

Conservative SPDEs as Fluctuating Mean Field Limits of Stochastic Gradient Descent

Vitalii Konarovskiy

Bielefeld University

SFB-Kolloquium — Potsdam

joint work with Benjamin Gess and Rishabh Gvalani



National Academy of Sciences of Ukraine
INSTITUTE OF MATHEMATICS

Table of Contents

- 1 Motivation and derivation of the SPDE
- 2 Quantified Mean-Field Limit
- 3 Well-posedness and superposition principle

Supervised Learning

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

Supervised Learning

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

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

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

Example: $\Phi(\theta, x_k) = c_k \cdot h(A_k \theta + b_k)$, $x_k = (A_k, b_k, c_k)$

Supervised Learning

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

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

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

Example: $\Phi(\theta, x_k) = c_k \cdot h(A_k \theta + b_k)$, $x_k = (A_k, b_k, c_k)$

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

$$\mathcal{L}(x) := \frac{1}{2} \mathbb{E}_{\vartheta} |f(\theta) - f_n(\theta; x)|^2 = \frac{1}{2} \int_{\Theta} |f(\theta) - f_n(\theta; x)|^2 \vartheta(d\theta),$$

where ϑ is the distribution of θ_i .

Stochastic gradient descent

Let $x_k(0) \sim \mu_0$ – i.i.d.

Stochastic gradient descent

Let $x_k(0) \sim \mu_0$ – i.i.d.

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

$$x_k(t_{i+1}) = x_k(t_i) - \nabla_{x_k} \left(\frac{1}{2} |f(\theta_i) - f_n(\theta_i; x)|^2 \right) \Delta t$$

where Δt – **learning rate**, $t_i = i\Delta t$, $\theta_i \sim \vartheta$ – i.i.d.,

Stochastic gradient descent

Let $x_k(0) \sim \mu_0$ – i.i.d.

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

$$\begin{aligned} x_k(t_{i+1}) &= x_k(t_i) - \nabla_{x_k} \left(\frac{1}{2} |f(\theta_i) - f_n(\theta_i; x)|^2 \right) \Delta t \\ &= x_k(t_i) - (f_n(\theta_i; x) - f(\theta_i)) \nabla_{x_k} \Phi(\theta_i, x_k(t_i)) \Delta t \end{aligned}$$

where Δt – **learning rate**, $t_i = i\Delta t$, $\theta_i \sim \vartheta$ – i.i.d.,

Stochastic gradient descent

Let $x_k(0) \sim \mu_0$ – i.i.d.

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

$$\begin{aligned}
 x_k(t_{i+1}) &= x_k(t_i) - \nabla_{x_k} \left(\frac{1}{2} |f(\theta_i) - f_n(\theta_i; x)|^2 \right) \Delta t \\
 &= x_k(t_i) - (f_n(\theta_i; x) - f(\theta_i)) \nabla_{x_k} \Phi(\theta_i, x_k(t_i)) \Delta t \\
 &= x_k(t_i) + \left(\nabla F(x_k(t_i), \theta_i) - \frac{1}{n} \sum_{l=1}^n \nabla_{x_k} K(x_k(t_i), x_l(t_i), \theta_i) \right) \Delta t
 \end{aligned}$$

where Δt – **learning rate**, $t_i = i\Delta t$, $\theta_i \sim \vartheta$ – i.i.d.,
 $F(x, \theta) = f(\theta)\Phi(\theta, x)$ and $K(x, y, \theta) = \Phi(\theta, x)\Phi(\theta, y)$.

Stochastic gradient descent

Let $x_k(0) \sim \mu_0$ – i.i.d.

