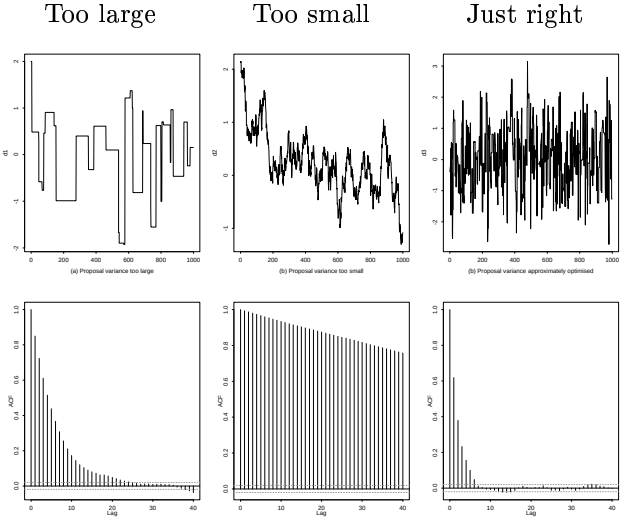
In either case, how to choose σ^2??

If σ^2 too small, the chain never goes anywhere.

If σ^2 too large, the chain usually rejects.

The Goldilocks Principle: Need the proposal scaling to be "just right".

The Goldilocks Principle illustrated:



[Trace plots (top) and auto-correlation plots]

How do we make a theory out of this?

For simplicity, assume (for now) that

$$\pi(\mathbf{x}) \;=\; \prod_{i=1}^{d} f(x_i)\,,$$

i.e. that the density π factors into i.i.d. components, each with (smooth) density f.

Also, assume that chain is in stationarity, i.e. that $X_0 \sim \pi(\cdot)$.

Also assume that either

$$Q(\mathbf{x}, \cdot) \sim N(\mathbf{x}, \sigma_d^2 I_d)$$

for RWM, or

$$Q(\mathbf{x}, \cdot) \sim N(\mathbf{X}_n + \frac{\sigma^2}{2}\nabla \log \pi(\mathbf{X}_n),\ \sigma^2 I_d)$$

for Langevin.

For RWM, let $I = \mathbf{E}[((\log f(Z))')^2]$ where $Z \sim f(z)\,dz$. Then as $d \to \infty$, it is optimal to choose $\sigma^2 \doteq 2.38/I^{1/2}d$, leading to an asymptotic acceptance rate $\doteq 0.234$.

More formally, set $\sigma_d^2 = \ell^2/d$, and let

$$Z_t^d = X_{[dt]}^{(1)}\,.$$

Thus, $\{Z_t\}$ follows the first component of $\{X_n\}$, with time speeded up by a factor of d.

Then as $d \to \infty$,

$$Z_d \Rightarrow Z$$

where Z satisfies the SDE,

$$dZ_t = h(\ell)^{1/2} dB_t + \frac{h(\ell)\nabla \log \pi(Z_t)}{2} dt\,,$$

where

$$h(\ell) = \ell^2 \times 2\Phi\left(-\frac{\sqrt{I}\ell}{2}\right) = \ell^2 \times A(\ell)\,,$$

where $A(\ell)$ is the asymptotic acceptance rate of the algorithm. Above choice of ℓ maximises $h(\ell)$.

For Langevin, again take $Z \sim f(z)\,dz$, and let let $J = \mathbf{E}[(5((\log f(Z))''')^2 - 3((\log f(Z))'')^3)/48]$.

Then as $d \to \infty$, it is optimal to choose $\sigma^2 \doteq 0.825/J^{1/2}d^{1/3}$, leading to an asymptotic acceptance rate $\doteq 0.574$.

More formally, set $\sigma_d^2 = \ell^2/d^{1/3}$, and let

$$Z_t^d = X_{[d^{1/3}t]}^{(1)}\,.$$

Then as $d \to \infty$,

$$Z_d \Rightarrow Z$$

where Z satisfies the SDE,

$$dZ_t = g(\ell)^{1/2} dB_t + \frac{g(\ell)\nabla \log \pi(Z_t)}{2} dt\,,$$

where

$$g(\ell) = 2\ell^2 \Phi\left(-J\ell^3\right) = \ell^2 \times A(\ell)\,,$$

where $A(\ell)$ is the asymptotic acceptance rate of the algorithm. Above choice of ℓ maximises $g(\ell)$.

Hence, for either RWM or Langevin algorithm, can determine optimal scaling just in terms of asymptotic acceptance rate! (0.234 for RWM, 0.574 for Langevin) Good! (Compare with "trial and error".)

