Pseudo-marginal Piecewise-Deterministic Monte Carlo.

G. Vasdekis
Joint work with R. Everitt

UCL

30 June 2023, Monte Carlo Methods 2023, Paris

An Example: ABC

- Approximate Bayesian Computation is used when one cannot evaluate the likelihood but can sample from it.

An Example: ABC

- Approximate Bayesian Computation is used when one cannot evaluate the likelihood but can sample from it.

Algorithm

- *Sample* $\theta \sim pr(\cdot)$
- *Sample* $y \sim \pi(\cdot|\theta)$
- *If* $dist(y, y^*) < \epsilon$ *then accept* θ.

An Example: ABC

- Approximate Bayesian Computation is used when one cannot evaluate the likelihood but can sample from it.

Algorithm

- *Sample* $\theta \sim pr(\cdot)$
- *Sample* $y \sim \pi(\cdot|\theta)$
- *If* $dist(y, y^*) < \epsilon$ *then accept* θ.

- This draws samples from

$$\pi_{ABC}(\theta|y^*) = \int_y pr\ (\theta)\ 1_{dist(y,y^*)<\epsilon}\ \pi(y|\theta)dy.$$

An Example: ABC

- Approximate Bayesian Computation is used when one cannot evaluate the likelihood but can sample from it.

Algorithm

- *Sample* $\theta \sim pr(\cdot)$
- *Sample* $y \sim \pi(\cdot|\theta)$
- *If* $dist(y, y^*) < \epsilon$ *then accept* θ.

- This draws samples from

$$\pi_{ABC}(\theta|y^*) = \int_y pr\ (\theta)\ 1_{dist(y,y^*)<\epsilon}\ \pi(y|\theta)dy.$$

$$\pi_{ABC}(\theta|y^*) = \int_y pr\ (\theta)\ K(dist(y, y^*))\ \pi(y|\theta)dy$$

Pseudo-marginal Metropolis Hastings [AR09]

- When targeting π with Metropolis Hastings, given a current state θ we generate a $\theta' \sim q(\theta, \cdot)$ and the we accept or reject this new state with probability

$$a(\theta, \theta') = \min\left\{1, \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')}\right\}.$$

Pseudo-marginal Metropolis Hastings [AR09]

- When targeting π with Metropolis Hastings, given a current state θ we generate a $\theta' \sim q(\theta, \cdot)$ and the we accept or reject this new state with probability

$$a(\theta, \theta') = \min\left\{1, \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')}\right\}.$$

- No access to $\pi(\theta)$, but we access to an unbiased estimator $\tilde{\pi}(\theta)$ for all θ (i.e. for all θ, $\mathbb{E}[\tilde{\pi}(\theta)] = \pi(\theta)$).

Pseudo-marginal Metropolis Hastings [AR09]

- When targeting π with Metropolis Hastings, given a current state θ we generate a $\theta' \sim q(\theta, \cdot)$ and the we accept or reject this new state with probability

$$a(\theta, \theta') = \min\left\{1, \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')}\right\}.$$

- No access to $\pi(\theta)$, but we access to an unbiased estimator $\tilde{\pi}(\theta)$ for all θ (i.e. for all θ, $\mathbb{E}[\tilde{\pi}(\theta)] = \pi(\theta)$).
- Given a current state θ and a value of the estimator $\tilde{\pi}(\theta)$, we generate a $\theta' \sim q(\theta, \cdot)$ and a random value of the estimator $\tilde{\pi}(\theta')$. We accept or reject this new state θ' with probability

$$a(\theta, \theta') = \min\left\{1, \frac{\tilde{\pi}(\theta')q(\theta', \theta)}{\tilde{\pi}(\theta)q(\theta, \theta')}\right\}.$$

Pseudo-marginal Metropolis Hastings [AR09]

- When targeting π with Metropolis Hastings, given a current state θ we generate a $\theta' \sim q(\theta, \cdot)$ and the we accept or reject this new state with probability

$$a(\theta, \theta') = \min\left\{1, \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')}\right\}.$$

- No access to $\pi(\theta)$, but we access to an unbiased estimator $\tilde{\pi}(\theta)$ for all θ (i.e. for all θ, $\mathbb{E}[\tilde{\pi}(\theta)] = \pi(\theta)$).
- Given a current state θ and a value of the estimator $\tilde{\pi}(\theta)$, we generate a $\theta' \sim q(\theta, \cdot)$ and a random value of the estimator $\tilde{\pi}(\theta')$. We accept or reject this new state θ' with probability

$$a(\theta, \theta') = \min\left\{1, \frac{\tilde{\pi}(\theta')q(\theta', \theta)}{\tilde{\pi}(\theta)q(\theta, \theta')}\right\}.$$

- It works!

