MARKOV CHAIN MONTE CARLO: METHODS AND PROPERTIES

Gareth Roberts

Luminy

September 2001

1. INTRODUCTION
2. MARKOV CHAINS AND MCMC
3. THE GIBSB SAMPLER
4. SCALING OF METROPOLIS-HASTINGS ALGORITHMS

1 INTRODUCTION TO ALGORITHMS

- The problem, examples.
- The algorithms
- Some examples.

The problem

We're given a (possibly un-normalised) density function π_u. We want to

- estimate expectations with respect to π:

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

- simulate a sample of points from distribution π.

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 $|B|$.

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),

Types of MCMC

Currently, all important MCMC algorithms are of the Metropolis-Hastings type. Important algorithms include:

- Gibbs sampler
- Random walk Metropolis algorithm
- Multiplicative random walk Metropolis
- The independence sampler
- Langevin methods
- Hybrid Monte Carlo
- Auxiliary variable methods

Metropolis-Hastings algorithm

Given a target denstity $\pi(\cdot)$ that we wish to sample from, and a Markov chain transition kernel density $q(\cdot,\cdot)$, we construct a Markov chain as follows. Given X_n, generate Y_{n+1} from $q(X_n,\cdot)$. Now set $X_{n+1} = Y_{n+1}$ with probability

$$\alpha(X_n, Y_{n+1}) = 1 \wedge \frac{\pi(Y_{n+1})q(Y_{n+1}, X_n)}{\pi(X_n)q(X_n, Y_{n+1})} .$$

Otherwise set $X_{n+1} = X_n$.

- **Symmetric random walk Metropolis algorithm**

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

  The acceptance probability simplifies to

  $$\alpha(x,y) = 1 \wedge \frac{\pi(y)}{\pi(x)}$$

- **Independence sampler**

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

- **Gibbs sampler**

  Suppose π is k-dimensional and write $\mathbf{x} = (x_1, \ldots, x_k)$

  $$q_i(\mathbf{x}, \mathbf{y}) = \pi(y^{(i)}|\mathbf{y}^{(-i)})$$

  for $x^{(j)} = y^{(j)}, j \neq i$, and $= 0$ otherwise.

- **Langevin methods**

  Motivated by diffusion processes. Given a suitably differentiable π, the diffusion

  $$dX_t = dB_t + \frac{1}{2}\nabla \log \pi(X_t)\ dt$$

  has stationary distribution π. It can be thought of as the continuous time limit of discrete approximations of the form:

  $$X_{t+\epsilon} - X_t = N(0,\epsilon) + \frac{1}{2}\nabla \log \pi(X_t)\ \epsilon\ .$$

  What about the stationary properties of such an approximation for "small enough" ϵ?

Suppose $\pi(x) \propto \exp\{-x^4/4\}$, then Langevin diffusion is

$$dX_t = dB_t - X_t^3\ dt$$

and the discrete approximation is therefore

$$X_{t+\epsilon} = X_t + N(0,\epsilon) - \epsilon X_t^3$$

which is transient for all $\epsilon > 0$.

Instead, use MALA (Metropolis adjusted Langevin algorithm for example Besag 1994, R and Tweedie, 1996):

Take $Y_{n+\epsilon} \sim N(X_n + \epsilon\nabla \log \pi(X_n), \epsilon)$ as the proposed new move. Use accept/reject to make sure π is stationary.

More sophisticated Langevin algorithms do exist.

- **Auxiliary variable methods**

Auxiliary variable methods adds extra variables to the state space of the Markov chain. These extra variables are of no intrinsic interest in their own right, but allow the construction of more effective Markov chain dynamics.

A simple example is **slice sampling**.

An auxiliary variable method considering densities as mixtures of uniforms. The following is the simplest possible example. Let $(U|X) \sim U(0, \pi(X))$, so that the joint distribution of (U, X) is uniform on the region under the graph of $\pi(X)$. Consider Gibbs sampling on U, X, all conditionals are uniform and easily sampled as long as we can "approximately invert" π.

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.

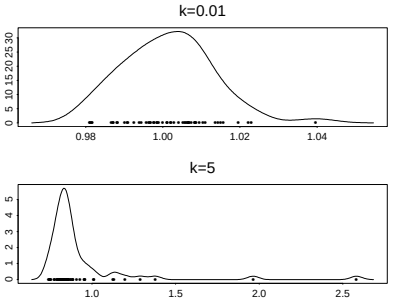

Figure 1: Sample means of each of the 110 runs. Monte carlo error in the case $k = 0.01$ is fairly small and symmetric about zero. However in the $k = 5$ case, the error is massively skewed, and most runs give very biased resuts.

