I said I would not write on SVGD anymore...

Adil Salim

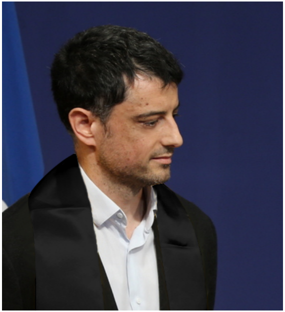
Microsoft

ICERM

May 2024

Joint work with

Victor Priser



Pascal Bianchi

Ongoing work. References, suggestions, comments are welcome!

Sampling

- Bayesian inference in machine learning
- Generative models

"Pikachu eating a sandwich"

Outline

Outline

Stein Variational Gradient Descent

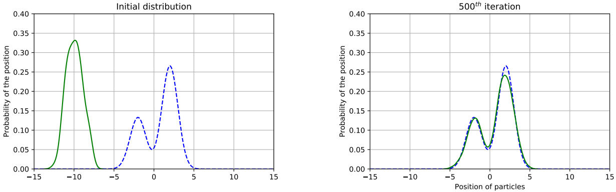
SVGD [Liu and Wang, 2016] is an algorithm to **sample from $\mu_\star \propto \exp(-F)$, where F L-smooth and nonconvex.**
SVGD maintains a set of N particles $x^1, \ldots, x^N$.

$$x^i_{k+1} = x^i_k - \frac{\gamma}{N} \sum_{j=1}^{N} \nabla F(x^j_k) K(x^i_k, x^j_k) - \nabla_2 K(x^i_k, x^j_k),$$

where $K(x, y)$ is a kernel associated to a Reproducing Kernel Hilbert Space H.

Low dimension vs high dimension

$$\mu_\star(x) \propto \exp(-F(x))$$



Simulation from [KSA$^+$20] (Code from Q. Liu)

However, in higher dimension SVGD particles can collapse due to the deterministic updates [Ba et al., 2021].

What do we know about the convergence of SVGD?

Let μ_k^N be the empirical measure of SVGD at iteration k, *i.e.*,

$$\mu_k^N = \frac{1}{N} \sum_{j=1}^{N} \delta_{x_k^j} \tag{1}$$

	k small	**k large**
N large	$\mathrm{KSD}(\mu_k^\infty \vert \mu_\star) < \frac{C}{k}$ [KSA+20]	$W_1(\mu_k^\infty, \mu_\star) \to 0$ [SSR22]
N small	$\mathrm{KSD}(\mu_k^N \vert \mu_\star) < \frac{C'}{k}$, $k < \log\log(N)$ [Shi and Mackey, 2024]	**???**

Remarks on the asymptotics of μ_k^N when $k \to \infty$

When $N < \infty$,

1. $(\mu_k^N)_k$ does not converge to $\mu_\star$ as $k \to \infty$.
   Because μ_k^N is discrete with $N < \infty$ masses whereas $\mu_\star$ has a continuous density.
2. The best hope is for $(\mu_k^N)_k$ to converge to something that converges to $\mu_\star$ as N grows.
3. Even if we were able to show that (μ_k^N) converges to some $\mathscr{L}^N$ as $k \to \infty$ (already non trivial, the particles could diverge), $\mathscr{L}^N$ would probably not converge to $\mu_\star$ as $N \to \infty$.
   Because of particles collapse in SVGD.

Outline

We study noisy SVGD

Noisy SVGD is SVGD with Langevin regularization.

$$x_{k+1}^i = x_k^i - \frac{\gamma_k}{N} \sum_{j=1}^{N} \nabla F(x_k^j) K(x_k^i, x_k^j) - \nabla_2 K(x_k^i, x_k^j) - \varepsilon \gamma_k \nabla F(x_k^i) + \sqrt{2\gamma_k \varepsilon} \xi_k^i$$

where $\varepsilon > 0$ is noise parameter, $\gamma_k \to 0$ and $(\xi_k^i)_{i,k}$ i.i.d standard Gaussian.
Our goal: describe the "limit" $\mathscr{L}^N$ of noisy SVGD as $k \to \infty$. First remark: $\mu_\star \notin \mathscr{L}^N$.

Results

Let μ_k^N be the empirical measure of noisy SVGD and $\overline{\mu_k^N}$ its empirical average over time k. We view them as random variables in the metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$.

1. **"Limit" of noisy SVGD**:
   For every $N > 0$, the sequence of r.v. $(\overline{\mu_k^N})_k$ is tight. Therefore $(\overline{\mu_k^N})_k$ converges in distribution as $k \to \infty$ to the set of its cluster points $\mathscr{L}^N$.
2. **Description of the "limit" I**:
   The set of r.v. $\cup_{N>0}\mathscr{L}^N$ is tight. Therefore $(\mathscr{L}^N)_N$ "converges" in distribution as $N \to \infty$ to the set of its cluster points $\mathscr{L}^\infty$.
3. **Description of the "limit" II**: $\mathscr{L}^\infty = \{\mu_\star\}$ a.s.

