Stochastic Proximal Langevin Algorithm: Potential Splitting and Nonasymptotic Rates

Adil Salim   Dmitry Kovalev   Peter Richtárik*

KAUST, Thuwal, Saudi Arabia. * Also affiliated with Moscow Institute of Physics and Technology, Dolgoprudny, Russia.

Goal

Sample from a target distribution $\mu^\star(x) \propto \exp(-U(x))$ where U convex.

The proposed method is a generalization of the Langevin algorithm to potentials U expressed as the sum of one stochastic smooth term and multiple stochastic nonsmooth terms.

We provide nonasymptotic rates for this method.

Background

KL divergence.
If $\mu \ll \pi$, then

$$\mathrm{KL}(\mu|\pi) := \int \log\left(\frac{d\mu}{d\pi}(x)\right) d\mu(x)$$

and $\mathrm{KL}(\mu|\pi) := +\infty$ else.

Wasserstein distance.
Let μ, ν probability measures with finite second moments.

$$W^2(\nu,\mu) := \inf\{\mathbb{E}\|Y - X\|^2, X \sim \mu, Y \sim \nu\}.$$

Langevin algorithm

If U is smooth, Langevin algorithm:

$$x^{k+1} = x^k - \gamma \nabla U(x^k) + \sqrt{2\gamma} W^k, \qquad (1)$$

where $\gamma > 0$ and $(W^k)_{k \geq 0}$ is a sequence of i.i.d. standard Gaussian random variables.
Typical nonasymptotic result:
$\mathrm{KL}(\overline{\mu_{\hat{x}^k}}|\mu^\star) = \mathcal{O}(1/\sqrt{k})$.

Another look at Langevin algorithm

The target $\mu^\star$ is the minimizer of $\mathcal{F} : \mu \mapsto \mathrm{KL}(\mu|\mu^\star)$ and Langevin algorithm can be seen as an (inexact) gradient descent applied to $\mathcal{F}$ ([1]).

Mathematical Problem

$$\text{Sample from} \quad \mu^\star(x) \propto \exp(-U(x)), \quad \text{where} \quad U(x) := F(x) + \sum_{i=1}^n G_i(x), \qquad (2)$$

where $F : \mathbb{R}^d \to \mathbb{R}$ is a smooth convex function and $G_1, \ldots, G_n : \mathbb{R}^d \to \mathbb{R}$ are (possibly nonsmooth) convex functions. Further, F and G written as expectations

$$F(x) = \mathbb{E}_\xi(f(x,\xi)), \quad \text{and} \quad G_i(x) = \mathbb{E}_\xi(g_i(x,\xi)). \qquad (3)$$

Stochastic Proximal Langevin Algorithm

Stochastic Proximal Langevin Algorithm (SPLA):

$$\begin{aligned} z^k &= x^k - \gamma \nabla f(x^k, \xi^k) \\ y_0^k &= z^k + \sqrt{2\gamma} W^k \\ y_i^k &= \mathrm{prox}_{\gamma g_i(\cdot,\xi^k)}(y_{i-1}^k) \quad \text{for} \quad i = 1, \ldots, n \\ x^{k+1} &= y_n^k, \end{aligned} \qquad (4)$$

where ξ^k i.i.d. copies of ξ.
Important instance: $U(x) = \mathbb{E}(g(x,\xi))$, $g(\cdot,\xi) : \mathbb{R}^d \to \mathbb{R}$ nonsmooth,

$$x^{k+1} = \mathrm{prox}_{\gamma g(\cdot,\xi^k)}(x^k) + \sqrt{2\gamma} W^k$$

Results

Table 1: Obtained complexity results

F	Rate	Nonasymptotic result
convex	$\mathrm{KL}(\mu_{\hat{x}^k} \mid \mu^\star) \leq \frac{1}{2\gamma(k+1)} W^2(\mu_{x^0}, \mu^\star) + \mathcal{O}(\gamma)$	$\mathrm{KL}(\overline{\mu_{\hat{x}^k}}\|\mu^\star) = \mathcal{O}(1/\sqrt{k})$
α-strongly convex	$W^2(\mu_{x^k}, \mu^\star) \leq (1-\gamma\alpha)^k W^2(\mu_{x^0}, \mu^\star) + \mathcal{O}\left(\frac{\gamma}{\alpha}\right)$	$W^2(\mu_{x^k}, \mu^\star) = \mathcal{O}(1/k)$
α-strongly convex	$\mathrm{KL}(\mu_{\tilde{x}^k} \mid \mu^\star) \leq \alpha(1-\gamma\alpha)^{k+1} W^2(\mu_{x^0}, \mu^\star) + \mathcal{O}(\gamma)$	$\mathrm{KL}(\mu_{\tilde{x}^k}\|\mu^\star) = \mathcal{O}(1/k)$

