Long run convergence of discrete-time interacting particles system

Victor Priser[1], Pascal Bianchi[1] and Walid Hachem[2]

[1]Télécom Paris — LTCI; [2] LIGM, CNRS, Univ. Gustave Eiffel

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

Objective

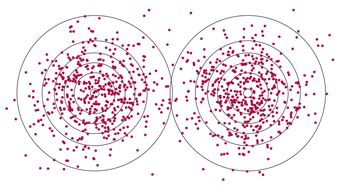
Target density: $\pi \propto \exp(-F)$.
Objective: Generate n *particles* $(X^1, \dots, X^n)$ s.t.

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

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 *red*).



Noisy Stein Variational Gradient Descent (NSVGD) Algorithm

NSVGD algorithm

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

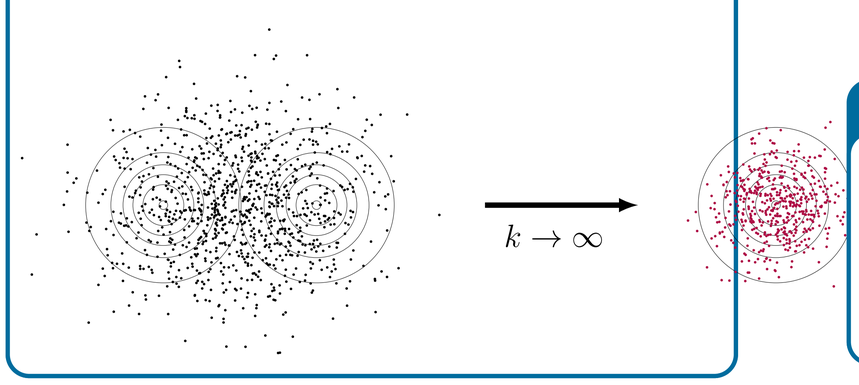
$$X_{k+1}^{i,n} = X_k^{i,n} \overbrace{-\lambda\gamma_k \nabla F(X_k^i) + \sqrt{2\lambda\gamma_k}\xi_k^i}^{\text{Langevin regularisation}} \underbrace{-\frac{\gamma_k}{n}\sum_{j\in[n]}\Big(K(X_k^i, X_k^j)\nabla F(X_k^j) - \nabla_2 K(X_k^i, X_k^j)\Big)}_{\text{True SVGD}}$$

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

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

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 *red*).



Particles system

n *interacting particles* $X_1(k), \dots, X_n(k) \in \mathbb{R}^d$ in *discrete-time*:

$$X_i(k+1) = X_i(k) + \frac{\gamma_k}{n}\sum_{i=1}^{n} b(X_i(k), X_j(k)) + \sqrt{\gamma_k}\xi_i(k+1)$$

- $\xi_i(k) \sim \mathcal{N}(0, \sigma I_d)$: i.i.d. sequence
- γ_k: *step size*

Empirical measure

$$\mu_k^n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(k)}$$

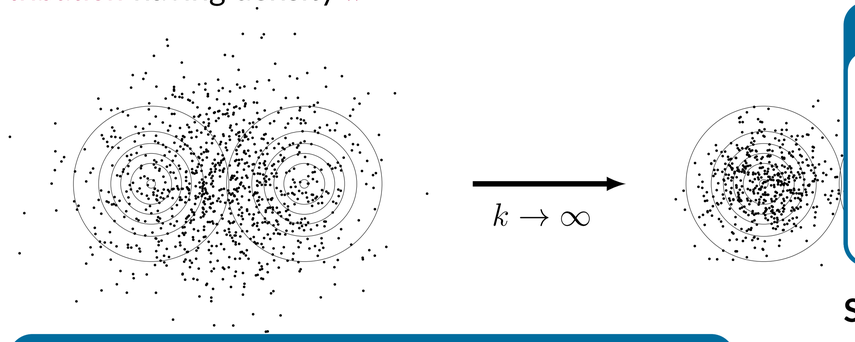
δ_x: dirac measure in x.

