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

21 March 2023

Outline

1. Markov chain Monte Carlo
2. Proximal MCMC
3. Optimal Scaling
4. Optimal scaling for Proximal MALA

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

Markov chain Monte Carlo

Idea: build a Markov chain $(X_k)_{k\geq 0}$ such that

- π is an invariant measure for the Markov chain
- a law of large numbers hold

$$\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} f(X_i) = \int f(x)\pi(x)\mathrm{d}x$$

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 $X_{k+1} = X_k$.

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 $X_{k+1} = X_k$.

- ▶ does not require normalising constant of π!

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

1. Markov chain Monte Carlo
2. Proximal MCMC
3. Optimal Scaling
4. Optimal scaling for Proximal MALA

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.

Moreau-Yoshida envelope

Take $\pi(\boldsymbol{x}) \propto \exp(-g(\boldsymbol{x}))$. We can define a family of distributions

$$\pi_\lambda(\boldsymbol{x}) \propto \exp\left[-g\left(\operatorname{prox}_g^\lambda(\boldsymbol{x})\right)\right] \exp\left[-\|\operatorname{prox}_g^\lambda(\boldsymbol{x}) - \boldsymbol{x}\|^2/(2\lambda)\right]$$

Moreau-Yoshida envelope

Take $\pi(\boldsymbol{x}) \propto \exp(-g(\boldsymbol{x}))$. We can define a family of distributions

$$\pi_\lambda(\boldsymbol{x}) \propto \exp\left[-g\left(\operatorname{prox}_g^\lambda(\boldsymbol{x})\right)\right] \exp\left[-\|\operatorname{prox}_g^\lambda(\boldsymbol{x}) - \boldsymbol{x}\|^2/(2\lambda)\right]$$

- converge (pointwise, in TV, ...) to π as $\lambda \to 0$
- π_λ is continuously differentiable with $\nabla \log \pi_\lambda(\boldsymbol{x}) = \lambda^{-1}(\boldsymbol{x} - \operatorname{prox}_g^\lambda(\boldsymbol{x}))$
- π_λ has at most Gaussian tails

Moreau-Yoshida envelope

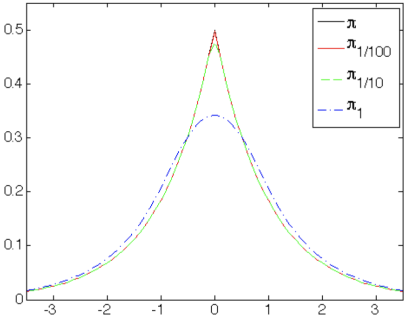

Figure: Moreau-Yoshida envelope for the Laplace distribution $\pi(x) \propto \exp(-|x|)$ (Pereyra, 2016).

Moreau-Yoshida envelope

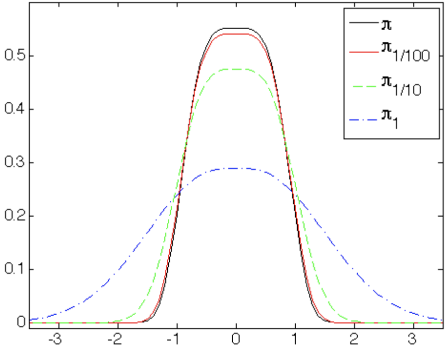

Figure: Moreau-Yoshida envelope for $\pi(x) \propto \exp(-x^4)$ (Pereyra, 2016).

Proximal MCMC

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

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

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

1. Markov chain Monte Carlo
2. Proximal MCMC
3. Optimal Scaling
4. Optimal scaling for Proximal MALA

Main Idea

Given a proposal

$$q(X_k, \cdot) = \mathcal{N}(\cdot; f(X_k), \sigma^2 I_d)$$

how should we chose σ to obtain good performances?

In particular, how should σ scale with the dimension d of the support of π?

