Long run convergence of discrete-time interacting particle systems of the McKean-Vlasov type

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

[1]LTCI, Télécom Paris, IP Paris, France

[2]CNRS, Laboratoire d'informatique Gaspard Monge (LIGM / UMR 8049), Université Gustave Eiffel, ESIEE Paris, France

https://arxiv.org/pdf/2403.17472.pdf

Discrete-Time Particle System

particles $(X_k^{1,n}, \ldots, X_k^{n,n})$: n sequences of random variables in $\mathbb{R}^d$,

$$X_{k+1}^{i,n} = X_k^{i,n} + \frac{\gamma_{k+1}}{n} \sum_{j=1}^{n} b(X_k^{i,n}, X_k^{j,n}) + \gamma_{k+1}\zeta_{k+1}^{i,n} + \sqrt{2\gamma_{k+1}}\xi_{k+1}^{i,n},$$

for all $k \geq 0$ and all $i = 1, \ldots, n$.

- $\gamma_k \to 0$ and $\sum \gamma_k = \infty$: step size
- b: continuous with linear growth i.e. $b(x, y) \leq C(1 + \|x\| + \|y\|)$.
- $(X_k^{i,n})_{i \leq n}$: exchangeable n–uple for every $k \in \mathbb{N}$.

Empirical measure

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

Application(1): Granular Media

Useful for modeling granular material (sand, snow, ...)

- U: models the collision of the material
- V: models the friction

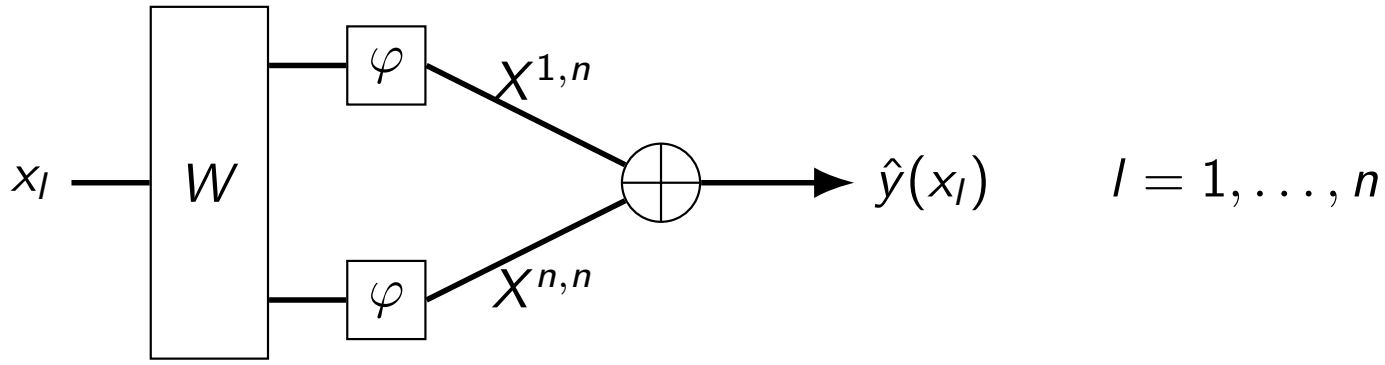
Granular media

$$X_{k+1}^{i,n} = X_k^{i,n} - \frac{\gamma_{k+1}}{n} \sum_{j=1}^{n} (\nabla U(X_k^{i,n} - X_k^{i,n}) + \nabla V(X_k^{i,n})) + \sqrt{2\gamma_{k+1}}\xi_{k+1}^{i,n}$$

Particle system with

$$b(x,y) = -\nabla U(x-y) - \nabla V(x) \text{ and } \zeta_k^{i,n} = 0\,.$$

Application (2): Neural Network



Risk: $R(X^{1,n}, \ldots, X^{n,n}) = \sum_{l=1}^{n} (\hat{y}(x_l) - \hat{y}_l)^2$

Stochastic Gradient Descent: Particle system $(X_k^{i,n})_{i \leq n, k \in \mathbb{N}}$

Risk as function in $\mathcal{P}(\mathbb{R}^d)$: $\bar{R}(\mu_k^n) := R(X_k^{1,n}, \ldots, X_k^{n,n})$

Question

$$\mu_k^n \xrightarrow[(k,n)\to(\infty,\infty)]{} \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\min}\, \bar{R}(\mu)?$$

$\lim_{k\to\infty} \lim_{n\to\infty} \mu_k^n = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\min}\, \bar{R}(\mu)$ is not pratical.

