Analysis of Proximal Markov chain Monte Carlo algorithms

Sullivan Castro
ENS Paris-Saclay
sullivan.castro@eleves.enpc.fr

Matthieu Mérigot-Lombard
ENS Paris-Saclay
merigotmatthieu@gmail.com

1 Introduction

In 1953, Rosenbluth et al. [1] introduced a novel approach for sampling thermodynamic equilibrium in physical systems using the Monte Carlo Markov Chain (MCMC) method. Their work laid the foundation for stochastic sampling techniques that are now widely used in statistical physics. In 1970, Hastings [2] extended this approach to accommodate more general probability distributions, including those known only up to a normalization constant. This generalization led to the development of the Metropolis-Hastings class of algorithms, which have since become fundamental in Bayesian statistics and probabilistic inference.

For high-dimensional models, one of the most effective extensions is the Metropolis-adjusted Langevin algorithm (MALA) [3], which leverages gradient information to capture local geometric properties and improve exploration of the parameter space. Another major advancement in high-dimensional inference is the introduction of log-concave distributions, commonly referred to as "convex models." These models have driven the development of proximal algorithms [4], which rely on proximity mappings of concave functions to construct fixed-point schemes and facilitate optimization through implicit updates rather than explicit gradient steps.

In this work, we will analyze a Proximal Langevin MCMC algorithm for log-concave distributions that are *not necessarily continuously differentiable* [5]. In addition, we propose an adaptive proximal parameter tuning scheme that automatically adjusts the step size during the burn-in phase to balance stability and exploration efficiency. The code for our experiments is available on Github.

2 Proximal Metropolis-Adjusted Langevin Algorithm

Definitions

Let $\boldsymbol{x} \in \mathbb{R}^n$ and let $\pi(d\boldsymbol{x})$ be a probability distribution that admits a density $\pi(\boldsymbol{x})$. We consider distributions of the form:

$$\pi(\boldsymbol{x}) \propto \exp\left(g(\boldsymbol{x})\right) \tag{1}$$

where $g : \mathbb{R}^n \to \mathbb{R}$ is concave and upper semicontinuous, $\lim_{\|\boldsymbol{x}\|\to\infty} g(\boldsymbol{x}) = -\infty$, and g can be evaluated pointwise. Furthermore, it is assumed that g has a *proximity mapping* that can be explicitly evaluated or efficiently approximated.

Definition 2.1 (Proximity Mapping)**.** *For any* $\lambda > 0$*, the* λ-proximity mapping *(or proximal operator) of the function* g *is defined as* $prox_g^\lambda(\boldsymbol{x}) = arg \max_{\boldsymbol{u}\in\mathbb{R}^n} g(\boldsymbol{u}) - \frac{\|\boldsymbol{u}-\boldsymbol{x}\|^2}{2\lambda}$

It can be shown that if g is strictly concave and $\lim_{\|\boldsymbol{x}\|\to\infty} g(\boldsymbol{x}) = -\infty$, then prox_g^λ is well defined for all $\lambda > 0$. Moreover, as $\lambda \to \infty$, the quadratic term vanishes and $\text{prox}_g^\lambda(\boldsymbol{x})$ maps to the set of maximizers of g, while when $\lambda \to 0$, the proximal operator tends to the identity operator.

Definition 2.2 (Moreau Approximation). *For any $\lambda > 0$, the λ-Moreau approximation of π is defined as:*

$$\pi_\lambda(\boldsymbol{x}) = \frac{1}{\kappa'} \sup_{\boldsymbol{u} \in \mathbb{R}^n} \pi(\boldsymbol{u}) \exp\left(-\frac{\|\boldsymbol{u} - \boldsymbol{x}\|^2}{2\lambda}\right)$$

where κ' is a normalizing constant.

An important property is that π_λ can be expressed (up to proportionality) using the proximal operator such that $\pi_\lambda(\boldsymbol{x}) \propto \exp\left(g\left(\text{prox}_g^\lambda(\boldsymbol{x})\right)\right) \exp\left(-\frac{\|\text{prox}_g^\lambda(\boldsymbol{x}) - \boldsymbol{x}\|^2}{2\lambda}\right)$

Definition 2.3 (Class of Distributions $\mathcal{E}(\beta, \gamma)$). *A density $\pi(\boldsymbol{x})$ belongs to the class of distributions $\mathcal{E}(\beta, \gamma)$ if, for some threshold u, there exist constants $\beta > 0$ and $\gamma > 0$ such that for $\|\boldsymbol{x}\| > u$,*

$$\pi(\boldsymbol{x}) \propto \exp\left(-\gamma \|\boldsymbol{x}\|^\beta\right)$$