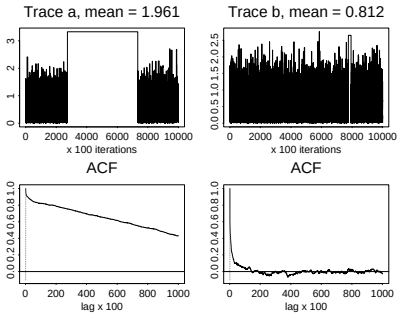

Figure 2: Two sample paths from the $k = 5$ simulation study.

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 tis moves accepted.

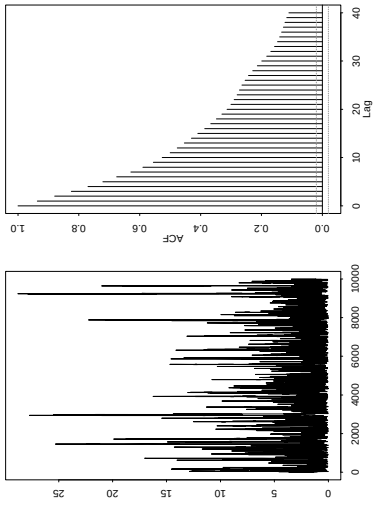

Figure 3: Does this look like reasonable output for a Cauchy distribution?

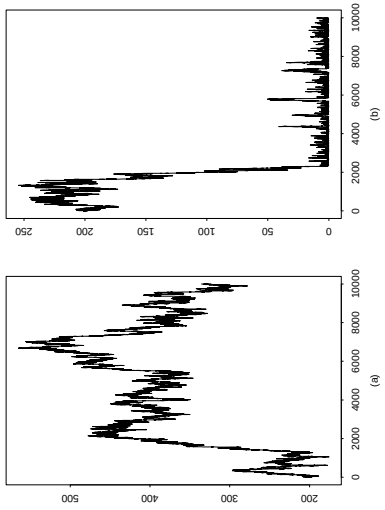

Figure 4: Random walk Metropolis on the Cauchy started out in the tail, showing the unstable heavy tailed nature of its excursions.

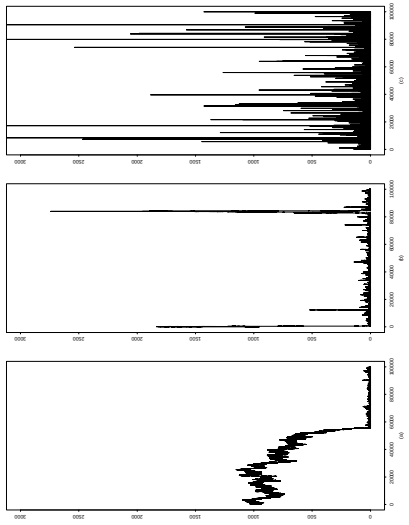

Figure 5: Three random walk Metropolis algorithms on the Cauchy example: (a) Gaussian proposal, (b) Cauchy proposal, (c) multiplicative random walk with Gaussian proposal.

2 MARKOV CHAINS AND MCMC

- Basic concepts
- General state space Markov chains
- Examples

Countable state space chains

Let $\mathcal{X} = \{0, 1, 2, \ldots\}$. X_n is a Markov chain on $\mathcal{X}$ if it satisfies the Markov property:

$$P[X_n = j | X_{n-1} = i, X_{n-2}, .., X_0] = P[X_n = j | X_{n-1} = i]$$
$$= P(i, j) \ .$$

We also write $P^n(i, j)$ for the n-step transition law of the Markov chain.

- Irreducibility:

$$\forall i, j, \ \ \exists n(i, j) : \ \ P^{n(i,j)}(i, j) > 0$$

- Periodicity:

$$\exists D_1, \ldots, D_d : \ P(x, D_i) = 1, \ x \in D_{i-1 (\mathrm{mod}\ d)}$$

this means that the Markov chain sample path exhibits cycles of period d.

Invariant (stationary) measures

We call π **invariant** if

$$\pi(x) = \sum_{y \in \mathcal{X}} \pi(y) P(y, x) \quad \forall x \in \mathcal{X}$$

Theorem 2.1 *If P is irreducible,*

- *there exists at most one probability measure π which is invariant;*
- $$\lim_{n \to \infty} \frac{\sum_{i=1}^n h(X_i)}{n} = E_\pi[h(X)] \quad a.s.$$
- *if P is also aperiodic, and π is an invariant probability measure, then*

$$P^n(x, y) \to \pi(y) \quad \forall x, y \in \mathcal{X} \ .$$

Key idea of MCMC: construct P so that π is invariant for P.

