Conservative SPDEs as fluctuating mean field limits of stochastic gradient descent

Vitalii Konarovskyi

Bielefeld University

Malliavin Calculus and its Applications — Kyiv

joint work with Benjamin Gess and Rishabh Gvalani

Table of Contents

Supervised learning

- Having a large sets of data $\{(\theta_i, \gamma_i),\ i \in I\}$, one needs to find a function $f : \Theta \to \mathbb{R}$ such that $f(\theta_i) = \gamma_i$.

Supervised learning

- Having a large sets of data $\{(\theta_i, \gamma_i),\ i \in I\}$, one needs to find a function $f : \Theta \to \mathbb{R}$ such that $f(\theta_i) = \gamma_i$.
- Usually one approximates f by

$$f_n(\theta) = \frac{1}{n} \sum_{k=1}^{n} U(\theta, x_k),$$

where $x_k \in \mathbb{R}^d$, $k \in \{1, \dots, n\}$, are parameters which have to be found.
Example: $U(\theta, x) = c \cdot h(a \cdot \theta + b), \quad x = (a, b, c)$

Supervised learning

- Having a large sets of data $\{(\theta_i, \gamma_i),\ i \in I\}$, one needs to find a function $f : \Theta \to \mathbb{R}$ such that $f(\theta_i) = \gamma_i$.
- Usually one approximates f by

$$f_n(\theta) = \frac{1}{n} \sum_{k=1}^{n} U(\theta, x_k),$$

where $x_k \in \mathbb{R}^d$, $k \in \{1, \dots, n\}$, are parameters which have to be found.
Example: $U(\theta, x) = c \cdot h(a \cdot \theta + b), \quad x = (a, b, c)$

- We measure the distance between f and f_n by the **generalization error**

$$\mathcal{L}[f_n] = \frac{1}{2} \mathbb{E}_m l(f(\theta), f_n(\theta)) = \frac{1}{2} \int_\Theta l(f(\theta), f_n(\theta)) \mathrm{m}(d\theta),$$

where m is the distribution of θ_i.

Supervised learning

- Having a large sets of data $\{(\theta_i, \gamma_i),\ i \in I\}$, one needs to find a function $f : \Theta \to \mathbb{R}$ such that $f(\theta_i) = \gamma_i$.
- Usually one approximates f by

$$f_n(\theta) = \frac{1}{n} \sum_{k=1}^{n} U(\theta, x_k),$$

where $x_k \in \mathbb{R}^d$, $k \in \{1, \dots, n\}$, are parameters which have to be found.
Example: $U(\theta, x) = c \cdot h(a \cdot \theta + b), \quad x = (a, b, c)$

- We measure the distance between f and f_n by the **generalization error**

$$\mathcal{L}[f_n] = \frac{1}{2}\mathbb{E}_m|f(\theta) - f_n(\theta)|^2 = \frac{1}{2}\int_\Theta |f(\theta) - f_n(\theta)|^2 \mathrm{m}(d\theta),$$

where m is the distribution of θ_i.

Stochastic gradient descent

The parameters x_k, $k \in \{1, \dots, n\}$ can be learned by stochastic gradient descent

$$\hat{x}_k(t_{i+1}) = \hat{x}_k(t_i) - \nabla_{x_k} l(f(\theta_i), f_n(\theta_i; x))\Delta t$$

where Δt is a **learning rate**, $t_i = i\Delta t$, $\{\theta_i, i \in \mathbb{N}\}$ are iid with distribution m,

Stochastic gradient descent

The parameters x_k, $k \in \{1, \dots, n\}$ can be learned by stochastic gradient descent

$$\begin{aligned} \hat{x}_k(t_{i+1}) &= \hat{x}_k(t_i) - \nabla_{x_k} l(f(\theta_i), f_n(\theta_i; x))\Delta t \\ &= \hat{x}_k(t_i) - (f_n(\theta_i; x) - f(\theta_i))\, \nabla_{x_k} U(\theta_i, \hat{x}_k(t_i))\Delta t \end{aligned}$$

where Δt is a **learning rate**, $t_i = i\Delta t$, $\{\theta_i, i \in \mathbb{N}\}$ are iid with distribution m,

Stochastic gradient descent

The parameters x_k, $k \in \{1, \dots, n\}$ can be learned by stochastic gradient descent

$$\begin{aligned}\hat{x}_k(t_{i+1}) &= \hat{x}_k(t_i) - \nabla_{x_k} l(f(\theta_i), f_n(\theta_i; x))\Delta t \\ &= \hat{x}_k(t_i) - (f_n(\theta_i; x) - f(\theta_i)) \nabla_{x_k} U(\theta_i, \hat{x}_k(t_i))\Delta t \\ &= \hat{x}_k(t_i) + \left(\nabla F_i(\hat{x}_k(t_i)) - \frac{1}{n} \sum_{l=1}^{n} \nabla_{x_k} K_i(\hat{x}_k(t_i), \hat{x}_l(t_i)) \right) \Delta t\end{aligned}$$

where Δt is a **learning rate**, $t_i = i\Delta t$, $\{\theta_i, i \in \mathbb{N}\}$ are iid with distribution m, $F_i(x) = f(\theta_i)U(\theta_i, x)$ and $K_i(x, y) = U(\theta_i, x)U(\theta_i, y)$.

Stochastic gradient descent

The parameters x_k, $k \in \{1, \ldots, n\}$ can be learned by stochastic gradient descent

$$\begin{aligned}\hat{x}_k(t_{i+1}) &= \hat{x}_k(t_i) - \nabla_{x_k} l(f(\theta_i), f_n(\theta_i; x))\Delta t \\ &= \hat{x}_k(t_i) - (f_n(\theta_i; x) - f(\theta_i)) \nabla_{x_k} U(\theta_i, \hat{x}_k(t_i))\Delta t \\ &= \hat{x}_k(t_i) + \left(\nabla F_i(\hat{x}_k(t_i)) - \langle \nabla_x K_i(\hat{x}_k(t_i), \cdot), \hat{\mu}^n_{t_i} \rangle \right) \Delta t\end{aligned}$$

where Δt is a **learning rate**, $t_i = i\Delta t$, $\{\theta_i, i \in \mathbb{N}\}$ are iid with distribution m, $\hat{\mu}^n_t = \frac{1}{n}\sum_{l=1}^n \delta_{\hat{x}_l(t)}$, $F_i(x) = f(\theta_i)U(\theta_i, x)$ and $K_i(x, y) = U(\theta_i, x)U(\theta_i, y)$.

Stochastic gradient descent

The parameters x_k, $k \in \{1, \dots, n\}$ can be learned by stochastic gradient descent

$$\begin{aligned}\hat{x}_k(t_{i+1}) &= \hat{x}_k(t_i) - \nabla_{x_k} l(f(\theta_i), f_n(\theta_i; x))\Delta t \\ &= \hat{x}_k(t_i) - (f_n(\theta_i; x) - f(\theta_i))\, \nabla_{x_k} U(\theta_i, \hat{x}_k(t_i))\Delta t \\ &= \hat{x}_k(t_i) + \left(\nabla F_i(\hat{x}_k(t_i)) - \langle \nabla_x K_i(\hat{x}_k(t_i), \cdot), \hat{\mu}^n_{t_i} \rangle \right) \Delta t \\ &= \hat{x}_k(t_i) + V_i(\hat{x}_k(t_i), \hat{\mu}^n_{t_i})\Delta t\end{aligned}$$

where Δt is a **learning rate**, $t_i = i\Delta t$, $\{\theta_i, i \in \mathbb{N}\}$ are iid with distribution m, $\hat{\mu}^n_t = \frac{1}{n} \sum_{l=1}^n \delta_{\hat{x}_l(t)}$, $F_i(x) = f(\theta_i) U(\theta_i, x)$ and $K_i(x, y) = U(\theta_i, x) U(\theta_i, y)$.

