Stochastic Proximal Langevin Algorithm

Adil Salim
Joint work with Dmitry Kovalev and Peter Richtarik

KAUST

August 6, 2019

Outline

Introduction

Consider a probability density over Euclidean space X :

$$\mu^\star(x) \propto \exp(-U(x))$$

where U is convex.

- Goal ? Sample from the distribution $\mu^\star$.
- Why ? Machine learning/ Signal processing/ Bayesian statistics problems.
- How ? Generate a sequence of random variables (x_n) in X s.t.

$$\mu_n \longrightarrow \mu^\star$$

where $x_n \sim \mu_n$.

Langevin Monte Carlo

Langevin Monte Carlo (LMC) is a sampling algorithm :

$$x_{n+1} = x_n - \gamma \nabla U(x_n) + \sqrt{2\gamma} B_{n+1}$$

where $(B_n)_n$ i.i.d r.v with standard gaussian distribution.

Intuition : (x_n) is a discretization of the (continuous time) Langevin equation

$$dX_t = -\nabla U(X_t) dt + \sqrt{2} dB_t$$

and it is well known that $X_t \longrightarrow \mu^\star(x) \propto \exp(-U(x))$.

Analysis of LMC

- Asymptotic theory : Well known
- Non-asymptotic theory :

$$\mathrm{KL}(\bar{\mu}_n \mid \mu^\star) \leq \frac{1}{2\gamma(n+1)} W^2(\mu_0, \mu^\star) + \mathcal{O}(\gamma)$$

If U α-strongly convex,

$$W^2(\mu_n, \mu^\star) \leq (1-\gamma\alpha)^n W^2(\mu_0, \mu^\star) + \mathcal{O}\left(\frac{\gamma}{\alpha}\right)$$

1. Last 5 years (Dalalyan, Durmus, Moulines, ...) :
   Based on Langevin equation
2. Last year (Wibisono, Bernton, Durmus *et. al.*, Jordan *et al.*, ...) :
   Based on convex optimization (in a measure space) — much "simpler" proofs.
   Intuition : (μ_n) is a discretization of the (continuous time) Wasserstein Gradient Flow of $\mathrm{KL}(\cdot|\mu^\star)$.

Outline

Introduction

Results

Gradient Flows

Experiments

Conclusion

Problem

We consider the case where U is nonsmooth and stochastic.

Why? SVM, logistic regression, structured priors/regularizations: overlapping group lasso, total variation regularization...

Sample from $\mu^\star \propto \exp(-U)$ where

$$U(x) = F(x) + \sum_{i=1}^{N} G_i(x) \tag{1}$$

- $F(x) = \mathbb{E}_\xi(f(x,\xi))$, α-**strongly convex** ($\alpha \geq 0$), smooth, bounded variance of stochastic gradients.
- $G_i(x) = \mathbb{E}_\xi(g_i(x,\xi))$, Lipschitz.

Current approach: Stochastic Subgradient Langevin Algorithm [Durmus *et al.*'18]

Algorithm

Stochastic Proximal Langevin Algorithm (SPLA):

$$x_{n+1} = T_\gamma(x_n - \gamma \nabla f(x_n, \xi_{n+1}), \xi_{n+1}) + \sqrt{2\gamma} B_{n+1},$$

where

$$T_\gamma(x, \xi) = \mathrm{prox}_{\gamma g_N(\cdot,\xi)} \circ \ldots \circ \mathrm{prox}_{\gamma g_1(\cdot,\xi)}(x),$$

where

$$\mathrm{prox}_g(x) = \arg\min_y \frac{1}{2}\|x - y\|^2 + g(y).$$

Related to **Stochastic Passty Algorithm** [Passty'79], [S. *et al.*'18].
Splitting algorithm.

KL divergence, Wasserstein distance

Kullback-Leibler divergence KL: $\mathrm{KL}(\mu|\nu) := \int \mu(x) \log(\frac{\mu(x)}{\nu(x)}) dx$.
Not a distance but $\mathrm{KL}(\mu|\nu) \geq 0$ with equality iff $\mu = \nu$.