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

$$\begin{aligned} x_k(t_{i+1}) &= x_k(t_i) - \nabla_{x_k} \left(\frac{1}{2} |f(\theta_i) - f_n(\theta_i; x)|^2 \right) \Delta t \\ &= x_k(t_i) - (f_n(\theta_i; x) - f(\theta_i)) \nabla_{x_k} \Phi(\theta_i, x_k(t_i)) \Delta t \\ &= x_k(t_i) + \left(\nabla F(x_k(t_i), \theta_i) - \langle \nabla_x K(x_k(t_i), \cdot, \theta_i), \nu_{t_i}^n \rangle \right) \Delta t \end{aligned}$$

where Δt – **learning rate**, $t_i = i\Delta t$, $\theta_i \sim \vartheta$ – i.i.d., $\nu_t^n = \frac{1}{n} \sum_{l=1}^n \delta_{x_l(t)}$,
 $F(x, \theta) = f(\theta)\Phi(\theta, x)$ and $K(x, y, \theta) = \Phi(\theta, x)\Phi(\theta, y)$.

Stochastic gradient descent

Let $x_k(0) \sim \mu_0$ – i.i.d.

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

$$\begin{aligned}
 x_k(t_{i+1}) &= x_k(t_i) - \nabla_{x_k} \left(\frac{1}{2} |f(\theta_i) - f_n(\theta_i; x)|^2 \right) \Delta t \\
 &= x_k(t_i) - (f_n(\theta_i; x) - f(\theta_i)) \nabla_{x_k} \Phi(\theta_i, x_k(t_i)) \Delta t \\
 &= x_k(t_i) + \left(\nabla F(x_k(t_i), \theta_i) - \langle \nabla_x K(x_k(t_i), \cdot, \theta_i), \nu_{t_i}^n \rangle \right) \Delta t \\
 &= x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i) \Delta t
 \end{aligned}$$

where Δt – **learning rate**, $t_i = i\Delta t$, $\theta_i \sim \vartheta$ – i.i.d., $\nu_t^n = \frac{1}{n} \sum_{l=1}^n \delta_{x_l(t)}$,
 $F(x, \theta) = f(\theta)\Phi(\theta, x)$ and $K(x, y, \theta) = \Phi(\theta, x)\Phi(\theta, y)$.

Continuous Dynamics of Parameters

Recall that $x_k(0) \sim \mu_0$ – i.i.d., Δt – learning rate, $t_i = i\Delta t$, $\theta_i \sim \vartheta$ – i.i.d.

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i)\Delta t, \quad k \in \{1, \dots, n\},$$

where $\nu_t^n = \frac{1}{n} \sum_{k=1}^n \delta_{x_k(t)}$.

Continuous Dynamics of Parameters

Recall that $x_k(0) \sim \mu_0$ – i.i.d., Δt – learning rate, $t_i = i\Delta t$, $\theta_i \sim \vartheta$ – i.i.d.

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i)\Delta t, \quad k \in \{1, \dots, n\},$$

where $\nu_{t_i}^n = \frac{1}{n} \sum_{k=1}^n \delta_{x_k(t)}$.

Considering the empirical distribution $\nu^n = \frac{1}{n} \sum_{k=1}^n \delta_{x_k}$, one has

$$f_n(\theta; x) = \frac{1}{n} \sum_{k=1}^n \Phi(\theta, x_k) = \langle \Phi(\theta, \cdot), \nu^n \rangle.$$

Continuous Dynamics of Parameters

Recall that $x_k(0) \sim \mu_0$ – i.i.d., Δt – learning rate, $t_i = i\Delta t$, $\theta_i \sim \vartheta$ – i.i.d.

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i)\Delta t, \quad k \in \{1, \dots, n\},$$

where $\nu_t^n = \frac{1}{n} \sum_{k=1}^n \delta_{x_k(t)}$.

Considering the empirical distribution $\nu^n = \frac{1}{n} \sum_{k=1}^n \delta_{x_k}$, one has

$$f_n(\theta; x) = \frac{1}{n} \sum_{k=1}^n \Phi(\theta, x_k) = \langle \Phi(\theta, \cdot), \nu^n \rangle.$$

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

$$dX_k(t) = V(X_k(t), \mu_t)dt,$$

$$\mu_t = \frac{1}{n} \sum_{k=1}^n \delta_{X_k(t)}, \quad V(x, \mu) = \mathbb{E}_\theta V(x, \mu, \theta).$$

Convergence to deterministic SPDE

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

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

where μ_t solves

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

with

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

and

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

[Mei, Montanari, Nguyen '18]

Convergence to deterministic SPDE

If $x_k(0) \sim \mu_0$ - i.i.d. and $\Delta t = \frac{1}{n}$, then

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

where μ_t solves

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

with

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

and

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

[Mei, Montanari, Nguyen '18]

\implies 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.