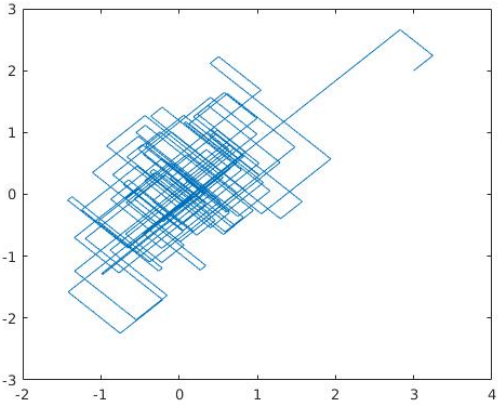


Figure: Zig-Zag Targeting a 2 dimensional Normal with positive correlations.
Run up to time 100

The Zig-Zag Sampler

Algorithm (Zig-Zag Sampler [BFR19])

1. *Start from point* $(\theta, v) = (\theta_1, ..., \theta_d; v_1, ..., v_d) \in \mathbb{R}^d \times \{-1, 1\}^d$.
2. *The process* (Θ_t, V_t) *moves according to the deterministic dynamics:* $\left\{ \frac{d}{dt}\Theta_t = v, t \geq 0 \right\}, \Theta_0 = \theta$ *and* $\{V_t = v \;\; , \; t \geq 0\}$.
3. *For all* $i \in \{1, ..., d\}$ *consider a non-homogeneous Poisson Process with intensity* $\{m_i(t) = \lambda_i(\Theta_t, v), t \geq 0\}$. *Suppose that the first arrival time is* T_i.
4. *Let* $T = \min\{T_i, i = 1, ..., d\}$ *and* $j = argmin\{T_i, i = 1, ..., d\}$.
5. *Set* $x = X_T$ *and* $v_j = -v_j$.
6. *Repeat from the first step.*

How to choose the rate λ?

Proposition (Bierkens-Fearnhead-Roberts 2019 [BFR19])

If $\mu(d\theta, dv) = \frac{1}{Z}\exp\{-U(\theta)\}(d\theta, dv)$

and consider a Zig-Zag process with rates

$$\lambda_i(\theta, v) = \max\{0, v_i \cdot \partial_i U(\theta)\} + \gamma_i(\theta)$$

where $v = (v_1, ..., v_d)$, ∂_i *the i-partial derivative and* γ_i *non-negative functions. The process has* μ *as unique invariant measure.*

Idea

- Target is

$$\pi(\theta) = Z^{-1} \int_{\Omega} \exp\{-U(\theta, \omega)\} p(\omega) d\omega.$$

Idea

- Target is

$$\pi(\theta) = Z^{-1} \int_{\Omega} \exp\{-U(\theta, \omega)\} p(\omega) d\omega.$$

- Joint Space

$$\pi(\theta, \omega) = \tilde{Z}^{-1} \exp\{-U(\theta, \omega)\} p(\omega).$$

Idea

- Target is

$$\pi(\theta) = Z^{-1} \int_{\Omega} \exp\{-U(\theta,\omega)\} p(\omega) d\omega.$$

- Joint Space

$$\pi(\theta,\omega) = \tilde{Z}^{-1} \exp\{-U(\theta,\omega)\} p(\omega).$$

- Keep ω fixed and explore the target $\exp\{-U(\cdot,\omega)\}$ with Zig-Zag.

Idea

- Target is

$$\pi(\theta) = Z^{-1} \int_{\Omega} \exp\{-U(\theta,\omega)\} p(\omega) d\omega.$$

- Joint Space

$$\pi(\theta,\omega) = \tilde{Z}^{-1} \exp\{-U(\theta,\omega)\} p(\omega).$$

- Keep ω fixed and explore the target $\exp\{-U(\cdot,\omega)\}$ with Zig-Zag.
- Add an extra refresh rate and when that clock rings update $\omega \sim p$.