Reversibility

We say X (or P) is reversible with respect to π if it satisfies the **detailed balance** relation:

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

If X (or P) is reversible with respect to π, the π is invariant for P

Reversibility is much easier to handle than invariance, so we try to construct MCMC algorithms to satisfy detailed balance.

Metropolis-Hastings

Proposition 2.2 *Metropolis-Hastings chains are reversible with respect to π.*

Proof

$$\pi(x) P(x, y) = \pi(x) q(x, y) \left[1 \wedge \frac{\pi(y) q(y, x)}{\pi(x) q(x, y)} \right]$$
$$= \pi(y) q(y, x) \wedge \pi(x) q(x, y)$$
$$= \pi(y) P(y, x)$$

So all we need to check is irreducibility and aperiodicity.

This proof can more or less be carried over to the general state space case where (perhaps) P becomes a transition density. But the general case does need a little extra care...

The Gibbs sampler

The Gibbs sampler is **NOT** reversible with respect to π. However each of its k constituent component moves are.

Proposition 2.3 *π is an invariant probability density for the Gibbs sampler on π.*

Proof

$$\pi(x) q_i(x, y) = \pi(x) \pi_i(y_i | y_{-i}) \mathbf{1}(x_{-i} = y_{-i})$$
$$= \frac{\pi(x) \pi(y)}{\pi_{-i}(y_{-i})} \mathbf{1}(x_{-i} = y_{-i})$$
$$= \frac{\pi(x) \pi(y)}{\pi_{-i}(x_{-i})} \mathbf{1}(x_{-i} = y_{-i})$$
$$= \pi(y) q_i(y, x) \ .$$

So each of the k components is reversible for π and so have invariant distribution π. Therefore the Gibbs sampler has invariant distribution π.

General chains

For a comprehensive treatment of general state space Markov chains, see *Markov chains and Stochastic stability* by Sean Meyn and Richard Tweedie (Springer, 1995).

Let the general state space $\mathcal{X}$ be endowed with σ-algebra $\mathcal{B}$. We write $P(x, A)$ for the one step probability of entering A started at x, for any $x \in \mathcal{X}$, $A \in \mathcal{B}$. The n-step transition probabilities satisfy the **Chapman-Kolmogorov** equations:

$$P^n(x, A) = \int_{\mathcal{X}} P(x, dy) P^{n-1}(y, A)$$
$$= \int_{\mathcal{X}} P^{n-1}(x, dy) P(y, A) \ .$$

Fundamental concepts for general chains

A chain is **ϕ-irreducible** if there exists a non-trivial measure ϕ on $\mathcal{B}$ such that for all $A \in \mathcal{B}$ with $\pi(A) > 0$, and for all $x \in \mathcal{X}$, there exists a positive integer $n(x)$ such that

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

For ϕ-irreducible chains, we can do (virtually) everything that's possible in discrete state spaces in a general measure space.

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 2.4 *(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 2.5 *(Shallow!) If X is ϕ-irreducible and π is invariant, then X is π-irreducible.*

Theorem 2.6 *If X is ϕ-irreducible, there exists a* **finite** *cycle $D_1, \ldots D_d$ with*

- $P(x, D_i) = 1, \ x \in D_{i-1 mod\ d}$;
- $\phi(\mathcal{X} - \cup_i D_i) = 0$.

We say X is **aperiodic** if $d = 1$.

- 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** $\int P(x,dy)f(y)$ is continuous in x $\forall x$, $\forall$ continuous bounded functions f (the **weak Feller** property).

Stability of Markov chains

Stability can be formulated in terms of

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

We say X is **Harris recurrent** if

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

form SOME small set C.

Theorem 2.7 *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 2.8 *(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*

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

- *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*

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

Define the **total variation distance** between two probability measures:

$$\|\nu_1 - \nu_2\| = 2 \sup_A |\nu_1(A) - \nu_2(A)| \ .$$

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

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

- $\forall x \in V_0$,

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

- If X is Harris recurrent then (??) holds for all x.
- If $V < \infty$ everywhere, then (??) 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 2.9 *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 minimal requirement for any sensible algorithm with transition probabilities P is that of ergodicity: that is for all $x \in \mathcal{X}$,

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

In fact it is easy to demonstrate that all three of the examples considered in the Cauchy example satisfy this requirement. We need to consider more refined conditions on $r(\cdot,\cdot)$ in order to compare these methods.

