Stochastic Modified Flows, Mean-Field Limits and Dynamics of Stochastic Gradient Descent

Vitalii Konarovskyi

University of Hamburg

Berlin Probability Colloquium
Berlin – 2026

joint work with Benjamin Gess and Rishabh Gvalani and Sebastian Kassing

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

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

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

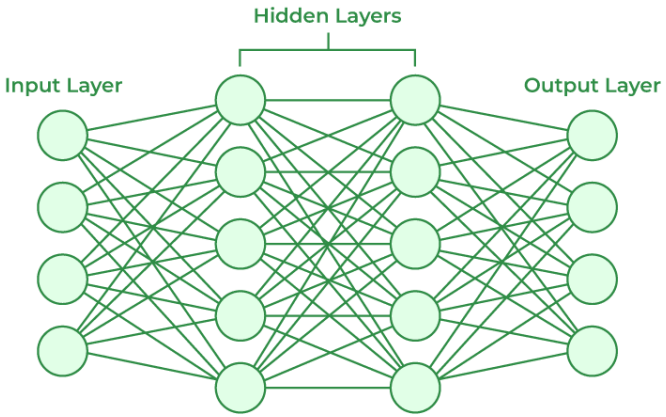
(Stochastic) Gradient Descent:

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

$$z(t_{i+1}) = z(t_i) - \frac{\alpha}{B}\sum_{j=1}^{B} \nabla_z \tilde{R}(z(t_i), \xi_{j,i})$$

for learning rate α, $t_i = \alpha i$, $\xi_{j,i} \sim P$ i.i.d. and mini-batch size B.

Neural Network



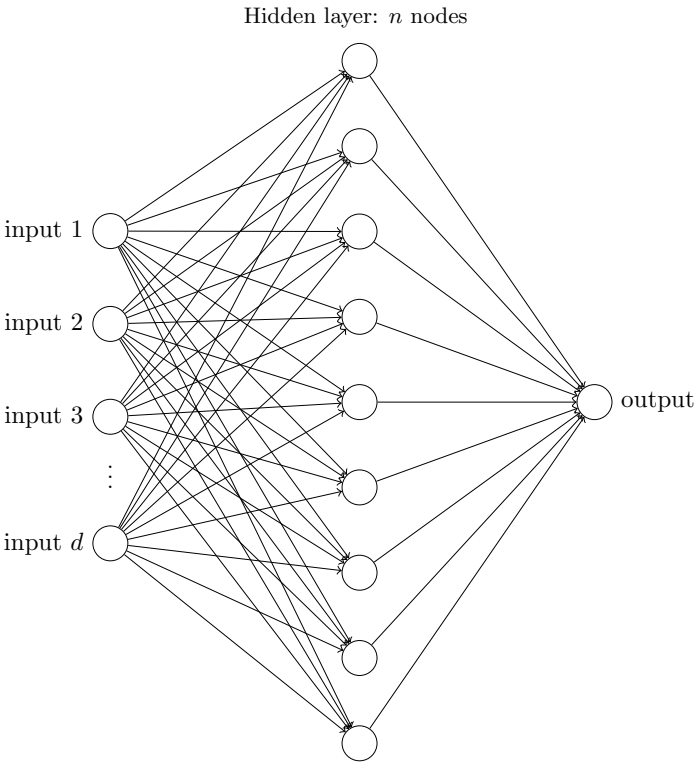
- $L \in \mathbb{N}$ – number of layers
- $d_1, \ldots, d_L$ – dimension of each layer
- σ – activation functions like $\sigma(x) = (1 + e^{-x})$, $\sigma(x) = \max(x, 0)$,...

Motion of "signals" from layer to layer:

$$\xi \mapsto (\sigma((W\xi + b)_i)),$$

output $\gamma = f(\xi)$ is approximated by $f_z(\xi)$ for $z = (W_1, b_1, W_2, b_2, \ldots)$

