Markov chain questions motivated by MCMC I: Complexity and Optimality

Gareth Roberts

University of Warwick

Durham, July/August 2017

including work mainly with many people ....

Contents

Metropolis-Hastings algorithm

Given a target density $\pi(\cdot)$ 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 $X_{n+1} = X_n$.

Two first scaling problems

- RWM

$$q(\mathbf{x}, \mathbf{y}) = q(|\mathbf{y} - \mathbf{x}|)$$

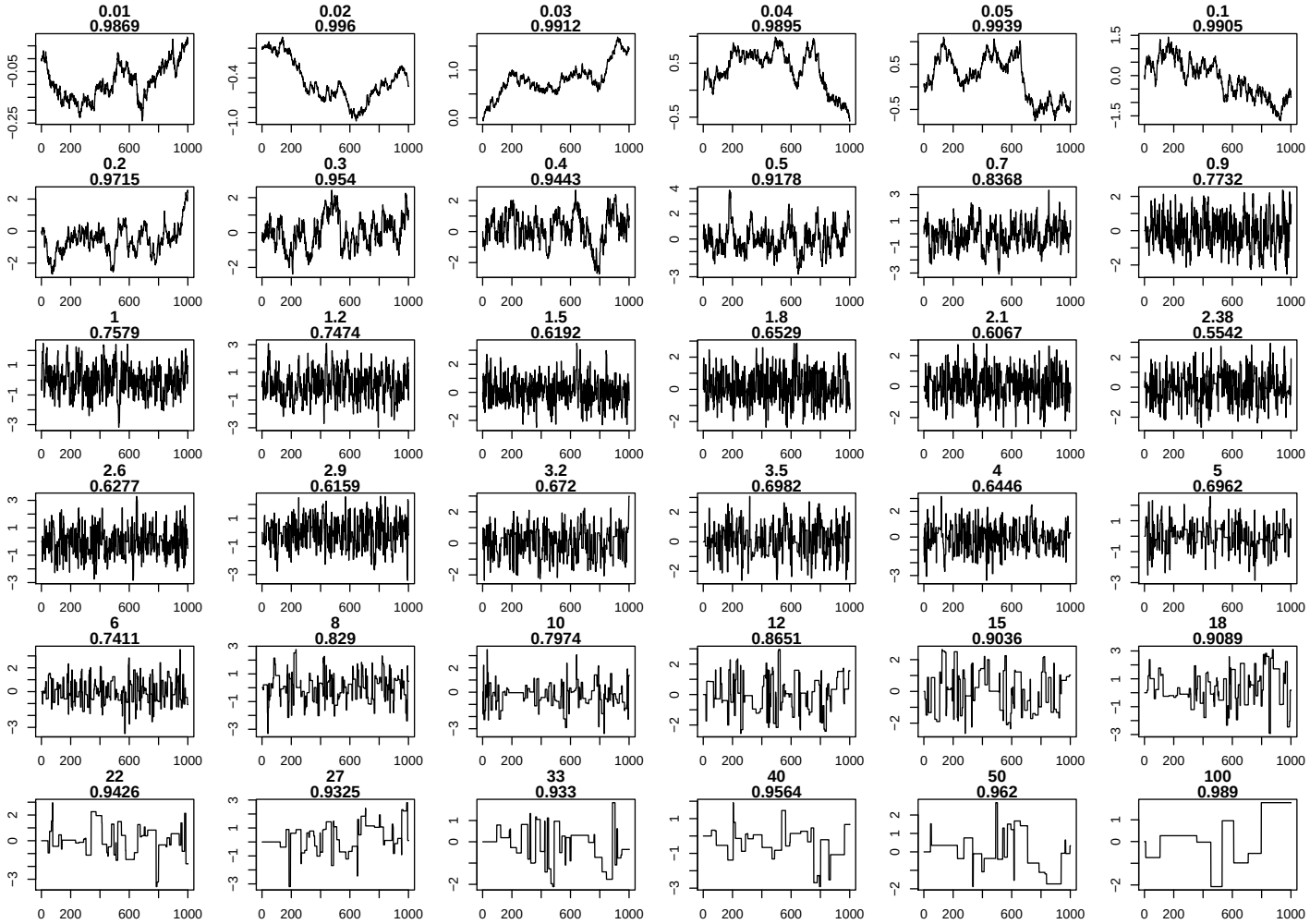
The acceptance probability simplifies to

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

For example $q \sim MVN_d(\mathbf{x}, \sigma^2 I_d)$, but also more generally.

- MALA

$$Y \sim MVN(x^{(k)} + \frac{hV\nabla \log \pi(x^{(k)})}{2}, hV) \ .$$



The Goldilocks dilemma

Scaling problems and diffusion limits

Choosing σ in the above algorithms to optimise efficiency. For 'appropriate choices' the d-dimensional algorithm has a limit which is a diffusion. The faster the diffusion the better!

- How should σ_d depend on d for large d?
- What does this tell us about the efficiency of the algorithm?
- Can we optimise σ_d in some sensible way?
- Can we characterise optimal (or close to optimal) values of σ_d in terms of observable properties of the Markov chain?

For RWM and MALA (and some other local algorithms such as Hamiltonian Monte Carlo) and for some simple classes of target distributions, a solution to the above can be obtained by considering a diffusion limit (for high dimensional problems).

Are lower dimensional updates better?

At each iteration, choose $d \times c_d$ components at random, and update these components according to a Metropolis algorithm which preseves the conditional distribution of those co-ordinates given the rest. The remaining $d(1 - c_d)$ components stay unchanged.

This is not really a generalisation of the Metropolis algorithm, but sometimes called Metropoplis-within-Gibbs.

How should be jointly choose (c_d, σ^2) to optimise the Markov chain?

Simulated tempering

Consider a d-dimensional target density $f_d(x)$, and suppose it is possible to construct MCMC on $f_{d,\beta} = f_d^\beta$, $0 \leq \chi \leq \beta \leq 1$.

Simulated tempering produces a Markov chain on $\mathcal{X} \times \mathcal{B}$ where $\mathcal{B}$ denotes a finite collection of temperatures.

But which temperatures should we choose?

This typically would mix better for small β. However we are interested in $f_{d,1}$.

Problem: Choose a finite collection of inverse temperatures, $B = \{\beta_i\}$ such that we can construct a Markov chain on $\mathbf{R}^d \times B$ which "optimally" permits the exploration of $f_{d,1}$.

This is also a scaling problem: chosing how large to make $\beta_i - \beta_{i-1}$ for each i.

What is "efficiency"?

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

$$\sigma^2(g, P) = \lim_{n\to\infty} n\mathrm{Var}\left(\frac{\sum_{i=1}^n g(X_i)}{n}\right) \ .$$

Under weak(ish) regularity conditions

$$\sigma^2(g, P) = \mathrm{Var}_\pi(g) + 2\sum_{i=1}^{\infty} \mathrm{Cov}_\pi(g(X_0), g(X_i))$$

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

How do we measure "efficiency" efficiently?

What is a good way to compare MCMC algorithms, and to identify whether they can be improved by changing proposal distribution scaling?

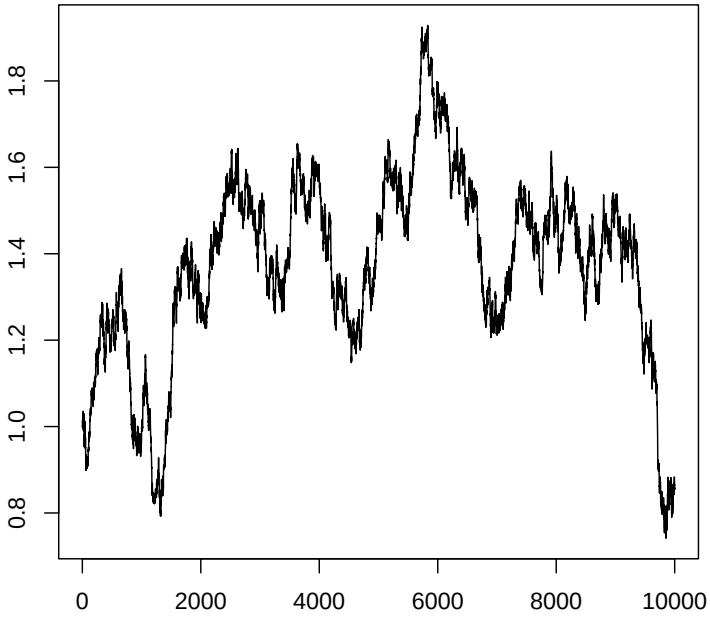
Common to resort to monitoring ESJD (as advocated strongly by Andrew Gelman)

$$ESJD = \mathbf{E}((X_{t+1} - X_t)^2)$$

What is the justification for this?

Optimising this is just like considering only linear functions g and ignoring all but the first term in

$$\sum_{i=1}^{\infty} \mathrm{Cov}_\pi(g(X_0), g(X_i))$$



MCMC sample paths and diffusions.

Here ESJM is the quadratic variation

$$\lim_{\epsilon \to 0} \sum_{i=1}^{[t\epsilon^{-1}]} (X_{i\epsilon} - X_{(i-1)\epsilon})^2$$

Reminicent of diffusion quadratic variation.

It will turn out that in fact many MCMC algorithms in high dimensions can be well-approximated by diffusion like the Langevin diffusion

$$dX_t^i = \sigma dB_t + \sigma^2 \nabla \log \pi(X_t^i)/2, \quad i = 1, 2,$$

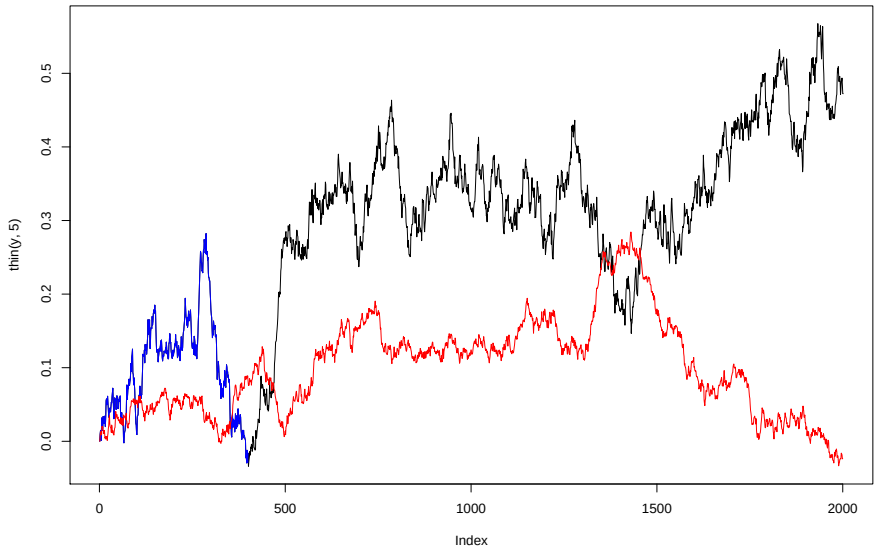
giving us a natural framework to consider optimisation (in this case for σ).

"Efficiency" for diffusions

Consider two Langevin diffusions, both with stationary distribution π.

$$dX_t^i = h_i^{1/2} dB_t + h_i \nabla \log \pi(X_t^i)/2, \quad i = 1, 2,$$

with $h_1 < h_2$.



X^2 is a "speeded-up" version of X^1.

A more powerful diffusion comparison result

R + Rosenthal 2012

Consider two Langevin diffusions, both with stationary distribution π.

$$dX_t^i = h_i(X_t^i)^{1/2} dB_t + V_i(X_t^i)dt, \quad i = 1, 2,$$

with $h_1(x) \leq h_2(x)$ for all x. (Here $V_i(x) = (h_i(x)\nabla \log \pi(x) + h_i'(x))/2$.)

Then X^2 dominates X^1 in Peskun order sense:

$$\lim_{t\to\infty} t\mathrm{Var}\left(\frac{\int_0^t g(X_s^1)ds}{t}\right) \geq \lim_{t\to\infty} t\mathrm{Var}\left(\frac{\int_0^t g(X_s^2)ds}{t}\right)$$

The first diffusion comparison result (R Gelman Gilks, 1997)

Consider the Metropolis case.

Theorem 1 *Suppose* $\pi \sim \prod_{i=1}^{d} f(x_i)$, $q(\mathbf{x}, \cdot) \sim N(\mathbf{x}, \sigma_d^2 I_d)$, $\mathbf{X}_0 \sim \pi$.

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

$$Z_t^d = X_{[td]}^{(1)} \ . \qquad \textit{Speed up time by factor } d$$

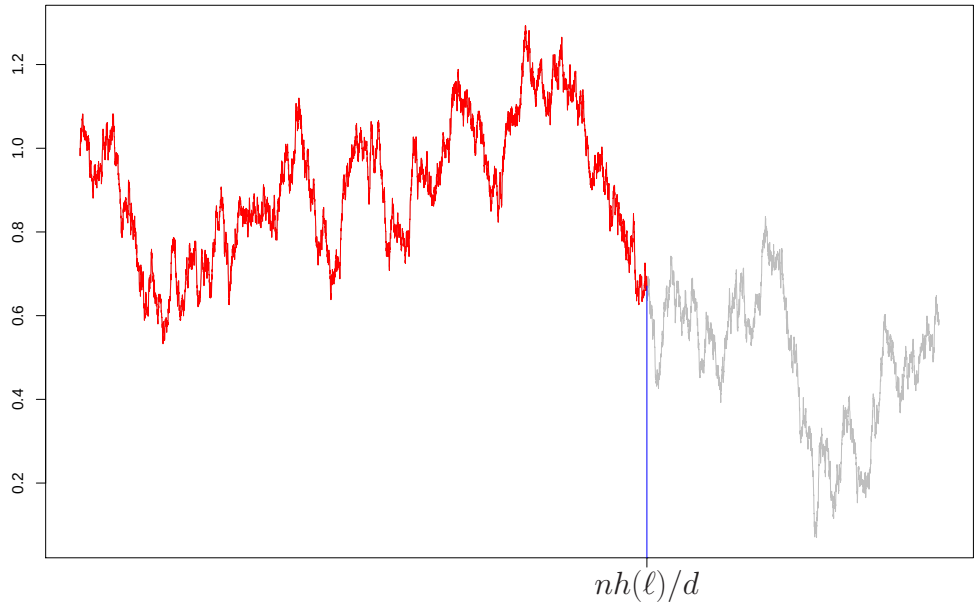
Z^d *is* **not** *a Markov chain, however in the limit as* d *goes to* ∞, *it is Markov:*

$$Z_d \Rightarrow Z$$

where Z *satisfies the SDE,*

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

for some function $h(\ell)$.



How much diffusion path do we get for our n iterations?

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

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

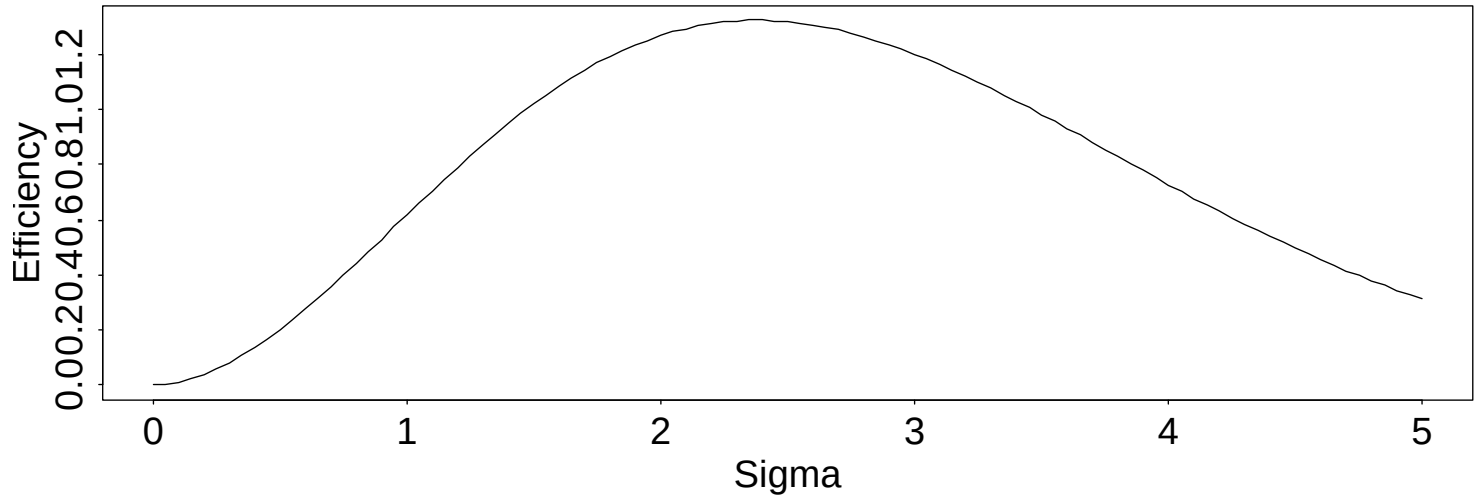
$$h(\ell) = \ell^2 \times A(\ell) \ ,$$

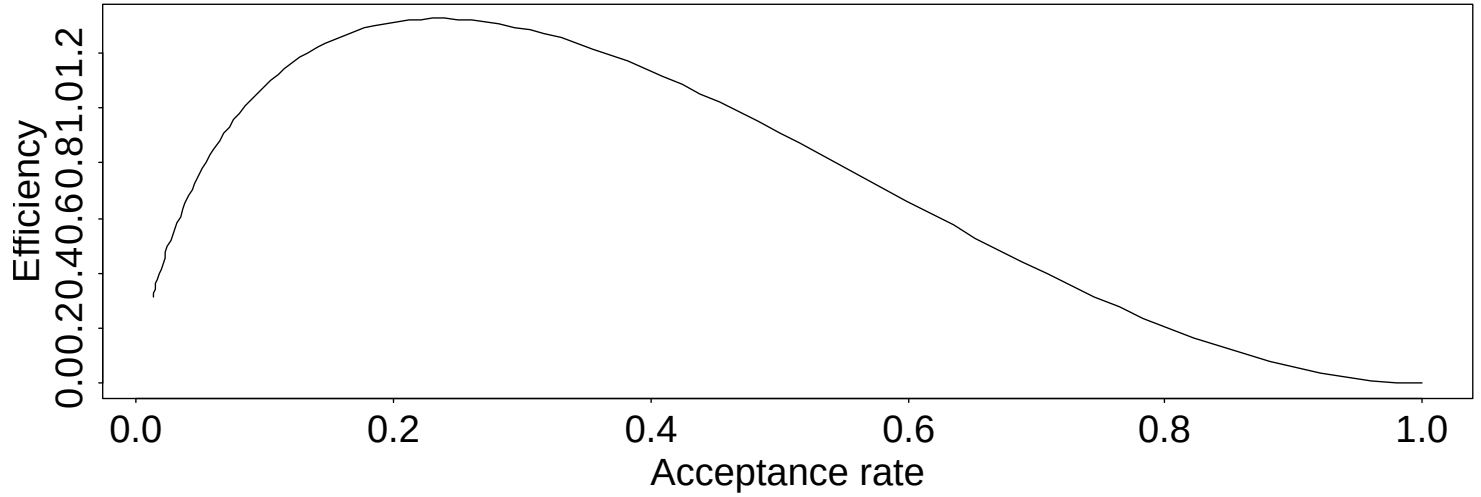
where $A(\ell)$ is the limiting overall acceptance rate of the algorithm, ie the proportion of proposed Metropolis moves ultimately accepted. So

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

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

Efficiency as a function of scaling and acceptance rate





When can we 'solve' the scaling problem for Metropolis?

