Enabling Quantification of

EQUIP

Uncertainty for Inverse Problems

When to jump off the edge of a cliff

Gareth Roberts

University of Warwick

Equip Launch, October 2013

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

Optimal scaling for MCMC and heterogeneity

Statistical inverse problems,typically ill-posed, and characterised by non-identifiability.

In a Bayesian context, this manifests itself by posterior distributions with very different scales.

Scale separation gets worse rather than better with more data.

Identifiable components are typically complex, highly non-linear, and in practice unknown.

So, often very difficult to define clever algorithms to exploit specific structure (though see Mark's talk!).

How does generic MCMC fair when we have scale separation?

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

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

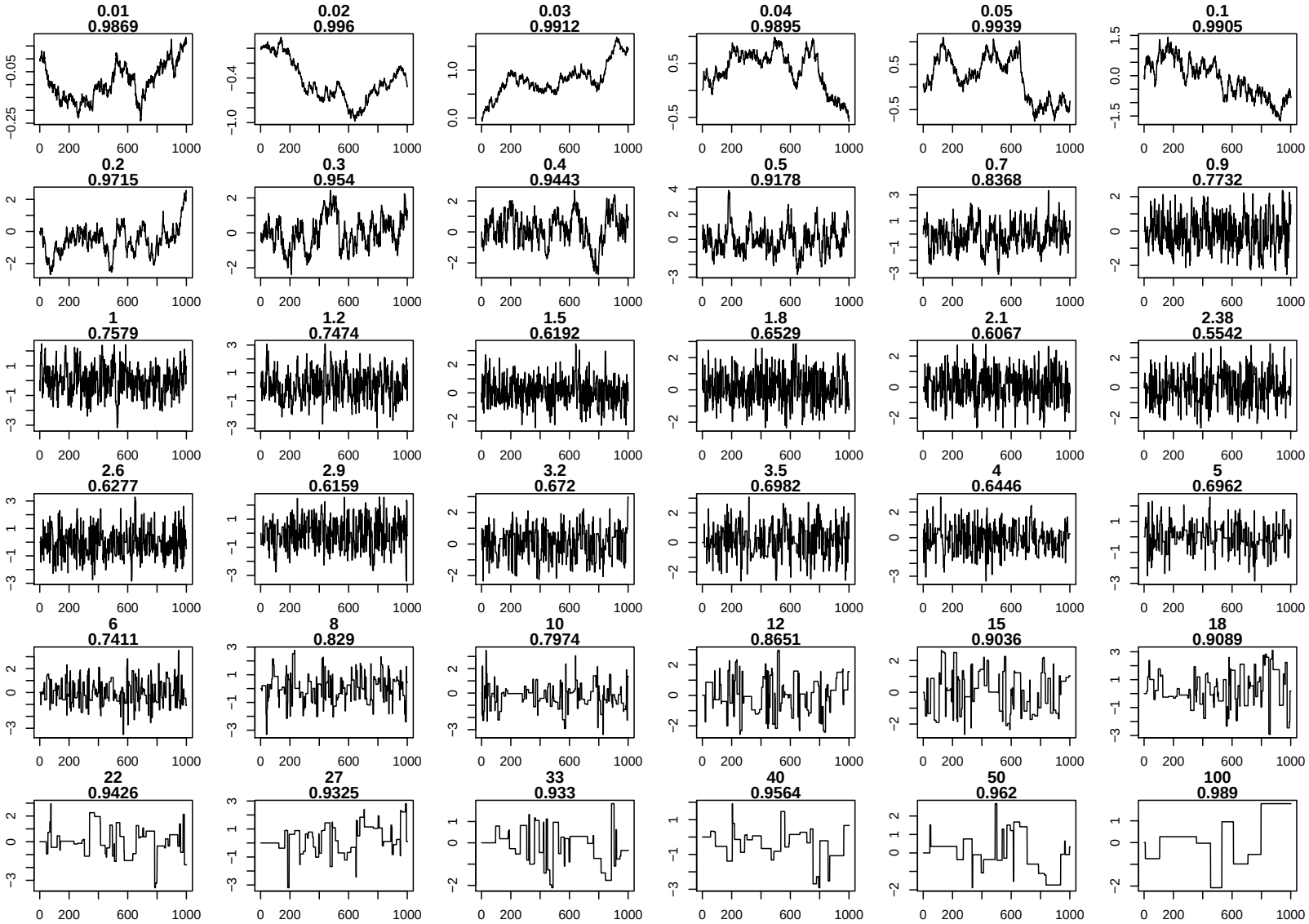
The acceptance probability simplifies to

