

The zig-zag and super-efficient sampling for Bayesian analysis of big data

LMS-CRiSM Summer School on Computational Statistics
15th July 2018

Gareth Roberts, University of Warwick

Joint work with Joris Bierkens, Paul Fearnhead and
Pierre-Andre Zitt



CoSInES project

CoSInES: Computational Statistical Inference for Engineering and Security

Scalable (mainly) Bayesian computational statistics methodology for complex problems.

Applications in Data-centric Engineering and Defence and Security.

Each project is recruiting 5 4-year postdocs working in theory, methodology and applications.

Starts in **October, 2018**.

Launch day on **10th September** at Warwick.

Closing day for **postdocs** (5 or 6) **13th September**.

Talk outline

- Introduce PDMP
- The zig-zag as an alternative to MCMC. “The Zig-Zag Process and Super-Efficient Sampling for Bayesian Analysis of Big Data” [arXiv:1607.03188](#)
- Some theoretical questions and answers. A paper on ergodicity with Bierkens and Zitt, to be on arxiv [very soon](#).

A review paper covering much of this material is to appear *Statistical Science*. “Piecewise Deterministic Markov Processes for Continuous-Time Monte Carlo” [arXiv:1611.07873](#)

Talk outline

- Introduce PDMP
- The zig-zag as an alternative to MCMC. “The Zig-Zag Process and Super-Efficient Sampling for Bayesian Analysis of Big Data” [arXiv:1607.03188](#)
- Some theoretical questions and answers. A paper on ergodicity with Bierkens and Zitt, to be on arxiv [very soon](#).

A review paper covering much of this material is to appear *Statistical Science*. “Piecewise Deterministic Markov Processes for Continuous-Time Monte Carlo” [arXiv:1611.07873](#)

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

$$\frac{dz_t}{dt} = \Phi(z_t), \quad (1)$$

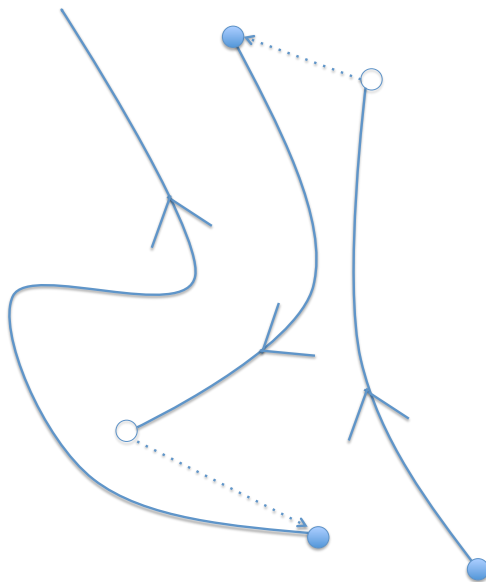
So

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

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)

Intrinsically continuous in time unlike (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)

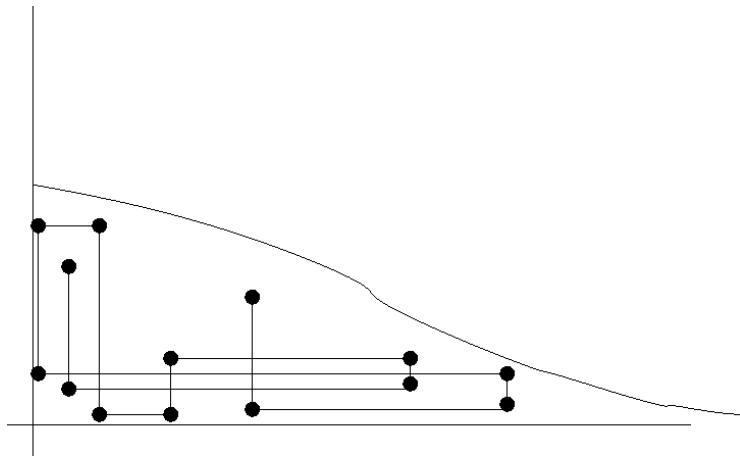
Intrinsically continuous in time unlike (almost all) algorithms. Why would they ever be useful for simulation?