Pseudo-marginal Zig-Zag

Algorithm

- *Assume we want to target* $\pi(\theta) = Z^{-1} \int_\Omega \exp\{-U(\theta,\omega)\} p(\omega) d\omega$, *where* $\theta \in \mathbb{R}^d$, $\omega \in \Omega$.
- *The state space is* $\mathbb{R}^d \times \Omega \times \{-1,+1\}^d$. *When the process is at* (θ,ω,v), *the* θ*-component tends to move along the line* $\{\theta + tv, t \geq 0\}$. *The velocity* v *and the variable* ω *remain fixed.*
- *We sample* T *the first arrival time of a non-homogeneous Poisson process with rate*

$$m(t) = \sum_{i=1}^{d} \lambda_i(\theta + tv, \omega, v)$$

where

$$\lambda_i(\theta,\omega,v) = \underbrace{\max\{0, v_i \cdot \partial_i U(\theta,\omega)\}}_{\textit{drift rate}} + \underbrace{\exp\{U(\theta,\omega)\} \cdot a(\theta) \cdot d^{-1}}_{\textit{refresh rate } \gamma}.$$

Here $a > 0$ *can be* **any** *function independent of* ω.

Pseudo-marginal Zig-Zag

Algorithm (Continued)

- *We move the process parallel to* v *until it reaches* $(\theta + Tv, \omega, v)$*. Set the new* $\theta = \theta + Tv$*.*
- *With probability*

$$\frac{\max\{0, v_i \cdot \partial_i U(\theta, \omega)\}}{\sum_{k=1}^{d} \lambda_k(\theta, \omega, v)}$$

we flip the sign of the i *coordinate of* v*.*

- *Or, with probability*

$$\frac{\exp\{U(\theta, \omega)\} \cdot a(\theta)}{\sum_{k=1}^{d} \lambda_k(\theta, \omega, v)},$$

we draw a new $\omega \sim p$*. We then start over from the second bullet.*

Invariance

Proposition ((Everitt-V. 2023+))

The pseudo-marginal Zig-Zag with rates given as before has the measure $\mu(dx, d\omega, dv) = \frac{1}{Z'} \exp\{-U(\theta, \omega)\} p(\omega) dx \; d\omega \; dv$ *as invariant.*

Ergodicity

Proposition (Ergodicity of Pseudo-marginal Zig-Zag)

Assume that $p(\omega) > 0$ for all ω. Assume that $U \in C^3$, and that for all ω,

$$\lim_{\|\theta\| \to \infty} U(\theta, \omega) = +\infty$$

and for some $\bar{\omega}$ the function $U(\cdot, \bar{\omega})$ has a non-degenerate local minimum. Then the Pseudo-marginal Zig-Zag is ergodic, i.e. for any $\theta \in \mathbb{R}^d, \omega \in \Omega, v \in \{-1, 1\}^d$,

$$\left\| \mathbb{P}_{\theta,\omega,v}\left((\Theta_t, \Omega_t, V_t) \in \cdot\right) - \mu(\cdot) \right\|_{TV} \xrightarrow{t \to \infty} 0.$$

Other type of Updates

- Instead of access to ω samples from p, one has access to samples from a Kernel $P(\theta, \omega, \cdot)$ that leaves p invariant. Then one can update the value of ω according to this P.
- All the previous results still hold.

Other type of Updates

- Instead of updating ω according to p, if one has access to a measure $p_{ref}(\theta, \cdot)$ on Ω, one can refresh

$$\omega \sim p_{ref}(\theta, \cdot).$$

Other type of Updates

- Instead of updating ω according to p, if one has access to a measure $p_{ref}(\theta, \cdot)$ on Ω, one can refresh

$$\omega \sim p_{ref}(\theta, \cdot).$$

The refresh rate for this scheme should be

$$\gamma(\theta, \omega) = \frac{p_{ref}(\theta, \omega)}{\exp\{-U(\theta, \omega)\}\, p(\omega)} a(\theta).$$

for any function $a \geq 0$ not depending on ω.

Other type of Updates

- Instead of updating ω according to p, if one has access to a measure $p_{ref}(\theta, \cdot)$ on Ω, one can refresh

$$\omega \sim p_{ref}(\theta, \cdot).$$

The refresh rate for this scheme should be

$$\gamma(\theta, \omega) = \frac{p_{ref}(\theta, \omega)}{\exp\{-U(\theta, \omega)\}\, p(\omega)} a(\theta).$$

for any function $a \geq 0$ not depending on ω.

