Langevin Monte Carlo Beyond Lipschitz Gradient Continuity

Matej Benko [1], Iwona Chlebicka [2], Jørgen Endal [3], Błażej Miasojedow [2]

[1]Brno University of Technology
[2]University of Warsaw
[3]Norwegian University of Science and Technology

Main Goal

Scope

- We provide an algorithm to sample from the probability measure with the density

$$\mu^*(x) = \frac{\exp(-V(x))}{\int_{\mathbb{R}^d} \exp(-V(y))\,\mathrm{d}y}.$$

- Sharp convergence rates in terms of the Wasserstein distance and Kullback-Leibler divergence are provided.
- We do not need to assume that ∇V is globally Lipschitz.
- **Our approach cover light tail distributions.**

Applications

- Bayesian statistics
- Machine Learning

Langevin Monte Carlo

Langevin Diffusion Equation

$$\begin{cases} \mathrm{d}Y_t = -\nabla V(Y_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, & \\ Y_t \sim \mu_t\,, & t > 0\,, \\ Y_0 \sim \mu_0\,, & \end{cases}$$

Standard (Explicit) Langevin Monte Carlo

$$\begin{aligned} &X_0 \sim \mu_0 \\ &X_k = X_{k-1} - \tau \nabla V(X_{k-1}) + \sqrt{2\tau} Z_k \qquad k = 1, 2, \ldots \end{aligned}$$

where $\{Z_k\}_{k \geq 1}$ i.i.d standard Gaussian distribution independent on history of chain.

State of the art

Typical assumptions

- global Lipschitz continuity of ∇V
- convexity (strong convexity) of V
- convexity can be relaxed if log-Sobolev inequality is satisfied

Existing results

- Finite time convergence bounds in terms of KL-divergence or Wasserstein distance.
- Explicit and polynomial dependence on dimension of bounds

Methods

- Convergence of diffusion process and comparison with its discretization.
- Apply gradient flow formulation and tools from convex optimization theory.

Literature

S. Chewi, M. A. Erdogdu, M. Li, et al., “Analysis of langevin monte carlo from poincare to log-sobolev,”

A. Dalalyan, “Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent,”

A. S. Dalalyan, “Theoretical guarantees for approximate sampling from smooth and log-concave densities,”,

A. Durmus, S. Majewski, and B. Miasojedow, “Analysis of Langevin Monte Carlo via convex optimization,”,

A. S. Dalalyan, A. Karagulyan, and L. Riou-Durand, “Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets,”,

A. Durmus and É. Moulines, “Nonasymptotic convergence analysis for the unadjusted Langevin algorithm,”,

A. Durmus and É. Moulines, “High-dimensional Bayesian inference via the unadjusted Langevin algorithm,”,

M. A. Erdogdu and R. Hosseinzadeh, “On the convergence of langevin monte carlo: The interplay between tail growth and smoothness,”

M. A. Erdogdu, R. Hosseinzadeh, and S. Zhang, “Convergence of langevin monte carlo in chi-squared and rényi divergence,”

A. Mousavi-Hosseini, T. K. Farghly, Y. He, et al., “Towards a complete analysis of langevin monte carlo: Beyond poincaré inequality,”

Why we need gradient Lipschitz V?

- **Euler–Maruyama dicretization is transient if ∇V is not Lipschitz**
  - G. O. Roberts and R. L. Tweedie, "Exponential convergence of Langevin distributions and their discrete approximations,",
  - J. C. Mattingly, A. M. Stuart, and D. J. Higham, "Ergodicity for SDEs and approximations: Locally Lipschitz vector fields and degenerate noise,",

Explicit Euler scheme:

$$x_{k+1} = x_k - \tau \nabla V(x_k)$$

We have $V(x_{k+1}) \leq V(x_k)$ if ∇V Lipschitz, otherwise $V(x_{k+1})$ could be arbitrarily large.

Implicit Euler scheme:

$$\color{red}{x_{k+1}} = x_k - \tau \nabla V(\color{red}{x_{k+1}})$$

We have $V(x_{k+1}) \leq V(x_k)$ always even if V is not convex.