Application(3): Stein Variational Gradient Descent (SVGD)

Target density: π
K: kernel (ex: gaussian kernel)

Stein Variational Gradient Descent algorithm with step size γ_k

$$X_{k+1}^{i,n} = X_{k+1} + \frac{\gamma_{k+1}}{n} \sum_{j=1}^{n} \Big(K(X_k^{i,n}, X_k^{j,n}) \nabla \log \pi(X_k^{j,n}) + \nabla_2 K(X_k^{i,n}, X_k^{j,n}) \Big)$$

Particle system with:

$$b(x,y) = K(x,y) \nabla \log \pi(y) + \nabla_2 K(x,y) \text{ and } \zeta_k^{i,n} = 0 \text{ and } \sigma = 0$$

Question

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

Main result

Definition: W_2 is the Wasserstein distance 2.
Assumption:

- Uniform integrability: $\lim_{a\to\infty} \sup_{k,n} \mathbb{E}\left[(X_k^{1,n})^2 \mathbb{1}_{\|X_k^{1,n}\|\geq a}\right] = 0.$

Main theorem

$$\frac{\sum_{l=1}^{k} \gamma_l W_2(\mu_l^n, \mathcal{S})}{\sum_{l=1}^{k} \gamma_l} \xrightarrow[(k,n)\to(\infty,\infty)]{\mathbb{P}} 0$$

$\mathcal{S}$ is a limit set defined hereafter.

- **Neural Networks:** $\mathcal{S} = \left\{ \arg\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \bar{R}(\mu) \right\}.$
- noisy **SVGD:** $\mathcal{S} = \{\pi\}$.

Review of the litterature

Strength of our result:

- General b continuous with linear growth.
- Discrete-time.
- Double limit.
- Slowly decreasing step size γ_k.

Literature on the general case:

- Double limit unknown without strong convexity assumption on b. One has to take $n \to \infty$ and after $t \to \infty$.
- Discrete-time rarely treated (when treated, it considers a bounded b).

Literature on applications:

- **Neural Networks:** (Chizat, 2018)$\to$ double limit in continuous time; (Montanari, 2018) $\to$ convergence of SGD but not in the double limit.
- **SVGD:** (Salim, 2021) $\to$ convergence when $t \to \infty$ and $n = \infty$.

Sketch of proof

Notation:

- $\mathcal{C}$: space of continuous function in $\mathbb{R}_+$.
- $\mathcal{P}_2(\mathcal{C}), \mathcal{P}_2(\mathbb{R}^d)$: probability measures on $\mathcal{C}$,$\mathbb{R}^d$ (equipped with the Wasserstein distance).
- Interpolated process of $(X_k^{i,n})_{k\in\mathbb{N}}$: $(\bar{X}_t^{i,n})_{t\in\mathbb{R}_+}$.
- Occupation measure: $m_t^n := \frac{1}{n}\sum_{i=1}^n \delta_{\bar{X}_{t+\cdot}^{i,n}} \in \mathcal{P}(\mathcal{C})$.

Steps:

1. Introduction on $V_2 \subset \mathcal{P}_2(\mathcal{C})$: set of McKean-Vlasov distributions.
2. Introduction of recurrent measures in V_2.
3. Convergence of m_t^n to recurrent measures.

Propagation of Chaos

Notation: $b(x,\mu) := \int b(x,y)d\mu(y)$.
B_t, $B_t^{i,n}$: independant Brownian motion in $\mathbb{R}^d$.

$$\mathrm{d}X_t^{i,n} = b(X_t^{i,n}, \frac{1}{n}\sum_{i=1}^{n}\delta_{X_t^{i,n}})\mathrm{d}t + \mathrm{d}B_t^{i,n}\,.$$

Nonlinear McKean-Vlasov SDE:

$$\mathrm{d}X_t = b(X_t, \mathscr{L}(X_t))\mathrm{d}t + \mathrm{d}B_t$$

Propagation of chaos

$$\sup_{t\in[0,T]} W_2(\mathscr{L}(X_t^{1,n},\dots,X_t^{k,n}), \mathscr{L}(X_t)^{\otimes k}) \leq C\frac{k}{n}\mathrm{e}^{CT}$$

Meaning: "Particles $X_k^{i,n}$ become independents and follow a McKean-Vlasov distribution as $n \to \infty$"

McKean-Vlasov distribution

$(X_t)_{t\in\mathbb{R}_+}$: the canonical process; $(\mathcal{F}_t^X)_{t\in\mathbb{R}_+}$: the natural filtration

McKean-Vlasov distributions (V_2)

$\rho \in \mathcal{P}_2(\mathcal{C})$ belongs to the class V_2 if, for every $g \in C_c^2(\mathbb{R}^d, \mathbb{R})$,

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

is a $(\mathcal{F}_t^X)_{t\geq 0}$-martingale on the probability space $(\mathcal{C}, \mathcal{B}(\mathcal{C}), \rho)$.

With $\rho \in \mathcal{P}(\mathcal{C})$, $\rho_t := (\pi_t)_{\#}\rho \in \mathcal{P}(\mathbb{R}^d)$.

