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 (NTNU)

Scope

- We provide an algorithm to sample from the probability measure with the density
$$\mu^*(x) = \frac{\exp(-V(x))}{\int_{\mathbb{R}^d} \exp(-V(y))\,\mathrm{d}y}.$$
- **We do not need to assume that ∇V is globally Lipschitz and moreover does not have to exists.**
- **Our approach cover light tail distributions.**

Assumptions

- 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 $q_V + 1$.

Gradient Flows

- We consider μ^* as a result of variational problem
$$\mu^* = \operatorname*{arg\,min}_{\mu \in \mathscr{P}_2(\mathbb{R}^d)} \mathcal{F}[\mu] = \operatorname*{arg\,min}_{\mu \in \mathscr{P}_2(\mathbb{R}^d)} \{\mathcal{F}_V[\mu] + \mathcal{F}_{\mathcal{E}}[\mu]\},$$
such that
$$\mathcal{F}_V[\mu] := \int_{\mathbb{R}^d} V(x)\,\mu(\mathrm{d}x) \qquad \text{and} \qquad \mathcal{F}_{\mathcal{E}}[\mu] := \begin{cases} \int_{\mathbb{R}^d} \mu(x)\log\mu(x)\,\mathrm{d}x, & \text{iff } \mu \ll \text{Leb}, \\ +\infty, & \text{otherwise}. \end{cases}$$

Algorithm

- We define ***Proximal operator:***
$$\operatorname{prox}_V^\tau(x) := \operatorname*{arg\,min}_{y \in \mathbb{R}^d} \{V(y) + \tfrac{1}{2\tau}|y-x|^2\}.$$
- We consider numerical approximation of the minimizer with known precision δ.
- Proximal step does not require existence of ∇V.
- The approach with proximal step allows us to sample from density with non-smooth potential.

Inexact Proximal Langevin Algorithm (IPLA)

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

  Step 1: Run routine for computing with an output
  $$X_{k+\frac{2}{3}} = \operatorname{prox}_V^\tau(X_k) + \Theta_{k+\frac{2}{3}}; \qquad |\Theta_{k+\frac{2}{3}}| \leq \delta,$$
  Step 2: Add Gaussian noise, i.e
  $$X_{k+1} = X_{k+\frac{2}{3}} + Z_{k+1}; \qquad Z_{k+1} \sim \mathcal{N}(0, 2\tau\mathbf{Id}).$$

- It is implicit version of Unadjusted Langevin Algorithm.
- Thanks to proximal step instead of gradient step we are beyond gradient Lipschitz continuity.

Theoretical Analysis of the Algorithm

- We split the inexact proximal step into two substeps: exact proximal step and additive error:

  (i) $X_{k+\frac{1}{3}} := \operatorname{prox}_V^\tau(X_k)$,

  (ii) $X_{k+\frac{2}{3}} := X_{k+\frac{1}{3}} + \Theta_{k+\frac{2}{3}}; \qquad \Theta_{k+\frac{2}{3}} \sim \xi_\delta, \quad \xi_\delta \in \mathscr{P}(\mathbb{R}^d);$ supported on $B(0,\delta)$,

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

- Step (i) is a step of length τ of gradient flow along the λ_V-convex potential functional $\mathcal{F}_V$; $\lambda_V > 0$.
- Step (ii) introduce approximation error from inexact proximal step.
- Step (iii) is a step of length τ of gradient flow along the convex entropy functional $\mathcal{F}_{\mathcal{E}}$.

Auxiliary Result (Moment Bound)

- Theorem: Let $m \geq 0$. There exists a constant $C > 0$ such that we have
$$\sup_k \mathbb{E}|X_k|^m \leq C d^{\frac{m}{2}}.$$
- We need this result since we allow ∇V to be beyond gradient Lipschitz.

