Fluctuation Analysis of Mean-Field Limits in Overparameterized SGD

Vitalii Konarovskyi

University of Hamburg

Probabilistic Perspectives in Neural Network-Based Machine Learning
Oberwolfach – 2025

joint work with Benjamin Gess and Rishabh Gvalani

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

National Academy of Sciences of Ukraine
INSTITUTE OF MATHEMATICS

Supervised learning and SGD dynamics

Give some data $\{(\xi_i, \gamma_i),\ i \in I\}$, the goal is to predict a new label γ given a new input ξ.

For simplicity assume that $\xi_i \sim P$ and $\gamma_i = f(\xi_i)$.

Supervised learning and SGD dynamics

Give some data $\{(\xi_i, \gamma_i),\ i \in I\}$, the goal is to predict a new label γ given a new input ξ.

For simplicity assume that $\xi_i \sim P$ and $\gamma_i = f(\xi_i)$.

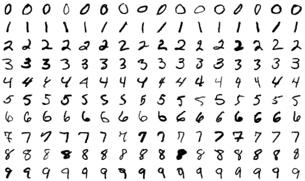
The output $\gamma = f(\xi)$ is approximated by $f_z(\xi)$, where $z \in \mathbb{R}^d$ are parameters that have to be learned from

$$R(z) := \mathbb{E}_P\left[l\left(f(\xi), f_z(\xi)\right)\right] = \mathbb{E}_P\left[\tilde{R}(z, \xi)\right] \to \min$$

Supervised learning and SGD dynamics

Give some data $\{(\xi_i, \gamma_i),\ i \in I\}$, the goal is to predict a new label γ given a new input ξ.

For simplicity assume that $\xi_i \sim P$ and $\gamma_i = f(\xi_i)$.



The output $\gamma = f(\xi)$ is approximated by $f_z(\xi)$, where $z \in \mathbb{R}^d$ are parameters that have to be learned from

$$R(z) := \mathbb{E}_P \left[l\left(f(\xi), f_z(\xi) \right) \right] = \mathbb{E}_P \left[\tilde{R}(z, \xi) \right] \to \min$$

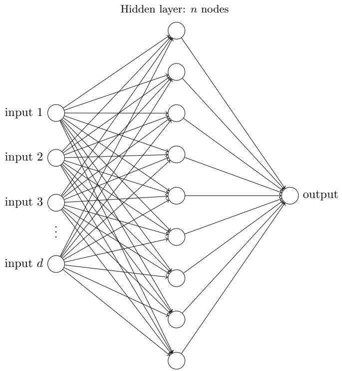
Stochastic Gradient Descent:

taking $z(0) \in \mathbb{R}^d$ define

$$z(t_{i+1}) = z(t_i) - \alpha \nabla_z \tilde{R}(z(t_i), \xi_i)$$

for learning rate α, $t_i = \alpha i$ and $\xi_i \sim P$ – i.i.d.

Neural network with one hidden layer



Network with a single hidden layer:

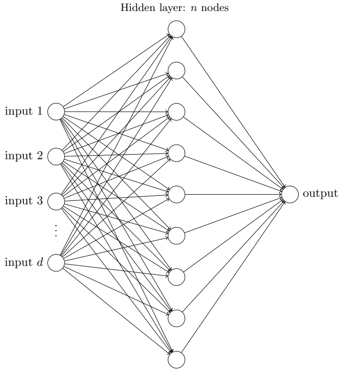
$$f_n(\mathbf{z},\xi) = \frac{1}{n}\sum_{k=1}^{n}\Phi(z_k,\xi)$$
$$= \int \Phi(z,\xi)\nu^n(dz) = \langle \Phi(\cdot,\xi),\nu^n\rangle\,,$$

where $z_k \in \mathbb{R}^d$, $k = 1,\dots,n$, are parameters which have to be found and

$$\nu^n = \frac{1}{n}\sum_{k=1}^{n}\delta_{z_k}$$

[Chizat, Bach, Mei, Nguye, Rotskoff, Sirignano, Vanden-Eijnden...]