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

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?
- How is this story affected by heterogeneity of scale?

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).

What is "efficiency"?

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

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

Under weak(ish) regularity conditions

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

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.

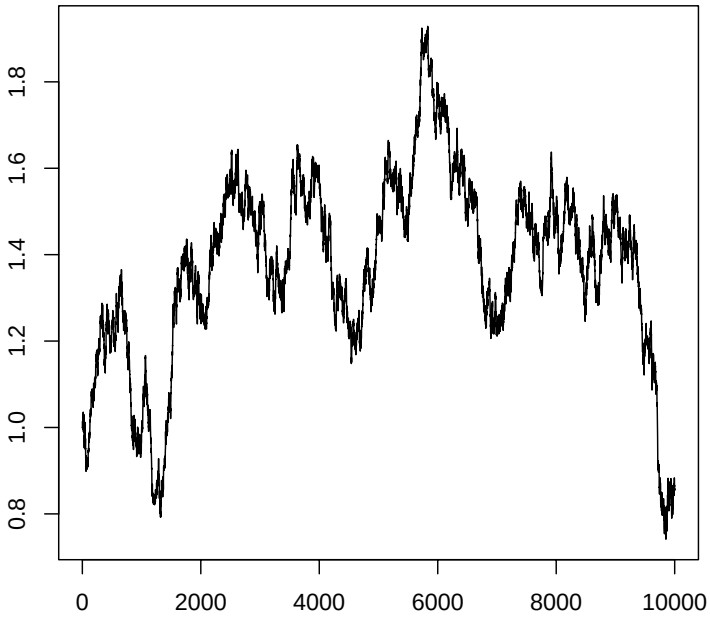
"It's easy, just measure ESJD instead!" Andrew Gelman, 1993

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

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:

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

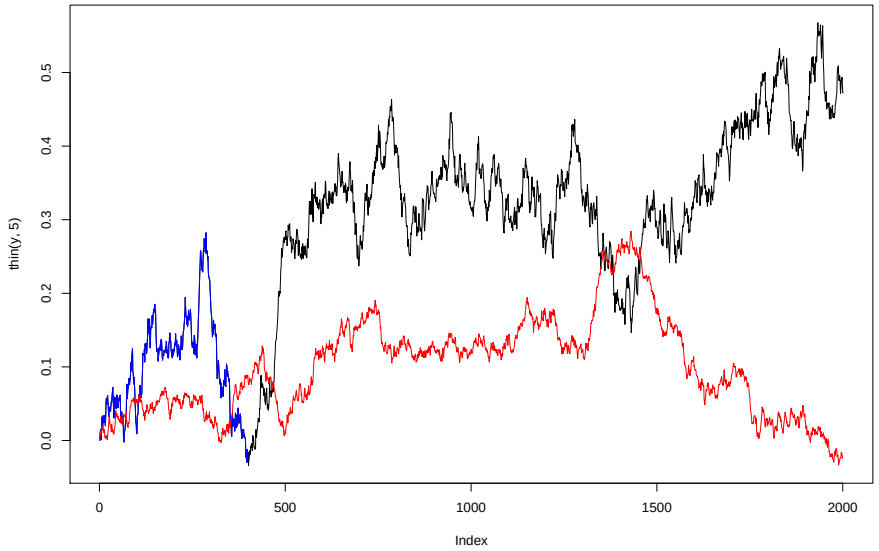
where Z is a d-dimensional standard normal random variable.

