Langevin Monte Carlo as an Optimization Algorithm

Adil Salim

Based on:
Durmus, Majewski, Miasojedow, JMLR 2019
S., Kovalev, Richtárik, NeurIPS 2019 (Spotlight)

KAUST

Outline

Optimization vs. Simulation

Consider U convex function. Two important problems:

1. [Optimization Literature] Find

$$x^\star = \arg\min_x U(x) = \arg\max \exp(-U(x))$$

2. [Sampling Literature] Sample

$$\pi(x) \propto \exp(-U(x))$$

$\sim$ Maximum a Posterori vs. Sampling a Posteriori.

Optimization

Smooth convex function $U : \mathbb{R}^d \to \mathbb{R}$.

Problem:

$$x_\star = \arg\min_x U(x)$$

Algorithm:

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

Or,

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

Euler discretization of the **Gradient Flow** of U

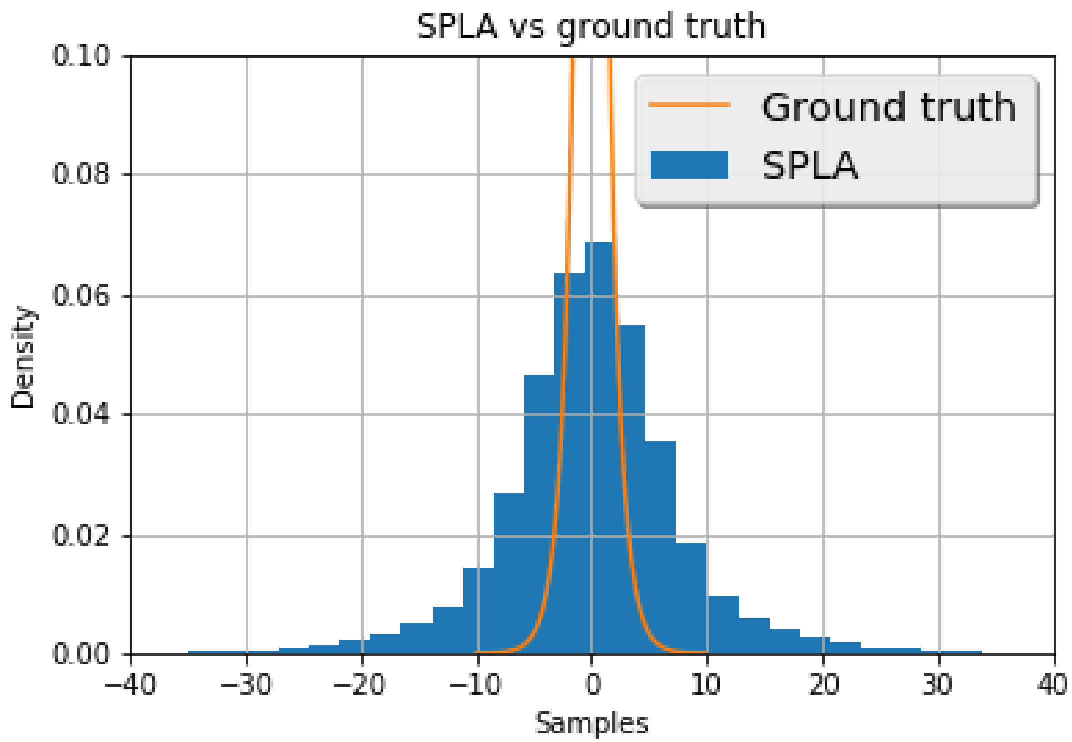
$$\mathrm{x}'(t) = -\nabla U(\mathrm{x}(t)),$$

Typically $U(\mathrm{x}(t)) - U(x_\star) = \mathcal{O}(1/t)$.

Sampling

Problem:

$$\pi(x) \propto \exp(-U(x)).$$



Langevin Monte Carlo

Algorithm: Langevin Monte Carlo (LMC)

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

where $(B_n)_n$ i.i.d standard gaussian **random variables**.

Looks like Gradient Descent!

Euler discretization of **Langevin equation**: (B_t) Brownian motion,

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

Typically $\mathrm{KL}(\mu(t)|\pi) = \mathcal{O}(1/t)$, where $X_t \sim \mu(t)$.