Also, in high dimensions, efficiency of RWM scales like d^{-1}, while efficiency of Langevin scales like $d^{-1/3}$ (better). [Note: scaling of $\sigma_d^2 = O(1/d^{1/3})$ was suggested in some contexts by physicists Kennedy and Pendleton (1991).]

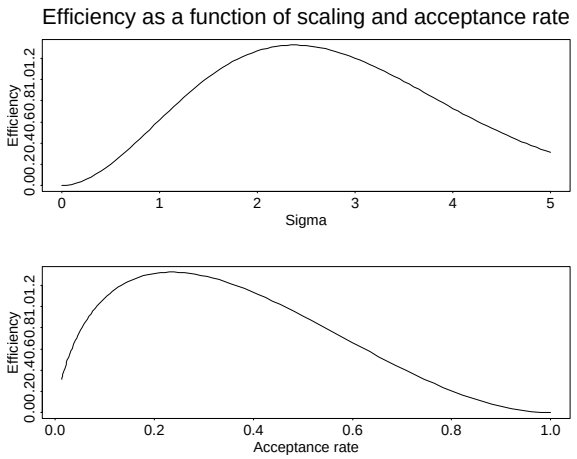
Result still holds if instead

$$\pi(\mathbf{x}) = \prod_{i=1}^{d} f_i(x_i)\,,$$

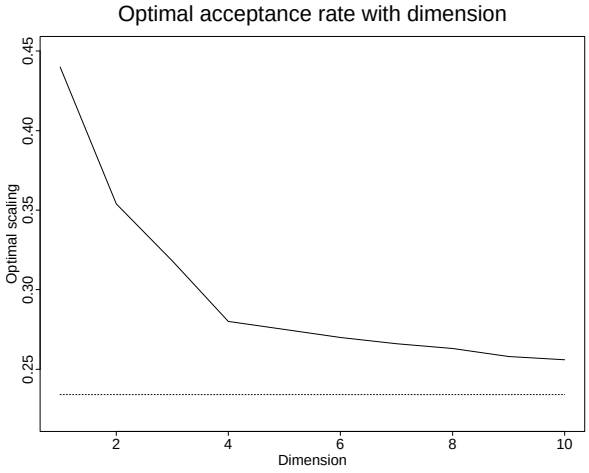
provided that (say) $f_i \to f_\infty$.

Similar results for random walk on discrete hypercube (Roberts), and finite-range homogeneous Markov random fields (Breyer and Roberts).

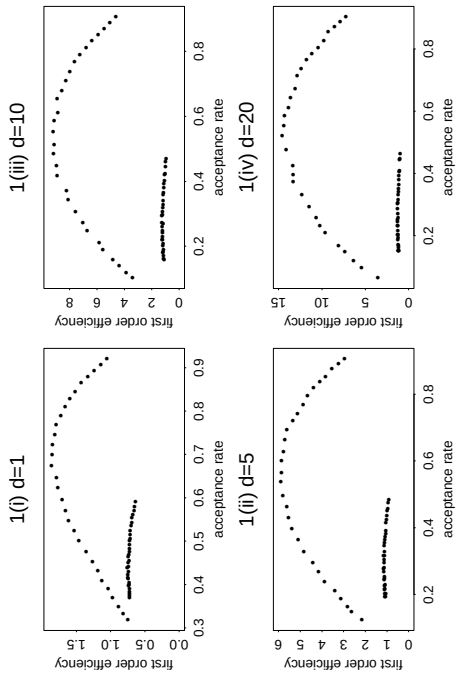
Note that we don't need the acceptance rate to be exactly 0.234 (or 0.574), just close:



Also, note that dimension doesn't have to be too large before asymptotics kick in:



Also, Langevin algorithms are significantly more efficient than RWM algorithms:



What is "efficiency"?

Let $\{X_n\}$ be a Markov chain.

Then for a π-integrable function f, "efficiency" is achieved by minimising

$$\mathrm{Var}_f \;=\; \lim_{n\to\infty} \frac{1}{n} \mathrm{Var}\left(\sum_{i=1}^{n} f(X_i)\right) .$$

In general relative efficiency between two possible Markov chains varies depending on what function of interest f is being considered.

The above results say that in various cases, as $d \to \infty$, the dependence on f disappears.

Hence, the optimal proposal scaling also leads to mimising Var_f, for any function $f : \mathcal{X} \to \mathbf{R}$ with $\pi(f^2) < \infty$.

Targets with heterogenous scaling

Above results only proved if

$$\pi(\mathbf{x}) = \prod_{i=1}^{d} f_i(x_i)\,,$$

where $f_i \to f_\infty$. What if not?

Suppose

$$\pi(\mathbf{x}) = \prod_{i=1}^{d} C_i\, f(C_i x_i)\,,$$

where $C_1 = 1$, and $\{C_i\}_{i=2}^{\infty}$ are i.i.d. positive r.v. with $E(C_i^2)/E(C_i)^2 \equiv b < \infty$.