Approach

Following [1], we prove that the iterates shadow a discretized gradient flow of $\mathcal{F}$:

$$2\gamma\left\{\mathcal{F}(\mu_{y_0^k}) - \mathcal{F}(\mu^\star)\right\} \leq (1-\gamma\alpha) W^2(\mu_{x^k}, \mu^\star) - W^2(\mu_{x^{k+1}}, \mu^\star) + \gamma^2 C. \qquad (5)$$

The results follows from $\mathcal{F}(\mu) - \mathcal{F}(\mu^\star) = \mathrm{KL}(\mu|\mu^\star)$.

Bayesian Trend Filtering on Graphs

Graph $G = (V,E)$. Total variation over G: for all $x \in \mathbb{R}^V$,

$$\mathrm{TV}(x,G) := \sum_{i,j \in V, \{i,j\} \in E} |x(i) - x(j)|,$$

Target ([2]):

$$\mu^\star(x) \propto \exp\left(-\tfrac{1}{2\sigma^2}\|x-y\|^2 + \lambda\, \mathrm{TV}(x,G)\right)$$

where $\sigma, \lambda > 0$.

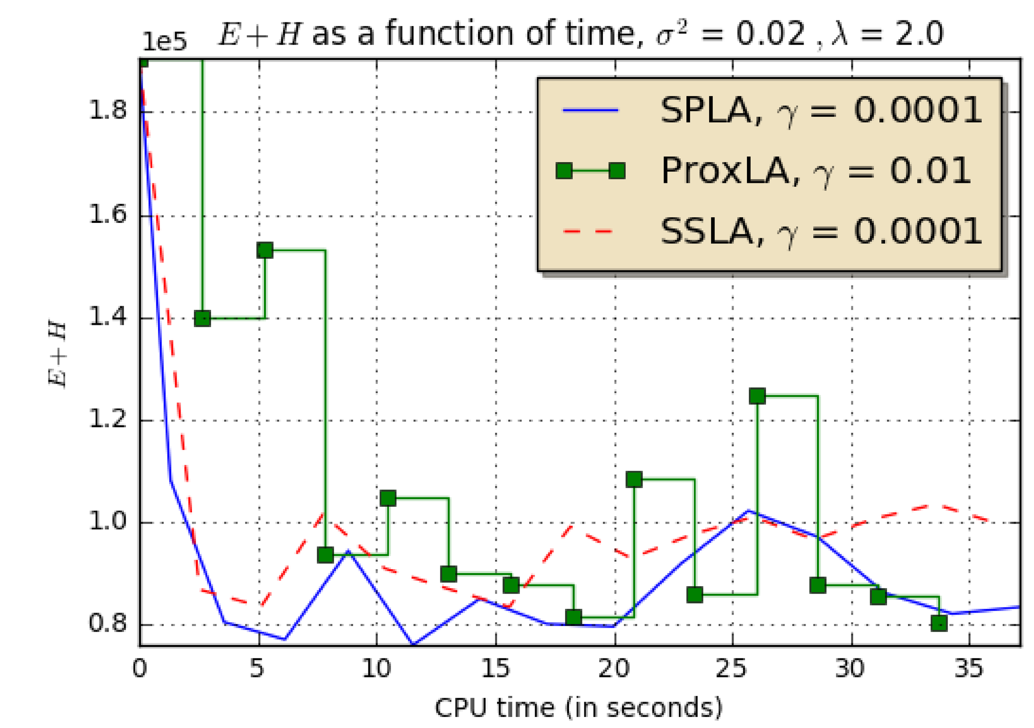

Figure 1: The functional $\mathcal{F} = E + H$ as a function of CPU time over the Facebook graph ($y \sim N(0, I_V)$).

References

[1] A. Durmus, S. Majewski, and B. Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *The Journal of Machine Learning Research*, 20(73):1–46, 2019.

[2] Y.-X. Wang, J. Sharpnack, A. J Smola, and R. J Tibshirani. Trend filtering on graphs. *The Journal of Machine Learning Research*, 17(1):3651–3691, 2016.

Contact

adil.salim, dmitry.kovalev@kaust.edu.sa, richtarik@gmail.com