Conservative SPDEs as fluctuating mean field limits of stochastic gradient descent

Vitalii Konarovskyi

Bielefeld University

DMV Annual Meeting 2022 — Berlin

joint work with Benjamin Gess and Rishabh Gvalani

UNIVERSITÄT BIELEFELD

NATIONAL ACADEMY OF SCIENCES OF UKRAINE
INSTITUTE OF MATHEMATICS

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Note that $f_n(\theta) = \frac{1}{n}\sum_{k=1}^n U(\theta, x_k(t)) = \int_{\mathbb{R}^d} U(\theta,x)\mu_t^n(dx)$ **should approximate the true function** f **for large** t**.**

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

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

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

Related works

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

Well-posedness results for similar SPDEs:

- **Continuity equation in the fluid dynamics and optimal transportation** [Ambrosio, Trevisan, Crippa. . .]. There $A = G = 0$.

Related works

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

Well-posedness results for similar SPDEs:

- **Continuity equation in the fluid dynamics and optimal transportation** [Ambrosio, Trevisan, Crippa...]. There $A = G = 0$.
- **Stochastic nonlinear Fokker-Planck equation** [Coghi, Gess '19]. The covariance A has more general structure (i.e. $A(x,\mu) - \tilde{A}(x,x,\mu) \geq 0$) but the noise is finite-dimensional.

Related works

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

Well-posedness results for similar SPDEs:

- **Continuity equation in the fluid dynamics and optimal transportation** [Ambrosio, Trevisan, Crippa...]. There $A = G = 0$.
- **Stochastic nonlinear Fokker-Planck equation** [Coghi, Gess '19]. The covariance A has more general structure (i.e. $A(x,\mu) - \tilde{A}(x,x,\mu) \geq 0$) but the noise is finite-dimensional.
- **Strong superposition solutions of SDEs** [Flandoli '09]. Only the existence of solutions. The coefficients are independent of μ and the noise is additive (G does not depend on μ and x.)

Related works

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

Well-posedness results for similar SPDEs:

- **Continuity equation in the fluid dynamics and optimal transportation** [Ambrosio, Trevisan, Crippa...]. There $A = G = 0$.
- **Stochastic nonlinear Fokker-Planck equation** [Coghi, Gess '19]. The covariance A has more general structure (i.e. $A(x,\mu) - \tilde{A}(x,x,\mu) \geq 0$) but the noise is finite-dimensional.
- **Strong superposition solutions of SDEs** [Flandoli '09]. Only the existence of solutions. The coefficients are independent of μ and the noise is additive (G does not depend on μ and x.)
- **Particle representations for a class of nonlinear SPDEs** [Kurtz, Xiong '99]. The equation has more general form but the initial condition μ_0 must have an L_2-density w.r.t. the Lebesgue measure.

Related works

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

Well-posedness results for similar SPDEs:

- **Continuity equation in the fluid dynamics and optimal transportation** [Ambrosio, Trevisan, Crippa...]. There $A = G = 0$.
- **Stochastic nonlinear Fokker-Planck equation** [Coghi, Gess '19]. The covariance A has more general structure (i.e. $A(x,\mu) - \tilde{A}(x,x,\mu) \geq 0$) but the noise is finite-dimensional.
- **Strong superposition solutions of SDEs** [Flandoli '09]. Only the existence of solutions. The coefficients are independent of μ and the noise is additive (G does not depend on μ and x.)
- **Particle representations for a class of nonlinear SPDEs** [Kurtz, Xiong '99]. The equation has more general form but the initial condition μ_0 must have an L_2-density w.r.t. the Lebesgue measure.

The results from [Kurtz, Xiong] can be applied to our equation if μ_0 has L_2-density!

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 '95)

$$dX(u,t) = V(X(u,t),\bar{\mu}_t)dt + \int_\Theta G(X(u,t),\bar{\mu}_t,\theta)W(d\theta,dt),$$

$$X(u,0) = u, \quad \bar{\mu}_t = \mu_0 \circ X^{-1}(\cdot,t), \quad u \in \mathbb{R}^d, \quad t \geq 0.$$

SDE with interaction

The SMFE has a connection with the SDE with interaction (Kotelenez '95)

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

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.

SMFE and SDE with interaction

Lemma

Let X be a solution to the SDE with interaction with $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Then $\bar{\mu}_t = \mu_0 \circ X^{-1}(\cdot, t)$, $t \geq 0$, is a solution to the SMFE.

SMFE and SDE with interaction

Lemma

Let X be a solution to the SDE with interaction with $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Then $\bar{\mu}_t = \mu_0 \circ X^{-1}(\cdot, t)$, $t \geq 0$, is a solution to the SMFE.

Definition: We will say that $\bar{\mu}_t$, $t \geq 0$, is a superposition solution to the stochastic mean-field equation.

SMFE and SDE with interaction

Lemma

Let X be a solution to the SDE with interaction with $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Then $\bar{\mu}_t = \mu_0 \circ X^{-1}(\cdot, t)$, $t \geq 0$, is a solution to the SMFE.

Definition: We will say that $\bar{\mu}_t$, $t \geq 0$, is a superposition solution to the stochastic mean-field equation.

Corollary

Let V, G be Lipschitz continuous. 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 iff it has **only** superposition solutions.

Uniqueness of solutions to SMFE

- To prove the uniqueness, we show that every solution to the (nonlinear) SMFE is a superposition solution.

