Northwestern University

Bregman Proximal Langevin Monte Carlo via Bregman–Moreau Envelopes

Joint work with Han Liu

International Conference on Machine Learning (ICML) 2022

Tim Tsz-Kit Lau
Department of Statistics and Data Science
Northwestern University
timlautk@u.northwestern.edu; https://timlautk.github.io

Problem

- (Approximately) sample from a probability distribution with density π

$$(\forall x \in \mathbb{R}^d) \quad \pi(x) = \mathrm{e}^{-U(x)} \Big/ \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)}\,\mathrm{d}y \propto \mathrm{e}^{-U(x)},$$

  where the *potential* $U \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is measurable and $0 < \int_{\operatorname{dom} U} \mathrm{e}^{-U(y)}\,\mathrm{d}y < +\infty$
- Usually, the number of dimensions $d \gg 1$
- Possibly nonsmooth composite potential U

$$(\forall x \in \mathbb{R}^d) \quad U(x) := f(x) + g(x)$$

- f is continuously differentiable but possibly not globally Lipschitz smooth (i.e., do not admit a globally Lipschitz gradient)
- g is possibly nonsmooth
- f and g are both convex, proper and lower semicontinuous

Langevin Monte Carlo Algorithms

- The Langevin Monte Carlo (LMC) algorithm (see e.g., Dalalyan, 2017) is arguably the most widely-studied *gradient-based MCMC algorithm*, which takes the form

$$(\forall k \in \mathbb{N}) \quad x_{k+1} = x_k - \gamma \nabla U(x_k) + \sqrt{2\gamma}\, \xi_k,$$

  where $\xi_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}_d(0, I)$ for all $k \in \mathbb{N}$ and $\gamma \in]0, 1]$ is a step size
- Possibly with varying step sizes, the LMC algorithm is also referred to as the *unadjusted* Langevin algorithm (ULA; Durmus and Moulines, 2017)
- Applying a Metropolis–Hastings correction step at each iteration of ULA, the algorithm is often referred to as the *Metropolis-adjusted* Langevin algorithm (MALA; Roberts and Tweedie, 1996).

Mirror-Langevin Algorithm

- Mirror-Langevin Algorithm (MLA; Hsieh et al., 2018; Zhang et al., 2020; Ahn and Chewi, 2021; Li et al., 2022, cf. *mirror descent*):

$$(\forall k \in \mathbb{N}) \quad x_{k+1} = \nabla\varphi^*\Big(\nabla\varphi(x_k) - \gamma\nabla U(x_k) + \sqrt{2\gamma}\,[\nabla^2\varphi(x_k)]^{1/2}\xi_k\Big)$$

- φ is a *Legendre* function
- E.g., **Hyperbolic entropy** (hypent) which interpolates between the squared Euclidean distance and the Boltzmann−Shannon entropy as β varies:

$$\varphi_\beta(x) = \sum_{i=1}^{d}\left[x_i \operatorname{arsinh}\left(\frac{x_i}{\beta_i}\right) - \sqrt{x_i^2 + \beta_i^2}\right]$$

$$\nabla\varphi_\beta(x) = \left(\operatorname{arsinh}\left(\frac{x_i}{\beta_i}\right)\right)_{1\leqslant i\leqslant d}$$

$$\nabla\varphi_\beta^*(x) = (\beta_i \sinh(x_i))_{1\leqslant i\leqslant d}$$

Bregman–Moreau Envelopes

- Smooth envelopes of the nonsmooth part g of the potential U
- Extending Moreau envelopes with Bregman divergences instead of squared Euclidean distances, the (left and right) *Bregman–Moreau envelopes* are

$$\overleftarrow{\mathrm{env}}^{\psi}_{\lambda,g}(x) := \inf_{y\in\mathbb{R}^d} \left\{ g(y) + \frac{1}{\lambda} D_\psi(y,x) \right\}$$

$$\overrightarrow{\mathrm{env}}^{\psi}_{\lambda,g}(x) := \inf_{y\in\mathbb{R}^d} \left\{ g(y) + \frac{1}{\lambda} D_\psi(x,y) \right\}$$

where $D_\psi(x,y)$ is the *Bregman divergence* between x and y associated with a Legendre function ψ and $\lambda > 0$

Bregman–Moreau Envelopes

- Extending Moreau proximity operators with Bregman divergences, the (left and right) *Bregman proximity operators* are

$$\overleftarrow{\mathrm{P}}^{\psi}_{\lambda,g}(x) := \operatorname*{argmin}_{y\in\mathbb{R}^d} \left\{ g(y) + \frac{1}{\lambda} D_{\psi}(y,x) \right\},$$

$$\overrightarrow{\mathrm{P}}^{\psi}_{\lambda,g}(x) := \operatorname*{argmin}_{y\in\mathbb{R}^d} \left\{ g(y) + \frac{1}{\lambda} D_{\psi}(x,y) \right\}$$