We need a sequence of target densities π_d which are sufficiently regular as $d \to \infty$ in order that meaningful (and optimisable) limiting distributions exist. Eg.

1. $\pi \sim \prod_{i=1}^{d} f(x_i)$.(NB for discts f, mixing is $O(d^2)$, rate 0.13, (Neal).)
2. $\prod_{i=1}^{d} f(c_i x_i)$, $q(\mathbf{x}, \cdot) \sim N(\mathbf{x}, \sigma_d^2 I_d)$. for scales c_i. (Bedard, Rosenthal, Voss).
3. Elliptically symmetric target densities (Sherlock, Bedard).
4. The components form a homogeneous Markov chain.
5. π is a Gibbs random field with finite range interactions (Breyer).
6. Discretisations of an infinite-dimensional system absolutely cts wrt a Gaussian measure (eg Pillai, Stuart, Thiery).
7. Purely discrete product form distributions.
8. New methodology based on Dirichlet forms (Zanella, Kendall).

A basic analysis of Metropolis

Write $\mathbf{Y}^{(d)} = \mathbf{X}^{(d)} + h^{1/2}\mathbf{Z}^{(d)}$.

$$\alpha(\mathbf{X}^{(d)}, \mathbf{Y}^{(d)}) = 1 \wedge \beta(\mathbf{X}^{(d)}, \mathbf{Y}^{(d)}) = 1 \wedge \frac{\pi^{(d)}(\mathbf{Y}^{(d)})}{\pi^{(d)}(\mathbf{X}^{(d)})}$$

Now in the case of product form IID target densities:

$$\log \beta((\mathbf{X}^{(d)}, (\mathbf{X}^{(d)} + h^{1/2}(\mathbf{Z}^{(d)})) = \sum_{i=1}^{d} \log f(X_i^{(d)} + h^{1/2} Z_i^{(d)}) - \log f(X_i^{(d)})$$

so components decouple, we can take a Taylor series expansion and as long as certain moments are finite, we can get a CLT for $\log \beta$. See later.

What about more generally when π does not take the IID product form case?

$$\log \beta((\mathbf{X}^{(d)}, (\mathbf{X}^{(d)} + h^{1/2}(\mathbf{Z}^{(d)}))$$

$$\approx h^{1/2} \nabla \log \pi(\mathbf{X}^{(d)}) \cdot \mathbf{Z}^{(d)} + \frac{1}{2} h \mathbf{Z}^{(d)\prime} \nabla \nabla' \log \pi(\mathbf{X}^{(d)}) \mathbf{Z}^{(d)}$$

$$\beta \approx \exp\{h^{1/2} G^{(d)} - h V^{(d)}/2\}$$

G is Gaussian and if $V^{(d)}$ converges in probability to a constant V (with h scaling appropriately with d), then the variance of G is V (as β has to have unit mean.).

In fact that is all we need for the 0.234 framework to hold!

Why?

Suppose now G is a standard Gaussian and ℓ incorporates scaling choice:

$$\text{ESJD} \approx \ell^2 \mathbf{E}\left(1 \wedge \exp\left\{\ell V^{1/2} G - \ell^2 V/2\right\}\right)$$

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

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

which is maximised by taking $A(\ell) = 0.234$.

Some remarks about the proof

Use generator approach to weak convergence.

First component is not Markov but has Markov limit. Standard approach (Ethier and Kurtz, 1986) requires that for the approximate generator applied to all functions in a core for the generator of the limiting process (in our case smooth functions with bounded support) converges to its appropriate limit in supremum norm.

How to cope with dependence on other $d-1$ components?

Let V be a test function (of first component only) and let G_d be the generator of the continuous time process which makes a Metropolis move at times generated through a Poisson process of rate d:

$$\begin{aligned} G_d V(\mathbf{x}^d) &= d\mathbb{E}\left[(V(Y_1) - V(x_1))\left\{1 \wedge \frac{\pi(\mathbf{Y})}{\pi(\mathbf{x}^d)}\right\}\right] \\ &= d\mathbb{E}\left[(V(Y_1) - V(x_1))\mathbb{E}\left[\left\{1 \wedge e^{B_d}\right\}|Y_1\right]\right], \end{aligned}$$

where

$$B_d = \log f(Y_1) - \log f(x_1) + \sum_{i=2}^{d} (\log f(Y_i) - \log f(x_i)).$$

But the blue term can be approximated by a Taylor expansion, LLN, and CLT (setting $g = \log f$):

$$\sum_{i=2}^{d} (\log f(Y_i) - \log f(x_i)) \approx \sum_{i=2}^{d} g'(x_i)(Y_i - x_i) + \frac{1}{2} g''(x_i)(Y_i - x_i)^2$$

$$\approx N(H_d, K_d)$$

where

$$H_d = \frac{\sum_{i=2}^{d} g''(x_i)\ell^2}{2d}$$

and

$$K_d = \frac{\sum_{i=2}^{d} (g'(x_i))^2 \ell^2}{d} \ .$$

For large d, H_d and K_d are close to their respective limits H and K with the simple identities $H = -K/2$ and $K = I\ell^2$.

Now need to use a large deviation bound to show that H_d and K_d are uniformly close to their limits for a time period which is at least $O(d)$ at least when the chain is started in stationarity.

Then we can approximate:

$$G_d V(\mathbf{x}^d) = d\mathbb{E}\left[(V(Y_1) - V(x_1))\mathbb{E}\left[\left\{1 \wedge e^{B_d}\right\}|Y_1\right]\right]$$

$$\approx d\mathbb{E}\left[(V(Y_1) - V(x_1))\mathbb{E}\left[\left\{1 \wedge \exp\{N(D_1 - I\ell^2/2, I\ell^2)\}\right\}|Y_1\right]\right]$$

where $D_1 = \log f(x_1 + d^{-1/2}\ell Z_1) - \log f(x_1)$.

Therefore to compute the inner expectation, we need the following calculation. (See, R Gelman and Gilks (1997)).

Proposition 2 *If* $A \sim N(\mu, \sigma^2)$, *then*

$$\mathbb{E}[1 \wedge e^A] = \Phi(\frac{\mu}{\sigma}) + \exp(\mu + \frac{\sigma^2}{2})\Phi(-\sigma - \frac{\mu}{\sigma})$$

and

$$\mathbb{E}[e^A; A < 0] = \exp(\mu + \frac{\sigma^2}{2})\Phi(-\sigma - \frac{\mu}{\sigma}).$$

Now we can perform a Taylor series expansion for the first component.

$$\begin{aligned} G_d V(\mathbf{x}^d) &\approx \frac{l^2}{2} V''(x_1)\mathbb{E}[1 \wedge e^{B_d}] + l^2 g'(x_1)V'(x_1)\mathbb{E}[e^{B_d}; B_d < 0] \\ &\approx \frac{l^2}{2} V''(x_1)\left(2\Phi\left(-\frac{l\sqrt{I}}{2}\right)\right) + l^2 g'(x_1)V'(x_1)\Phi\left(-\frac{l\sqrt{I}}{2}\right) \\ &= h(\ell)\left\{\frac{1}{2}V''(x_1) + \frac{1}{2}g'(x_1)V'(x_1)\right\} = GV(x_1) \end{aligned}$$

which is the generator of the required limiting diffusion.

Complexity

Need to unify limit algorithm theory with Markov chain complexity theory.

Cannot (for instance) use

$$T_d = \sup_{x \in \mathbb{R}^d} \{t; \|P_d^t(x, \cdot) - \pi\| < \epsilon\}$$

For any probability measures μ and π, let

$$\|\mu - \pi\|_{KR} = \sup_{f \in \mathrm{Lip}_1^1} (\mu(f) - \pi(f)).$$

The Kantorovich-Rubinstein metric metricises weak convergence.

The following comes from R+Rosenthal, 2016, JAP.

Theorem 3 *Let $X^d = \{X_t^d\}_{t\geq 0}$ be a stochastic process on $(\mathcal{X}, \mathcal{F}, \rho)$, for each $d \in \mathbb{N}$, which converges weakly to $X^\infty = \{X_t^\infty\}_{t\geq 0}$. Assume these processes all have the same stationary probability distribution π. Then for any $\epsilon > 0$, there are $D < \infty$ and $T < \infty$ such that*

$$\mathbb{E}_{X_0^d \sim \pi} \| \mathcal{L}_{X_0^d}(X_t^d) - \pi \|_{KR} \; < \; \epsilon \,, \quad t \geq T, \quad d \geq D \,.$$

The main point here is that for Markov chains the KR distance from stationarity is not necessarily non-increasing, but IS INCREASING when integrated with respect to π.

Can we use this to provide a formal complexity result for RWM?

Theorem 4 *Let X^d be a RWM algorithm satisfying technical conditions slightly stronger than R et al. (1997). Then for any $\epsilon > 0$, there is $D < \infty$ and $T < \infty$ such that*

$$\mathbb{E}_{X_0^d \sim \pi} \| \mathcal{L}_{X_0^d}(X^d_{\lfloor dt \rfloor, 1}) - h \|_{KR} \; < \; \epsilon \,, \quad t \geq T, \quad d \geq D \,.$$

Hence, the RWM algorithm takes $O(d)$ iterations to converge to within ϵ of stationarity in any one coordinate.

Can produce analogous statements for the subsequent diffusion scaling results.

Metropolis-within-Gibbs

Update $c_d d$ components at a time, components picked by random scan at each iteration.

How should be jointly choose (c_d, σ^2) to optimise the Markov chain?

This is a slightly simplified version of the commonly used strategy to update (usually deterministically) a subset of the d coordinates.

We again turn to the IID target distribution case for a rigorous result.

Theorem 5 *Suppose that* $c_d \to c$, *as* $d \to \infty$, *for some* $0 < c \leq 1$. *Then, as* $d \to \infty$,

$$U^d \Rightarrow U,$$

where U_0 *is distributed according to* f *and* U *satisfies the Langevin SDE*

$$dU_t = \sqrt{h_c(l)} dB_t + h_c(l) \frac{f'(U_t)}{2f(U_t)} dt$$

with $h_c(l) = 2cl^2\Phi(-\frac{l\sqrt{cI}}{2})$ *and* $I \equiv \mathbb{E}_f[(\frac{f'(X)}{f(X)})^2]$.

The important part of the theorem is $h_c(l)$, the speed of the limiting diffusion. The aim is to maximise $h_c(l)$.

Note that

$$h_c(l) = c \times l^2 \times 2\Phi(-\frac{l\sqrt{cI}}{2}).$$

Corollary 6 *Let* $c_d \to c$ *as* $d \to \infty$ *for some* $0 < c \leq 1$. *Then,*

(i) $\lim_{d\to\infty} a_d^{c_d}(l) = a^c(l) = 2\Phi(-\frac{l\sqrt{cI}}{2})$.

(ii) *Let* $\hat{l}$ *be the unique value of* l *which maximises* $h_1(l) = 2l^2\Phi(-\frac{l\sqrt{I}}{2})$, *and let* $\hat{l}_c$ *be the unique value of* l *which maximises* $h_c(l)$. *Then* $\hat{l}_c = c^{-\frac{1}{2}}\hat{l}$ *and* $h_c(\hat{l}_c) = h_1(\hat{l})$.

(iii) *For all* $0 < c \leq 1$, *the optimal acceptance rate* $a^c(\hat{l}_c) = 0.234$ *(to three decimal places).*

The key consequence of Corollary 6 is that updating any proportion $c > 0$ of the components is asymptotically equivalent to full RWM (*i.e.* $c = 1$).

Let V be any bounded, continuous, sufficiently differentiable, real valued function.

Let

$$GV(x) = h_c(l)\{\frac{1}{2}V''(x) + \frac{1}{2}g'(x)V'(x)\},$$

Then G is the generator of the Langevin SDE described in Theorem 5.

Let V be a function of the first component only and

$$G_dV(\mathbf{x}^d) = d\mathbb{E}[(V(\mathbf{Y}) - V(\mathbf{x}^d))\{1 \wedge \frac{\pi(\mathbf{Y})}{\pi(\mathbf{x}^d)}\}].$$

Then to prove the Theorem 5, it is sufficient to show that

$$\sup_{\mathbf{x}^d \in F_d} |G_dV(\mathbf{x}^d) - GV(x_1)| \to 0 \quad \text{as } d \to \infty$$

where $d\mathbb{P}(F_d^c) \to 0$ as $d \to \infty$.

Is this just a continuous state space phenomenon?

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 7 *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 $B \sim Binomial(r-1, p)$),

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

What happens in the smooth limit?

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

The overall acceptance rate of the algorithm is given by

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

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

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

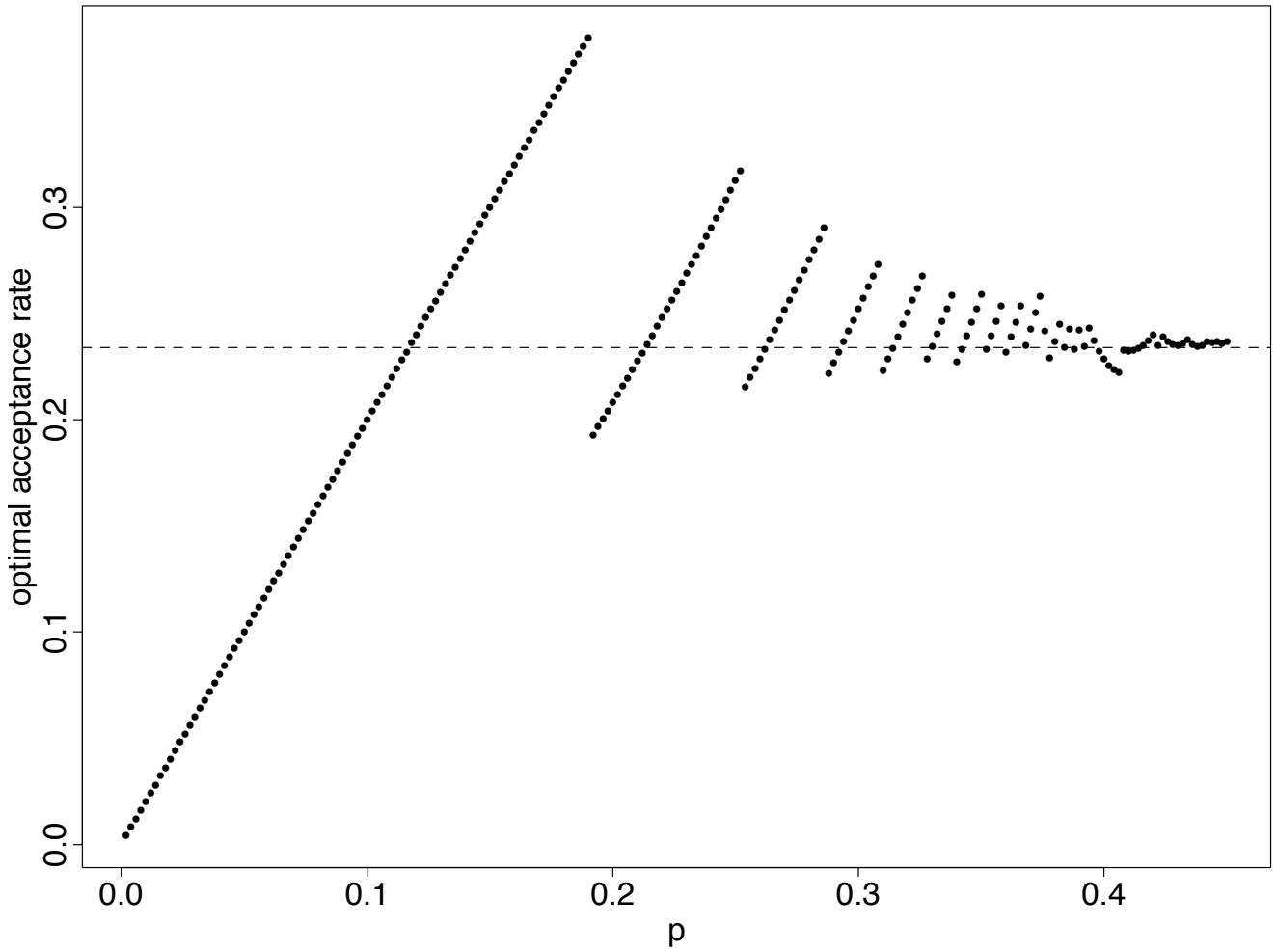
and

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

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

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

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 the continuous problem.



