Some recent advances in MCMC methods

Gareth Roberts, University of Warwick

OBayes 2019, Warwick

WARWICK
THE UNIVERSITY OF WARWICK

CRiSM

i-like.org.uk

MCMC school report

MCMC has been the workhorse of Bayesian Statistics for 30 years now. How is it doing?

GOOD

- It is still the gold standard.
- Huge number of innovative MCMC algorithms have been devised to capitalise on the flexibility of the Metropolis-Hastings framework.
- Major advances in theory to underpin the use of MCMC on Bayesian problems
- Lots of excellent software is enabling clever algorithms to be used by non-computational statistics specialists.

COULD DO BETTER

- Computationally slow in comparison to many approximate (but often very accurate) methods.
- The curse of dimensionality still afflicts many methods.
- Struggles for problems where the target density function is either extremely expensive to compute (eg big data, or completely intractable (for either computational or data confidentiality reasons).

COULD DO BETTER

- Computationally slow in comparison to many approximate (but often very accurate) methods.
- The curse of dimensionality still afflicts many methods.
- Struggles for problems where the target density function is either extremely expensive to compute (eg big data, or completely intractable (for either computational or data confidentiality reasons).

So major emphasis in MCMC methodological research is on methodology which

- is scalable (scaling well with dimension, data size and sometimes other);

COULD DO BETTER

- Computationally slow in comparison to many approximate (but often very accurate) methods.
- The curse of dimensionality still afflicts many methods.
- Struggles for problems where the target density function is either extremely expensive to compute (eg big data, or completely intractable (for either computational or data confidentiality reasons).

So major emphasis in MCMC methodological research is on methodology which

- is scalable (scaling well with dimension, data size and sometimes other);
- robust, for example to different models, data sets, etc;

COULD DO BETTER

- Computationally slow in comparison to many approximate (but often very accurate) methods.
- The curse of dimensionality still afflicts many methods.
- Struggles for problems where the target density function is either extremely expensive to compute (eg big data, or completely intractable (for either computational or data confidentiality reasons).

So major emphasis in MCMC methodological research is on methodology which

- is scalable (scaling well with dimension, data size and sometimes other);
- robust, for example to different models, data sets, etc;
- has associated software enabling its accessibility to a much wider audience than MCMC experts.

Talk outline

- Introduce PDMP, including the zig-zag and the bouncy particle sampler
- Super-efficiency and other theoretical results.
- Bayesian fusion
- The RESTORE algorithm

Talk outline

- Introduce PDMP, including the zig-zag and the bouncy particle sampler
- Super-efficiency and other theoretical results.
- Bayesian fusion
- The RESTORE algorithm

Super-Efficiency:

$$\frac{\text{computational cost of running algorithm}}{\text{cost of one single likelihood evaluation}} \longrightarrow 0$$

in the big data asymptotic.

Piecewise-deterministic Markov processes

Continuous time stochastic process, denote by Z_t.

The dynamics of the PDP involves random events, with deterministic dynamics between events and possibly random transitions at events.

(i) The deterministic dynamics. eg specified through an ODE

$$\frac{\mathrm{d}z_t}{\mathrm{d}t} = \Phi(z_t), \tag{1}$$

So

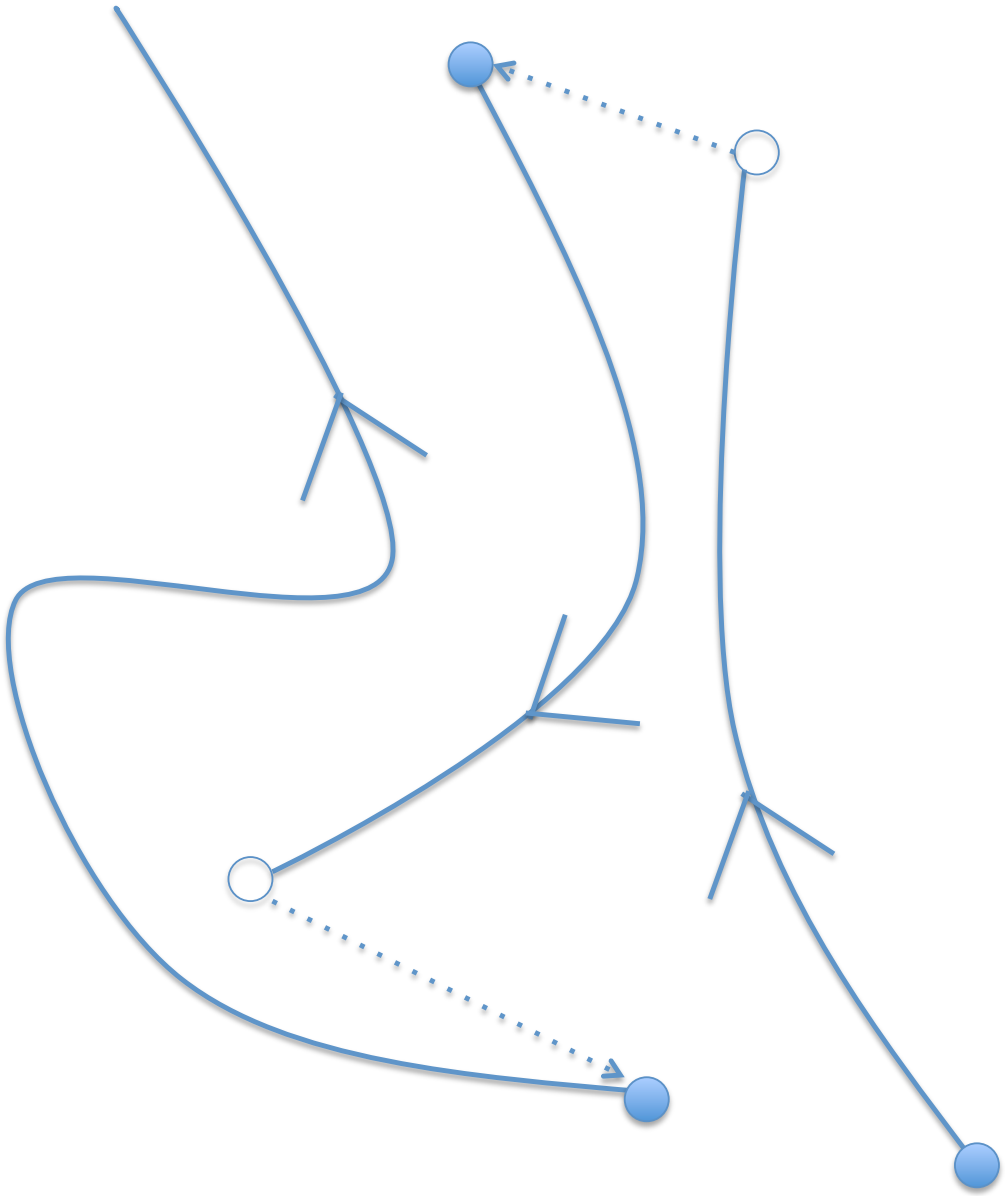
$$z_{s+t} = \Psi(z_t, s)$$

for some function Ψ.

(ii) The event rate. Events occur at rate, $\lambda(z_t)$,

(iii) The transition distribution at events. At each event time τ, Z changes according to some transition kernel

PDMP



PDMP

Date back to 1951 paper by Mark Kac on the telegraph process.

Mathematical foundations: Davis (1984, JRSS B)

Instrinsically continuous in time unike (almost all) algorithms. Why would they ever be useful for simulation?

Unlike diffusion processes they are comparatively understudied, and underused (either for models or in stochastic simulation).

PDMP

Date back to 1951 paper by Mark Kac on the telegraph process.

Mathematical foundations: Davis (1984, JRSS B)

Instrinsically continuous in time unike (almost all) algorithms. Why would they ever be useful for simulation?