Analysis of LMC

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

$$D(x_n, \pi) \leq \frac{C}{n^\alpha}$$

where $D(x_n, p)$ is some "distance" between π and the distribution of x_n.

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

Goal of this talk : Analysis of LMC using convex optimization.

Outline

Wasserstein Space

Space of probability distribution

$$\mathcal{P}(\mathsf{X}) := \{\mu : \int \|x\|^2 d\mu(x) < \infty\}$$

Wasserstein distance over this space

$$W^2(\mu, \nu) := \inf \mathbb{E}(\|X - Y\|^2), \quad \forall \mu, \nu \in \mathcal{P}_2(\mathsf{X}),$$

where the inf is w.r.t. all r.v (X, Y) such that $X \sim \mu$ and $Y \sim \nu$.
Example: $W^2(\delta_x, \delta_y) = \|x - y\|^2$.

Optimization problem in Wasserstein space

Smooth "convex" function $\mathcal{F} : \mathcal{P}(\mathsf{X}) \to \mathbb{R}$.

Problem:

$$\mu_\star = \arg\min_\mu \mathcal{F}(\mu)$$

Gradient Flow of $\mathcal{F}$ [Ambrosio *et al.*'08]

$$\mu'(t) = -\nabla_W \mathcal{F}(\mu(t))$$

Typically, $\mathcal{F}(\mu(t)) - \mathcal{F}(\mu_\star) = \mathcal{O}(1/t)$.

Examples of Wasserstein Gradient Flows: I. Entropy

Let (B_t) Brownian motion, $\sqrt{2}B_t \sim \mu(t)$. Then, GF $(\mu(t))$ associated to

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

Examples of Wasserstein Gradient Flows: II. Potential

Let $(\mathrm{x}(t))$ (classical) GF of U:

$$\mathrm{x}'(t) = -\nabla U(\mathrm{x}(t)), \quad \mathrm{x}(t) \sim \mu(t)$$

Then, GF $(\mu(t))$ associated to

$$\mathcal{E}(\mu) := \int U(x) d\mu(x).$$

III. Combination of the two last

Let (X_t) solution to Langevin equation

$$dX_t = \underbrace{-\nabla U(X_t)dt}_{\text{GF of } \mathcal{E}} + \underbrace{\sqrt{2}dB_t}_{\text{GF of } \mathcal{H}}, \quad X_t \sim \mu(t).$$

Then, GF $(\mu(t))$ associated to [Jordan *et al.*'98]

$$\mathcal{F}(\mu) := \mathcal{H}(\mu) + \mathcal{E}(\mu).$$

What is $\mathcal{F}$?

Recall $\pi \propto \exp(-U)$, $\mathcal{F}(\mu) = \mathcal{H}(\mu) + \int U d\mu$.

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 $\mu = \nu$.

Then,

$$\mathrm{KL}(\mu|\pi) = \mathcal{F}(\mu) - \mathcal{F}(\pi) = \mathcal{F}(\mu) + C.$$

Summary: Langevin is GF of KL

Let $\pi \propto \exp(-U)$.
Smooth "convex" function $\mathrm{KL}(\cdot|\pi) : \mathcal{P}(\mathsf{X}) \to \mathbb{R}$.

Problem:

$$\pi = \arg\min_{\mu} \mathrm{KL}(\mu|\pi) = \arg\min_{\mu} \mathcal{F}(\mu).$$

Gradient Flow of KL (= Continuous time Gradient Descent): $(\mu(t))$ such that $X_t \sim \mu(t)$ where

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

Typically, $\mathcal{F}(\mu(t)) - \mathcal{F}(\pi) = \mathrm{KL}(\mu(t)|\pi) = \mathcal{O}(1/t)$.

What about LMC?

Discrete Gradient Flow of KL (=Gradient Descent): Langevin Monte Carlo

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

Not just an analogy : One actually prove convergence rates for KL by imitating the proof of Gradient Descent. [Durmus *et al.*'19]

Table: Complexity results for Langevin algorithm.