Optimal limiting acceptance rate as a function of p.

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 8 *Consider a random-walk Metropolis algorithm, with target density of this form, where* $\{C_i\}$ *are i.i.d. with* $E(C_i) = 1$, *and set* $b \equiv E(C_i^2) < \infty$. *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*

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

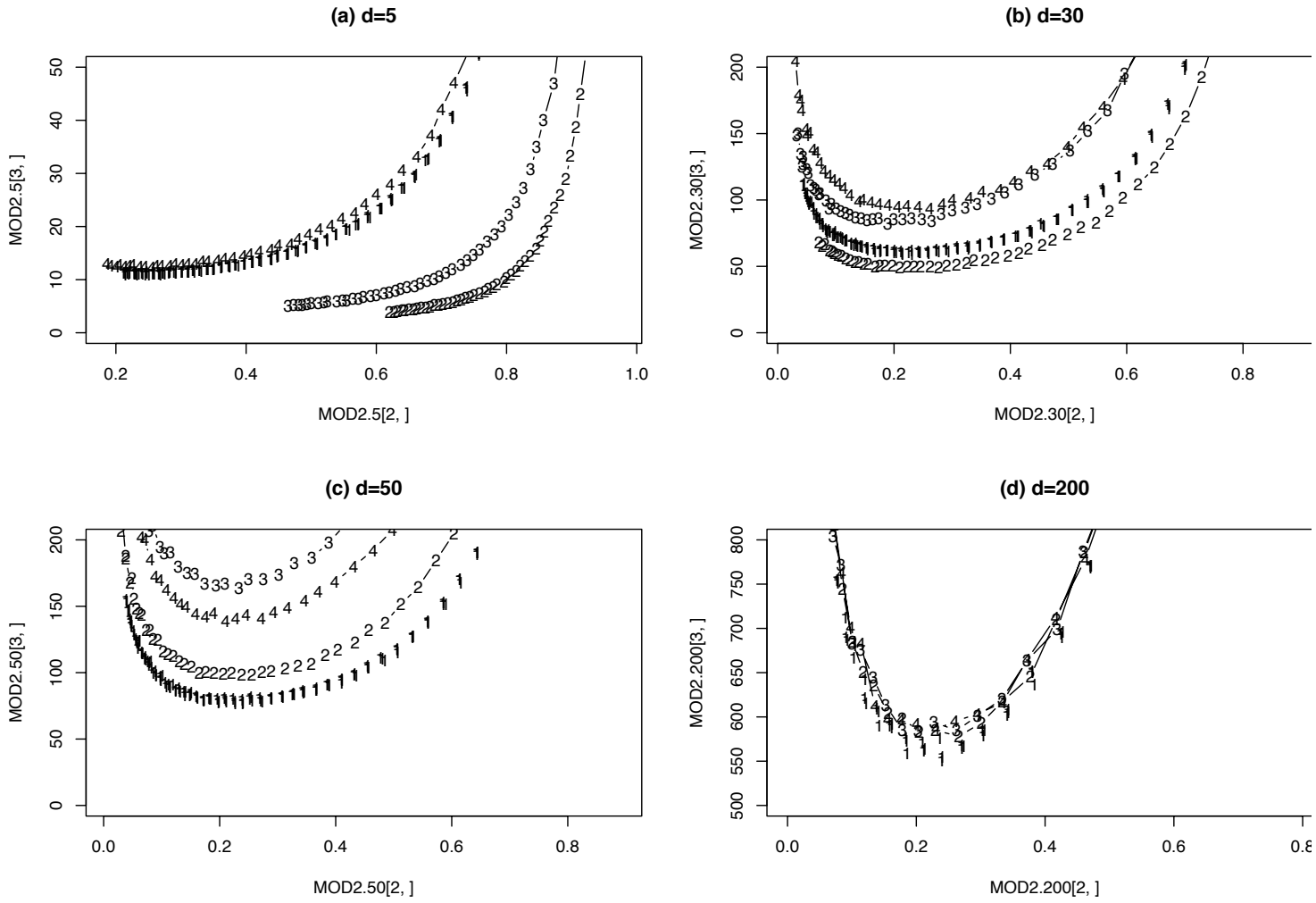
where B_t *is standard Brownian motion, and where*

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

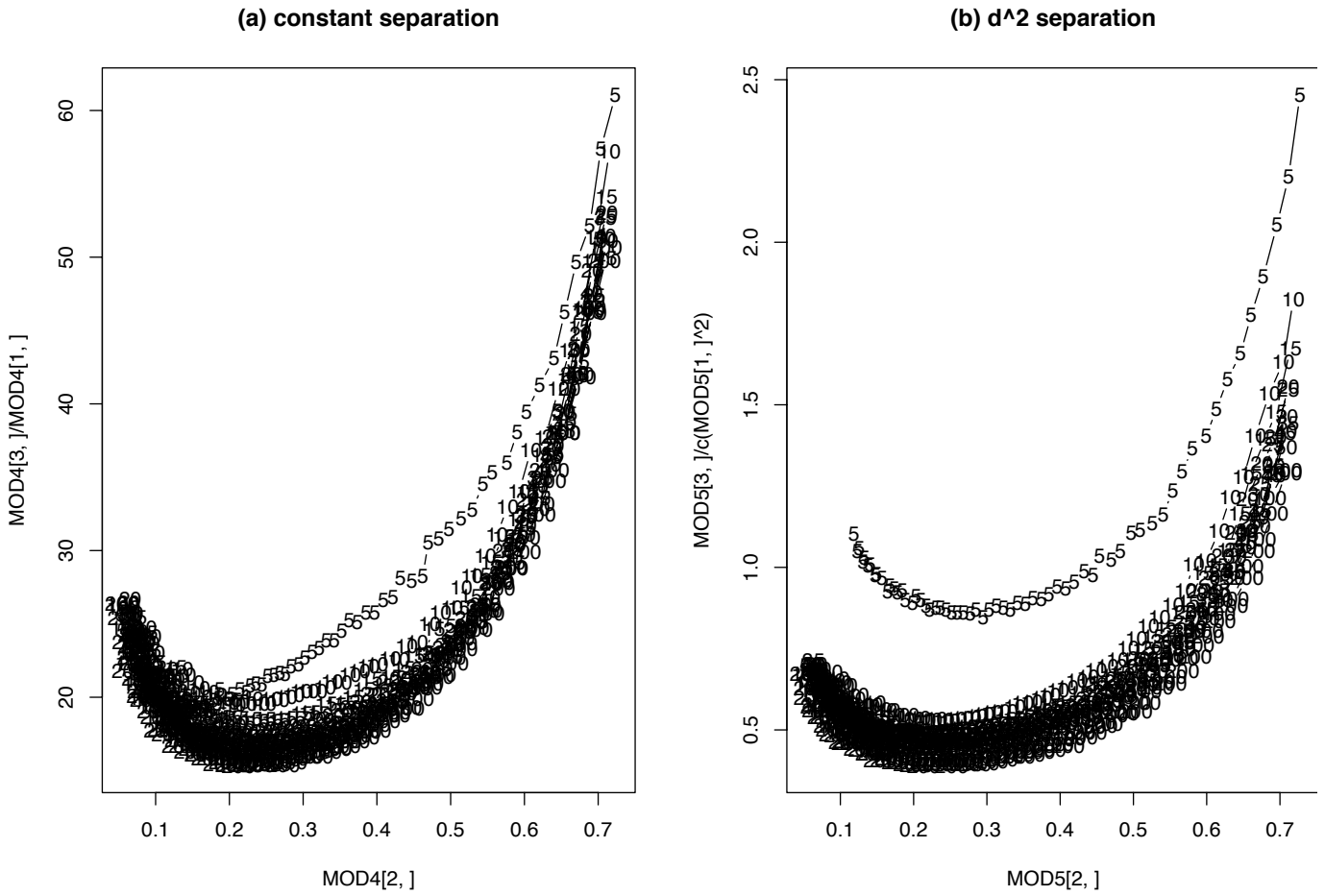
with $I = E_f[(g'(X))^2]$.

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. target 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 $\frac{C_1^2}{bd}$.

b (in general defined as $E(C_i^2)/E(C_i)^2$) acts as a term to measure average costs of heterogeneity



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.

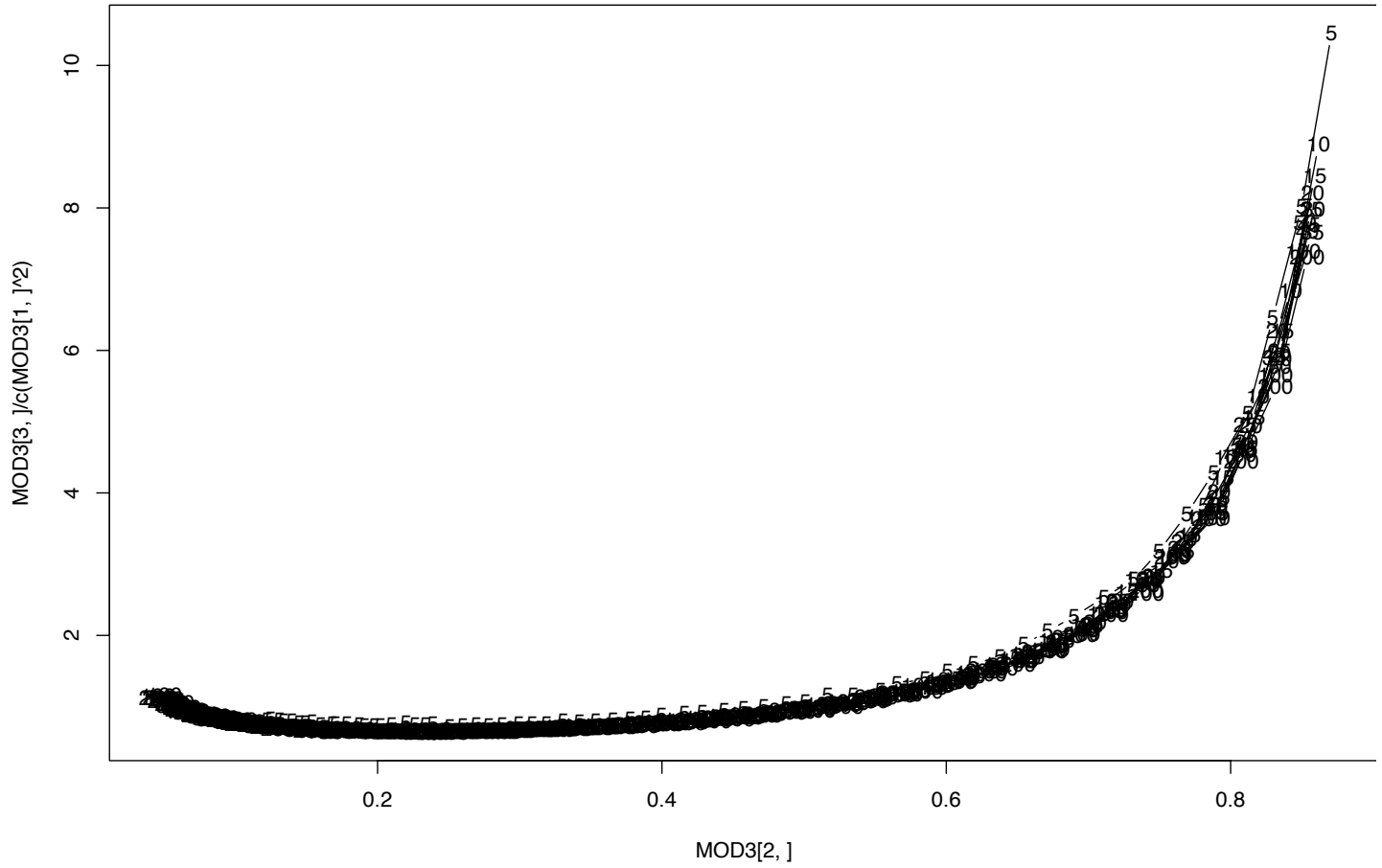


(a) A bimodal (mixture of normals) example, with modes separated by 3 sds in the first component (b) $O(d^{1/2})$ separation (where diffusion limits exist but are reducible). In this case, the 0.234 guideline is misleading!

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 $d\rho + 1 - \rho$.

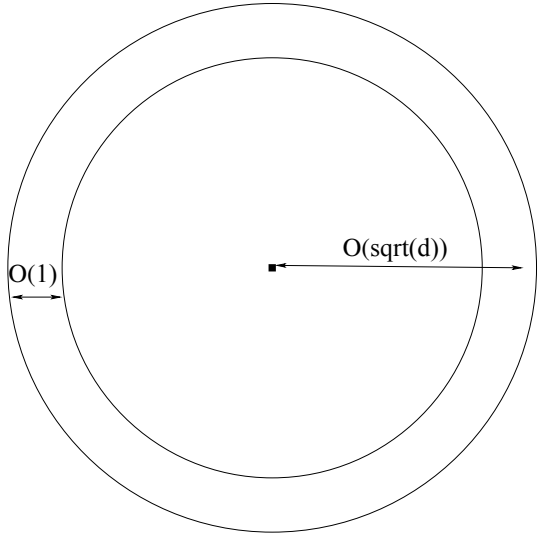
By an orthogonal transformation, we can write π as a collection of d independent normals with variances equal to the eigenvalues of Σ. Therefore can put in above framework with $C_1 = (d\rho + 1 - \rho)^{-1/2}$ and $C_i = (1 - \rho)^{-1/2}$ otherwise. The principle eigenfunction here is $\bar{x}$ so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



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

Picturing RWM in high dimensions

eg consider $\mathbf{X} \sim N(\mathbf{0}, I_d)$: $\mathbf{X}'\mathbf{X}$ has mean d and s.d. $(2d)^{1/2}$, so



Target distribution lies concentrated around the surface of a d-dim hyper-sphere.

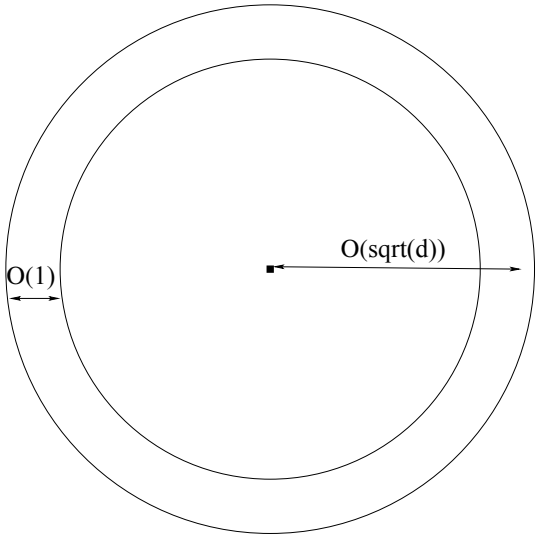
Two independent processes, the radial process (1-dimensional), needing to move $O(1)$) and the angular one (with a need to move distances $O(d^{1/2})$). Which process converges quickest?

Spherical symmetry (Sherlock and R, 2009, Bernoulli)

Theorem 9 *Let* $\{\mathbf{X}^{(d)}\}$ *be a sequence of* $d-$*dimensional spherically symmetric unimodal target distributions and let* $\{\mathbf{Y}^{(d)}\}$ *be a sequence of jump proposal distributions. If there exist sequences* $\{k_x^{(d)}\}$ *and* $\{k_y^{(d)}\}$ *such that the marginal radial distribution function of* $\mathbf{X}^{(d)}$ *satisfies* $|\mathbf{X}^{(d)}|/k_d \xrightarrow{D} R$ *where* R *is a non-negative random variable with no point mass at* 0, $|\mathbf{Y}^{(d)}|/k_y^{(d)} \xrightarrow{m.s.} 1$, *and provided there is a solution to an explicit integral equation involving the distribution of* R, *then suppose that* α_d *denotes the optimal acceptance probability (in the sense of minimising the expected squared jumping distance satisfies*

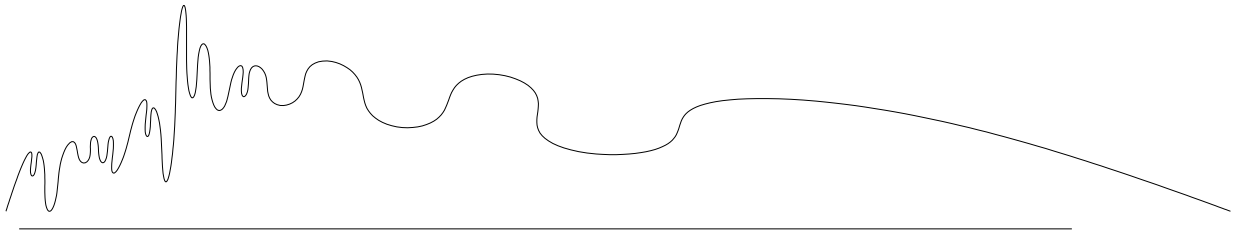
$$0 < \lim_{d\to\infty} \alpha_d = \alpha_\infty \leq 0.234$$

with $\alpha_\infty = 0.234$ *if and only if* R *equals some fixed positive constant with probability* 1. *If* R *does have a point mass at* 0, *OR the integral condition does not hold (essentially* R *has a heavy tailed distribution) then* $\alpha_\infty = 0$.



