# Gradient-Based Markov Chain Monte Carlo for Bayesian Inference With Non-Differentiable Priors

*Torben Sell*

University of Edinburgh

June 30, 2023

# About me

Further research interests:
Function space inference (Bayesian neural networks), statistical learning (classification), missing data, filtering

https://www.maths.ed.ac.uk/~tsell/

# Co-authors

Jacob Vorstrup Goldman

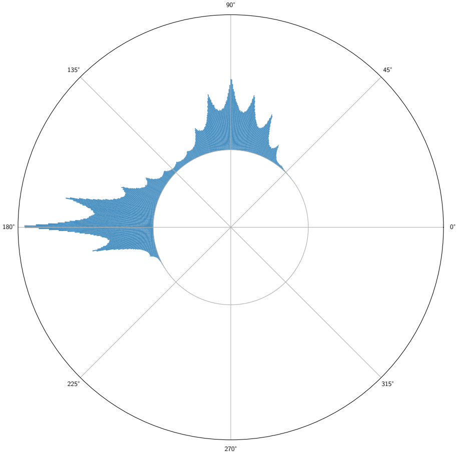
Sumeetpal Sidhu Singh

# Motivation

Non-differentiable distributions appear in e.g. imaging, genetics, biology.

Consider for example $\pi(x) \propto \exp\{-|x|\}$ or $\pi(x) \propto |x+\epsilon|^{p-1/2} K_{p-1/2}(|x|+\epsilon)$.

# Motivation



# Motivation

Problem: Can't easily use gradient-based sampling methods.

# Solution I

Target a different distribution.

# Solution I

Target a different distribution.

Instead of $\pi \propto \exp\{-g(x)\}$, target $\pi^\lambda \propto \exp\{-g^\lambda(x)\}$, where $g^\lambda(x) = \inf_z \left[g(z) + \frac{1}{2\lambda}\|x - z\|^2\right]$ is the Moreau-Yosida envelope of $g$.

# Solution I

Target a different distribution.

Instead of $\pi \propto \exp\{-g(x)\}$, target $\pi^\lambda \propto \exp\{-g^\lambda(x)\}$, where $g^\lambda(x) = \inf_z \left[g(z) + \frac{1}{2\lambda}\|x-z\|^2\right]$ is the Moreau-Yosida envelope of $g$.

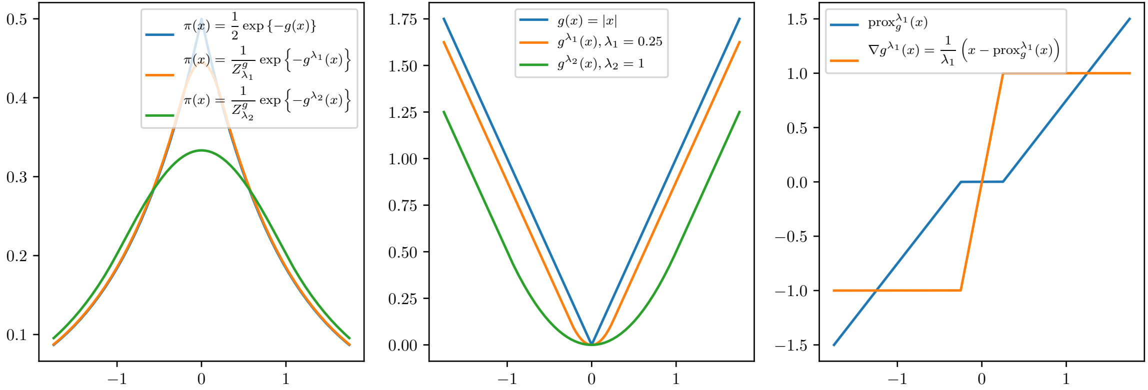
If $g$ is lower semi-continuous and convex, then for $\lambda > 0$, $\nabla g^\lambda$ is $1/\lambda$-Lipschitz continuous, and given by

$$\nabla g^\lambda(x) = \frac{1}{\lambda}\left(x - \mathrm{prox}_g^\lambda(x)\right),$$

where

$$\mathrm{prox}_g^\lambda(x) = \arg\min_u \left[g(u) + \frac{1}{2\lambda}\|x-u\|^2\right].$$

# Solution I


$\pi(x) = \frac{1}{2}\exp\{-g(x)\}$
$\pi(x) = \frac{1}{Z^g_{\lambda_1}}\exp\left\{-g^{\lambda_1}(x)\right\}$
$\pi(x) = \frac{1}{Z^g_{\lambda_2}}\exp\left\{-g^{\lambda_2}(x)\right\}$
$g(x) = |x|$
$g^{\lambda_1}(x), \lambda_1 = 0.25$
$g^{\lambda_2}(x), \lambda_2 = 1$
$\mathrm{prox}_g^{\lambda_1}(x)$
$\nabla g^{\lambda_1}(x) = \frac{1}{\lambda_1}\left(x - \mathrm{prox}_g^{\lambda_1}(x)\right)$
0.1
0.2
0.3
0.4
0.5
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
−1
0
1

# Solution I

Can correct with Metropolis-Hastings or accept the error.