- Existence: For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\exists \rho \in V_2$, $s.t.\ \rho_0 = \mu$
- Uniqueness: For $\rho, \rho' \in V_2$, $\rho_0 = \rho_0' \implies \rho = \rho'$

Existence and uniqueness hold if b is Lipschitz for instance.

Birkhoff center

Semi-flow

$\Phi : (\rho, t) \in V_2 \times \mathbb{R}_+ \mapsto (x \in \mathcal{C} \mapsto x(t + \cdot))_{\#}\rho$

Birkhoff Center

BC: "Points ρ that appear an infinite number of times in the trajectory of the flow Φ starting at ρ"

BC is the subset of a Lyapunov function's equilibria points.

Sketch of proof: Convergence to McKean-Vlasov processes

Step 1: Tightness

There exists (t_n, φ_n) such that $m_{t_n}^{\varphi_n} \to M \in \mathcal{P}(\mathcal{P}_2(\mathcal{C}))$ in distribution.

$\mathcal{M}$: the set accumulation points M.

Step 2: Convergence

$M(\mathrm{V}_2) = 1$ for every $M \in \mathcal{M}$.

Meaning:

$$W_2(m_t^n, \mathrm{V}_2) \xrightarrow[(t,n)\to(\infty,\infty)]{\mathbb{P}} 0\,.$$

Sketch of proof: Ergodic convergence

$M_t^n := \frac{1}{t}\int_0^t \delta_{m_t^n} ds \in \mathcal{P}(\mathcal{P}_2(\mathcal{C}))$
$\mathscr{M}$: accumulation points of $(M_n^t)_{t,n}$ (random measure in $\mathcal{P}(\mathcal{P}_2(\mathcal{C}))$)

Step 3: Invariance by translation

Let $\Upsilon \in \mathscr{M}$,

- $\Upsilon(V_2) = 1$ a.s.
- $\forall t > 0$, $(\Phi_t)_\# \Upsilon = \Upsilon$ a.s.

Poincaré recurrence theorem: $\Upsilon(BC_\Phi) = 1$, a.s.
Meaning:

$$\frac{1}{t}\int_0^t W_2(m_t^n, BC) ds \xrightarrow[(t,n)\to(\infty,\infty)]{\mathbb{P}} 0\,.$$

Marginalizing: $\mathcal{S} = (\pi_0)_\# BC$

Pointwise convergence: When do we have the convergence of the non-averaged empirical measure?

$\Psi : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}_+ \to \mathbb{R}$: semi-flow associated to the McKean-Vlasov equation.

Global Attractor $A \subset \mathcal{P}_2(\mathbb{R}^d)$

- Compact subspace of $\mathcal{P}_2(\mathbb{R}^d)$.
- $\forall \mu \in \mathcal{P}_2(\mathbb{R}^d)$, $W_2(\Psi_t(\mu), A) \to 0$.
- $\forall t > 0$; $\Psi_t(A) = A$.
- There exists a neighborhood $A \subset \mathcal{W} \subset \mathcal{P}_2(\mathbb{R}^d)$ such that $\sup_{\mu \in \mathcal{W}} W_2(\Psi_t(\mu), A) \to 0$.

Assumptions:

- Existence of a semi-flow and a global attractor A.
- Assumptions of the main theorem.

Pointwise convergence

$$W_2\left(\mu_k^n, A\right) \xrightarrow[(k,n)\to(\infty,\infty)]{\mathbb{P}} 0$$

Pointwise convergence vs Ergodic convergence

- Better convergence with the pointwise result.
- Pointwise result requires additional assumptions.
- The existence of a global attractor A is not trivial and $(\pi_0)_{\#}BC \subset A$.
- Gradient flow structure on a strictly convex Lyapunov function $\rightarrow A$ exists and $A = (\pi_0)_{\#}BC$.

Application

- Granular Media
  - **Lyapunov function**: Helmholtz energy with set of equilibria $\mathcal{E}$.
  - $(\pi_0)_{\#}BC \subset \mathcal{E} \subset A$.
  - Equality if the Lyapunov function is strictly convex.
- Neural Network
  - **Lyapunov function**: Risk $\bar{R}$.
  - $(\pi_0)_{\#}BC = A = \{\underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\min}\, \bar{R}(\mu)\}$
- SVGD
  - **Lyapunov function**: Kullback-Leibler divergence.
  - $(\pi_0)_{\#}BC = A = \{\pi\}$

Conclusion

Contribution:

- Novel way to obtain a double limit with weaker assumptions.
- We propose a discrete-time result, which is the type of result useful in practice.

Extension/application of this result:

- Convergence speed (what is the convergence speed when $t = \log(n)$ or $t = \exp(n)$?).
- Other classes of particle systems (transformers, ...).
- Applying this result to other models.
- Change the dynamics of the particle system by projecting onto a compact set $\rightarrow$ avoid stability problems.