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

.... until recently

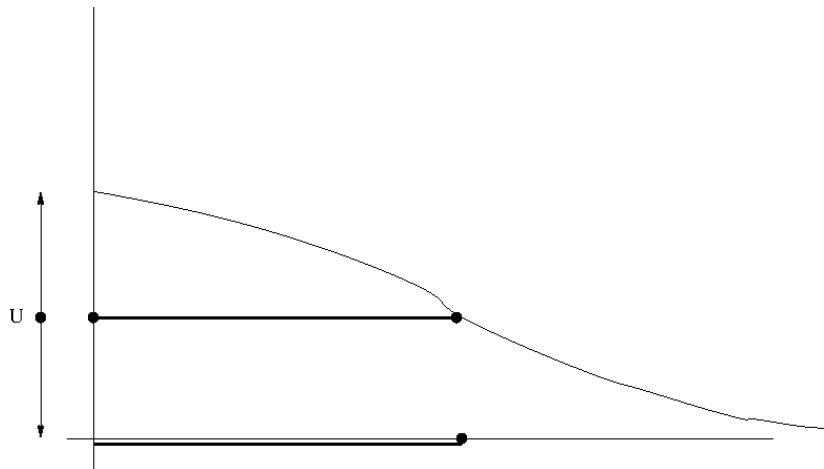
The slice sampler

An MCMC method for simulating from a target density by introducing an **auxiliary variable**.



Well understood theoretically, eg R + Rosenthal (1999)

An alternative



Metropolis-Hastings

[Metropolis et al. 1953, Hastings 1970]

- S finite set (*state space*)
- $Q(x, y)$ transition probabilities on S (*proposal chain*)
- $\pi(x)$ a probability distribution on S (*target distribution*)

Define **acceptance probabilities**

$$A(x, y) = \min \left(1, \frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)} \right).$$

The **Metropolis-Hastings (MH) chain** is

$$P(x, y) = \begin{cases} Q(x, y)A(x, y) & y \neq x, \\ 1 - \sum_{z \neq x} Q(x, z)A(x, z) & y = x. \end{cases}$$

The MH chain is **reversible**:

$$\pi(x)P(x, y) = \pi(y)P(y, x) \quad \forall x, y \in S.$$

In particular, π is **invariant** for P .

Non-reversibility for MCMC?

The fact that MH is reversible is **good** because

- Beautiful clean mathematical theory: Markov chain transition operator is **self-adjoint**, spectrum is **real**, if **geometrically ergodic** then CLTs hold for **all** L^2 functions ...

Non-reversibility for MCMC?

The fact that MH is reversible is **good** because

- Beautiful clean mathematical theory: Markov chain transition operator is **self-adjoint**, spectrum is **real**, if **geometrically ergodic** then CLTs hold for **all** L^2 functions ...
- Detailed balance is a **local condition** - crucial for implementability, don't need to do an integral to decide whether to accept or reject.

Non-reversibility for MCMC?

The fact that MH is reversible is **good** because

- Beautiful clean mathematical theory: Markov chain transition operator is **self-adjoint**, spectrum is **real**, if **geometrically ergodic** then CLTs hold for **all** L^2 functions ...
- Detailed balance is a **local condition** - crucial for implementability, don't need to do an integral to decide whether to accept or reject.

So why should we bother to look further?

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.

Metropolis-Hastings

[Metropolis et al. 1953, Hastings 1970]

- S finite set (*state space*)
- $Q(x, y)$ transition probabilities on S (*proposal chain*)
- $\pi(x)$ a probability distribution on S (*target distribution*)

Define **acceptance probabilities**

$$A(x, y) = \min \left(1, \frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)} \right).$$

The **Metropolis-Hastings (MH) chain** is

$$P(x, y) = \begin{cases} Q(x, y)A(x, y) & y \neq x, \\ 1 - \sum_{z \neq x} Q(x, z)A(x, z) & y = x. \end{cases}$$

The MH chain is **reversible**:

$$\pi(x)P(x, y) = \pi(y)P(y, x) \quad \forall x, y \in S.$$

In particular, π is **invariant** for P .

Non-Reversible Metropolis-Hastings

[Bierkens, 2015]

- S finite set (*state space*)
- $Q(x, y)$ transition probabilities on S (*proposal chain*)
- $\pi(x)$ a probability distribution on S (*target distribution*)
- $\Gamma \in \mathbb{R}^{S \times S}$: skew-symmetric matrix with zero row-sums (*vorticity matrix*)

Define *acceptance probabilities*

$$A(x, y) = \min \left(1, \frac{\pi(y)Q(y, x) + \Gamma(x, y)}{\pi(x)Q(x, y)} \right).$$

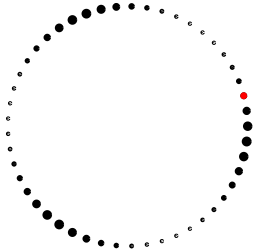
The *non-reversible* Metropolis-Hastings (MH) chain is

$$P(x, y) = \begin{cases} Q(x, y)A(x, y) & y \neq x, \\ 1 - \sum_{z \neq x} Q(x, z)A(x, z) & y = x. \end{cases}$$

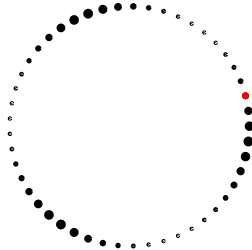
The NRMH chain is *non-reversible*:

$$\pi(x)P(x, y) \neq \pi(y)P(y, x) \quad \exists x, y \in S.$$

But π is *invariant* for P .