Let $W_t^d = X_{[td]}^{(1)}$.

What does $\{W_t\}$ converge to as $d \to \infty$?

Theorem 7.1 *Let $\{X_n\}$ be RWM, and let $W_t^d = X_{[td]}^{(1)}$. Then as $d \to \infty$, W_t^d converges weakly to a limiting diffusion process W_t satisfying*

$$dW_t = \frac{1}{2} g'(W_t)(C_1 s)^2 dt + (C_1 s) dB_t\,,$$

where B_t is standard Brownian motion, and where

$$s^2 = 2\ell^2 \Phi(-\ell b^{1/2} I^{1/2}/2) = \frac{1}{b} \times 2(\ell^2 b)\Phi(-(\ell^2 b)^{1/2} I^{1/2}/2)\,,$$

with $I = E_f[((\log f(Z))')^2]$.

Hence, the efficiency of the algorithm (when considering functionals of the first coordinate only), as a function of acceptance rate, is identical to that for i.i.d. traget densities, except multiplied by the global factor of $1/b$.

In particular, the optimal acceptance rate is still equal to 0.234.

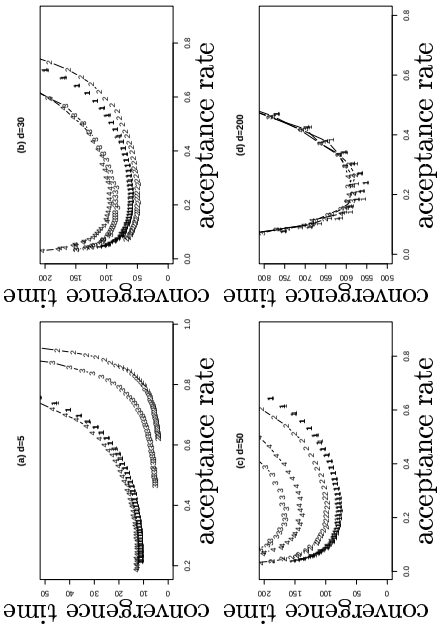
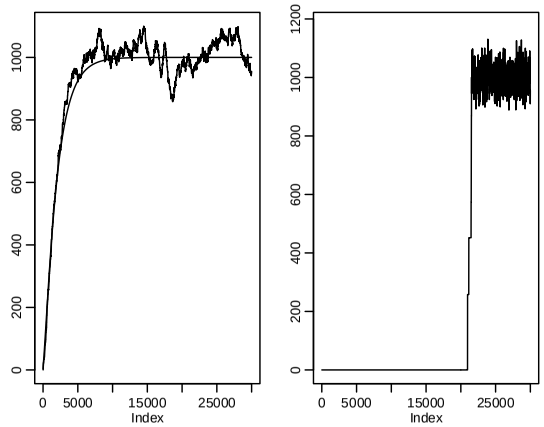


Figure 8: The convergence time of RWM in a hereogeneous environment, in dimensions 5, 30, 50 and 200. Here the plotting number indicates a particular random collection of C_i's.

Effect of Starting Value

Above results are only valid if $X_0 \sim \pi(\cdot)$, i.e. once the chain is stationary. What about in the transient phase?

Here are simulations of $\|X_n\|^2$ for RWM (left) and Langevin (right), with $\pi(\cdot) = N(0, I_d)$, with "optimally" scaled proposals, with $d = 1000$, when started far from stationarity:



What happened??

RWM moves in deterministically, towards the "center" of $\pi(\cdot)$:

THM (Christensen + R + R): For RWM with scaling $\sigma^2 = \ell^2/d$, let $W_t^d = (1/d)\|\mathbf{X}_{\lfloor td \rfloor}\|^2$, with $W_0^d = w_0 \neq 1$. Then W^d converges weakly to the function f satisfying $f(0) = w_0$ and

$$f'(t) = \left[\ell^2 + \exp((\ell^2/2)(f(t)-1)) \times \right. \\ \left. \times(1-2f(t))\ell^2\right] \Phi(-\ell f(t)^{1/2}/2)\,.$$

Meanwhile, Langevin gets really stuck:

THM: For Langevin, acceptance probability for first move is $O\big(\exp(-C\ d^{1/3})\big)$; very small. [Reason: $q(\mathbf{y}, \mathbf{x})$ is usually small.]

On the other hand, with scaling $\sigma^2 = O(d^{-1/2})$ instead of $\sigma^2 = O(d^{-1/3})$, Langevin also moves to center deterministically. Good! [So, perhaps best to <u>alternate</u> $O(d^{-1/2})$ and $O(d^{-1/3})$ moves.]

