Centre for Research in Statistical Methodology

`http://go.warwick.ac.uk/crism`

- Conferences and workshops (**including general calls for workshops to be organised primarily outside Warwick, calls every 6 months, next in June 2013**)
- Research Fellow positions
- PhD studentships
- Academic visitor programme.

i-like.org.uk

A 5 year project funded through a *Programme Grant* from EPSRC.

Led by me from Warwick. Also involving Bristol, Lancaster, Oxford. Collaborative project Principal Investigators Gareth Roberts, David Firth, Christophe Andrieu, Paul Fearnhead, Chris Holmes.

5 postocs directly funded under grant, plus many more researchers infomally connected with the research activity.

Some recent advance in scaling in MCMC

Gareth Roberts

University of Warwick

For Oxwaspers, October 2014

including work mainly with Jeffrey Rosenthal, Chris Sherlock, Alex Beskos, Pete Neal and Alex Thiery

Plan

1. The beginning of the scaling story
2. Why 0.234?
3. Some recent problems
   (a) divergence of target scale
   (b) spacing of temperatures in simulated tempering
4. Conclusions

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

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

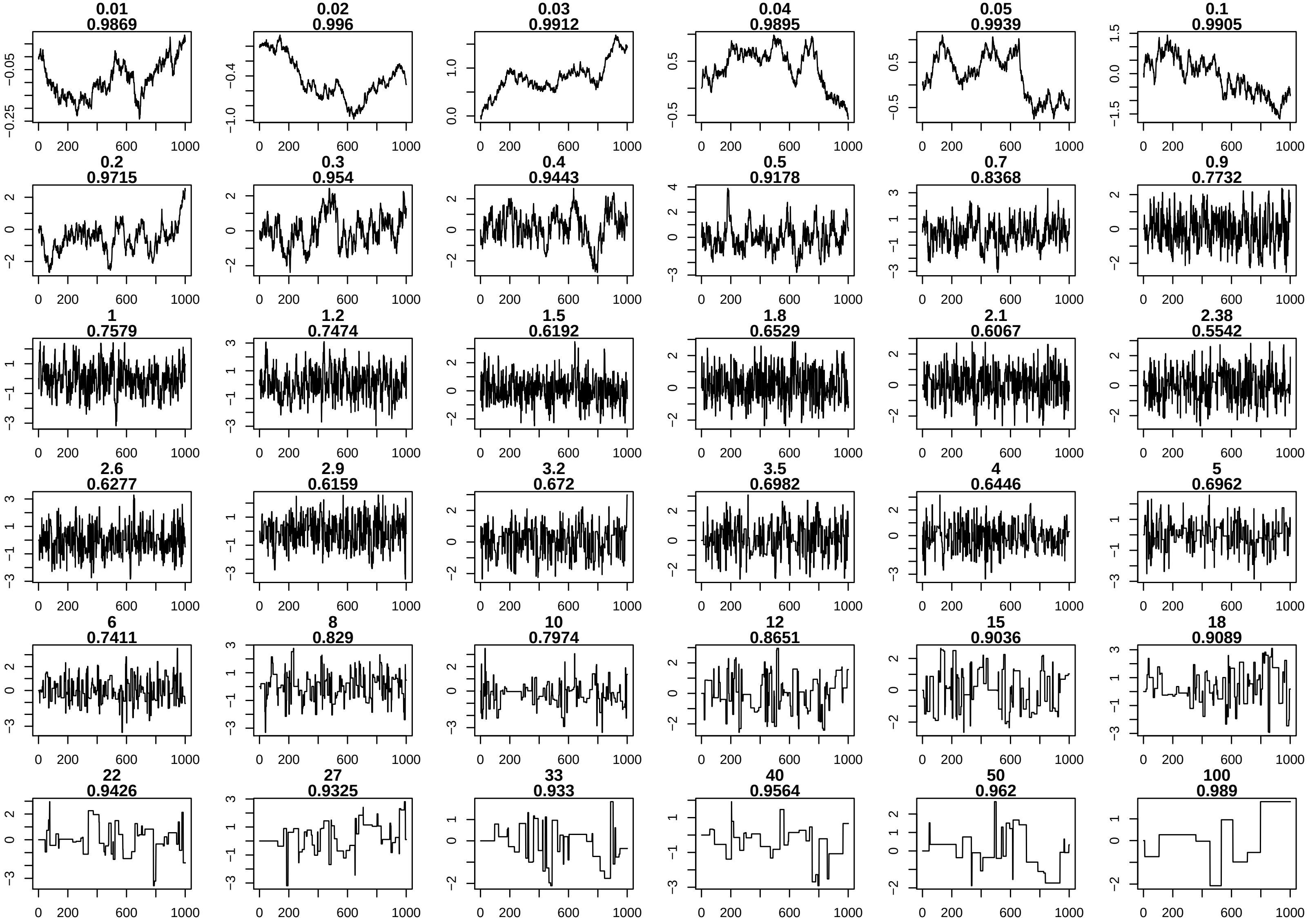
  The acceptance probability simplifies to