Inexact Proximal Langevin Algorithm

- **Proximal operator:**

$$\mathrm{prox}_V^{\tau}(x) := \arg\min_{y\in\mathbb{R}^d} \left\{ V(y) + \frac{1}{2\tau}|y-x|^2 \right\}.$$

- We consider numerical approximation of the minimizer with known precision δ

Inexact Proximal Langevin Algorithm (IPLA)

- Sample initial distribution $X_0 \sim \mu_0$
- For $k = 0, \ldots, n-1$:

Step 1: Run routine for computing with an output

$$X_{k+\frac{2}{3}} = \mathrm{prox}_V^{\tau}(X_k) + \Theta_{k+\frac{2}{3}}\,; \qquad |\Theta_{k+\frac{2}{3}}| \leq \delta\,,$$

Step 2: Add Gaussian noise, i.e

$$X_{k+1} = X_{k+\frac{2}{3}} + Z_{k+1}\,; \qquad Z_{k+1} \sim \mathcal{N}(0, 2\tau \mathrm{Id})\,.$$

Assumptions

Assumptions for IPLA

- V is λ-convex for $\lambda \in \mathbb{R}$ outside of a ball $B \subset \mathbb{R}^d$:

$$V(x) \geq V(y) + \nabla V(y) \cdot (x - y) + \tfrac{\lambda}{2} \mathbb{1}_{\mathbb{R}^d \setminus B}(y) \, |x - y|^2 \, ; \qquad \forall x, y \in \mathbb{R}^d \, .$$

- V is gradient locally Lipschitz with polynomial q-growth:

$$|\nabla V(x) - \nabla V(y)| \leq L_q \min\{|x - y|, 1\}(|x|^{q-1} + |y|^{q-1}) \quad \forall x, y \in \mathbb{R}^d \, .$$

- Initial distribution $\varrho_0 \in \mathscr{P}_{q_V+1}(\mathbb{R}^d)$, i.e. has finite moment of order $q_V + 1$.

Theoretical results

KL Error Bound

Let $\tau < 1/\lambda_V$, $\kappa, \alpha > 0$, and $\delta \leq \kappa\tau^{1+\alpha}$. Assume further that

$$0 < \tau_\varepsilon \leq \min\left\{\left(\frac{\varepsilon}{3C(\mu^*)\kappa}\right)^{\frac{1}{\alpha}}, 1\right\}$$

and $K(\tau_\varepsilon) \leq \varepsilon/3$. Let the number of iterations n_ε be such that $n_\varepsilon \geq 3W_2^2(\varrho_0, \mu^*)/(2\varepsilon\tau_\varepsilon)$. Then

$$\mathrm{KL}(\varrho_{n_\varepsilon}|\mu^*) \leq \varepsilon\,.$$

Moreover, for computing one sample in terms of KL with precision ε, in a case of warm start ($W_2^2(\varrho_0, \mu^*) \leq C$ for some absolute constant C), we need

(i) $d^{\frac{q_V+1}{2}}\mathcal{O}(\varepsilon^{-2})$ iterations, if $\alpha \geq 1$;

(ii) $d^{\frac{q_V+1}{2}}\mathcal{O}(\varepsilon^{-1-\alpha^{-1}})$ iterations, if $\alpha < 1$.

Theoretical Results

W_2 Error Bound (for strongly convex case)

Suppose that $\boldsymbol{R_V} = \boldsymbol{0}$ and $\boldsymbol{\lambda_V} > \boldsymbol{0}$. Let $\tau < 1/\lambda_V$, $\alpha \geq 0$, and $\delta \leq \kappa\tau^{1+\alpha}$. Assume further that

$$\tau_\varepsilon{}^{2\alpha} \leq \frac{\lambda_V^2 \varepsilon}{96\kappa^2 \log^2(6W_2^2(\varrho_0, \mu^*)\varepsilon^{-1})}$$

and $K(\tau_\varepsilon) \leq \frac{1}{12}\lambda_V \varepsilon$. Let the number of iterations n_ε be such that

$$2\log(6W_2^2(\varrho_0, \mu^*)\varepsilon^{-1})\tau_\varepsilon{}^{-1}\lambda_V^{-1} \leq n_\varepsilon \leq 4\log(6W_2^2(\varrho_0, \mu^*)\varepsilon^{-1})\tau_\varepsilon{}^{-1}\lambda_V^{-1}\,.$$

Then

$$W_2^2(\varrho_{n_\varepsilon}, \mu^*) \leq \varepsilon\,.$$

Moreover, for computing one sample in terms of the Wasserstein distance with precision ε, in the case of warm start, up to logarithmic terms we need

(i) $d^{\frac{q_V+1}{2}}\mathcal{O}(\varepsilon^{-2})$ iterations, if $\alpha \geq \frac{1}{2}$;

(ii) $d^{\frac{q_V+1}{2}}\mathcal{O}(\varepsilon^{-\alpha^{-1}})$ iterations, if $\alpha < \frac{1}{2}$.

