Long-time asymptotics of noisy SVGD outside the population limit

Victor Priser[1], Pascal Bianchi[1] and Adil Salim[2]
[1]Télécom Paris —LTCI, France; [2] Microsoft Research, USA

This research was funded by the chair DSAIDIS (Data Science & Artificial Intelligence for Digitalized Industry & Services)

Objective

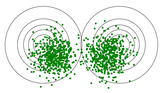
Target density: $\pi \propto \exp(-F)$.
Objective: Generate n *particles* $(X^1, \dots, X^n)$ s.t.

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{X^i} \simeq \pi$$

δ_x: dirac measure in x.

Example with a Gaussian mixture centered in m_1, m_2:

We use $F(x) = \|x - m_1\|^2 + \|x - m_2\|^2$ and generates n particles (in green).



Noisy Stein Variational Gradient Descent (NSVGD) Algorithm

NSVGD algorithm

n *interacting particles* $X_k^1, \dots, X_k^n \in \mathbb{R}^d$. For every $i \le n$ and $k \in \mathbb{N}$:

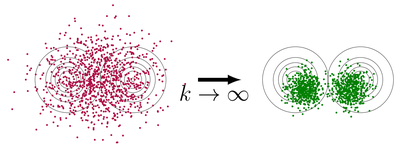
$$X_{k+1}^{i,n} = X_k^{i,n} \overbrace{-\lambda\gamma_k \nabla F(X_k^i) + \sqrt{2\lambda\gamma_k}\xi_k^i}^{\text{Langevin regularisation}} \underbrace{-\frac{\gamma_k}{n}\sum_{j\in[n]}\Big(K(X_k^i, X_k^j)\nabla F(X_k^j) - \nabla_2 K(X_k^i, X_k^j)\Big)}_{\text{True SVGD}}.$$

- $(\xi_k^i) \sim_{i.i.d.} \mathcal{N}(0, I_d)$
- K: *kernel*, for instance, $K(x,y) = \exp(-\|x-y\|^2)$
- $\gamma_k \to 0$: *step size*
- $\lambda \ge 0$: *mixture weight*

Stein Variational Gradient Descent (SVGD, Liu (2016)) = NSVGD with $\lambda = 0$.

Example with a Gaussian mixture centered in m_1, m_2:

Initialize the particles $(X_0^1, \dots, X_0^n)$ from a Gaussian distribution (in *red*).



Algorithm convergence

Empirical measure of NSVGD at time k

$$\mu_k^n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_k^i}$$

We want $\pi \simeq \mu_k^n$.

- μ_k^n: n *atoms*
- π: *density*

$\Big\}\ n \to \infty$.

Gradient algorithm: $k \to \infty$.

Question

At what *rate* must $(k, n) \to (\infty, \infty)$ so that $\mu_k^n \to \pi$?

What does the user do?

- They *fix* n and let $k \to \infty$.
- They *increase* n if the result is not good enough.

Regime of interest: $(k, n) \to (\infty, \infty)$ with $n \ll k$.

Literature review

For SVGD (NSVGD with $\lambda = 0$):

	Low value of k	**Large value of k**
$n = \infty$	$d(\mu_k^\infty, \pi) \le \frac{C}{k}$ Korba (2020)	$d(\mu_k^\infty, \pi) \to 0$ Salim (2022)
$n < \infty$	$d(\mu_k^n, \pi) \le \frac{C}{k}$ with $k \le \phi(n)$	*???* *Regime of interest*

Shi (2024): $\phi(n) = \log\log(n)$.
Carrillo (2023), Banerjee (2024): ϕ polynomial in n.
d: metric on measures (Kernel Stein Discrepancy, Wasserstein distance W_2)

Main results

$\mathscr{L}^n := \{$*Accumulation points of* $(\mu_k^n)_k\}$. That is, for every $\mu \in \mathscr{L}^n$, there exists a sequence $(k_i) \to \infty$ such that $W_2(\mu_{k_i}^n, \mu) \to 0$ as $i \to \infty$.

Theorem 1

F strongly convex and $\lambda > 0$:

$$\mathscr{L}^n \underset{n\to\infty}{\longrightarrow} \pi\,.$$

In other words:

$$\lim_{n\to\infty} \mathbb{P}\Big(\sup_{\mu^n \in \mathscr{L}^n} W_2(\mu^n, \pi) \ge \varepsilon\Big) = 0\,.$$

Averaged empirical measure:

$$\bar{\mu}_k^n := \frac{1}{k}\sum_{i=1}^{k} \mu_i^n\,.$$

$\overline{\mathscr{L}^n} = \{$Accumulation points of $(\bar{\mu}_k^n)_k\}$.

Theorem 2

Assume that $\lambda > 0$:

$$\overline{\mathscr{L}^n} \underset{n\to\infty}{\longrightarrow} \pi\,.$$

NSVGD converges in the regime $k \gg n$.

The regularization

NSVGD: Mixture of the Langevin and SVGD algorithm:

- $\lambda \ll 1$: NSVGD $\simeq$ SVGD.
- $\lambda \gg 1$: NSVGD $\simeq$ Langevin algorithm.