8 SOME CONVERGENCE RESULTS FOR THE GIBBS SAMPLER

- Irreducibility, aperiodicity
- Informal discussion of efficiency
- The Gaussian case
- Gibbs sampler variants
  - random scan
  - random permutation
  - reversible Gibbs sampler
- Blocking
- Positive association class
- Parameterisation issues for linear models
- Non-Gaussian case?

Much of this is in Roberts and Sahu (JRSSB, 1997, pp. 291–317). See also Roberts and Sahu, 'Rate of convergence of Gibbs sampler by Gaussian approximation' (see MCMC preprint server).

<u>Irreducibility, aperiodicity</u>

The Gibbs sampler is not necessarily π-irreducible. However, given suitable conditions on π, usually easily checked, π-irreducibility can be assured.

Theorem 8.1 *If*

1. $\mathcal{X}^+ \doteq \{x \in \mathcal{X}; \pi(x) > 0\}$ *is connected*
2. π *is continuous*
3. $\int \pi(x) dx^{(i)}$ *is a locally bounded function.*

Then the Gibbs sampler is π-irreducible.

Proved in Roberts and Smith (1994).

N.B. Extensions possible

The Gibbs sampler is also (essentially) aperiodic and (as long as it's carefully defined) Harris recurrent.

How much can we say about its speed of convergence?

$$\|P^n(x,\cdot) - \pi(\cdot)\| \downarrow 0 \quad \text{as } n \to \infty$$

But how quickly?

For a two-dimensional Gibbs sampler (Yali Amit) there exists a constant ρ and a function $A(x)$ such that

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq A(x)\rho^n$$

and ρ can be characterised as

$$\rho^{1/2} = \sup_{\text{functions } f,g} \text{Corr}\big(f(X^{(1)}), g(X^{(2)})\big)$$

where here $(X^{(1)}, X^{(2)})$ is distributed as π.

But in itself this is of little use. Look at important special case...

The Gaussian case

$$\pi \sim N(0, Q^{-1}),$$

$$Q = \begin{pmatrix} Q_{11} & & Q_{1k} \\ \vdots & & \vdots \\ Q_{k1} & \dots & Q_{kk} \end{pmatrix}$$

Given $\Phi_0 = (X_0^{(1)} \dots X_0^{(k)})$

$$X_1^{(1)} \sim N\left(\frac{-\sum_{j\neq 1} Q_{1j} X_0^{(j)}}{Q_{11}}, \frac{1}{Q_{11}}\right)$$

Linear and Gaussian.

So by iterating

$$\Phi_1 = (X_1^{(1)} \dots X_1^{(k)})$$

is linear and Gaussian in Φ_0.

More precisely...

Theorem 8.2 *Let $\pi \sim N(0, Q^{-1})$, with*

$$Q = \begin{pmatrix} Q_{11} & Q_{12} & \dots & Q_{1k} \\ \vdots & & & \vdots \\ Q_{k1} & \dots & \dots & Q_{kk} \end{pmatrix}$$

The Gibbs sampler on π produces a multivariate AR(1) process with $\Phi_{n+1} \sim N(B\Phi_n, \Sigma - B\Sigma B')$ where $\Sigma = Q^{-1}$. Let

$$A = I - \text{diag}(Q_{11}^{-1} \dots Q_{kk}^{-1})Q$$

and set $A = L + U$ where L is the lower block triangular matrix of A

$$L = \begin{pmatrix} 0 & \dots & \dots & 0 \\ A_{21} & \ddots & & \vdots \\ \ddots & \ddots & \ddots & \vdots \\ A_{11} & \dots & A_{kk-1} & 0 \end{pmatrix}$$

Then

$$B = (I - L)^{-1}U$$

In fact...

The rate of convergence of the Gibbs sampler on Gaussian densities can be written as

$$\rho = \rho(B)$$

i.e. the radius of convergence of B (maximum modulus eigenvalue).

The Random Scan Gibbs sampler

With probability $\frac{1}{k}$, replace ith component by a random draw from $\pi_i(\cdot \mid X_n^{(-i)})$. For fair comparison, we carry out k iterations of the random scan sampler in order to compare with the usual (deterministic scan) sampler.

Theorem 8.3 *The rate of convergence of the RSGS is given by*

$$\rho_{\text{RSGS}} = \left[k^{-1}\big(k - 1 + \lambda(A)\big)\right]^k$$

where $\lambda(A)$ is the maximum eigenvalue of A.

Note Unlike B, A is similar to a symmetric matrix, so has real (though not necessarily positive) eigenvalues.

Other forms of Gibbs sampler

Random permutation Gibbs sampler

Choose a permutation of $\{1 \dots k\}$, $\sigma(\cdot)$ say.