Rates of convergence of Markov chains

(a) *A Markov chain is uniformly ergodic if $r(\mathbf{x},n) \downarrow 0$ uniformly in $\mathbf{x}$ as $n \to \infty$.*

(b) *A Markov chain is geometrically ergodic if there exists a real valued function $V(\mathbf{x}), \mathbf{x} \in \mathcal{X}$ and a positive constant $\rho < 1$ such that $r(\mathbf{x},n) \leq V(\mathbf{x})\rho^n$ for all $\mathbf{x} \in \mathcal{X}$, and all $b \in \mathbb{Z}_+$. The infimum of all values ρ for which we can do this is called the (geometric) rate of convergence of the Markov chain.*

(c) *A Markov chain is polynomially ergodic at rate p for some $p > 0$ if there exists a real valued function $h(\mathbf{x}) \geq 1, \mathbf{x} \in \mathcal{X}$ such that $r(\mathbf{x},n) \leq h(\mathbf{x})n^{-r}$ for all $\mathbf{x} \in \mathcal{X}$, $r < p$ and $n \in \mathbb{Z}_+$.*

Drift and minorisation conditions

To describe excursions from the set C, use Foster-Lyapunov *drift* conditions. Suppose that $V : \mathcal{X} \to [1,\infty)$ such that P satisfies

$$\mathbf{E}[V(\mathbf{X}_1)|\mathbf{X}_0 = \mathbf{x}] \leq V(\mathbf{x}) - (1-\lambda)V(\mathbf{x})^\alpha + b\mathbf{1}_C(\mathbf{x}) \tag{3}$$

for constants $b > 0$, $0 < \lambda < 1$, and $0 \leq \alpha \leq 1$. α is the most important parameter here, measuring the extent of the "drift" back towards the small set C.

Let τ denote the first time the chain enters the set C.

Theorem 2.10 *We suppose throughout that C is a small set and the Markov chain is aperiodic and satisfies the drift condition.*

(a) For all $\mathbf{x} \in \mathcal{X}$, $\mathbf{P}(\tau < \infty | \mathbf{X}_0 = \mathbf{x}) = 1$ and $\mathbf{E}(\tau | \mathbf{X}_0 = \mathbf{x}) < \infty$. Furthermore, the Markov chain in positive recurrent and irreducible.

(b) Suppose that $\alpha = 1$, then there is a positive number s_0 such that $\mathbf{E}(e^{s\tau} | \mathbf{X}_0 = \mathbf{x}) < \infty$ for all $s < s_0, \mathbf{x} \in \mathcal{X}$ and the Markov chain is geometrically ergodic.

(c) If $\alpha = 1$ and V is bounded, then the chain is uniformly ergodic.

(d) Suppose that $0 < \alpha < 1$, then $\mathbf{E}(\tau^p | \mathbf{X}_0 = \mathbf{x}) < \infty$ for all $\mathbf{x} \in \mathcal{X}$ and for all $p \leq 1/(1-\alpha)$. Moreover in this case, the chain is polynomially ergodic with rate $\alpha/(1-\alpha)$.

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

$$n^{1/2} \left(\frac{\sum_{i=1}^n f(\mathbf{X}_i)}{n} - \mathsf{E}_\pi(f(X)) \right) \Rightarrow \tag{4}$$

$$N(0, \tau_f \mathrm{Var}_\pi(f(X))) \tag{5}$$

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

Theorem 2.11 *CLT for Markov chains.*

(a) 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.

(b) Suppose that the drift condition holds for some $\alpha < 1$, and there exists a positive constant η with $1 - \alpha \leq \eta \leq 1$ and such that $\pi(V^{2\eta}) < \infty$. Then a $\sqrt{n}$-CLT exists for all functions f with $|f| \leq V^{\alpha+\eta-1}$.

For the polynomially ergodic case, CLTs can be shown to exist only for functions which do not grow too rapidly.

Eg, CLTs hold for all bounded functions if $\pi(V^{2(1-\alpha)}) < \infty$, or $\alpha \geq 1/2$.

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.

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 $r/2$. However, 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:

$$P(\mathbf{x}_i, \{\mathbf{x}_i\}) \to 1 \ , \tag{6}$$

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.

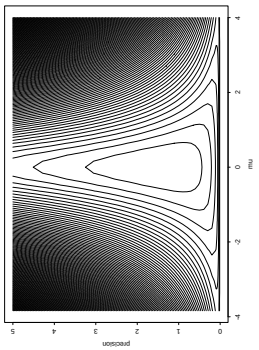

Figure 6: A contour polot of the contours of the posterior distribution in the normal-Gamma example.