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

  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) 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, and suppose it is possible to construct MCMC on $f_{d,\beta} = f_d^{\beta}$, $0 \leq \chi \leq \beta \leq 1$.

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

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

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?

It is well-established that estimating limiting variance is hard.

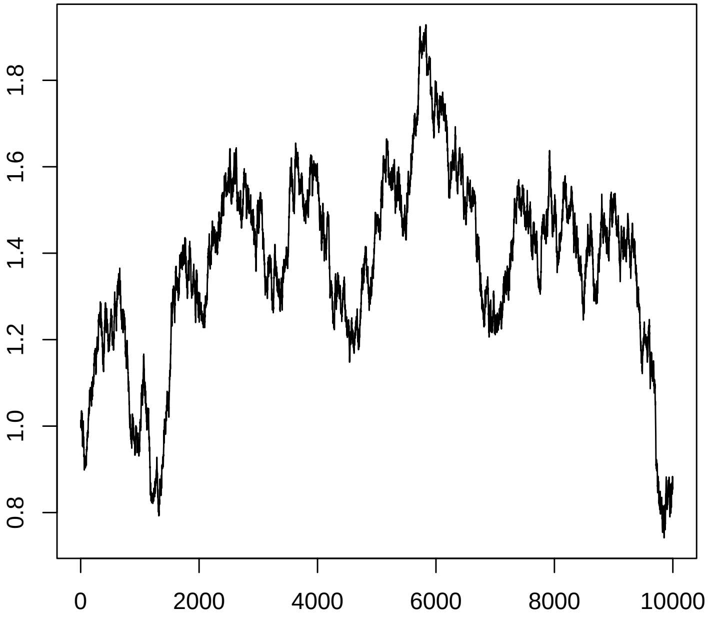
“It’s easy, just measure ESJD instead!” Andrew Gelman, 1993

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

Why? “It’s obvious!” Andrew Gelman 2011

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

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

Diffusions

A d-dimensional diffusion is a continuous-time strong Markov process with continuous sample paths. We can define a diffusion as the solution of the Stochastic Differential Equation (SDE):

$$\mathrm{d}X_t = \mu(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}B_t.$$

where B denotes d-dimensional Brownian motion, σ is a $d \times d$ matrix and μ is a d-vector.