Update component $\sigma(1)$according to $\pi_{\sigma(1)}(\cdot \mid \cdot)$
Update component $\sigma(2)$according to $\pi_{\sigma(2)}(\cdot \mid \cdot)$
$\vdots$ $\vdots$
Update component $\sigma(k)$according to $\pi_{\sigma(k)}(\cdot \mid \cdot)$.

Reversible Gibbs sampler

(Recall the Gibbs sampler is **NOT** reversible.)

Updatecomponent 1
$\vdots$ $\vdots$
Updatecomponent $k-1$
Updatecomponent k
Updatecomponent $k-1$
$\vdots$ $\vdots$
Updatecomponent 1

Similar results for the rate of convergence of RPGS and REGS exist.

Blocking schemes

Suppose we update components $1 \dots r_1$ together. This the blocked Gibbs sampler.

Update $X_n^{(1)} \dots X_n^{(r_1)}$ according to $\pi(X^{(1)} \dots X^{(r_1)} \mid X^{(r_1+1)} \dots X^{(k)})$

Update $X_n^{(r_1+1)}$ ..... $\pi_{r_1+1}(\cdot \mid \cdot)$

$\vdots$

Update $X_n^{(r)}$ ..... $\pi_k(\cdot \mid \cdot)$

Theorem 8.2 still holds in a slightly modified form:

$$\begin{aligned} \rho &= \rho(B) \\ B &= (I - L)^{-1} U \\ L &= \text{lower triangular matrix of } A \\ A &= I - \text{blockdiag}(Q^{-1})Q \end{aligned}$$

where

$$\text{blockdiag}(Q^{-1}) = \begin{pmatrix} \begin{pmatrix} Q_{11} & \dots & Q_{1r_1} \\ \vdots & & \vdots \\ Q_{r_1 1} & \dots & Q_{r_1 r_1} \end{pmatrix}^{-1} & & & \\ & Q_{r_1+1,r_1+1}^{-1} & & 0 \\ & & \ddots & \\ & 0 & & 0 \quad Q_{kk}^{-1} \end{pmatrix}$$

Positive association

Off diagonal elements of Q are non-positive

$$\Leftrightarrow$$

Partial correlations for π are non-negative, ie

$$\big(\text{Corr}(X^{(i)}, X^{(j)} \mid X^{(\ell)} \rho \neq i, j) \geq 0\big)$$

And either of these conditions implies that

$$\Rightarrow \Sigma \text{ is non-negative (elementwise)}$$

Many nice clean results can be proved for distributions within this class.

Theorem 8.4 *If π exhibits* **positive association** *(all partial correlations are non-negative) then blocking schemes* **improve** *the rate of convergence.*

Remark

This is NOT true in general!! i.e. there are examples where blocking **SLOWS DOWN** convergence.

Example

$$
\begin{aligned}
X^{(1)} &\sim N(0,1) \\
X^{(i)} &\sim N(\eta X^{(i-1)}, 1-\eta^2) \\
Q &=
\end{aligned}
$$

$$
\frac{1}{1-\eta^2}
\begin{pmatrix}
1 & -\eta & 0 & \dots & & & \\
-\eta & 1+\eta^2 & -\eta & 0 & & & \\
0 & -\eta & \ddots & & \ddots & & \\
\vdots & \ddots & & & & 0 & \vdots \\
\vdots & & \ddots & & & & 0 \\
 & & & & 1+\eta^2 & \ddots & -\eta \\
0 & \dots & \dots & \dots & 0 & -\eta & 1
\end{pmatrix}
$$

If $\eta \geq 0$, π exhibits positive association. So blocking schemes will only aid convergence.

What happens for $\eta < 0$? Set

$$
\begin{aligned}
Z^{(i)} &= X^{(i)} \quad i \text{ even} \\
&= -X^{(i)} \quad i \text{ odd}
\end{aligned}
$$

Then Z is AR(1) with parameter $-\eta$. So blocking aids convergence here also.

Example: Gaussian image analysis

$$
\pi(x) \propto \exp\Big\{-\beta \sum_{i\sim j} (x^{(i)} - x^{(j)})^2\Big\} \times \exp\Big\{-\gamma \sum_i (x^{(i)})^2\Big\}
$$

where $\sim$ denotes a neighbourhood relationship.

Gaussian with positive association. So

- Blocking schemes improve things
- If π is the prior and the likelihood is proportional to

$$
\prod_i \exp\left\{\frac{-1}{2\sigma^2}(y_i - x_i)^2\right\} \qquad (12)
$$

Then

$$
B_{\text{prior}} \geq B_{\text{post}}
$$

for **ANY** data set $\mathbf{y}$, and $\rho(B_{\text{prior}}) \geq \rho(B_{\text{post}})$.