- If we have access to the conditional measure

$$p_{ref}(\theta, \cdot) = \pi(\cdot|\theta) = \frac{1}{Z_\theta} \exp\{-U(\theta, \cdot)\} p(\cdot)$$

then (Gibb's Zig-Zag [SSLD22]) we can use it with refresh rate

$$\gamma(\theta, \omega) = constant.$$

Example 1, ABC

- In Approximate Bayesian Computation (ABC), the target is

$$\pi_{ABC}(\theta|y^*) = \int_{\omega\in\Omega} pr(\theta)K(\|f\ (\theta,\omega) - y^*\|_2)p(\omega)d\omega,$$

for some kernel function K, where we assume that for all θ,

$$f\ (\theta,\omega) \sim \pi(\cdot|\theta), \text{ when } \omega \sim p.$$

Example 1, ABC

- In Approximate Bayesian Computation (ABC), the target is

$$\pi_{ABC}(\theta|y^*) = \int_{\omega\in\Omega} pr(\theta)K(\|f\ (\theta,\omega) - y^*\|_2)p(\omega)d\omega,$$

  for some kernel function K, where we assume that for all θ,

$$f\ (\theta,\omega) \sim \pi(\cdot|\theta),\ \text{when } \omega \sim p.$$

- If we have access to f and $\nabla_\theta f$ we can run Pseudo-marginal Zig-Zag and target $\pi_{ABC}(\cdot|y^*)$.

Example 1, ABC

- In Approximate Bayesian Computation (ABC), the target is

$$\pi_{ABC}(\theta|y^*) = \int_{\omega\in\Omega} pr(\theta) K(\|f\ (\theta,\omega) - y^*\|_2) p(\omega) d\omega,$$

  for some kernel function K, where we assume that for all θ,

$$f\ (\theta,\omega) \sim \pi(\cdot|\theta), \text{ when } \omega \sim p.$$

- If we have access to f and $\nabla_\theta f$ we can run Pseudo-marginal Zig-Zag and target $\pi_{ABC}(\cdot|y^*)$.
- The rate of switching the direction from v_i to $-v_i$ will be

$$\max\{0, v_i \cdot \partial_i\left(-\log pr(\theta) - \log K(\|f\ (\theta,\omega) - y^*\|_2)\right)\}$$

Example 1, ABC

- In Approximate Bayesian Computation (ABC), the target is

$$\pi_{ABC}(\theta|y^*) = \int_{\omega\in\Omega} pr(\theta)K(\|f\ (\theta,\omega) - y^*\|_2)p(\omega)d\omega,$$

for some kernel function K, where we assume that for all θ,

$$f\ (\theta,\omega) \sim \pi(\cdot|\theta), \text{ when } \omega \sim p.$$

- If we have access to f and $\nabla_\theta f$ we can run Pseudo-marginal Zig-Zag and target $\pi_{ABC}(\cdot|y^*)$.
- The rate of switching the direction from v_i to $-v_i$ will be

$$\max\{0, v_i \cdot \partial_i\left(-\log pr(\theta) - \log K(\|f\ (\theta,\omega) - y^*\|_2)\right)\}$$

and the rate to update $\omega \sim p$ will be

$$\gamma(\theta) = \frac{1}{K(\|f\ (\theta,\omega) - y^*\|_2)} \cdot a(\theta).$$

Specific Kernels K

Exponential Kernel: $K(s) = \exp\{-\frac{1}{\epsilon}s\}$.

Specific Kernels K

Exponential Kernel: $K(s) = \exp\{-\frac{1}{\epsilon}s\}$.
Then the rate of switching v_i to $-v_i$ is

$$\max\left\{0, v_i\left(-\frac{pr\,'(\theta)}{pr(\theta)} + \frac{1}{\epsilon}\left\langle \frac{f\,(\theta,\omega) - y^*}{\|f\,(\theta,\omega) - y^*\|_2}, \partial_{\theta_i} f\,(\theta,\omega)\right\rangle\right)\right\}$$

Specific Kernels K

Exponential Kernel: $K(s) = \exp\{-\frac{1}{\epsilon}s\}$.
Then the rate of switching v_i to $-v_i$ is

$$\max\left\{0, v_i\left(-\frac{pr\,'(\theta)}{pr(\theta)} + \frac{1}{\epsilon}\left\langle\frac{f\ (\theta,\omega)-y^*}{\|f\ (\theta,\omega)-y^*\|_2}, \partial_{\theta_i} f\ (\theta,\omega)\right\rangle\right)\right\}$$

and the rate to update ω

$$\gamma(\theta,\omega) = \exp\left\{\frac{1}{\epsilon}\left(\|f\ (\theta,\omega)-y^*\|_2 - \|f\ (\theta,0)-y^*\|_2\right)\right\} a(\theta)$$

a any function.

