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 Summer 2014)
- Research Fellow positions: next advertising 2 positions around February 2014
- PhD studentships
- Academic visitor programme.
From Peskun Ordering to Optimal Simulated Tempering

Gareth Roberts
University of Warwick
MCMSki, Chamonix, January 2014

mainly joint work with Jeffrey Rosenthal, but with aspects of joint work with Yves Atchade
Plan for talk

1. Why 0.234 is natural in many problems
2. Comparisons of algorithms based on their diffusion limits, links to Peskun ordering
3. A heterogenous scaling problem: spacing of temperatures in simulated tempering
4. Local 0.234 story for simulated tempering
5. 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**

  \[ q(x, y) = q(|y - x|) \]

  The acceptance probability simplifies to

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

  For example \( q \sim MVN_d(x, \sigma^2 I_d) \), 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 $\sigma$ 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 $\sigma_d$ depend on $d$ for large $d$?
• What does this tell us about the efficiency of the algorithm?
• Can we optimise $\sigma_d$ in some sensible way?
• Can we characterise optimal (or close to optimal) values of $\sigma_d$ 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).
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 $\beta$. 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 $\mathbb{R}^d \times B$ which “optimally” permits the exploration of $f_{d,1}$.

This is also a scaling problem: choosing how large to make $\beta_i - \beta_{i-1}$ for each $i$. 
What is “efficiency”? 

Let $X$ be a Markov chain. Then for a $\pi$-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

\[ ESJD = \mathbb{E}((X_{t+1} - X_t)^2) \]

Why is this a good idea?

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

\[ \sum_{i=1}^{\infty} \text{Cov}_{\pi}(g(X_0), g(X_i)) \]
Diffusion limits: a framework for studying algorithm optimality

Many MCMC algorithms are well-approximated by diffusions (usually in the sense of diffusion limit results for high-dimensional algorithms, or other limiting regimes).

Examples include Random walk Metropolis, various versions of Langevin algorithm, Gibbs samplers simulated tempering.

Provides a natural framework for studying algorithm complexity and optimisation.

But why is the comparison of diffusion limits any easier than comparing the finite-dimensional algorithms?
MCMC sample paths and diffusions.

Here ESJM is the quadratic variation

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

$$dX_t = \mu(X_t)dt + \sigma(X_t)dB_t.$$  

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

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

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

where $Z$ is a $d$-dimensional standard normal random variable.
“Efficiency” for diffusions

Consider two Langevin diffusions, both with $\pi$ invariant. for $h_1 < h_2$,

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

$X^2$ is a “speeded-up” version of $X^1$.

But how can we compare diffusions which have non-constant diffusion coefficient?
Peskun ordering

Peskun (1973). Many uses in MCMC theory (eg see work by Mira, Tierney and co-workers). $P_1$ and $P_2$ are two Markov chain kernels with invariant distribution $\pi$. We say $P_2$ dominates $P_1$ in the Peskun sense, and write $P_2 \succeq P_1$ if for all $x$, and sets $A$ not containing $x$

$$P_1(x, A) \leq P_2(x, A).$$

Peskun ordering implies an ordering on asymptotic variances of ergodic estimates:

$$\lim_{n \to \infty} \left( n \text{Var} \left( \frac{\sum_{i=1}^{n} g(X_{i}^{(1)})}{n} \right) \right) \geq \lim_{n \to \infty} \left( n \text{Var} \left( \frac{\sum_{i=1}^{n} g(X_{i}^{(2)})}{n} \right) \right)$$

where $X^{(i)}$ moves according to $P_i$.

Surprisingly many applications in MCMC!

But are proposal scaling problems completely incompatible with Peskun??
Peskun ordering in continuous time

Peskun ordering of $P_1$ and $P_2$ does not imply, nor is implied by ordering of the 2 step transition kernels.

How about continuous time? (See work of Leisen + Mira 2008.)

But diffusions satisfy $P(x, \{x\}) = 0$ so there can never be any interesting Peskun orderings in the original sense.

HOWEVER MANY discrete time processes possess the SAME diffusion limit. Maybe we can consider the limit along sequences of chains which do satisfy Peskun ordering.
A more powerful diffusion comparison result

Consider two Langevin diffusions, both with stationary distribution $\pi$.

$$dX^i_t = h_i(X^i_t)^{1/2}dB_t + V_i(X^i_t)dt, \quad i = 1, 2,$$

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

Under regularity conditions on the tails of $\pi$ (which have to decay exponentially, or $\pi$ needs to have bounded support)

Then $X^2$ dominates $X^1$ in covariance ordering sense:

$$\lim_{t \to \infty} t \text{Var} \left( \frac{\int_0^t g(X^1_s)ds}{t} \right) \geq \lim_{t \to \infty} t \text{Var} \left( \frac{\int_0^t g(X^2_s)ds}{t} \right)$$
Diffusion comparison result (ctd)

Proof by finding suitable Peskun ordered birth and death processes with the given diffusion limit.