Wasserstein distance W : $W^2(\mu,\nu) := \inf \mathbb{E}(\|X - Y\|^2)$ where the inf (in fact a min) is w.r.t. all r.v (X, Y) such that $X \sim \mu$ and $Y \sim \nu$.
Wasserstein space $(\mathcal{P}_2(\mathrm{X}), W)$ metric space.

Reformulation of the problem

$$\mathrm{KL}(\mu|\mu^\star) = (\mathcal{E}(\mu) + \mathcal{H}(\mu)) - (\mathcal{E}(\mu^\star) + \mathcal{H}(\mu^\star))$$

where Potential energy

$$\mathcal{E}(\mu) := \int U(x) d\mu(x) = \int F(x) d\mu(x) + \sum_{i=1}^{N} \int G_i(x) d\mu(x),$$

and Entropy

$$\mathcal{H}(\mu) := \int \mu(x) \log(\mu(x)) dx.$$

SPLA solves

$$\min_{\mu \in \mathcal{P}_2(\mathsf{X})} \mathcal{F}(\mu) := \mathcal{E}(\mu) + \mathcal{H}(\mu).$$

Results

Theorem 1

$$2\gamma\left(\mathcal{F}(\tilde{\mu_n}) - \mathcal{F}(\mu^\star)\right) \le (1-\gamma\alpha)W^2(\mu_n,\mu^\star) - W^2(\mu_{n+1},\mu^\star) + \gamma^2 C. \quad (2)$$

Corollary 2

If $\alpha = 0$

$$\mathrm{KL}(\bar{\mu_n} \mid \mu^\star) \le \frac{1}{2\gamma(n+1)} W^2(\mu_0,\mu^\star) + \mathcal{O}(\gamma).$$

If $\alpha > 0$,

$$W^2(\mu_n,\mu^\star) \le (1-\gamma\alpha)^n W^2(\mu_0,\mu^\star) + \mathcal{O}\left(\frac{\gamma}{\alpha}\right)$$

$$\mathrm{KL}(\tilde{\mu_n} \mid \mu^\star) \le \alpha(1-\gamma\alpha)^{n+1} W^2(\mu_0,\mu^\star) + \mathcal{O}(\gamma).$$

Outline

Gradient Flow (GF) in Euclidean space

The Gradient Flow (GF) associated to U is the solution to the Differential Inclusion

$$\dot{\mathrm{x}}(t) \in -\partial U(\mathrm{x}(t)), \quad t \geq 0. \tag{3}$$

Equivalently, it is the solution to

$$\{U(\mathrm{x}(t)) - U(a)\} \leq -\frac{1}{2}\frac{d}{dt}\|\mathrm{x}(t) - a\|^2, \quad \forall a \in \mathrm{X}, \forall t \geq 0.$$

(Euclidean space - Continuous time)

Stochastic Passty Algorithm

Stochastic Passty Algorithm can be seen as a discretization of the Differential Inclusion.
Easier to see of the (particular case of) Gradient Descent algorithm:

$$\frac{x_{n+1} - x_n}{\gamma} = -\nabla U(x_n). \tag{4}$$

Analysis:

$$\{U(x_{n+1}) - U(x_\star)\} \leq \frac{\|x_n - x_\star\|^2 - \|x_{n+1} - x_\star\|^2}{2\gamma}.$$

(Euclidean space - Discrete time)

GF in Wasserstein space

In the Wasserstein space, a GF $(\mu_t)_{t\geq 0}$ associated to a "convex" function $\mathcal{F} : \mathcal{P}_2(\mathsf{X}) \to \mathbb{R}$ is defined as solution to

$$\{\mathcal{F}(\mu_t) - \mathcal{F}(\nu)\} \leq -\frac{1}{2}\frac{d}{dt}W^2(\mu_t, \nu), \quad \forall \nu \in \mathcal{M}(\mathsf{X}), \forall t \geq 0.$$

(Measure space - Continuous time)

Examples of GF[1]