Where the radial component does not converge to a point mass, the target distribution has heterogenous roughness. See also recent work by Kamatani.

Does this happen in other situations?



Eccentricity

Theorem 10 *Suppose we can write* $\mathbf{X}^{(d)} = T_d^{-1/2}\mathbf{Z}^{(d)}$ *for matrices* $\{T_d\}$ *each having collections of (positive) eigenvalues* $\{\nu_i^{(d)}; 1 \leq i \leq d\}$, *and where* $\{\mathbf{Z}^{(d)}\}$ *be a sequence of* $d-$*dimensional spherically symmetric unimodal target distributions and let* $\{\mathbf{Y}^{(d)}\}$ *be a sequence of jump proposal distributions. If the conditions of previous theorem hold (on* $\mathbf{Z}^{(d)}$ *rather than* $\mathbf{Z}^{(d)}$ *this time). Suppose that* $\{T_d\}$ *are not* too eccentric*:*

$$\lim_{d\to\infty} \frac{\sup_{1\leq i\leq d} \nu_i^{(d)}}{\sum_1^d \nu_i^{(d)}} = 0 \ ,$$

then suppose that α_d *denotes the optimal acceptance probability (in the sense of minimising the expected squared jumping distance satisfies*

$$0 < \lim_{d\to\infty} \alpha_d = \alpha_\infty \leq 0.234$$

with $\alpha_\infty = 0.234$ *if and only if* R *equals some fixed constant with probability* 1.

See also work by Mylene Bedard.

Another example of different speeds

A caricature of MCMC on models with unidentifiable parameters (eg certain inverse problems).

Consider the target distribution $\pi_\varepsilon : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \mapsto \mathbb{R}$:

$$\pi_\varepsilon(x, y) = \pi(x)\, \pi_\varepsilon(y|x) = \frac{1}{\varepsilon^{d_y}}\, e^{A(x)+B(x,y/\varepsilon)} \ ,$$

with $\varepsilon > 0$ being 'small'. Propose

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} + \ell\, h(\varepsilon) \begin{pmatrix} Z_x \\ Z_y \end{pmatrix} \ , \tag{1}$$

for constant $\ell > 0$, scaling factor $h(\varepsilon)$ and noise $(Z_x, Z_y)^\top \sim N(0, I_{d_x+d_y})$.

$$\alpha = \alpha(x, Y, Z_x, Z_y) = 1 \wedge e^{A(x')-A(x)+B(x',Y')-B(x,Y)} \tag{2}$$

where we have set:

$$Y = y/\varepsilon \ ; \quad Y' = Y + \ell\, \tfrac{h(\varepsilon)}{\varepsilon}\, Z_y \ .$$

Theorem 11 *Consider the continuous-time process:*

$$x_{\varepsilon,t} = x_{\lfloor t/h(\varepsilon)^2 \rfloor} \ , \quad t \geq 0 \ , \tag{3}$$

started in stationarity, $\bar{x}_0 \sim \pi(x)$. *Assume that* $h(\varepsilon) \to 0$ *as* $\varepsilon \to 0$. *Then, as* $\varepsilon \to 0$, *we have that* $x_{\varepsilon,t} \Rightarrow x_t$ *with* x_t *the diffusion process specified as the solution of the stochastic differential equation:*

$$dx_t = \tfrac{\ell^2}{2} \left(a_0(x_t, \ell) \, \nabla A(x_t) dt + \nabla a_0(x_t, \ell) \right) + \sqrt{a_0(x_t, \ell) \ell^2} \, dW_t \ ,$$

where a_0 *denotes the acceptance probability of moves around* x*:*

$$a_0(x, \ell) = \tfrac{1}{(2\pi)^{dy/2}} \int_{\mathbb{R}^{dy} \times \mathbb{R}^{dy}} \left(1 \wedge e^{B(x, Y + \ell Z) - B(x,Y)} \right) e^{B(x,Y)} dY \, dZ \ . \tag{4}$$

Optimal scaling for the diverging scales problem

By analysing the form of the acceptance probability a_0, we get a surprise!

- If $d_Y = 1$, it is optimal to propose jumps of size $0(1)$, the limiting optimal algorithmis a continuous time pure jump process. Cost of heterogeneity $= \varepsilon^{-1/2}$. optimal acceptance probability is 0!
- If $d_Y \geq 3$, the diffusion regime is optimal, cost of heterogeneity is $O(\varepsilon^{-1})$, optimal acceptance probability can be anything.
- If $d_y = 2$, anything can happen ..

Discontinuous target density

Can significantly weaken the smoothness assumptions needed for the original result (Theorem 1), eg to π is differentiable in L^p mean, (Durmus et. al., 2016). But smoothness is not completely redundant ..

Suppose $\pi \sim \prod_{i=1}^{d} f(x_i)$ is supported on the unit hypercube, with

$$f(x) = \exp(g(x)), \quad 0 < x < 1$$

where $g \in C^1[0,1]$. We assume that g is suitable differentiable and at least one of $g(0)$ or $g(1)$ is not 0.

$q(\mathbf{x}, \cdot) \sim \prod_{i=1}^{d} U(x_i - \sigma_d, x_i + \sigma_d)$, $\mathbf{X}_0 \sim \pi$.

Is there any penalty for the discontinuity?

Theorem 12 *Set* $\sigma_d^2 = \ell^2/d^2$. *Consider*

$$Z_t^d = X^{(1)}_{[td^2]} \ . \qquad \textit{Speed up time by factor } d^2$$

$$Z_d \Rightarrow Z$$

where Z *satisfies the reflected Langevin SDE on* $[0,1]$,

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

with

$$h(\ell) = \frac{2\ell^2}{3} \exp\left(-\frac{f^*\ell}{2}\right)$$

and $f^* = \lim_{x \downarrow 0} \left(\frac{f(x)+f(1-x)}{2}\right)$

So algorithm is now $O(d^2)$.

Optimal limiting acceptance probability is now 0.13

Suppose $\pi \sim \prod_{i=1}^{d} f(x_i)$, and we assume that f is 'sufficiently' smooth.

We consider the MALA proposal given by

$$\mathbf{Y}_{n+1} \sim N(\mathbf{X}_n + \frac{\sigma_d^2}{2}\nabla \log \pi(\mathbf{X}_n),\ \sigma_d^2 I_d). \tag{5}$$

The scaling $\sigma_d^2 = O(d^{-1/3})$ was suggested in the physics literature The following result is taken from R + Rosenthal, 1998.

Theorem 13 *Consider a Metropolis-Hastings chain $\mathbf{X}_0, \mathbf{X}_1, \ldots$ for a target distribution having density π, and with MALA proposals as above.*

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

$$Z_t^d = X^{(1)}_{[d^{1/3}t]},$$

where $X_n^{(1)}$ is the first component of $\mathbf{X}_n$. Then, assuming various regularity conditions on the densities f (described in detail in R + Rosenthal, 1998), as $d \to \infty$, we have the following:

1. *Z^d converges weakly to the continuous-time process Z satisfying*

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

where

$$g(\ell) = 2\ell^2 \Phi(-J\ell^3), \tag{6}$$

and J is given by

$$J = \sqrt{\mathbf{E}\left(\frac{5(\log f)'''(X)^2 - 3(\log f)''(X)^3}{48}\right)} \ , \tag{7}$$

where the expectation is with respect to f.

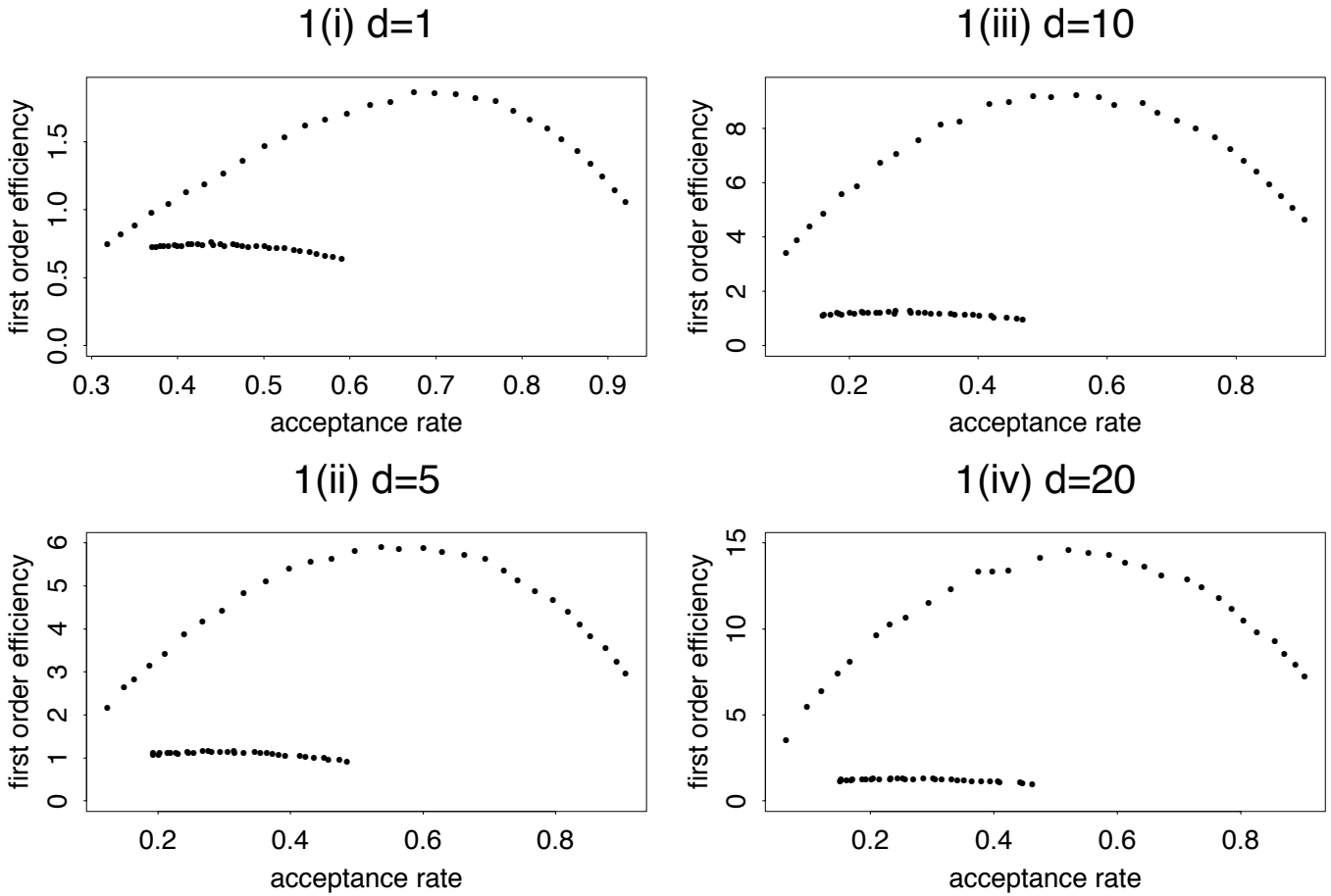
2. *The acceptance rate, a, of the algorithm is given by $2\Phi(-J\ell^3)$, and the scaling which gives optimal asymptotic efficiency is that having asymptotic acceptance rate equal to* 0.574.

For componentwise Langevin strategies, the conclusion is the opposite of that for Metropolis: the optimal is to update as many components as possible.

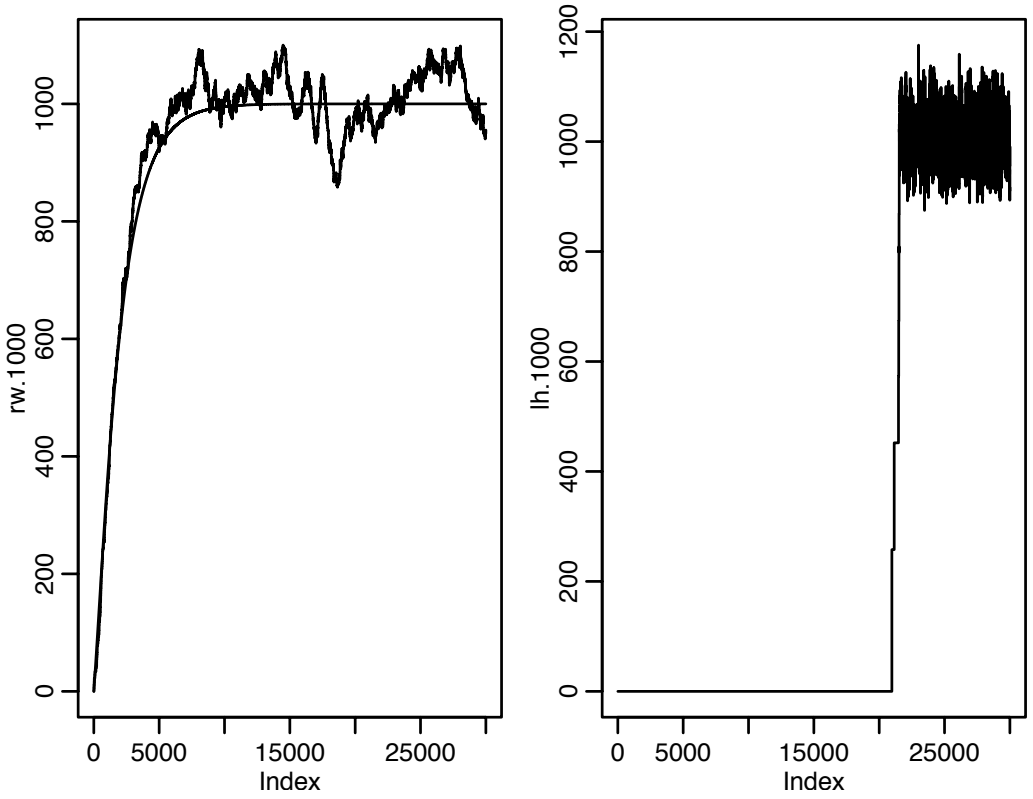
Similar limits exist for other Langevin schemes.

Hamiltonian Monte Carlo has pure jump process limit.

A comparison of MALA to RWM



Non-stationary initial distribution



Gaussian example

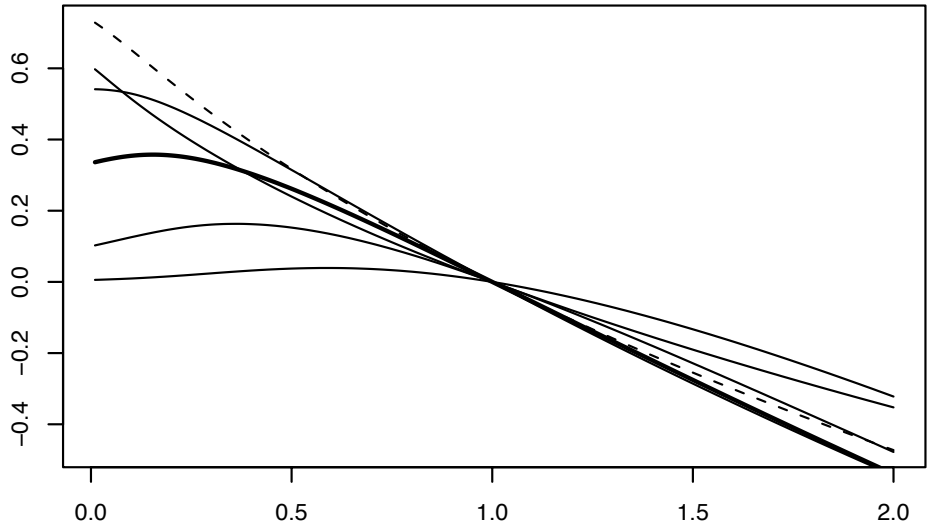
Set $\pi \sim MVN_d(\mathbf{0}, I_d)$. Suppose we apply 'optimally scaled' RWM.

