TELECOM Paris

IP PARIS

SVGD: Convergence and New Algorithms

Aymane El Firdoussi
Tutor: Pascal Bianchi

June 21, 2023

Sampling algorithms

Our task is the following:

Input : target density π

$\downarrow$ *algorithm*

Output : $X_1, X_2, \ldots, X_n$ realizations of $X \sim \pi$.

There are already some algorithms: Markov Chain Monte Carlo: Metropolis-Hastings, Langevin, ...

Subject and contributions

- **What:** We study the SVGD algorithm.
- **How:** gradient flows, Wasserstein spaces.
- **Why:** New approach to solve the sampling problem
- **Contributions:** understand its convergence with much more simple conditions / Proposed new algorithms derived from it that can increase its performance

Brief review on Langevin

Langevin algorithm

We want to sample from

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

where $F\ \mathbb{R}^d \longrightarrow \mathbb{R}$ is **convex !** The Langevin algorithm gives a Markov Chain, for which the stationary distribution tends to the target:

$$X_{n+1} = X_n - \epsilon \nabla F(X_k) + \sqrt{2\epsilon}\xi_n \qquad \text{(LMC)}$$

where $\xi_n \sim \mathcal{N}(0,1)$ and $\epsilon > 0$ is a step-size, and X_0 is a random variable of distribution μ_0. This equation is the Euler discretization of the SDE:

$$\mathrm{d}X_t = -\nabla F(X_t)\mathrm{d}t + \sqrt{2}dB_t$$

where B_t is a Brownian motion.

Interesting result by JKO

JKO (see [1]) proved that the well-known Physics' equation called Fokker-Planck can be seen as the continuity equation for the following flow:

$$\dot{\mu_t} = -\nabla_W KL(\mu_t|\pi)$$

Hence, Langevin performs a gradient descent of the KL divergence in the Wasserstein space !

$$\dot{X_t} = -\nabla_W KL(\mu_t|\pi)(X_t) \qquad \text{(WGF)}$$

where μ_t is the density of X_t.

Construction of SVGD

Stein Variational Gradient Descent

SVGD was inspired from the Gradient Descent of the KL on the Wasserstein space.
Let us consider some initial particles $(X_i^0)_{i=1}^N$ following a common density μ_0. We want to apply a mapping T of the form:

$$T(x) = x + \epsilon\phi(x)$$

on each particle $(X_{n+1}^i = T(X_n^i))$ so that to maximally decrease the KL between $\widehat{\mu}_n^N = \frac{1}{N}\sum_{i=1}^N \delta_{X_i^n}$ and π. This can be done by solving:

$$\max_{\phi \in \mathcal{H}} \left\{ -\frac{\mathrm{d}}{\mathrm{d}\epsilon} KL(T_{\#}\mu_n|\pi)|_{\epsilon=0}, \;\; s.t \;\; \|\phi\|_{\mathcal{H}} \leq 1 \right\} \tag{1}$$

- If $\mathcal{H}$ is the L^2 space, then: we the Wasserstein Gradient descent. - If $\mathcal{H}$ is a RKHS, we get SVGF:

$$\phi^*_{\mu,\pi}(y) \propto -\int \nabla F(x)K(x,y) + \nabla_1 K(x,y)\mathrm{d}\mu(x)$$

Algorithm : SVGD

Algorithm 1 Stein Variational Gradient descent (Liu and Wang, 2016)

Require: a set $X_0^1, ..., X_0^N \in \mathcal{X}$ of N particles, a kernel K, the number of iterations M, and a step size $\epsilon > 0$

1: **for** n = 1, 2, ..., M **do**

2: **for** i = 1,2,..,N **do**

3: $X_{n+1}^i = X_n^i - \frac{\epsilon}{N} \sum_{j=1}^{j=N} K(X_n^j, X_n^i) \nabla F(X_n^j) - \nabla_1 K(X_n^j, X_n^i)$

Main contributions

Simple Proof of convergence

Inspired by Salim et al. [11], and MACS203b (Asymptotic statistics).