Unlike diffusion processes they are comparatively understudied, and underused (either for models or in stochastic simulation).

.... until recently

PDMP

Date back to 1951 paper by Mark Kac on the telegraph process.

Mathematical foundations: Davis (1984, JRSS B)

Instrinsically continuous in time unike (almost all) algorithms. Why would they ever be useful for simulation?

Unlike diffusion processes they are comparatively understudied, and underused (either for models or in stochastic simulation).

.... until recently

A review paper covering much of this material is Fearnhead, Bierkens, Pollock and R (2018) *Statistical Science*.

Non-reversibility for MCMC?

BUT it has long been known in probability that non-reversible chains can sometimes converge much more rapidly than reversible ones (see for instance Hwang, Hwang-Ma and Sheu (1993), Chen Lovasz and Pak (1999), Diaconis, Holmes and Neal (2000).

Non-reversibility for MCMC?

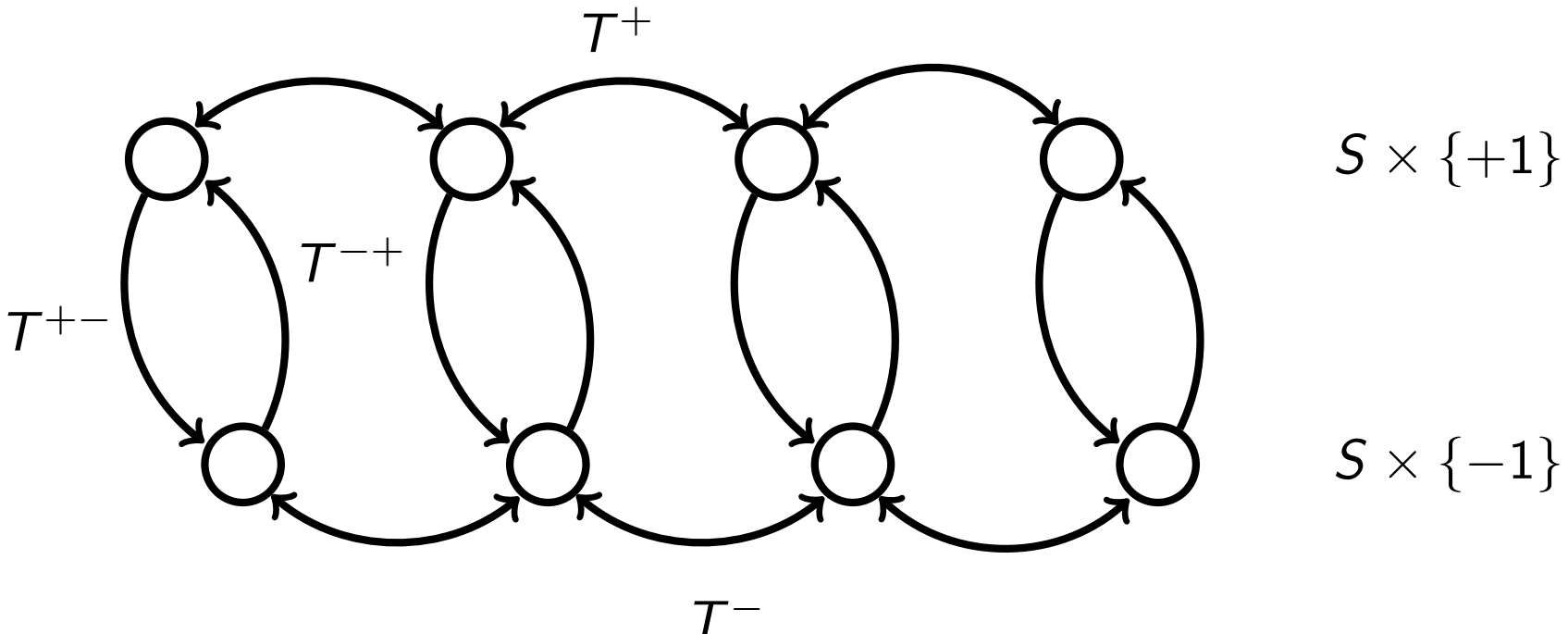
BUT it has long been known in probability that non-reversible chains can sometimes converge much more rapidly than reversible ones (see for instance Hwang, Hwang-Ma and Sheu (1993), Chen Lovasz and Pak (1999), Diaconis, Holmes and Neal (2000).

Hamiltonian MCMC (Hybrid Monte Carlo) tries to construct chains with non-reversible character, but ultimately it is also reversible because of the accept/reject step.

A general lifted Markov chain [Turitsyn, Chertkov, Vucelja, 2011]

- State space S augmented to $S^\sharp = S \times \{-1, +1\}$.
- T^+, T^- are sub-Markov transition matrices on S.
- $T^\pm$ satisfy skew-detailed balance: for all $x, y \in S$, $\pi(x)T^+(x,y) = \pi(y)T^-(y,x)$.
- T^{-+}, T^{+-} transitions between replicas, e.g.

$$T^{-+}(x) = \max\left(0, \sum_{y \neq x}(T^+(x,y) - T^-(x,y))\right).$$



Lifted Metropolis-Hastings [Turitsyn, Chertkov, Vucelja, 2011]

How to choose T^+ and T^-?

Introduce a quantity of interest: $\eta : S \to \mathbb{R}$

Take (Q, π) reversible, e.g. Metropolis-Hastings chain.

Define

$$T^+(x,y) := \begin{cases} Q(x,y) & \text{if } \eta(y) \geq \eta(x) \\ 0 & \text{if } \eta(y) < \eta(x). \end{cases}$$

$$T^-(x,y) := \begin{cases} Q(x,y) & \text{if } \eta(y) \leq \eta(x) \\ 0 & \text{if } \eta(y) > \eta(x). \end{cases}$$

Then skew-detailed balance is satisfied:

$$\pi(x)T^+(x,y) = \pi(y)T^-(y,x) \quad \text{for all } x, y.$$

In practice, Lifted Metropolis-Hastings algorithm:

- Propose according to proposal chain Q
- If move is allowed, accept with MH acceptance probability
- If move is not allowed, possibly switch replica.

Does lifting solve the non-reversible MCMC problem?

The problem is that we need to know the switching probabilities, eg

$$T^{-+}(x) = \max\left(0, \sum_{y \neq x} (T^{+}(x,y) - T^{-}(x,y))\right).$$

This will typically be difficult to calculate, usually impossible in continuous state spaces.

Does lifting solve the non-reversible MCMC problem?

The problem is that we need to know the switching probabilities, eg

$$T^{-+}(x) = \max\left(0, \sum_{y \neq x} (T^{+}(x,y) - T^{-}(x,y))\right).$$

This will typically be difficult to calculate, usually impossible in continuous state spaces.

So lifting is not generally applicable

But, mathematically we can take a limit of smaller proposed moves and speed up the process to obtain a continuous time limit.

We initially did this for the Curie-Weiss model in statistical physics (http://arxiv.org/abs/1509.00302. *Annals of Applied Probability*, 2017).

But, mathematically we can take a limit of smaller proposed moves and speed up the process to obtain a continuous time limit.

We initially did this for the Curie-Weiss model in statistical physics (http://arxiv.org/abs/1509.00302. *Annals of Applied Probability*, 2017).

This was purely for mathematical reasons to understand lifting for the Curie-Weiss model.