Goal: Propose an SPDE which would capture the fluctuations of the SGD dynamics and also would give its better approximation.

Classical SDE for SGD Dynamics

Stochastic gradient descent

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i) \Delta t$$

Classical SDE for SGD Dynamics

Stochastic gradient descent

$$\begin{aligned}x_k(t_{i+1}) &= x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i) \Delta t \\ &= x_k(t_i) + \mathbb{E}_\theta V(\dots) \Delta t + \sqrt{\Delta t} (V(\dots) - \mathbb{E}_\theta V(\dots)) \sqrt{\Delta t}\end{aligned}$$

Classical SDE for SGD Dynamics

Stochastic gradient descent

$$\begin{aligned}
 x_k(t_{i+1}) &= x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i) \Delta t \\
 &= x_k(t_i) + \underbrace{\mathbb{E}_\theta V(\dots)}_{=V(x_k(t_i), \nu_{t_i}^n)} \Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}} \underbrace{(V(\dots) - \mathbb{E}_\theta V(\dots))}_{=G(x_k(t_i), \nu_{t_i}^n, \theta_i)} \sqrt{\Delta t}
 \end{aligned}$$

Classical SDE for SGD Dynamics

Stochastic gradient descent

$$\begin{aligned}
 x_k(t_{i+1}) &= x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i) \Delta t \\
 &= x_k(t_i) + \underbrace{\mathbb{E}_\theta V(\dots)}_{=V(x_k(t_i), \nu_{t_i}^n)} \Delta t + \underbrace{\sqrt{\Delta t} (V(\dots) - \mathbb{E}_\theta V(\dots))}_{=\sqrt{\alpha} G(x_k(t_i), \nu_{t_i}^n, \theta_i)} \sqrt{\Delta t}
 \end{aligned}$$

is the Euler-Maruyama scheme for the SDE

$$dX_k(t) = V(X_k(t), \mu_t^n) dt + \sqrt{\alpha} (\Sigma^{\frac{1}{2}})_k(X(t)) dB(t), \quad k \in \{1, \dots, n\}$$

where $\mu_t^n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i(t)}$, $\Sigma_{k,l}(x) = \mathbb{E}_\theta G(x_k, \mu, \theta) \otimes G(x_l, \mu, \theta)$ and B – n -dim Brownian motion.

Classical SDE for SGD Dynamics

Stochastic gradient descent

$$\begin{aligned}
 x_k(t_{i+1}) &= x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i) \Delta t \\
 &= x_k(t_i) + \underbrace{\mathbb{E}_\theta V(\dots)}_{=V(x_k(t_i), \nu_{t_i}^n)} \Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}} \underbrace{(V(\dots) - \mathbb{E}_\theta V(\dots))}_{=G(x_k(t_i), \nu_{t_i}^n, \theta_i)} \sqrt{\Delta t}
 \end{aligned}$$

is the Euler-Maruyama scheme for the SDE

$$dX_k(t) = V(X_k(t), \mu_t^n) dt + \sqrt{\alpha} (\Sigma^{\frac{1}{2}})_k(X(t)) dB(t), \quad k \in \{1, \dots, n\}$$

where $\mu_t^n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i(t)}$, $\Sigma_{k,l}(x) = \mathbb{E}_\theta G(x_k, \mu, \theta) \otimes G(x_l, \mu, \theta)$ and B – n -dim Brownian motion.