Neural network with one hidden layer



Network with a single hidden layer:

$$f_n(\mathbf{z},\xi) = \frac{1}{n}\sum_{k=1}^{n}\Phi(z_k,\xi)$$
$$= \int \Phi(z,\xi)\nu^n(dz) = \langle \Phi(\cdot,\xi),\nu^n\rangle\,,$$

where $z_k \in \mathbb{R}^d$, $k = 1,\dots,n$, are parameters which have to be found and

$$\nu^n = \frac{1}{n}\sum_{k=1}^{n}\delta_{z_k}$$

[Chizat, Bach, Mei, Nguye, Rotskoff, Sirignano, Vanden-Eijnden...]

Risk Function: for $\mathbf{z} = (z_1,\dots,z_n)$

$$R(\mathbf{z}) := \frac{1}{2}\mathbb{E}_P\left[(f(\xi) - f_n(\mathbf{z},\xi))\,|^2\right] \to \min$$

SGD for shallow network

The parameters $\mathbf{z} = (z_k)_{k=1,\dots,n}$ can be learned by stochastic gradient descent

$$
\begin{aligned}
z_k(t_{i+1}) &= z_k(t_i) - \alpha \nabla_{z_k} \left(\frac{1}{2} \left| f(\xi_i) - f_n(\mathbf{z}, \xi_i) \right|^2 \right) \\
&= z_k(t_i) - \alpha \left(f_n(\mathbf{z}, \xi_i) - f(\xi_i) \right) \nabla_{z_k} \Phi(z_k(t_i), \xi_i) \\
&= z_k(t_i) + \alpha \left[\nabla F(z_k(t_i), \xi_i) - \langle \nabla_z K(z_k(t_i), \cdot, \xi_i), \nu^n_{t_i} \rangle \right] \\
&= z_k(t_i) + \alpha \tilde{V}(z_k(t_i), \nu^n_{t_i}, \xi_i),
\end{aligned}
$$

where α is a learning rate, $t_i = i\alpha$, $\xi_i \sim P$ – i.i.d., $\nu_t = \frac{1}{n} \sum_{l=1}^n \delta_{z_l(t)}$, and

$$
F(z, \xi) = f(\xi)\Phi(z, \xi) \quad \text{and} \quad K(z, y, \xi) = \Phi(z, \xi)\Phi(y, \xi).
$$

$$
\rightsquigarrow \qquad z_k(t_{i+1}) = z_k(t_i) + \alpha \tilde{V}(z_k(t_i), \nu^n_{t_i}, \xi_i)
$$

SGD and distribution dependent stochastic flow

The dynamics of the SGD

$$\begin{aligned} z_k(t_{i+1}) &= z_k(t_i) + \alpha \tilde{V}(z_k(t_i), \nu^n_{t_i}, \xi_i) \\ &= z(t_i) + \underbrace{\mathbb{E}_\xi \tilde{V}(\dots)}_{=: V(z_k(t_i), \nu^n_{t_i})} \alpha + \sqrt{\alpha} \underbrace{\left(\tilde{V}(z(t_i), \nu^n_{t_i}, \xi_i) - \mathbb{E}_\xi \tilde{V}(\dots)\right)}_{=: G(z(t_i), \nu^n_{t_i}, \xi_i)} \sqrt{\alpha} \end{aligned}$$

can be captured by **Distribution Dependent Stochastic Flow:**

$$dZ_t(u) = V(Z_t(u), \mu_t)dt + \sqrt{\alpha} \underbrace{\int_\Theta G(Z_t(u), \mu_t, \xi) W(d\xi, dt)}_{= \sum_i \langle G, e_i \rangle dw_i},$$

$$Z_0(u) = u, \quad \mu_t = \mu_0 \circ Z_t^{-1},$$

where W is a cylindrical Wiener process on $L_2(\Xi, \mathrm{Law}\xi)$, i.e.