U	Rate
convex	$\mathrm{KL}(\mu_{\hat{x}_n} \mid \pi) \leq \frac{1}{2\gamma(n+1)} W^2(\mu_{x_0}, \pi) + \mathcal{O}(\gamma)$
α-strongly convex	$W^2(\mu_{x_n}, \pi) \leq (1-\gamma\alpha)^n W^2(\mu_{x_0}, \pi) + \mathcal{O}\left(\frac{\gamma}{\alpha}\right)$

Outline

Nonsmooth optimization

Convex optimization goes far beyond Gradient Descent, *e.g.* nonsmooth optimization

Problem:

$$\min_x U(x) := F(x) + G(x)$$

where F smooth, G nonsmooth.

Algorithm:

$$x_{n+1} = \mathrm{prox}_{\gamma G}(x_n - \gamma \nabla F(x_n))$$

where $\mathrm{prox}_{\gamma G}(x) := \arg\min_y G(y) + \frac{1}{2\gamma}\|y - x\|^2$.

Nonsmooth **and** Stochastic optimization

Convex optimization goes far beyond Gradient Descent, *e.g.* stochastic optimization

Problem:

$$\min_x U(x) := F(x) + G(x)$$

where $F(x) = \mathbb{E}_\xi(f(x,\xi))$ smooth, $G(x) = \mathbb{E}(g(x,\xi))$ nonsmooth, ξ random variable.

Algorithm:[Bianchi *et al.*'17]

$$x_{n+1} = \mathrm{prox}_{\gamma g(\cdot,\xi_{n+1})}(x_n - \gamma \nabla f(x_n, \xi_{n+1}))$$

where (ξ_n) i.i.d.

Stochastic Proximal Langevin Algorithm

Let $\pi \propto \exp(-U) = \exp(-F)\exp(-G)$.

Problem:

$$\pi = \arg\min_{\mu} \mathrm{KL}(\mu|\pi) = \arg\min_{\mu} \mathcal{F}(\mu),$$

where $\mathcal{F}(\mu) = \mathcal{H}(\mu) + \int U d\mu = \mathcal{H}(\mu) + \int F d\mu + \int G d\mu$.

Stochastic Proximal Langevin Algorithm:[S.*et al*'19]:

$$x_{n+1} = \mathrm{prox}_{\gamma g(\cdot,\xi_{n+1})}(x_n - \gamma \nabla f(x_n, \xi_{n+1})) + \sqrt{2\gamma} B_{n+1}$$

Convergence rates

We see SPLA as an optimization algorithm in Wasserstein space. Recall $U(x) = F(x) + G(x) = \mathbb{E}(f(x,\xi)) + \mathbb{E}(g(x,\xi))$.

Table: Complexity results for SPLA.

F	Rate
convex	$\mathrm{KL}(\mu_{\hat{x}_n} \mid \pi) \leq \frac{1}{2\gamma(n+1)} W^2(\mu_{x_0}, \pi) + \mathcal{O}(\gamma)$
α-strongly convex	$W^2(\mu_{x_n}, \pi) \leq (1-\gamma\alpha)^n W^2(\mu_{x_0}, \pi) + \mathcal{O}\left(\frac{\gamma}{\alpha}\right)$
α-strongly convex	$\mathrm{KL}(\mu_{\tilde{x}_n} \mid \pi) \leq \alpha(1-\gamma\alpha)^{n+1} W^2(\mu_{x_0}, \pi) + \mathcal{O}(\gamma)$