Neural network with one hidden layer



Network with a single hidden layer:

$$f_n(\mathbf{z}, \xi) = \frac{1}{n} \sum_{k=1}^{n} c_k \sigma \left(a_k \cdot \xi + b_k \right)$$

$$= \frac{1}{n} \sum_{k=1}^{n} \Phi(z_k, \xi)$$

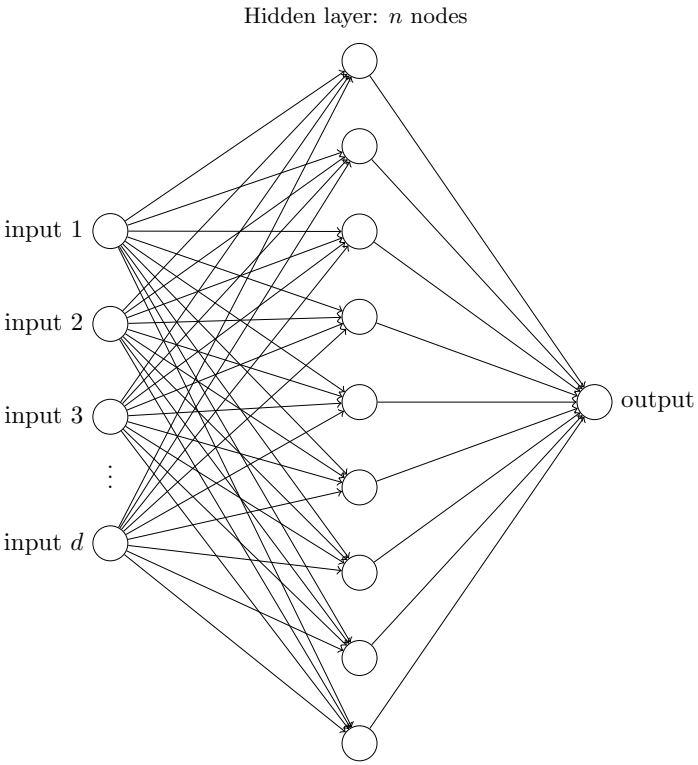
$$= \int \Phi(z, \xi) \nu^n(dz) = \langle \Phi(\cdot, \xi), \nu^n \rangle ,$$

where $z_k \in \mathbb{R}^d$, $k = 1, \ldots, n$, 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:

$$f_n(\mathbf{z}, \xi) = \frac{1}{n} \sum_{k=1}^{n} c_k \sigma \left(a_k \cdot \xi + b_k \right)$$

$$= \frac{1}{n} \sum_{k=1}^{n} \Phi(z_k, \xi)$$

$$= \int \Phi(z, \xi) \nu^n(dz) = \langle \Phi(\cdot, \xi), \nu^n \rangle ,$$

where $z_k \in \mathbb{R}^d$, $k = 1, \ldots, n$, 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 $\mathbf{z} = (z_1, \ldots, z_n)$ $\quad R(\mathbf{z}) := \frac{1}{2} \mathbb{E}_\xi \left[\left(f(\xi) - f_n(\mathbf{z}, \xi) \right)^2 \right] \to \min$

SGD for shallow network

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

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

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

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

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

Continuous Dynamics of Parameters

Take $z_k(0) \sim \mu_0$ – i.i.d., α – learning rate, $t_i = i\alpha$, $\xi_i \sim P$ i.i.d. and

$$z_k(t_{i+1}) = z_k(t_i) + \alpha \tilde{V}(z_k(t_i), \nu^n_{t_i}, \xi_i), \quad k \in \{1, \dots, n\},$$

where $\nu^n_t = \frac{1}{n} \sum_{l=1}^{n} \delta_{z_l(t)}$.

Continuous Dynamics of Parameters

Take $z_k(0) \sim \mu_0$ – i.i.d., α – learning rate, $t_i = i\alpha$, $\xi_i \sim P$ i.i.d. and

$$z_k(t_{i+1}) = z_k(t_i) + \alpha \tilde{V}(z_k(t_i), \nu^n_{t_i}, \xi_i), \quad k \in \{1, \dots, n\},$$

where $\nu^n_t = \frac{1}{n} \sum_{l=1}^n \delta_{z_l(t)}$.

The expression for $z_k(t)$ looks as an Euler scheme for

$$dZ_k(t) = V(Z_k(t), \mu^n_t)dt,$$

$$\mu^n_t = \frac{1}{n} \sum_{k=1}^n \delta_{Z_k(t)},$$

for $V(z, \mu) = \mathbb{E}_\xi \tilde{V}(z, \mu, \xi)$

Convergence to PDE

If $z_k(0) \sim \mu_0$ – i.i.d. and $\alpha = \frac{1}{n}$, then

$$d(\nu_t^n, \mu_t) = O\left(\frac{1}{\sqrt{n}}\right),$$

where μ_t solves

$$d\mu_t = -\nabla \left(V(\cdot, \mu_t)\mu_t\right) dt$$

with initial condition μ_0.