(a) Metropolis-Hastings



(b) Non-reversible Metropolis-Hastings

Cycles and lifting

Recall Γ skew-symmetric with zero row sums.

Also want acceptance probability

$$A(x, y) = \min \left(1, \frac{\pi(y)Q(y, x) + \Gamma(x, y)}{\pi(x)Q(x, y)} \right)$$

to be non-negative.

A 4-state example illustrates: **No cycles \Rightarrow no non-reversible Markov chain.**



Cycles and lifting

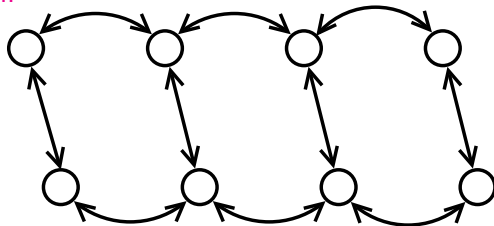
Recall Γ skew-symmetric with zero row sums.

Also want acceptance probability

$$A(x, y) = \min \left(1, \frac{\pi(y)Q(y, x) + \Gamma(x, y)}{\pi(x)Q(x, y)} \right)$$

to be non-negative.

A 4-state example illustrates: **No cycles \Rightarrow no non-reversible Markov chain.**



How to construct lifted MCMC algorithms?

A general lifted Markov chain

[Turitsyn, Chertkov, Vucelja, 2011]

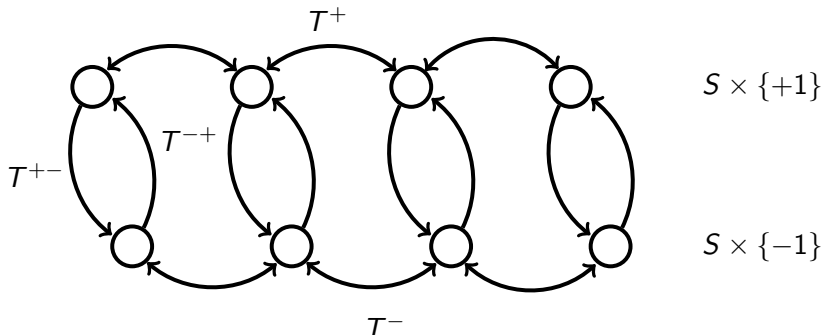
- State space S augmented to $S^\sharp = S \times \{-1, +1\}$.
- T^+, T^- are sub-Markov transition matrices on S .
- T^\pm satisfy **skew-detailed balance**: for all $x, y \in S$,
 $\pi(x) T^+(x, y) = \pi(y) T^-(y, x)$.

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: $\eta : S \rightarrow \mathbb{R}$

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>. to appear in *Annals of Applied Probability*).

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>. to appear in *Annals of Applied Probability*).

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

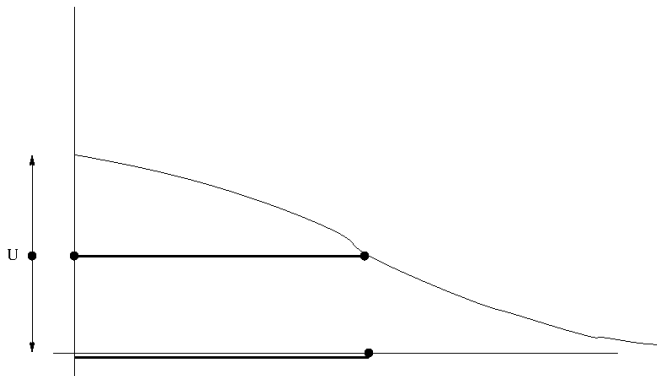
But, mathematically we can take a limit of smaller proposed moves and **speed up** the process to obtain a **continuous time limit**.

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

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.

Another look at our initial example ...