Simulations: Toy model

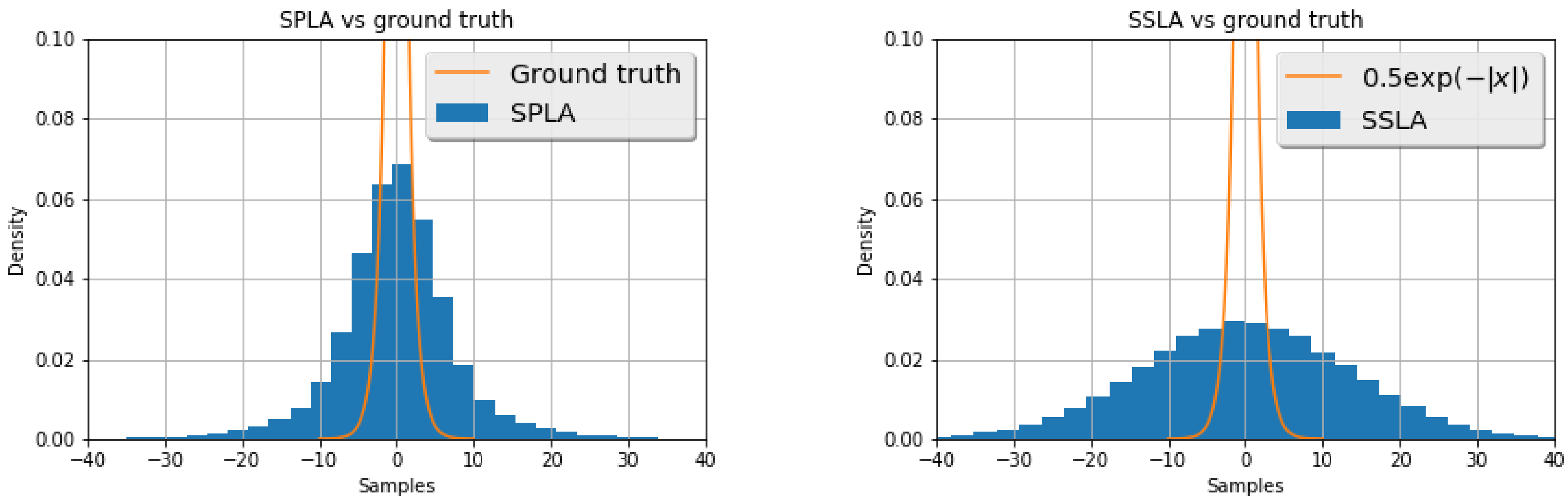


Figure: Comparison between histograms of SPLA and SSLA and the true density $0.5\exp(-|x|)$.

Simulations: Trend filtering on graphs

Let $G = (V, E)$ graph.