[Mei, Montanari, Nguyen '18]

Convergence to PDE

If $z_k(0) \sim \mu_0$ – i.i.d. and $\alpha = \frac{1}{n}$, then

$$d(\nu_t^n, \mu_t) = O\left(\frac{1}{\sqrt{n}}\right),$$

where μ_t solves

$$d\mu_t = -\nabla \left(V(\cdot, \mu_t)\mu_t\right) dt$$

with initial condition μ_0.

[Mei, Montanari, Nguyen '18]

$\rightsquigarrow$ **The mean behavior of the SGD dynamics can then be analysed by considering μ_t**

Main goal

Problem. After passing to the deterministic gradient flow μ, all of the information about the inherent fluctuations of the stochastic gradient descent dynamics is lost.

Main goal

Problem. After passing to the deterministic gradient flow μ, all of the information about the inherent fluctuations of the stochastic gradient descent dynamics is lost.

Propose an S(P)DE which would capture the fluctuations of the SGD dynamics and also would give its better approximation.

SDE for SGD dynamics (classical approach)

Stochastic gradient descent

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

is the Euler-Maruyama scheme for the SDE

$$dZ_k(t) = V(Z_k(t), \mu^n_t)dt + \sqrt{\alpha}(\Sigma^{\frac{1}{2}})_k(\mathbf{Z}(t))dB(t), \quad k \in \{1, \dots, n\},$$

where $\mu^n_t = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i(t)}$, $\Sigma_{k,l}(\mathbf{z}) = \mathbb{E}_\xi\left[G(z_k, \mu, \xi) \otimes G(z_l, \mu, \xi)\right]$ and

B is n-dim Brownian motion.

SDE for SGD dynamics (classical approach)

Stochastic gradient descent

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

is the Euler-Maruyama scheme for the SDE

$$dZ_k(t) = V(Z_k(t), \mu^n_t)dt + \sqrt{\alpha}(\Sigma^{\frac{1}{2}})_k(\mathbf{Z}(t))dB(t), \quad k \in \{1, \dots, n\},$$

where $\mu^n_t = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i(t)}$, $\Sigma_{k,l}(\mathbf{z}) = \mathbb{E}_\xi\left[G(z_k, \mu, \xi) \otimes G(z_l, \mu, \xi)\right]$ and

B is n-dim Brownian motion.

$\rightsquigarrow$ $\Sigma^{\frac{1}{2}}$ is $dn \times dn$ matrix!

Martingale problem for empirical distribution

$$dZ_k(t) = V(Z_k(t), \mu_t^n)dt + \sqrt{\alpha}(\Sigma^{\frac{1}{2}})_k(\mathbf{Z}(t))dB(t), \quad k \in \{1, \dots, n\},$$

where $\mu_t^n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i(t)}$, and $\tilde{A}(z_k, z_l, \mu) := \Sigma_{k,l}(\mathbf{z}) = \mathbb{E}_\xi G(z_k, \mu, \xi) \otimes G(z_l, \mu, \xi)$

Martingale problem for empirical distribution

$$dZ_k(t) = V(Z_k(t), \mu_t^n)dt + \sqrt{\alpha}(\Sigma^{\frac{1}{2}})_k(\mathbf{Z}(t))dB(t), \quad k \in \{1, \dots, n\},$$

where $\mu_t^n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i(t)}$, and $\tilde{A}(z_k, z_l, \mu) := \Sigma_{k,l}(\mathbf{z}) = \mathbb{E}_\xi G(z_k, \mu, \xi) \otimes G(z_l, \mu, \xi)$

[Rotskoff, Vanden-Eijnden, CPAM, 2022]:

Taking $\varphi \in \mathrm{C}_c^2(\mathbb{R}^d)$, we get for the empirical measure μ_t^n

$$\langle \varphi, \mu_t^n \rangle = \langle \varphi, \mu_0^n \rangle + \frac{\alpha}{2}\int_0^t \left\langle \nabla^2 \varphi : A(\cdot, \mu_s^n), \mu_s^n \right\rangle ds + \int_0^t \langle \nabla \varphi \cdot V(\cdot, \mu_s^n), \mu_s^n \rangle ds$$
$$+ \text{Mart.},$$

where $A(z, \mu) = \tilde{A}(z, z, \mu)$ and

Martingale problem for empirical distribution

$$dZ_k(t) = V(Z_k(t), \mu_t^n)dt + \sqrt{\alpha}(\Sigma^{\frac{1}{2}})_k(\mathbf{Z}(t))dB(t), \quad k \in \{1, \dots, n\},$$

where $\mu_t^n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i(t)}$, and $\tilde{A}(z_k, z_l, \mu) := \Sigma_{k,l}(\mathbf{z}) = \mathbb{E}_\xi G(z_k, \mu, \xi) \otimes G(z_l, \mu, \xi)$

[Rotskoff, Vanden-Eijnden, CPAM, 2022]:

Taking $\varphi \in \mathrm{C}_c^2(\mathbb{R}^d)$, we get for the empirical measure μ_t^n

$$\langle \varphi, \mu_t^n \rangle = \langle \varphi, \mu_0^n \rangle + \frac{\alpha}{2}\int_0^t \left\langle \nabla^2\varphi : A(\cdot, \mu_s^n), \mu_s^n \right\rangle ds + \int_0^t \langle \nabla\varphi \cdot V(\cdot, \mu_s^n), \mu_s^n \rangle ds$$
$$+ \text{Mart.},$$

where $A(z, \mu) = \tilde{A}(z, z, \mu)$ and

$$[\text{Mart.}]_t = \alpha \int_0^t \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (\nabla\varphi(x) \otimes \nabla\varphi(y)) : \tilde{A}(x, y, \mu_s^n)\mu_s^n(dx)\mu_s^n(dy)ds$$

SGD and distribution dependent stochastic flow

The dynamics of the SGD

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

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.

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

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

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]

Our goal

Goal: Compare the SGD dynamics $\nu_t^n = \frac{1}{n}\sum_{k=1}^n \delta_{z_k(t)}$ defined via

$$z_k(t_{i+1}) = z_k(t_i) + \alpha \tilde{V}(z_k(t_i), \nu_{t_i}^n, \xi_i)$$

with $z_k(0) \sim \mu_0$ i.i.d. and mean-field dynamics

Our goal

Goal: Compare the SGD dynamics $\nu_t^n = \frac{1}{n}\sum_{k=1}^n \delta_{z_k(t)}$ defined via

$$z_k(t_{i+1}) = z_k(t_i) + \alpha \tilde{V}(z_k(t_i), \nu_{t_i}^n, \xi_i)$$

with $z_k(0) \sim \mu_0$ i.i.d. and mean-field dynamics

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

started from μ_0.

Our goal

Goal: Compare the SGD dynamics $\nu_t^n = \frac{1}{n}\sum_{k=1}^n \delta_{z_k(t)}$ defined via

$$z_k(t_{i+1}) = z_k(t_i) + \alpha \tilde{V}(z_k(t_i), \nu_{t_i}^n, \xi_i)$$

with $z_k(0) \sim \mu_0$ i.i.d. and mean-field dynamics

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

started from μ_0.

Let

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

Idea: We will focus on the case $\alpha = \frac{1}{n}$ and compare the fluctuation fields for both dynamics:

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

CLT for the SGD

We set

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

where

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

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 $D([0,\infty), H_{-J})$ 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 $(V(z,\mu) = \nabla F(z) - \langle \nabla K(z,\cdot), \mu\rangle)$

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

Quantified CLT for SMFE

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

Let μ_t^n be a solution to

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

started from $\mu_0^n = \nu_0^n = \frac{1}{n}\sum_{k=1}^n \delta_{z_k(0)}$ with $z_k(0) \sim \mu_0$ i.i.d. Then

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

where ζ is a Gaussian process solving

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

Moreover, $\left(\mathbb{E} \sup_{t\in[0,T]} \|\eta_t^n - \zeta_t\|_{H_{-J}}^2\right)^{\frac{1}{2}} \leq \frac{C}{\sqrt{n}}$.