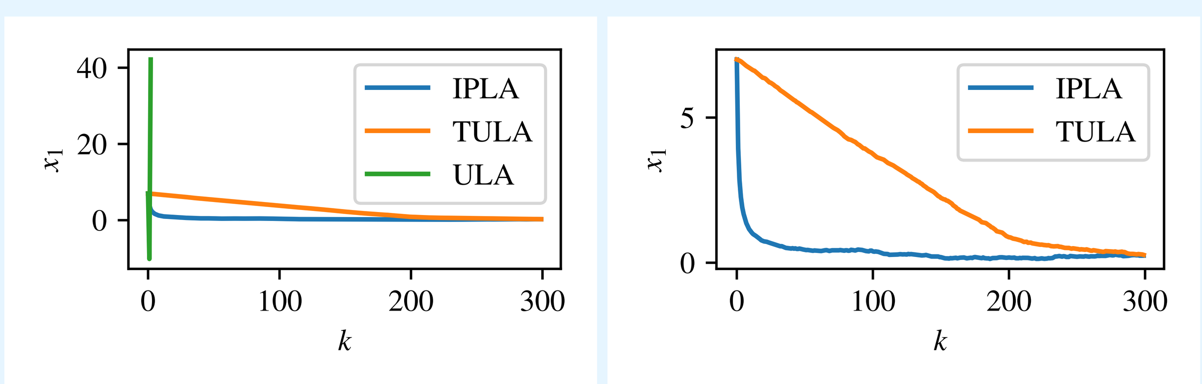

Figure 1. Trajectory of the 1st coordinate from sampling from distribution with the light tails (Example 1) starting in a tail. Both plots are based on the same data.

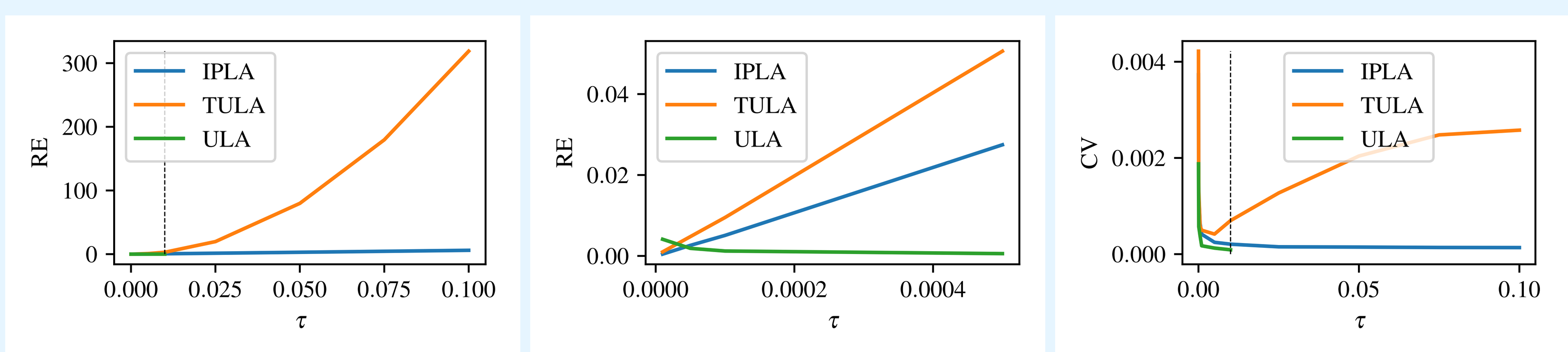

Figure 2. Example 1, starting in minimizer. Dependence of RE and CV of IPLA, TULA and ULA on stepsize τ. ULA gives results only for smaller values of τ. Dashed line represents the observed edge of the area of ULA stability ($\tau \approx 0.01$).

Results: KL Error Bound

- We state bound for a sequence of probability measures $\{\nu_n^N\}_{n\in\mathbb{N}}$, called average measures, defined for every $n, N \in \mathbb{N}$, $n \geq 1$ by
$$\nu_n^N := \tfrac{1}{n}\textstyle\sum_{k=N+1}^{N+n} \varrho_k,$$
while N is a burn-in time.
- Theorem: *Let $\tau < 1/\lambda_V$, $\kappa, \alpha > 0$, and $\delta \leq \kappa\tau^{1+\alpha}$. Then it holds:*
$$\mathrm{KL}(\nu_n^N|\mu^*) \leq \tfrac{1}{2n\tau}W_2^2(\varrho_N, \mu^*) - \tfrac{1}{2n\tau}W_2^2(\varrho_{N+n}, \mu^*) + C_{\mu^*}\kappa\tau^\alpha + C_{q_V}\tau d^{\frac{q_V+1}{2}}.$$
- Corollary: *Assume further that*
$$0 < \tau_\varepsilon \leq C_V\,\varepsilon$$
and let the number of iterations n_ε be large enough – we derived explicit condition. Then
$$\mathrm{KL}(\nu_{n_\varepsilon}^0|\mu^*) \leq \varepsilon.$$