Often understood intuitively and constructively via its dynamics over small time intervals. Approximately for small h:

$$X_{t+h}|X_t = x_t \quad \sim \quad x_t + h\mu(x_t) + h^{1/2}\sigma(x_t)Z$$

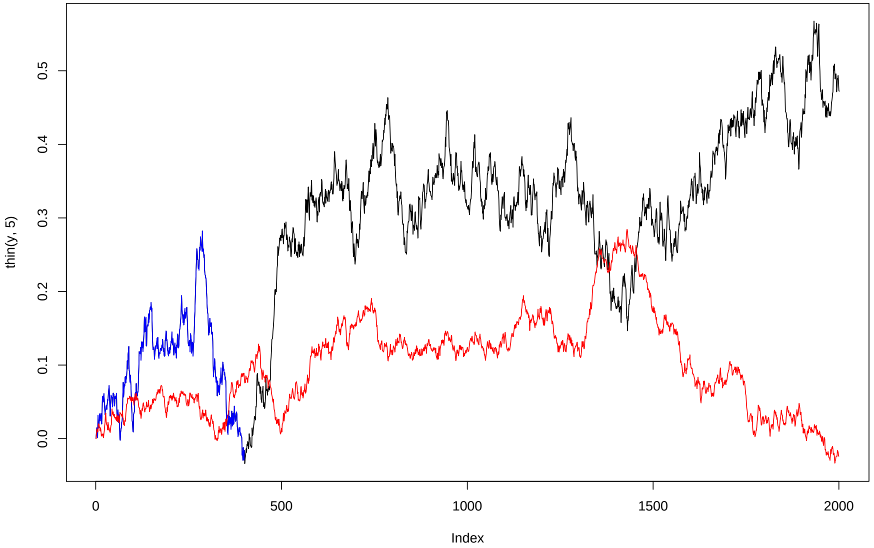
where Z is a d-dimensional standard normal random variable.

“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 $h_1(x) \le h_2(x)$ 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:

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

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

Consider the Metropolis case.

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

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

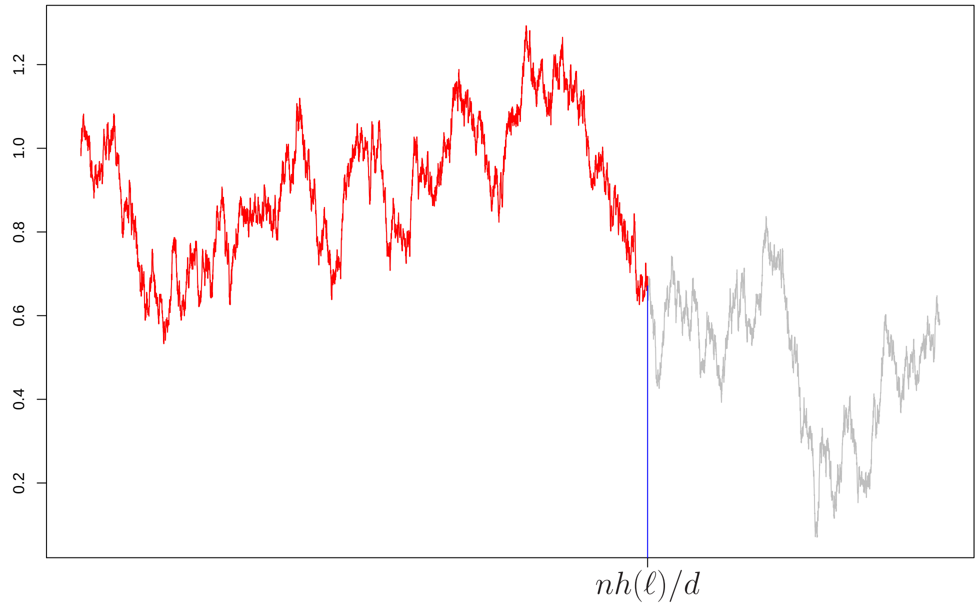
Z^d is **not** a Markov chain, however in the limit as d goes to ∞, it is Markov:

$$Z_d \Rightarrow Z$$

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?