Example 1: Sampling from Distribution with Light Tail

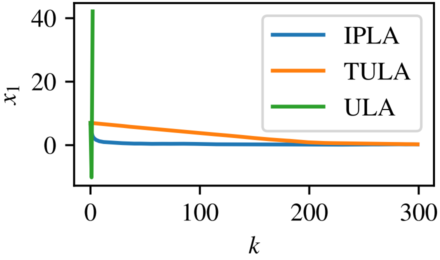
Density of form

$$\mu^*(x) \propto \exp\left(-\frac{|x|^4}{4}\right).$$

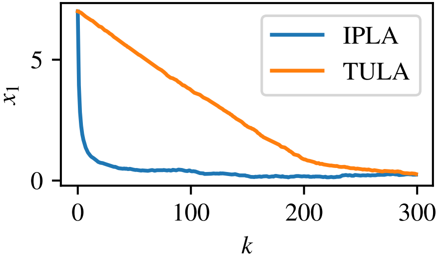
Mom.	Method	Start in tail RE	Start in tail CV	Start in $x_0 = 0$ RE	Start in $x_0 = 0$ CV
$\mathbb{E}\|Y\|^2$	IPLA	0.0027	0.0019	0.0006	0.0018
	TULA	0.0047	0.0016	0.0030	0.0019
	ULA	NaN	NaN	0.0020	0.0018
$\mathbb{E}\|Y\|^4$	IPLA	0.0054	0.0039	0.0025	0.0036
	TULA	0.0095	0.0032	0.0073	0.0039
	ULA	NaN	NaN	0.0028	0.0035
$\mathbb{E}\|Y\|^6$	IPLA	0.0081	0.0058	0.0047	0.0054
	TULA	0.0144	0.0047	0.0120	0.0058
	ULA	NaN	NaN	0.0032	0.0053

TULA: N. Brosse, A. Durmus, É. Moulines, et al., The tamed unadjusted Langevin algorithm, 2019

Example 1: Sampling from Distribution with Light Tail cont.



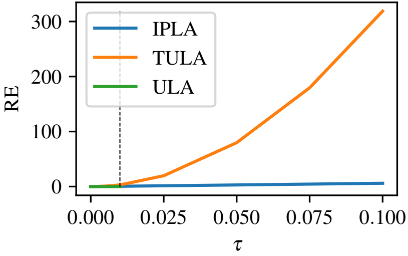
(a)



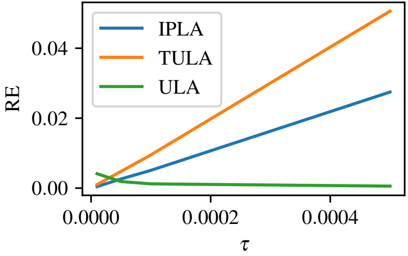
(b)

Figure: Trajectory of the 1^{st} coordinate from Example 1 starting in a tail. Both plots are based on the same data.

Example 1: Sampling from Distribution with Light Tail cont.



(a) RE



(b) RE (detail)

Figure: Example 1, starting in minimizer. Dependence of relative error (RE) of IPLA, TULA and ULA on stepsize τ. ULA gives results only for smaller values of τ. Dashed line represents the observed edge of the area of ULA stability ($\tau \approx 0.01$).

Example 2: Ginzburg–Landau model

Density

$$V(x) = \sum_{i,j,k=1}^{q} \tfrac{1-\upsilon}{2} x_{ijk}^2 + \tfrac{\upsilon\varkappa}{2} |\widetilde{\nabla} x_{ijk}|^2 + \tfrac{\upsilon\varsigma}{4} x_{ijk}^4 \,,$$

where $\widetilde{\nabla} x_{ijk} = \big(x_{i+jk} - x_{ijk}, x_{ij+k} - x_{ijk}, x_{ijk+} - x_{ijk}\big)\,.$

Mom.	Method	Start in tail		Start in $x_0 = 0$	
		RE	CV	RE	CV
$\mathbb{E}\|Y\|^2$	IPLA	0.0025	0.0244	0.0748	0.0786
	TULA	0.0067	0.0213	0.0859	0.0739
	ULA	NaN	NaN	0.0727	0.0697
$\mathbb{E}\|Y\|^4$	IPLA	0.0053	0.0491	0.1425	0.1558
	TULA	0.0134	0.0425	0.1635	0.1473
	ULA	NaN	NaN	0.1385	0.1399
$\mathbb{E}\|Y\|^6$	IPLA	0.0083	0.0742	0.2040	0.2323
	TULA	0.0199	0.0638	0.2337	0.2208
	ULA	NaN	NaN	0.1980	0.2117

Bayesian Image Deconvolution

Density

$$\mu^* \propto \exp\left(-\frac{1}{2\sigma^2}|y - Hx|^2 - \beta\, TV(x)\right).$$

- TV is not smooth function (violate assumption on ULA and TULA)

(a) Original image (b) Blurred (c) Processed ($\beta = 0.03$)

Figure: Result of the Bayesian Image Denoising. The original photo by Zbyszko Siemaszko 1955-56.