SVGD's Spurious measure

x_* s.t. $\nabla F(x_*) = 0$:

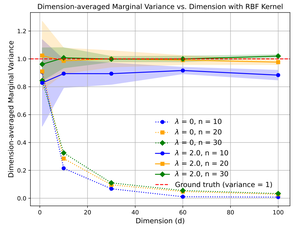
$$\mu_k^n = \delta_{x_*} \implies \forall p \quad \mu_{k+p}^n = \mu_k^n\,.$$

With $\lambda > 0$, such spurious stationary distributions *do not exist*.

Particle Collapse

In *high dimensions* $d > n$:

- The empirical measure μ_k^n *converges to a Dirac measure for SVGD*.
- There is *no particle collapse for NSVGD* when $\lambda > 0$.



Sketch of proof

X_t^i: *linear interpolation* of the discrete-time process X_k^i.

We study the *occupation measure* of the particles:

$$m_t^n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_{l+\cdot}^i} \in \mathcal{P}(C([0,\infty[, \mathbb{R})).$$

Stability condition $\implies$ the distributions of $(m^n)_{t,n}$ are *tight* in $\mathcal{P}(\mathcal{P}(C([0,\infty), \mathbb{R})))$

Tightness

From every sequence $(\tilde{t}_n, \tilde{\varphi}_n) \to (\infty, \infty)$, there exists a subsequence $(t_n, \varphi_n) \to (\infty, \infty)$ and a measure $M \in \mathcal{P}(\mathcal{P}(C([0,\infty[, \mathbb{R}]))$ such that:

$$m_{t_n}^{\varphi_n} \xrightarrow{\text{distribution}} M\,.$$

To characterize the limiting measures, we introduce the set V_2 of McKean-Vlasov distributions.

McKean-Vlasov Distribution

V_2 is the set of probability measures $\rho \in \mathcal{P}(C([0,\infty), \mathbb{R}))$ that are the *pathwise laws* of strong solutions X_t to the *Stochastic Differential Equation* (SDE):

$$\mathrm{d}X_t = \overbrace{-\lambda\nabla F(X_t)\,\mathrm{d}t + \sqrt{2\lambda}\,\mathrm{d}B_t}^{\text{Langevin regularisation}} \underbrace{-\int [K(X_t, y)\nabla F(y) - \nabla_2 K(X_t, y)]\,\mathrm{d}\rho_t(y)}_{\text{True SVGD}},$$

where ρ_t is the time-t marginal of ρ (i.e., the law of X_t), and $(B_t)_{t\ge0}$ is a Brownian motion.

We characterize the limiting measure M.

Propagation of chaos

$$M(\mathsf{V}_2) = 1\,.$$

In other words, the accumulation points of $(m_t^n)_{t,n}$ exists in every regime $(t, n) \to (\infty, \infty)$ and are McKean-Vlasov distributions.

The lemma that states that the *McKean–Vlasov distributions converge to* π

Descent Lemma

For $\rho \in \mathsf{V}_2$, if ρ_t has a density for every t:

$$\mathrm{KL}(\rho_t||\pi) - \mathrm{KL}(\rho_0||\pi) = -\int_0^t (\mathcal{I}_{\mathsf{stein}}(\rho_s||\pi) + \lambda\mathcal{I}(\rho_s||\pi))\,\mathrm{d}s\,,$$

where

$$\mathcal{I}(\mu||\pi) := \int \left\|\nabla \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)\right\|^2 \mathrm{d}\mu\,,$$

and

$$\mathcal{I}_{\mathsf{stein}}(\mu||\pi) := \int \left\|P_\mu \nabla \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)\right\|_{\mathcal{H}}^2 \mathrm{d}\mu\,,$$

where $P_\mu f := \int K(\cdot, y) f(y)\,\mathrm{d}\mu(y)$.
$\|\cdot\|_{\mathcal{H}}$: norm in the RKHS generated by K.

$\lambda > 0 \implies$ *ρ_t has a density for every $t > 0$.*
The *ergodic occupation measure* is defined as

$$M_t^n = \frac{1}{t}\int_0^t \delta_{m_s^n}\,\mathrm{d}s \in \mathcal{P}(\mathcal{P}(C([0,\infty), \mathbb{R})))\,.$$

For a bounded continuous function f and $\tau > 0$,

$$\int f(\rho_\tau)\,\mathrm{d}M_t^n(\rho) \simeq \int f(\rho_0)\,\mathrm{d}M_t^n(\rho)\,,$$

for large t. This means that the limiting measures of the ergodic occupation measures are *time-shift recurrent*.
The Descent Lemma implies:

$$0 = \int \mathrm{KL}(\rho_t||\pi) - \mathrm{KL}(\rho_0||\pi)\,\mathrm{d}M(\rho) = \int\int_0^t (\mathcal{I}_{\mathsf{stein}}(\rho_s||\pi) + \lambda\mathcal{I}(\rho_s||\pi))\,\mathrm{d}s\,\mathrm{d}M(\rho)\,,$$

almost surely for any accumulation point M of $(M_t^n)_{t,n}$.
Since,

$$\mathcal{I}_{\mathsf{stein}}(\mu||\pi) \ge 0 \quad \text{and} \quad \mathcal{I}(\mu||\pi) \ge 0,$$

with equality if and only if $\mu = \pi$, the measure M is supported on a unique measure ρ such that $\rho_t = \pi$ for every $t > 0$.
In other words, the *accumulation points* of $(\bar{\mu}_k^n)_{k,n}$ are, almost surely, *equal to* π.