Comparing GS updating schemes

Which is bigger, ρ_{DUGS} or ρ_{RSGS}?

Theorem 8.5 *For Gaussian π with positive association,*

$$
\rho_{\text{DUGS}} \leq \rho_{\text{RSGS}}
$$

Corollary 8.6 *For Gaussian π with $Q_{ij} = 0$ for $|i-j| \geq 1$*

$$
\rho_{\text{DUGS}} \leq \rho_{\text{RSGS}}
$$

Hierarchical models Gibbs sampling

$f(y \mid \theta_1)$	likelihood
$h_1(\theta_1 \mid \theta_2)$	1st stage prior
$h_2(\theta_2 \mid \theta_3)$	
$\vdots$	
$h_k(\theta_k \mid \theta_{k+1})$	θ_{k+1} known kth stage prior

Conditional independence graph has a linear structure fitting into the conditional independence structure needed for the application of Corollary 8.6.

Positive association equivalence class

$$Q = \begin{pmatrix} 1 & - & - \\ & 1 & - \\ & & 1 \end{pmatrix}$$ positive association

---

$$Q = \begin{pmatrix} 1 & - & + \\ & 1 & + \\ & & 1 \end{pmatrix}$$

In positive association equivalence class (PAEC)

$$X^{(1)} = Z^{(1)}$$
$$X^{(2)} = Z^{(2)}$$
$$X^{(3)} = -Z^{(3)}$$

$$Q = \begin{pmatrix} 1 & - & - \\ & 1 & + \\ & & 1 \end{pmatrix}$$ NOT in PAEC

There are 'two types' of 3-dimensional distribution:

1. Rugby ball PAEC
2. Flying saucer

In higher dimensions, there are more equivalence classes:

$$2^{(k-1)(k-2)/2} \quad \text{in } k\text{-dimensions.}$$

However, by conditional independence, many encountered distributions fall into the PAEC.

Parameterisation of linear models

Model:

$Y_i = \mu + \alpha_i + \varepsilon_i$ non-centered parameterisation

where

$$\varepsilon_i \quad \text{IID} \quad N(0, \sigma_e^2)$$
$$\alpha_i \quad \text{IID} \quad N(0, \sigma_\alpha^2)$$
$$\mu \text{ has 'flat prior'}$$

Model:

$$Y_i = \gamma_i + \varepsilon_i$$

where ε_i IID $N(0, \sigma_e^2)$ hierarchically centred parameterisation

$$\gamma_i = \mu + \alpha_i$$

$$\gamma_i \mid \mu \sim N(\mu, \sigma_\alpha^2)$$

μ has flat prior.

$$\rho(\text{non-centered}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_e^2}$$

$$\rho(\text{centered}) = \frac{\sigma_e^2}{\sigma_\alpha^2 + \sigma_e^2}$$

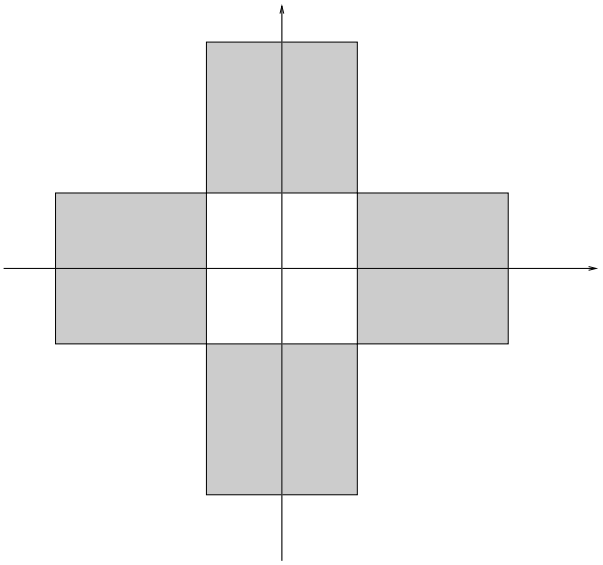
So the best choice of parameterisation depends on the relative sizes of $\sigma_e^2, \sigma_\alpha^e$.

In practice, this might well be performed as part of a larger Gibbs sampler in which $\sigma_e^2, \sigma_\alpha^2$ are also unknown. However, in this case we can always alternate between the two parameterisations depending on the relative sizes of $\sigma_e^2, \sigma_\alpha^2$.

However even better choices exist.... see for example Papaspiliopoulos + R (Valencia VII, OUP, 2003).

The non-Gaussian case?

An analysis of correlation structure is clearly **NOT** sufficient here!



Consider Gibbs sampling on the shaded region.

Use the sampling directions $y, x + \rho y$. Call the components $X^{(1,\rho)}$ and $X^{(2,\rho)}$. Then

$$\text{Cov}(X^{(1,\rho)}, X^{(2,\rho)}) = 0 \Leftrightarrow \rho = 0$$

However...

