Conservative SPDEs as Fluctuating Mean Field Limits of Stochastic Gradient Descent

Vitalii Konarovskyi

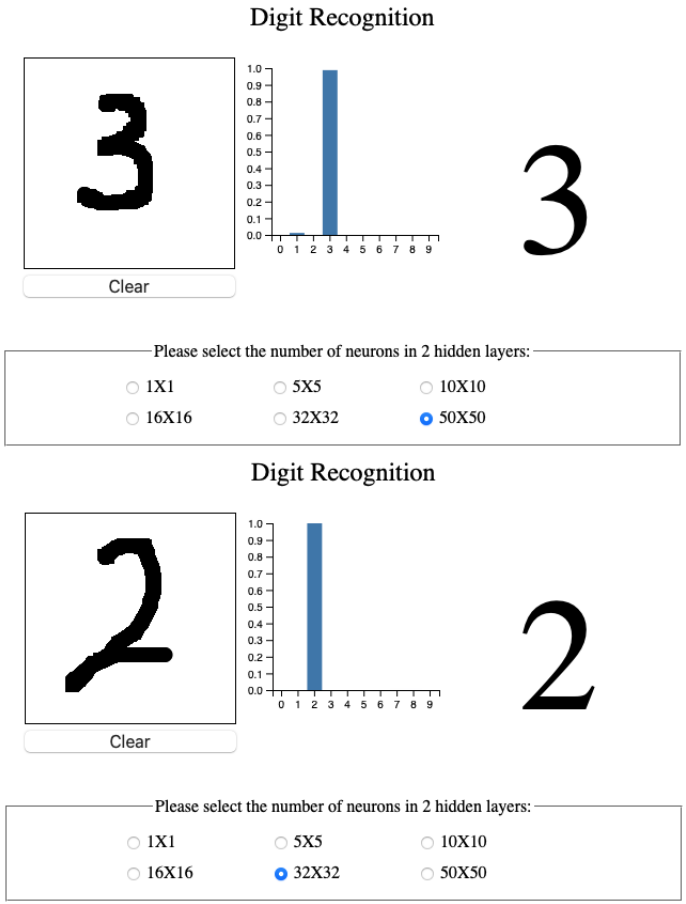
Hamburg University

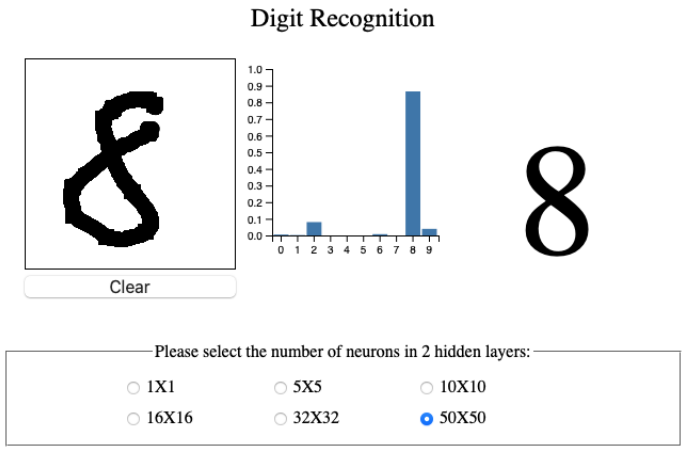
Mathematischen Kolloquium —TU Clausthal 2024

joint work with Benjamin Gess and Rishabh Gvalani

Table of Contents

Toy simulation: Recognizing hand-written digits





- MNIST database (used 15 000 data for training)
- Neural network with two hidden layers
- (stochastic) gradient descent
- accuracy – 93% on testing data, 68% – on hand-writing inputs