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

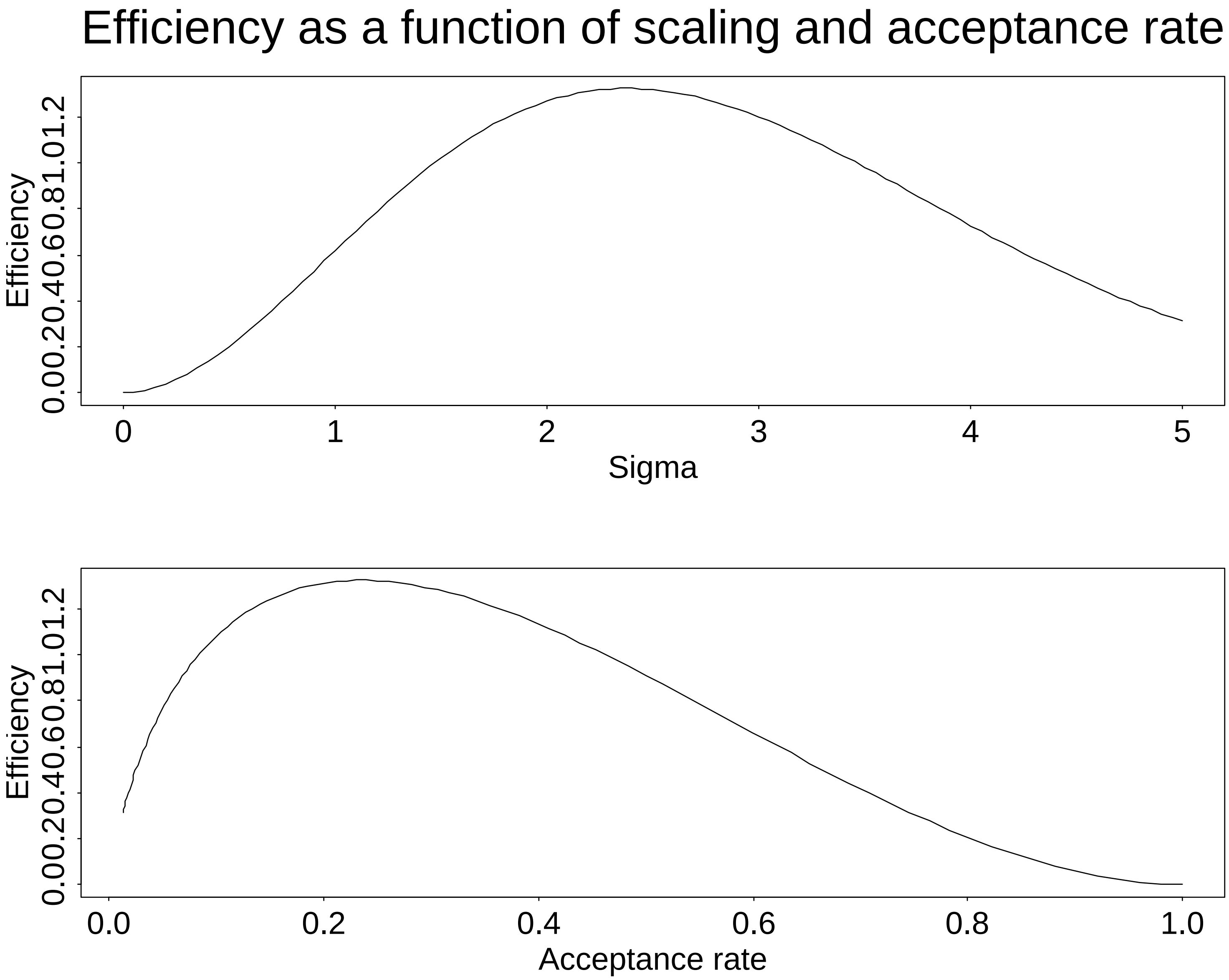
and $I = E_f[((\log f(X))')^2]$. So

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

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
Efficiency
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
1
2
3
4
5
Sigma
Efficiency
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.2
0.4
0.6
0.8
1.0
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. $\pi \sim \prod_{i=1}^{d} f(c_i x_i)$, $q(\mathbf{x}, \cdot) \sim N(\mathbf{x}, \sigma_d^2 I_d)$. for some inverse 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.