$$W(\xi, t) = \sum_l e_l(\xi) w_l(t).$$

Recall the SGD:

$$z_k(t_{i+1}) = z_k(t_i) + V(z_k(t_i), \nu^n_{t_i})\alpha + \sqrt{\alpha} G(z(t_i), \nu^n_{t_i}, \xi_i)\sqrt{\alpha}$$

and the Distribution Dependent Stochastic Flow:

$$\begin{aligned} dZ_t(u) &= V(Z_t(u), \mu_t)dt \\ &\qquad + \sqrt{\alpha} \int_\Theta G(Z_t(u), \mu_t, \xi) W(d\xi, dt), \\ Z_0(u) &= u, \quad \mu_t = \mu_0 \circ Z_t^{-1}, \end{aligned}$$

Recall the SGD:

$$z_k(t_{i+1}) = z_k(t_i) + V(z_k(t_i), \nu^n_{t_i})\alpha + \sqrt{\alpha} G(z(t_i), \nu^n_{t_i}, \xi_i)\sqrt{\alpha}$$

and the Distribution Dependent Stochastic Flow:

$$\begin{aligned} dZ_t(u) &= V(Z_t(u), \mu_t)dt \\ &\qquad + \sqrt{\alpha}\int_\Theta G(Z_t(u), \mu_t, \xi) W(d\xi, dt), \\ Z_0(u) &= u, \quad \mu_t = \mu_0 \circ Z_t^{-1}, \end{aligned}$$

Theorem [Gess, Kassing, K. '24, JMLR]

Let $\mu_0 \in \mathcal{P}_2$ and $\nabla V(z, \nu, \xi)$ be regular enough in z. Then for every $\Phi \in \mathrm{C}^4_b(\mathcal{P}_2)$

$$\sup_{t_i \le T} |\mathbb{E}\Phi(\nu^n_{t_i}) - \mathbb{E}\Phi(\mu_{t_i})| \le C\alpha + C\sqrt{\mathbb{E}W_2^2(\mu_0, \nu^n_0)}$$

where $W_2^2(\rho_1, \rho_2) = \inf_{\eta_j \sim \rho_j} \mathbb{E}\left[|\eta_1 - \eta_2|^2\right]$

Recall the SGD:

$$z_k(t_{i+1}) = z_k(t_i) + V(z_k(t_i), \nu^n_{t_i})\alpha + \sqrt{\alpha} G(z(t_i), \nu^n_{t_i}, \xi_i)\sqrt{\alpha}$$

and the Distribution Dependent Stochastic Modified Flow:

$$\begin{aligned} dZ_t(u) &= V(Z_t(u), \mu_t)dt - \frac{\alpha}{4}\nabla|V(Z_t(u), \mu_t)|^2 dt - \frac{\alpha}{4}\langle D|V(Z_t(u), \mu_t)|^2, \mu_t\rangle dt \\ &\quad + \sqrt{\alpha}\int_\Theta G(Z_t(u), \mu_t, \xi) W(d\xi, dt), \\ Z_0(u) &= u, \quad \mu_t = \mu_0 \circ Z_t^{-1}, \end{aligned}$$

Theorem [Gess, Kassing, K. '24, JMLR]

Let $\mu_0 \in \mathcal{P}_2$ and $\nabla V(z, \nu, \xi)$ be regular enough in z. Then for every $\Phi \in \mathrm{C}^4_b(\mathcal{P}_2)$

$$\sup_{t_i \le T} |\mathbb{E}\Phi(\nu^n_{t_i}) - \mathbb{E}\Phi(\mu_{t_i})| \le C\alpha^2 + C\sqrt{\mathbb{E}W_2^2(\mu_0, \nu^n_0)}$$

where $W_2^2(\rho_1, \rho_2) = \inf_{\eta_j \sim \rho_j} \mathbb{E}\left[|\eta_1 - \eta_2|^2\right]$

Weak Topology vs. Strong Topology

Observe that for a 1-d Brownian motion w the error between $\sqrt{\alpha}w_t$ and 0 can is different in different topologies

$$W_1\left(\mathrm{Law}\left(\sqrt{\alpha}w_t\right), \delta_0\right) = \mathbb{E}\left[\left|\sqrt{\alpha}w_t - 0\right|\right] = O\left(\sqrt{\alpha}\right)$$