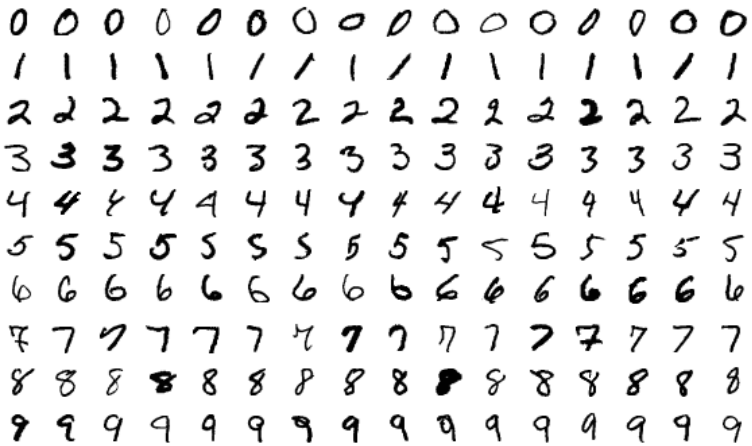
Simulation done by **Bohdan Tkachuk**
(student at Applied College of Yuriy Fedkovych Chernivtsi National University, Ukraine)

Supervised Learning

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

Supervised Learning

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



MNIST database

ξ_i – pictures (vector with coordinates coding every pixel)
γ_i – corresponding digit

Supervised Learning

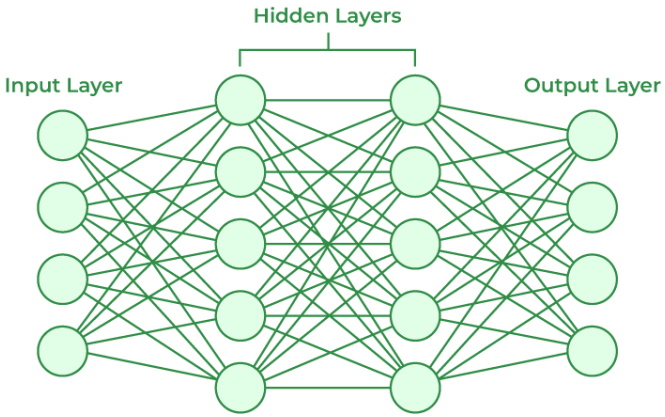
Give some data $\{(\xi_i, \gamma_i),\ i \in I\}$, the main goal of supervision learning is to predict a new γ given a new ξ.

MNIST database

ξ_i – pictures (vector with coordinates coding every pixel)
γ_i – corresponding digit

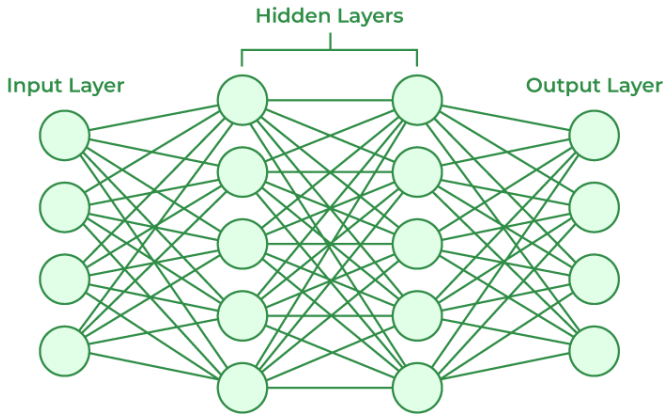
Assume that $\xi_i \sim P$ i.i.d. and $f(\xi_i) = \gamma_i$ (in general: γ_i may not be a deterministic function of ξ)

Neural Network



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

Neural Network

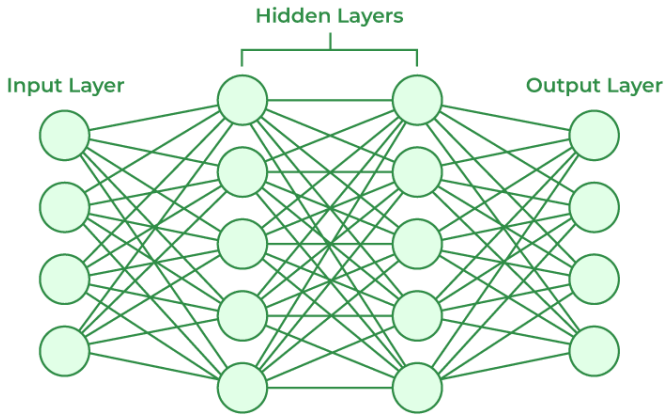


- $L \in \mathbb{N}$ – number of layers
- $d_1, \ldots, d_L$ – dimension of each layer
- σ – activation functions ($\sigma(x) = (1 + e^{-x})$, $\sigma(x) = \max(x, 0)$,...)