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

SDE Driven by Inf-Dim Noise for SGD Dynamics

Stochastic gradient descent

$$\begin{aligned}
 x_k(t_{i+1}) &= x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i) \Delta t \\
 &= x_k(t_i) + \underbrace{\mathbb{E}_\theta V(\dots)}_{=V(x_k(t_i), \nu_{t_i}^n)} \Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}} \underbrace{(V(\dots) - \mathbb{E}_\theta V(\dots))}_{=G(x_k(t_i), \nu_{t_i}^n, \theta_i)} \sqrt{\Delta t}
 \end{aligned}$$

is the Euler-Maruyama scheme for the SDE

$$dX_k(t) = V(X_k(t), \mu_t^n) dt + \sqrt{\alpha} \int_{\Theta} G(X_k(t), \mu_t^n, \theta) W(d\theta, dt), \quad k \in \{1, \dots, n\}$$

where $\mu_t^n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i(t)}$, W – white noise on $L_2(\Theta, \vartheta)$.

[Gess, Kassing, K. '23]

Stochastic Mean-Field Equation

$$dX_k(t) = V(X_k(t), \mu_t^n)dt + \sqrt{\alpha} \int_{\Theta} G(X_k(t), \mu_t^n, \theta)W(d\theta, dt), \quad k \in \{1, \dots, n\}$$

where $\mu_t^n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i(t)}$, W – white noise on $L_2(\Theta, \vartheta)$.

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

Stochastic Mean-Field Equation

$$dX_k(t) = V(X_k(t), \mu_t^n)dt + \sqrt{\alpha} \int_{\Theta} G(X_k(t), \mu_t^n, \theta)W(d\theta, dt), \quad k \in \{1, \dots, n\}$$

where $\mu_t^n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i(t)}$, W – white noise on $L_2(\Theta, \vartheta)$.

Using Itô 's formula, we come to the **Stochastic Mean-Field Equation**:

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

Stochastic Mean-Field Equation

$$dX_k(t) = V(X_k(t), \mu_t^n)dt + \sqrt{\alpha} \int_{\Theta} G(X_k(t), \mu_t^n, \theta)W(d\theta, dt), \quad k \in \{1, \dots, n\}$$

where $\mu_t^n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i(t)}$, W – white noise on $L_2(\Theta, \vartheta)$.

Using Itô 's formula, we come to the **Stochastic Mean-Field Equation**:

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

where $A(x_k, \mu) = \mathbb{E}_{\theta} G(x_k, \mu) \otimes G(x_k, \mu)$.

Stochastic Mean-Field Equation

$$dX_k(t) = V(X_k(t), \mu_t^n)dt + \sqrt{\alpha} \int_{\Theta} G(X_k(t), \mu_t^n, \theta)W(d\theta, dt), \quad k \in \{1, \dots, n\}$$

where $\mu_t^n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i(t)}$, W – white noise on $L_2(\Theta, \vartheta)$.

Using Itô 's formula, we come to the **Stochastic Mean-Field Equation**:

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

where $A(x_k, \mu) = \mathbb{E}_{\theta} G(x_k, \mu) \otimes G(x_k, \mu)$.

The martingale problem for this equation was considered in [Rotskoff, Vanden-Eijnden, CPAM, '22]

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

Related Works

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

Well-posedness results for similar SPDEs:

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

Related Works

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

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. $A - \mathbb{E}G \otimes G \geq 0$) but the noise is **finite-dimensional**.

Related Works

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

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. $A - \mathbb{E}G \otimes G \geq 0$) but the noise is **finite-dimensional**.

Related Works

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

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. $A - \mathbb{E}G \otimes G \geq 0$) but the noise is **finite-dimensional**.
- **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

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

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. $A - \mathbb{E}G \otimes G \geq 0$) but the noise is **finite-dimensional**.
- **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!

Table of Contents

- 1 Motivation and derivation of the SPDE
- 2 Quantified Mean-Field Limit**
- 3 Well-posedness and superposition principle

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

$$\mathcal{W}_p^p(\rho_1, \rho_2) = \inf \left\{ \mathbb{E} d^p(\xi_1, \xi_2) : \xi_i \sim \rho_i \right\}$$

Higher Order Approximation of SGD

Stochastic Mean-Field Equation:

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

where $A(x_k, \mu) = \mathbb{E}_{\theta} G(x_k, \mu) \otimes G(x_k, \mu)$.

Higher Order Approximation of SGD

Stochastic Mean-Field Equation:

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

where $A(x_k, \mu) = \mathbb{E}_{\theta} G(x_k, \mu) \otimes G(x_k, \mu)$.