1. (B_t) Brownian motion, $\sqrt{2}B_t \sim \mu_t$. (μ_t) GF associated to $\mathcal{H}(\mu)$.
2. More generally, (X_t) solution to Langevin equation

$$dX_t = -\nabla U(X_t)dt + \sqrt{2}dB_t,$$

$X_t \sim \mu_t$. (μ_t) GF associated to

$$\mathcal{F}(\mu) = \mathcal{H}(\mu) + \mathcal{E}(\mu) = \mathrm{KL}(\mu|\mu^\star) + C$$

[1] see [Ambrosio *et al.*'08]

What we prove

Inspired from [Durmus *et al.*'18], we prove:

$$\{\mathcal{F}(\mu_n) - \mathcal{F}(\mu^\star)\} \leq \frac{W^2(\mu_n, \mu^\star) - W^2(\mu_{n+1}, \mu^\star)}{2\gamma} + C.$$

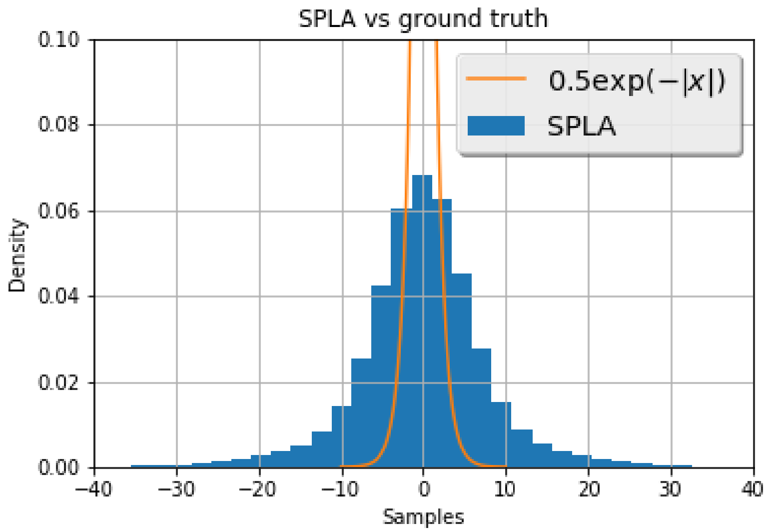
(Measure space - Discrete time)

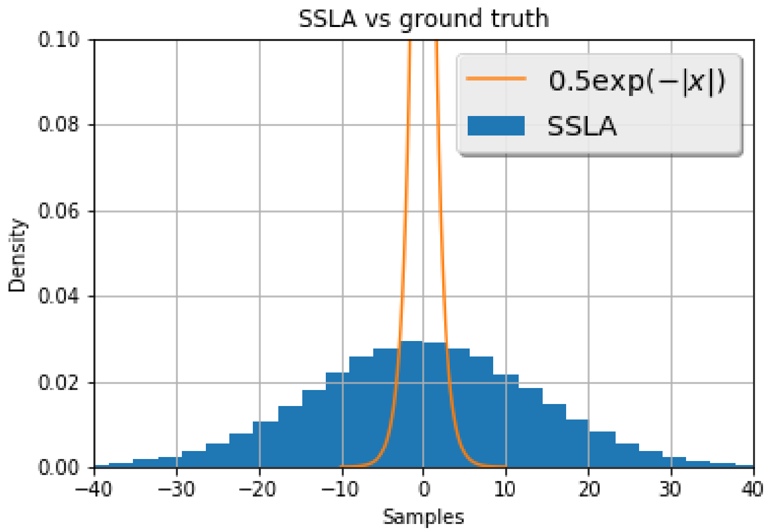
1. Prove it for $\mathcal{E}(\mu) - \mathcal{E}(\mu^\star)$ (**Optimization: Stochastic Passty**)
2. Prove it for $\mathcal{H}(\mu) - \mathcal{H}(\mu^\star)$ (Gradient flow)
3. Sum the inequalities.