Gibbs sampler is reducible $\Leftrightarrow \rho = 0$!!

Bayesian posteriors

$$Y_1 \ldots Y_m \sim \text{IID} \quad f(\cdot \mid \tilde{\theta})$$

Prior on θ: $p(\theta)$

Let

$$\pi_m \propto \prod_{i=1}^{m} f(y_i \mid \theta) p(\theta)$$

For large m, we commonly have

$$\pi_m \overset{\text{some sense}}{\longrightarrow} N\left(\tilde{\theta}, \frac{I^{-1}(\tilde{\theta})}{m}\right)$$

where $I(\theta)$ is Fisher Information,

$$-E\left[\frac{\partial^2}{d\theta^2} \log f(X \mid \theta)\right]$$

What can we say about the Gibbs sampler on π_m?

Theorem 8.6 *(Roberts and Sahu, J. Comp. Graph. Stats., 2001) Under the 'usual' conditions that ensure asymptotic normality for the posterior distribution,*

Convergence time $(\pi_m) \longrightarrow$ *Convergence time* (π)

However...

Nothing rigorously known about how quickly this convergence is achieved.

9 ROUND-OFF ERROR FOR MCMC

MCMC motivation

Throughout this course, we have been assuming that the Markov chains under consideration move around in a continuous state space. On a computer, an approximation to the 'true' algorithm is carried out, perhaps by means of a 'round-off' error.

Whilst it seems reasonable to assume that such discrete approximations should have a negligible impact on individual iterations of the sampler, it's not so clear that convergence properties of the algorithms will be as well-approximated.

In this last part of the course, we'll examine robustness properties of algorithms under perturbations of the Markov chain transition kernel. We'll focus here on geometric ergodicity and on the stationary measure of the Markov chain.

General setup

We're given a Markov chain transition kernel $P(x, \cdot)$ on state space $\mathcal{X}$, with stationary distribution $\pi(\cdot)$.

Generate X_0 from some initial distribution.

Then for $k = 1, 2, 3, \ldots$, generate

$$X_k \sim P(X_{k-1}, \cdot)$$

on a *computer*.

As we've seen, under mild assumptions (e.g. ϕ-irreducibility and aperiodicity), as $k \to \infty$, distribution of X_k converges to $\pi(\cdot)$.

Hence, for "large enough" k, can take X_k as an approximate sample from $\pi(\cdot)$.

PROBLEM:

Computers don't run the Markov chain exactly!

e.g. pseudo-randomness, roundoff errors, finite range, approximate algorithms, etc.

Thus, instead have

$$X_k \sim \widetilde{P}(X_{k-1}, \cdot)$$

where $\widetilde{P}(x, \cdot)$ is similar to $P(x, \cdot)$, but not identical.

HOW is convergence of $\mathcal{L}(X_k)$ to $\pi(\cdot)$ affected??

For example, perhaps $X_k = h(Y_k)$, where

$$Y_k \sim P(X_{k-1}, \cdot)$$

and where $h : \mathcal{X} \to \mathcal{X}$ with $h(x)$ "near" x.

Then $\widetilde{P}(x, A) = P(x, h^{-1}(A)) \equiv P_h(x, A)$.

[e.g. $\mathcal{X} = \mathbf{R}$, $h(x) = \delta \lfloor \frac{1}{2} + \delta^{-1} x \rfloor$ where $\delta = 2^{-31}$; "roundoff function".]

GEOMETRIC ERGODICITY

Recall a Markov chain is *geometrically ergodic* if

$$\|P^n(x, \cdot) - \pi(\cdot)\| \;\leq\; M(x)\, \rho^n$$

for some $\rho < 1$, where

$$\|P^n(x, \cdot) - \pi(\cdot)\| \;\equiv\; \sup_{A \subseteq \mathcal{X}} |P^n(x, A) - \pi(A)| \,.$$

We'd like to be able to say that Markov chains actually implemented on the computer bear some resemblance to the 'true' Markov chain.

Is geometric ergodicity preserved under small perturbation??

NOT ALWAYS!

e.g. $X = [0, \infty)$, with (for $x > 2$)

$$P(x, \cdot) = \mathbf{Unif}\left[x - \frac{4}{x},\; x + \frac{1}{x}\right]$$

This chain is geometrically ergodic with the choice $V(x) = xe^{x^2} + 1$, since we compute that

$$PV(x) = \left(\frac{e^{2+x^{-2}} - e^{-8+16x^{-2}}}{10}\right)(V(x) - 1) + 1$$

$$< 0.95\, V(x)\,, \qquad x > 2\,.$$

However, with roundoff function $h(x) = x + \delta$,

$$P_h(x, \cdot) = \mathbf{Unif}\left[x - \frac{4}{x} + \delta,\; x + \frac{1}{x} + \delta\right]$$

which is *transient* (the chain drifts off to $+\infty$, with no convergence at all), for any $\delta > 0$.