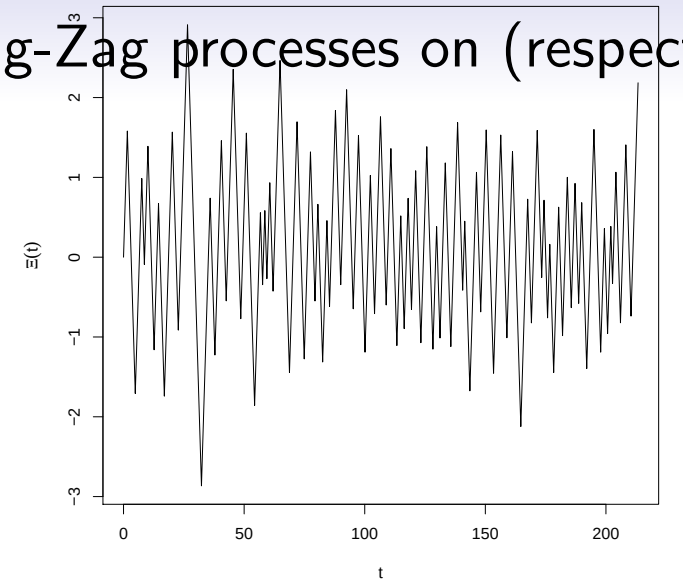
But, mathematically we can take a limit of smaller proposed moves and speed up the process to obtain a continuous time limit.

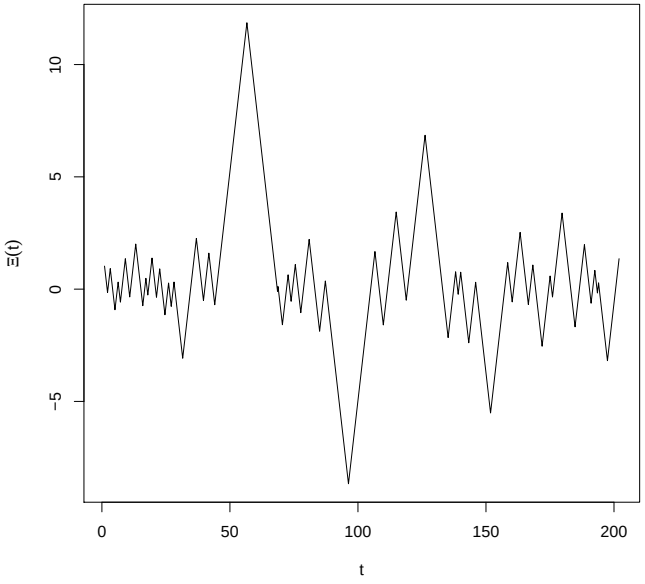
We initially did this for the Curie-Weiss model in statistical physics (http://arxiv.org/abs/1509.00302. *Annals of Applied Probability*, 2017).

This was purely for mathematical reasons to understand lifting for the Curie-Weiss model.

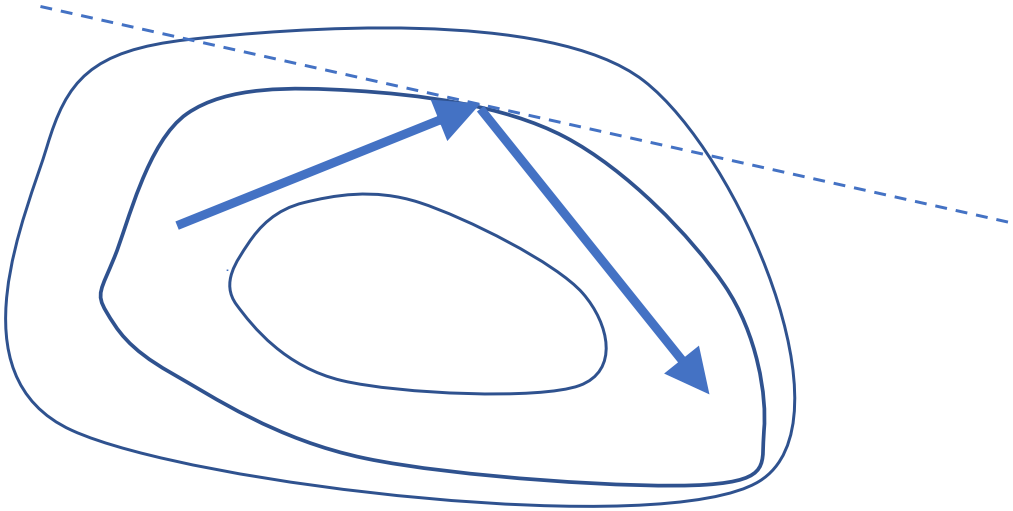
But the continuous-time limit argument extends easily to general target densities.

One-dimensional Zig-Zag processes on (respectively) Gaussian and Cauchy targets.





Closely related to another PDMP scheme, the bouncy particle sampler (BPS), [Bouchard-Côté et al., 2015].



Many alternatives/variants available.

Canonical Zig-Zag

State (X_t, V_t) in dimension d.

$$dX_t = V_t \; dt$$

$V_{t-}^{(i)} \to 1 - V_{t-}^{(i)}$ at rate

$$\lambda_i(X_t, V_t) = \lambda_i^0(X_t, V_t) \equiv \max\left\{0, -V_{t-}^{(i)} \frac{\partial \log \pi(X_{t-})}{\partial X^{(i)}}\right\}$$

Invariant distribution is

$$\pi_E(x, v) \propto \pi(x)$$

ie in stationarity X and V are independent with V being uniform over all configerations: $(\pm 1, \pm 1, \ldots \pm 1)$

Refreshment

But there is a lot more flexibility!

For instance, can take

$$\lambda_i(x, v) = \lambda_i^0(x, v) + \nu(x)$$

for any function ν.

Why might we do this?

Refreshment

But there is a lot more flexibility!

For instance, can take

$$\lambda_i(x, v) = \lambda_i^0(x, v) + \nu(x)$$

for any function ν.

Why might we do this?

To help visit different parts of the state space.

But the larger ν is, the *closer* to reversibility.

The canonical Zig-Zag is the *most non-reversible*.

Implementation

How do we simulate continuous time stochastic process like this?

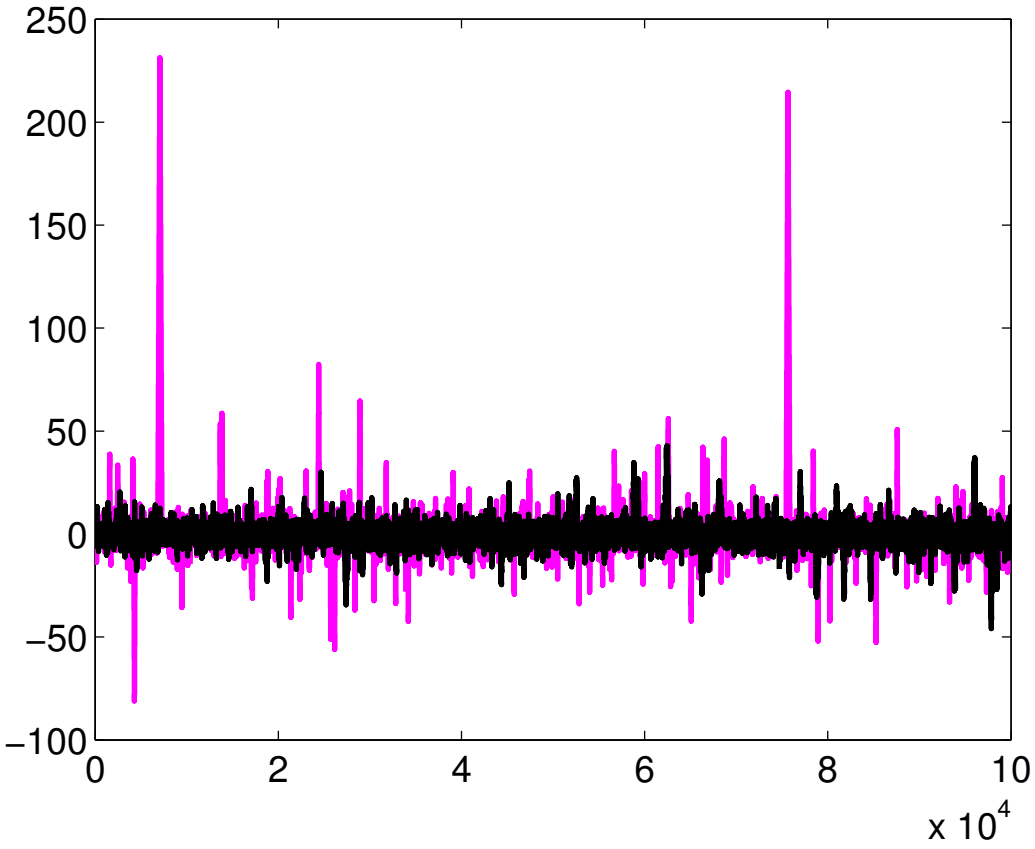
By using thinned poisson processes

For example, if $|(\log \pi)'(x)| < c$, simulate a Poisson process of rate c (by simulating the exponential inter-arrival times). Then at each poisson time, we accept as a direction change with probability $\max(-(\log \pi)'(x), 0)/c$.