Set up

- $\pi_d(\boldsymbol{x}) = \prod_{i=1}^{d} \pi(x_i) \propto \exp\left(-\sum_{i=1}^{d} g(x_i)\right)$
- g is sufficiently differentiable
- g has finite moments, $\int_{\mathbb{R}} x^k \exp(-g(x))\mathrm{d}x < \infty$ for all $k \in \mathbb{N}$
- g' is Lipschitz
- $X_0 \sim \pi_d$

Set up is unrealistic but results have proven to be useful outside simple scenarios

Optimal scaling

- **RWM**: $q(X_k, \cdot) = \mathcal{N}(\cdot; X_k, \sigma^2 I_d)$
  - if $\sigma_d^2 \propto d^{-1}$
  - convergence to a Langevin diffusion
  - asymptotic optimal acceptance ratio ≈ 0.234
- **MALA**: $q(X_k, \cdot) = \mathcal{N}(\cdot; X_k + \frac{\sigma^2}{2}\nabla \log \pi(X_k), \sigma^2 I_d)$
  - if $\sigma_d^2 \propto d^{-1/3}$
  - convergence to a Langevin diffusion
  - asymptotic optimal acceptance ratio ≈ 0.574

Outline

1. Markov chain Monte Carlo
2. Proximal MCMC
3. Optimal Scaling
4. Optimal scaling for Proximal MALA

Our aim

$$q(X_k, Y) = \mathcal{N}\left(Y; X_k + \frac{\sigma^2}{2\lambda}\left(\mathrm{prox}_g^{\lambda}(X_k) - X_k\right), \sigma^2 I_d\right),$$

Investigate the optimal scaling properties of proximal MCMC 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 $\alpha, \beta > 0$.
Then we can write

$$\lambda = \frac{\sigma^{2m} r}{2}$$

where $r := c^2/\ell^{2m} > 0$, $m := \beta/\alpha$.

Differentiable targets

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

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

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

Differentiable targets – g' Lipschitz

Under appropriate condition we have

Acceptance rate

If $\alpha = 1/3$, $\beta = m/3$ for $m > 1$ and $r > 0$, the asymptotic average acceptance rate converges to

$$a(\ell, r) = 2\Phi\left(-\frac{\ell^3 K_1}{2}\right),$$

$$K_1^2 = K_{MALA}^2 = \frac{1}{16}\mathbb{E}_X\left[g''(X)^3\right] + \frac{5}{48}\mathbb{E}_X\left[g'''(X)^2\right].$$

Differentiable targets – g' Lipschitz

Under appropriate condition we have

Acceptance rate

If $\alpha = 1/3$, $\beta = m/3$ for $m = 1$ and $r > 0$, the asymptotic average acceptance rate converges to

$$a(\ell, r) = 2\Phi\left(-\frac{\ell^3 K_2(r)}{2}\right),$$

$$K_2^2(r) = K_{MALA}^2 + \frac{1}{8}\left(r + 2r^2\right)\mathbb{E}_X\left[g''(X)g'(X)\right]^2 + \frac{r}{8}\mathbb{E}_X\left[g''(X)^3\right].$$

Th result in the case of a Gaussian distribution and $r = 1$ was also established in Pillai (2022).

Differentiable targets – g' Lipschitz

Convergence to diffusion

As $d \to \infty$ we have convergence to a Langevin diffusion

$$\mathrm{d}Y_t = h(\ell, r)^{1/2}\mathrm{d}B_t + \frac{h(\ell, r)}{2} g'(x)\mathrm{d}t,$$

where $h(\ell, r) = \ell^2 a(\ell, r)$ is the speed of the diffusion. The speed $h(\ell, r)$ is maximized at the unique value of ℓ such that $a(\ell, r) = 0.574$.