Theorem (Convergence in population limit)

*Let $(\mu_n)_n$ the sequence of measures associated to the particles generated by SVGD at each time step $n \in \mathbb{N}$. If F is **coercive** and $KL(\mu_0|\pi) < +\infty$, then μ_n converges weakly to π, i.e*

$$\mu_n \Longrightarrow \pi \tag{2}$$

An interesting topic

To our best knowledge, it was not proved that we can interchange the order of the limits:

$$\lim_{l \to +\infty} \lim_{n \to +\infty} \widehat{\mu}_l^n = \pi \tag{3}$$

New algorithms

Noisy SVGD

Algorithm 2 Noisy SVGD

Require: a set $X_0^1, ..., X_0^N \in \mathcal{X}$ of N particles, a kernel K, the number of iterations M, and a step size $\epsilon > 0$

1: **for** n = 1, 2, ..., M **do**

2: **for** i = 1,2,..,N **do**

3: $X_{n+1}^i = X_n^i - \frac{\epsilon}{N} \sum_{j=1}^{j=N} K(X_n^j, X_n^i) \nabla F(X_n^j) - \nabla_1 K(X_n^j, X_n^i) + B_n^i$

Which type of noise ?

Of course, not any noise should improve the algorithm !

If we take a noise that depends on all data points, for example:

$$B_n^i = \frac{\epsilon}{N}\sum_{j=1}^{N} K(X_n^j, X_n^i)\nabla F(X_n^j) - \nabla_1 K(X_n^j, X_n^i) - \epsilon(K(X_n^l, X_n^i)\nabla F(X_n^l) - \nabla_1 K(X_n^l, X_n^i))$$

We obtain the following algorithm, which I call **Stochastic SVGD** (pure invention):

$$X_{n+1}^i = X_n^i - \epsilon(K(X_n^l, X_n^i)\nabla F(X_n^l) - \nabla_1 K(X_n^l, X_n^i)) \qquad \text{(SSVGD)}$$

Very good results !

- SVGD: $\mathcal{O}(N^2 M)$
- Stochastic SVGD: $\mathcal{O}(NM)$

And can simulate densities with non-convex potentials:

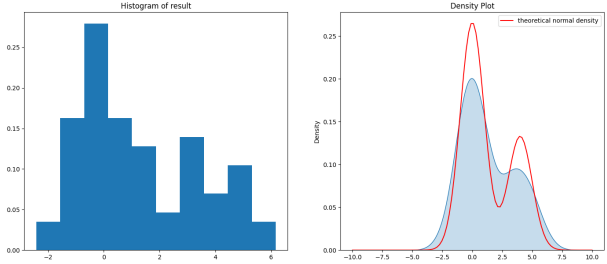

Figure 1: Gaussian mixture $\frac{2}{3}\mathcal{N}(0,1) + \frac{1}{3}\mathcal{N}(4,1)$ with Stochastic SVGD

Decreasing the KL using Langevin and SVGD

Algorithm 3 Langevin SVGD

Require: a set $X_0^1, ..., X_0^N \in \mathcal{X}$ of N particles, a kernel K, the number of iterations M, and a step size $\epsilon > 0$

1: **for** n = 1, 2, ..., M **do**

2: **for** i = 1,2,..,N **do**

3: $$X_{n+1}^i = X_n^i - \frac{\epsilon}{2}\nabla F(X_n^i) + \sqrt{2\epsilon}\xi_n^i - \frac{\epsilon}{2N}(\sum_{j=1}^N \nabla F(X_n^j)K(X_n^j, X_n^j) - \nabla_1 K(X_n^j, X_n^i))$$

Even faster with **Stochastic Langevin SVGD**

Good results too !

$$B_n^i = -\frac{\epsilon}{2}\nabla F(X_n^i) + \sqrt{2\epsilon}\xi_n^i - \frac{\epsilon}{2N}\left(\sum_{j=1}^{N} \nabla F(X_n^j)K(X_n^j, X_n^j) - \nabla_1 K(X_n^j, X_n^i)\right)$$

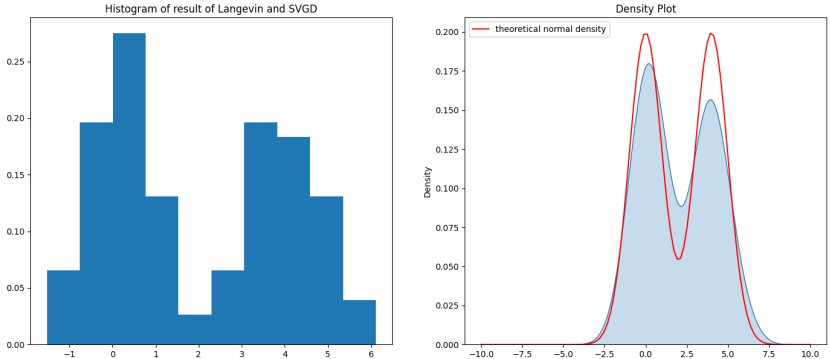

Figure 2: Langevin SVGD

Kernel density estimation: SD-TSIA205

Let now $X_1, ..., X_N$ be random vectors with the same density f. The kernel estimator of f is defined by:

$$\hat{f}(t, X_1, ..., X_N) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{X_i - t}{h}\right) \tag{4}$$

where K is a positive definite kernel.