Convergence to deterministic SPDE

According to [Mei, Montanarib, Nguyen. A mean field view of the landscape of two-layer neural networks]

$$d(\hat{\mu}_t^n, \mu_t) = O\left(\frac{1}{\sqrt{n}}\right) + O\left(\sqrt{\Delta t}\right),$$

where μ_t solves

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

with

$$V(x, \mu) = \mathbb{E}V_i(x, \mu) = \nabla F(x) - \langle \nabla_x K(x, \cdot), \mu\rangle$$

and

$$F(x) = \mathbb{E}_m f(\theta)U(\theta, x), \quad K(x, y) = \mathbb{E}_m[U(\theta, x)U(\theta, y)].$$

Main goal

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

Goal: To identify a class of nonlinear conservative SPDEs which serve as such a fluctuating continuum model and show that those equations give a better approximation of the SGD dynamics than the deterministic SDE in the overparametrised regime.

SDE for SGD

Stochastic gradient descent

$$\hat{x}_k(t_{i+1}) = \hat{x}_k(t_i) + V_i(\hat{x}_k(t_i), \hat{\mu}^n_{t_i})\Delta t$$
$$= \hat{x}_k(t_i) + V(\hat{x}_k(t_i), \hat{\mu}^n_{t_i})\Delta t + \sqrt{\Delta t}\left(V_i(\hat{x}_k(t_i), \hat{\mu}^n_{t_i}) - V(\hat{x}_k(t_i), \hat{\mu}^n_{t_i})\right)\sqrt{\Delta t}$$

is the Euler-Maruyama scheme for the SDE

$$dx_k(t) = V(x_k(t), \mu^n_t)dt + \sqrt{\Delta t} dB_k(t), \quad k \in \{1, \dots, n\}$$

$$d[B_k, B_l]_t = \mathrm{Cov}\,(V_i, V_i)\, dt = \tilde{A}(x_k(t), x_l(t), \mu^n_t)dt,$$

where $\mu^n_t = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$, $k, l \in \{1, \dots, n\}$.

Equation for empirical measure μ_t^n

We came to the SDE

$$dx_k(t) = V(x_k(t), \mu_t^n)dt + \sqrt{\alpha}dB_k(t)$$
$$d[B_k, B_l]_t = \tilde{A}(x_k(t), x_l(t), \mu_t^n)dt,$$

where $\mu_t^n = \frac{1}{n}\sum_{l=1}^n \delta_{x_l(t)}$, $\tilde{A}(x, y, \mu) = \left(\mathbb{E}_{\mathrm{m}} G_k(x, \mu, \theta) G_l(y, \mu, \theta)\right)_{i,j\in[d]}$.

[1] $B : C = \sum_{i,j=1}^d B_{i,j} C_{i,j}$

Equation for empirical measure μ_t^n

We came to the SDE

$$dx_k(t) = V(x_k(t), \mu_t^n)dt + \sqrt{\alpha}dB_k(t)$$
$$d[B_k, B_l]_t = \tilde{A}(x_k(t), x_l(t), \mu_t^n)dt,$$

where $\mu_t^n = \frac{1}{n}\sum_{l=1}^n \delta_{x_l(t)}$, $\tilde{A}(x, y, \mu) = (\mathbb{E}_{\mathrm{m}} G_k(x, \mu, \theta) G_l(y, \mu, \theta))_{i,j\in[d]}$.

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

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

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

[1] $B : C = \sum_{i,j=1}^d B_{i,j} C_{i,j}$

Equation for empirical measure μ_t^n

We came to the SDE

$$dx_k(t) = V(x_k(t), \mu_t^n)dt + \sqrt{\alpha} dB_k(t)$$
$$d[B_k, B_l]_t = \tilde{A}(x_k(t), x_l(t), \mu_t^n)dt,$$

where $\mu_t^n = \frac{1}{n}\sum_{l=1}^n \delta_{x_l(t)}$, $\tilde{A}(x,y,\mu) = \left(\mathbb{E}_{\mathrm{m}} G_k(x,\mu,\theta)G_l(y,\mu,\theta)\right)_{i,j\in[d]}$.

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

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

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

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

[1] $B : C = \sum_{i,j=1}^d B_{i,j}C_{i,j}$

Overparametrised limit ($n \to \infty$)

Assuming that the number of parameters $n \to \infty$ and $x_i(0) \sim \mu_0$ are i.i.d., the limit $\mu_t = \lim_{n\to\infty} \mu_t^n$ solves the SPDE: $\forall \varphi \in \mathcal{C}_c^2(\mathbb{R}^d)$

$$\langle \varphi, \mu_t \rangle = \langle \varphi, \mu_0 \rangle + \frac{\alpha}{2} \int_0^t \langle \nabla^2 \varphi : A(\cdot, \mu_s), \mu_s \rangle \, ds + \int_0^t \langle \nabla \varphi \cdot V(\cdot, \mu_s), \mu_s \rangle \, ds + M_\varphi(t),$$

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

where $\tilde{A}(x, y, \mu) = (\mathbb{E}_{\mathrm{m}} G_k(x, \mu, \theta) G_l(y, \mu, \theta))_{k,l \in [d]}$ and $A(x, \mu) = \tilde{A}(x, x, \mu)$.

For more details regarding derivation of the martingale problem above see [Rotskoff, Vanden-Eijnden *Trainability and accuracy off neural networks: an interacting particle system approach* (to appear in CPAM)]

Stochastic mean-field equation

We will assume the noise of equation has a special structure:
we will take a cylindrical Wiener process W on $L_2(\Theta, \mathrm{m})$ and assume

$$M_\varphi(t) = \sqrt{\alpha} \int_0^t \int_\Theta \langle \nabla \varphi \cdot G(\cdot, \mu_s, \theta), \mu_s \rangle W(d\theta, ds)$$

Stochastic mean-field equation

We will assume the noise of equation has a special structure:
we will take a cylindrical Wiener process W on $L_2(\Theta, \mathrm{m})$ and assume

$$M_\varphi(t) = \sqrt{\alpha} \int_0^t \int_\Theta \langle \nabla\varphi \cdot G(\cdot, \mu_s, \theta), \mu_s \rangle W(d\theta, ds)$$

then

$$\begin{aligned}[M_\varphi]_t &= \alpha \int_0^t \int_\Theta \langle \nabla\varphi \cdot G(\cdot, \mu_s, \theta), \mu_s \rangle \langle \nabla\varphi \cdot G(\cdot, \mu_s, \theta), \mu_s \rangle \, \mathrm{m}(d\theta) ds \\ &= \int_0^t \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (\nabla\varphi(x) \otimes \nabla\varphi(y)) : \tilde{A}(x, y, \mu_s) \mu_s(dx) \mu_s(dy) ds\end{aligned}$$

Stochastic mean-field equation

We will assume the noise of equation has a special structure:
we will take a cylindrical Wiener process W on $L_2(\Theta, \mathrm{m})$ and assume

$$M_\varphi(t) = \sqrt{\alpha} \int_0^t \int_\Theta \langle \nabla\varphi \cdot G(\cdot, \mu_s, \theta), \mu_s \rangle W(d\theta, ds)$$

then

$$[M_\varphi]_t = \alpha \int_0^t \int_\Theta \langle \nabla\varphi \cdot G(\cdot, \mu_s, \theta), \mu_s \rangle \langle \nabla\varphi \cdot G(\cdot, \mu_s, \theta), \mu_s \rangle \, \mathrm{m}(d\theta) ds$$
$$= \int_0^t \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (\nabla\varphi(x) \otimes \nabla\varphi(y)) : \tilde{A}(x, y, \mu_s) \mu_s(dx) \mu_s(dy) ds$$