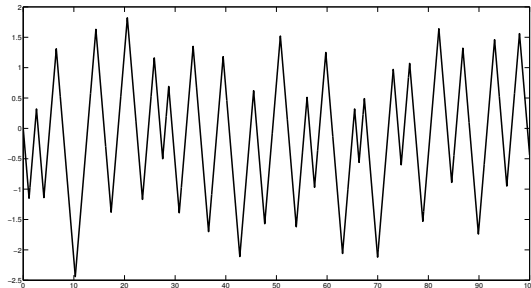


Instead of apriori drawing the uniform random variable, **change direction** with hazard rate

$$\max\{0, -(\log \pi)'(x)\}$$

One-dimensional zig zag process

Extend over the whole real line to a unimodal density with mode at 0 give trajectories:



Implementation

How do we simulate continuous time stochastic process like this?

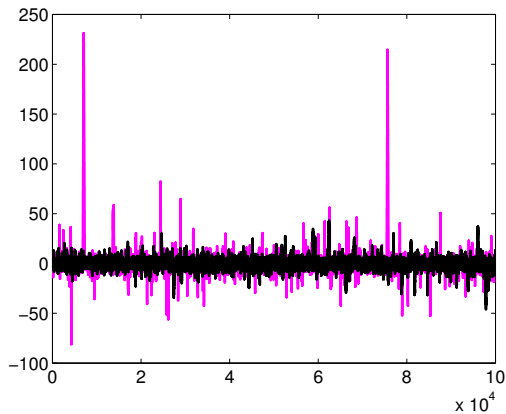
By using **thinned poisson processes**

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 $\theta = (1, -1, -1, 1, 1, -1, 1, 1)$.

Switching rate in j th direction is just

$$\max \{ -\theta_j (\nabla \log \pi(\mathbf{x}))_j, 0 \}$$



PDMP for Monte Carlo

This is now a very active area, for example:

- [Goldstein, 1951],[Kac, 1974]: constant jump rate λ , relation to *telegraph equation*
- [Peters, de With, Physical Review E, 2012]: first occurrence of the zig zag process for sampling from general targets, with multi-dimensional, non-ergodic extension.
- [Fontbana, Guérin, Malrieu, 2012, 2015]: convergence analysis of zig zag process under convex potential.
- [Monmarché, 2014]: simulated annealing using “run-and-tumble” process.
- [Bouchard-Côté et al., 2015]: bouncy particle sampler.

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

Subsampling for zig-zag

Sub-sampling

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

Theorem: This works! (invariant distribution π)

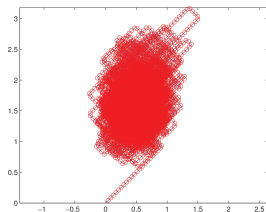
Subsampling + control variates

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

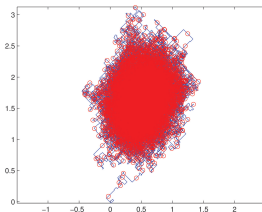
- Without subsampling, $O(N)$ computations required
- Using **subsampling**, gain **factor $N^{1/2}$** \rightsquigarrow complexity $O(N^{1/2})$ per step
- Using **control variates**, gain **additional factor $N^{1/2}$** \rightsquigarrow complexity $O(1)$ per step

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.

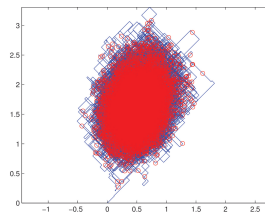
Subsampling + control variates – Logistic growth



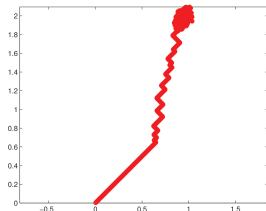
(a) $N = 100$



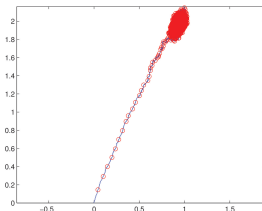
(b) $N = 100$



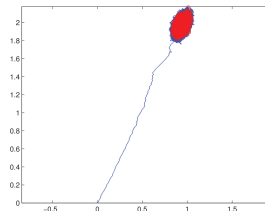
(c) $N = 100$



(d) $N = 10,000$

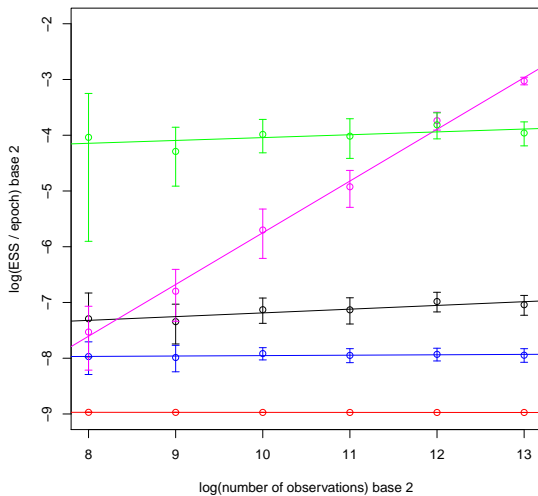


(e) $N = 10,000$



(f) $N = 10,000$

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 $\gamma_i(x) = \tilde{\gamma} > 0$.