Weak Topology vs. Strong Topology

Observe that for a 1-d Brownian motion w the error between $\sqrt{\alpha} w_t$ and 0 can is different in different topologies

$$W_1\left(\mathrm{Law}\left(\sqrt{\alpha} w_t\right), \delta_0\right) = \mathbb{E}\left[\left|\sqrt{\alpha} w_t - 0\right|\right] = O\left(\sqrt{\alpha}\right)$$

but

$$\mathbb{E}\left[f\left(\sqrt{\alpha} w_t\right)\right] - \mathbb{E}\left[f(0)\right] = \mathbb{E}\left[f'(0)\sqrt{\alpha} w_t + \frac{1}{2} f''(\theta) \alpha w_t^2\right] = O\left(\|f''\| \alpha\right)$$

Main goal

Goal: Compare the SGD and mean-field dynamics in strong (Wasserstein topology).

Main goal

Goal: Compare the SGD and mean-field dynamics in strong (Wasserstein topology).

Note that

$$\nu_t^n - \rho_t = O\left(\sqrt{\alpha} + \frac{1}{\sqrt{n}}\right)$$

[Mei, Montanari, Nguyen '18, PNAS] and

$$\mu_t - \rho_t = \left(\sqrt{\alpha}\right),$$

where

$$\partial_t \rho_t = -\nabla\left(V(\cdot, \rho_t)\rho_t\right), \quad \rho_0 = \mu_0.$$

Idea: We will focus on the case $\alpha = \frac{1}{n}$, (where the SGD and the initial noises are balanced) and compare the fluctuation field for both dynamics:

$$\sqrt{n}\left(\nu^n - \rho\right) \quad \text{and} \quad \sqrt{n}\left(\mu - \rho\right)$$

CLT for the SGD

We set

$$\zeta_t^n := \sqrt{n}\left(\nu_t^n - \rho_t\right),$$

where

$$z_k(t_{i+1}) = z_k(t_i) + \frac{1}{n}\tilde{V}(z_k(t_i), \nu_{t_i}^n, \xi_i) \quad \text{and} \quad \nu_t^n = \frac{1}{n}\sum_{k=1}^{n} \delta_{z_k(t)}$$

Theorem [Sirignano, Spiliopoulos '20, SPA]

Let μ_0 has a compact support and V be regular enough in z. Then

$$\zeta_t^n \to \zeta_t, \quad t \geq 0,$$

in $D\left([0,\infty), H_{-J}\right)$ in distribution and

$$\mathbb{E} \sup_{t\in[0,T]} \|\zeta_t^n\|_{H_{-J}}^2 \leq C,$$

where ζ is a Gaussian process solving $(V(z,\mu) = \nabla F(z) - \langle \nabla K(z,\cdot), \mu\rangle)$

$$d\zeta_t = -\nabla\cdot\left(V(\cdot,\rho_t)\zeta_t + \langle \nabla K(x,\cdot), \zeta_t\rangle \rho_t(dx)\right) dt - \nabla\cdot \int_{\Xi} G(\cdot,\rho_t,\xi)\rho_t W(d\xi, dt).$$

Stochastic mean-field equation

Considering

$$dZ_t(u) = V(Z_t(u), \mu_t)dt + \sqrt{\alpha}\int_\Xi G(Z_t(u), \mu_t, \xi)W(d\xi, dt)$$
$$Z_0(u) = u, \quad \mu_t = \mu_0 \circ Z_t^{-1}$$

and applying Ito's formula, we get the **Stochastic Mean-Field Equation:**

$$\begin{aligned} d\mu_t = -\nabla \cdot (V(\cdot, \mu_t)\mu_t)dt + \frac{\alpha}{2}\nabla^2 : (A(\cdot, \mu_t)\mu_t)dt \\ - \sqrt{\alpha}\nabla \cdot \int_\Xi G(\cdot, \mu_t, \xi)\mu_t \, W(d\xi, dt) \end{aligned}$$

where $A(z, \mu) = \mathbb{E}G(z, \mu, \xi) \otimes G(z, \mu, \xi)$.