We come to the **Stochastic Mean-Field Equation** (SMFE):

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

Table of Contents

Definition of solutions to SMFE

$$d\mu_t = \frac{1}{2}\nabla^2 : (A(\cdot,\mu_t)\mu_t)\,dt - \nabla\cdot(V(\cdot,\mu_t)\mu_t)\,dt - \nabla\cdot\int_\Theta G(\cdot,\mu_t,\theta)\mu_t W(d\theta,dt)$$

Definition of (weak-strong) solution

A continuous $(\mathcal{F}_t^W)$-adapted process μ_t, $t \geq 0$, in $\mathcal{P}_2(\mathbb{R}^d)$ is a *solution to SMFE* started from μ_0 if $\forall\ \varphi \in \mathcal{C}_c^2(\mathbb{R}^d)$ a.s. $\forall t \geq 0$

$$\langle \varphi,\mu_t\rangle = \langle\varphi,\mu_0\rangle + \frac{1}{2}\int_0^t \left\langle \nabla^2\varphi : A(\cdot,\mu_s),\mu_s\right\rangle ds + \int_0^t \langle\nabla\varphi\cdot V(\cdot,\mu_s),\mu_s\rangle\, ds$$
$$+\int_0^t\int_\Theta \langle\nabla\varphi\cdot G(\cdot,\mu_s,\theta),\mu_s\rangle\, W(d\theta,ds)$$

SDE with interaction

The SMFE has a connection with the SDE with interaction (Kotelenez, Dorogovtsev)

$$dX(u,t) = V(X(u,t),\bar{\mu}_t)dt + \int_\Theta G(X(u,t),\bar{\mu}_t,\theta)W(d\theta,dt),$$
$$X(u,0) = u, \quad \bar{\mu}_t = \mu_0 \circ X^{-1}(\cdot,t), \quad u \in \mathbb{R}^d, \quad t \geq 0.$$

SDE with interaction

The SMFE has a connection with the SDE with interaction (Kotelenez, Dorogovtsev)

$$dX(u,t) = V(X(u,t),\bar{\mu}_t)dt + \int_\Theta G(X(u,t),\bar{\mu}_t,\theta)W(d\theta,dt),$$
$$X(u,0) = u, \quad \bar{\mu}_t = \mu_0 \circ X^{-1}(\cdot,t), \quad u \in \mathbb{R}^d, \quad t \geq 0.$$

Theorem (Dorogovtsev' 07)

Let V, G be Lipschitz continuous, i.e. $\exists L > 0$ such that a.s.

$$|V(x,\mu) - V(y,\nu)| + |||G(x,\mu,\cdot) - G(y,\nu,\cdot)|||_{\mathrm{m}} \leq L\left(|x-y| + \mathcal{W}_2(\mu,\nu)\right).$$

Then for every $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ the SDE with interaction has a unique solution started from μ_0.

Well-posedness of SMFE

Theorem (Gess, Gvalani, K. 2022)

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

$$d\mu_t = \frac{1}{2}\nabla^2 : (A(\cdot, \mu_t)\mu_t)\, dt - \nabla \cdot (V(\cdot, \mu_t)\mu_t)\, dt - \nabla \cdot \int_\Theta G(\cdot, \mu_t, \theta)\mu_t W(d\theta, dt)$$

has a unique solution. Moreover, μ_t is a **superposition solution**, i.e.,

$$\mu_t = \mu_0 \circ X^{-1}(\cdot, t), \quad t \geq 0,$$

where X solves

$$dX(u,t) = V(X(u,t), \mu_t)dt + \int_\Theta G(X(u,t), \mu_t, \theta)W(d\theta, dt), \quad X(u,0) = u.$$

Table of Contents

Wasserstein distance

Let (E, d) be a Polish space, and for $p \geq 1$ $\mathcal{P}_p(E)$ be a space of all probability measures ρ on E with

$$\int_E d^p(x, o)\rho(dx) < \infty.$$

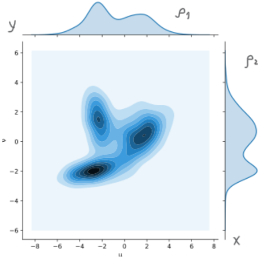
Wasserstein distance

Let (E, d) be a Polish space, and for $p \geq 1$ $\mathcal{P}_p(E)$ be a space of all probability measures ρ on E with

$$\int_E d^p(x, o)\rho(dx) < \infty.$$

For $\rho_1, \rho_2 \in \mathcal{P}_p(E)$ we define the **Wasserstein distance** by

$$\mathcal{W}_p^p(\rho_1, \rho_2) = \inf \left\{ \int_{E^2} d^p(x, y)\chi(dx, dy) : \begin{array}{l} \chi(\cdot \times E) = \rho_1, \\ \chi(E \times \cdot) = \rho_2 \end{array} \right\}$$



Wikipedia

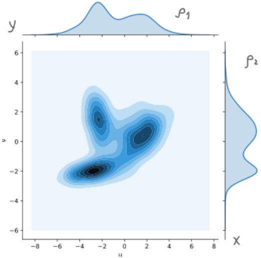
Wasserstein distance

Let (E, d) be a Polish space, and for $p \geq 1$ $\mathcal{P}_p(E)$ be a space of all probability measures ρ on E with

$$\int_E d^p(x, o)\rho(dx) < \infty.$$

For $\rho_1, \rho_2 \in \mathcal{P}_p(E)$ we define the **Wasserstein distance** by

$$\mathcal{W}_p^p(\rho_1, \rho_2) = \inf \left\{ \int_{E^2} d^p(x, y)\chi(dx, dy) : \quad \begin{array}{l} \chi(\cdot \times E) = \rho_1, \\ \chi(E \times \cdot) = \rho_2 \end{array} \right\}$$
$$= \inf \left\{ \mathbb{E} d^p(\xi_1, \xi_2) : \quad \xi_i \sim \rho_i \right\}$$



Wasserstein distance

Let (E, d) be a Polish space, and for $p \geq 1$ $\mathcal{P}_p(E)$ be a space of all probability measures ρ on E with

$$\int_E d^p(x, o)\rho(dx) < \infty.$$

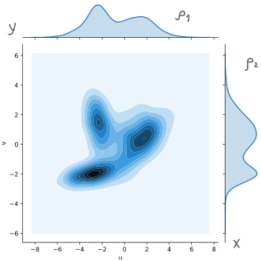
For $\rho_1, \rho_2 \in \mathcal{P}_p(E)$ we define the **Wasserstein distance** by

$$\mathcal{W}_p^p(\rho_1, \rho_2) = \inf\left\{\int_{E^2} d^p(x,y)\chi(dx,dy): \quad \begin{array}{l}\chi(\cdot \times E) = \rho_1, \\ \chi(E \times \cdot) = \rho_2\end{array}\right\}$$

$$= \inf\left\{\mathbb{E}d^p(\xi_1, \xi_2): \quad \xi_i \sim \rho_i\right\}$$

Proposition

$(\mathcal{P}_p(E), \mathcal{W}_p)$ is a Polish space.



Wikipedia

Convergence of the empirical measure

Theorem **(Gess, Gvalani, K. 2022)**

Let A, V, G be Lipschitz continuous and let $\mu^{n,\alpha}$ and μ^{α} be superposition solutions to the SMFE

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

started from $\mu_0^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and μ_0, respectively, where $x_i \sim \mu_0$ are independent. Then

$$\mathbb{E} \sup_{t\in[0,T]} \mathcal{W}_2^2(\mu_t^{n,\alpha}, \mu_t^{\alpha}) \leq C\mathbb{E}\mathcal{W}_2^2(\mu_0^n, \mu_0) \leq C' n^{-1},$$

where the constants C, C' are independent of α.