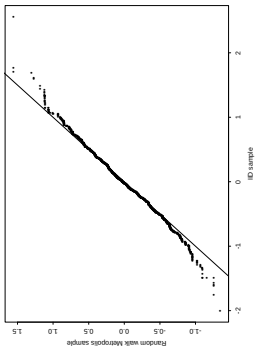

Figure 7: A q-q plot illustrating that the random walk Metropolis algorithm fails to adequately explore the tails of the normal-Gamma distribution.

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 extentds to other algorithms such as Langevin, hybrid Metropolis methods, Gibbs, Independence sampler, etc.

Theorem 2.12 *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

$$\lim_{K \to \infty} \frac{\log \pi(X_i \in (K, \infty))}{K} = 0 \ , \qquad (7)$$

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.

3 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).

Gareth Roberts / MCMC: Introduction to algorithms and theory / Rebild 1998

Irreducibility, aperiodicity

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

Theorem 3.1 *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

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

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 3.2 *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

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

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 3.3 *The rate of convergence of the RSGS is given by*

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

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

Reversible Gibbs sampler

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

Updatecomponent 1
$\quad\vdots \qquad \vdots$
Updatecomponent $k-1$
Updatecomponent k
Updatecomponent $k-1$
$\quad\vdots \qquad \vdots$
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 4.2 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

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

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 3.4 *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

$$\prod_i \exp\left\{\frac{-1}{2\sigma^2}(y_i - x_i)^2\right\} \tag{8}$$

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 3.5 *For Gaussian π with positive association,*

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

Corollary 4.6 *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 4.6.

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)

$$\begin{aligned} X^{(1)} &= Z^{(1)} \\ X^{(2)} &= Z^{(2)} \\ X^{(3)} &= -Z^{(3)} \end{aligned}$$

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

$2^{(k-1)(k-2)/2}$ in k-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

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

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.

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

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....

The non-Gaussian case?

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

[figure]

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 $\Leftrightarrow \rho = 0$!!

Theorem 3.6 *(Roberts and Sahu, 1998) 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.

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?

4 SCALING OF METROPOLIS-HASTINGS ALGORITHMS

Some of this is joint work with Andrew Gelman and Wally Gilks, and with Jeff Rosenthal.

- Metropolis on continuous spaces
- metropolis on discrete spaces
- Langevin

Types of MCMC

Currently, all important MCMC algorithms are of the Metropolis-Hastings type. Important algorithms can be loosely classified as follows:

	Local	Global
Vanilla	Random walk Metropolis	Independence sampler
Problem specific	Langevin algorithms	Gibbs sampler/ IID
Auxiliary	Hybrid Monte Carlo	Slice sampling

Scaling problems arise for local algorithms.

Hastings-Metropolis algorithm

Given a target denstity $\pi(\cdot)$ that we wish to sample from, and a Markov chain transition kernel density $q(\cdot,\cdot)$, we construct a Markov chain as follows. Given X_n, generate Y_{n+1} from $q(X_n,\cdot)$. Now set $X_{n+1} = Y_{n+1}$ with probability

$$\alpha(X_n, Y_{n+1}) = 1 \wedge \frac{\pi(Y_{n+1})q(Y_{n+1}, X_n)}{\pi(X_n)q(X_n, Y_{n+1})} .$$

Otherwise set $X_{n+1} = X_n$.

- **Symmetric random walk Metropolis algorithm**

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

The acceptance probability simplifies to

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)}$$

Algorithms is geometrically ergodic "most of the time" when the tails of the target density are no heavier than exponential.

Problem 1

Given a (Lebesgue) density π, and a symmetric proposal

$$q_\sigma = \frac{1}{q} q(\cdot/\sigma)$$

for the RWM algorithm, how should we choose σ to optimise the efficiency of the algorithm?

More specifically...

Let

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

and suppose that

$$q \sim N(0, I_d \sigma_d^2)$$

.

- How should σ_d depend on d for large d?
- Can we optimise σ_d in some sensible way?
- Can we charactise optimal (or close to optimal) values of σ_d in terms of observable properties of the Markov chain?
- How are the answers affected by choice of q? f?
- What happens if π is not product form?

Problem 2

A discrete problem. Let π be a density on the vertices of an d-dimensional hypercube with

$$\pi(i_1, i_2, \ldots, i_d) = p^{|\{j;\ i_j=0\}|}(1-p)^{|\{j;\ i_j=1\}|} .$$

Assume wlog that $p < 1/2$.