Multiple ω's

- Instead of one ω, we can use multiple ω's:

$$\pi_{ABC}(\theta|y^*) = \int_{\omega_1,..,\omega_N} pr(\theta) K\left(\left\|\frac{1}{N}\sum_{k=1}^{N} f\left(\theta, \omega_k\right) - y^*\right\|_2\right) p(\omega) d\omega,$$

- In the previous equations $f\left(\theta, \omega\right)$ is replaced by $\frac{1}{N}\sum_{k=1}^{N} f\left(\theta, \omega_k\right)$

Simulations

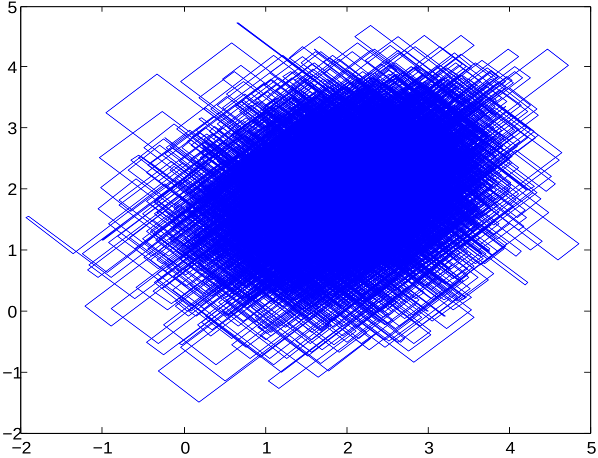


Figure: Traceplot of Pseudo-marginal ABC Zig-Zag. $N(0, I_2)$ prior and the model is $N(\theta_1, \theta_2, (1\ 0.2; 0.2\ 1))$. Observed data $y^* = [4\ 4]$ and $\epsilon = 0.1$. $N = 10^4$ switches of direction.

Simulations

Algorithm	ESS/minute
MALA	971.1
PM-ZZ	858.3
PM-SUZZ(0)	**1159.7**

- Effective Sample size of Pseudo-marginal Zig-Zag (PM-ZZ) and PM-Random Walk Metropolis (PM-RWM). $N(0, I_2)$ prior and the model is $N(\theta_1, \theta_2, (1\ 0.2; 0.2\ 1))$. Observed data $y^* = [4\ 4]$ and $\epsilon = 0.1$. $N = 10^4$ switches of direction

Example 2

- **Model:** $y_i \sim N(\theta_1 + \theta_2 + \beta\theta_2^2, 1),\ i = 1, ..., n.$
  Prior: $\theta_1, \theta_2 \sim N(0, 5)$ i.i.d.

Example 2

- **Model:** $y_i \sim N(\theta_1 + \theta_2 + \beta\theta_2^2, 1)$, $i = 1, ..., n$.
  Prior: $\theta_1, \theta_2 \sim N(0, 5)$ i.i.d.
- If $a = \theta_1 + \theta_2$ and $b = \theta_1 - \theta_2$ then a is much more informative than b.

Example 2

- **Model:** $y_i \sim N(\theta_1 + \theta_2 + \beta\theta_2^2, 1)$, $i = 1, ..., n$.
  Prior: $\theta_1, \theta_2 \sim N(0, 5)$ i.i.d.
- If $a = \theta_1 + \theta_2$ and $b = \theta_1 - \theta_2$ then a is much more informative than b.
- Use Pseudo-marginal Zig-Zag, with b **in the place of** ω. For fixed b/ω **explore the informative** a **using Zig-Zag**.