Idea of the proof

Since $\mu^{n,\alpha}$ and μ^{α} are superposition solutions,

$$\mu_t^{n,\alpha} = \mu_0^n \circ X_{n,\alpha}^{-1}(\cdot, t), \quad \mu^{\alpha} = \mu_0 \circ X_{\alpha}^{-1}(\cdot, t),$$

where $X_{n,\alpha}$ and X_{α} are solutions to

$$dX(u,t) = V(X(u,t), \mu_t)dt + \sqrt{\alpha} \int_{\Theta} G(X(u,t), \mu_t, \theta) W(d\theta, dt), \quad X(u,0) = u.$$

Idea of the proof

Since $\mu^{n,\alpha}$ and μ^{α} are superposition solutions,

$$\mu_t^{n,\alpha} = \mu_0^n \circ X_{n,\alpha}^{-1}(\cdot, t), \quad \mu^{\alpha} = \mu_0 \circ X_{\alpha}^{-1}(\cdot, t),$$

where $X_{n,\alpha}$ and X_{α} are solutions to

$$dX(u,t) = V(X(u,t), \mu_t)dt + \sqrt{\alpha} \int_{\Theta} G(X(u,t), \mu_t, \theta) W(d\theta, dt), \quad X(u,0) = u.$$

$$\begin{aligned}
\mathbb{E} \sup_{s \in [0,t]} |X_{n,\alpha}(u,s) - X_{\alpha}(v,s)|^2 &\leq 3|u-v|^2 \\
&+ 3t\mathbb{E} \int_0^t \left| V(X_{n,\alpha}(u,s), \mu_s^{n,\alpha}) - V(X_{\alpha}(v,s), \mu_s^{\alpha}) \right|^2 ds \\
&+ 3\alpha\mathbb{E} \sup_{s \in [0,t]} \left| \int_0^s \int_{\Theta} \left(G(X_{n,\alpha}(u,r), \mu_r^{n,\alpha}, \theta) - G(X_{\alpha}(v,r), \mu_r^{\alpha}, \theta) \right) W(d\theta, dr) \right|^2 \\
&\leq 3|u-v|^2 + C \int_0^t \left(\mathbb{E} \sup_{r \in [0,s]} |X_{n,\alpha}(u,s) - X_{\alpha}(v,s)|^2 + \mathbb{E}\mathcal{W}_2^2\left(\mu_s^{n,\alpha}, \mu_s^{\alpha}\right) \right) ds
\end{aligned}$$

Idea of the proof

Since $\mu^{n,\alpha}$ and μ^{α} are superposition solutions,

$$\mu_t^{n,\alpha} = \mu_0^n \circ X_{n,\alpha}^{-1}(\cdot, t), \quad \mu^{\alpha} = \mu_0 \circ X_{\alpha}^{-1}(\cdot, t),$$

where $X_{n,\alpha}$ and X_{α} are solutions to

$$dX(u,t) = V(X(u,t), \mu_t)dt + \sqrt{\alpha}\int_{\Theta} G(X(u,t), \mu_t, \theta)W(d\theta, dt), \quad X(u,0) = u.$$

$$\begin{aligned}
\mathbb{E}\sup_{s\in[0,t]}|X_{n,\alpha}(u,s) - X_{\alpha}(v,s)|^2 &\leq 3|u-v|^2 \\
&+ 3t\mathbb{E}\int_0^t \left|V(X_{n,\alpha}(u,s), \mu_s^{n,\alpha}) - V(X_{\alpha}(v,s), \mu_s^{\alpha})\right|^2 ds \\
&+ 3\alpha\mathbb{E}\sup_{s\in[0,t]}\left|\int_0^s\int_{\Theta}\left(G(X_{n,\alpha}(u,r), \mu_r^{n,\alpha}, \theta) - G(X_{\alpha}(v,r), \mu_r^{\alpha}, \theta)\right)W(d\theta, dr)\right|^2 \\
&\leq 3|u-v|^2 + C\int_0^t \left(\mathbb{E}\sup_{r\in[0,s]}|X_{n,\alpha}(u,s) - X_{\alpha}(v,s)|^2 + \mathbb{E}\mathcal{W}_2^2\left(\mu_s^{n,\alpha}, \mu_s^{\alpha}\right)\right) ds
\end{aligned}$$

Idea of the proof

Hence, for any χ with marginals μ_0^n and μ_0, we get

$$\mathbb{E} \sup_{s\in[0,t]} \mathcal{W}_2^2(\mu_s^{n,\alpha}, \mu_s^{\alpha}) \leq \mathbb{E} \sup_{s\in[0,t]} \int_{\mathbb{R}^{2d}} |X_{n,\alpha}(u,s) - X_{\alpha}(v,s)|^2 \chi(du,dv)$$
$$\leq C \int_{\mathbb{R}^{2d}} |u-v|^2 \chi(du,dv) + C \int_0^t \mathbb{E}\mathcal{W}_2^2(\mu_s^{n,\alpha}, \mu_s^{\alpha}) ds.$$

Idea of the proof

Hence, for any χ with marginals μ_0^n and μ_0, we get

$$\mathbb{E} \sup_{s\in[0,t]} \mathcal{W}_2^2(\mu_s^{n,\alpha}, \mu_s^{\alpha}) \le \mathbb{E} \sup_{s\in[0,t]} \int_{\mathbb{R}^{2d}} |X_{n,\alpha}(u,s) - X_\alpha(v,s)|^2 \chi(du,dv)$$

$$\le C\mathcal{W}_2^2(\mu_0^n, \mu_0) + C\int_0^t \mathbb{E} \sup_{r\in[0,s]} \mathcal{W}_2^2(\mu_r^{n,\alpha}, \mu_r^{\alpha}) ds.$$

Idea of the proof

Hence, for any χ with marginals μ_0^n and μ_0, we get

$$\mathbb{E}\sup_{s\in[0,t]}\mathcal{W}_2^2(\mu_s^{n,\alpha},\mu_s^{\alpha}) \le \mathbb{E}\sup_{s\in[0,t]}\int_{\mathbb{R}^{2d}}|X_{n,\alpha}(u,s)-X_{\alpha}(v,s)|^2\chi(du,dv)$$

$$\le C\mathcal{W}_2^2(\mu_0^n,\mu_0)+C\int_0^t\mathbb{E}\sup_{r\in[0,s]}\mathcal{W}_2^2(\mu_r^{n,\alpha},\mu_r^{\alpha})ds.$$

For the control

$$\mathbb{E}\mathcal{W}_2^2(\mu_0^n,\mu_0)\le Cn^{-1}$$

see e.g. [Bolley, Guillin, Villani '07, in PTRF]

Law of large numbers behavior for $\alpha \to 0$

Theorem (Gess, Gvalani, K. 2022)

If μ^α is a superposition solution to

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

and $d\mu_t^0 = -\nabla \cdot (V(\cdot, \mu_t^0)\mu_t^0)dt$. Then

$$\mathbb{E} \sup_{t\in[0,T]} \mathcal{W}_2^2(\mu_t^\alpha, \mu_t^0) \le C\alpha.$$

Law of large numbers behavior for $\alpha \to 0$

Theorem (Gess, Gvalani, K. 2022)

If μ^α is a superposition solution to

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

and $d\mu_t^0 = -\nabla\cdot(V(\cdot,\mu_t^0)\mu_t^0)dt$. Then

$$\mathbb{E}\sup_{t\in[0,T]}\mathcal{W}_2^2(\mu_t^\alpha,\mu_t^0) \le C\alpha.$$

Corollary

$$\mathbb{E}\sup_{t\in[0,T]}\mathcal{W}_2^2(\mu_t^{n,\frac{1}{n}},\mu_t^0) \le Cn^{-1}$$

or formally $\mu_t^{n,\frac{1}{n}} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)} = \mu_t^0 + O(n^{-1/2})$.