Consider the following *random scan* version of the RWM algorithm. Suppose that at each iteration, the algorithm picks $r(d)$ sites at random, $S = \{i_{j_1}, i_{j_2}, \ldots, i_{j_{r(d)}}\}$ say. The proposed move is therefore to change the value at each state in S and to leave all states in $\{1, \ldots, d\} - S$ unaltered. The acceptance probability is given by the usual formula:

$$\alpha = 1 \wedge \frac{\pi(\text{proposed})}{\pi(\text{current})} .$$

- How should we choose $r(d)$?
- How does the optimal scaling, and the choice of $r(d)$ vary as a function of d?
- Can we understand this in terms of observable properties of the Markov chain?
- Is this discrete case fundamentally different from the continuous one in Problem 1?

Langevin methods

MALA, Hastings adjusted Langevin algorithm,.... (for example Besag, 1994, R and Tweedie, 1996)

Motivated by diffusion processes. Given a suitably differentiable π, the diffusion

$$dX_t = dB_t + \frac{1}{2}\nabla \log \pi(X_t)\ dt$$

has stationary distribution π. Propose:

$$Y_{n+1} - X_n \sim N(0, \epsilon) + \frac{1}{2}\nabla \log \pi(X_n)\ \epsilon\ .$$

Use accept/reject to make sure π is stationary.

Problem 3

Suppose $\pi(\mathbf{x}) = \prod_{i=1}^d f(x_i)$. Let the Langevin proposal be $N(\mathbf{X}_n + \frac{\sigma_d^2}{2}\nabla \log \pi(\mathbf{X}_n), \sigma_d^2 I_d)$.

- How should σ_d depend on d for large d?
- Can we optimise σ_d in some sensible way?
- Can we charactise optimal (or close to optimal) values of σ_d in terms of observable properties of the Markov chain?
- What happens if π is not product form?
- What about scaling for the more sophisticated algorithms?

What is "efficiency" in small dimensional problems?

Let X be a Markov chain. Then for a π-integrable function f, efficiency can be described by

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

In general relative efficiency between two possible Markov chains varies depending on what function of interest f is being considered. As $d \to \infty$ the dependence on f disappears, at least in cases where we have a diffusion limit as we will....

Return to problem 1

Recall $\pi \sim \prod_{i=1}^d f(x_i)$, $q(\mathbf{x}, \cdot) \sim N(\mathbf{x}, \sigma_d^2 I_d)$.

Set $\sigma_d^2 = \ell^2/d$. Consider

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

Z^d is **not** a Markov chain, however in the limit as d goes to ∞, it is Markov:

$$Z_d \Rightarrow Z$$

where Z satisfies the SDE,

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

where

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

and $I = E_f[((\log f(X))')^2]$.

Write the above decomposition of h as

$$h(\ell) = \ell^2 \times A(\ell)\ ,$$

where $A(\ell)$ is the limiting overall acceptance rate of the algorithm, ie the proportion of proposed Metropolis moves ultimately accepted. So

$$h(\ell) = \frac{4}{I}\left(\Phi^{-1}(A(\ell))\right)^2 A(\ell)\ ,$$

and so the maximisation problem can be written entirely in terms of the algorithm's acceptance rate.

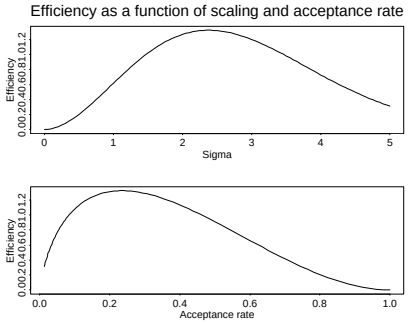

Figure 8:

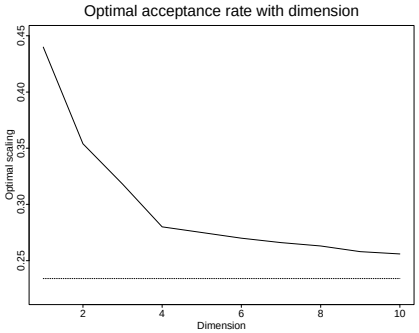

Figure 9:

Extensions

Problem is to find sequences of target densities where we can consider sensible limits for high dimensional problems. In ALL these extensions, expressions for the optimal scaling are complex, but correspond to the acceptance rate of 0.234.

- $\pi = \prod_{i=1}^d f_i(x_i)$, where the functions f_i "converge" to a limiting function f_∞ say.
- $\pi = f(x_1)\prod_{i=2}^d p(x_{i-1}, x_i)$ where p is the transition kernel of a positive recurrent Markov chain.
- π is a finite range homogeneous Markov random field (see Breyer and Roberts, 1998).
- π_d is multivariate normal with pretty much any covariance structure for which a meaningful limit algorithm can be defined.