Stochastic mean-field equation

Considering

$$dZ_t(u) = V(Z_t(u), \mu_t)dt + \sqrt{\alpha}\int_\Xi G(Z_t(u), \mu_t, \xi)W(d\xi, dt)$$
$$Z_0(u) = u, \quad \mu_t = \mu_0 \circ Z_t^{-1}$$

and applying Ito's formula, we get the **Stochastic Mean-Field Equation:**

$$d\mu_t = -\nabla \cdot (V(\cdot, \mu_t)\mu_t)dt + \frac{\alpha}{2}\nabla^2 : (A(\cdot, \mu_t)\mu_t)dt - \sqrt{\alpha}\nabla \cdot \int_\Xi G(\cdot, \mu_t, \xi)\mu_t \, W(d\xi, dt)$$

where $A(z, \mu) = \mathbb{E}G(z, \mu, \xi) \otimes G(z, \mu, \xi)$.

$\rightsquigarrow$ The martingale problem for this equation is the same as in [Rotskoff, Vanden-Eijnden, '22, CPAM]

Well-posedness of SMFE

Theorem [Gess, Rishabh, K. '25, PTRF]

Let the coefficients V, G be Lipschitz continuous and smooth enough w.r.t. special variable. Then the SMFE

$$d\mu_t = -\nabla \cdot (V(\cdot, \mu_t)\mu_t)dt + \frac{\alpha}{2}\nabla^2 : (A(\cdot, \mu_t)\mu_t)dt \\ - \sqrt{\alpha}\nabla \cdot \int_\Xi G(\cdot, \mu_t, \xi)\mu_t \, W(d\xi, dt)$$

has a unique solution. Moreover, it is a superposition solution, i.e., $\mu_t = \mu_0 \circ Z_t^{-1}$

Well-posedness of SMFE

Theorem [Gess, Rishabh, K. '25, PTRF]

Let the coefficients V, G be Lipschitz continuous and smooth enough w.r.t. special variable. Then the SMFE

$$d\mu_t = -\nabla \cdot (V(\cdot, \mu_t)\mu_t)dt + \frac{\alpha}{2}\nabla^2 : (A(\cdot, \mu_t)\mu_t)dt$$
$$- \sqrt{\alpha}\nabla \cdot \int_{\Xi} G(\cdot, \mu_t, \xi)\mu_t\, W(d\xi, dt)$$

has a unique solution. Moreover, it is a superposition solution, i.e., $\mu_t = \mu_0 \circ Z_t^{-1}$

Idea of Proof: We use the transformation

$$\gamma_t = \mu_t \circ \psi_t^{-1}$$

with

$$\psi_t^k(z) = z^k - \int_0^t \nabla\psi_s^k(z) \cdot M(z, \circ ds) \quad \text{and} \quad M(z,t) = \sqrt{\alpha}\int_0^t G(z, \mu_t, \xi)W(d\xi, ds)$$

to reduce the SMFE to the continuity equation

$$d\gamma_t = -\nabla(b(t,\cdot)\gamma_t)dt, \quad \gamma_0 = \mu_0 \quad \rightsquigarrow \quad \gamma_t = \mu_0 \circ Y_t^{-1},$$

where $dY_t(x) = b(t, Y_t(x))dx$

Quantified CLT for SMFE

Theorem **[Gess, Rishabh, K. '25, PTRF]**

Let μ_t^n be a solution to the

$$d\mu_t = -\nabla \cdot (V(\cdot,\mu_t)\mu_t)dt + \frac{1}{2}\nabla^2 : (A(\cdot,\mu_t)\mu_t)dt - \frac{1}{\sqrt{n}}\nabla \cdot \int_\Xi G(\cdot,\mu_t,\xi)\mu_t\, W(d\xi,dt)$$

started from $\mu_0^n = \nu_0^n = \frac{1}{n}\sum_{k=1}^n \delta_{z_k(0)}$ with $z_k(0) \sim \mu_0$ i.i.d. Then under the assumptions of the previous theorem,

$$\eta_t^n := \sqrt{n}\,(\mu_t^n - \rho_t) \to \zeta_t, \quad t \geq 0,$$

where ζ is a Gaussian process solving

$$d\zeta_t = -\nabla\cdot\Big(V(\cdot,\mu_t^0)\zeta_t + \langle \nabla K(x,\cdot),\zeta_t\rangle \mu_t^0(dx)\Big)\,dt - \nabla\cdot\int_\Xi G(\cdot,\mu_t^0,\xi)\mu_t^0 W(d\xi,dt).$$

Moreover, $\mathbb{E}\sup_{t\in[0,T]}\|\eta_t^n - \zeta_t\|_{H_{-J}}^2 \leq \frac{C}{n}$.