Differentiable targets – g' Lipschitz

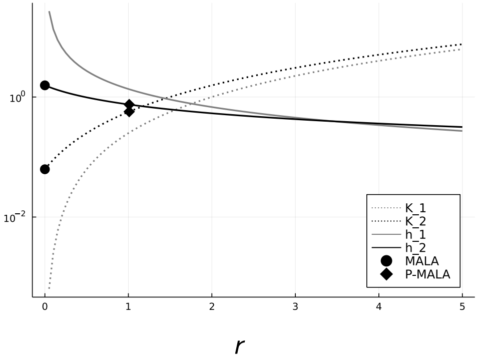

Figure: Speed of Langevin diffusion as a function of $r = c^2/\ell^{2m}$ for a Gaussian target when $m = 1$.

Differentiable targets – g' **not** Lipschitz

Under appropriate condition we have

Acceptance rate

If $\alpha = 1/2$, $\beta = 1/4$ and $r > 0$, the asymptotic average acceptance rate converges to

$$a(\ell, r) = 2\Phi\left(-\frac{\ell^2 K_3(r)}{2}\right), \quad K_3^2(r) = r^2 \mathbb{E}_X\left[\frac{[g''(X)g'(X)]^2}{4}\right].$$

Differentiable targets – g' **not** Lipschitz

Convergence to diffusion

As $d \to \infty$ we have convergence to a Langevin diffusion

$$\mathrm{d}Y_t = h(\ell, r)^{1/2}\mathrm{d}B_t + \frac{h(\ell, r)}{2} g'(x)\mathrm{d}t,$$

where $h(\ell, r) = \ell^2 a(\ell, r)$ is the speed of the diffusion. The speed $h(\ell, r)$ is maximized at the unique value of ℓ such that $a(\ell, r) = 0.452$.

Differentiable targets – g' **not** Lipschitz

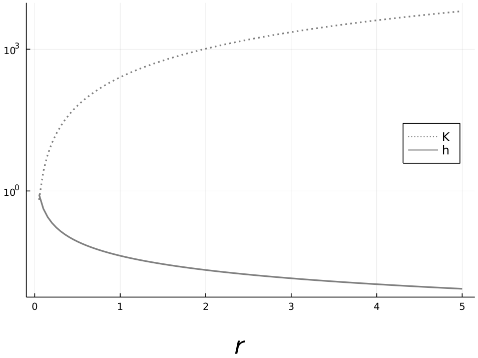

Figure: Speed of Langevin diffusion as a function of $r = c^2/\ell^{2m}$ for a light tail target $\pi(x) \propto \exp\left(-x^6\right)$.

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

Acceptance rate

If $\alpha = 2/3$, $\beta = 2m/3$ for $m \geq 1$ and $r \geq 0$, the asymptotic average acceptance rate does not depend on c and converges to

$$a(\ell) = 2\Phi\left(-\frac{\ell^{3/2}}{(72\pi)^{1/4}}\right).$$

Laplace target

Converge to diffusion

As $d \to \infty$ we have convergence to a Langevin diffusion

$$\mathrm{d}Y_t = h(\ell, c)^{1/2}\mathrm{d}B_t - \frac{h(\ell, c)}{2}\,\mathrm{sgn}(Y_t)\mathrm{d}t,$$

where $h(\ell, c) = \ell^2 a(\ell, c)$ is the speed of the diffusion. In addition, h does not depend on c and is maximized at the unique value of ℓ such that $a(\ell) = 0.360$.

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

Ideas of proof

- **differentiable targets:** the proof follows the structure of Roberts and Rosenthal (1998), Taylor expand the log-acceptance ratio and show convergence to the appropriate objects
- **Laplace target:** the proof is similar to that of Durmus et al. (2017)
  - show convergence of the log-acceptance ratio using Linderberg's CLT
  - show convergence to a diffusion using Kolmogorov's criterion and the corresponding martingale problem

Thank you!

Bibliography I

Alain Durmus, Sylvain Le Corff, Eric Moulines, and Gareth O Roberts. Optimal scaling of the random walk Metropolis algorithm under Lp mean differentiability. *Journal of Applied Probability*, 54(4):1233–1260, 2017.

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.

Natesh S. Pillai. Optimal scaling for the proximal Langevin algorithm in high dimensions. *arXiv preprint arXiv:2204.10793*, 2022.

Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.