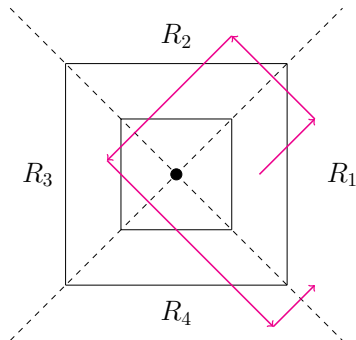
This makes proving ergodicity easy under minimal assumptions on π (eg that it is C^1 and positive everywhere).

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.

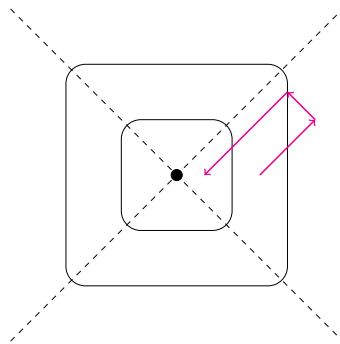
Can we establish an ergodicity result for the canonical zig-zag, ie $\tilde{\gamma} = 0$?

A counter example

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

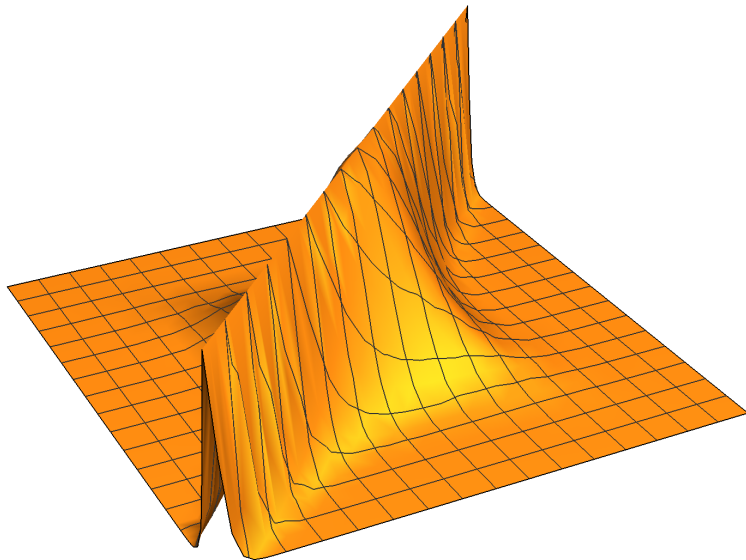


(a) Contour lines, the regions R_1, R_2, R_3 and R_4 , and a typical trajectory for the potential function $U(x) = \max(|x_1|, |x_2|)$. From the displayed starting position it is impossible to reach a point in R_1 with direction $(-1, -1)$.



(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 evanescence



Theorem

Assume that

1. π is positive and \mathcal{C}^3
2. $\lim_{|x| \rightarrow \infty} \pi(x) = 0$, 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. $\lim_{|x| \rightarrow \infty} \pi(x) = 0$, 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.

Complexity

Recent work with Joris Bierkens and Kengo Kamatani (work in progress)

For certain stylised examples (eg with independent components or spherical symmetry), mixing times of Zig-Zag scale as $O(d)$ (with time scaling according to unit computer implementation time).

Beats, RWM, MALA, HMC, Bouncy Particle Sampler, etc.

Will these results be robust to more complex problems?

Final remarks

- PDMPs have many uses for simulation of stochastic processes (even those very different from PDMPs) as well as **steady state** simulation.

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.

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.
- More work is needed on studying the theoretical and empirical properties of these algorithms, and exploiting their flexibility.

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.
- More work is needed on studying the theoretical and empirical properties of these algorithms, and exploiting their flexibility.
- Zigzag is a flexible and usually easy-to-implement method for simulating from a target distribution.

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.
- More work is needed on studying the theoretical and empirical properties of these algorithms, and exploiting their flexibility.
- Zigzag is a flexible and usually easy-to-implement method for simulating from a target distribution.
- Can zigzag be a competitor to Hamiltonian MCMC?

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.
- More work is needed on studying the theoretical and empirical properties of these algorithms, and exploiting their flexibility.
- Zigzag is a flexible and usually easy-to-implement method for simulating from a target distribution.
- Can zigzag be a competitor to Hamiltonian MCMC?
- R package, see **RZigZag** which can be found on CRAN.