Theorem 1 (Gess, Gvalani, K. 2022)

- V, G – Lipschitz cont. and diff. w.r.t. the special variable with bdd deriv.;
- ν_t^n – the empirical process associated to the SGD dynamics with $\alpha = \frac{1}{n}$;
- μ_t^n – a (unique) solution to the SMFE started from

$$\mu_0^n = \nu_0^n = \frac{1}{n} \sum_{k=1}^n \delta_{x_k(0)}$$

with $x_k(0) \sim \mu_0$ i.i.d.

Then all $p \in [1, 2)$

$$\mathcal{W}_p(\text{Law } \mu^n, \text{Law } \nu^n) = o(n^{-1/2}).$$

Quantified Central Limit Theorem for SMFE

Theorem 2 (Gess, Gvalani, K. 2022)

Under the assumptions of the previous theorem, $\eta_t^n := \sqrt{n} (\mu_t^n - \mu_t^0) \rightarrow \eta_t$ where η_t is a Gaussian process solving

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

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

Quantified Central Limit Theorem for SMFE

Theorem 2 (Gess, Gvalani, K. 2022)

Under the assumptions of the previous theorem, $\eta_t^n := \sqrt{n}(\mu_t^n - \mu_t^0) \rightarrow \eta_t$ where η_t is a Gaussian process solving

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

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

Remark. [Sirignano, Spiliopoulos, '20]

For $\tilde{\eta}_t^n := \sqrt{n}(\nu_t^n - \mu_t^0)$

$$\mathbb{E} \sup_{t \in [0, T]} \|\tilde{\eta}_t^n\|_{-J}^2 \leq C \quad \text{and} \quad \tilde{\eta}^n \rightarrow \eta.$$

CLT for SMFE + CLT for SGD \implies Higher Order Approx.

Note that

$$\mu_t^n = \mu_t^0 + n^{-1/2}\eta + O(n^{-1}).$$

CLT for SMFE + CLT for SGD \implies Higher Order Approx.

Note that

$$\mu_t^n = \mu_t^0 + n^{-1/2}\eta + O(n^{-1}).$$

$$\nu_t^n = \mu_t^0 + n^{-1/2}\eta + o(n^{-1/2}).$$

CLT for SMFE + CLT for SGD \implies Higher Order Approx.

Note that

$$\begin{aligned}\mu_t^n &= \mu_t^0 + n^{-1/2}\eta + O(n^{-1}). \\ \nu_t^n &= \mu_t^0 + n^{-1/2}\eta + o(n^{-1/2}).\end{aligned}$$

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

CLT for SMFE + CLT for SGD \implies Higher Order Approx.

Note that

$$\begin{aligned}\mu_t^n &= \mu_t^0 + n^{-1/2}\eta + O(n^{-1}). \\ \nu_t^n &= \mu_t^0 + n^{-1/2}\eta + o(n^{-1/2}).\end{aligned}$$

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

$$\begin{aligned}\sqrt{n^p} \mathcal{W}_p^p(\text{Law}(\mu^n), \text{Law}(\nu^n)) &= \sqrt{n^p} \inf \mathbb{E} \left[\sup_{t \in [0, T]} \|\mu_t^n - \nu_t^n\|_{-J}^p \right] \\ &= \inf \mathbb{E} \left[\sup_{t \in [0, T]} \|\sqrt{n}(\mu_t^n - \mu_t^0) - \sqrt{n}(\nu_t^n - \mu_t^0)\|_{-J}^p \right] \\ &= \mathcal{W}_p^p(\text{Law}(\eta^n), \text{Law}(\tilde{\eta}^n)) \rightarrow 0.\end{aligned}$$

Table of Contents

- 1 Motivation and derivation of the SPDE
- 2 Quantified Mean-Field Limit
- 3 Well-posedness and superposition principle**

Continuity Equation

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

Continuity Equation

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

$$\implies \mu_t = \mu_0 \circ X(\cdot, t),$$

where

$$dX(u, t) = V(X(u, t))dt, \quad X(u, 0) = u.$$

[Ambrosio, Trevisan, Lions, ...]