Outline

Stochastic proximal vs Stochastic subgradient

$$U(x) = G_1(x) = |x|, \quad G_1(x) = \mathbb{E}(|x| + x\xi), \quad \xi \sim N(0,1)$$

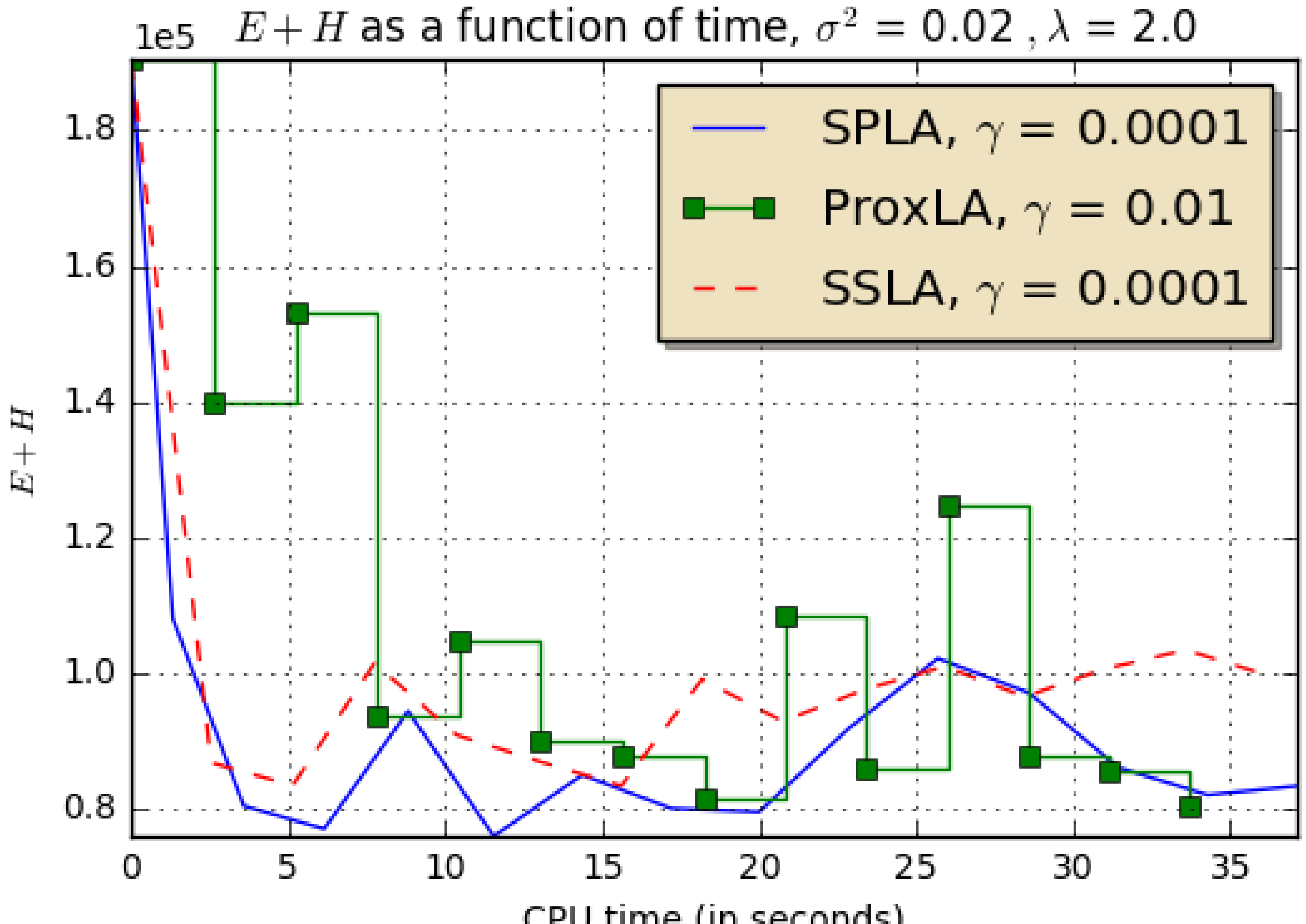




Bayesian trend filtering on graphs

$G = (V, E)$ graph, $y \in \mathbb{R}^V$.

$$U(x) = \frac{1}{2}\|x - y\|^2 + \lambda \mathrm{TV}(x, G), \quad \mathrm{TV}(x, G) = \sum_{\{i,j\} \in E} |x(i) - x(j)|$$



Outline

Main ideas

- Langevin as discretization of Wasserstein GF
- Discretization using splitting and **stochastic proximity operators**
- Generalization of previous results.