Motion of "signals" from layer to layer:

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

Neural Network



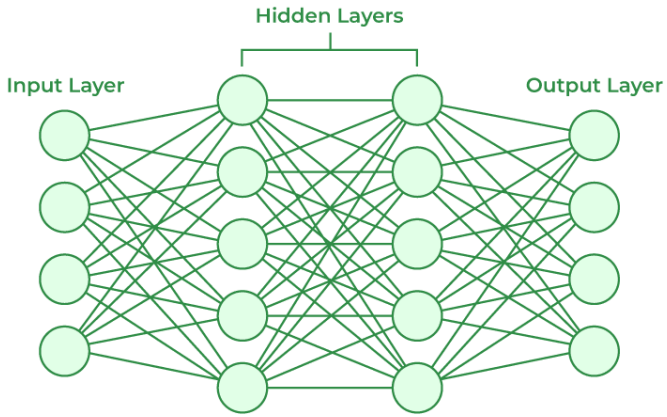
- $L \in \mathbb{N}$ – number of layers
- $d_1, \ldots, d_L$ – dimension of each layer
- σ – activation functions ($\sigma(x) = (1 + e^{-x})$, $\sigma(x) = \max(x, 0)$,...)

Motion of "signals" from layer to layer:

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

Output $\gamma = f(\xi)$ is approximated by $f(\xi; z)$ (in this example $z = (W_1, b_1, W_2, b_2, \ldots)$)

Neural Network



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

Motion of "signals" from layer to layer:

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

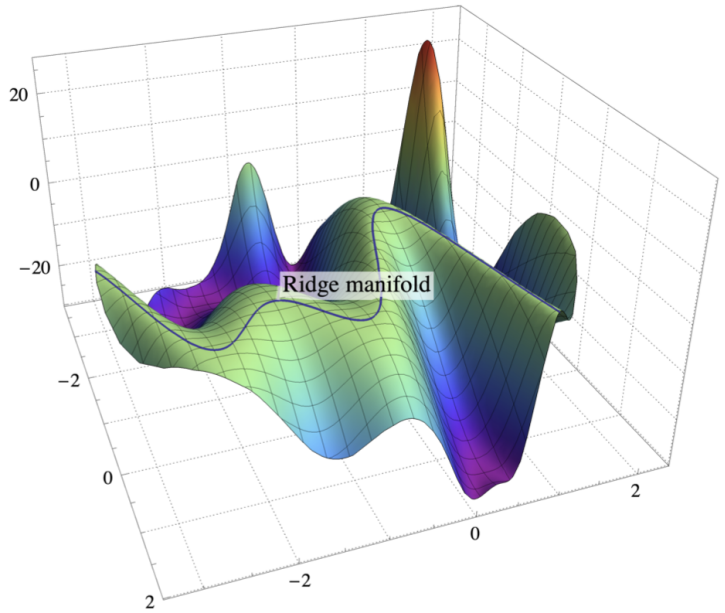
Output $\gamma = f(\xi)$ is approximated by $f(\xi; z)$ (in this example $z = (W_1, b_1, W_2, b_2, \ldots)$)
We measure the distance between f and $f(\cdot; z)$ by the **risk function**

$$R(z) := \mathbb{E}_P l(f(\xi), f(\xi; z))$$

Stochastic Gradient Descent

Set

$$\tilde{R}(z,\xi) := l(f(\xi), f(\xi; z)), \quad R(z) = \mathbb{E}_P \tilde{R}(z,\xi) \to \min$$