Results: W_2 Error Bound

- Theorem: *Suppose that $R_V = 0$ and $\lambda_V > 0$. Let $\tau < 1/\lambda_V$, $\alpha \geq 0$, and $\delta \leq \kappa\tau^{1+\alpha}$. Iff τ_ε is small enough and n_ε satisfy derived condition then*
$$W_2^2(\varrho_{n_\varepsilon}, \mu^*) \leq \varepsilon.$$
- We give explicit conditions on τ_ε and n_ε depending on geometry of V and ϱ_0 and μ^*.

Sketch of the Proof

- We prove inequality
$$2\tau(\mathcal{F}[\varrho_{k+1}] - \mathcal{F}[\nu]) \leq W_2^2(\varrho_k, \nu) - W_2^2(\varrho_{k+1}, \nu) + C_\nu\delta + C_{q_V} d^{\frac{q_V+1}{2}}\tau^2.$$
- Proven inequality is combination of gradient flow characterization and our moment bound auxiliary result.
- We use that for $\mu \in \mathscr{P}_2(\mathbb{R}^d)$ and $\mathcal{F}_{\mathcal{E}}[\mu] < +\infty$ it satifies
$$\mathcal{F}[\mu] - \mathcal{F}[\mu^*] = \mathrm{KL}(\mu|\mu^*),$$
where KL stands for Kullback-Leibler divergence and it is convex.
- We consider exponential decay of gradient flow λ_V-convex functional $\mathcal{F}_V$, $\lambda_V > 0$ and nonincreasing decay along convex $\mathcal{F}_{\mathcal{E}}$.

Experiments

- Example 1: Sampling from distribution with light tails
$$\mu^*(x) \propto \exp\left(-\tfrac{|x|^4}{4}\right).$$
- Example 2: Ginzburg-Landau model with the potential
$$V(x) = \sum_{i,j,k=1}^{q} \tfrac{1-\upsilon}{2}x_{ijk}^2 + \tfrac{\upsilon\varkappa}{2}|\widetilde{\nabla} x_{ijk}|^2 + \tfrac{\upsilon\varsigma}{4}x_{ijk}^4.$$
- Example 3: Bayesian image deconvolution: Bayesian estimate of sharp picture x from observed blured picture (y) with prior distribution given by nonsmooth isotropic 2D total variation (TV) function. We sample from posterior distribution in the form
$$\mu^* \propto \exp\left(-\tfrac{1}{2\sigma^2}|y - Hx|^2 - \beta\ \mathrm{TV}(x)\right).$$

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

Figure 3. Result of the Bayesian image denoising problem. The original photo by Zbyszko Siemaszko 1955-56.

- We compared IPLA with explicit Unadjusted Langevin Algorithm (ULA) and Tamed Unadjusted Langevin Algorithm (TULA). Abbreviations: RE $\equiv$ relative error, CV $\equiv$ coefficient of variance.

Table 1. Estimation of the moments of light tails distribution from Example 1.

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

Table 2. Estimation of moments of Ginzburg–Landau model from Example 2.

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	0.0025	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

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[2] Matej Benko, Iwona Chlebicka, Jørgen Endal, and Błażej Miasojedow. Convergence rates of particle approximation of forward-backward splitting algorithm for granular medium equations, arxiv: 2405.18034, 2024.

[3] Nicolas Brosse, Alain Durmus, Éric Moulines, and Sotirios Sabanis. The tamed unadjusted Langevin algorithm. *Stochastic Process. Appl.*, 129(10):3638–3663, 2019.

[4] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20:Paper No. 73, 46, 2019.

[5] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.

 Matej.Benko@vut.cz