Higher Order Approximation

Recall:

$$z_k(t_{i+1}) = z_k(t_i) + \frac{1}{n} V(z_k(t_i), \nu^n_{t_i}, \xi_i) \quad \text{and} \quad \nu^n_t = \frac{1}{n} \sum_{k=1}^{n} \delta_{z_k(t)},$$

μ^n_t is a solution to the

$$d\mu_t = -\nabla \cdot (V(\cdot, \mu_t)\mu_t)dt + \frac{1}{2n} \nabla^2 : (A(\cdot, \mu_t)\mu_t)dt - \frac{1}{\sqrt{n}} \nabla \cdot \int_{\Xi} G(\cdot, \mu_t, \xi)\mu_t \, W(d\xi, dt)$$

started from $\mu^n_0 = \nu^n_0 = \frac{1}{n} \sum_{k=1}^{n} \delta_{z_k(0)}$ with $z_k(0) \sim \mu_0$ i.i.d and $\alpha = \frac{1}{n}$, where

$$V(z, \mu) = \mathbb{E}\tilde{V}(z, \mu, \xi), \quad \text{and} \quad G(z, \mu, \xi) = \tilde{V}(z, \nu, \xi) - V(z, \nu).$$

Higher Order Approximation

Recall:

$$z_k(t_{i+1}) = z_k(t_i) + \frac{1}{n} V(z_k(t_i), \nu^n_{t_i}, \xi_i) \quad \text{and} \quad \nu^n_t = \frac{1}{n}\sum_{k=1}^n \delta_{z_k(t)},$$

μ^n_t is a solution to the

$$d\mu_t = -\nabla \cdot (V(\cdot, \mu_t)\mu_t)dt + \frac{1}{2n}\nabla^2 : (A(\cdot, \mu_t)\mu_t)dt - \frac{1}{\sqrt{n}}\nabla \cdot \int_\Xi G(\cdot, \mu_t, \xi)\mu_t \, W(d\xi, dt)$$

started from $\mu^n_0 = \nu^n_0 = \frac{1}{n}\sum_{k=1}^n \delta_{z_k(0)}$ with $z_k(0) \sim \mu_0$ i.i.d and $\alpha = \frac{1}{n}$, where

$$V(z, \mu) = \mathbb{E}\tilde{V}(z, \mu, \xi), \quad \text{and} \quad G(z, \mu, \xi) = \tilde{V}(z, \nu, \xi) - V(z, \nu).$$

Theorem [Gess, Gvalani, K. '25, PTRF]

For V quite regular in z one has

$$\sup_{t_i \le T} W_p\left(\mathrm{Law}\mu^n_{t_i}, \mathrm{Law}\nu_{t_i}\right) = o\left(\sqrt{\alpha}\right)$$

for all $p \in [1, 2)$.

Idea of proof

We observe that

$$\mu_t^n = \rho_t + \frac{1}{\sqrt{n}}\zeta + O\left(\frac{1}{n}\right)$$

$$\nu_t^n = \rho_t + \frac{1}{\sqrt{n}}\zeta + o\left(\frac{1}{\sqrt{n}}\right).$$

Therefore,

$$\mu^n - \nu^n = o\left(\frac{1}{\sqrt{n}}\right)$$

Idea of proof

We observe that

$$\mu_t^n = \rho_t + \frac{1}{\sqrt{n}}\zeta + O\left(\frac{1}{n}\right)$$
$$\nu_t^n = \rho_t + \frac{1}{\sqrt{n}}\zeta + o\left(\frac{1}{\sqrt{n}}\right).$$