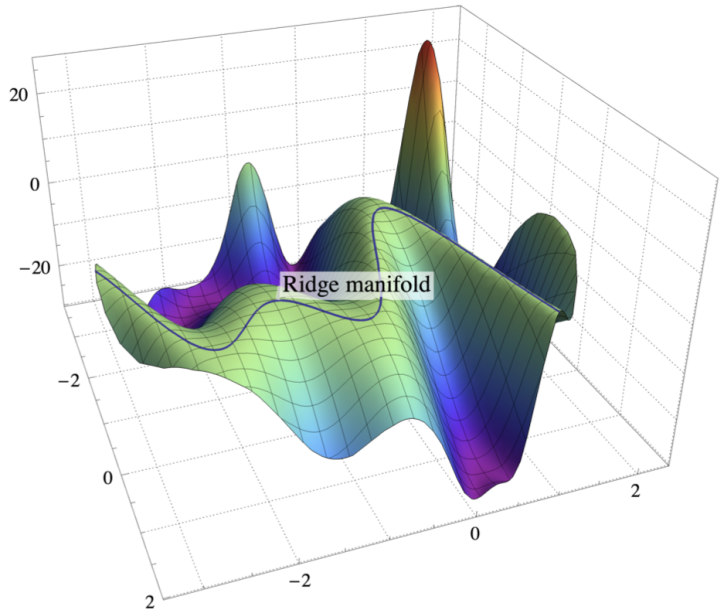


P. Mertikopoulos, N. Hallak, A. Kavis, V. Cevher '20

Stochastic Gradient Descent

Set

$$\tilde{R}(z,\xi) := l(f(\xi), f(\xi; z)), \quad R(z) = \mathbb{E}_P \tilde{R}(z,\xi) \to \min$$



P. Mertikopoulos, N. Hallak, A. Kavis, V. Cevher '20

Stochastic Gradient Descent: taking $z(0) \in \mathbb{R}^d$ define

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

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

Stochastic Differential Equation as Limiting Dynamics

Strategy: The systematic understanding of SGD dynamics has to rely on the identification of universal structures that are invariant to many degrees of freedoms (choice of loss function, architecture of network ...), while retaining the essential properties of SGD.

$$z(t_{i+1}) = z(t_i) - \nabla R(z(t_i), \xi_i)\Delta t$$

Stochastic Differential Equation as Limiting Dynamics

Strategy: The systematic understanding of SGD dynamics has to rely on the identification of universal structures that are invariant to many degrees of freedoms (choice of loss function, architecture of network ...), while retaining the essential properties of SGD.

$$
\begin{aligned}
z(t_{i+1}) &= z(t_i) - \nabla R(z(t_i), \xi_i)\Delta t \\
&= z(t_i) - \nabla \underbrace{\mathbb{E}_\xi R(\dots)}_{R(z(t_i))} \Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}} \underbrace{(\nabla \mathbb{E}_\xi R(\dots) - \nabla R(z(t_i), \xi_n))}_{=G(z(t_i),\xi_i)} \sqrt{\Delta t}
\end{aligned}
$$

Stochastic Differential Equation as Limiting Dynamics

Strategy: The systematic understanding of SGD dynamics has to rely on the identification of universal structures that are invariant to many degrees of freedoms (choice of loss function, architecture of network ...), while retaining the essential properties of SGD.

$$\begin{aligned} z(t_{i+1}) &= z(t_i) - \nabla R(z(t_i), \xi_i)\Delta t \\ &= z(t_i) - \nabla \underbrace{\mathbb{E}_\xi R(\dots)}_{R(z(t_i))} \Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}} \underbrace{(\nabla \mathbb{E}_\xi R(\dots) - \nabla R(z(t_i), \xi_n))}_{=G(z(t_i),\xi_i)} \sqrt{\Delta t} \end{aligned}$$

is the Euler scheme ($\Delta t \to 0$) for the SDE

$$dZ_t = -\nabla R(Z_t)dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(Z_t)dw_t,$$

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

Stochastic Differential Equation as Limiting Dynamics

Strategy: The systematic understanding of SGD dynamics has to rely on the identification of universal structures that are invariant to many degrees of freedoms (choice of loss function, architecture of network ...), while retaining the essential properties of SGD.