Property 1. *If $\pi \in \mathcal{E}(\beta, \gamma)$, then the Moreau approximation π_λ is continuously differentiable even if π is not, with gradient:*

$$\nabla \log \pi_\lambda(\boldsymbol{x}) = \frac{1}{\lambda}\left(prox_g^\lambda(\boldsymbol{x}) - \boldsymbol{x}\right)$$

Algorithm Explanation

Definition 2.4 (Monte Carlo Sampling using Langevin Dynamics). *Let π be differentiable and strictly positive everywhere, ensuring that $\nabla \log(\pi)$ is well defined. Consider an n-dimensional Brownian motion $W = (W_t)_{t \geq 0}$, and define a Langevin diffusion process $\{Y_t : 0 \leq t \leq T\}$ in $\mathbb{R}^n$ with π as its stationary distribution, governed by the stochastic differential equation:*

$$dY(t) = \frac{1}{2} \nabla \log \pi(Y(t))dt + dW(t), Y(0) = y_0$$

Unfortunately, direct simulation from $Y(t)$ is generally not possible. Therefore, a discrete-time approximation of the Langevin diffusion process is required. Using the Euler-Maruyama approximation with discrete-time increment δ, known as the Unadjusted Langevin Algorithm (ULA), we obtain the iterative scheme:

$$Y^{(m+1)} = Y^{(m)} + \frac{\delta}{2} \nabla \log \pi(Y^{(m)}) + \sqrt{\delta} Z^{(m)}, \quad Z^{(m)} \sim \mathcal{N}(0, I_n)$$

When π is not differentiable, the gradient $\nabla \log \pi$ is not well defined. One can thus approximate:

$$\nabla \log \pi \approx \nabla \log \pi_\lambda$$

where π_λ is a continuously differentiable approximation of π.

Definition 2.5 (P-ULA). *Using Property 1:*

$$Y^{(m+1)} = (1 - \frac{\delta}{2\lambda}) Y^{(m)} + \frac{\delta}{2\lambda} prox_g^\lambda(Y^{(m)}) + \sqrt{\delta} Z^{(m)}, \quad Z^{(m)} \sim \mathcal{N}(0, I_n)$$

From a perspective of proximal point optimization, λ must be greater than $\frac{\delta}{2}$ to ensure stability. However, to minimize the effect of the approximation of π by π_λ, λ should be as close as possible to 0. Therefore, $\lambda = \frac{\delta}{2}$.

To account for the bias introduced by approximating π with π_λ, a Metropolis-Hastings correction is applied, leading to the Proximal Metropolis-Adjusted Langevin Algorithm (P-MALA).

Definition 2.6 (P-MALA). *Let $(X^{(m)})_{m \in \mathbb{N}}$ be a Markov chain that uses P-ULA as a proposal. Let Y^* be a candidate, and $q(\cdot|\cdot)$ be the proposal density that follows $Y^* | X^{(m)} \sim \mathcal{N}\left(prox_g^{\delta/2}(X^{(m)}),\, \delta\, I_n\right)$. The acceptance rate α is defined as:*

$$\alpha(X^{(m)}, Y^*) = \min\left\{1,\, \frac{\pi(Y^*)\, q(X^{(m)}|Y^*)}{\pi(X^{(m)})\, q(Y^*|X^{(m)})}\right\}$$

Theorem 2.1 (Geometric ergodicity). *Let $\pi \in \mathcal{E}(\beta, \gamma)$ such that Hypothesis 1 holds. Then, P-MALA is geometrically ergodic for all $\delta > 0$.*

3 Comparison with Metropolis-Adjusted Langevin Algorithm

Stronger Convergence Guarantees

Consider a target density π defined as in Hypothesis 1. Unlike standard MALA, which depends on the gradient $\nabla \log \pi$ and thus requires differentiability of π, the P-MALA algorithm relaxes this condition by approximating π with the continuously differentiable density π_λ of Defintion 2.2.

In practice, replacing the gradient with the proximal operator prox_g^λ allows P-MALA to be applied to models that include non-smooth components, such as the total-variation (TV) prior model:

$$\pi(\boldsymbol{x}|\boldsymbol{y}) \propto \exp\left(-\alpha\|\nabla_d \boldsymbol{x}\|_1\right)$$

where ∇_d denotes the discrete gradient operator.

Moreover, if π belongs to the class $\mathcal{E}(\beta, \gamma)$, P-MALA is geometrically ergodic, meaning there exist constants $C > 0$ and $0 < \rho < 1$ such that:

$$\|P^m(x, \cdot) - \pi(\cdot)\| \leq C\rho^m$$

This guarantees that P-MALA converges at a geometric rate, achieving robustness in situations where standard MALA might fail. Indeed, to ensure geometric ergodicity, MALA generally requires that the target density be at least twice continuously differentiable [6].