Therefore,

$$\mu^n - \nu^n = o\left(\frac{1}{\sqrt{n}}\right)$$

More precisely,

$$\sqrt{n^p}\, W_p^p\left(\mathrm{Law}(\mu^n), \mathrm{Law}(\nu^n)\right) = \sqrt{n^p}\inf \mathbb{E}\left[\sup_{t\in[0,T]} \|\mu_t^n - \nu_t^n\|_{H_{-J}}^p\right]$$
$$= \inf \mathbb{E}\left[\sup_{t\in[0,T]} \|\sqrt{n}(\mu_t^n - \rho_t) - \sqrt{n}(\nu_t^n - \rho_t)\|_{H_{-J}}^p\right]$$
$$= W_p^p\left(\mathrm{Law}(\eta^n), \mathrm{Law}(\zeta^n)\right) \to 0.$$

Open problem and optimal rate of error bound

We recall that

$$\zeta_t^n := \sqrt{n}\,(\nu_t^n - \rho_t) \to \zeta_t$$

where

$$z_k(t_{i+1}) = z_k(t_i) + \frac{1}{n}\tilde{V}(z_k(t_i), \nu_{t_i}^n, \xi_i) \quad \text{and} \quad \nu_t^n = \frac{1}{n}\sum_{k=1}^{n} \delta_{z_k(t)}.$$

Open problem and optimal rate of error bound

We recall that

$$\zeta_t^n := \sqrt{n}\,(\nu_t^n - \rho_t) \to \zeta_t$$

where

$$z_k(t_{i+1}) = z_k(t_i) + \frac{1}{n}\tilde{V}(z_k(t_i), \nu_{t_i}^n, \xi_i) \quad \text{and} \quad \nu_t^n = \frac{1}{n}\sum_{k=1}^n \delta_{z_k(t)}.$$

Open problem:

$$W_p\left(\mathrm{Law}(\zeta^n), \mathrm{Law}(\zeta)\right) = \left(\mathbb{E} \sup_{t\in[0,T]} \|\zeta_t^n - \zeta_t\|^p\right)^{\frac{1}{p}} = O\left(\frac{1}{n}\right)?$$

for optimal coupling (ζ^n, ζ).

Open problem and optimal rate of error bound

We recall that

$$\zeta_t^n := \sqrt{n}\,(\nu_t^n - \rho_t) \to \zeta_t$$

where

$$z_k(t_{i+1}) = z_k(t_i) + \frac{1}{n}\tilde{V}(z_k(t_i), \nu_{t_i}^n, \xi_i) \quad \text{and} \quad \nu_t^n = \frac{1}{n}\sum_{k=1}^{n} \delta_{z_k(t)}.$$

Open problem:

$$W_p\left(\mathrm{Law}(\zeta^n), \mathrm{Law}(\zeta)\right) = \left(\mathbb{E} \sup_{t\in[0,T]} \|\zeta_t^n - \zeta_t\|^p\right)^{\frac{1}{p}} = O\left(\frac{1}{n}\right)?$$

for optimal coupling (ζ^n, ζ).

Then

$$\mu_t^n = \rho_t + \frac{1}{\sqrt{n}}\zeta + O\left(\frac{1}{n}\right)$$

$$\nu_t^n = \rho_t + \frac{1}{\sqrt{n}}\zeta + \cancel{o\left(\frac{1}{\sqrt{n}}\right)}\; O\left(\frac{1}{n}\right).$$

Therefore, $\mu^n - \nu^n = \cancel{o\left(\frac{1}{\sqrt{n}}\right)}\; O\left(\frac{1}{n}\right)$.

References

[1] Gess, Kassing, Konarovskyi,
Stochastic Modified Flows, Mean-Field Limits and Dynamics of Stochastic Gradient Descent
Journal of Machine Learning Research 25 (2024)

[2] Gess, Gvalani, Konarovskyi,
Conservative SPDEs as Fluctuating Mean Field Limits of Stochastic Gradient Descent
Probability Theory and Related Fields 192 (2025)

Thank you!