Targets with heterogenous scaling

Suppose

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

and $q(\mathbf{y}) \sim N(0, I_d\sigma_d^2)$, with $\sigma_d^2 = \ell^2/d$ for some $\ell > 0$.

Theorem 4.1 *Consider a random-walk Metropolis algorithm, with target density of this form, where $\{C_i\}$ are i.i.d. with $E(C_i^2)/E(C_i)^2 \equiv b < \infty$. Let $W_t^d = C_1 X_{[td]}^{(1)}$. Then as $d \to \infty$, W_t^d converges 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 $I = E_f[(g'(X))^2]$.

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 $\frac{C_1^2}{b}$. In particular, the optimal acceptance rate is still equal to 0.234. For a fixed function f, the optimal asymptotic efficiency is proportional to $\frac{C_1^2}{bd}$.

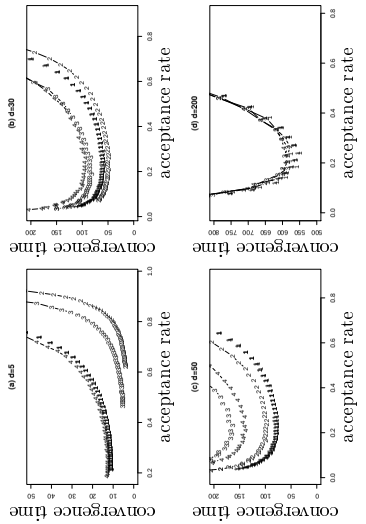

Figure 10: The convergence time of RWM in a hereogeneous environment, in dimensions 5, 30, 50 and 200. Here the plotting number indicates a particular random collection of C_i's.

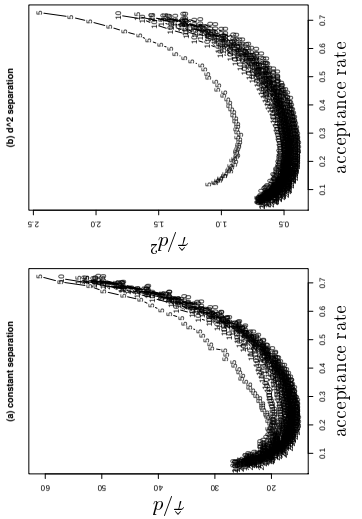

Figure 11: (a) A bimodal (mixture of normals) example, with modes separated by 3 sds in the first component (b)$\hat{\tau}/d^2$ as a function of acceptance rate in the case of $O(d^{1/2})$ separation (where diffusion limits fail).

Example: exchangeable MVN

$\pi \sim MVN(0, \Sigma)$ where $\Sigma_{ii} = 1$, $\Sigma_{ij} = \rho > 0$, $i \neq j$. The eigenvalues of Σ are just $1-\rho$, $d-1$ times and $d\rho + 1 - \rho$.

By an orthogonal transformation, we can write π as a collection of d independent normals with variances equal to the eigenvalues of Σ. Therefore can put in above framework with $C_1 = (d\rho + 1 - \rho)^{-1/2}$ and $C_i = (1-\rho)^{-1/2}$ otherwise. The principle eigenfunction here is $\bar{x}$ so we'd expect this to coverge in time $O(db/C_1^2)$ $= O(d^2)$. Orthogonal functions (like $\bar{x}$ converge like d^2).

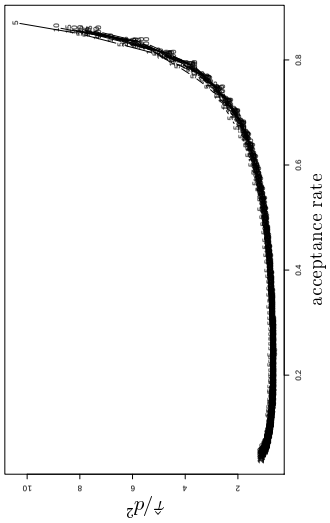

Figure 12: The convergence of $\bar{x}$ for the exchangeable normal example, demonstrating that it is in fact $O(d^2)$.

Return to problem 2

Recall $\pi(\mathbf{x}) = p^{\#0's}(1-p)^{\#1's}$.

Fix r and let d go to infinity along coprime (to r) values. Let $S_t^d = X_1^{[dt]}$. So S^d is a continuous time binary process (not Markov) making jumps at times which are integer multiples of $1/d$.