- $\overleftarrow{\mathrm{env}}^{\psi}_{\lambda,g}$ and $\overrightarrow{\mathrm{env}}^{\psi}_{\lambda,g}$ are **differentiable**
- Gradients of (left and right) Bregman–Moreau envelopes

$$\nabla \overleftarrow{\mathrm{env}}^{\psi}_{\lambda,g}(x) = \frac{1}{\lambda} \nabla^2 \psi(x) \left(x - \overleftarrow{\mathrm{P}}^{\psi}_{\lambda,g}(x) \right)$$

$$\nabla \overrightarrow{\mathrm{env}}^{\psi}_{\lambda,g}(x) = \frac{1}{\lambda} \left(\nabla \psi(x) - \nabla \varphi\left(\overrightarrow{\mathrm{P}}^{\psi}_{\lambda,g}(x) \right) \right)$$

Bregman Proximal LMC Algorithms

- Instead of directly sampling from π, we propose to sample from distributions whose potentials being smooth surrogates of U, defined by

$$\overleftarrow{U}_\lambda^\psi := f + \overleftarrow{\mathrm{env}}_{\lambda,g}^\psi \quad \text{and} \quad \overrightarrow{U}_\lambda^\psi := f + \overrightarrow{\mathrm{env}}_{\lambda,g}^\psi$$

- ψ is a Legendre function possibly different from the Legendre function φ in MLA to allow full flexibility
- The corresponding surrogate target densities are

$$\overleftarrow{\pi}_\lambda^\psi \propto \exp\left(-\overleftarrow{U}_\lambda^\psi\right) \quad \text{and} \quad \overrightarrow{\pi}_\lambda^\psi \propto \exp\left(-\overrightarrow{U}_\lambda^\psi\right).$$

The Bregman–Moreau Unadjusted Mirror-Langevin Algorithm

- The *Bregman–Moreau unadjusted mirror-Langevin algorithm* (BMUMLA) iterates, for $k \in \mathbb{N}$,

$$x_{k+1} = \nabla\varphi^*\left(\nabla\varphi(x_k) - \gamma\nabla U_\lambda^\psi(x_k) + \sqrt{2\gamma}\left[\nabla^2\varphi(x_k)\right]^{1/2}\xi_k\right).$$

- When $\varphi = \psi = \|\cdot\|^2/2$, then BMUMLA reduces to MYULA (Durmus et al., 2018)
- Sampling analogue of the *Bregman proximal gradient algorithm* via right BMUMLA with $\varphi = \psi$

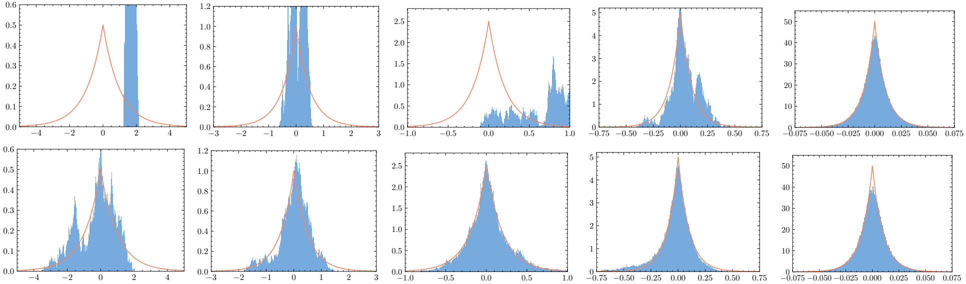
Numerical Experiments

- Nonsmooth sampling (anisotropic Laplace distribution):

$$f = 0 \quad \text{and} \quad g(\boldsymbol{x}) = \|\boldsymbol{\alpha} \odot \boldsymbol{x}\|_1 = \sum_{i=1}^{d} \alpha_i |x_i| \quad \text{with} \quad \boldsymbol{\alpha} = (1, 2, \ldots, d)^\top$$

- MYULA is known to perform poorly due to the *anisotropy*: with a relatively small step size, MYULA mixes fast for the *narrow* marginals, whereas it mixes slowly in the *wide* ones

MYULA vs BMUMLA



(a) $d = 1$ (b) $d = 2$ (c) $d = 5$ (d) $d = 10$ (e) $d = 100$

The End

Thank you!

References

Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-Langevin algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 3(79):651–676, 2017.

Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 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.

Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored Langevin dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Ruilin Li, Molei Tao, Santosh S. Vempala, and Andre Wibisono. The mirror Langevin algorithm converges with vanishing bias. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2022.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. Wasserstein control of mirror Langevin Monte Carlo. In *Proceedings of the Conference on Learning Theory (COLT)*, 2020.