This makes the algorithm inexpensive to implement as we only need to calculate $(\log \pi)'(x)$ occasionally.

There are many other details .... though the method is not so complicated.

Zig zag process for sampling the Cauchy distribution

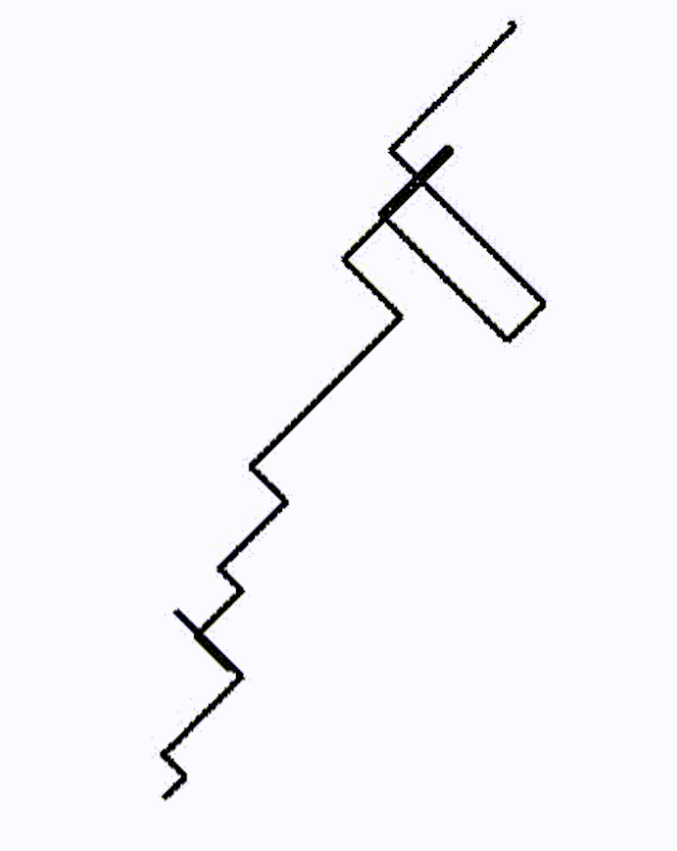


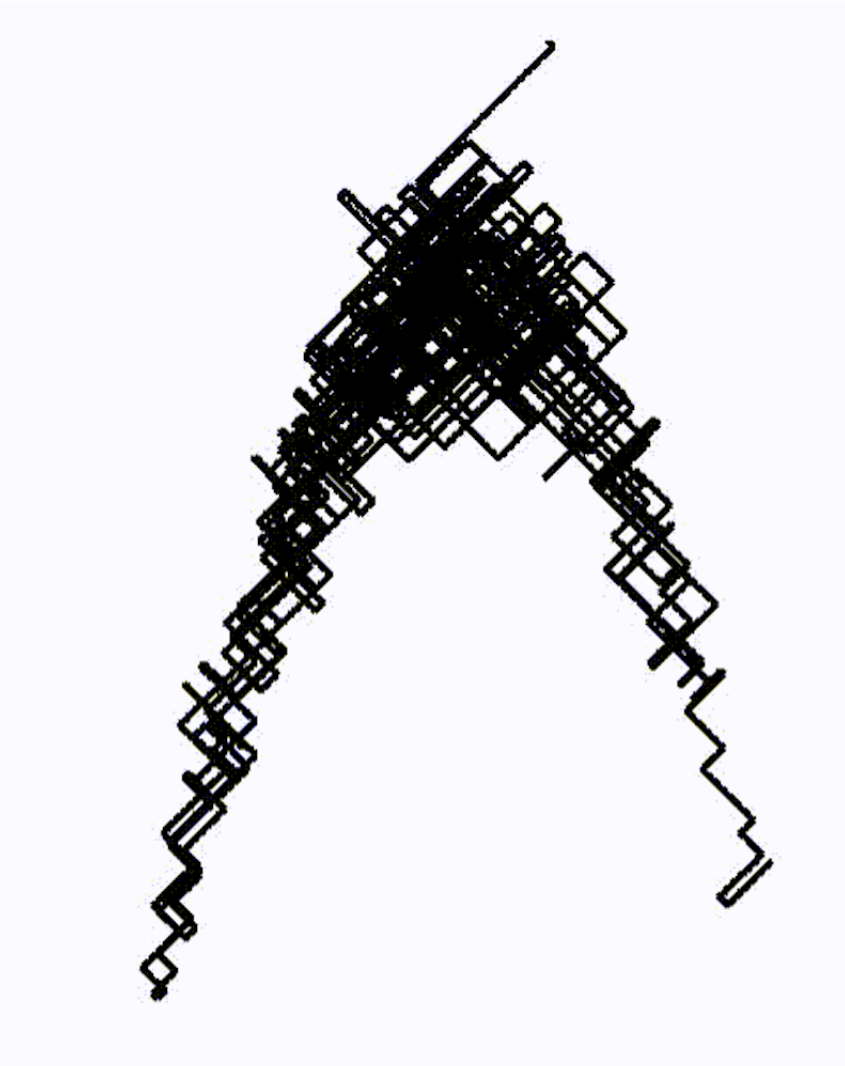
$T = 10,000$

Multi-dimensional zig zag process

Multi-dimensional zig zag process: here we have a multi-dimensional binary velocity, eg $(1, -1, -1, 1, 1, -1, 1, 1)$.

Efficient sampling (currently for potentials with locally Lipschitz gradients in multiple dimensions, but with obvious ways to extend).





Subsampling

Motivation: intractable likelihood problems where calculating π at any one fixed location is prohibitively expensive (given that very many evaluations will be required to run the algorithm. For this talk, concentrate on the Bayesian setting:

$$\pi(x) = \prod_{i=1}^{N} \pi_i(x)$$

Eg we have N observations (but this method is not in any way restricted to the independent data case).

Subsampling

Motivation: intractable likelihood problems where calculating π at any one fixed location is prohibitively expensive (given that very many evaluations will be required to run the algorithm. For this talk, concentrate on the Bayesian setting:

$$\pi(x) = \prod_{i=1}^{N} \pi_i(x)$$

Eg we have N observations (but this method is not in any way restricted to the independent data case).

Aim to be lazy and only use a small number of the terms in the product.

Subsampling

Motivation: intractable likelihood problems where calculating π at any one fixed location is prohibitively expensive (given that very many evaluations will be required to run the algorithm. For this talk, concentrate on the Bayesian setting:

$$\pi(x) = \prod_{i=1}^{N} \pi_i(x)$$

Eg we have N observations (but this method is not in any way restricted to the independent data case).

Aim to be lazy and only use a small number of the terms in the product.