Fluctuations around mean-field limit

Since $\mu_t^{n,\frac{1}{n}} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)} = \mu_t^0 + O(n^{-1/2})$, we consider

$$\eta_t^n = \sqrt{n}\left(\mu_t^{n,\frac{1}{n}} - \mu_t^0\right).$$

Assumptions

1. $V(x,\mu) = \bar{V}(x) + \langle \tilde{V}(x,\cdot), \mu\rangle$
2. For some $J \geq \frac{d}{2}+4$ one has that $\bar{V} \in \mathcal{C}_b^J(\mathbb{R}^d)$, $\tilde{V} \in \mathcal{C}^J(\mathbb{R}^d \times \mathbb{R}^d)$ and for every compact set $K \in \mathcal{P}_2(\mathbb{R}^d)$ and $i \in [d]$

$$\|\bar{V}_i\|_{\mathcal{C}^J} + \|\tilde{V}_i\|_{\mathcal{C}^J \times H^J} + \sup_{\mu\in K}\|A_{i,i}(\cdot,\mu)\|_{\mathcal{C}} < \infty.$$

where

$$\|f\|^2_{\mathcal{C}^m \times H^J} = \sum_{|\alpha|\leq m}\sum_{|\beta|\leq J}\sup_{x\in\mathbb{R}^d}\int_{\mathbb{R}^d}\left(D_x^\alpha D_y^\beta f(x,y)\right)^2 dy.$$

Quantified center limit theorem for SMFE

Theorem (Gess, Gvalani, K. 2022)

Let A, V, G be Lipschitz continuous and satisfy the assumptions above. There exists the Gaussian fluctuation field η, which is a solution to the linear SPDE

$$d\eta_t = -\nabla \cdot \left(V(\cdot, \mu_t^0)\eta_t + \langle \tilde{V}(\cdot, y), \eta_t(y) \rangle \mu_t^0 \right) dt \\ - \nabla \cdot \int_\Theta G(\cdot, \mu_t^0, \theta)\mu_t^0 W(d\theta, dt).$$

Moreover,

$$\mathbb{E} \sup_{t \in [0,T]} \|\eta_t^n - \eta_t\|_{H^{-J}}^2 \leq C n^{-1}.$$

Idea of proof. Equation for η^n

Remind that

$$d\mu_t^{n,\frac{1}{n}} = \frac{1}{2n}\nabla^2 : \left(A(\cdot,\mu_t^{n,\frac{1}{n}})\mu_t^{n,\frac{1}{n}}\right) dt - \nabla\cdot\left(V(\cdot,\mu_t^{n,\frac{1}{n}})\mu_t^{n,\frac{1}{n}}\right) dt \\ - \frac{1}{\sqrt{n}}\nabla\cdot\int_\Theta G(\cdot,\mu_t^{n,\frac{1}{n}},\theta)\mu_t^{n,\frac{1}{n}} W(d\theta,dt),$$

$$d\mu_t^0 = -\nabla\cdot\left(V(\cdot,\mu_t^0)\mu_t^0\right) dt$$

and

$$V(x,\mu) = \bar{V}(x) + \langle\tilde{V}(x,\cdot),\mu\rangle$$

Idea of proof. Equation for η^n

Remind that

$$d\mu_t^{n,\frac{1}{n}} = \frac{1}{2n}\nabla^2 : \left(A(\cdot,\mu_t^{n,\frac{1}{n}})\mu_t^{n,\frac{1}{n}}\right) dt - \nabla\cdot\left(V(\cdot,\mu_t^{n,\frac{1}{n}})\mu_t^{n,\frac{1}{n}}\right) dt - \frac{1}{\sqrt{n}}\nabla\cdot\int_\Theta G(\cdot,\mu_t^{n,\frac{1}{n}},\theta)\mu_t^{n,\frac{1}{n}} W(d\theta,dt),$$

$$d\mu_t^0 = -\nabla\cdot\left(V(\cdot,\mu_t^0)\mu_t^0\right) dt$$

and

$$V(x,\mu) = \bar{V}(x) + \langle\tilde{V}(x,\cdot),\mu\rangle$$

Then $\eta_t^n = \sqrt{n}\left(\mu_t^{n,\frac{1}{n}} - \mu_t^0\right)$ solves

$$d\eta_t^n = \frac{1}{2\sqrt{n}}\nabla^2 : \left(A(\cdot,\mu_t^{n,\frac{1}{n}})\mu_t^{n,\frac{1}{n}}\right) dt - \nabla\cdot\left(V(\cdot,\mu_t^{n,\frac{1}{n}})\eta_t^n + \left\langle\tilde{V}(\cdot,y),\eta_t^n(y)\right\rangle\mu_t^0\right) dt - \nabla\cdot\int_\Theta G(\cdot,\mu_t^{n,\frac{1}{n}},\theta)\mu_t^{n,\frac{1}{n}} W(d\theta,dt)$$

Idea of proof. Equation for η

$$d\eta_t^n = \frac{1}{2\sqrt{n}}\nabla^2 : \left(A(\cdot,\mu_t^{n,\frac{1}{n}})\mu_t^{n,\frac{1}{n}}\right)dt - \nabla\cdot\left(V(\cdot,\mu_t^{n,\frac{1}{n}})\eta_t^n + \left\langle \tilde{V}(\cdot,y),\eta_t^n(y)\right\rangle\mu_t^0\right)dt$$
$$- \nabla\cdot\int_\Theta G(\cdot,\mu_t^{n,\frac{1}{n}},\theta)\mu_t^{n,\frac{1}{n}}W(d\theta,dt)$$

Idea of proof. Equation for η

$$d\eta_t^n = \frac{1}{2\sqrt{n}}\nabla^2 : \left(A(\cdot, \mu_t^{n,\frac{1}{n}})\mu_t^{n,\frac{1}{n}}\right) dt - \nabla \cdot \left(V(\cdot, \mu_t^{n,\frac{1}{n}})\eta_t^n + \left\langle \tilde{V}(\cdot, y), \eta_t^n(y) \right\rangle \mu_t^0\right) dt$$
$$- \nabla \cdot \int_\Theta G(\cdot, \mu_t^{n,\frac{1}{n}}, \theta)\mu_t^{n,\frac{1}{n}} W(d\theta, dt)$$

Formally passing to the limit, we get

$$d\eta_t = 0dt - \nabla \cdot \left(V(\cdot, \mu_t^0)\eta_t + \left\langle \tilde{V}(\cdot, y), \eta_t(y) \right\rangle \mu_t^0\right) dt$$
$$- \nabla \cdot \int_\Theta G(\cdot, \mu_t^0, \theta)\mu_t^0 W(d\theta, dt)$$