Continuity Equation

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

$$\implies \mu_t = \mu_0 \circ X(\cdot, t),$$

where

$$dX(u, t) = V(X(u, t))dt, \quad X(u, 0) = u.$$

[Ambrosio, Trevisan, Lions, . . .]

The Stochastic Mean-Field Equation was derived from:

$$dX_k(t) = V(X_k(t), \mu_t^n)dt + \sqrt{\alpha} \int_{\Theta} G(X_k(t), \mu_t^n, \theta)W(d\theta, dt),$$

$$X_k(0) = x_k(0), \quad \mu_t^n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i(t)}.$$

Well-Posedness of SMFE

Theorem 3 (Gess, Gvalani, K. 2022)

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

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

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

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

SDE with Interaction

SDE with interaction:

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

SDE with Interaction

SDE with interaction:

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

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

Theorem (Kotelenez '95, Dorogovtsev' 07)

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

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

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

Definition of Solutions to SMFE

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

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. $\forall t \geq 0$

$$\begin{aligned} \langle \varphi, \mu_t \rangle &= \langle \varphi, \mu_0 \rangle + \int_0^t \langle \nabla \varphi \cdot V(\cdot, \mu_s), \mu_s \rangle ds + \frac{\alpha}{2} \int_0^t \langle \nabla^2 \varphi : A(\cdot, \mu_s), \mu_s \rangle ds \\ &\quad + \sqrt{\alpha} \int_0^t \int_{\Theta} \langle \nabla \varphi \cdot G(\cdot, \mu_s, \theta), \mu_s \rangle W(d\theta, ds) \end{aligned}$$

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 $\mu_t = \mu_0 \circ X^{-1}(\cdot, t)$, $t \geq 0$, 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 $\mu_t = \mu_0 \circ X^{-1}(\cdot, t)$, $t \geq 0$, is a solution to the SMFE.

Definition: We will say that μ_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 $\mu_t = \mu_0 \circ X^{-1}(\cdot, t)$, $t \geq 0$, is a solution to the SMFE.

Definition: We will say that μ_t , $t \geq 0$, is a **superposition solution** to the Stochastic Mean-Field equation.

Corollary

Let V, G be Lipschitz continuous. Then the SMFE

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

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:

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

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:

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

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 State Space

We introduce the field of martingales

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

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

$$\psi_t^k(x) = x^k - \int_0^t \nabla \psi_s^k(x) \cdot M(x, \circ ds).$$

Random Transformation of State Space

We introduce the field of martingales

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

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

$$\psi_t^k(x) = x^k - \int_0^t \nabla \psi_s^k(x) \cdot M(x, \circ ds).$$

Lemma (see Kunita Stochastic flows and SDEs)

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

Transformed SPDE

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

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

we define

$$\rho_t = \nu_t \circ \psi_t^{-1}.$$

Transformed SPDE

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

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

we define

$$\rho_t = \nu_t \circ \psi_t^{-1}.$$

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 .

^aAmbrosio, Lions, Trevisan, . . .

Comparison in Strong Topology

$x_k(0) \sim \mu_0$ – i.i.d., α – learning rate, $t_i = i\alpha$, $\theta_i \sim \vartheta$ – i.i.d.

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i)\alpha, \quad k \in \{1, \dots, n\},$$

where $\nu_t^n = \frac{1}{n} \sum_{k=1}^n \delta_{x_k(t)}$.

Comparison in Strong Topology

$x_k(0) \sim \mu_0$ – i.i.d., α – learning rate, $t_i = i\alpha$, $\theta_i \sim \vartheta$ – i.i.d.

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i)\alpha, \quad k \in \{1, \dots, n\},$$

where $\nu_t^n = \frac{1}{n} \sum_{k=1}^n \delta_{x_k(t)}$.

$$\begin{aligned} dX(u, t) &= V(X(u, t), \mu_t^n)dt \\ &\quad + \sqrt{\alpha} \int_{\Theta} G(X(u, t), \mu_t^n, \theta) W(d\theta, dt), \\ X(u, 0) &= u, \quad \mu_t^n = \nu_0^n \circ X^{-1}(\cdot, t), \end{aligned}$$

where W is a cylindrical Wiener process on $L_2(\Theta, P)$.