# Solution I

Can correct with Metropolis-Hastings or accept the error.

## Theorem

*Let $g = -\log \pi$ be the negative logarithm of a probability density function, with $g$ being a proper lower semi-continuous convex function, and $L$-Lipschitz. Let $g^\lambda$ be the Moreau-Yosida envelope to $g$, and let $\pi^\lambda(x) = \exp(-g^\lambda(x))/(\int \exp(-g^\lambda(z))dz)$ be a probability density function. Then for any $\pi$- and $\pi^\lambda$-integrable $f : \mathcal{X} \to \mathbb{R}$:*

$$|\mathbb{E}_{\pi^\lambda}(f) - \mathbb{E}_\pi(f)| \leq (\exp(L^2\lambda) - 1)\mathbb{E}_{\pi^\lambda}(|f|) \tag{1}$$

$$|\mathbb{E}_{\pi^\lambda}(f) - \mathbb{E}_\pi(f)| \leq (\exp(L^2\lambda) - 1)\mathbb{E}_\pi(|f|). \tag{2}$$

*The same inequalities hold if $g = g_1 + g_2$ with a convex and Lipschitz-continuous $g_1$ and a differentiable (but not necessarily Lipschitz-continuous) $g_2$.*

# Solution I

Various algorithms exist:

- MY-ULA (= Unadjusted Langevin Algorithm targeting $\pi^\lambda$)
- MY-UULA (= Unadjusted Underdamped Langevin Algorithm targeting $\pi^\lambda$)
- SK-ROCK (= stabilised integrator targeting $\pi^\lambda$)
- pMALA (= proximal MALA, targeting $\pi$, MY-ULA + Metropolis Hastings)

# Solution II

Can we use MYEs in PDMPs?

# Solution II

Can we use MYEs in PDMPs? Even better...

# Solution II

Can we use MYEs in PDMPs? Even better...

E.g. Zig-Zag Sampler (ZZ): augment state space $\mathcal{X} \subset \mathbb{R}^d$ with $v \in \{-1, 1\}^d$; target the joint distribution $p(x, v) = \pi(x)\mathcal{U}(v)$.

# Solution II

Can we use MYEs in PDMPs? Even better...

E.g. Zig-Zag Sampler (ZZ): augment state space $\mathcal{X} \subset \mathbb{R}^d$ with $v \in \{-1, 1\}^d$; target the joint distribution $p(x, v) = \pi(x)\mathcal{U}(v)$.

$(z_t)_{t\geq 0} = (x_t, v_t)_{t\geq 0}$ follows ODE $(\dot{x}_t, \dot{v}_t) = (v_t, 0)$ and switches $i$th velocity with rate

$$\rho^i_{ZZ}(t) := \rho^i_{ZZ}(t; x, v) = \max\left\{0, \frac{\partial}{\partial x_i} U(x + v \cdot t) \cdot v_i\right\}.$$

# Solution II

$$A_0 = \left\{ x \in \mathcal{X} \mid \exists\ i \text{ such that } \frac{\partial U}{\partial x_i} \text{ does not exist.} \right\}$$

# Solution II

$$A_0 = \left\{ x \in \mathcal{X} \mid \exists\ i \text{ such that } \frac{\partial U}{\partial x_i} \text{ does not exist.} \right\}$$

## Lemma

*Consider a distribution $\pi(x)\mathcal{U}(v)$ that is differentiable in $x$ outside of $A_0$. If $A_0$ is a Lebesgue null-set, the Zig-Zag process with generator given by $\mathcal{L}_{ZZ} f(x,v) = \langle \nabla_x f(x), v \rangle + \sum_{i=1}^n \rho_{ZZ}^i(t)[f(x, \mathcal{F}_i v) - f(x,v)]$ has invariant distribution $\pi(x)\mathcal{U}(v)$.*