Idea of proof. Norm of difference

$$d\eta_t^n = \frac{1}{2\sqrt{n}}\nabla^2 : \left(A(\cdot,\mu_t^{n,\frac{1}{n}})\mu_t^{n,\frac{1}{n}}\right)dt - \nabla\cdot\left(V(\cdot,\mu_t^{n,\frac{1}{n}})\eta_t^n + \left\langle \tilde{V}(\cdot,y),\eta_t^n(y)\right\rangle \mu_t^0\right)dt$$

$$- \nabla\cdot\int_\Theta G(\cdot,\mu_t^{n,\frac{1}{n}},\theta)\mu_t^{n,\frac{1}{n}} W(d\theta,dt)$$

$$d\eta_t = 0dt - \nabla\cdot\left(V(\cdot,\mu_t^0)\eta_t + \left\langle \tilde{V}(\cdot,y),\eta_t(y)\right\rangle \mu_t^0\right)dt$$

$$- \nabla\cdot\int_\Theta G(\cdot,\mu_t^0,\theta)\mu_t^0 W(d\theta,dt)$$

Idea of proof. Norm of difference

$$d\eta_t^n = \frac{1}{2\sqrt{n}}\nabla^2 : \left(A(\cdot,\mu_t^{n,\frac{1}{n}})\mu_t^{n,\frac{1}{n}}\right) dt - \nabla\cdot\left(V(\cdot,\mu_t^{n,\frac{1}{n}})\eta_t^n + \left\langle \tilde{V}(\cdot,y),\eta_t^n(y)\right\rangle \mu_t^0\right) dt$$
$$- \nabla\cdot\int_\Theta G(\cdot,\mu_t^{n,\frac{1}{n}},\theta)\mu_t^{n,\frac{1}{n}} W(d\theta,dt)$$
$$d\eta_t = 0dt - \nabla\cdot\left(V(\cdot,\mu_t^0)\eta_t + \left\langle \tilde{V}(\cdot,y),\eta_t(y)\right\rangle \mu_t^0\right) dt$$
$$- \nabla\cdot\int_\Theta G(\cdot,\mu_t^0,\theta)\mu_t^0 W(d\theta,dt)$$

Setting $\zeta_t^n = \eta_t^n - \eta_t$, and using Itô's formula, we get

$$d\|\zeta_t^n\|_{H^{-J}}^2 = \frac{1}{\sqrt{n}}\langle R(\mu_t^{n,\frac{1}{n}}),\zeta_t^n\rangle_{H^{-J}} dt + 2\langle Q(t,\eta_t^n,\mu_t^{n,\frac{1}{n}}) - Q(t,\eta_t^0,\mu_t^0),\zeta_t^n\rangle_{H^{-J}} dt$$
$$+ \|B(\mu_t^{n,\frac{1}{n}}) - B(\mu_t^0)\|_{\mathrm{HS},H^{-J}}^2 dt + \langle\zeta_t^n,(B(\mu_t^{n,\frac{1}{n}}) - B(\mu_t^0))dW_t\rangle_{H^{-J}}$$

Idea of proof. Estimate of the most problematic term

Estimate of a part of $\langle Q(t,\eta_t^n,\mu_t^{n,\frac{1}{n}}) - Q(t,\eta_t^0,\mu_t^0),\zeta_t^n\rangle_{H^{-J}}$

$$\langle \nabla\cdot(\bar{V}\eta_t^n) - \nabla\cdot(\bar{V}\eta_t^0),\zeta_t^n\rangle_{H^{-J}} = \langle \nabla\cdot(\bar{V}\zeta_t^n),\zeta_t^n\rangle_{H^{-J}} \leq \|\nabla\cdot(\bar{V}\zeta_t^n)\|_{H^{-J}}\|\zeta_t^n\|_{H^{-J}}$$

Idea of proof. Estimate of the most problematic term

Estimate of a part of $\langle Q(t,\eta_t^n,\mu_t^{n,\frac{1}{n}}) - Q(t,\eta_t^0,\mu_t^0),\zeta_t^n\rangle_{H^{-J}}$

$$\langle \nabla\cdot(\bar{V}\eta_t^n) - \nabla\cdot(\bar{V}\eta_t^0),\zeta_t^n\rangle_{H^{-J}} = \langle \nabla\cdot(\bar{V}\zeta_t^n),\zeta_t^n\rangle_{H^{-J}} \le \|\nabla\cdot(\bar{V}\zeta_t^n)\|_{H^{-J}}\|\zeta_t^n\|_{H^{-J}}$$

$$\begin{aligned}\|\nabla\cdot(\bar{V}\zeta_t^n)\|_{H^{-J}} &= \sup_{\varphi\in\mathcal{C}_0^\infty}\frac{1}{\|\varphi\|_J}\langle\nabla\varphi\cdot\bar{V},\zeta_t^n\rangle \\ &\le \sup_{\varphi\in\mathcal{C}_0^\infty}\frac{1}{\|\varphi\|_J}\|\nabla\varphi\|_J\|\bar{V}\|_{\mathcal{C}^J}\|\zeta_t^n\|_{H^{-J}} = +\infty\end{aligned}$$

Idea of proof. Estimate of the most problematic term

Estimate of a part of $\langle Q(t,\eta_t^n,\mu_t^{n,\frac{1}{n}}) - Q(t,\eta_t^0,\mu_t^0),\zeta_t^n\rangle_{H^{-J}}$

$$\langle \nabla\cdot(\bar V\eta_t^n) - \nabla\cdot(\bar V\eta_t^0),\zeta_t^n\rangle_{H^{-J}} = \langle \nabla\cdot(\bar V\zeta_t^n),\zeta_t^n\rangle_{H^{-J}} \le \|\nabla\cdot(\bar V\zeta_t^n)\|_{H^{-J}}\|\zeta_t^n\|_{H^{-J}}$$

$$\|\nabla\cdot(\bar V\zeta_t^n)\|_{H^{-J}} = \sup_{\varphi\in\mathcal{C}_0^\infty}\frac{1}{\|\varphi\|_J}\langle\nabla\varphi\cdot\bar V,\zeta_t^n\rangle \le \sup_{\varphi\in\mathcal{C}_0^\infty}\frac{1}{\|\varphi\|_J}\|\nabla\varphi\|_J\|\bar V\|_{\mathcal{C}^J}\|\zeta_t^n\|_{H^{-J}} = +\infty$$

Lemma

Let $v=(v_i)_{i\in[d]}\in\mathcal{C}_b^J(\mathbb{R}^d)$. Then the map $\mathcal{G}: H^{-J+1}\to H^{-J}$ defined by

$$\langle\varphi,\mathcal{G}(f)\rangle = \langle\nabla\varphi\cdot v, f\rangle$$

satisfies

$$|\langle\mathcal{G}(f),f\rangle_{H^{-J}}| \le C\|v\|_{\mathcal{C}^J}\|f\|_{H^{-J}}^2$$

Idea of proof of Lemma

For $f \in H^{-J+1} \subset H^{-J}$ there exists $\tilde{f} \in H^J$ suhc that

$$\langle \varphi, f \rangle = \langle \varphi, \tilde{f} \rangle_J.$$

Idea of proof of Lemma

For $f \in H^{-J+1} \subset H^{-J}$ there exists $\tilde{f} \in H^J$ suhc that

$$\langle \varphi, f \rangle = \langle \varphi, \tilde{f} \rangle_J.$$

Assume additionally $\tilde{f} \in \mathcal{C}_c^J(\mathbb{R}^d)$. Then

$$\langle \mathcal{G}(f), f \rangle_{H^{-J}} = \langle \tilde{f}, \mathcal{G}(f) \rangle = \langle \nabla \tilde{f} \cdot v, f \rangle = \langle \nabla \tilde{f} \cdot v, \tilde{f} \rangle_J$$

Idea of proof of Lemma

For $f \in H^{-J+1} \subset H^{-J}$ there exists $\tilde{f} \in H^J$ suhc that

$$\langle \varphi, f \rangle = \langle \varphi, \tilde{f} \rangle_J.$$

Assume additionally $\tilde{f} \in \mathcal{C}_c^J(\mathbb{R}^d)$. Then

$$\begin{aligned}\langle \mathcal{G}(f), f \rangle_{H^{-J}} &= \langle \tilde{f}, \mathcal{G}(f) \rangle = \langle \nabla \tilde{f} \cdot v, f \rangle = \langle \nabla \tilde{f} \cdot v, \tilde{f} \rangle_J \\ &= \dots + \sum_{|\beta|=J} \int_{\mathbb{R}^d} D^\beta \left(\partial_i \tilde{f} v_i \right) D^\beta \tilde{f} dx\end{aligned}$$