Comparison in Strong Topology

$x_k(0) \sim \mu_0$ – i.i.d., α – learning rate, $t_i = i\alpha$, $\theta_i \sim \vartheta$ – i.i.d.

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i)\alpha, \quad k \in \{1, \dots, n\},$$

where $\nu_t^n = \frac{1}{n} \sum_{k=1}^n \delta_{x_k(t)}$.

$$\begin{aligned} dX(u, t) &= V(X(u, t), \mu_t^n)dt \\ &\quad + \sqrt{\alpha} \int_{\Theta} G(X(u, t), \mu_t^n, \theta) W(d\theta, dt), \\ X(u, 0) &= u, \quad \mu_t^n = \nu_0^n \circ X^{-1}(\cdot, t), \end{aligned}$$

where W is a cylindrical Wiener process on $L_2(\Theta, P)$.

\implies For $\alpha = \frac{1}{n}$,

$$\mathcal{W}_p(\text{Law } \mu^n, \text{Law } \nu^n) = o(n^{-1/2}).$$

Comparison in Strong Topology

$x_k(0) \sim \mu_0$ – i.i.d., α – learning rate, $t_i = i\alpha$, $\theta_i \sim \vartheta$ – i.i.d.

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i)\alpha, \quad k \in \{1, \dots, n\},$$

where $\nu_t^n = \frac{1}{n} \sum_{k=1}^n \delta_{x_k(t)}$.

$$\begin{aligned} dX(u, t) &= V(X(u, t), \mu_t^n)dt \\ &\quad + \sqrt{\alpha} \int_{\Theta} G(X(u, t), \mu_t^n, \theta) W(d\theta, dt), \\ X(u, 0) &= u, \quad \mu_t^n = \nu_0^n \circ X^{-1}(\cdot, t), \end{aligned}$$

where W is a cylindrical Wiener process on $L_2(\Theta, P)$.

\implies For $\alpha = \frac{1}{n}$, +Quantified CLT for SGD

$$\mathcal{W}_p(\text{Law } \mu^n, \text{Law } \nu^n) = O(n^{-1}).$$

Comparison in Strong Topology

$x_k(0) \sim \mu_0$ – i.i.d., α – learning rate, $t_i = i\alpha$, $\theta_i \sim \vartheta$ – i.i.d.

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i)\alpha, \quad k \in \{1, \dots, n\},$$

where $\nu_t^n = \frac{1}{n} \sum_{k=1}^n \delta_{x_k(t)}$.

$$\begin{aligned} dX(u, t) &= V(X(u, t), \mu_t^n)dt \\ &\quad + \sqrt{\alpha} \int_{\Theta} G(X(u, t), \mu_t^n, \theta) W(d\theta, dt), \\ X(u, 0) &= u, \quad \mu_t^n = \nu_0^n \circ X^{-1}(\cdot, t), \end{aligned}$$

where W is a cylindrical Wiener process on $L_2(\Theta, P)$.

\implies For $\alpha = \frac{1}{n}$, +Quantified CLT for SGD

$$\mathcal{W}_p(\text{Law } \mu^n, \text{Law } \nu^n) = O(n^{-1}) = O(\alpha).$$

Comparison in Weak Topology

$x_k(0) \sim \mu_0$ – i.i.d., α – learning rate, $t_i = i\alpha$, $\theta_i \sim \vartheta$ – i.i.d.

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i)\alpha, \quad k \in \{1, \dots, n\},$$

where $\nu_t^n = \frac{1}{n} \sum_{k=1}^n \delta_{x_k(t)}$.

$$\begin{aligned} dX(u, t) &= V(X(u, t), \mu_t^n)dt \\ &\quad + \sqrt{\alpha} \int_{\Theta} G(X(u, t), \mu_t^n, \theta) W(d\theta, dt), \\ X(u, 0) &= u, \quad \mu_t^n = \nu_0^n \circ X^{-1}(\cdot, t), \end{aligned}$$

where W is a cylindrical Wiener process on $L_2(\Theta, P)$.

