Sampling via Convex Optimization

Adil Salim

KAUST, Sping 2019

Outline

Introduction

Optimization in Euclidean space

Optimization in the space of probability measures

Analysis of Langevin Monte Carlo

Conclusion

Introduction

Consider a probability density over Euclidean space X :

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

where U is convex and smooth.

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

$$x_n \longrightarrow \pi$$

in distribution.

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 : LMC 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 $X_t \longrightarrow \pi(x) \propto \exp(-U(x))$.

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 (last part of the presentation)[1].

[1] Based on [Durmus *et al.*'18]

Outline

Gradient Flow (GF) in Euclidean space

Consider a smooth convex function $F : \mathsf{X} \to \mathbb{R}$. The Gradient Flow (GF) associated to F is the solution to the ODE

$$\dot{\mathsf{x}}(t) = -\nabla F(\mathsf{x}(t)), \quad t \geq 0. \tag{1}$$

Equivalently (prove it), it is the solution to

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

(Euclidean space - Continuous time)

Lyapunov functions

Three Lyapunov functions are usually used to study GF.
Let $x_\star \in \arg\min F$.

1. $L_1(t) = F(\mathrm{x}(t)) - F(x_\star)$. $\dot{L_1}(t) \leq 0$. Therefore, $F(\mathrm{x}(t)) \searrow$.
2. $L_2(t) = \frac{1}{2}\|\mathrm{x}(t) - x_\star\|^2$. Using (Euclidean space - Continuous time),

$$0 \leq \{F(\mathrm{x}(t)) - F(x_\star)\} \leq -\dot{L_2}(t).$$

   Moreover, using the convexity of F

$$F(\overline{\mathrm{x}}(t)) - F(x_\star) \leq \frac{\|\mathrm{x}(0) - x_\star\|^2 - \|\mathrm{x}(t) - x_\star\|^2}{2t}.$$

3. $L_3(t) = tL_1(t) + L_2(t)$. Using $\dot{L_3}(t) \leq 0$, using the convexity of F,

$$F(\mathrm{x}(t)) - F(x_\star) \leq \frac{\|\mathrm{x}(0) - x_\star\|^2 - \|\mathrm{x}(t) - x_\star\|^2}{2t}. \qquad (2)$$

Gradient Descent Algorithm

The Gradient algorithm with step $\gamma > 0$

$$\frac{x_{n+1} - x_n}{\gamma} = -\nabla F(x_n) \tag{3}$$

can be seen as a discretization of the GF. Therefore, its analysis follows the same lines.

For example, here is an analysis using a discrete version of L_2.

Analysis of Gradient Algorithm

$$
\begin{aligned}
\|x_{n+1} - x_\star\|^2 &= \|x_n - x_\star\|^2 + \gamma^2 \|\nabla F(x_n)\|^2 - 2\gamma \langle \nabla F(x_n), x_n - x_\star \rangle \\
&\leq \|x_n - x_\star\|^2 + \gamma^2 \|\nabla F(x_n)\|^2 - 2\gamma \left\{ F(x_n) - F(x_\star) \right\} \\
&\leq \|x_n - x_\star\|^2 + \gamma^2 \|\nabla F(x_n)\|^2 - 2\gamma \left\{ F(x_{n+1}) - F(x_\star) \right\} \\
&\quad - 2\gamma \left\{ F(x_n) - F(x_{n+1}) \right\} \\
&\leq \|x_n - x_\star\|^2 - \gamma^2 \left(1 - \gamma L\right) \|\nabla F(x_n)\|^2 - 2\gamma \left\{ F(x_{n+1}) - F(x_\star) \right\}
\end{aligned}
$$

where the last inequality comes from the smoothness of F :

$$
F(x_{n+1}) - F(x_n) \leq \langle \nabla F(x_n), x_{n+1} - x_n \rangle + \frac{L}{2} \|x_{n+1} - x_n\|^2 = -\gamma \left(1 - \frac{\gamma L}{2}\right) \|\nabla F(x_n)\|^2.
$$

Hence,

$$
\left\{ F(x_{n+1}) - F(x_\star) \right\} \leq \frac{\|x_n - x_\star\|^2 - \|x_{n+1} - x_\star\|^2}{2\gamma}
$$

(Euclidean space - Discrete time)

If $\overline{x}_n = \frac{1}{n} \sum_{k=1}^{n} x_k$, using the convexity of F,

$$
F(\overline{x}_n) - F(x_\star) \leq \frac{\|x_0 - x_\star\|^2 - \|x_n - x_\star\|^2}{2\gamma n}.
$$

Outline

GF in the space of probability measures

In the sequel, we assume all the measures μ we consider to have a positive density $\mu(x)$ w.r.t Lebesgue. Let $\mu, \nu \in \mathcal{M}(\mathrm{X})$ probability measures.

Wasserstein distance W_2 : $W_2^2(\mu, \nu) := \inf \mathbb{E}(\|X - Y\|^2)$ where the inf (in fact a min) is w.r.t. all r.v (X, Y) such that $X \sim \mu$ and $Y \sim \nu$. Similar to $\|\cdot\|^2$.