Idea of proof of Lemma

For $f \in H^{-J+1} \subset H^{-J}$ there exists $\tilde{f} \in H^J$ suhc that

$$\langle \varphi, f \rangle = \langle \varphi, \tilde{f} \rangle_J.$$

Assume additionally $\tilde{f} \in \mathcal{C}_c^J(\mathbb{R}^d)$. Then

$$\begin{aligned}
\langle \mathcal{G}(f), f \rangle_{H^{-J}} = \langle \tilde{f}, \mathcal{G}(f) \rangle = \langle \nabla \tilde{f} \cdot v, f \rangle &= \langle \nabla \tilde{f} \cdot v, \tilde{f} \rangle_J \\
&= \cdots + \sum_{|\beta|=J} \int_{\mathbb{R}^d} D^\beta \left(\partial_i \tilde{f} v_i \right) D^\beta \tilde{f} dx \\
&= \cdots + \sum_{|\beta|=J} \int_{\mathbb{R}^d} v_i D^\beta \left(\partial_i \tilde{f} \right) D^\beta \tilde{f} dx
\end{aligned}$$

Idea of proof of Lemma

For $f \in H^{-J+1} \subset H^{-J}$ there exists $\tilde{f} \in H^J$ suhc that

$$\langle \varphi, f \rangle = \langle \varphi, \tilde{f} \rangle_J.$$

Assume additionally $\tilde{f} \in \mathcal{C}_c^J(\mathbb{R}^d)$. Then

$$\begin{aligned}
\langle \mathcal{G}(f), f \rangle_{H^{-J}} &= \langle \tilde{f}, \mathcal{G}(f) \rangle = \langle \nabla \tilde{f} \cdot v, f \rangle = \langle \nabla \tilde{f} \cdot v, \tilde{f} \rangle_J \\
&= \cdots + \sum_{|\beta|=J} \int_{\mathbb{R}^d} D^\beta \left(\partial_i \tilde{f} v_i \right) D^\beta \tilde{f} dx \\
&= \cdots + \sum_{|\beta|=J} \int_{\mathbb{R}^d} v_i D^\beta \left(\partial_i \tilde{f} \right) D^\beta \tilde{f} dx \\
&= \cdots + \frac{1}{2} \sum_{|\beta|=J} \int_{\mathbb{R}^d} v_i \partial_i \left(D^\beta \tilde{f} \right)^2 dx
\end{aligned}$$

Idea of proof of Lemma

For $f \in H^{-J+1} \subset H^{-J}$ there exists $\tilde{f} \in H^J$ suhc that

$$\langle \varphi, f \rangle = \langle \varphi, \tilde{f} \rangle_J.$$

Assume additionally $\tilde{f} \in \mathcal{C}_c^J(\mathbb{R}^d)$. Then

$$\begin{aligned}
\langle \mathcal{G}(f), f \rangle_{H^{-J}} = \langle \tilde{f}, \mathcal{G}(f) \rangle = \langle \nabla \tilde{f} \cdot v, f \rangle &= \langle \nabla \tilde{f} \cdot v, \tilde{f} \rangle_J \\
&= \dots + \sum_{|\beta|=J} \int_{\mathbb{R}^d} D^\beta \left(\partial_i \tilde{f} v_i \right) D^\beta \tilde{f} dx \\
&= \dots + \sum_{|\beta|=J} \int_{\mathbb{R}^d} v_i D^\beta \left(\partial_i \tilde{f} \right) D^\beta \tilde{f} dx \\
&= \dots + \frac{1}{2} \sum_{|\beta|=J} \int_{\mathbb{R}^d} v_i \partial_i \left(D^\beta \tilde{f} \right)^2 dx \\
&= \dots - \frac{1}{2} \sum_{|\beta|=J} \int_{\mathbb{R}^d} \partial v_i \left(D^\beta \tilde{f} \right)^2 dx.
\end{aligned}$$

Higher order approximation of the SGD dynamics

The quantified CLT gives us that

$$\mu_t^{n,\frac{1}{n}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i(t)} = \mu_t^0 + n^{-1/2}\eta + O(n^{-1}).$$

Higher order approximation of the SGD dynamics

The quantified CLT gives us that

$$\mu_t^{n,\frac{1}{n}} = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i(t)} = \mu_t^0 + n^{-1/2}\eta + O(n^{-1}).$$

On the other hand, the empirical distribution of SGD with n parameters and learning rate $\alpha = \frac{1}{n}$ satisfies[2]

$$\hat{\mu}_t^{n,\frac{1}{n}} = \frac{1}{n}\sum_{i=1}^{n}\delta_{\hat{x}_i(\lfloor nt\rfloor)} = \mu_t^0 + n^{-1/2}\eta + o(n^{-1/2})$$

[2] see Sirignano, Spiliopoulos '20

Higher order approximation of the SGD dynamics

The quantified CLT gives us that

$$\mu_t^{n,\frac{1}{n}} = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i(t)} = \mu_t^0 + n^{-1/2}\eta + O(n^{-1}).$$

On the other hand, the empirical distribution of SGD with n parameters and learning rate $\alpha = \frac{1}{n}$ satisfies[2]

$$\hat{\mu}_t^{n,\frac{1}{n}} = \frac{1}{n}\sum_{i=1}^{n}\delta_{\hat{x}_i(\lfloor nt\rfloor)} = \mu_t^0 + n^{-1/2}\eta + o(n^{-1/2})$$

Therefore, $\hat{\mu}^{n,\frac{1}{n}} - \mu^{n,\frac{1}{n}} = o(n^{-1/2})$.

[2] see Sirignano, Spiliopoulos '20

Higher order approximation of the SGD dynamics

Theorem (Gess, Gvalani, K. 2022)

Let $\mu^{n,\frac{1}{n}}$ be a superposition solution to the SMFE with learning rate $\alpha = \frac{1}{n}$ started from $\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$. Let also $\hat{\mu}^{n,\frac{1}{n}}$ be the empirical process associated to the SGD with $\alpha = \frac{1}{n}$. Then

$$\mathcal{W}_p\left(\mathrm{Law}(\mu^{n,\frac{1}{n}}), \mathrm{Law}(\hat{\mu}^{n,\frac{1}{n}})\right) = o(n^{-1/2})$$

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

Idea of proof

$$\sqrt{n}\mathcal{W}_p\left(\mathrm{Law}(\mu^{n,\frac{1}{n}}), \mathrm{Law}(\hat{\mu}^{n,\frac{1}{n}})\right)$$

Idea of proof

$$\sqrt{n}\mathcal{W}_p\left(\mathrm{Law}(\mu^{n,\frac{1}{n}}), \mathrm{Law}(\hat{\mu}^{n,\frac{1}{n}})\right)$$
$$= \sqrt{n}\inf\left\{\mathbb{E}\sup_{t\in[0,T]}\|\nu_t^{n,\frac{1}{n}} - \hat{\nu}_t^{n,\frac{1}{n}}\|_{H^{-J}}^p, \quad \begin{array}{l} \nu^{n,\frac{1}{n}} \sim \mu^{n,\frac{1}{n}}, \\ \hat{\nu}^{n,\frac{1}{n}} \sim \hat{\mu}^{n,\frac{1}{n}} \end{array}\right\}^{1/p}$$