Quantified CLT for SMFE

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

Let μ_t^n be a solution to

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

started from $\mu_0^n = \nu_0^n = \frac{1}{n}\sum_{k=1}^n \delta_{z_k(0)}$ with $z_k(0) \sim \mu_0$ i.i.d. Then

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

where ζ is a Gaussian process solving

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

Moreover, $\left(\mathbb{E} \sup_{t\in[0,T]} \|\eta_t^n - \zeta_t\|_{H_{-J}}^2\right)^{\frac{1}{2}} \leq \frac{C}{\sqrt{n}}$.

Stochastic mean-field equation captures the fluctuations of the SGD dynamics.

Higher Order Approximation

Let

$$W_p(\rho_1, \rho_2) = \inf_{\xi_i \sim \rho_i} \mathbb{E}\left[\|\xi_1 - \xi_2\|^p\right]^{\frac{1}{p}}$$

be the p-Wasserstein distance

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

Let V and G be Lipschitz continuous and V, G be quite regular in z. Then one has

$$W_p\left(\mathrm{Law}\mu_\cdot^n, \mathrm{Law}\nu_\cdot^n\right) = o\left(\frac{1}{\sqrt{n}}\right)$$

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,

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

Open problem

What is the rate of convergence in CLT for SGD:

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

Open problem

What is the rate of convergence in CLT for SGD:

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

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

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

$$W_1\left(\mathrm{Law}\left(\frac{1}{n} w_t\right), \delta_0\right) = \mathbb{E}\left[\left|\frac{1}{n} w_t - 0\right|\right] = O\left(\frac{1}{n}\right)$$

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

$$W_1\left(\mathrm{Law}\left(\frac{1}{n} w_t\right), \delta_0\right) = \mathbb{E}\left[\left|\frac{1}{n} w_t - 0\right|\right] = O\left(\frac{1}{n}\right)$$

but

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

Comparison in weak topology

Recall the SGD:

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

and the Distribution Dependent Stochastic Flow:

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

Comparison in weak topology

Recall the SGD:

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

and the Distribution Dependent Stochastic Flow:

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

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 $\Phi \in \mathrm{C}^4_b(\mathcal{P}_2)$

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

Comparison in weak topology

Recall the SGD:

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

and the Distribution Dependent Stochastic Modified Flow:

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

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 $\Phi \in \mathrm{C}^4_b(\mathcal{P}_2)$

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

Idea of proof. SGD as a flow

The SGD

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

where $\nu^n_t = \frac{1}{n}\sum_{k=1}^n \delta_{z_k(t)}$ can be build as follows:

$$z(u, t_{i+1}) = z(u, t_i) + \alpha \tilde{V}(z(u, t_i), \nu_{t_i}, \xi_i),$$
$$z(u, 0) = u, \quad \nu_{t_i} = \nu_0^{-1} \circ z(\cdot, t_i)$$

by taking $\nu_0 := \nu^n_0$.

Idea of proof. SGD as a flow

The SGD

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

where $\nu^n_t = \frac{1}{n}\sum_{k=1}^n \delta_{z_k(t)}$ can be build as follows:

$$z(u, t_{i+1}) = z(u, t_i) + \alpha \tilde{V}(z(u, t_i), \nu_{t_i}, \xi_i),$$
$$z(u, 0) = u, \quad \nu_{t_i} = \nu_0^{-1} \circ z(\cdot, t_i)$$

by taking $\nu_0 := \nu^n_0$.