$$\begin{aligned} z(t_{i+1}) &= z(t_i) - \nabla R(z(t_i), \xi_i)\Delta t \\ &= z(t_i) - \nabla \underbrace{\mathbb{E}_\xi R(\dots)}_{R(z(t_i))} \Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}} \underbrace{(\nabla \mathbb{E}_\xi R(\dots) - \nabla R(z(t_i), \xi_n))}_{=G(z(t_i),\xi_i)} \sqrt{\Delta t} \end{aligned}$$

is the Euler scheme ($\Delta t \to 0$) for the SDE

$$dZ_t = -\nabla R(Z_t)dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(Z_t)dw_t,$$

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

[Li, Tai, E, Malladi, Arora, Wang, Kassing, Gess,...]

Our Goal

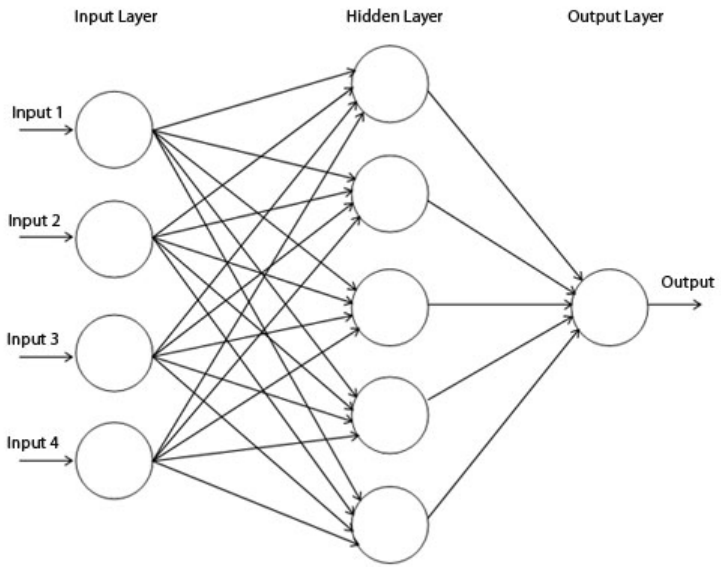
1. Learning rate goes to zero;

Our Goal

1. Learning rate goes to zero;
2. Number of neurons goes to infinity;

Our Goal

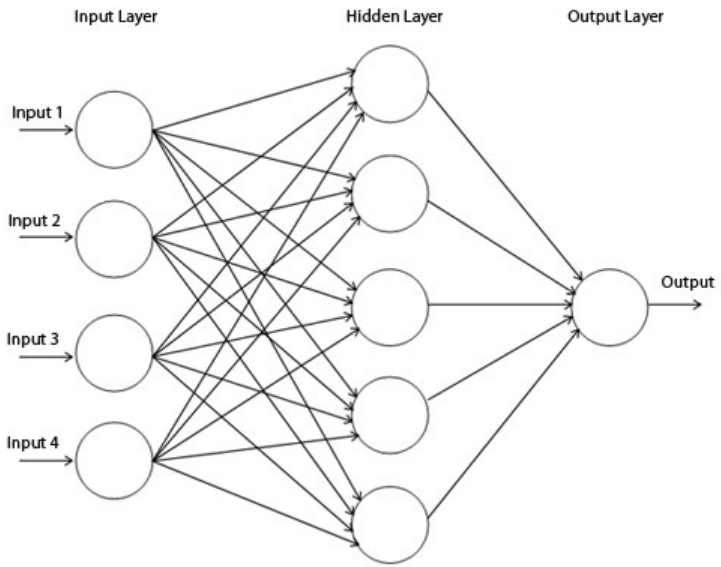
1. Learning rate goes to zero;
2. Number of neurons goes to infinity;
3. Neural network has only one hidden layer;