Idea of proof

$$
\sqrt{n}\mathcal{W}_p\left(\mathrm{Law}(\mu^{n,\frac{1}{n}}),\mathrm{Law}(\hat{\mu}^{n,\frac{1}{n}})\right)
$$

$$
= \sqrt{n}\inf\left\{\mathbb{E}\sup_{t\in[0,T]}\|\nu_t^{n,\frac{1}{n}}-\hat{\nu}_t^{n,\frac{1}{n}}\|_{H^{-J}}^p,\quad \begin{array}{l}\nu^{n,\frac{1}{n}}\sim\mu^{n,\frac{1}{n}},\\ \hat{\nu}^{n,\frac{1}{n}}\sim\hat{\mu}^{n,\frac{1}{n}}\end{array}\right\}^{1/p}
$$

$$
= \inf\left\{\mathbb{E}\sup_{t\in[0,T]}\|\sqrt{n}(\nu_t^{n,\frac{1}{n}}-\mu_t^0)-\sqrt{n}(\hat{\nu}_t^{n,\frac{1}{n}}-\mu_t^0)\|_{H^{-J}}^p,\quad \begin{array}{l}\nu^{n,\frac{1}{n}}\sim\mu^{n,\frac{1}{n}},\\ \hat{\nu}^{n,\frac{1}{n}}\sim\hat{\mu}^{n,\frac{1}{n}}\end{array}\right\}^{1/p}
$$

Idea of proof

$$\sqrt{n}\mathcal{W}_p\left(\mathrm{Law}(\mu^{n,\frac{1}{n}}),\mathrm{Law}(\hat{\mu}^{n,\frac{1}{n}})\right)$$

$$= \sqrt{n}\inf\left\{\mathbb{E}\sup_{t\in[0,T]}\|\nu_t^{n,\frac{1}{n}}-\hat{\nu}_t^{n,\frac{1}{n}}\|_{H^{-J}}^p,\quad \begin{array}{l}\nu^{n,\frac{1}{n}}\sim\mu^{n,\frac{1}{n}},\\ \hat{\nu}^{n,\frac{1}{n}}\sim\hat{\mu}^{n,\frac{1}{n}}\end{array}\right\}^{1/p}$$

$$= \inf\left\{\mathbb{E}\sup_{t\in[0,T]}\|\sqrt{n}(\nu_t^{n,\frac{1}{n}}-\mu_t^0)-\sqrt{n}(\hat{\nu}_t^{n,\frac{1}{n}}-\mu_t^0)\|_{H^{-J}}^p,\quad \begin{array}{l}\nu^{n,\frac{1}{n}}\sim\mu^{n,\frac{1}{n}},\\ \hat{\nu}^{n,\frac{1}{n}}\sim\hat{\mu}^{n,\frac{1}{n}}\end{array}\right\}^{1/p}$$

$$= \mathcal{W}_p\left(\mathrm{Law}(\eta^{n,\frac{1}{n}}),\mathrm{Law}(\hat{\eta}^{n,\frac{1}{n}})\right)$$

Idea of proof

$$
\begin{aligned}
&\sqrt{n}\mathcal{W}_p\left(\mathrm{Law}(\mu^{n,\frac{1}{n}}), \mathrm{Law}(\hat{\mu}^{n,\frac{1}{n}})\right)\\
&= \sqrt{n}\inf\left\{\mathbb{E}\sup_{t\in[0,T]}\|\nu_t^{n,\frac{1}{n}} - \hat{\nu}_t^{n,\frac{1}{n}}\|_{H^{-J}}^p, \quad \begin{array}{l}\nu^{n,\frac{1}{n}} \sim \mu^{n,\frac{1}{n}},\\ \hat{\nu}^{n,\frac{1}{n}} \sim \hat{\mu}^{n,\frac{1}{n}}\end{array}\right\}^{1/p}\\
&= \inf\left\{\mathbb{E}\sup_{t\in[0,T]}\|\sqrt{n}(\nu_t^{n,\frac{1}{n}} - \mu_t^0) - \sqrt{n}(\hat{\nu}_t^{n,\frac{1}{n}} - \mu_t^0)\|_{H^{-J}}^p, \quad \begin{array}{l}\nu^{n,\frac{1}{n}} \sim \mu^{n,\frac{1}{n}},\\ \hat{\nu}^{n,\frac{1}{n}} \sim \hat{\mu}^{n,\frac{1}{n}}\end{array}\right\}^{1/p}\\
&= \mathcal{W}_p\left(\mathrm{Law}(\eta^{n,\frac{1}{n}}), \mathrm{Law}(\hat{\eta}^{n,\frac{1}{n}})\right)\\
&\le \mathcal{W}_p\left(\mathrm{Law}(\eta^{n,\frac{1}{n}}), \mathrm{Law}(\eta)\right) + \mathcal{W}_p\left(\mathrm{Law}(\eta), \mathrm{Law}(\hat{\eta}^{n,\frac{1}{n}})\right)
\end{aligned}
$$

Idea of proof

$$
\begin{aligned}
&\sqrt{n}\mathcal{W}_p\left(\mathrm{Law}(\mu^{n,\frac{1}{n}}), \mathrm{Law}(\hat{\mu}^{n,\frac{1}{n}})\right)\\
&= \sqrt{n}\inf\left\{\mathbb{E}\sup_{t\in[0,T]}\|\nu_t^{n,\frac{1}{n}} - \hat{\nu}_t^{n,\frac{1}{n}}\|_{H^{-J}}^p,\quad \begin{array}{l}\nu^{n,\frac{1}{n}} \sim \mu^{n,\frac{1}{n}},\\ \hat{\nu}^{n,\frac{1}{n}} \sim \hat{\mu}^{n,\frac{1}{n}}\end{array}\right\}^{1/p}\\
&= \inf\left\{\mathbb{E}\sup_{t\in[0,T]}\|\sqrt{n}(\nu_t^{n,\frac{1}{n}} - \mu_t^0) - \sqrt{n}(\hat{\nu}_t^{n,\frac{1}{n}} - \mu_t^0)\|_{H^{-J}}^p,\quad \begin{array}{l}\nu^{n,\frac{1}{n}} \sim \mu^{n,\frac{1}{n}},\\ \hat{\nu}^{n,\frac{1}{n}} \sim \hat{\mu}^{n,\frac{1}{n}}\end{array}\right\}^{1/p}\\
&= \mathcal{W}_p\left(\mathrm{Law}(\eta^{n,\frac{1}{n}}), \mathrm{Law}(\hat{\eta}^{n,\frac{1}{n}})\right)\\
&\le \mathcal{W}_p\left(\mathrm{Law}(\eta^{n,\frac{1}{n}}), \mathrm{Law}(\eta)\right) + \mathcal{W}_p\left(\mathrm{Law}(\eta), \mathrm{Law}(\hat{\eta}^{n,\frac{1}{n}})\right)\\
&\le \left[\mathbb{E}\sup_{t\in[0,T]}\|\eta_t^{n,\frac{1}{n}} - \eta_t\|_{H^{-J}}^p\right]^{1/p} + \left[\mathbb{E}\sup_{t\in[0,T]}\|\hat{\eta}_t^{n,\frac{1}{n}} - \eta_t\|_{H^{-J}}^p\right]^{1/p} \to 0.
\end{aligned}
$$

Conclusion

Conclusion

The **Stochastic Mean-Field Equation** provides a higher order approximation to the SGD dynamics than the approximation by the non-fluctuation limit μ^0 which give the order $O(n^{-1/2})$.

Reference

Gess, Gvalani, Konarovskyi,
Conservative SPDEs as fluctuating mean field limits of stochastic gradient descent
(arXiv:2207.05705)

Thank you!