Recall

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

Idea of proof. Difference of semigroups

Set $t_1 := \Delta t := \alpha$ and define

$$\mathcal{S}\Psi(\mu_0) := \mathbb{E}_{\mu_0}\Psi(\nu_{t_1})$$

and

$$\mathcal{T}_t\Psi(\mu_0) := \mathbb{E}_{\mu_0}\Psi(\mu_t).$$

Idea of proof. Difference of semigroups

Set $t_1 := \Delta t := \alpha$ and define

$$\mathcal{S}\Psi(\mu_0) := \mathbb{E}_{\mu_0}\Psi(\nu_{t_1})$$

and

$$\mathcal{T}_t\Psi(\mu_0) := \mathbb{E}_{\mu_0}\Psi(\mu_t).$$

Then for $t_n = n\alpha = n\Delta t$

$$\begin{aligned}\mathbb{E}_{\mu_0}\Phi\left(\nu_{t_n}\right) - \mathbb{E}_{\mu_0}\Phi\left(\mu_{t_n}\right) &= \mathcal{S}^n\Phi(\mu_0) - \mathcal{T}_{t_n}\Phi(\mu_0) \\ &= \sum_{i=0}^{n-1}\left(\mathcal{S}^{n-i}\mathcal{T}_{t_i}\Phi(\mu_0) - \mathcal{S}^{n-i-1}\mathcal{T}_{t_i}\Phi(\mu_0)\right) \\ &= \sum_{i=0}^{n-1}\mathcal{S}^{n-i-1}\left(\mathcal{S}\mathcal{T}_{t_i}\Phi(\mu_0) - \mathcal{T}_\alpha\mathcal{T}_{t_i}\Phi(\mu_0)\right).\end{aligned}$$

Idea of proof. Difference of semigroups

Set $t_1 := \Delta t := \alpha$ and define

$$\mathcal{S}\Psi(\mu_0) := \mathbb{E}_{\mu_0}\Psi(\nu_{t_1})$$

and

$$\mathcal{T}_t\Psi(\mu_0) := \mathbb{E}_{\mu_0}\Psi(\mu_t).$$

Then for $t_n = n\alpha = n\Delta t$

$$\begin{aligned}\mathbb{E}_{\mu_0}\Phi\left(\nu_{t_n}\right) - \mathbb{E}_{\mu_0}\Phi\left(\mu_{t_n}\right) &= \mathcal{S}^n\Phi(\mu_0) - \mathcal{T}_{t_n}\Phi(\mu_0)\\ &= \sum_{i=0}^{n-1}\left(\mathcal{S}^{n-i}\mathcal{T}_{t_i}\Phi(\mu_0) - \mathcal{S}^{n-i-1}\mathcal{T}_{t_i}\Phi(\mu_0)\right)\\ &= \sum_{i=0}^{n-1}\mathcal{S}^{n-i-1}\left(\mathcal{S}\mathcal{T}_{t_i}\Phi(\mu_0) - \mathcal{T}_\alpha\mathcal{T}_{t_i}\Phi(\mu_0)\right).\end{aligned}$$

Thus

$$\sup_{\mu_0\in\mathcal{P}}\left|\mathbb{E}_{\mu_0}\Phi\left(\nu_{t_n}\right) - \mathbb{E}_{\mu_0}\Phi\left(\mu_{t_n}\right)\right| \le \sum_{i=0}^{n-1}\sup_{\mu_0\in\mathcal{P}_2}\left|\mathcal{S}\left[\mathcal{T}_{t_i}\Phi(\mu_0)\right] - \mathcal{T}_\alpha\left[\mathcal{T}_{t_i}\Phi(\mu_0)\right]\right|.$$

Idea of proof. Expansions of semigroups

Expansion in Taylor's series w.r.t α

$$\mathcal{S}\Psi(\mu_0) = \Psi(\mu_0) + \alpha \int_{\mathbb{R}^d} D\Psi(z, \mu_0) \cdot V(z, \mu_0)\mu_0(dz) + \alpha^2(\ldots) + \alpha^3 R_1(\Psi, \mu_0),$$

where $\sup_{\mu_0 \in \mathcal{P}_2} |R_1| \leq C\|\Psi\|_{C_b^3}$.

Idea of proof. Expansions of semigroups

Expansion in Taylor's series w.r.t α

$$\mathcal{S}\Psi(\mu_0) = \Psi(\mu_0) + \alpha \int_{\mathbb{R}^d} D\Psi(z, \mu_0) \cdot V(z, \mu_0)\mu_0(dz) + \alpha^2(\ldots) + \alpha^3 R_1(\Psi, \mu_0),$$

where $\sup_{\mu_0 \in \mathcal{P}_2} |R_1| \leq C\|\Psi\|_{C_b^3}$.

$$\mathcal{T}_\alpha \Psi(\mu_0) = \Psi(\mu_0) + \int_0^\alpha \mathcal{L}\mathcal{T}_s \Psi(\mu_0) ds,$$

where $\mathcal{L} = \mathcal{L}_1 + \alpha\mathcal{L}_2$ and

$$\mathcal{L}_1\Psi(\mu_0) = \int_{\mathbb{R}^d} D\Psi(z, \mu_0) \cdot V(z, \mu_0)\mu_0(dz), \quad \mathcal{L}_2\Psi(\mu_0) = \ldots$$

Iterating the equality above, one gets

$$\mathcal{T}_\alpha \Psi(\mu_0) = \Psi(\mu_0) + \alpha\mathcal{L}_1\Psi(\mu_0) + \alpha^2 \left(\mathcal{L}_2 + \frac{1}{2}\mathcal{L}_1^2\right) \Psi(\mu_0) + \alpha^3 R_2(\Psi, \mu_0),$$

where $\sup_{\mu_0 \in \mathcal{P}_2} |R_2| \leq C\|\Psi\|_{C_b^4}$.