by Nicola Manzini

Our Goal

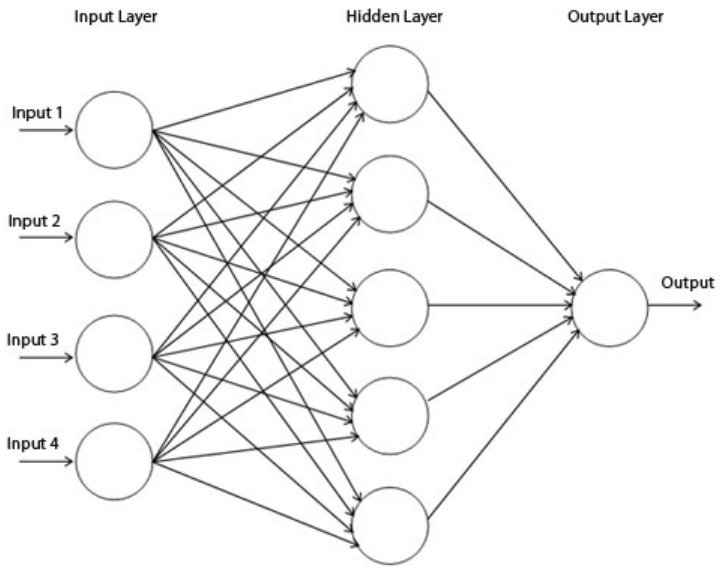
1. Learning rate goes to zero;
2. Number of neurons goes to infinity;
3. Neural network has only one hidden layer;



by Nicola Manzini

4. Propose a continuous model that also catches the fluctuations of SGD.

Neural network with one hidden layer



Network with a single hidden layer:

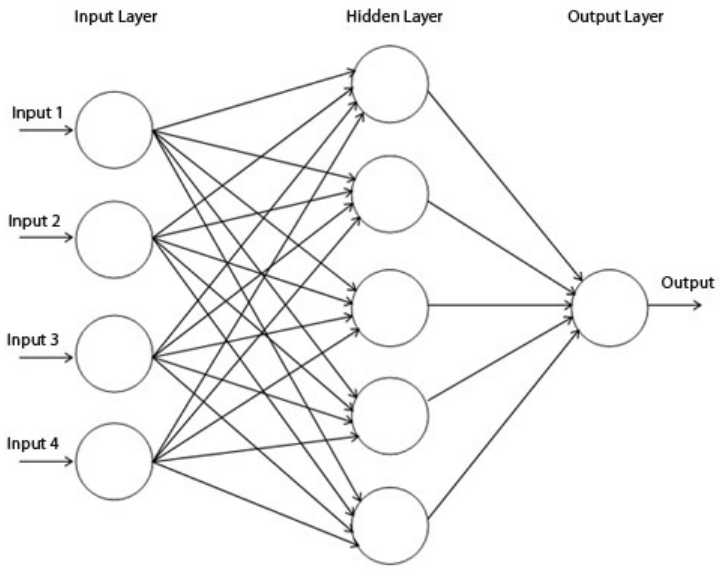
$$f_n(\xi; x) = \frac{1}{n}\sum_{k=1}^{n} \Phi(\xi, x_k)$$
$$= \langle \Phi(\xi, \cdot), \nu^n \rangle,$$

where $x_k \in \mathbb{R}^d$, $k \in \{1, \dots, n\}$, are parameters which have to be found, $\nu^n = \frac{1}{n}\sum_{k=1}^{n} \delta_{x_k}$

by Nicola Manzini

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

Neural network with one hidden layer



Network with a single hidden layer:

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

where $x_k \in \mathbb{R}^d$, $k \in \{1, \dots, n\}$, are parameters which have to be found, $\nu^n = \frac{1}{n}\sum_{k=1}^{n} \delta_{x_k}$

by Nicola Manzini

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

Generalization error

$$\mathcal{L}(x) := \frac{1}{2}\mathbb{E}_P |f(\xi) - f_n(\xi; x)|^2 = \frac{1}{2}\int_{\xi} |f(\xi) - f_n(\xi; x)|^2 P(d\xi),$$

where P is the distribution of ξ_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

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

where Δt – **learning rate**, $t_i = i\Delta t$, $\xi_i \sim P$ – 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

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

where Δt – **learning rate**, $t_i = i\Delta t$, $\xi_i \sim P$ – 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