In the space of probability measures, a GF $(\mu_t)_{t \geq 0}$ associated to a "convex" function $\mathcal{F} : \mathcal{M}(\mathrm{X}) \to \mathbb{R}$ is defined to be a solution to

$$\{\mathcal{F}(\mu_t) - \mathcal{F}(\nu)\} \leq -\frac{1}{2}\frac{d}{dt}W_2^2(\mu_t, \nu), \quad \forall \nu \in \mathcal{M}(\mathrm{X}), \forall t \geq 0.$$

(Measure space - Continuous time)

Examples of GF[2]

1. (B_t) Brownian motion, $\sqrt{2}B_t \sim \mu_t$. (μ_t) GF associated to
$$\mathcal{H}(\mu) := \int \mu(x)\log(\mu(x))dx.$$
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 $\mathcal{H}(\mu) + \mathcal{E}(\mu)$ where
$$\mathcal{E}(\mu) := \int U(x)d\mu(x).$$

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

Kullback-Liebler

Kullback-Liebler 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 $\mu = \nu$.

From now on, let $\pi(x) \propto \exp(-U(x))$, where $U : \mathsf{X} \to \mathbb{R}$ convex smooth.

Let $\mathcal{F}(\mu) := \mathrm{KL}(\mu|\pi)$. Then,

$$\mathcal{F}(\mu) = \mathcal{F}(\mu) - \mathcal{F}(\pi) = \mathcal{H}(\mu) + \mathcal{E}(\mu) - (\mathcal{H}(\pi) + \mathcal{E}(\pi)). \tag{4}$$

In other words, Langevin is the GF associated to $\mathcal{F}$.

Outline

LMC algorithm

Recall LMC algorithm

$$x_{n+1} = x_n - \gamma \nabla U(x_n) + \sqrt{2\gamma} B_{n+1} \tag{5}$$

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

$$x_n \sim \mu_n$$

and

$$\widetilde{x_{n+1}} := x_n - \gamma \nabla U(x_n) \sim \widetilde{\mu_{n+1}}.$$

We shall prove

$$\{\mathcal{F}(\mu_{n+1}) - \mathcal{F}(\pi)\} \leq \frac{W_2^2(\mu_n, \pi) - W_2^2(\mu_{n+1}, \pi)}{2\gamma} + L\gamma d.$$

(Measure space - Discrete time)

Step 1

Denote d the dimension of X.

Convexity + smoothness :

$$0 \leq U(x_{n+1}) - U(\widetilde{x_{n+1}}) - \langle \nabla U(\widetilde{x_{n+1}}), x_{n+1} - \widetilde{x_{n+1}} \rangle \quad \leq \frac{L}{2} \|\widetilde{x_{n+1}} - x_{n+1}\|^2$$

$$0 \leq U(x_{n+1}) - U(\widetilde{x_{n+1}}) - \langle \nabla U(\widetilde{x_{n+1}}), \sqrt{2\gamma} B_{n+1} \rangle \qquad \leq \frac{L}{2} \|\sqrt{2\gamma} B_{n+1}\|^2.$$

Taking the expectation :

$$\{\mathcal{E}(\mu_{n+1}) - \mathcal{E}(\widetilde{\mu_{n+1}})\} \leq L\gamma d.$$

Step 2 : "Gradient Descent"

First, for every $y \in \mathrm{X}$,

$$U(\widetilde{x_{n+1}}) - U(y) \le \frac{\|x_n - y\|^2 - \|\widetilde{x_{n+1}} - y\|^2}{2\gamma}$$

(This is Eq. (Euclidean space - Discrete time))
Then, taking the expectation and then the inf over couplings,

$$\{\mathcal{E}(\widetilde{\mu_{n+1}}) - \mathcal{E}(\pi)\} \le \frac{W_2^2(\mu_n, \pi) - W_2^2(\widetilde{\mu_{n+1}}, \pi)}{2\gamma}.$$

Step 3

Consider the GF (ν_t) associated to $\mathcal{H}$ starting at $\nu_0 = \widetilde{\mu_{n+1}}$. Then,

$$\mathcal{H}(\nu_t) - \mathcal{H}(\pi) \leq \frac{W_2^2(\nu_0, \pi) - W_2^2(\nu_t, \pi)}{2t}.$$

(This is Eq. (2) but in a measure space)
Moreover, $\mu_{n+1} = \nu_\gamma$ because the Brownian motion is the GF associated to $\mathcal{H}$ (up to a factor $\sqrt{2}$, see Slide 13).

$$\{\mathcal{H}(\mu_{n+1}) - \mathcal{H}(\pi)\} \leq \frac{W_2^2(\widetilde{\mu_{n+1}}, \pi) - W_2^2(\mu_{n+1}, \pi)}{2\gamma}.$$

End of the proof

Summing the three inequalities

$$\{\mathcal{F}(\mu_{n+1}) - \mathcal{F}(\pi)\} \leq \frac{W_2^2(\mu_n, \pi) - W_2^2(\mu_{n+1}, \pi)}{2\gamma} + L\gamma d.$$

(Measure space - Discrete time)

Using the convexity of $\mathcal{F}(\mu) = \mathrm{KL}(\mu|\pi)$,

$$\mathrm{KL}(\overline{\mu}_{n+1}|\pi) \leq \frac{W_2^2(\mu_0, \pi)}{2\gamma n} + L\gamma d.$$

Take $\gamma = \mathcal{O}(1/\sqrt{n})$.

Outline

Main ideas

- Gradient Descent as a discretization of Euclidean GF
- Langevin as discretization of measure-valued GF
- Langevin as Gradient algorithm in measure space.

Related topics

- Nesterov acceleration of GF
- Langevin for non convex optimization
- Stein Variational Gradient Descent.