Consider $W_t^d = |\mathbf{X}_{[td]}|^2/\mathrm{d}$

Theorem 14 *When $W_0^d = w_0 \neq 1$, then as $d \to \infty$, we have $W^d \Rightarrow f$, where f is a deterministic function satisfying $f(0) = w_0$ and*

$$f'(t) \;=\; a_\ell(f(t))$$

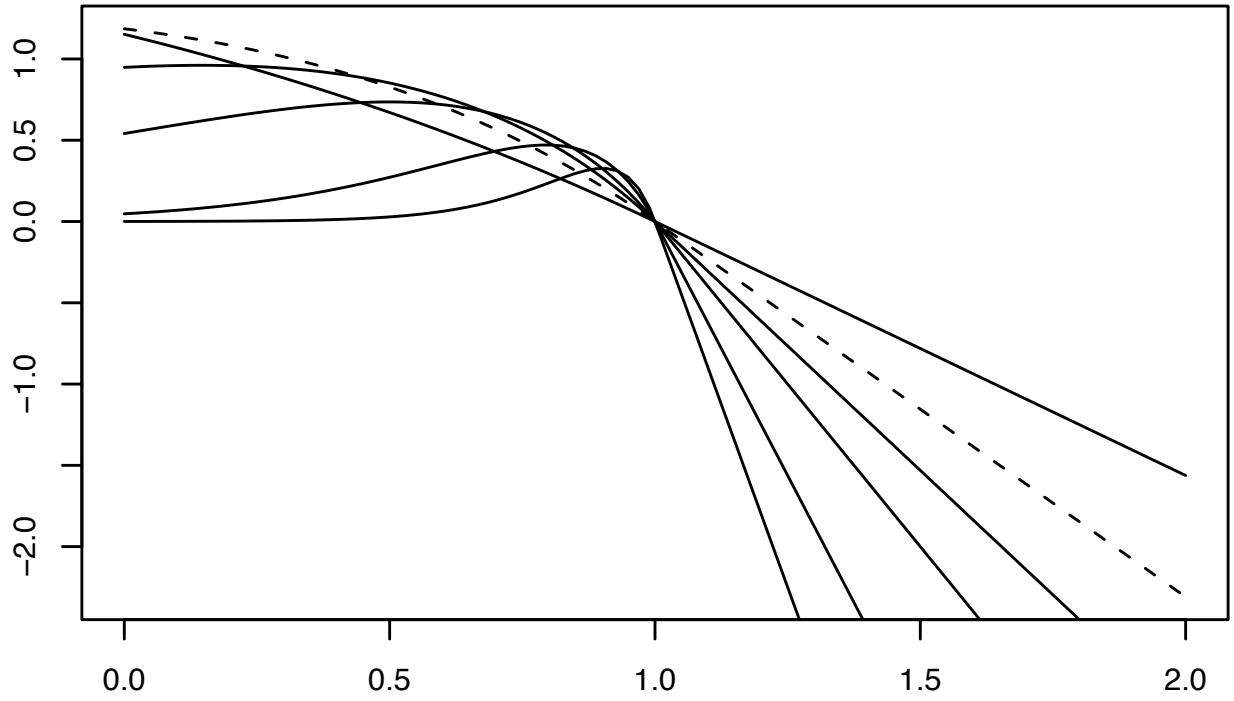
with function $a_\ell(\cdot)$ which can be explicitly calculated.



Langevin case

Using the 'optimal' scaling it gets stuck...

Though using the scaling $\sigma_d^2 = \ell^2/d^{1/2}$, we get a similar deterministic limit result.



Deterministic convergence speed, $a_\ell(\cdot)$, the Langevin case.

These ideas recently generalised significanty by Jourdain, Lelievre and Miasojedow

A Point Process Example

From Møller, Syversveen and Waagepetersen (1998 Sc. J. Stat.) Locations of 126 Scots pine saplings in a Finnish forest

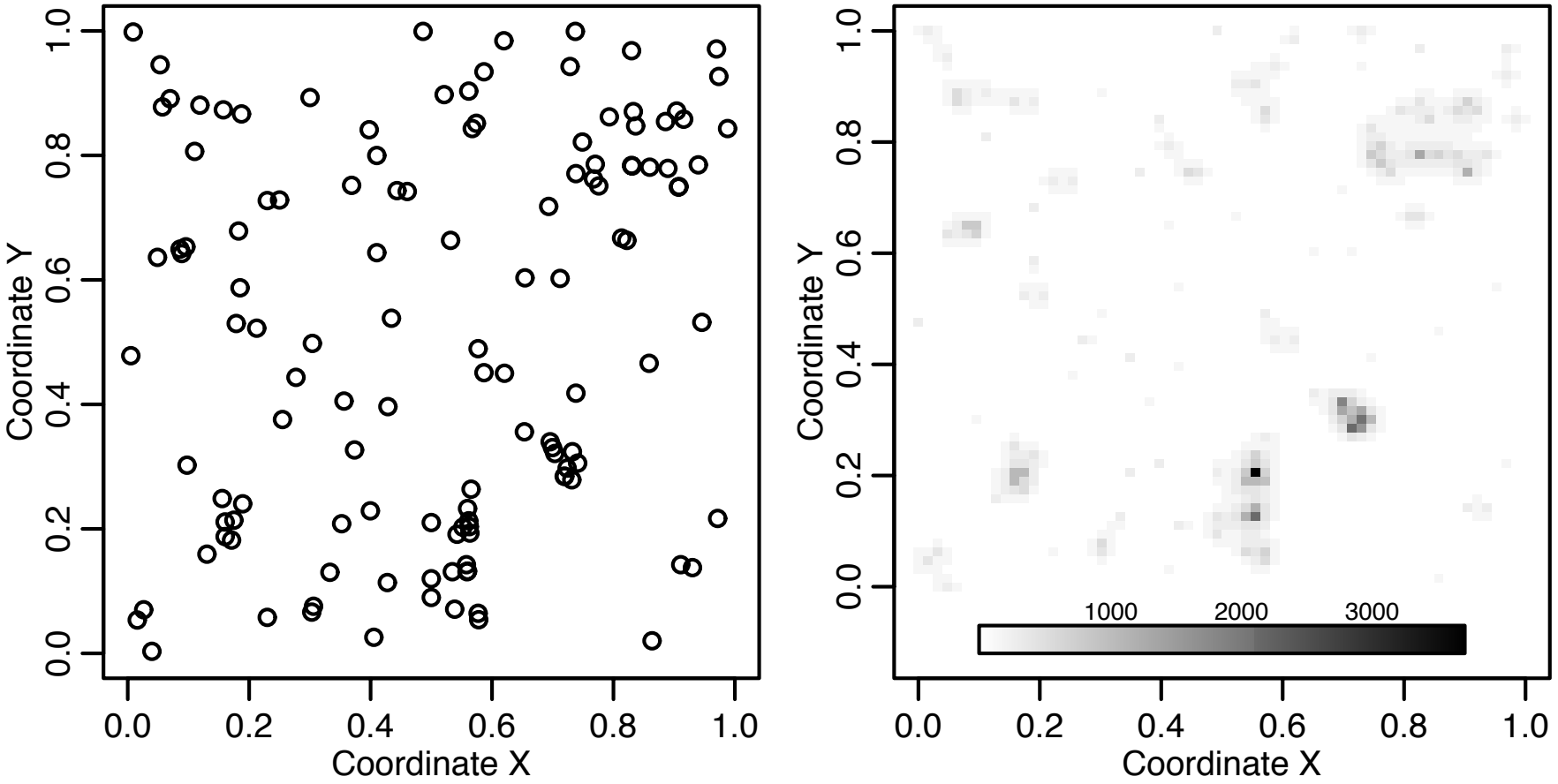
Observed point pattern modelled as a Poisson point process X with intensity

$$\Lambda(s) = \exp(Y(s)),$$

where $Y(\cdot) = \{Y(s) \mid s \in \mathbf{R}^2\}$ is a Gaussian process with mean $\mathrm{E}[Y(s)] = \mu$ and covariance

$$\mathrm{Cov}(Y(s), Y(s')) = \sigma^2 \exp(-\|s - s'\|/\beta).$$

The latent Gaussian process is discretised on a 64×64 regular grid.



Scotish pine saplings. Left : locations of trees. Right : the estimated intensity $\mathbf{E}[\Lambda(s) \mid x]$.

Updating latent Gaussian field requires MALA updates.

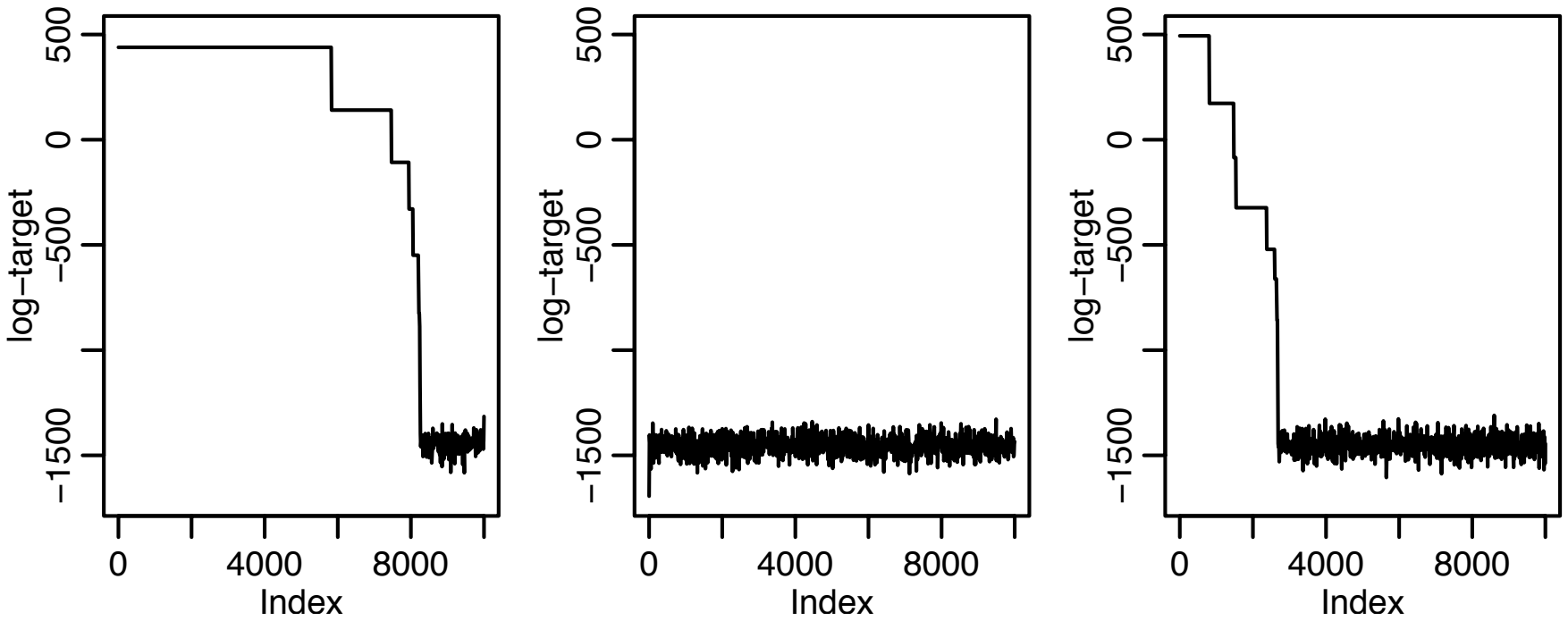
Compare the performance of the algorithm for three different starting values. The starting values expressed in terms of Y (which have to be transformed to starting values for Γ) are

I : $Y_{i,j} = \mu$ for $i, j = 1, \ldots, 64$.

II : a random starting value, simulated from the prior $Y \sim \mathrm{N}(\mu, \Sigma)$.

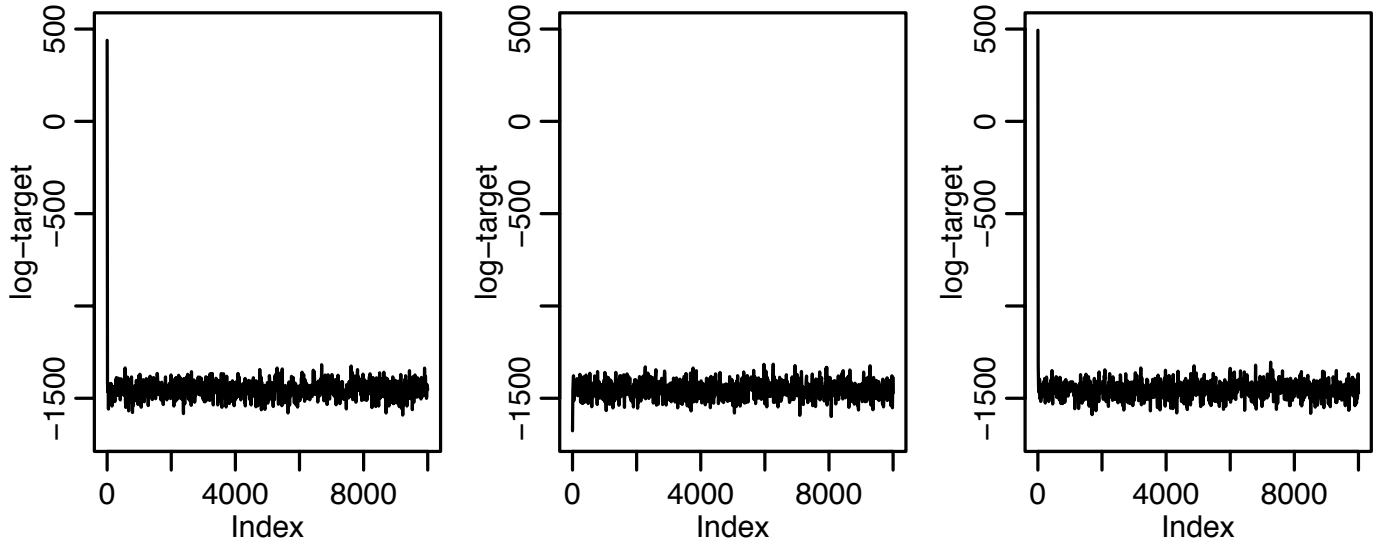
III : a starting value near the posterior mode. Let $Y_{i,j}$ solve the equation $0 = x_{i,j} - \exp(Y_{i,j}) - (Y_{i,j} - \beta)/\sigma^2$.

In all three cases we use the scaling $\hat{\ell}^2/(4096)^{1/3} = 0.16$ where $\hat{\ell} = 1.6$ is derived using 'optimal scaling' criteria.



Scots pine saplings. Traceplots $\log(\gamma \mid x)$ when using the scaling 0.16. Left : starting value I. Middle : starting value II. Right : starting value III.

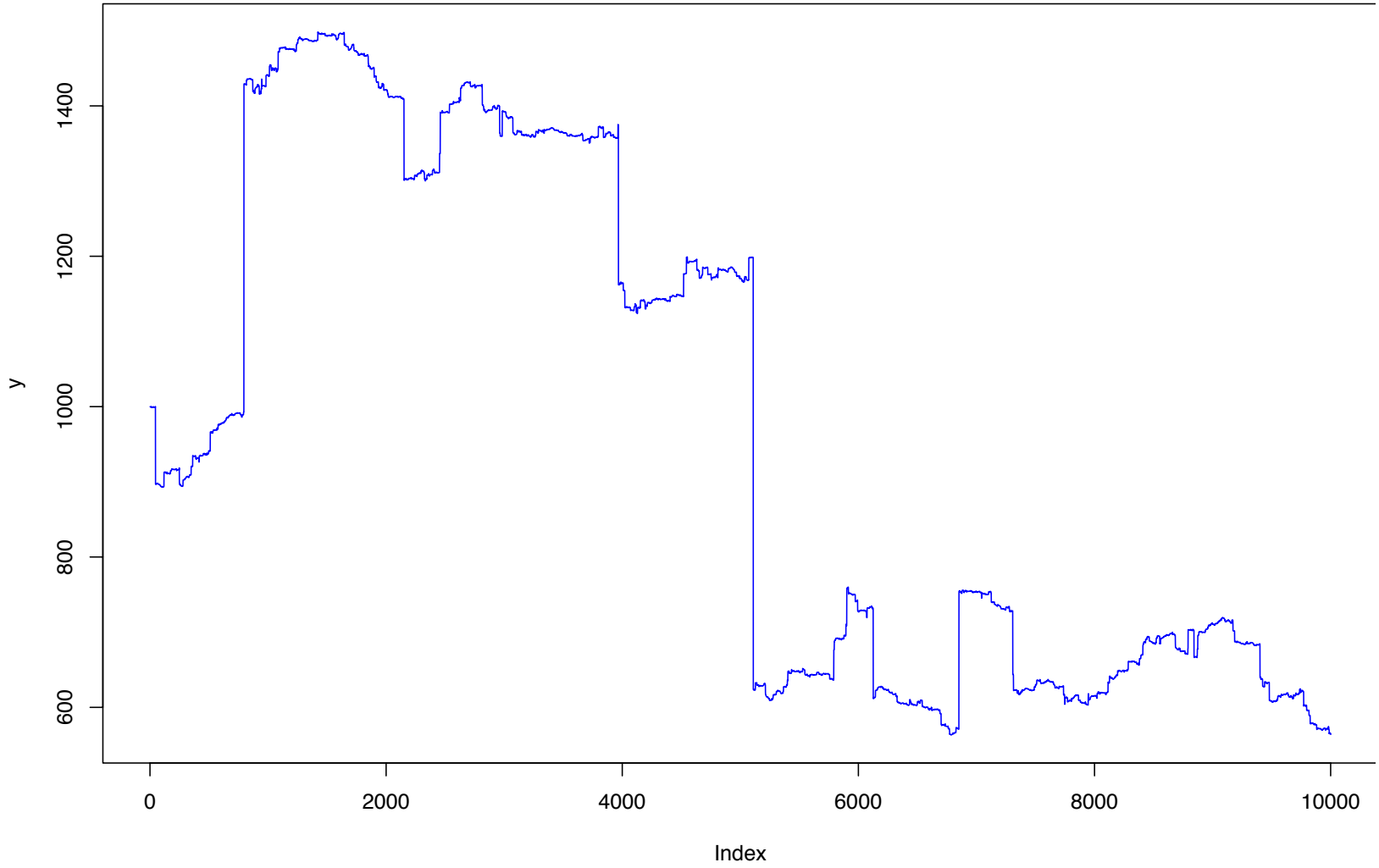
Now using the scaling $\hat{\ell}^2/(4096)^{1/2} = 0.034$. The acceptance rate for all algorithms was around 95%.