Comparison in Weak Topology

$x_k(0) \sim \mu_0$ – i.i.d., α – learning rate, $t_i = i\alpha$, $\theta_i \sim \vartheta$ – i.i.d.

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i)\alpha, \quad k \in \{1, \dots, n\},$$

where $\nu_t^n = \frac{1}{n} \sum_{k=1}^n \delta_{x_k(t)}$.

$$\begin{aligned} dX(u, t) &= V(X(u, t), \mu_t^n)dt \\ &\quad + \sqrt{\alpha} \int_{\Theta} G(X(u, t), \mu_t^n, \theta) W(d\theta, dt), \\ X(u, 0) &= u, \quad \mu_t^n = \nu_0^n \circ X^{-1}(\cdot, t), \end{aligned}$$

where W is a cylindrical Wiener process on $L_2(\Theta, P)$.

For $U \in \mathcal{C}^4(\mathcal{P}_2)$

$$\sup_{t \leq T} |\mathbb{E}U(\mu_t^n) - \mathbb{E}U(\nu_t^n)| =$$

like in [Li, Tai, E, JMLR, '19] for SGD dynamics.

Comparison in Weak Topology

$x_k(0) \sim \mu_0$ - i.i.d., α - learning rate, $t_i = i\alpha$, $\theta_i \sim \vartheta$ - i.i.d.

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i)\alpha, \quad k \in \{1, \dots, n\},$$

where $\nu_t^n = \frac{1}{n} \sum_{k=1}^n \delta_{x_k(t)}$.

$$\begin{aligned} dX(u, t) &= V(X(u, t), \mu_t^n)dt \\ &\quad + \sqrt{\alpha} \int_{\Theta} G(X(u, t), \mu_t^n, \theta) W(d\theta, dt), \\ X(u, 0) &= u, \quad \mu_t^n = \nu_0^n \circ X^{-1}(\cdot, t), \end{aligned}$$

where W is a cylindrical Wiener process on $L_2(\Theta, P)$.

For $U \in \mathcal{C}^4(\mathcal{P}_2)$, and $n \geq 1/\alpha^{4d}$

$$\sup_{t \leq T} |\mathbb{E}U(\mu_t^n) - \mathbb{E}U(\nu_t^n)| = O(\alpha)$$

like in [Li, Tai, E, JMLR, '19] for SGD dynamics.

Comparison in Weak Topology

$x_k(0) \sim \mu_0$ – i.i.d., α – learning rate, $t_i = i\alpha$, $\theta_i \sim \vartheta$ – i.i.d.

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n, \theta_i)\alpha, \quad k \in \{1, \dots, n\},$$

where $\nu_t^n = \frac{1}{n} \sum_{k=1}^n \delta_{x_k(t)}$.

$$dX(u, t) = V(X(u, t), \mu_t^n)dt - \frac{\alpha}{4} \nabla |V(X(u, t), \mu_t^n)|^2 dt - \frac{\alpha}{4} \langle D|V(X(u, t), \mu_t^n)|^2, \mu_t^n \rangle dt \\ + \sqrt{\alpha} \int_{\Theta} G(X(u, t), \mu_t^n, \theta) W(d\theta, dt),$$

$$X(u, 0) = u, \quad \mu_t^n = \nu_0^n \circ X^{-1}(\cdot, t),$$

where W is a cylindrical Wiener process on $L_2(\Theta, P)$.

For $U \in \mathcal{C}^4(\mathcal{P}_2)$, and $n \geq 1/\alpha^{4d}$

$$\sup_{t \leq T} |\mathbb{E}U(\mu_t^n) - \mathbb{E}U(\nu_t^n)| = O(\alpha^2)$$

like in [Li, Tai, E, JMLR, '19] for SGD dynamics.

See [Gess, Kassing, K. '23].

Reference



Gess, Gvalani, Konarovskiy,

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



Gess, Kassing, Konarovskiy,

Stochastic Modified Flows, Mean-Field Limits and Dynamics of Stochastic Gradient Descent
(arXiv:2302.07125)

Thank you!