A basic analysis of Metropolis

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

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

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

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

G is Gaussian and if $V^{(d)}$ converges in probability to constants, then the variance of G is V.

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:

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

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

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

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

Metropolis-within-Gibbs

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

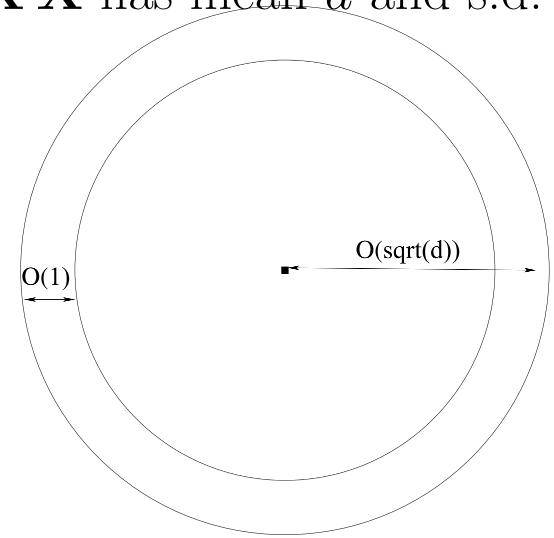
How should be jointly choose (c_d, σ^2) to optimise the Markov chain?

(Neal and R, 2008) Suppose $c_d \to c \in (0, 1]$, then we get the same optimal efficiency curve,independently of c.

So 0.234 still holds, and all values of c are equally efficient!

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. $(2d)^{1/2}$



Target distribution lies concentrated around the surface of a d-dimensional hypersphere.

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?

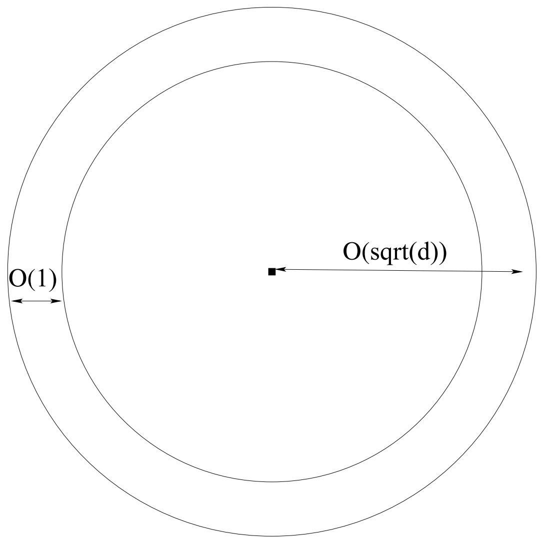
Spherical symmetry (Sherlock and R, 2009, Bernoulli)

Theorem 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

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

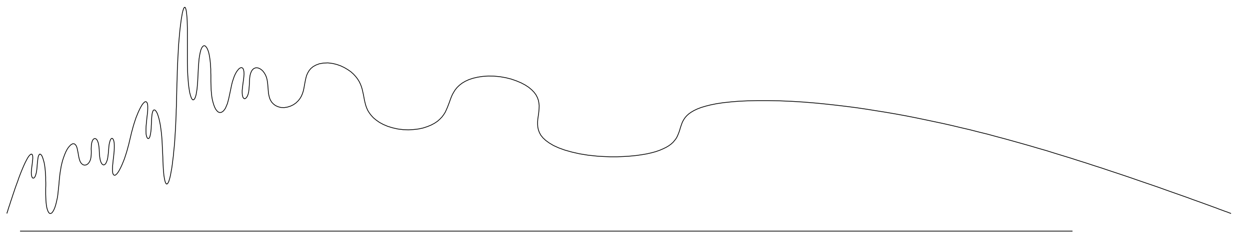
with $\alpha_\infty = 0.234$ if and only if R equals some fixed positive constant with probability 1.