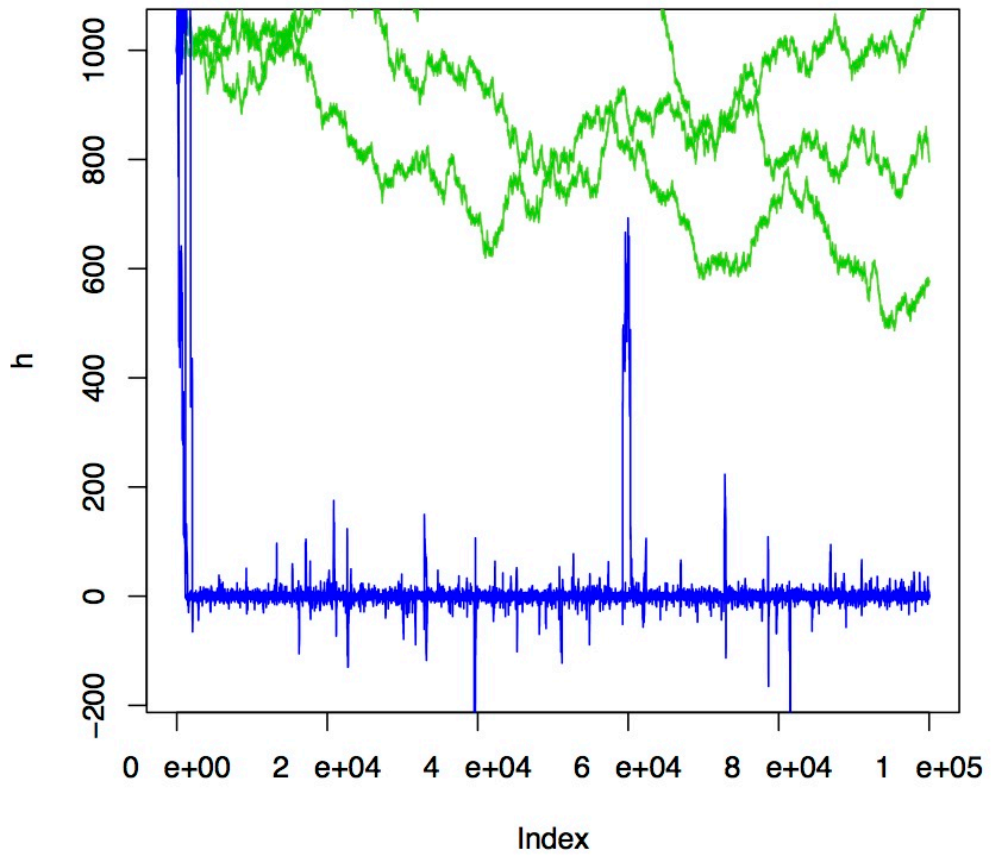


Scots pine saplings. Traceplots $\log(\gamma \mid x)$ when using the scaling 0.034. Left : starting value I. Middle : starting value II. Right : starting value III.

Heavy-tailed proposals

If proposal variance is infinite, all the above theory fails and diffusion limits cannot exist!





To fix ideas, consider RWM, and replace independent Gaussian proposals in each direction by independent Cauchy proposals in each direction.

Evidence from other results that heavy-tailed proposals improve mixing (eg Jarner and R, 2003, 2006)

Discontinuous targets, heavy-tailed proposals

Suppose $\pi \sim \text{Unif}(0,1)^d$.

$q(\mathbf{x}, \cdot) \sim \text{Cauchy}(\mathbf{x}, \sigma_d^2 I_d)$, $\mathbf{X}_0 \sim \pi$.

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

$$Z_t^d = X^{(1)}_{[td\log(d)]} \,. \qquad \text{Speed up time by factor } d\log d$$

$$Z_d \Rightarrow \text{ a scaled truncated Cauchy process}$$

with an associated explicit optimal scaling problem.

Here, light-tailed proposals are $O(d^2)$ while Cauchy proposals are $O(d\log d))$.

Simulated tempering

Consider a d-dimensional target density

$$f_d(x) \;=\; e^{dK} \prod_{i=1}^{d} f(x_i)\,,$$

for some unnormalised one-dimensional density function $f : \mathbf{R} \to [0, \infty)$, where $K = -\log(\int f(x)dx)$.

Consider simulated tempering in d dimensions, with inverse-temperatures chosen as follows: $\beta_0^{(d)} = 1$, and $\beta_{i+1}^{(d)} = \beta_i^{(d)} - \frac{\ell(\beta_i^{(d)})}{d^{1/2}}$ for some fixed C^1 function $\ell : [0, 1] \to \mathbf{R}$.

To stop adding new temperature values, we fix some $\chi \in (0, 1)$ and keep going until the inverse temperatures drop below χ, i.e. we stop at temperature $\beta_{k(d)}^{(d)}$ where $k(d) = \sup\{i : \beta_i^{(d)} \geq \chi\}$.

The optimal temperature spacing problem asks what is the optimal choice of the function ℓ.

We shall consider a joint process $(y_n^{(d)}, X_n)$, with $X_n \in \mathbf{R}^d$, and with $y_n^{(d)}\{\beta_i^{(d)};\ 0 \leq i \leq k(d)\}$ defined as follows.

Choose $X_{n-1} \sim f^\beta$, then proposing Z_n to be β_{i+1} or β_{i-1} with probability $1/2$ each, and then accepting Z_n with the usual Metropolis acceptance probability. We assume (unrealistically!) that the chain then immediately jumps to stationary at the new temperature, i.e. that mixing within a temperature is infinitely more efficient than mixing between temperatures.

The process $(y_n^{(d)}, X_n)$ is thus a Markov chain with stationary density

$$f_d(\beta, x) \;=\; e^{dK(\beta)} \prod_{i=1}^{d} f^\beta(x_i)\,,$$

where $K(\beta) = -\log \int f^\beta(x)dx$ is the normalising constant.

A diffusion limit for inverse temperature

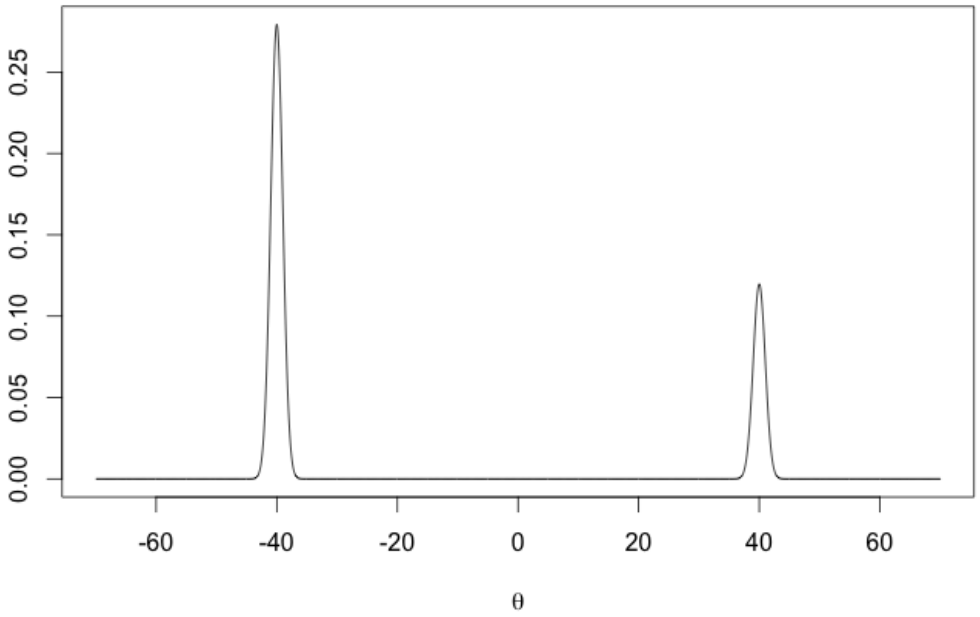
Theorem 15 *$\{y_n^{(d)}\}$ speeded up by a factor of d, converges weakly as $d \to \infty$ to a diffusion limit $\{X_t\}_{t \geq 0}$ satisfying*

$$dX_t = \left[2\ell^2 \Phi\left(\frac{-\ell I^{1/2}}{2}\right)\right]^{1/2} dB_t + \left[\ell(X)\ell'(X)\Phi\left(\frac{-I^{1/2}\ell}{2}\right) - \ell^2\left(\frac{\ell I^{1/2}}{2}\right)'\varphi\left(\frac{-I^{1/2}\ell}{2}\right)\right] dt\,,$$

for X_t in $(\chi, 1)$ with reflecting boundaries at both χ and 1.

Theorem 16 *The speed of this diffusion is maximised, and the asymptotic variance of all L^2 functionals is minimised, when the ℓ is chosen so that the asymptotic temperature acceptance probability at each and every temperature is equal to 0.234.*

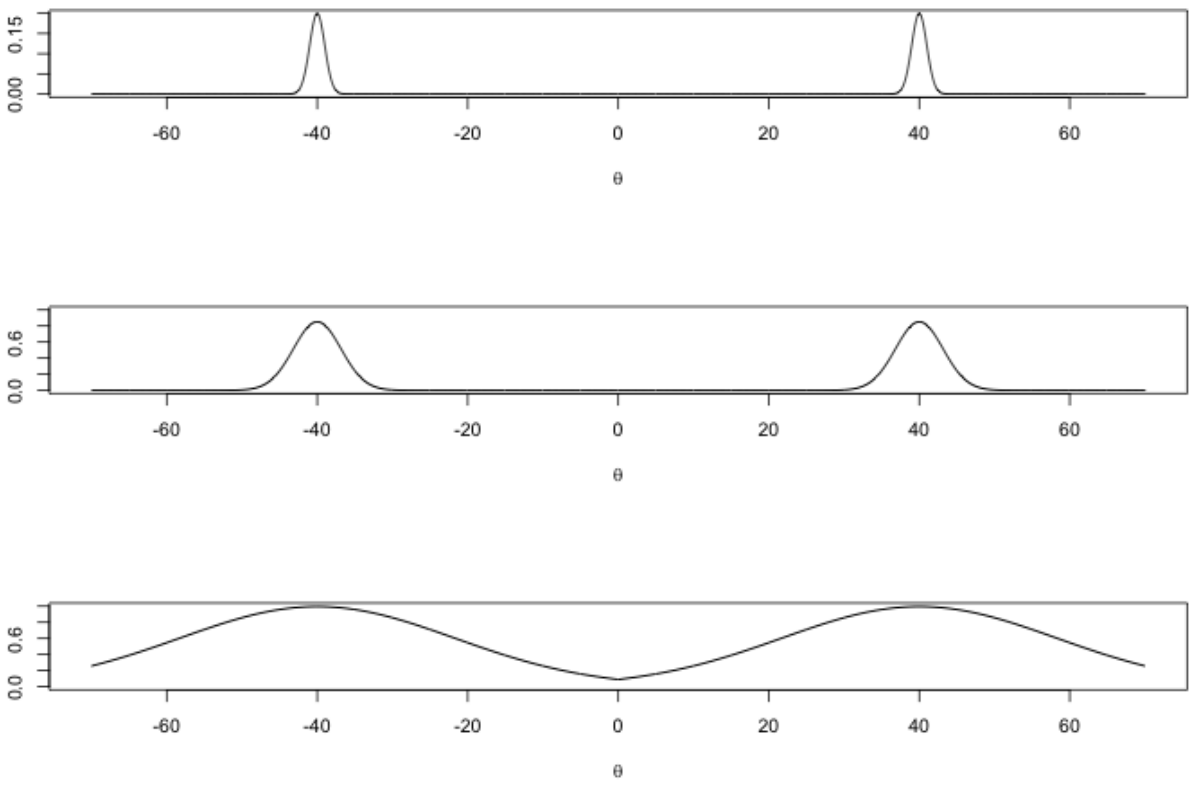
Simulating from high-dimensional multi-modal distributions



Hard even for some relatively small-dimensional problems.

Very hard for moderate to high-dimensional problems.

Simulating from high-dimensional multi-modal distributions



Does it work?

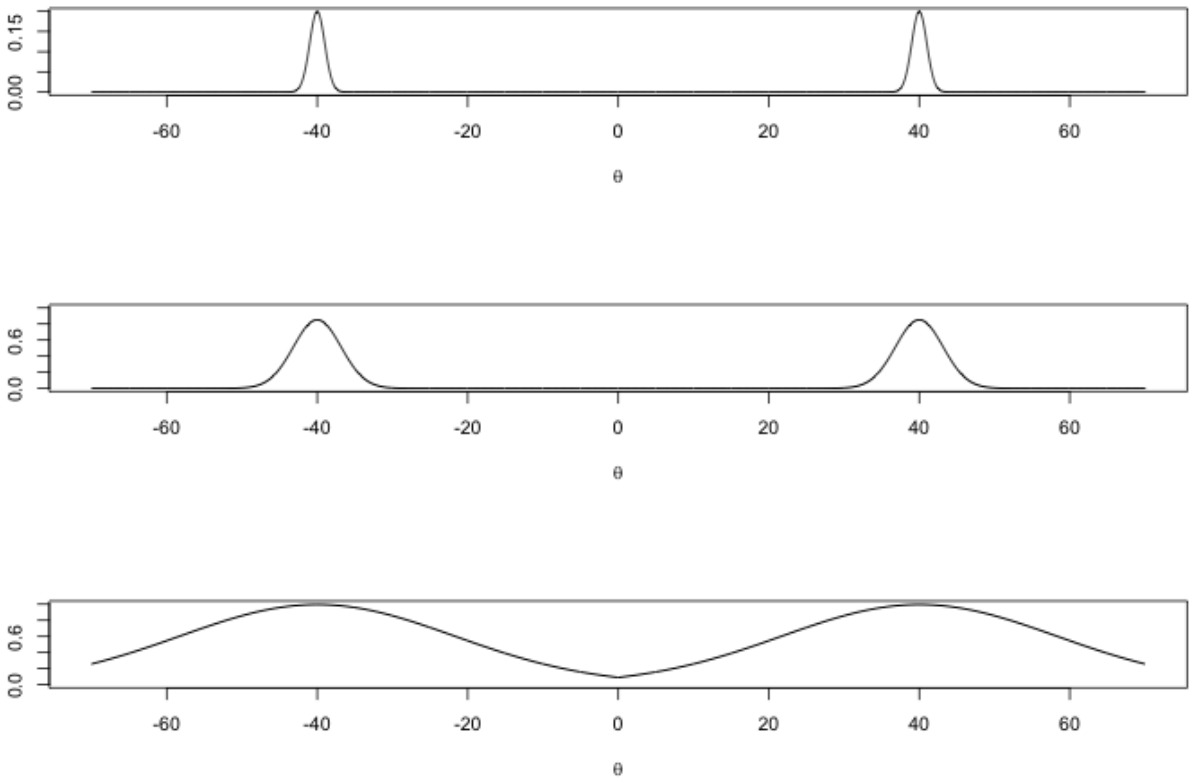
Simulated tempering has been extremely successful on many hard problems, although mostly for relatively low-dimensional distributions.

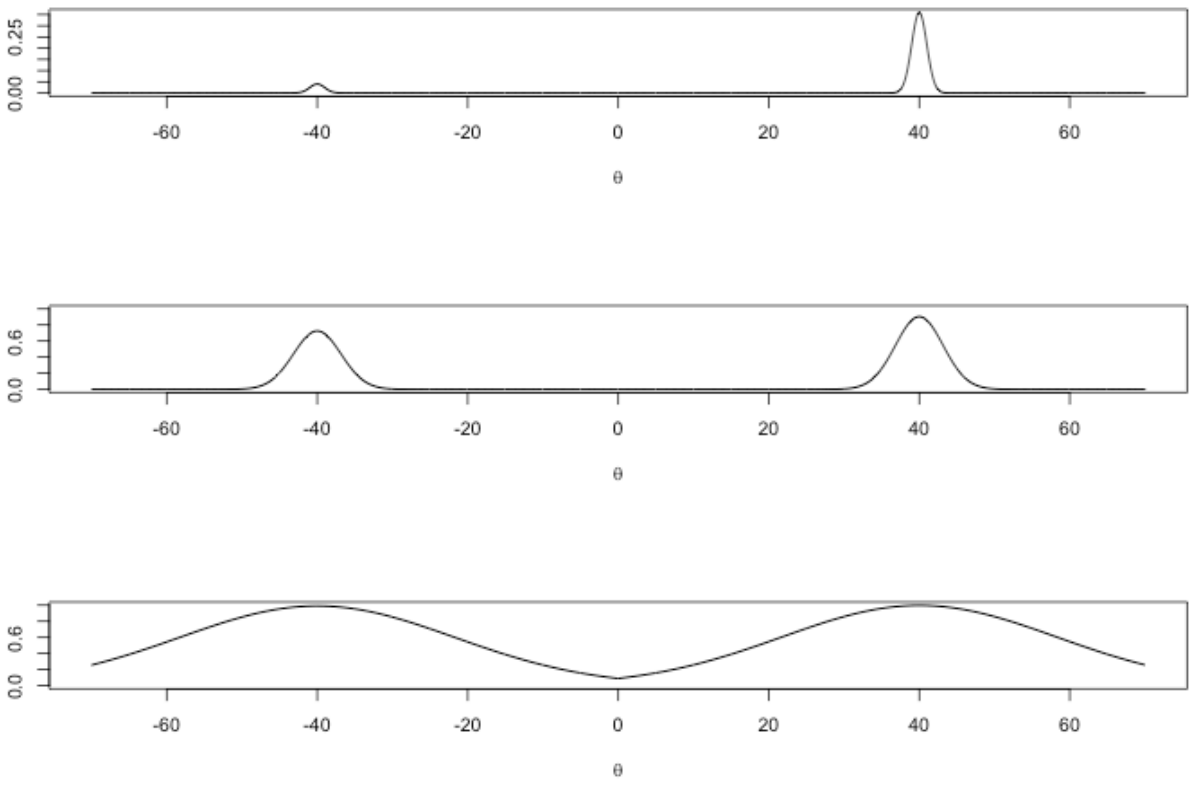
Woodard, Schmidler and Huber (2009) show negative results to show that convergence times are usually exponential in dimension.

WHY?