A dominated convergence argument is needed to extend the covariance ordering to the limiting diffusions.

Can extend argument to give approximate covariance ordering for ANY processes with these respective limits.
The first diffusion comparison result (R Gelman Gilks, 1997)

Consider the Metropolis case.

Suppose \( \pi \sim \prod_{i=1}^{d} f(x_i) \), \( q(x, \cdot) \sim N(x, \sigma_d^2 I_d) \), \( X_0 \sim \pi \).

Set \( \sigma_d^2 = \ell^2 / d \). Consider

\[
Z^d_t = X^{(1)}_{[td]}.
\]

Speed up time by factor \( d \)

\( Z^d \) is not a Markov chain, however in the limit as \( d \) goes to \( \infty \), 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 \( I = Ef[((\log f(X))')^2] \). 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

- Sigma Efficiency
- Acceptance rate Efficiency

![Graph of 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 $\pi_d$ 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, (Peter Neal).)
2. $\pi \sim \prod_{i=1}^{d} f(c_i x_i), \ q(x, \cdot) \sim N(x, \sigma_d^2 I_d)$. 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. $\pi$ 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 $Y^{(d)} = X^{(d)} + h^{1/2}Z^{(d)}$.

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

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

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

$G$ 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 $\ell$ incorporates scaling choice:

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

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

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

which is maximised by taking $A(\ell) = 0.234$. 
**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 $f : \mathbb{R} \to [0, \infty)$, 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 $\ell : [0, 1] \to \mathbb{R}$.

To stop adding new temperature values, we fix some $\chi \in (0, 1)$ and keep going until the inverse temperatures drop below $\chi$, 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 $\ell$.

We shall consider a joint process $(y^{(d)}_n, X_n)$, with $X_n \in \mathbb{R}^d$, and with $y^{(d)}_n$ taking values in $\{\beta^{(d)}_i; 0 \leq i \leq k(d)\}$ defined as follows.

Choose $X_{n-1} \sim f^\beta$, then proposing $Z_n$ to be $\beta_{i+1}$ or $\beta_{i-1}$ 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 $(y^{(d)}_n, X_n)$ 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 $\{Y_t\}_{t \geq 0}$ satisfying

$$dY_t = \left[2\ell^2(Y_t)\Phi \left(\frac{-\ell(Y(t))I^{1/2}(Y_t)}{2}\right)\right]^{1/2} dB_t$$

$$+ \left[\ell(Y)\ell'(Y)\Phi \left(\frac{-I^{1/2}(Y_t)\ell}{2}\right) - \ell^2 \left(\frac{\ell I^{1/2}(Y_t)}{2}\right) \phi \left(\frac{-I^{1/2}(Y_t)\ell}{2}\right)\right] dt,$$

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

**Corollary** The speed of this diffusion is maximised, and the asymptotic variance of all $L^2$ functionals is minimised, when the $\ell(y)$ is chosen so that the asymptotic temperature acceptance probability at each and every inverse temperature $y$ is equal to 0.234.
**Final comments**

The simulated tempering temperature spacing problem can be considered as an example of a *state-dependent proposal variance* problem.

Peskun ordering can be used in a quite general way for high-dimensional processes which exhibit limiting diffusion behavior (very many do!).

Interesting extensions of this work will consider whether the temperature scaling problem really interacts with the *within-temperature* mixing of the algorithm.

The 0.234 story is very natural in many situations. It is closely linked to the question of a concentration of measure phenomenon in high-dimensional problems.

0.234 can be a local phenomenon, not just an average over the whole state space.

There are lots of things happening in scaling problems for MCMC (eg my session tomorrow) but still many rich and interesting scaling problems out there.
Conditions for diffusion comparison result

Suppose that the following hold: (A1) \( \pi \) is log-Lipschitz, i.e. for all \( x, y \), there is \( L < \infty \) with

\[ |\log \pi(y) - \log \pi(x)| \leq L |y - x| \]

(A2) Either (a) the support of \( \pi \) is a bounded interval \([a, b]\), and the diffusions \( X^\sigma \) have reflecting boundaries at \( a \) and \( b \); or (b) the support of \( \pi \) is \( \mathbb{R} \), and \( \pi \) has exponentially-bounded tails, i.e. there is \( 0 < K < \infty \) and \( r > 0 \) such that

\[ \pi(x + y) \leq \pi(x)e^{-ry}, \quad x > K, \ y > 0 \]

and

\[ \pi(x - y) \leq \pi(x)e^{-ry}, \quad x < -K, \ y > 0. \]