Bad!

Problem for running MCMC on a computer!!

Are any positive results possible??

NOTE THAT in the example, we had

$$V(x) = xe^{x^2} + 1$$

which grows very quickly for large x.

In fact, $\log V(x) \sim x^2$, so that $\log V$ is not uniformly continuous on $\mathcal{X} = [0, \infty)$.

(Since $V \geq 1$, $\log V$ is uniformly continuous whenever V is uniformly continuous.)

On the positive side, we have:

Theorem 9.1 *(R,R + Schwartz, JAP 1998): Suppose P is geometrically ergodic, with $\log V$ (or V) uniformly continuous on $\mathcal{X}$. Then there is $\delta > 0$ such that the modified chain P_h is geometrically ergodic whenever* $\sup_{x \in \mathcal{X}} \text{dist}\big(h(x),\, x\big) < \delta$.

Idea of proof:

Show that C is small for P_h, too. Then show that P_h satisfies similar drift condition to P, for same V and C, and only slightly worse β and b.

Note that since $\log V$ uniformly continuous,

$$\frac{V\big(h(y)\big)}{V(y)} \;\approx\; 1\,, \qquad \text{uniformly over } y \in \mathcal{X}\,.$$

This is the computational key.

REMARK: Can similarly obtain "ϵ-versions" of quantitative convergence rate bounds, assuming $\log V$ is uniformly continuous on $\mathcal{X}$.

REMARK: Can give similar results for floating point type error functions which satisfy

$$\text{dist}(h(x), x) \le \delta |x|$$

See Breyer, R + R (Stats. Prob. Letters, 2001)

WHAT ABOUT $\pi(\cdot)$?

Even if $\widetilde{P}$ is also geometrically ergodic, it will have a different stationary distribution, say $\widetilde{\pi}$.

Will $\widetilde{\pi}$ be close to π?

We will certainly need a condition to say that $P(x, \cdot)$ is 'close' to $P(h(x), \cdot)$. Such a property is catured by the notion of Feller continuity. We say that a Markov chain P is weakly Feller continuous if for all continuous bounded functions, g,

$$Pg(x) = \mathbf{E}_x(g(X_1))$$

is a continuous function of x.

It turns out that all Metropolis chains with proposal kernel density $q(x, y)$ continuous in x are weak Feller, so this is a very natural condition for MCMC application.

In fact the weak Feller condition turns out to be ALL we need in addition to the conditions we required for the preservation of geometric ergodicity.

Theorem 9.2 *Let P be geometrically ergodic for some V with $\log V$ uniformly continuous on $\mathcal{X}$, and with stationary distribution $\pi(\cdot)$. Consider a sequence of modified chains $P_{h_k}(x, \cdot)$, where*

$$\lim_{k \to \infty} \sup_{x \in \mathcal{X}} \text{dist}\big(h_k(x),\, x\big) = 0\,.$$

Assume each P_{h_k} is ϕ-irreducible, and that P is "weak Feller continuous". Then each P_{h_k} is geometrically ergodic with stationary distribution $\pi_k(\cdot)$, and furthermore $\{\pi_k\}$ converges weakly to $\pi(\cdot)$.

If furthermore

$$\lim_{k \to \infty} \|P_{h_k}(x, \cdot) - P(x, \cdot)\| = 0\,,$$

for all $x \in \mathcal{X}$, then we also have

$$\lim_{k \to \infty} \|\pi_k(\cdot) - \pi(\cdot)\| = 0\,.$$

(Don't even need Feller continuity for this.)

We conclude that, for sufficiently small perturbations of geometrically ergodic chains, new stationary distributions are close to original, assuming that $\log V$ is uniformly continuous on $\mathcal{X}$.

Of course this can be really tough to test in practise...

SUMMARY:

- Computers used to run Markov chains $P(x, \cdot)$.
- Due to various computer limitations (e.g. roundoff error), they instead run modified chain $\widetilde{P}(x, \cdot)$.
- In general, arbitrarily small modifications can create arbitrarily large problems, e.g. turn geometrically ergodic Markov chains into transient Markov chains.
- If the drift function V has $\log V$ uniformly continuous on $\mathcal{X}$, then we have positive results that for sufficiently small perturbations:
  - Geometric ergodicity is preserved;
  - Modification of stationary distributions is arbitrarily small.
  - Related resuts for floating point round-off.
- These results suggest that running Markov chains on computers can be done safely, but extra care must be taken to protect from arbitrarily large effects of arbitrarily small perturbations.