Further observations

Assume that the SGD dynamics

$$\begin{aligned} z_k(t_{i+1}) &= z_k(t_i) + \alpha \tilde{V}(z_k(t_i), \nu^n_{t_i}, \xi_i) \\ &= z_k(t_i) + \alpha \tilde{V}(z_k(t_i), \xi_i) \\ &= z_k(t_i) - \alpha \nabla \tilde{R}(z_k(t_i), \xi_i) \end{aligned}$$

is independent of $\nu^n_{t_i}$.

Further observations

Assume that the SGD dynamics

$$\begin{aligned} z_k(t_{i+1}) &= z_k(t_i) + \alpha \tilde{V}(z_k(t_i), \nu^n_{t_i}, \xi_i) \\ &= z_k(t_i) + \alpha \tilde{V}(z_k(t_i), \xi_i) \\ &= z_k(t_i) - \alpha \nabla \tilde{R}(z_k(t_i), \xi_i) \end{aligned}$$

is independent of $\nu^n_{t_i}$.

Then for each k, $z_k(t)$ is SGD dynamics for the minimization of

$$z \mapsto \mathbb{E}_\xi \tilde{R}(z, \xi)$$

with initialization $z_k(0)$.

Further observations

Assume that the SGD dynamics

$$\begin{aligned} z_k(t_{i+1}) &= z_k(t_i) + \alpha \tilde{V}(z_k(t_i), \nu^n_{t_i}, \xi_i) \\ &= z_k(t_i) + \alpha \tilde{V}(z_k(t_i), \xi_i) \\ &= z_k(t_i) - \alpha \nabla \tilde{R}(z_k(t_i), \xi_i) \end{aligned}$$

is independent of $\nu^n_{t_i}$.

Then for each k, $z_k(t)$ is SGD dynamics for the minimization of

$$z \mapsto \mathbb{E}_\xi \tilde{R}(z, \xi)$$

with initialization $z_k(0)$.

Then the stochastic flow captures n-point motion of the SGD

$$dZ_t(u) = V(Z_t(u))dt - \frac{\alpha}{4}\nabla |V(Z_t(u))|^2 dt + \sqrt{\alpha} \int_\Theta G(Z_t(u), \xi) W(d\xi, dt),$$

$$Z_0(u) = u,$$

$$\mu_t = \nu^n_0 \circ Z_t^{-1}, \qquad \nu^n_0 = \frac{1}{n} \sum_{k=1}^n \delta_{z_k(0)}$$

n -point motion for SGD

Consider SGD dynamics

$$z_k(t_{i+1}) = z_k(t_i) - \alpha \nabla R(z_k(t_i), \xi_i),$$

with initialization $z_k(0)$, $k \in \{1, \dots, n\}$.

Consider the Stochastic Modified Flow:

$$dZ_t(u) = V(Z_t(u))dt - \frac{\alpha}{4}\nabla |V(Z_t(u))|^2 dt + \sqrt{\alpha}\int_\Theta G(Z_t(u), \xi) W(d\xi, dt),$$

$$Z_0(u) = u,$$

$$\mu_t = \nu_0^n \circ Z_t^{-1}, \qquad \nu_0^n = \frac{1}{n}\sum_{k=1}^n \delta_{z_k(0)}$$

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

Define $Z_k(t) := Z(z_k(0), t)$, $k \in \{1, \dots, n\}$. Then for every $f \in \mathrm{C}_b^4(\mathbb{R}^{dn})$

$$\sup_{t_i \le T} |\mathbb{E} f(z_1(t_i), \dots, z_n(t_i)) - \mathbb{E} f(Z_1(t_i), \dots, Z_n(t_i))| \le C\alpha^2.$$

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!