For instance we might try pseudo-marginal MCMC (Beaumont, 2003, Andrieu and Roberts, 2009).But that would require an unbiased non-negative estimate of $\pi(x)$ with variance which is stable as a function of N.

Subsampling

Motivation: intractable likelihood problems where calculating π at any one fixed location is prohibitively expensive (given that very many evaluations will be required to run the algorithm. For this talk, concentrate on the Bayesian setting:

$$\pi(x) = \prod_{i=1}^{N} \pi_i(x)$$

Eg we have N observations (but this method is not in any way restricted to the independent data case).

Aim to be lazy and only use a small number of the terms in the product.

For instance we might try pseudo-marginal MCMC (Beaumont, 2003, Andrieu and Roberts, 2009).But that would require an unbiased non-negative estimate of $\pi(x)$ with variance which is stable as a function of N. But this is not possible for a product without computing cost which is at least $O(N)$.

Subsampling within PDMP

PDMP for the exploration of high-dimensional distributions (such as zig-zag or the ScaLE algorithm, Fearnhead, Johansen, Pollock and Roberts, 2016) typically use $\log \pi(x)$ rather than $\pi(x)$ and

$$\log \pi(x) = \sum_{i=1}^{N} \log \pi_i(x)$$

for which there are well-behaved $O(1)$ cost, $O(1)$ variance (or sometime a little worse). Can we use this?

Zig zag switching rate $\max\left(0, -j \sum_{i=1}^{N} (\log \pi)'_i(x)\right) \rightsquigarrow O(N)$ calculation at every switch

Subsampling for zig-zag

Sub-sampling

- Determine global upper bound M for switching rate
- Simulate Exponential(M) random variable T
- Generate $I \sim \text{discrete}(\{1, \ldots, N\})$
- Accept the generated T as a "switching time" with probability $N \max\left(0, -j(\log \pi_I)'(Y(T))\right)/M$

Theorem: This works! (invariant distribution π)

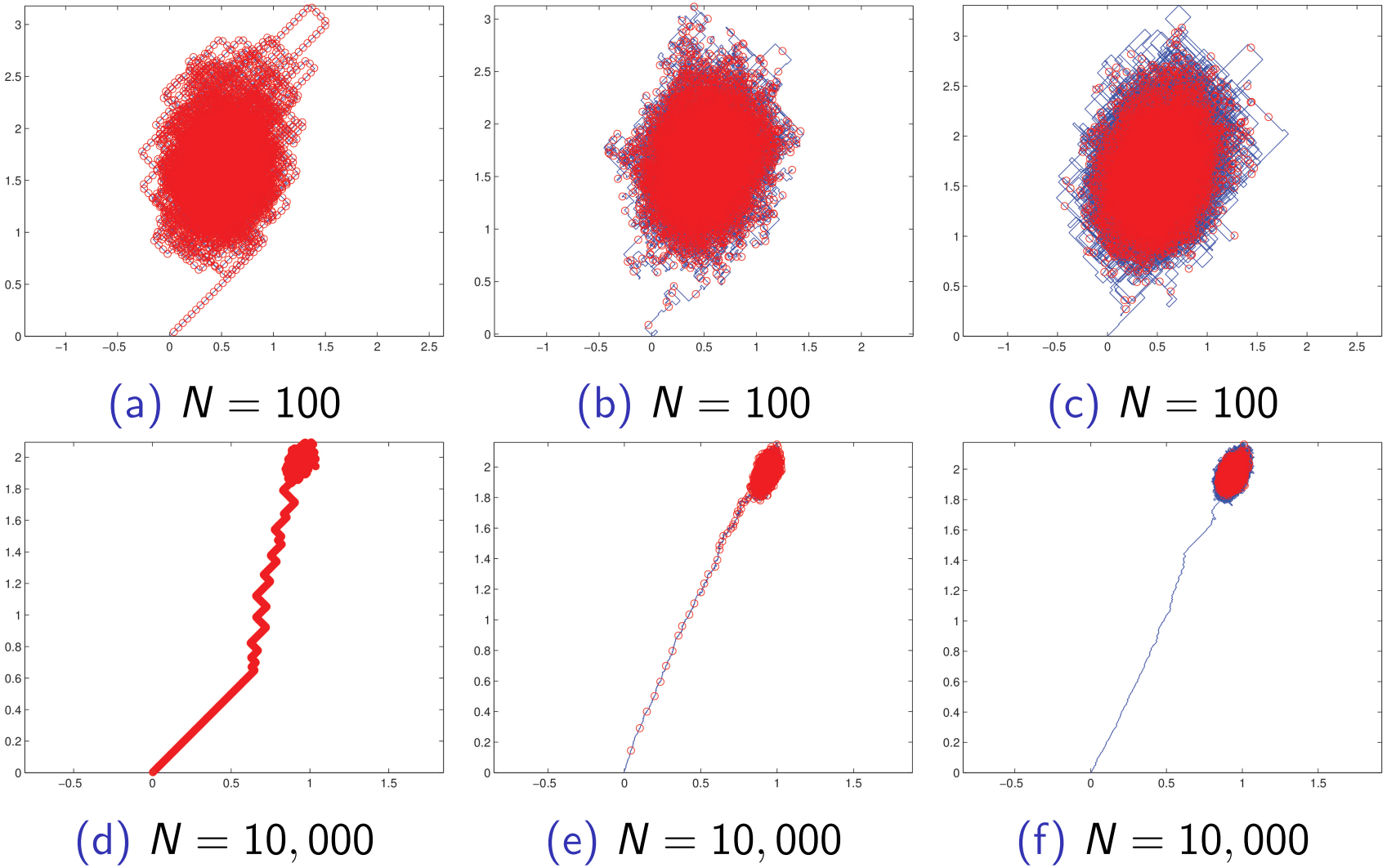
Subsampling + control variates

Crudely, for an $O(1)$ update in state space:

- Without subsampling, $O(N)$ computations required
- Using subsampling, gain factor $N^{1/2}$ $\rightsquigarrow$ complexity $O(N^{1/2})$ per step
- Using control variates, gain additional factor $N^{1/2}$ $\rightsquigarrow$ complexity $O(1)$ per step

Superefficiency We call an epoch the time taken to make one function evaluation of the target density π. The control variate subsampled zig-zag is superefficient in the sense that the effective sample size from running the algorithm per epoch diverges.