# Solution II

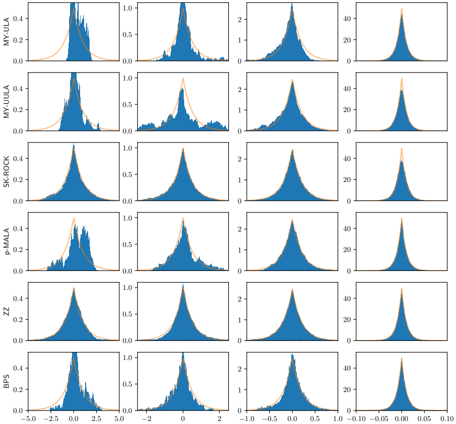
$$A_0 = \left\{ x \in \mathcal{X} \mid \exists\ i \text{ such that } \frac{\partial U}{\partial x_i} \text{ does not exist.} \right\}$$

## Lemma

*Consider a distribution $\pi(x)\mathcal{U}(v)$ that is differentiable in $x$ outside of $A_0$. If $A_0$ is a Lebesgue null-set, the Zig-Zag process with generator given by $\mathcal{L}_{ZZ} f(x,v) = \langle \nabla_x f(x), v \rangle + \sum_{i=1}^n \rho^i_{ZZ}(t)[f(x, \mathcal{F}_i v) - f(x,v)]$ has invariant distribution $\pi(x)\mathcal{U}(v)$.*

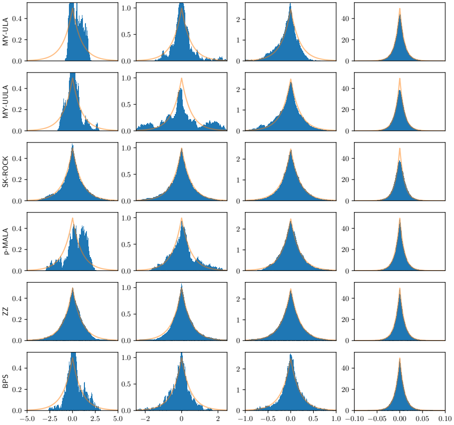
A similar result holds for the Bouncy Particle Sampler (BPS).

# Anisotropic Laplace




$$\pi(x) \propto \prod_{j=1}^{100} \exp(-j \times |x_j|)$$

# Anisotropic Laplace

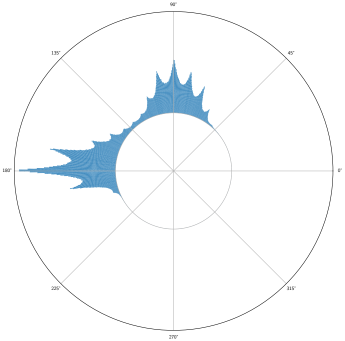



$$\pi(x) \propto \prod_{j=1}^{100} \exp(-j \times |x_j|)$$

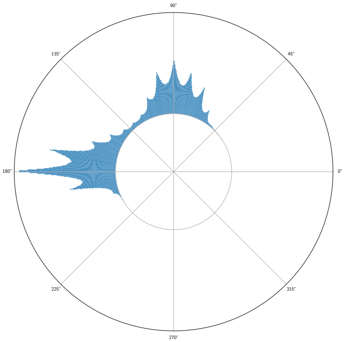
| Algorithm | MY-ULA | MY-UULA | SK-ROCK |
|---|---|---|---|
| $\beta = 1$ | 2.0 | 2.3 | 6.0 |
| $\beta = 100$ | 50.5 | 218.3 | 4197.4 |
| Algorithm | pMALA | BPS | ZZ |
| $\beta = 1$ | 1.7 | 3.0 | 24.9 |
| $\beta = 100$ | 182.9 | 755.5 | 2037.4 |

Effective sample size per second for the fastest and slowest mixing dimensions of the anisotropic Laplace obtained from long runs of the respective algorithms. Recall that the first three algorithms are asymptotically biased, while the last three are asymptotically exact.

# Ants



# Ants

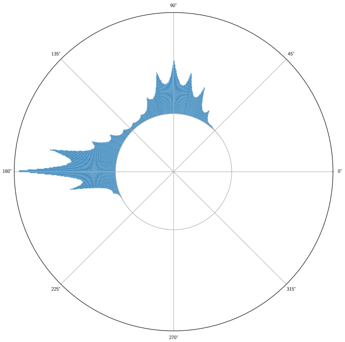



Observations: $y_i \in [0, 2\pi)$, $i = 1 \ldots 253$. Likelihood is a mixture of two wrapped asymmetric Laplace distributions.