Implementation of Wasserstein Gradient Descent

What if we do a direct gradient descent on the Kullback-Leibler divergence, we will get a process $(X_n)_n$ that follows:

$$X_{n+1} = X_n - \epsilon \nabla \log\left(\frac{\mu_n(X_n)}{\pi(X_n)}\right)$$

where μ_n is the density of the random variable X_n. Hence

$$X_{n+1} = X_n - \epsilon \left(\nabla F(X_n) + \frac{\nabla \mu_n(X_n)}{\mu_n(X_n)}\right)$$

Kernel estimator of a density :

$$X^i_{n+1} = X^i_n - \epsilon \left(\nabla F(X^i_n) + \frac{\frac{1}{h}\sum_{j=1}^N \nabla K\left(\frac{X^i_n - X^j_n}{h}\right)}{\sum_{j=1}^N K\left(\frac{X^i_n - X^j_n}{h}\right)}\right) \tag{5}$$

Excellent !

Unfortunatly, this was proposed in [4].

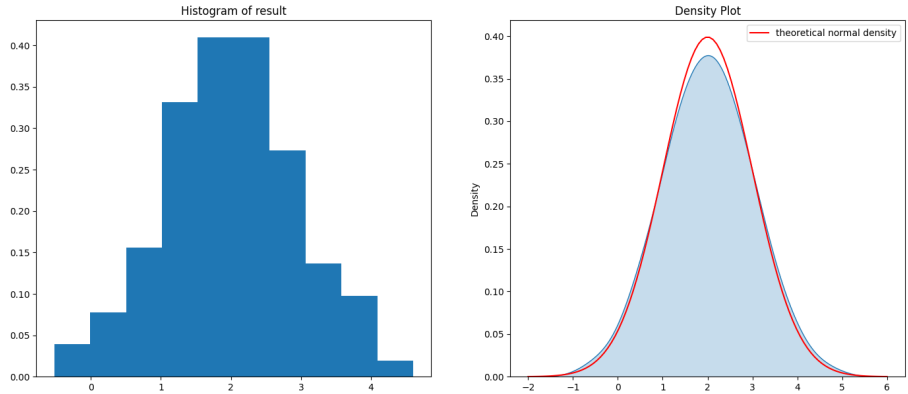

Figure 3: Kernelized WGD

Acknowledgment

First of all, I would like to express my deep gratitude to Prof. **Pascal Bianchi** for his unwavering support throughout this project, and to all members of the jury of Parcours Recherche who helped me make this achievement possible.

Bibliography

[1] Richard Jordan, David Kinderlehrer and Felix Otto, *The Variational Formulation of the Fokker-Planck Equation*, Society of industrial and Applied Mathematics, 1999.
[2] Anna Korba, Adil Salim, Michael Arbel, Giulia Luise and Arthur Gretton, *A Non-Asymptotic Analysis for Stein Variational Gradient Descent*, Advances in Neural Information Processing Systems 33 (NeurIPS 2020).
[3] Qiang Liu, *Stein Variational Gradient Descent as Gradient Flow*, NeurIPS Proceedings.
[4] Yifei Wang, Peng Chen, Wuchen Li, *Projected Wasserstein Gradient Descent for High-Dimensional Bayesian Inference*, Society for Industrial for Applied Mathematics.
[5] Léon Bottou, *Stochastic Gradient Descent tricks*, Springer.

Bibliography

[6] Ian Goodfellow, Yoshua Bengio and Aaron Courville, *Deep Learning*, The MIT Press
[7] J. Gorham and L. Mackey. Measuring sample quality with kernels. In International Conference on Machine Learning (ICML), 2017.
[8] Filippo Santambrogio, Optimal Transport for applied mathematicians, calculus of variations, PDEs and Modelling, Birkhäuser, June 2008
[9] Cédric Villani, Optimal transport, old and new, Springer, June 2008.
[10] Qiang Liu, Dilin Wang, *Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm*, NeurIPS Proceedings.
[11] Adil Salim, Lukang Sun, Peter Richtarik, A Convergence Theory for SVGD in the Population Limit under Talagrand's Inequality T1, Proceedings of Machine Learning Research.