Subsampling + control variates – Logistic growth



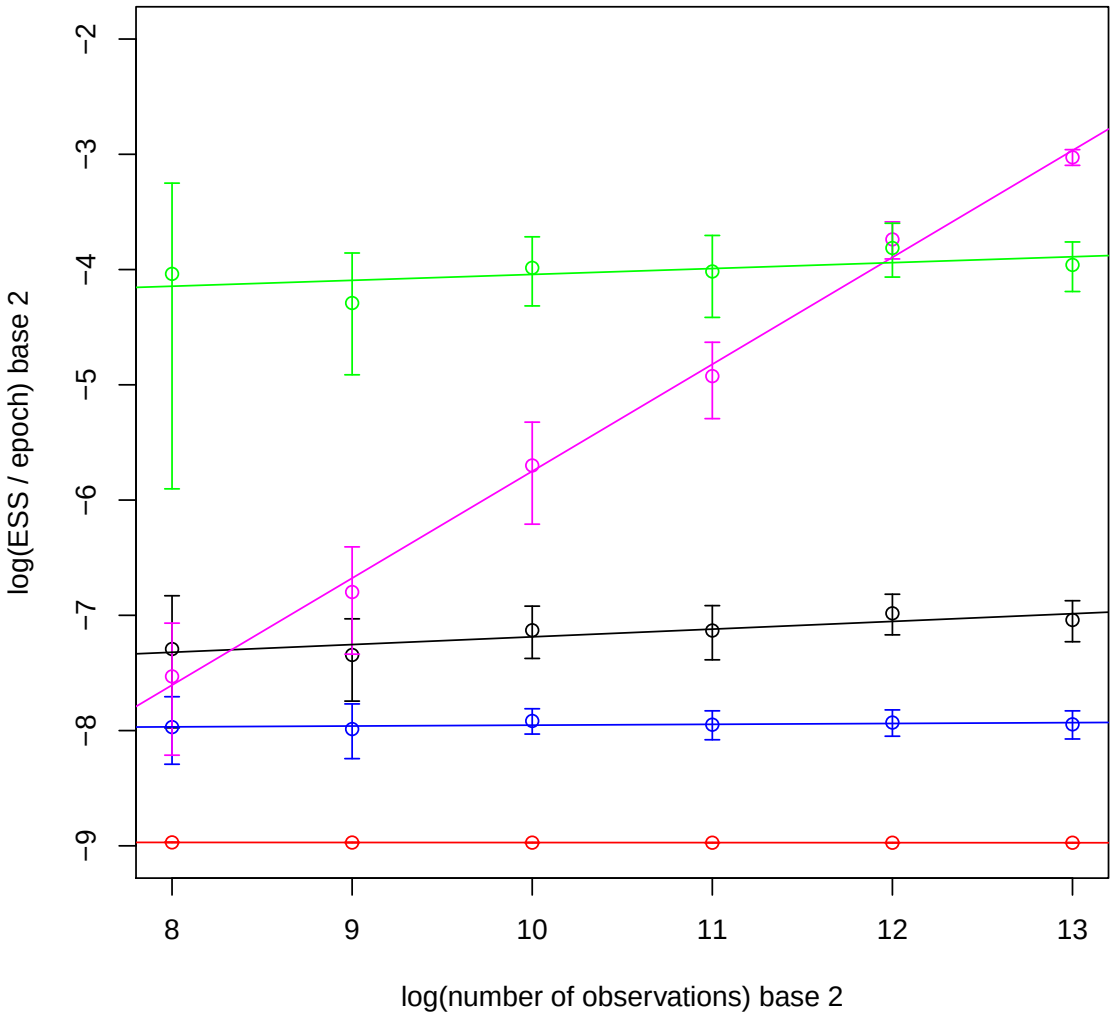
(a) $N = 100$ (b) $N = 100$ (c) $N = 100$

(d) $N = 10,000$ (e) $N = 10,000$ (f) $N = 10,000$

[Bierkens, Roberts, 2017, http://arxiv.org/abs/1509.00302]

[Bierkens, Fearnhead, Roberts, Ann Stat, 2019]

Effective Sample Size per epoch



Is the zig-zag ergodic?

An invariant distribution for (x, v) for the zig-zag is just

$$\pi_E(x, v) \propto \pi(x)$$

ie $X \sim \pi$ and independently the velocity v is uniformly distributed within $\{-1, 1\}^d$.

Ergodicity requires that we can reach all locations in (x, v) space. But can we ensure this?

Simple solution:

Include a residual jump rate γ_i which is uniformly positive, eg $\gamma_i(x) = \tilde{\gamma} > 0$.

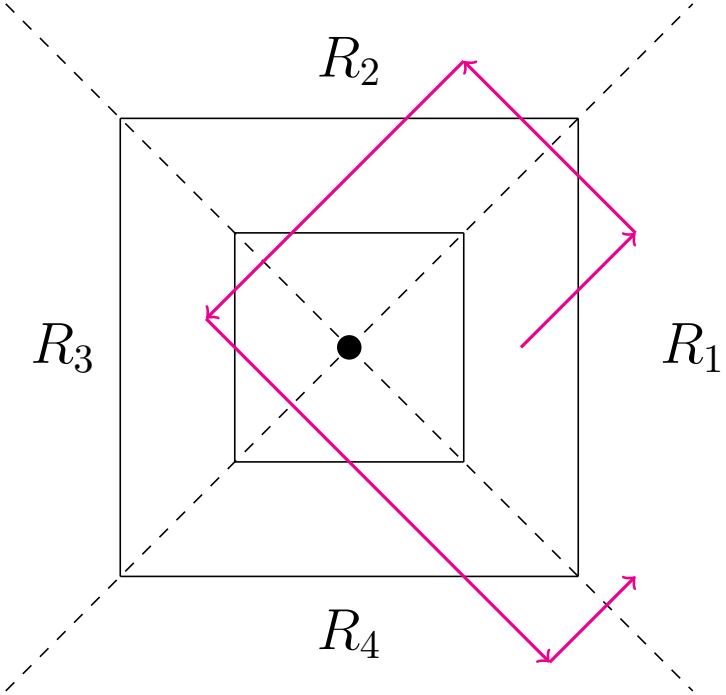
This makes proving ergodicity easy under minimal assumptions on π (eg that it is C^1 and positive everywhere).

But for large $\tilde{\gamma}$, the zig-zag then looks more and more like a Langevin diffusion which is reversible. Many of the advantages of non-reversibility are therefore lost.

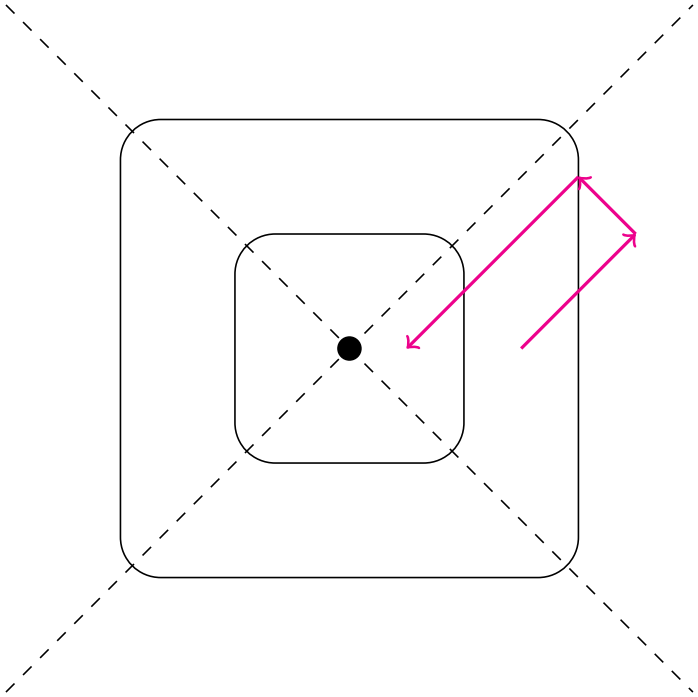
Can we establish an ergodicity result for the canonical zig-zag, ie $\tilde{\gamma} = 0$?