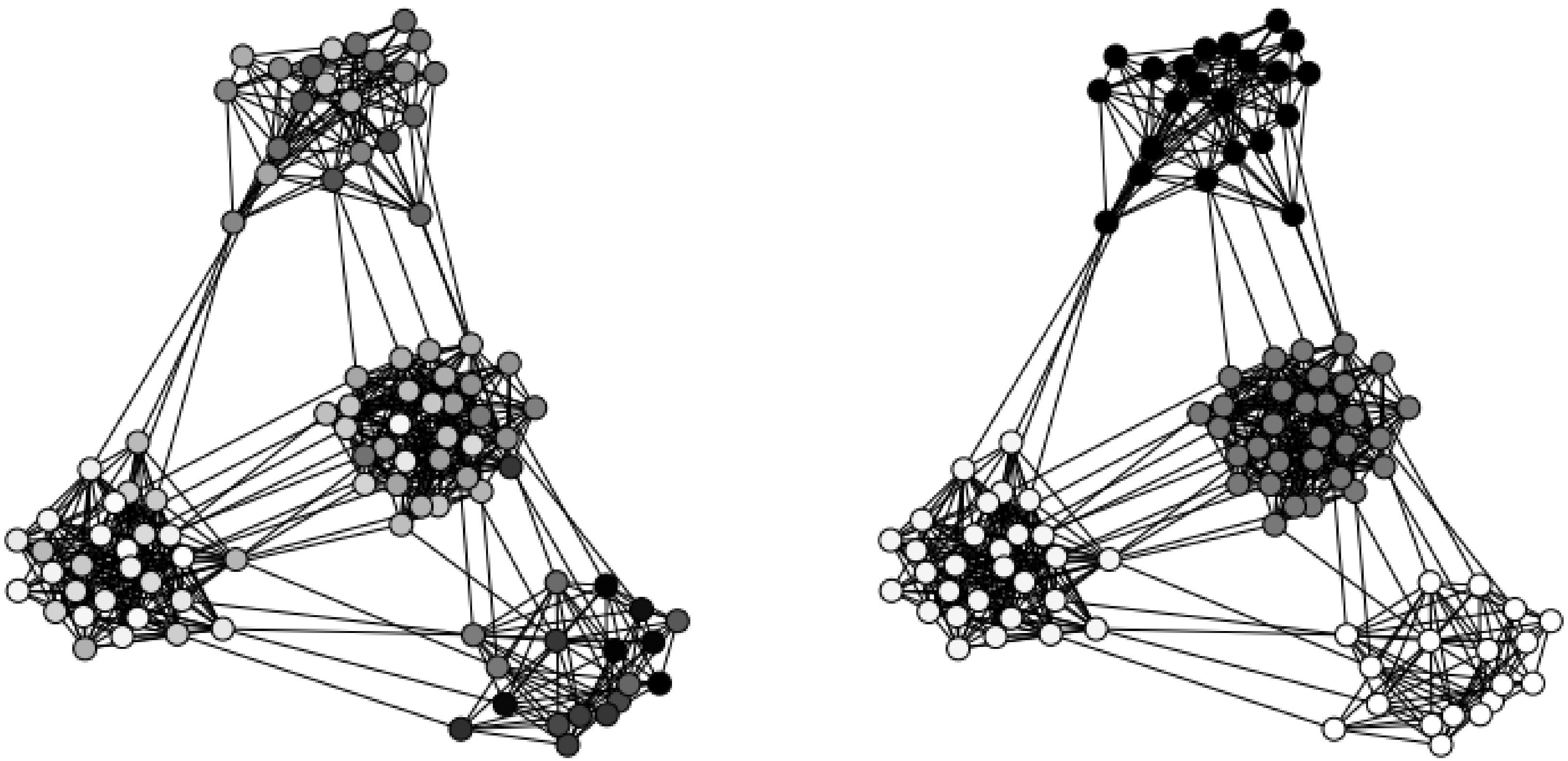


Figure: The signal is the grayscale of the node. Left: Noised signal over the nodes. Right: Sought signal.

Bayesian context

Trend filtering on graphs [Wang *et al.*'16]. Let

$$\pi \propto \exp(-U) = \underbrace{\exp(-F)}_{\text{likelihood}} \underbrace{\exp(-G)}_{\text{prior}},$$

where $F(x) = \frac{1}{2}\|x - a\|^2$ and

$$G(x) = \mathrm{TV}(x, G) = \sum_{\{i,j\} \in E} |x(i) - x(j)| \propto \mathbb{E}_e(|x(e_1) - x(e_2)|),$$

where e random edge.

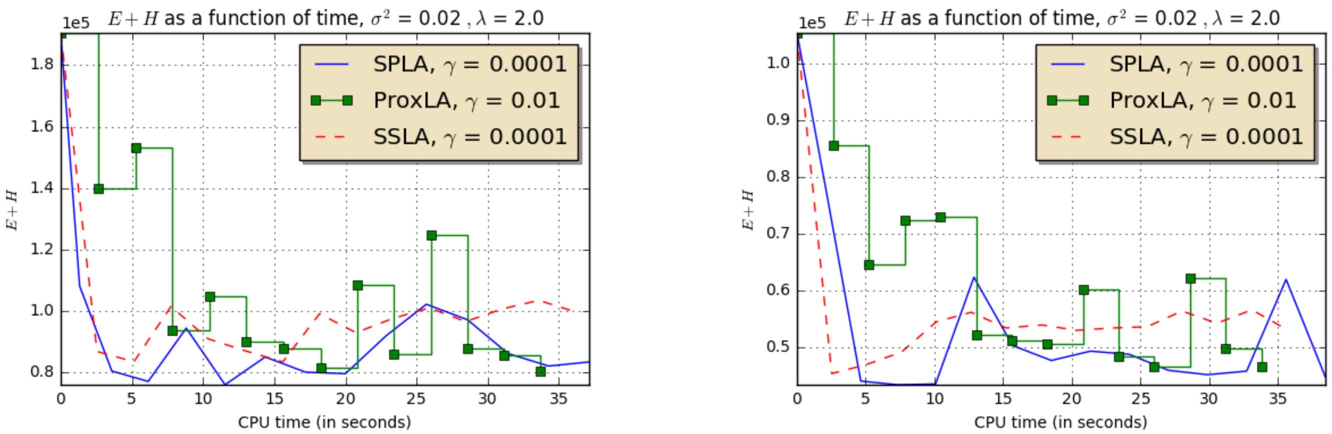


Figure: $\mathcal{F} = \mathcal{H} + \mathcal{E}_U$ as a function of CPU time over the Facebook graph.

Outline