"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.

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

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

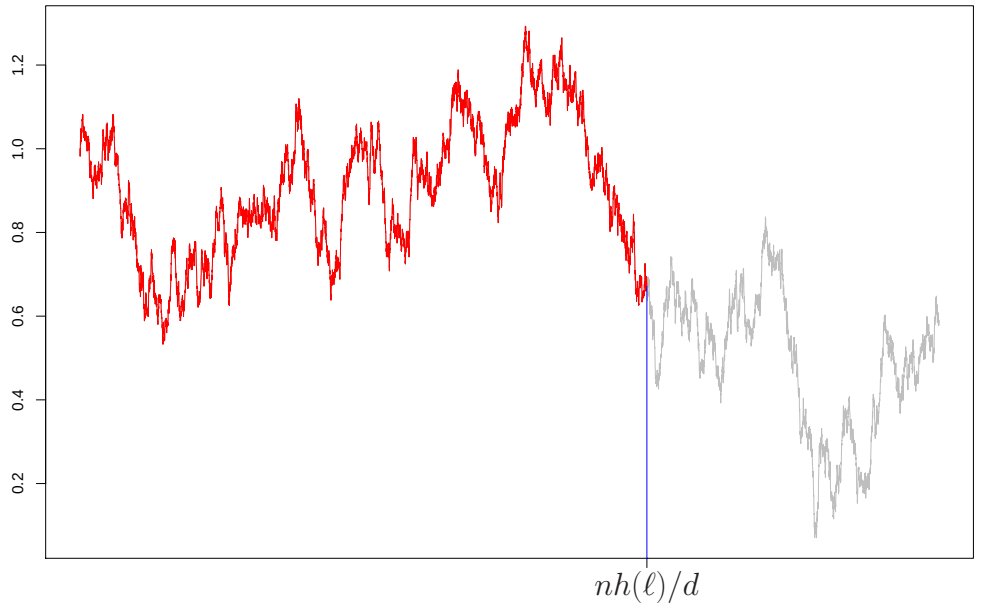
Z^d is **not** a Markov chain, however in the limit as d goes to ∞, it is Markov:

$$Z_d \Rightarrow Z$$

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

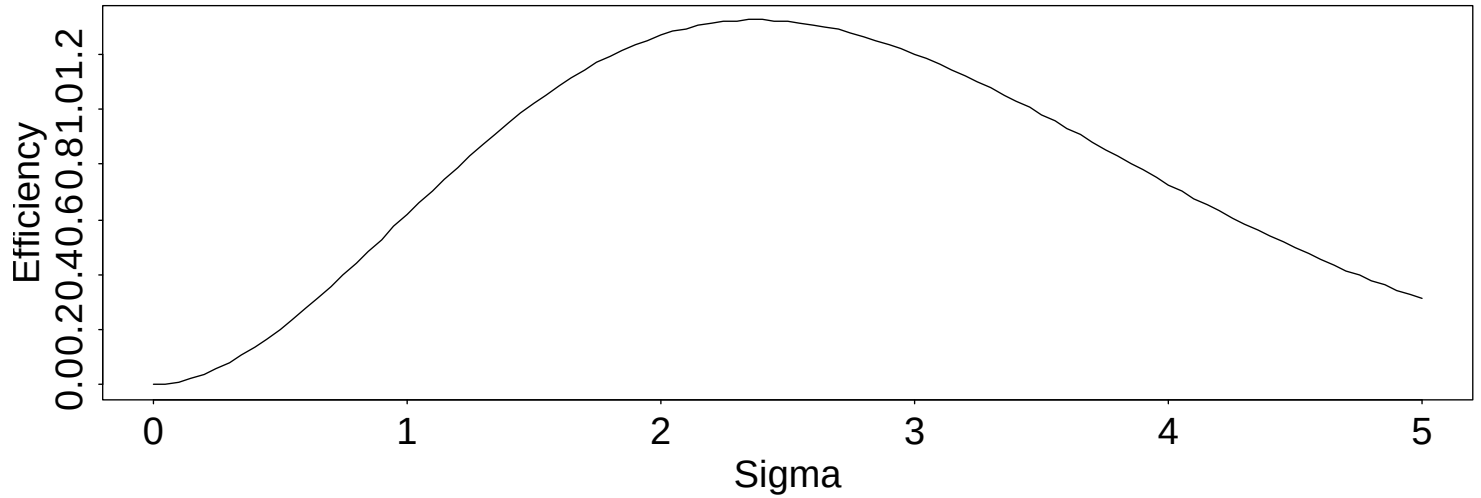
$$h(\ell) = \ell^2 \times A(\ell) \ ,$$

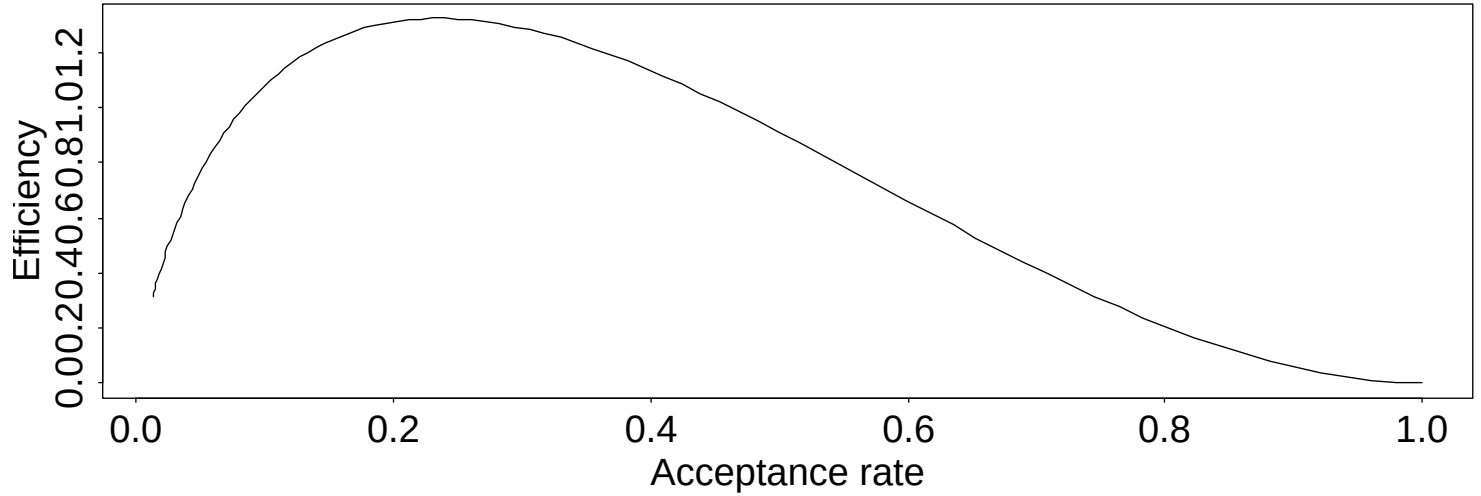
where $A(\ell)$ is the limiting overall acceptance rate of the algorithm, ie the proportion of proposed Metropolis moves ultimately accepted. So