Computational Overhead

The existence of a proximity mapping is a first constraint. However, the evaluation of the proximal operator may also become *computationally intensive* in high-dimensional settings. Thus, the per-iteration cost in P-MALA is higher than the cost of evaluating gradients in standard MALA. In fact, in the image denoising experiment described in the paper, the computation time for P-MALA was approximately twice that for MALA.

Parameter Sensitivity and Hyperparameter Tuning

Similarly to MALA, the efficiency of P-MALA is highly sensitive to the choice of the discretization step δ. An inappropriate selection of this parameters can lead to either slow exploration of the state space or an increased rejection rate in the Metropolis-Hastings correction step [7].

Restriction to Log-Concave Distributions

P-MALA is onl designed for log-concave target densities. As such, its theoretical guarantees and practical performance are not directly extendable to non-convex models where g lacks concavity.

4 Adaptive Proximal Metropolis-Adjusted Langevin Algorithm

Proposed Adaptive Scheme

In the original P-MALA, where the proximal parameter is fixed as $\lambda = \frac{\delta}{2}$ to ensure stability, the optimal value of δ may vary depending on the problem. If set too high, discretization errors become significant, and if too low, exploration becomes slow.

We propose to adapt the step size δ during the burn-in phase according to the observed acceptance rate. Let A_k be the empirical acceptance rate over a batch of iterations of size B, and A_{target} the target acceptance rate. Then, δ follows:

$$\delta_{k+1} = \delta_k \exp\left(\gamma\left(A_{\text{target}} - A_k\right)\right)$$

where $\gamma > 0$ controls the adaptation speed. This update aims to optimize the balance between discretization error and exploration efficiency, thus increasing the effective sample sizes (ESS).

Theoretical Considerations

It has been established that the optimal acceptance rate for MALA is approximately $A_{\text{opt}} \approx 0.574$ [8]. While an optimal is not formally proven for P-MALA, empirical findings suggest that an acceptance rate between 40% and 60% leads to the best results. Thus, we set $A_{\text{target}} = 0.5$.

Although the theoretical convergence proof assumes δ constant, the same convergence guarantees hold if adaptation is restricted to the burn-in phase and parameters remain fixed afterward.

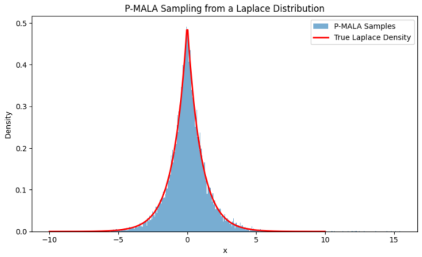

Figure 1: Sampling of the one-dimensiontal Laplace distribution ($\pi(x) \propto \exp(-|x|)$) obtained using P-MALA with $M = 20,000$ iterations, a burn-in period of 10%, step size $\delta = 1$, and initial guess $x_0 = 5.0$.

5 Application to Bayesian Denoising

Consider the Bayesian image deconvolution problem where the observed image y is modeled as $\boldsymbol{y} = H\boldsymbol{x_0} + \boldsymbol{w}$ with H being a known linear operator and $w \sim \mathcal{N}(0, \sigma^2 I_n)$. Using a Total-Variation (TV) prior leads to the posterior:

$$\pi(\boldsymbol{x}|\boldsymbol{y}) \propto \exp\left\{-\frac{\|\boldsymbol{y} - H\boldsymbol{x}\|^2}{2\sigma^2} - \alpha\|\nabla_d \boldsymbol{x}\|_1\right\}$$

A common approach to image denoising is to compute the maximum a posteriori (MAP) estimate, which maximizes the posterior distribution. However, assessing the confidence intervals of such predictions is challenging without employing MCMC methods. This is where P-MALA becomes particularly useful.

P-MALA is well-suited for this transition kernel because the posterior distribution is non-differentiable, rendering standard MALA inapplicable. Moreover, since the posterior is log-concave, its proximal operator is well-defined. Although the proximal operator associated with the TV norm lacks a closed-form solution, efficient computational methods exist, such as the proxTV package used in our implementation.