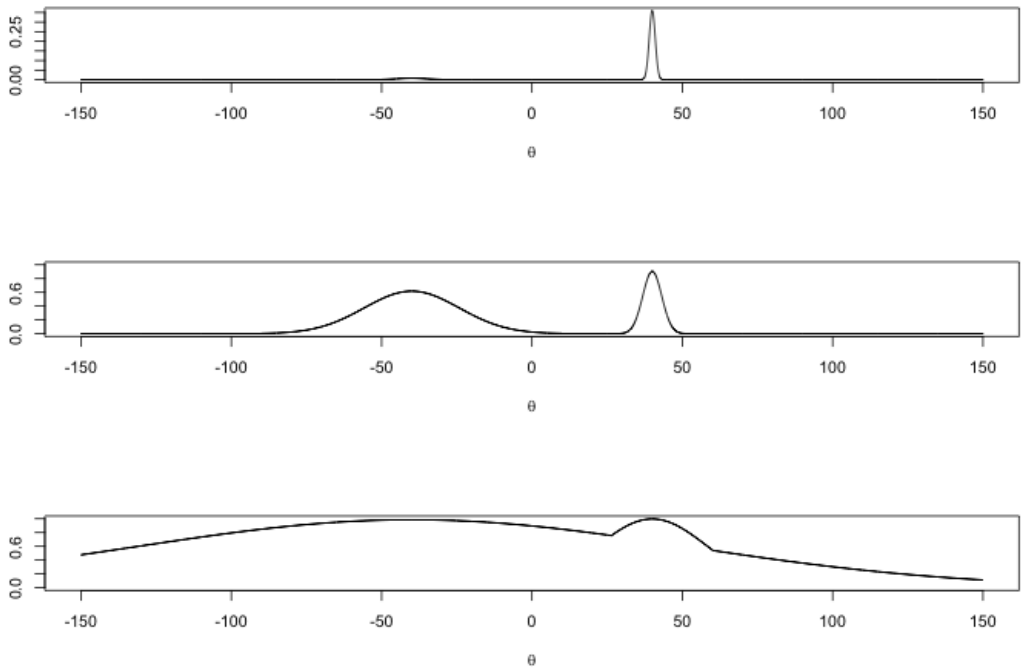
To understand consider some simple examples again.

The symmetric mode case:



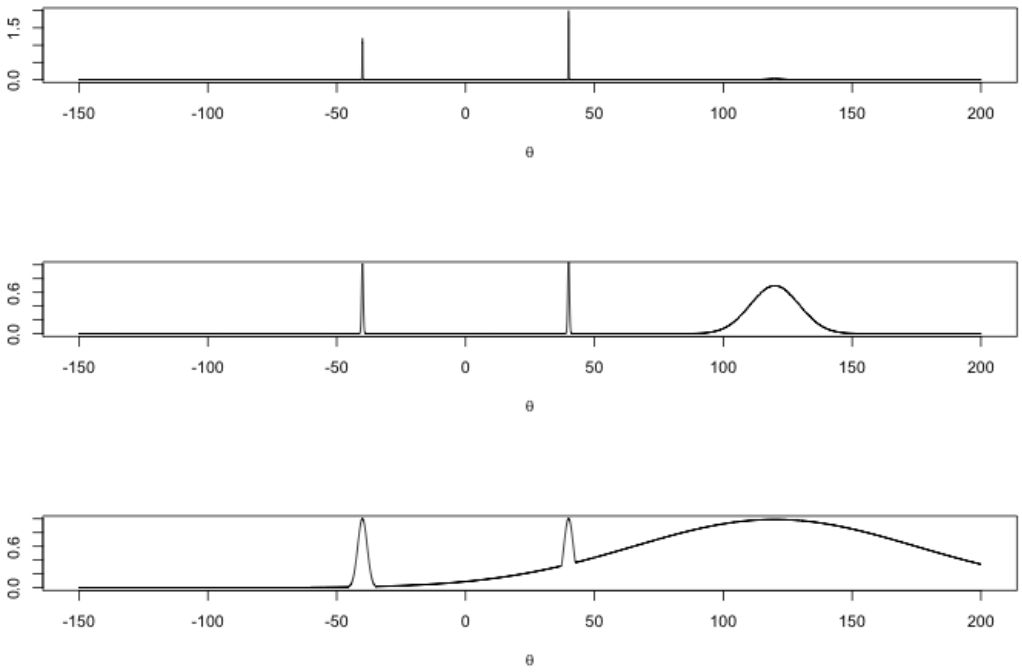

0.00
0.25
-60
-40
-20
0
20
40
60
θ
0.0
0.6
-60
-40
-20
0
20
40
60
θ
0.0
0.6
-60
-40
-20
0
20
40
60
θ

The non-symmetric case:



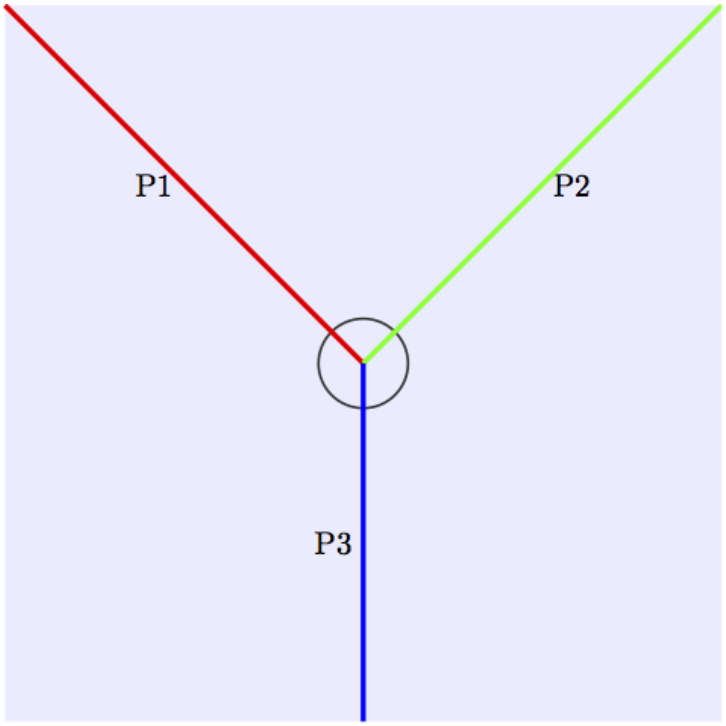
The weight in each mode is not (even approximately) stable to powering.

Another example ...



The problem is exacerbated in high-dimensional contexts.

Understanding how simulated tempering moves between modes



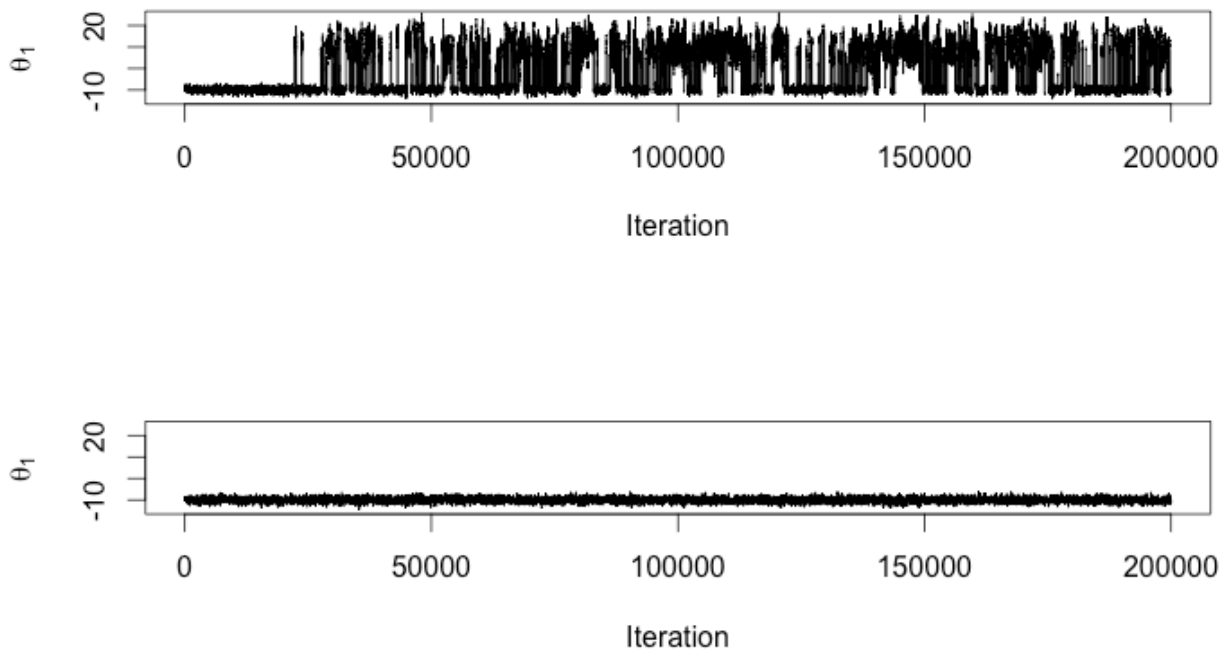
WeSt weight-stable tempering

Ordinary tempering

$$f_\beta(x) \propto f^\beta(x)$$

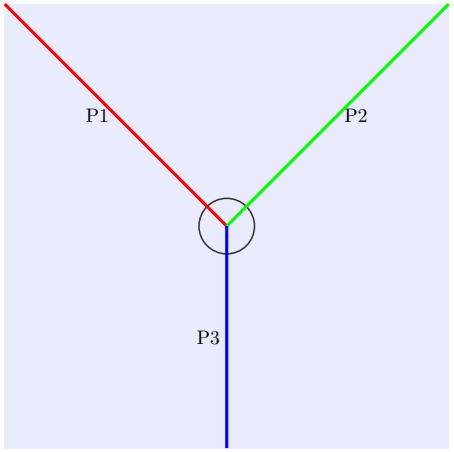
WeSt tempering:

$$f_\beta(x) \propto f(x) \exp\left\{-\frac{1-\beta}{2}(\nabla \log f(x))'(\nabla^2 \log f(x))^{-1}\nabla \log f(x)\right\}$$


θ1
20
-10
0
50000
100000
150000
200000
Iteration
θ1
20
-10
0
50000
100000
150000
200000
Iteration

High-dimensional behaviour

Can show that high-dimensional limit of the algorithm behaves like Walsh Brownian motion (or just skew-Brownian motion when there are 2 modes).



Conclusions

Active area with many directions

- Other algorithms: Hamiltonian Mone Carlo, Delayed Acceptance MCMC, Multiple-Try MCMC, more sophisticated Langevin schemes, pseudo-marginal MCMC ...
- Relaxing smoothness conditions
- Infinite dimensional algorithms
- Multi-scale limits
- methodology: heterogeneity of proposals, adaptive MCMC ...
- ....

Hopefully I have convinced you that this is an interesting area for probability!

An important infinite dimensional case

Target distribution π can be expressed as a change of measure from a Gaussian process on an (infinite-dimensional) Hilbert space $\mathcal{H}$

$$\frac{d\pi}{d\pi_0}(x) = \exp\big(-\Phi(x)\big), \quad \pi_0 \sim \mathcal{N}(0, \mathcal{C}).$$

This arises naturally in many situations.

Calculations generally involve approximation/truncation to some finite-dimensional problem.

big advantage to constructing algorithms in $\mathcal{H}$: robustness to the choice of truncation.

Bayesian density estimation

X is a Gaussian process on $\mathbb{R}^n$ with covariance operator $\mathcal{C}$.

Observations: $\{y_i\}$ from density proportional to e^X giving rise to log-likelihood $-\Phi(X, y)$.

Prior π_0, posterior π.

$$\frac{d\pi}{d\pi_0}(X) \propto \exp\{-\Phi(X, y)\}$$

Discretely observed diffusions

Eg

$$dX_t = dB_t + \alpha_\theta(X_t)dt$$

observations $Y_1, \ldots Y_n$ where $Y_i = X_{t_i}$.

Standard MCMC strategy alternates between updating

- $\alpha | X_{[0,t]}$; and
- $X[0,T]|\mathbf{Y}, \theta$.

The second step involves simulating from a collection of conditionally independent densities

$$\frac{d\pi}{d\pi_0}(x) \propto \exp\left(\int_{t_{i-1}}^{t_i} \alpha_\theta(X_s)dX_s - \frac{1}{2}\int_{t_{i-1}}^{t_i} \alpha_\theta^2(X_s)ds\right),$$

where π_0 denotes the Gaussian measure given by the law of a Brownian bridge conditioned to respect the endpoints prescribed by the data.

Data Assimilation in Fluid Mechanics

- Sample $x \in \mathcal{H} = L^2(\Omega, \mathbb{R}^2)$ initial condition for Navier-Stokes equation:

$$\frac{dX}{dt} + \nu AX + B(X, X) = f,\ X(z, 0) = x(z)$$

- Conditioned on noisy observations $y = \{z_j(t_k)\}$ of

$$y_{jk} = X(z_j, t),\ z_j(0) = z_{j,0} + \varepsilon_{jk}$$

- Given prior π_0, sample $x \in L^2(\Omega, \mathbb{R}^2)$ from posterior μ :

$$\frac{d\pi}{d\pi_0}(x) \propto \exp\Big(-\frac{1}{2}\big|\Sigma^{-\frac{1}{2}}(y - \mathcal{G}(x))\big|^2\Big).$$

Oil Recovery

- Sample permeability $k \in \mathcal{H} = L^2(\Omega, \mathbb{R}^3)$.

$$\nabla_z \cdot \Big(k \nabla_z p \Big) = 0,$$

- Conditioned on indirect observations y of p.
- Let $k(z) = \exp(x(z))$ and sample $x \in L^2(\Omega, \mathbb{R}^3)$ from posterior μ :

$$\frac{d\pi}{d\pi_0}(x) \propto \exp\Big(-\frac{1}{2} \big| \Sigma^{-\frac{1}{2}} (y - \mathcal{G}(x)) \big|^2 \Big).$$

Common Structure

- Change of Measure from Gaussian in $\mathcal{H}$

$$\frac{d\pi}{d\pi_0}(x) = \exp\bigl(-\Phi(x)\bigr), \quad \pi_0 \sim \mathcal{N}(0, \mathcal{C}).$$

- There exist constants $M^{\pm}$ and $k \geq 0$ such that the standard deviations λ_i in π_0 satisfy

$$M^- \leq i^k \lambda_i \leq M^+.$$

- Φ satisfies some kind of smoothness/boundedness conditions, frequently expressed in terms of an appropriate Sobolev norm.

The Karhunen-Loéve expansion

Hilbert space $\mathcal{H}$ containing X, where X is a Hilbert space on which π_0 is supported. The eigenpairs solve the problem

$$\mathcal{C}\varphi_i = \lambda_i^2 \varphi_i, \quad i = 1, 2, \ldots.$$

Let $\{\xi_i\}_{i=1}^{\infty}$ denote an IID $\mathcal{N}(0, \lambda_i^2)$ and

$$x = \sum_{i=1}^{\infty} \xi_i \varphi_i(x). \tag{8}$$

This series converges in $L^2(\Omega; \mathcal{H})$.

Expansion is useful conceptually as well as a guide to algorithm construction.

Improving on Euler-Maruyama

$$X_{t+h} - X_t = \int_t^{t+h} \frac{V\nabla \log \pi(X_s)}{2} ds + MVN(O, hV)$$

so the Euler-Maruyama approximation estimates the integral by its value at the left hand endpoint.

We introduce the (partially) implicit discretisation, which estimates the drift term by

$$(1-\theta)\frac{V\nabla \log \pi(x^{(k)})}{2} + \theta \frac{V\nabla \log \pi(x^{(k+1)})}{2}$$

$\theta = 1$ is called the *fully* implicit case, $\theta = 0$ is the *explicit* or *Euler* discretisation, and $\theta = 0.5$ is the Crank-Nicolson approach.

Partially implicit proposals for MCMC

Partially implicit discretisation schemes are widely known and used in deterministic and stochastic numerical analysis.

Consider implicit proposal on $\mathbb{R}^d$:

$$g(Y) = f(X)$$

where g is one-to-one then Y has density $f(g(y))|J(g^{-1}(y))|$ where J is the appropriate Jacobian matrix of partial derivatives of g.

- Need g to be one-to-one;
- need g^{-1} to be rapidly calculable, if necessary by a stable iterative algorithm.

Well-known in numerical analysis that partially implicit methods can be more stable than explicit ones.

Casella, R + Stramer (2011, MCAP) studies MCMC methods and their properties constructed in this way. A general theory is not available.

The Multivariate Gaussian case

$\pi \sim MVN(0, \Sigma)$. Can solve the implicit equation to give proposal

$$Y = \left(I + \frac{hV\Sigma^{-1}\theta}{2}\right)^{-1}\left(I - \frac{hV\Sigma^{-1}(1-\theta)}{2}\right)X +$$

$$MVN\left(0, \left(I + \frac{hV\Sigma^{-1}\theta}{2}\right)^{-1} hV\left(\left(I + \frac{hV\Sigma^{-1}\theta}{2}\right)^{-1}\right)^{T}\right)$$

This exactly preserves the invariance of π if and only if $\theta = 0.5$ (Crank-Nicolson).

"Error" is $O(h^2)$ for all other θ values in $[0, 1]$.

Optimality in general

- No unique measure of optimality!
- Optimise proposal variance to maximise:

$$M(d) = \mathbb{E}\left\| x^{(k+1)} - x^{(k)} \right\|^2$$

.

- Equivalent to minimising time one correlation.
- Formal justification critically requires diffusion and SPDE limit results.
- Where limiting behaviour is not 'diffusion-like', squared jumping distance criteria are not appropriate.