Uniqueness of solutions to SMFE

- To prove the uniqueness, we show that every solution to the (nonlinear) SMFE is a superposition solution.
- We first freeze the solution μ_t in the coefficients, considering the linear SPDE:

$$d\nu_t = \frac{1}{2}\nabla^2 : (a(t,\cdot)\nu_t)\, dt - \nabla \cdot (v(t,\cdot)\nu_t)\, dt \\ - \nabla \cdot \int_\Theta g(t,\cdot,\theta)\nu_t W(d\theta, dt),$$

where $a(t,x) = A(x,\mu_t)$, $v(t,x) = V(x,\mu_t)$ and $g(t,x,\theta) = G(x,\mu_t,\theta)$.

Uniqueness of solutions to SMFE

- To prove the uniqueness, we show that every solution to the (nonlinear) SMFE is a superposition solution.
- We first freeze the solution μ_t in the coefficients, considering the linear SPDE:

$$d\nu_t = \frac{1}{2}\nabla^2 : (a(t,\cdot)\nu_t)\,dt - \nabla\cdot(v(t,\cdot)\nu_t)\,dt - \nabla\cdot\int_\Theta g(t,\cdot,\theta)\nu_t W(d\theta,dt),$$

  where $a(t,x) = A(x,\mu_t)$, $v(t,x) = V(x,\mu_t)$ and $g(t,x,\theta) = G(x,\mu_t,\theta)$.
- We remove the second order term and the noise term from the linear SPDE by a (random) transformation of the space.

Random transformation of the space

We introduce the field of martingales

$$M(x,t) = \int_0^t g(s,x,\theta)W(d\theta, ds), \quad x \in \mathbb{R}^d, \quad t \geq 0.$$

and consider a solution $\psi_t(x) = (\psi_t^1(x), \ldots, \psi_t^d(x))$ to the stochastic transport equation

$$\psi_t^k(x) = x^k - \int_0^t \nabla\psi_s^k(x) \cdot M(x, \circ ds), \quad t \geq 0, \quad x \in \mathbb{R}^d, \quad k \in \{1, \ldots, d\}.$$

Random transformation of the space

We introduce the field of martingales

$$M(x,t) = \int_0^t g(s,x,\theta)W(d\theta, ds), \quad x \in \mathbb{R}^d, \quad t \geq 0.$$

and consider a solution $\psi_t(x) = (\psi_t^1(x), \dots, \psi_t^d(x))$ to the stochastic transport equation

$$\psi_t^k(x) = x^k - \int_0^t \nabla\psi_s^k(x) \cdot M(x, \circ ds), \quad t \geq 0, \quad x \in \mathbb{R}^d, \quad k \in \{1, \dots, d\}.$$

Lemma (see Kunita Stochastic flows and SDEs)

Under some smooth assumption on the coefficient g, the exists a field of diffeomorphisms $\psi(t,\cdot) : \mathbb{R}^d \to \mathbb{R}^d$, $t \geq 0$, which solves the stochastic transport equation.

Transformation of space

For the solution ν_t, $t \geq 0$, to the linear SPDE

$$d\nu_t = \frac{1}{2}\nabla^2 : (a(t,\cdot)\nu_t)\, dt - \nabla \cdot (v(t,\cdot)\nu_t)\, dt - \nabla \cdot \int_\Theta g(t,\cdot,\theta)\nu_t W(d\theta, dt),$$

we define

$$\rho_t = \nu_t \circ \psi_t^{-1}, \quad t \geq 0$$

Transformation of space

For the solution ν_t, $t \geq 0$, to the linear SPDE

$$d\nu_t = \frac{1}{2}\nabla^2 : (a(t,\cdot)\nu_t)\, dt - \nabla \cdot (v(t,\cdot)\nu_t)\, dt - \nabla \cdot \int_\Theta g(t,\cdot,\theta)\nu_t W(d\theta, dt),$$

we define

$$\rho_t = \nu_t \circ \psi_t^{-1}, \quad t \geq 0$$

Proposition

Let the coefficient g be smooth enough. Then ρ_t, $t \geq 0$, is a solution to the continuity equation[a]

$$d\rho_t = -\nabla(b(t,\cdot)\rho_t)dt, \quad \rho_0 = \nu_0 = \mu_0,$$

for some b depending on v and derivatives of a and ψ.

[a] Ambrosio, Lions, Trevisan,...

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

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

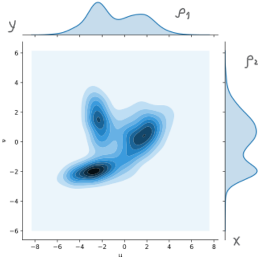
Wasserstein distance

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

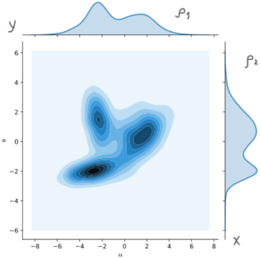
Wasserstein distance

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

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



Wikipedia

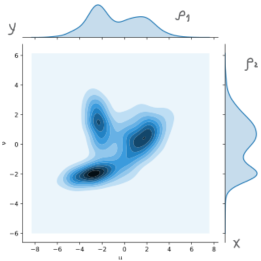
Wasserstein distance

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

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



Wikipedia

Proposition

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

Convergence of the empirical measure

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

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

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

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

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

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

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

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

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

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

Quantified central limit theorem for SMFE

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

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

Theorem (Gess, Gvalani, K. 2022)

There exists the Gaussian fluctuation field η, which is a solution to the linear SPDE

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

Moreover,

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

Higher order approximation of the SGD dynamics

The quantified CLT gives us that

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

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

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

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

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

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

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

$$\leq \left[\mathbb{E}\sup_{t\in[0,T]}\|\eta_t^{n,\frac{1}{n}}-\eta_t\|_{H^{-J}}^p\right]^{1/p}+\left[\mathbb{E}\sup_{t\in[0,T]}\|\hat{\eta}_t^{n,\frac{1}{n}}-\eta_t\|_{H^{-J}}^p\right]^{1/p}\to 0.$$

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!