Example 2

- **Model:** $y_i \sim N(\theta_1 + \theta_2 + \beta\theta_2^2, 1)$, $i = 1, ..., n$.
  Prior: $\theta_1, \theta_2 \sim N(0, 5)$ i.i.d.
- If $a = \theta_1 + \theta_2$ and $b = \theta_1 - \theta_2$ then a is much more informative than b.
- Use Pseudo-marginal Zig-Zag, with b **in the place of** ω. For fixed b/ω **explore the informative** a **using Zig-Zag**.
- However, conditional posterior

$$\pi(b|a) \propto \exp\left\{-\frac{b^2}{2 \cdot 5}\right\} \cdot \exp\left\{-\sum_{i=1}^{n}\left(a + \frac{\beta}{4}a^2 + \frac{\beta}{4}b^2 - \frac{\beta}{2}ab - y_i\right)^2\right.$$

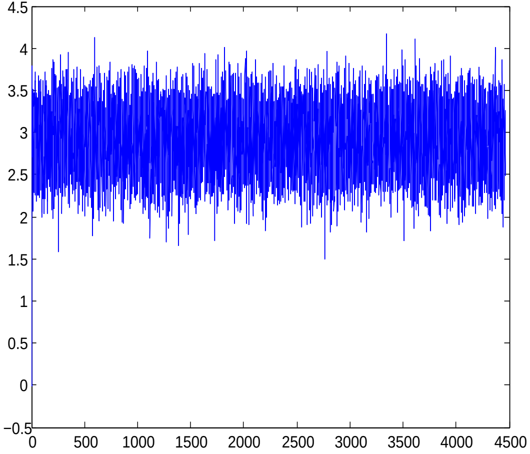
Example 2

- **Model:** $y_i \sim N(\theta_1 + \theta_2 + \beta\theta_2^2, 1)$, $i = 1, ..., n$.
  Prior: $\theta_1, \theta_2 \sim N(0, 5)$ i.i.d.
- If $a = \theta_1 + \theta_2$ and $b = \theta_1 - \theta_2$ then a is much more informative than b.
- Use Pseudo-marginal Zig-Zag, with b **in the place of** ω. For fixed b/ω **explore the informative** a **using Zig-Zag**.
- However, conditional posterior

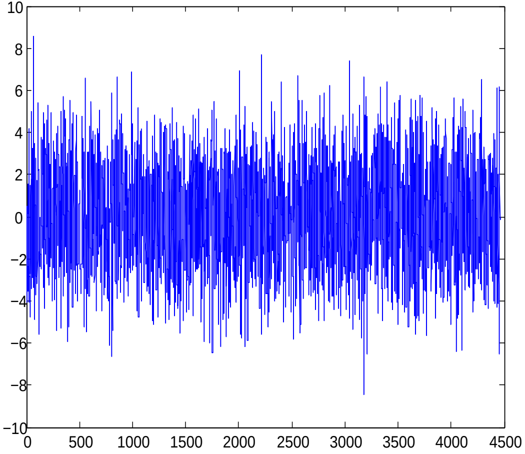
$$\pi(b|a) \propto \exp\left\{-\frac{b^2}{2 \cdot 5}\right\} \cdot \exp\left\{-\sum_{i=1}^{n}\left(a + \frac{\beta}{4}a^2 + \frac{\beta}{4}b^2 - \frac{\beta}{2}ab - y_i\right)^2\right.$$

- Instead one could update $b \sim N(0, 5)$.

Trace Plot for a



Trace Plot for b



ESS

Algorithm	ESS/minute
MALA	1972.0
PM-ZZ	4296.1
Gibbs-ZZ	4132.9
PM-SUZZ(0)	4569.0
Gibbs-SUZZ(0)	**4788.3**

- Effective Sample size of MALA, Pseudo-marginal Zig-Zag (PM-ZZ), Gibbs Zig-Zag, Pseudo-marginal Speed Up Zig-Zag (PM-SUZZ), Random Walk Metropolis (RWM) and Pseudo-marginal Speed Up Zig-Zag (PM-SUZZ), and Gibbs Speed Up Zig-Zag (Gibbs-SUZZ) for Example 2.

Thank you for your attention!

C. Andrieu and G. O. Roberts.
The pseudo-marginal approach for efficient monte carlo computations.
Ann. Statist., 37(2):697–725, 04 2009.

Joris Bierkens, Paul Fearnhead, and Gareth Roberts.
The zig-zag process and super-efficient sampling for bayesian analysis of big data.
Ann. Statist., 47(3):1288–1320, 06 2019.

Matthias Sachs, Deborshee Sen, Jianfeng Lu, and David Dunson.
Posterior Computation with the Gibbs Zig-Zag Sampler.
Bayesian Analysis, pages 1 – 19, 2022.

G. V. and G. O. Roberts.
Speed Up Zig-Zag, 2023+.
To appear *Annals of Applied Probability*. Available at https://arxiv.org/abs/2103.16620.

[VR23], [SSLD22]