A counter example

$$\pi(x, y) \propto \{-\max(|x|, |y|)\}$$

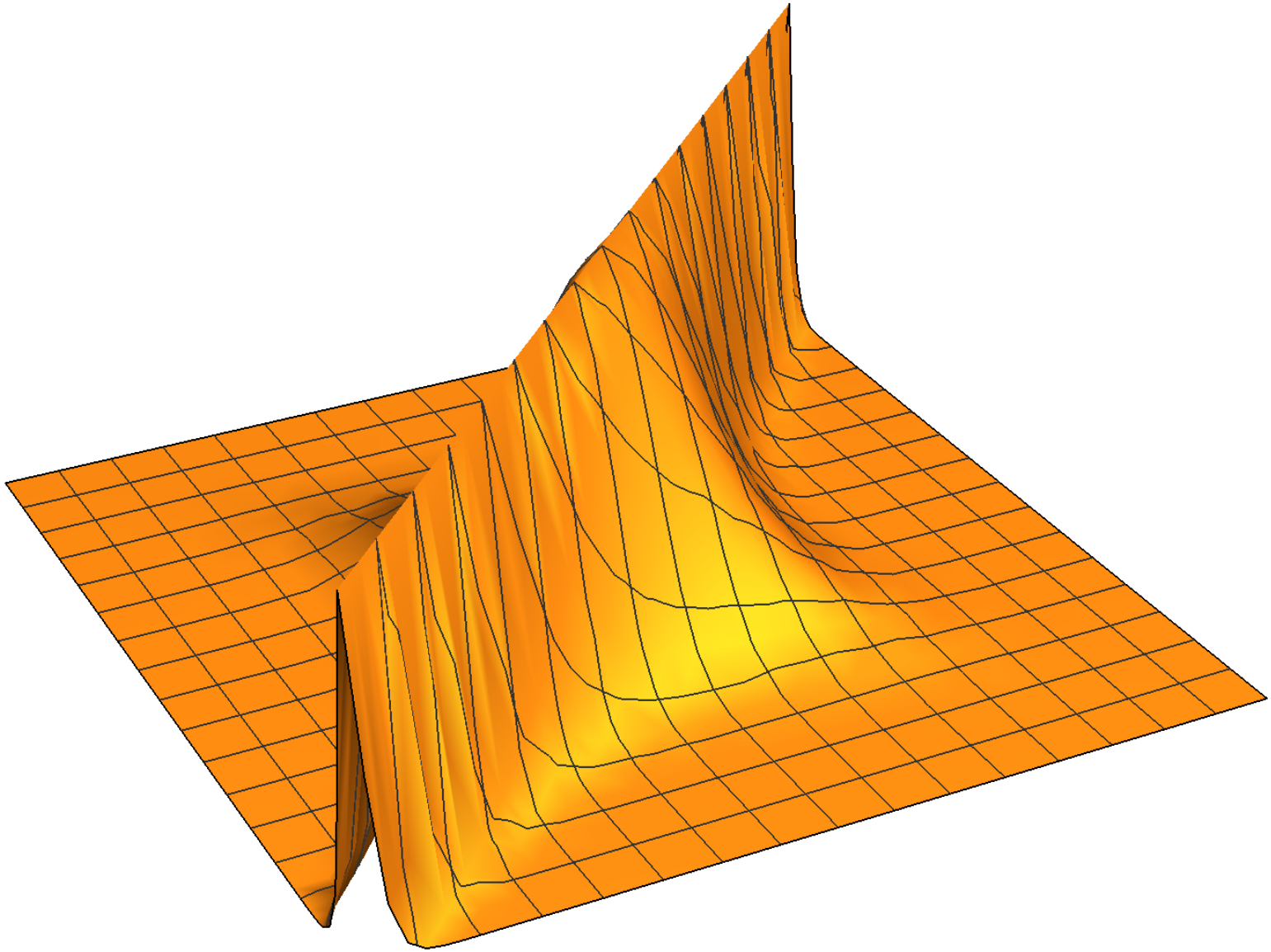


(a) Contour lines, the regions R_1, R_2, R_3 and R_4, and a typical trajectory for the potential function $U(x) = \max(|x_1|, |x_2|)$. From the displayed starting position it is impossible to reach a point in R_1 with direction $(-1, -1)$.



(b) Once we smooth the density function slightly, it becomes possible to switch the second coordinate of the direction vector, making the process irreducible.

We also need to preclude evanesence



Theorem

Assume that

1. π *is positive and* $\mathcal{C}^3$
2. $\lim_{|x|\to\infty} \pi(x) = 0$, *and*
3. *has a non-degenerate local maximum, ie the Hessian at the local maximum is strictly negative definite.*

Then the chain is irreducible and converges to π *from any starting distribution.*

Theorem

Assume that

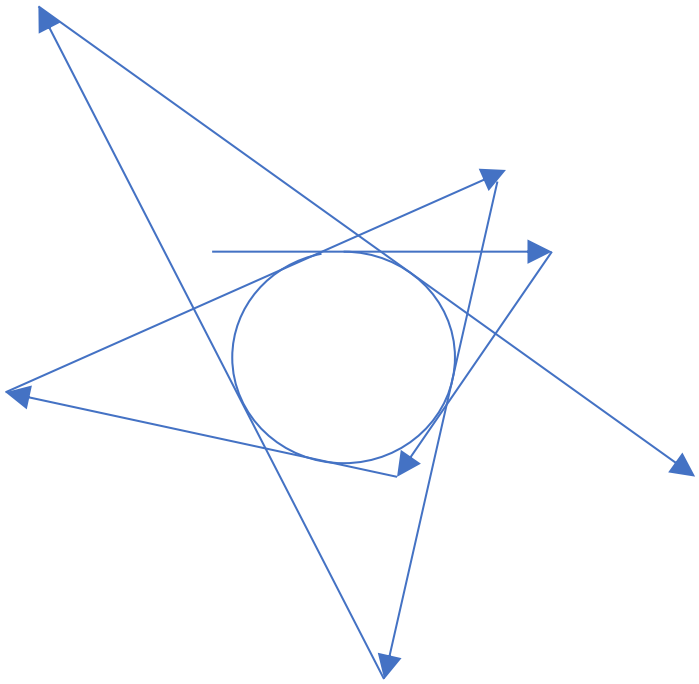
1. π *is positive and* $\mathcal{C}^3$
2. $\lim_{|x|\to\infty} \pi(x) = 0$, *and*
3. *has a non-degenerate local maximum, ie the Hessian at the local maximum is strictly negative definite.*

Then the chain is irreducible and converges to π *from any starting distribution.*

Method of proof relies heavily upon smoothness and the ability to approximate by a Gaussian around the local mode. (3) can no doubt be weakened.

Irreducibility of BPS?

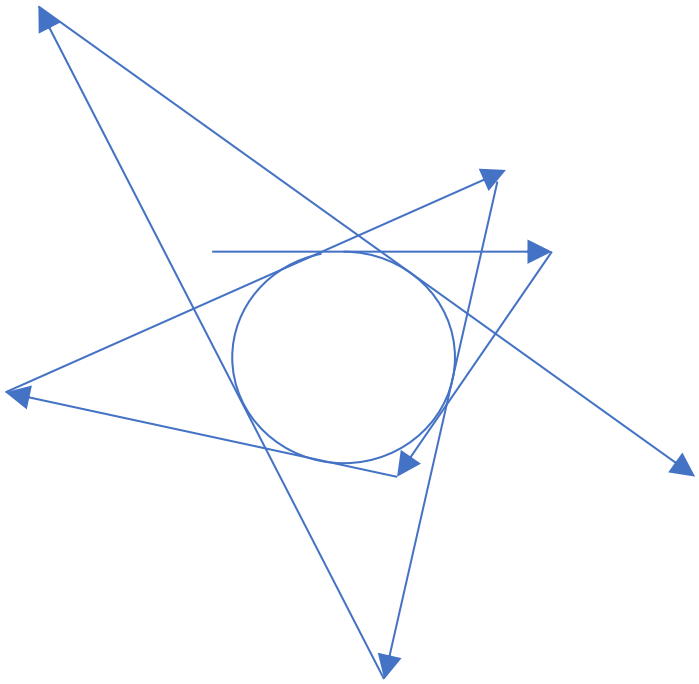
Without refreshment, canonical BPS can easily be irreducible. Eg for 2-dimensional isotropic Normal Distribution:



Irreducibility is restored under very mild regularity conditions under refreshment.

Irreducibility of BPS?

Without refreshment, canonical BPS can easily be irreducible. Eg for 2-dimensional isotropic Normal Distribution:



Irreducibility is restored under very mild regularity conditions under refreshment.

But how much refreshment?

High-dimensional PDMP

Exciting recent directions look at the behaviour of PDMPs in high-dimensional settings, eg Andrieu, Durmus, Nsken, and Roussel, Deligianidis, Durmus, others..

