Optimal scaling for Proximal MALA

Francesca R. Crucinio [1] Alain Durmus [2] Pablo Jiménez [3]
and Gareth O. Roberts [4]

[1]CREST, ENSAE Paris

[2]CMAP, École Polytechnique

[3]Sorbonne Université and Université Paris Cité

[4]Department of Statistics, University of Warwick

31 October 2023

Outline

Markov chain Monte Carlo

Aim: sample from a probability distribution π on $\mathbb{R}^d$ and approximate expectations w.r.t. π

$$\int f(x)\pi(x)\mathrm{d}x$$

Motivation: compute posterior expectations in Bayesian inference

Metropolis-Hastings

Build a Markov chain $(X_k)_{k\geq 0}$ such that, given the current state X_k

- Sample $Y \sim q(X_k, \cdot)$
- with probability

$$\frac{\pi(Y)q(Y, X_k)}{\pi(X_k)q(X_k, Y)} \wedge 1.$$

set $X_{k+1} = Y$, otherwise, set $X_{k+1} = X_k$.

- **RWM:** $q(X_k, \cdot) = \mathcal{N}(\cdot; X_k, \sigma^2 I_d)$
- **MALA:** $q(X_k, \cdot) = \mathcal{N}(\cdot; X_k + \frac{\sigma^2}{2}\nabla \log \pi(X_k), \sigma^2 I_d)$

Outline

Proximal MCMC: Idea

- Build a proposal q using a **proximity map**.
- Main idea introduced in (Pereyra, 2016) then refined in (Durmus et al., 2018).
- Can be applied to non-differentiable targets π

Proximity map

Proximity map

For g convex, proper and lower semi-continuous and $\lambda > 0$

$$\text{prox}_g^{\lambda}(\boldsymbol{x}) := \arg\min_{\boldsymbol{u}\in\mathbb{R}^d} \left[g(\boldsymbol{u}) + \frac{\|\boldsymbol{u}-\boldsymbol{x}\|^2}{2\lambda} \right].$$

Moves points in the direction of the minimum of g.

Take $\pi(\boldsymbol{x}) \propto \exp(-g(\boldsymbol{x}))$. For each $\lambda > 0$ we can build an approximation π_λ of π.

Moreau-Yoshida envelope

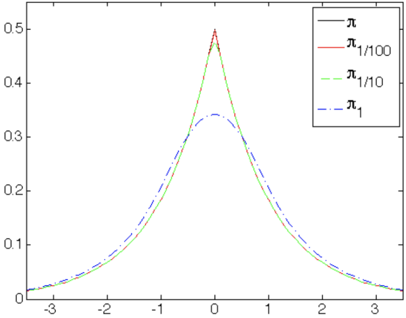

Figure: Moreau-Yoshida envelope for the Laplace distribution $\pi(x) \propto \exp(-|x|)$ (Pereyra, 2016).

Moreau-Yoshida envelope

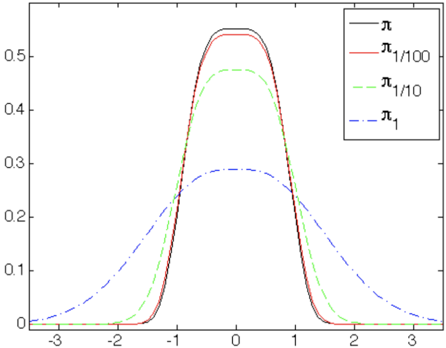

Figure: Moreau-Yoshida envelope for $\pi(x) \propto \exp(-x^4)$ (Pereyra, 2016).

Proximal MCMC

Since π_λ is differentiable for any $\lambda > 0$, we can build a MALA-like algorithm to sample from π_λ :

$$\begin{aligned} q(X_k, Y) &= \mathcal{N}\left(Y; X_k + \frac{\sigma^2}{2} \nabla \log \pi_\lambda(X_k), \sigma^2 I_d\right) \\ &= \mathcal{N}\left(Y; X_k + \frac{\sigma^2}{2\lambda}\left(\operatorname{prox}_g^\lambda(X_k) - X_k\right), \sigma^2 I_d\right), \end{aligned}$$

and then accept/reject using the usual MH step

$$\frac{\pi(Y) q(Y, X_k)}{\pi(X_k) q(X_k, Y)} \wedge 1.$$

Outline

Our aim

Investigate the optimal scaling properties of proximal MCMC for targets

$$\pi_d(\boldsymbol{x}) = \prod_{i=1}^{d} \pi(x_i) \propto \exp\left(-\sum_{i=1}^{d} g(x_i)\right)$$

when both $\sigma_d \to 0$ and $\lambda_d \to 0$ for

- differentiable g
- the Laplace distribution, $g(x) = |x|$

Notation

We consider

$$\sigma_d^2 = \frac{\ell^2}{d^\alpha}, \qquad \lambda_d = \frac{c^2}{2d^\beta}$$

for some $\alpha, \beta > 0$.