Objective

Find convergence of the *discrete-time* empirical measure μ_t^n to *McKean-Vlasov stationary measures*

Application 1: Monte Carlo methods

Aim: generate n particles according to a *target distribution* having density π



SVGD: Stein Variational Gradient Descent (Liu (2016))

- $b(x,y) = K(x,y)\nabla \log(\pi)(y) + \nabla_2 K(x,y)$
- $\sigma = 0$

K: *kernel*, for instance, $K(x,y) = \exp(-\|x-y\|^2)$

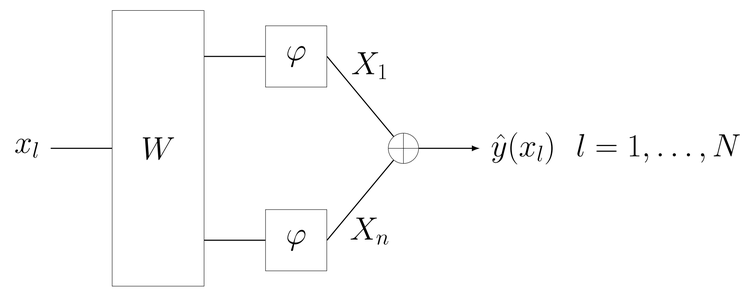
Result (Salim (2022)): $\mu_k^\infty \xrightarrow[k\to\infty]{} \pi$

Problem: Assuming $n = \infty$ is not practical

Question: Do we have convergence in the double regime $(k,n) \to (\infty,\infty)$?

$$\mu_k^n \xrightarrow[(k,n)\to(\infty,\infty)]{} \pi$$

Application 2: Neural Networks



Objective: find the minimizers of the *risk*: $\mathcal{R}(X_1, \dots, X_n) := \sum_{l=1}^{N}(\hat{y}(x_l) - y_l)^2$

(Stochastic) Gradient Descent: $X_i(k+1) = X_i(k) + \gamma_k \nabla_{X_i}\mathcal{R}(X_1(k), \dots, X_n(k))$

Interpret the risk as a *function on* $\mathcal{P}(\mathbb{R}^d)$ (probability measures on $\mathbb{R}^d$)

$$\mathcal{R}(\mu_k^n) = \mathcal{R}(X_1(k), \dots, X_n(k))$$

Idea (Chizat, Rotskoff, Mei (2018))

Continuous time: μ_t^n

μ_t^n shadows *gradient flow* of $\mathcal{R}$ on $\mathcal{P}(\mathbb{R}^d)$: $\partial_t \mu_t = -\nabla\mathcal{R}(\mu_t)$

Result: $\mu_t^n \xrightarrow[(t,n)\to(\infty,\infty)]{} \arg\min \mathcal{R}(\mu)$

Question: What about the convergence of the true Stochastic Gradient Descent:

$$\mu_k^n \xrightarrow[(k,n)\to(\infty,\infty)]{} \arg\min \mathcal{R}(\mu)$$

Result

L: *limit set* of McKean-Vlasov measures

$$L = \{\lim_{n\to\infty} \mathbb{P}_{x_{t_n}}, \quad (t_n) \to \infty\}$$

Theorem

Assumptions: *stability conditions*, $\gamma_k \to 0$ and $\sum_{k=1}^{\infty}\gamma_k = \infty$

$$\mu_k^n \xrightarrow[(k,n)\to(\infty,\infty)]{\mathbb{P}} L$$

SVGD: $L = \{\pi\}$

Neural networks: $L = \{\arg\min \mathcal{R}(\mu)\}$

Sketch of proof

$X_i(t)$: *linear interpolation* of the discrete-time process $X_i(k)$

Shifted occupation measure of the particles:

$$(\Theta_t)_{\#}m^n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(t+\cdot)} \in \mathcal{P}(C([0,\infty[,\mathbb{R}])).$$

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

Tightness

There exists a subequence $(t_n, \varphi_n) \to (\infty,\infty)$ and a measure $M \in \mathcal{P}(\mathcal{P}(C([0,\infty[,\mathbb{R}]))$ such that:

$$(\Theta_{t_n})_{\#}m^{\varphi_n} \xrightarrow{\text{distribution}} M\,.$$

To characterize the limiting measures as *McKean-Vlasov distributions*, we use the *martingale problem*.

Limiting measures

Assumption: $(\Theta_{t_n})_{\#}m^{\varphi_n} \xrightarrow{\text{distribution}} M$
For all $m \in \mathcal{P}(C([0,\infty[,\mathbb{R}])$ M-a.e., m is a *McKean-Vlasov Distribution*:

$$\int\left(\phi(x_t) - \phi(x_s) - \int_s^t L((\pi_u)_{\#}m)(\phi)(x_u)\right)\prod_{j=1}^{r} h_j(x_{t_j})dm(x)$$

for all functions $\phi, h_1, \dots, h_r$ bounded and $t_1, \dots, t_r < s < t$.

$(\pi_u)_{\#}m$: pushforward of the measure m by the coordinate function π_u

$L(\mu)(\phi)(x) = \int\langle b(x,y), \nabla\phi\rangle\, d\mu(y) + \sigma^2\Delta\phi(x)$ *(Infinitesimal generator)*