$$\theta = \begin{cases} y - \mu_i, & \text{for } y - \mu_i > 0 \\ y - \mu_i + 2\pi, & \text{for } y - \mu_i \leq 0 \end{cases}$$

$$L(\theta|\mu_i, \lambda_i, \kappa_i) = \frac{\lambda_i \kappa_i}{1 + \kappa_i^2} \left( \frac{e^{-\lambda_i \kappa_i \theta}}{1 - e^{-2\pi \lambda_i \kappa_i}} + \frac{e^{(\lambda_i/\kappa_i)\theta}}{e^{2\pi(\lambda_i/\kappa_i)} - 1} \right)$$

# Ants



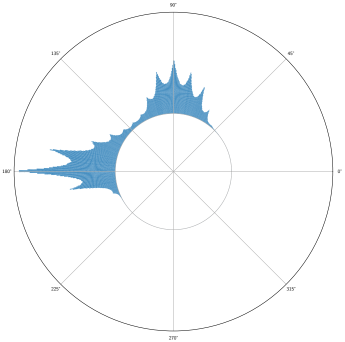

Observations: $y_i \in [0, 2\pi)$, $i = 1 \ldots 253$. Likelihood is a mixture of two wrapped asymmetric Laplace distributions.

$$\theta = \begin{cases} y - \mu_i, & \text{for } y - \mu_i > 0 \\ y - \mu_i + 2\pi, & \text{for } y - \mu_i \leq 0 \end{cases}$$

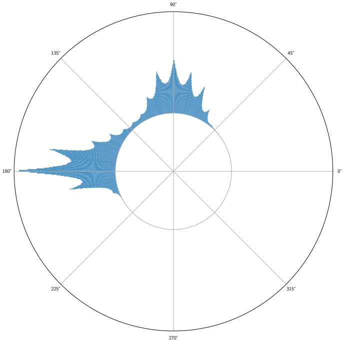
$$L(\theta|\mu_i, \lambda_i, \kappa_i) = \frac{\lambda_i \kappa_i}{1 + \kappa_i^2} \left( \frac{e^{-\lambda_i \kappa_i \theta}}{1 - e^{-2\pi\lambda_i\kappa_i}} + \frac{e^{(\lambda_i/\kappa_i)\theta}}{e^{2\pi(\lambda_i/\kappa_i)} - 1} \right)$$

Priors: $\mu_i \sim \mathcal{U}[0, 2\pi]$, $\lambda_i \sim$Exp$(1)$, $\kappa_i \sim$Gamma$(2, 1/2)$, $\rho \sim$Beta$(100, 100)$.

# Ants



# Ants



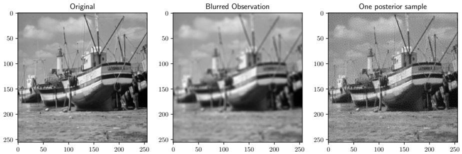

| Algorithm | $\mu_1$ | $\lambda_1$ | $\kappa_1$ | $\rho$ |
|---|---|---|---|---|
| BPS | 3.36 | 164.80 | 6.29 | 2095.44 |
| ZZ | 0.66 | 39.53 | 1.12 | 537.33 |
| RWMH | 2.88 | 37.01 | 4.52 | 1153.13 |

ESS/s for different variables. The ESS/s for the variables from the second mixture are similar, as is expected due to the mixture components being indistinguishable from one another.

# Imaging

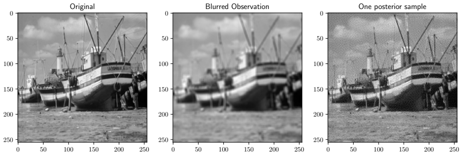
# Imaging

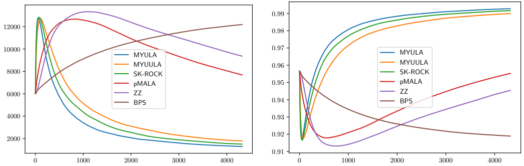
$$\pi(x) \propto \exp\left(-\frac{1}{2\sigma^2}\|Hx - y\|_2^2 - \alpha TV(x)\right)$$



# Imaging

$$\pi(x) \propto \exp\left(-\frac{1}{2\sigma^2}\|Hx - y\|_2^2 - \alpha TV(x)\right)$$




Left: The MSE of the mean estimates, estimated every $10$ seconds. Right: The SSIM of the mean estimates, estimated every $10$ seconds.

# Conclusions

**MYEs**

- are better at estimating high-dimensional distributions
- require log-concavity (of the non-diff. part)
- require calculating/ knowing the proximal operator
- can add a MH step

**PDMPs**

- can more easily adapt to anisotropic targets
- allow subsampling and parallelisation
- are inherently exact
- require calculating event rates

Thank you for listening!