Theorem 4.2 *Assume X^0 is distributed according to π. Then as $d \to \infty$*

$$S^d \Rightarrow S$$

where S is a two state continuous time Markov chain with stationary distribution $(p, 1-p)$. In fact the Q−matrix for S has the following form:

$$Q = e(r) \times \begin{pmatrix} -(1-p) & 1-p \\ p & -p \end{pmatrix}\,,$$

where with $B \sim$Binomial$(r-1, p))$,

$$e(r) = \frac{r}{1-p} \times \mathbf{E}\left[1 \wedge \left(\frac{1-p}{p}\right)^{2B+2-r}\right]\,.$$

$$= r \times \left[\sum_{i \geq r/2}^{r-1} \binom{r-1}{i} p^{i-1}(1-p)^{r-2-i} + \mathbf{1}_{r\ odd} \binom{r-1}{(r-1)/2} p^{(r-1)/2}(1-p)^{(r-3)/2}\right]\,.$$

The overall acceptance rate of the algorithm is given by

$$a(p, r) = 2p\mathbf{E}\left[1 \wedge \left(\frac{1-p}{p}\right)^{2B+2-r}\right]\,.$$

Set $\lambda = (1/2 - p)^2 r$ and let $p \uparrow 1/2$ in such a way that λ remains constant,

$$2B + 2 - r \Rightarrow N(-8\lambda, 16\lambda)$$

and

$$a(p, \lambda/(1/2-p)^2) \to 2\Phi(-2\sqrt{\lambda})\,.$$

Thus for p close to 1/2, we can write

$$e(r) \approx 2r \times a(p, r)$$
$$\approx \frac{2}{(1/2 - p)^2}\lambda \times 2\Phi(-2\sqrt{\lambda}) \ .$$

The optimal choice of r is thus approximately that achieving acceptance rate 0.234 and the efficiency curve $e(\cdot)$ against $a(\cdot)$ converges to that of problem 1.

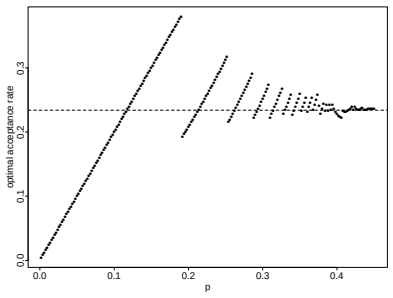

Figure 13: Optimal limiting acceptance rate as a function of p.

Return to problem 3

Recall: $\pi = \prod_{i=1}^{d} f(x_i)$, with proposal $N(\mathbf{X}_n + \frac{\sigma_d^2}{2}\nabla \log \pi(\mathbf{X}_n), \sigma_d^2 I_d)$.

Try analogous trick to that in Problem 1: set $\sigma_d^2 = \ell^2/d$. Consider

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

Again we get a limiting Langevin diffusion:

$$dZ_t = \ell dB_t + \frac{\ell^2 \nabla \log \pi(Z_t)}{2} dt \ .$$

The speed of this limiting algorithm as a function of ℓ is ℓ^2, unbounded. Therefore scaling of $\sigma_d^2 = O(1/d)$ is too small.

A scaling of $\sigma_d^2 = O(1/d^{1/3})$ was suggested in some contexts by Kennedy and Pendleton (1991). The following is due to R+Rosenthal (1995).

Let $\sigma_d^2 = \ell^2/d^{1/3}$ and set

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

Then as $d \to \infty$, Z^d converges weakly to

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

where $g(\ell) = 2\ell^2\Phi(J\ell^3)$, J is a complicated functional of f.

Again the acceptance probability of the algorithm is given by $2\Phi(J\ell^3)$ and the optimal scaling is characterised by that having acceptance rate 0.574.

Behaviour in high dimensions

Let T_d be the number of iterations needed to converge to stationarity for a problem in d dimensions.

- **Random Walk Metropolis** In the best case scenario, for large d need to take $\sigma_d^2 = O(d^{-1})$.

  $$T_d = O(d)$$

- **Langevin algorithms**

  For large d we need to take $\sigma_d^2 = O(d^{-1/3})$.

  $$T_d = O(d^{1/3})$$

  Similar complexity results are available for higher order Langevin algorithms, and possibly composite Langevin type algorithms such as Hybrid Monte Carlo.

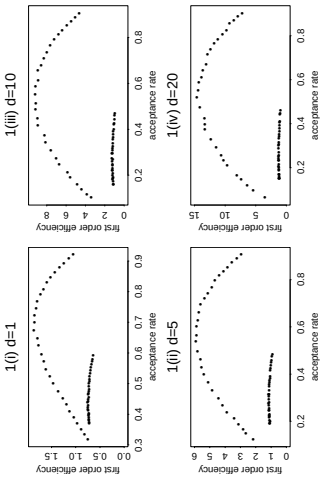

Figure 14: A comparison of Metropolis and Langevin algorithms in terms of efficiency.