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.

Does this happen in other situations?



Eccentricity

Theorem Suppose we can write $\mathbf{X}^{(d)} = T_d\mathbf{Z}^{(d)}$ for matrices $\{T_d\}$ each having collections of 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

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:

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

where a_0 denotes the acceptance probability of moves around x:

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

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 ..

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 $\{y_n^{(d)}\}$ 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 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.

Other things going on in scaling

- Most results need smoothness conditions on the target. What happens for discontinuous densities? (0.13)
- Results for MALA (0.574), ‘Metropolis within Gibbs’ (0.234), ‘Langevin within Gibbs’ (0.574, though optimal to do full updates), Hamiltonian MCMC (Hybrid Monte Carlo) (0.651), pseudo-marginal MCMC (0.07).
- What happens to algorithms started away from stationarity? (Christensen, Rosenthal and recent paper by Jourdain, Lelievre and Miasojedow).
- What happens if we use heavy-tailed proposals? (not 0.234 unless proposal has 2nd moments plus a little more)

Other things going on in scaling (continued)

- Scaling problems in function space (Pillai, Stuart, Thiery, Hairer....)
- Convergence of entire d-dimensional processes to infinite dimensional limit, often described by an SPDE.
- What about multivariate scaling problems?
- What about scaling in different ways in different parts of the space more generally?
- Integration into adaptive schemes.

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

$$x = \sum_{i=1}^{\infty} \xi_i \varphi_i(x). \tag{5}$$

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 proposal on $\mathbb{R}^d$: $g(Y)$ has d-dimensional density f 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 partialderivatives 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!
- Optimize proposal variance to maximize:

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

.

- Equivalent to minimizing 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

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

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

- Optimize over choice of discretization, $\mathcal{A}$ and Δt.
- It is usually the case that $M(d) \propto \Delta t_{\mathrm{opt}}$. (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

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

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

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

Theorem 2

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

Theorem 3

- $\mathcal{A} = I$ 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

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

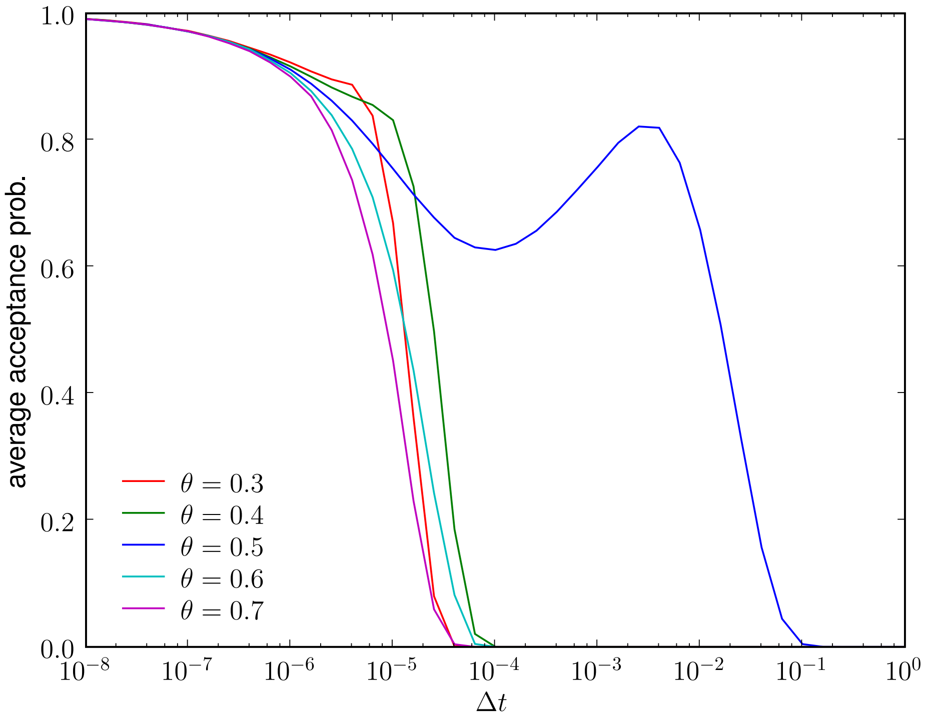
$$\frac{y-x}{\Delta t} + \frac{\mathcal{A}\Big(\theta\mathcal{C}^{-1}y + (1-\theta)\mathcal{C}^{-1}x\Big)}{2} = \sqrt{\frac{\mathcal{A}}{\Delta t}}\xi, \quad \xi \sim \mathcal{N}(0, I).$$