Optimality in the Gaussian case

(mainly from R + Rosenthal, 2001, Stat Sci)

It seems intuitively clear that the optimal shape for proposals should take $V = \Sigma$. But how bad is it when V and Σ have different shapes?

Let $\{\lambda_i\}$ denote the eigenvalues of $\Sigma^{-1/2}V^{1/2}$, and define

$$R = \frac{\sum_{i=1}^{d} \lambda_i^6/d}{(\sum_{i=1}^{d} \lambda_i/d)^6} \ .$$

Then R gives an inefficiency factor quantifying how much worse it is to use a shape V proposal rather than a shape Σ one.

Complexity

Consider simple Gaussian target: $\pi(x) \propto \exp\{-\sum_{i=1}^{d} x_i^2\}$, and consider best possible choice of proposal scaling h.

- Random walk Metropolis "Error" in proposal is $O(h)$ and cost of dimensionality is $O(d)$.
- (Fully explicit) Langevin "Error" in proposal is $O(h^2)$ and cost of dimensionality is $O(d^{1/3})$.
- Partially implicit Langevin, $\theta \neq 0.5$ "Error" in proposal is $O(h^2)$ and cost of dimensionality is $O(d^{1/3})$.
- Crank-Nicolson Langevin, $\theta = 0.5$ No "Error" in proposal, and no cost of dimensionality.

Moving away from Gaussian ....

Change of Measure from Gaussian in $\mathbb{R}^d$

$$\pi(x) = \exp\Big(-\Phi(x) - \frac{1}{2}\langle x, \Sigma^{-1}x\rangle\Big).$$

Langevin SDE:

$$dX_t = -\frac{V(\nabla\Phi(X_t) - \Sigma^{-1}X_t)}{2}dt + V^{1/2}dt$$

A tractable alternative to the pure implicit method is to be implicit only for the linear part of this SDE:

$$Y = \frac{-Vh\nabla\Phi(x)}{2} - \frac{V\Sigma^{-1}h(\theta Y + (1-\theta)x)}{2}$$

Random walk Metropolis

Consider sampling from

$$\pi \sim N\left(0, \mathcal{C}^{-1}\right)$$

Propose a move to

$$Y = X + N\left(0, h\mathcal{A}^{-1}\right)$$

Turns out that for all possible choices of h and $\mathcal{A}$ has acceptance probability $= 0$ for almost all proposed moves.

The same is true for all explicit Langevin schemes, higher order Langevin, Hybrid Monte Carlo, etc...

Langevin Proposals

- The Langevin SPDE is π_0-invariant:

$$\frac{dx}{dt} = \frac{\mathcal{A}\nabla \log \pi_0(x)}{2} + \sqrt{\mathcal{A}}\frac{dW}{dt},$$

  where dW/dt is space time white noise.

- Here $\mathcal{A}^* = \mathcal{A}$, $\mathcal{A} > 0$.

SPDE proposal

SPDE suggests candidate proposals $x \longrightarrow y$. Suppose we could implement proposals and accept-reject mechanisms without error on H. However lessons from finite dimensions:

- If we use an Euler-Maruyama approximation of SPDE, acceptance probability is identically 0.
- If we use a partially implicit Euler-Maruyama approximation of SPDE with $\theta \neq 1/2$ then acceptance probability is identically 0.
- If we use a partially implicit Euler-Maruyama approximation of SPDE with $\theta = 1/2$, acceptance probability may still be 0 if we do not match up $\mathcal{C}$ with $\mathcal{A}$ sufficiently well.

Discretisation

- Optimise over choice of discretisation, $\mathcal{A}$ and Δt.
- It is usually the case that $M(d) \propto \Delta t_{\text{opt}}$. (Related to diffusion process limits of algorithms in high-dimensions).
- Required number of steps is proportional to $M(d)^{-1}$.

(Recall $M(d) = \mathbb{E}\left\|x^{(k+1)} - x^{(k)}\right\|^2$.)

IID Products

$$\pi_0(x) = \Pi_{i=1}^d f(x_i).$$

$$\textbf{Proposal} \quad \frac{y-x}{\Delta t} = \frac{\beta \nabla \log \pi_0(x)}{2} + \sqrt{\frac{1}{\Delta t}}\xi, \quad \xi \sim \mathcal{N}(0, I).$$

Theorem 17 *(Roberts et al 97, Roberts/Rosenthal 98)*

- $\beta = 0$ *then* $M(d) = \mathcal{O}(d^{-1})$. *(*$\mathbb{E}\alpha = 0.234\ldots$*).*
- $\beta = 1$ *then* $M(d) = \mathcal{O}(d^{-1/3})$. *(*$\mathbb{E}\alpha = 0.574\ldots$*).*

Scaled Product

$$\pi(x) = \pi_0(x) = \Pi_{i=1}^{d} \frac{1}{\lambda_i} f\left(\frac{x_i}{\lambda_i}\right).$$

$$M^- \leq i^k \lambda_i \leq M^+.$$

Proposal $$\frac{y-x}{\Delta t} = \frac{\mathcal{A}\nabla \log \pi_0(x)}{2} + \sqrt{\frac{\mathcal{A}}{\Delta t}}\xi, \quad \xi \sim \mathcal{N}(0, I).$$

Theorem 18 • $\mathcal{A} = I$ *then* $M(d) = \mathcal{O}(d^{-(2k+1/3)})$. *(*$\mathbb{E}\alpha = 0.574\ldots$*).*

• $\mathcal{A} = \mathcal{C}$ *then* $M(d) = \mathcal{O}(d^{-1/3})$. *(*$\mathbb{E}\alpha = 0.574\ldots$*).*

Change of Measure

$$\pi(x) = \exp\Big(-\Phi_d(x)\Big)\pi_0(x) = \exp\Big(-\Phi_d(x)\Big)\Pi_{i=1}^d \frac{1}{\lambda_i} f\Big(\frac{x_i}{\lambda_i}\Big).$$

Proposal $\quad \dfrac{y-x}{\Delta t} = \dfrac{\mathcal{A}\nabla \log \pi_0(x)}{2} + \sqrt{\dfrac{\mathcal{A}}{\Delta t}}\xi, \quad \xi \sim \mathcal{N}(0, I).$

Theorem 19 • $\mathcal{A} = I$ *then* $M(d) = \mathcal{O}(d^{-(2k+1/3)})$. *(*$\mathbb{E}\alpha = 0.574\ldots$*).*

• $\mathcal{A} = \mathcal{C}$ *then* $M(d) = \mathcal{O}(d^{-1/3})$. *(*$\mathbb{E}\alpha = 0.574\ldots$*).*

Change of Measure Does Not Affect Optimality

Change of Measure from Gaussian

$$\pi(x) = \exp\Big(-\Phi_n(x) - \frac{1}{2}\langle x, \mathcal{C}^{-1}x\rangle\Big).$$

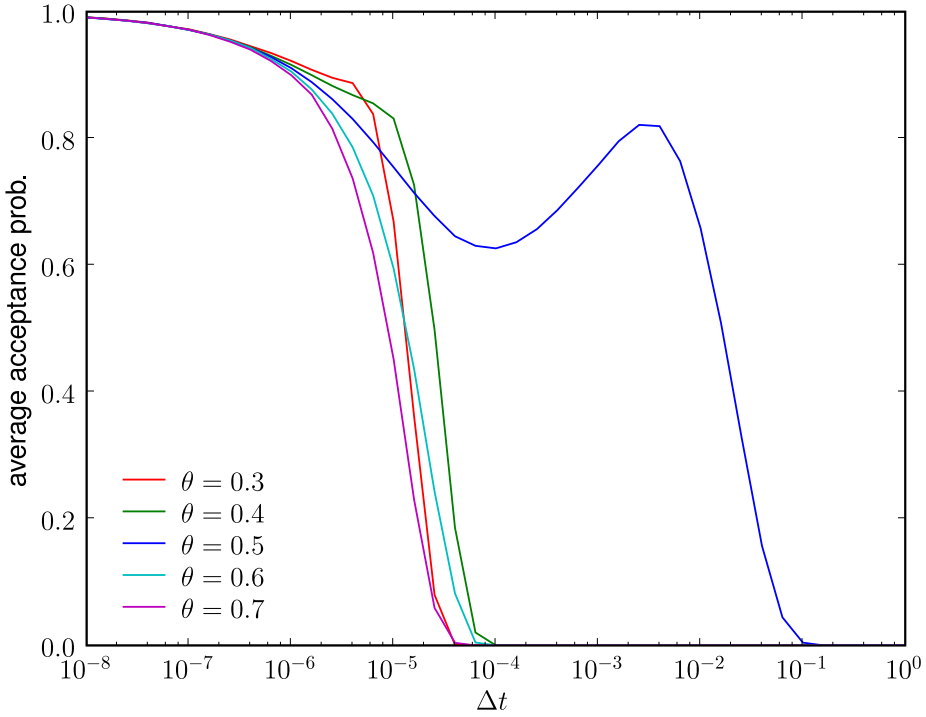
$$\frac{y-x}{\Delta t} + \frac{\mathcal{A}\Big(\theta\mathcal{C}^{-1}y + (1-\theta)\mathcal{C}^{-1}x\Big)}{2} = \sqrt{\frac{\mathcal{A}}{\Delta t}}\xi, \quad \xi \sim \mathcal{N}(0, I).$$

Theorem 20 • $\theta \neq \frac{1}{2}$ *and* $\mathcal{A} = \mathcal{C}$ *then* $M(d) = \mathcal{O}(d^{-1/3})$. *(*$\mathbb{E}\alpha = 0.574\ldots$*).*

• $\theta = \frac{1}{2}$ *and* $\mathcal{A} = I, \mathcal{C}$ *then* $M(n) = \mathcal{O}(1)$. *(*$\mathbb{E}\alpha$ *not identified).*

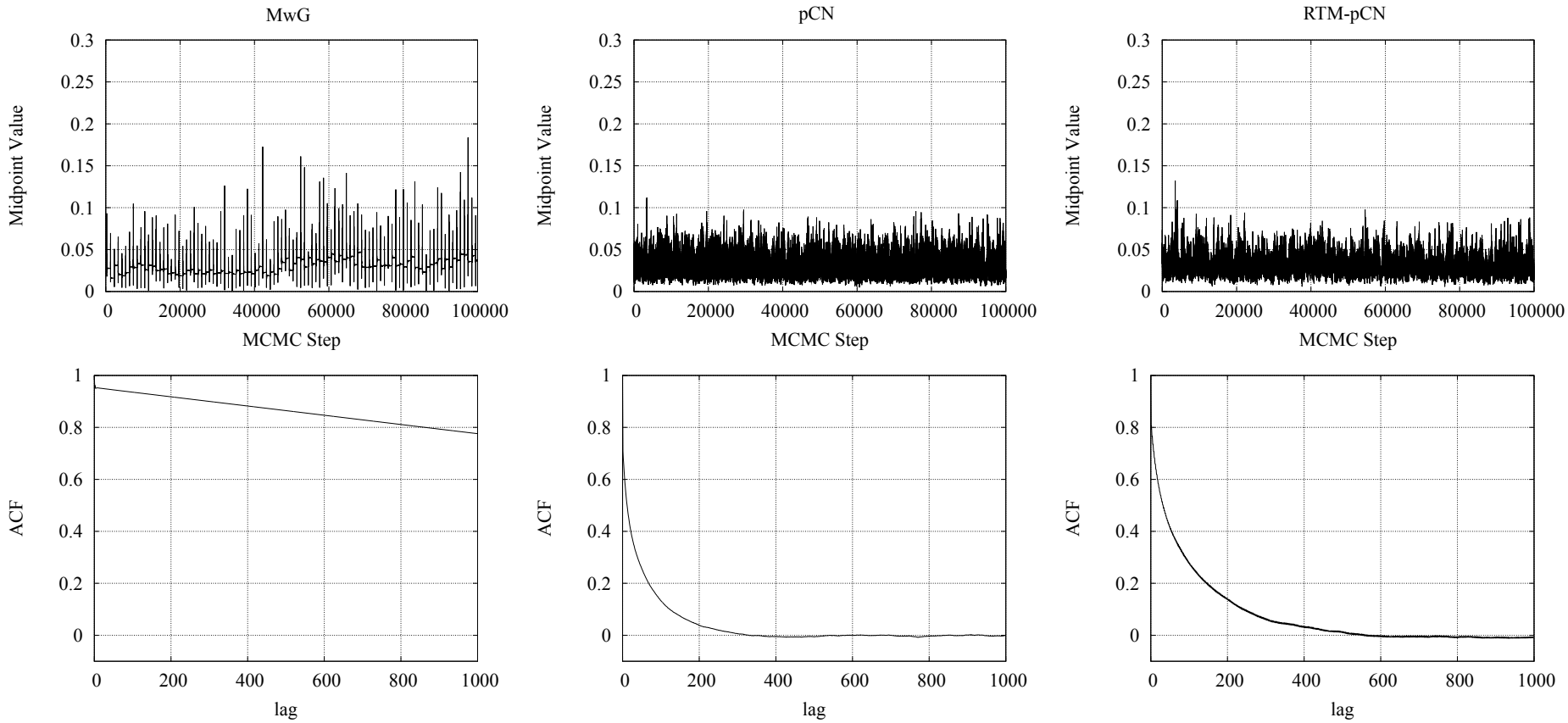
Implicitness Impacts Optimality

Example: Diffusion bridge Sampling



Average acceptance probability in equilibrium for the signal processing problem ($d = 1000$).

Comparison between RWM and pCN



Summary

We have shown that:

- **Applications:** Measures which have density with respect to a Gaussian arise naturally in many applications where the solution is a measure on functions.
- **Algorithms:** Optimised algorithms for the dominating Gaussian measure.
- **SPDEs:** Langevin SPDEs form natural basis for MCMC propoals.
- **Algorithms:** Using these SPDEs, MCMC methods can be constructed in function space.
- **Numerical Analysis:** Ideas such as steepest descents, preconditioning and implicitness have crucial impact on complexity of MCMC algorithms, corresponding to the multivariate proposal scaling problem well known in statistical MCMC.

Ongoing and future things ..

- **Priors:** More flexible families of priors, eg *sieve* priors can have theoretical and algorithmic advantages.
- **Applications:** are numerous in physics, data assimilation, signal processing and econometrics. Much needs to be done - hence EQUIP project (led by Andrew Stuart in Warwick).
- Robustification of algorithms for large datasets. Understanding the interplay between discrete approximation in the parameter space and size of data set.
- Working with exact algorithms on the Hilbert space using retrospective simulation methods. May be possible, might be computationally efficient ...

Other things going on in scaling

- Most results need smoothness conditions on the target. Need more precise results on relationship between smoothness and mixing rates (and algorithm limits) (Durmus, LeCorff, Moulines). (0.13)
- Hamiltonian MCMC (Hybrid Monte Carlo) (0.651), pseudo-marginal MCMC (0.07).
- Scaling in different ways in different parts of the space more generally.
- Integration into adaptive schemes.

Hopefully I have convinced you that this is an interesting area for probability!

THE END!

Thank you for listening

References

- M. Bédard, *Weak Convergence of Metropolis Algorithms for Non-iid Target Distributions*. Ann. Appl. Probab. **17**(2007), 1222-44.
- M. Bédard and J.S. Rosenthal, *Optimal Scaling of Metropolis Algorithms: Heading Towards General Target Distributions*. Can. J. Stat., 36, 4, 483–503, 2008.
- A. Beskos, G.O. Roberts, A.M. Stuart and J. Voss. "An MCMC Method for diffusion bridges." Stochastics and Dynamics, 8, 3, 319–350, 2008.
- A. Beskos, G.O. Roberts and A.M. Stuart. "Optimal scalings for local Metropolis-Hastings chains on non-product targets in high dimensions." Annals of Applied Probability, 19, 3, 863–898, 2009.
- B.Casella, GO Roberts and O. Stramer. Stability of partially implicit Langevin schemes and their MCMC variants, Methodology and Computing in Applied Probability, 13, 4, 835–854, 2011.

References (Continued)

- S Cotter, AM Stuart, GO Roberts and D White. MCMC methods for functions: modifying old algorithms to make them faster to appear in *Statistical Science*, 2013.
- A. Gelman, W.R. Gilks and G.O. Roberts, *Weak convergence and optimal scaling of random walk Metropolis algorithms*. Ann. Appl. Prob. **7**(1997), 110–120.
- M. Hairer, A.M.Stuart, P. Wiberg and J. Voss. "Analysis of SPDEs Arising in Path Sampling. Part 1: The Gaussian Case." Comm. Math. Sci. 3(2005), 587–603
- M. Hairer, A.M.Stuart and J. Voss. "Sampling the posterior: an approach to non-Gaussian data assimilation." Physica D, **230**(2007), 50–64.
- M. Hairer, A.M.Stuart and J. Voss. "Analysis of SPDEs Arising in Path Sampling. Part 2: The Nonlinear Case." Ann. Appl. Prob. 17(2007), 1657–1706.

References (Continued)

- A.M.Stuart, P. Wiberg and J. Voss. "Conditional Path Sampling of SDEs and the Langevin MCMC Method." Comm. Math. Sci. 2(2004), 685–697.
- NS Pillai, AM Stuart, and AN Thiery. Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. Annals of Applied Probability, 22, 2320-2356, 2012.
- G.O. Roberts and J. Rosenthal, *Optimal scaling of discrete approximations to Langevin diffusions*. JRSSB **60**(1998), 255–268.
- GO Roberts and JS Rosenthal. Optimal scaling of various Metropolis-Hastings algorithms, Statistical Science, 16, 4, 351–367, 2001.
- GO Roberts and O Stramer. Bayesian inference for incomplete observations of diffusion processes, Biometrika, 88, 603–221, 2001.