Report here joint work with Joris Bierkens and Kengo Kamatani.

A best case scenario analysis

$\mathbf{X} \sim \pi \equiv N(0, I_d)$.

Consider

1. coordinate process $X^{(1)}$;
2. radial process or log density, $|\mathbf{X}|$;
3. angular process: we consider $\langle \mathbf{X}, v \rangle$

Method	Angular momentum	Negative log-density	1st Coordinate
ZZ	$O(d^{1/2})$	$O(d^{1/2})$	$O(d^{1/2})$
BPS	$O(1)$	$O(d)$	$O(d)$

Table: Size of continuous time intervals required to obtain approximately independent samples for the piecewise deterministic processes.

Method	Angular momentum	Negative log-density	1st Coordinate
ZZ	$O(d)$	$O(d)$	$O(d)$
BPS	$O(d)$	$O(d^2)$	$O(d^2)$

Table: Algorithmic complexity (incorporating computational effort per unit time).

Method	Angular momentum	Negative log-density	1st Coordinate
ZZ	$O(d^{1/2})$	$O(d^{1/2})$	$O(d^{1/2})$
BPS	$O(1)$	$O(d)$	$O(d)$

Table: Size of continuous time intervals required to obtain approximately independent samples for the piecewise deterministic processes.

Method	Angular momentum	Negative log-density	1st Coordinate
ZZ	$O(d)$	$O(d)$	$O(d)$
BPS	$O(d)$	$O(d^2)$	$O(d^2)$

Table: Algorithmic complexity (incorporating computational effort per unit time).

As a comparison:

Method	Algorithmic complexity
RWM	$O(d^2)$
MALA	$O(d^{4/3})$
HMC	$O(d^{5/4})$

Table: Complexity of traditional MCMC

MCMC and Hybridisation

MCMC methods have often been hybridised, eg P_1 is good at mixing horizontally and P_2 is good at mixing vertically. So instead we use

$$\frac{P_1 + P_2}{2} \text{ random scan}$$

or

$$P_1 P_2 \text{ deterministic or systematic scan}$$

or other variants. The traditional approach uses $\pi = \pi P_i$, $i = 1, 2$.

A much more flexible framework could be achieved if we allow P_1 and P_2 to compensate for each other.

An example of this is the RESTORE algorithm.

RESTORE

The Randomly Exploring STOchastically REnewing algorithm.

A continuous time algorithm proceeding according to deterministic or stochastic generator $\mathcal{L}$ but reset to distribution with density μ at killing times with state dependent intensity $\kappa(x)$. We set

$$\kappa(x) = \frac{\mathcal{L}^*\pi(x)}{\pi(x)} + c\frac{\mu(x)}{\pi(x)} \ ,$$

for some positive constant c chosen to ensure the non-negativity of $\kappa(x)$. Here $\mathcal{L}^*$ denotes the adjoint of $\mathcal{L}$.

μ could be some approximation for π.

RESTORE (continued)

Suppose the $\mathcal{L}$ dynamics are deterministic satisfying the differential equation

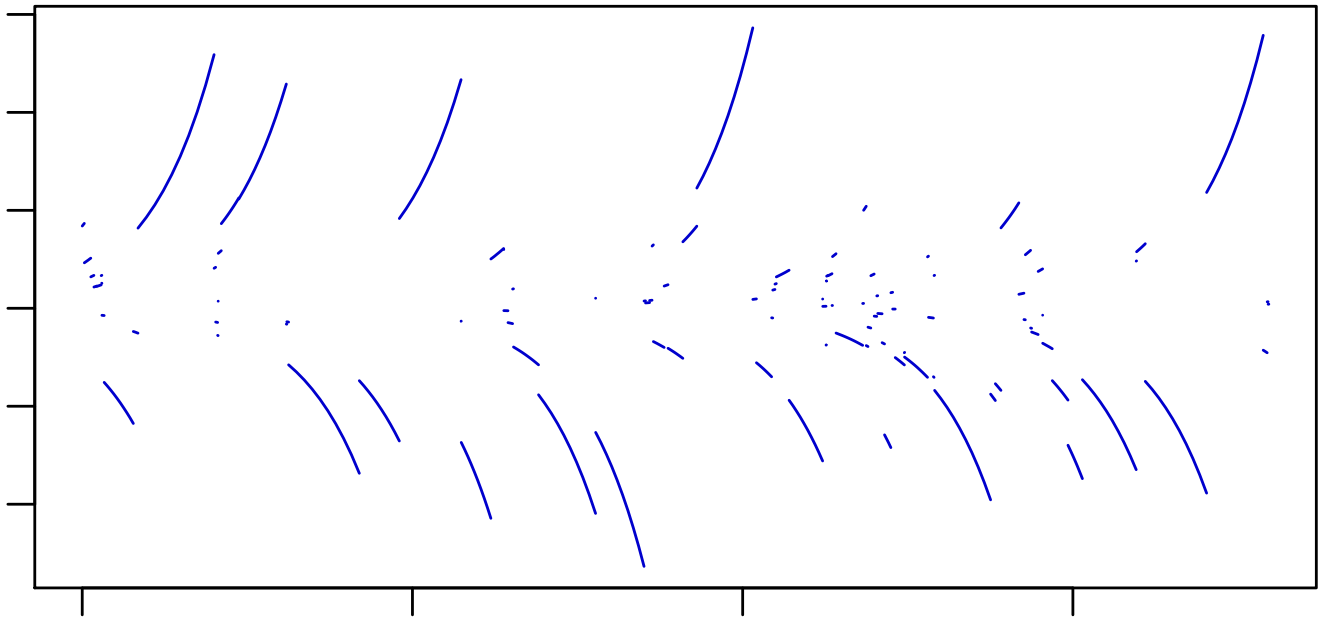
$$\frac{dy}{dt} = y$$

so that

$$\mathcal{L}f(y) = yf'(y) \ .$$

Can (for instance) be used to correct for under-dispersion in μ.

The following simulation was carried out by Hector McKimm.



Bayesian fusion

Imagine we have N data sets each sitting on different computer systems. Data cannot be shared for reasons of privacy or sheer size. Each computer gives rise to its own subposterior, $\pi_n(y)$, $1 \leq i \leq N$.

Samples from π_n can be shared, but how can we sample from

$$\pi(y) = \prod_{i=1}^{N} \pi_n(y) \ ?$$

Monte Carlo fusion provides a generic and exact solution. Set

$$g(x_1, \ldots x_N, y) = \prod_{i=1}^{N} \left[\frac{\pi_n^2(x_n) p_n(y|x_n)}{\pi_n(y)} \right]$$

where p_n is any Markov chain move which preserves π_n^2. Then $g(x_1, \ldots x_N, y)$ admits y marginal π.

If we take p_n to be the transition of a suitable Langevin diffusion, it turns out that we can sample from g by rejection sampling from

$$h(x_1, \ldots x_N, y) = \prod_{n=1}^{N} \pi_n(x) \exp\left(-\frac{N(y-\bar{x})^2}{2T}\right)$$

which is easily sampled by taking independent draws from each of the N servers, averaging the values and imposing a Gaussian perturbation to get y.

Note that this can all be done securely without even sharing simulated values by neat tricks from homomorphic secret sharing.

This is the basis of ongoing work with Hongsheng Dai, Murray Pollock and Louis Aslett.

Final remarks

- PDMPs have many uses for simulation of stochastic processes (even those very different from PDMPs) as well as steady state simulation.
- Subsampling and control-variate tweaks greatly improve efficiency in certain situations. PDMP are particularly amenable to this.
- Can zigzag be a competitor to Hamiltonian MCMC?
- RESTORE provides a very general framework of propose and improve algorithms with the potential to improve on approximations to the target density.
- Bayesian fusion allows fusion of samples without approximation. Currently not robust to large numbers of subsamples.