Theorem 4

- $\theta \neq \frac{1}{2}$ 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).

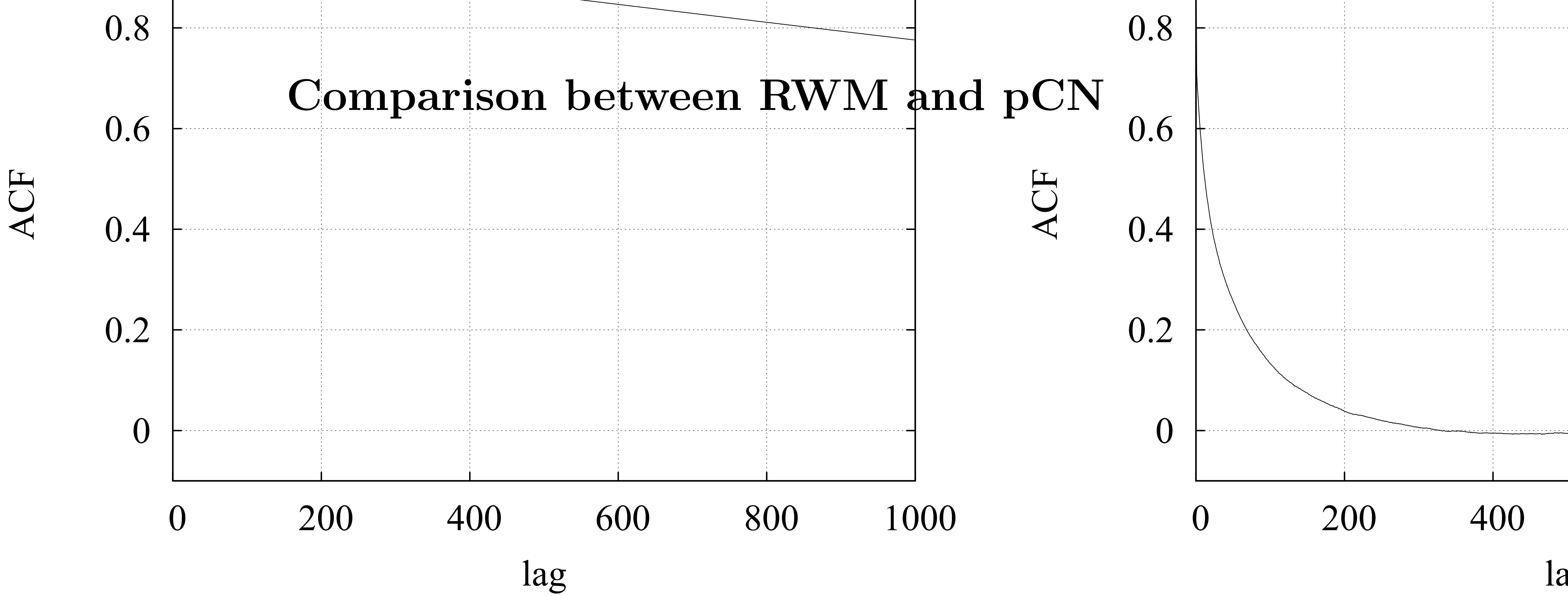
Implicitness Impacts Optimality

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!
- 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 ...

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.