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

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





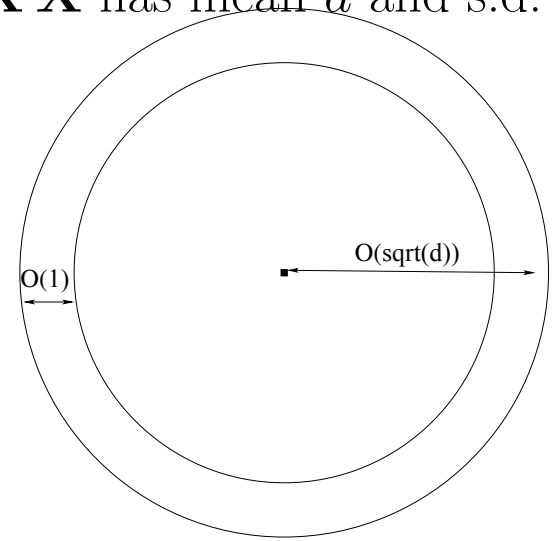
When can we 'solve' the scaling problem for Metropolis?

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.

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 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?

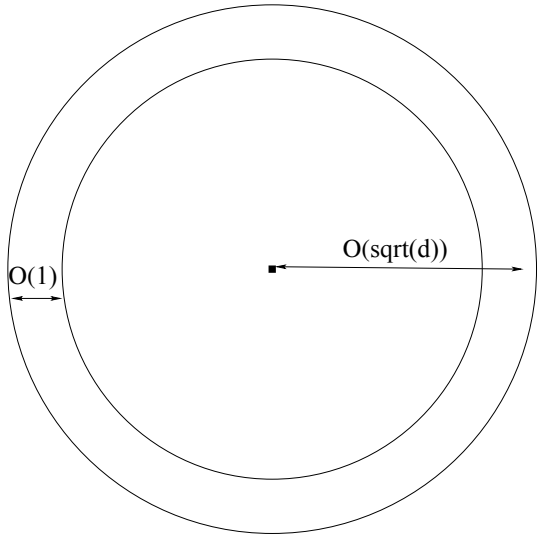
Spherical symmetry (Sherlock and R, 2009, Bernoulli)

Theorem 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

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

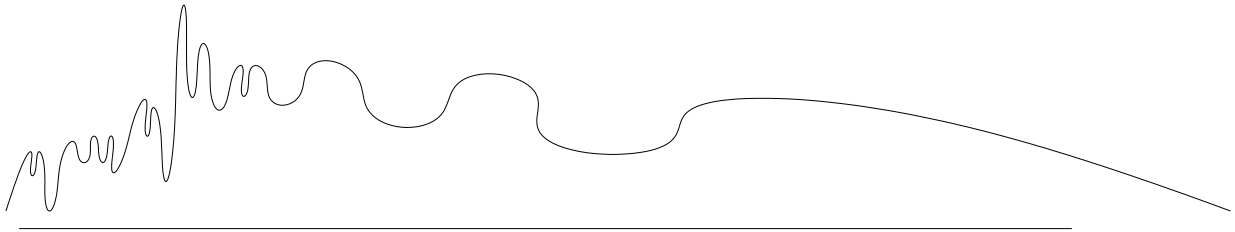
with $\alpha_\infty = 0.234$ if and only if R equals some fixed positive constant with probability 1.

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.

Can happen in many other situations.



What should be we do when we have heterogenous roughness?

Ideally choose a state-dependent proposal distribution.

Expressions for the cost of herogeneity can be obtained in some situations:

- Scale of different components drawn from some non-trivial distribution R + Rosenthal 2001.
- Scales of component i scales like i^{-k} for some $k > 0$. (Beskos, R. Stuart and Voss.)

All these results look at high-dimensional limits.

What can we do in finite-dimensions?

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.

Scaling with diverging scales

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:

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

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 ..

How does this fit with Equip?

In many ill-posed inverse problems, separation of scale is typically non-linear and difficult or impossible to really know about.

The above very simple result needs to be substantially generalised to understand the penalty intrinsic to the ill-posedness of the problem.

Gradient based MCMC algorithms are very promising (cue Mark!) though often have instabilities which are particularly sensitive to heterogeneity of scale. Need tounderstand this further!

Ideally need theory to underpin a robust MCMC methodology to cope with non-identifiability issues.