Differentiable targets

Given $\pi_d(\boldsymbol{x}) \propto \exp\left(-\sum_{i=1}^{d} g(x_i)\right)$, the proposal for proximal MCMC is

$$q(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{d} \mathcal{N}\left(y_i; x_i + \frac{\sigma_d^2}{2} g'\left(\mathrm{prox}_g^{\lambda_d}(x_i)\right), \sigma_d^2 I_d\right).$$

- $c = 0$ we get $\boldsymbol{x} + \frac{\sigma_d^2}{2} g'\left(\mathrm{prox}_g^{\lambda_d}(\boldsymbol{x})\right) = \boldsymbol{x} + \frac{\sigma_d^2}{2} g'(\boldsymbol{x})$, i.e. **MALA**

Differentiable targets – g' Lipschitz

- $\sigma_d^2 = \ell^2 d^{-1/3},\ \lambda_d = c^2 d^{-\beta}/2,\ \beta > 1/3$
  - acceptance rate $a(\ell) = 2\Phi\left(-\frac{\ell^3 K_1}{2}\right)$ and $K_1 = K_{MALA}$
  - speed $h(\ell) = \ell^2 a(\ell)$ maximized at $a(\ell) = 0.574$
- $\sigma_d^2 = \ell^2 d^{-1/3},\ \lambda_d = c^2 d^{-1/3}/2$
  - define $r := c^2/\ell^2 > 0$, acceptance rate $a(\ell, r) = 2\Phi\left(-\frac{\ell^3 K_2(r)}{2}\right)$ and $K_2^2(r) \geq K_{MALA}^2$ and increasing
  - speed $h(\ell, r) = \ell^2 a(\ell, r)$ maximized at $a(\ell, r) = 0.574$

Differentiable targets – g' **not** Lipschitz

- $\sigma_d^2 = \ell^2 d^{-1/2}, \; \lambda_d = c^2 d^{-1/4}/2$
  - define $r := c^2/\ell > 0$, acceptance rate $a(\ell, r) = 2\Phi\left(-\frac{\ell^2 K_3(r)}{2}\right)$ and $K_3^2(r) = r^2 \mathbb{E}_X\left[\frac{[g''(X)g'(X)]^2}{4}\right]$
  - speed $h(\ell, r) = \ell^2 a(\ell, r)$ maximized at $a(\ell) = 0.452$

Differentiable targets – g' Lipschitz

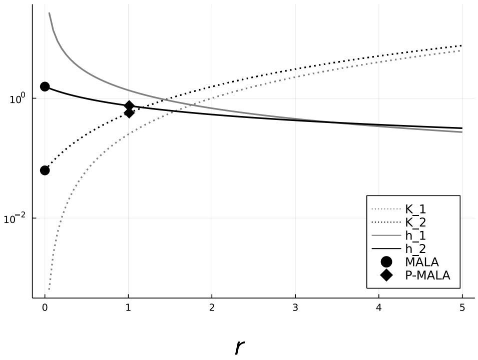

Figure: Speed of Langevin diffusion as a function of $r = c^2/\ell^2$ for a Gaussian target.

Differentiable targets – g' **not** Lipschitz

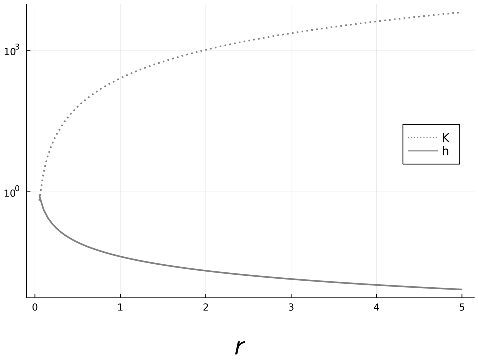

Figure: Speed of Langevin diffusion as a function of $r = c^2/\ell^2$ for a light tail target $\pi(x) \propto \exp\left(-x^6\right)$.

Laplace target

Given $\pi_d(\boldsymbol{x}) \propto \exp\left(-\sum_{i=1}^d |x_i|\right)$, the proposal for proximal MCMC is

$$q(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{d} \mathcal{N}\left(y_i; f(x_i), \sigma_d^2\right),$$

where

$$f(x_i) = x_i - \frac{\sigma_d^2}{2} \operatorname{sgn}(x_i) \mathbb{1}_{|x_i| \geq \lambda_d}(x_i) - \frac{\sigma_d^2}{2\lambda_d} x_i \mathbb{1}_{|x_i^d| < \lambda_d}(x_i)$$

Laplace target

- $\sigma_d^2 = \ell^2 d^{-2/3}, \; \lambda_d = c^2 d^{-\beta}/2, \; \beta > 2/3$
  - acceptance rate $a(\ell) = 2\Phi\left(-\frac{\ell^{3/2}}{(72\pi)^{1/4}}\right)$
  - speed $h(\ell) = \ell^2 a(\ell)$ maximized at $a(\ell) = 0.360$

The case $c = 0$ corresponds to a subgradient version of MALA.

Take home messages

When implementing proximal MALA we need to take into consideration: efficiency, robustness (i.e. when is the Markov chain geometrically ergodic?), cost of obtaining gradients

We have not explored the robustness of proximal MALA, however,

- if MALA is geometrically ergodic and $\nabla \log \pi(x)$ is cheaper than $\mathrm{prox}_g^{\lambda}(x) \rightarrow$ use MALA
- in cases in which $\nabla \log \pi(x)$ is more expensive than $\mathrm{prox}_g^{\lambda}(x)$ $\rightarrow$ use proximal MALA with λ as small as possible
- for light tail distributions use proximal MALA

Thank you!

Bibliography I

Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

Marcelo Pereyra. Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, 26(4):745–760, 2016.