Let $g(\boldsymbol{x}) = g_1(\boldsymbol{x}) + g_2(\boldsymbol{x})$ where $g_1 \in \mathcal{C}^1$ and ∇g_1 is Lipchitz continuous, one can approximate $\text{prox}_g^{\delta/2}(\boldsymbol{x}) \approx \text{prox}_{g_2}^{\delta/2}(\boldsymbol{x} + \frac{\delta}{2}\nabla g_1(\boldsymbol{x}))$. Therefore, with $g(\boldsymbol{x}) = -\frac{\|\boldsymbol{y}-H\boldsymbol{x}\|^2}{2\sigma^2} - \alpha\|\nabla_d \boldsymbol{x}\|_1$:

$$\text{prox}_g^{\delta/2}(\boldsymbol{x}) \approx \text{prox}_{-\alpha\|\nabla_d \cdot\|_1}^{\delta/2}(\boldsymbol{x} + \frac{\delta}{2\sigma^2}H^T(\boldsymbol{y} - H\boldsymbol{x})) \tag{2}$$

Then, to obtain the confidence intervals, a Markov chain $(Y^{(m)})_{m\in\{1,\dots,M\}}$ is generated using the P-MALA proposal, and pixel-wise 95% credibility intervals is computed from the chain by taking the 2.5^{th} and 97.5^{th} percentiles:

$$\text{CI}_{95\%}(x_i) = \left[q_{2.5}((x_i^{(m)})_{m\in\{1,\dots,M\}}),\ q_{97.5}((x_i^{(m)})_{m\in\{1,\dots,M\}})\right]$$

To evaluate the performance of the algorithm, it is first applied to a simple denoising case where $H = I$ on a one-dimensional signal. The corresponding results are presented in Figure 2. The MAP estimate produces a smoothed signal that is close to x_0. However, the results obtained using P-MALA show significant instability, even when tested with different values of δ and varying numbers of iterations. This instability is unlikely to originate from the P-MALA implementation itself, as Figure 1 demonstrates that sampling from the Laplace distribution functions correctly. Instead, these findings suggest that the instability may come from an incorrect implementation of the proximal operator, which we were unable to resolve. It is also improbable that the issue arises from Approximation 2 as it was tested with small values of δ. The effect of the adaptive scheme, as shown in Figure 2, confirms its intended impact on the acceptance rates. However, due to the instability of the original algorithm, the adaptive scheme only amplifies this behavior.

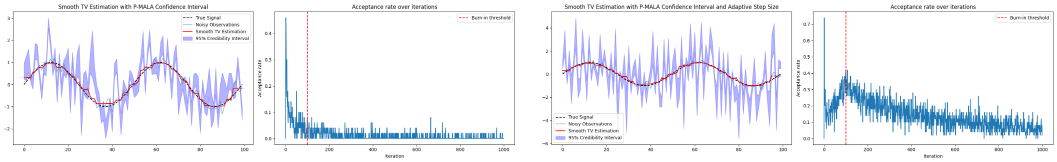

Figure 2: Comparison of the MAP estimate, the 95% credibility interval, and the empirical acceptance rates obtained using P-MALA without (left) and with (right) adaptive scheme for a one-dimensional sinusoidal signal corrupted by Gaussian noise $w \sim \mathcal{N}(0, 0.04 I_n)$. P-MALA was run with $M = 50,000$ iterations, a burn-in period of 10%, step size $\delta = 0.001$, a proposal standard deviation $\sigma = 0.2$, and hot-started with MAP estimate. For the adaptive P-MALA scheme, the adaptation speed is set to $\gamma = 0.1$, and batch size to $B = 50$.

Furthermore, the algorithm was tested for image denoising. However, given the existing instability, the results were not satisfactory and are therefore not presented here, but still provided in the code.

6 Conclusion

In this work, we have provided an analysis of the P-MALA algorithm. We first reviewed its theoretical foundations, focusing on the proximal operators and the Euler-Maruyama approximation. We then examined its strengths and weaknesses, particularly its geometric ergodicity, computationnal cost and sensitivity to parameter choices.

We then applied P-MALA to Bayesian denoising, to test its ability to provide uncertainty quantification for non-differentiable priors. Moreover, we introduced an adaptive tuning mechanism to optimize the step size during burn-in, potentially improving sampling efficiency and increasing effective sample sizes.

References

[1] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 06 1953.

[2] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 04 1970.

[3] George Casella Christian P. Robert. *Monte Carlo Statistical Methods*. Springer, 2004.

[4] Stephen Boyd Neal Parikh. *Proximal Algorithms*. Foundations and Trends® in Optimization, 2014.

[5] Marcelo Pereyra. Proximal markov chain monte carlo algorithms. *Statistics and Computing*, 2015.

[6] Alain Durmus and Éric Moulines. On the geometric convergence for mala under verifiable conditions. 2022.

[7] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2:341–363, 1996.

[8] 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.