Corollaries

All together our results imply

- $$\frac{\sum_{\ell=1}^{k} \gamma_\ell W_2(\mu_\ell^N, \mu_\star)}{\sum_{\ell=1}^{k} \gamma_\ell} \xrightarrow[k,N\to\infty]{\mathbb{P}} 0 \quad (2)$$

- Under Log Sobolev Inequality,

$$W_2(\mu_k^N, \mu_\star) \xrightarrow[k,N\to\infty]{\mathbb{P}} 0 \quad (3)$$

The regime $N \ll k$ is new.

Outline

Interpolated process level

For each particle trajectory $(x_k^i)_k$ we define the interpolated trajectory $x^i : \mathbb{R}_+ \to \mathbb{R}^d$.

Let $\mathcal{C}$ the set of continuous functions from $\mathbb{R}_+$ to $\mathbb{R}^d$ endowed with the topology of uniform convergence on compact sets. Then, $x^i \in \mathcal{C}$. Moreover, $x^i(t+\cdot) \in \mathcal{C}$ for every $t \geq 0$.

Next we define the empirical measure of the shifted interpolated trajectories

$$\mu^N(t) = \frac{1}{N}\sum_{j=1}^{N} \delta_{x^j(t+\cdot)} \tag{4}$$

We view $\mu^N(t)$ as a r.v. in the metric space $(\mathcal{P}_2(\mathcal{C}), W_2)$.

"Limit" of noisy SVGD

Under growth assumption on F:

For every $N > 0$, the sequence of r.v. $(\overline{\mu^N(t)})_t$ is tight. Therefore $(\overline{\mu^N(t)})_t$ converges in distribution as $k \to \infty$ to the set of its cluster points $\mathscr{L}^N$.

This time each element of $\mathscr{L}^N$ is a random measure supported by N continuous functions, *i.e.*, trajectories (instead of N points as before).

Description of the "limit" I

Under growth assumption on F:

The set of r.v. $\cup_{N>0}\mathscr{L}^N$ is tight. Therefore $(\mathscr{L}^N)_N$ "converges" in distribution as $N \to \infty$ to the set of its cluster points $\mathscr{L}^\infty$.

It remains to relate $\mathscr{L}^\infty$ and $\mu_\star$.

McKean Vlasov equation

$$dX_t = -b(X_t, \mathcal{L}(X_t))dt + \sqrt{2\varepsilon} dB_t,$$
$$b(x, \mu) := \int \mu(dy) \nabla F(y) K(x, y) - \nabla_2 K(x, y) + \varepsilon \nabla F(x)$$

where (B_t) standard Brownian motion.

The trajectory of one particle of Noisy SVGD is a discretization of MKV equation in time and space.

A McKean Vlasov measure = the law of a (weak) solution $(X_t)_t$ of the MKV equation.

McKean Vlasov measures

More precisely, $\rho \in \mathcal{P}_2(\mathcal{C})$ is a MKV measure if ρ solves the following martingale problem.
For every $g \in C_c^2(\mathbb{R}^d)$,

$$g(X_t) - \int_0^t \langle b(X_s, \rho_s), \nabla g(X_s) \rangle + \varepsilon^2 \Delta g(X_s) ds$$

is a martingale, where $(X_t)_t \sim \rho$.

Description of the "limit" II

Recall that the elements of $\mathscr{L}^\infty$ are random measures over the set of continuous functions.

Under boundedness assumption of the kernel:

The elements of $\mathscr{L}^\infty$ are a.s. McKean Vlasov measures.
To understand $\mathscr{L}^\infty$ elements (and relate them to $\mu_\star$), we need to understand MKV measures.

Asymptotics of MKV measures when $t \to \infty$

Let ρ a (deterministic) MKV measure and denote ρ_t its marginal distributions.

1. If ρ is stationary, then $\rho_t = \mu_\star$ for every $t > 0$
2. Under LSI, $\rho_t \to \mu_\star$ uniformly.

Basically, the realizations of $\mathscr{L}^\infty$ are measures ρ s.t. $W_2(\rho_t, \mu_\star) \to 0$. The various conclusions we obtained follow from this observation.

Conclusion

- We provide understanding of $\mathscr{L}^N$, the limit set of noisy SVGD as $k \to \infty$.
- We show a dynamical result (convergence to MKV measures) on the way.
- Quantification of the convergence to $\mathscr{L}^N$?
- Quantification of the convergence of $\mathscr{L}^N$ to $\mathscr{L}^\infty$?
- Role of ε on the convergence speed?

References I

[Ba et al., 2021] Ba, J., Erdogdu, M. A., Ghassemi, M., Sun, S., Suzuki, T., Wu, D., and Zhang, T. (2021).
Understanding the variance collapse of svgd in high dimensions.
In *International Conference on Learning Representations*.

[Liu and Wang, 2016] Liu, Q. and Wang, D. (2016).
Stein variational gradient descent: A general purpose Bayesian inference algorithm.
In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2378–2386.

[Shi and Mackey, 2024] Shi, J. and Mackey, L. (2024).
A finite-particle convergence rate for stein variational gradient descent.
Advances in Neural Information Processing Systems, 36.

Selected publications I

[KSA+20] Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton.
A non-asymptotic analysis for Stein variational gradient descent.
In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[SSR22] Adil Salim, Lukang Sun, and Peter Richtárik.
A convergence theory for SVGD in the population limit under Talagrand's inequality T1.
In *International Conference on Machine Learning (ICML)*, 2022.