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

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

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

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

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

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

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

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

Continuous Dynamics of Parameters

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

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

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

Continuous Dynamics of Parameters

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

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

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

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

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

Continuous Dynamics of Parameters

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

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

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

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

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

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

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

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

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

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

with

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

and

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

[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

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

with

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

and

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

[Mei, Montanari, Nguyen '18]

$\Longrightarrow$ 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

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i), \nu^n_{t_i}, \xi_i)\Delta t$$

Classical SDE for SGD Dynamics

Stochastic gradient descent

$$\begin{aligned} x_k(t_{i+1}) &= x_k(t_i) + V(x_k(t_i), \nu^n_{t_i}, \xi_i)\Delta t \\ &= x_k(t_i) + \mathbb{E}_\xi V(\dots)\Delta t + \sqrt{\Delta t}\,(V(\dots) - \mathbb{E}_\xi V(\dots))\sqrt{\Delta t} \end{aligned}$$

Classical SDE for SGD Dynamics

Stochastic gradient descent

$$
\begin{aligned}
x_k(t_{i+1}) &= x_k(t_i) + V(x_k(t_i), \nu^n_{t_i}, \xi_i)\Delta t \\
&= x_k(t_i) + \underbrace{\mathbb{E}_\xi V(\dots)}_{=V(x_k(t_i),\nu^n_{t_i})} \Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}} \underbrace{(V(\dots) - \mathbb{E}_\xi V(\dots))}_{=G(x_k(t_i),\nu^n_{t_i},\xi_i)} \sqrt{\Delta t}
\end{aligned}
$$

Classical SDE for SGD Dynamics

Stochastic gradient descent

$$\begin{aligned} x_k(t_{i+1}) &= x_k(t_i) + V(x_k(t_i), \nu^n_{t_i}, \xi_i)\Delta t \\ &= x_k(t_i) + \underbrace{\mathbb{E}_\xi V(\dots)}_{=V(x_k(t_i),\nu^n_{t_i})}\Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}}\underbrace{(V(\dots) - \mathbb{E}_\xi V(\dots))}_{=G(x_k(t_i),\nu^n_{t_i},\xi_i)}\sqrt{\Delta t} \end{aligned}$$

is the Euler-Maruyama scheme for the SDE

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

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

Classical SDE for SGD Dynamics

Stochastic gradient descent

$$\begin{aligned} x_k(t_{i+1}) &= x_k(t_i) + V(x_k(t_i), \nu^n_{t_i}, \xi_i)\Delta t \\ &= x_k(t_i) + \underbrace{\mathbb{E}_\xi V(\dots)}_{=V(x_k(t_i),\nu^n_{t_i})}\Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}}\underbrace{(V(\dots) - \mathbb{E}_\xi V(\dots))}_{=G(x_k(t_i),\nu^n_{t_i},\xi_i)}\sqrt{\Delta t} \end{aligned}$$

is the Euler-Maruyama scheme for the SDE

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

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

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

Martingale Problem for Empirical distribution

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

where $\mu_t^n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i(t)}$, $\Sigma_{k,l}(x) = \tilde{A}(x_k, x_l, \mu) := \mathbb{E}_\xi G(x_k, \mu, \xi) \otimes G(x_l, \mu, \xi)$

Martingale Problem for Empirical distribution

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

where $\mu_t^n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i(t)}$, $\Sigma_{k,l}(x) = \tilde{A}(x_k, x_l, \mu) := \mathbb{E}_\xi G(x_k, \mu, \xi) \otimes G(x_l, \mu, \xi)$

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

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

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

Martingale Problem for Empirical distribution

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

where $\mu_t^n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i(t)}$, $\Sigma_{k,l}(x) = \tilde{A}(x_k, x_l, \mu) := \mathbb{E}_\xi G(x_k, \mu, \xi) \otimes G(x_l, \mu, \xi)$

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

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

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

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

[Rotskoff, Vanden-Eijnden, CPAM, 2022]

SDE Driven by Inf-Dim Noise for SGD Dynamics

Stochastic gradient descent

$$\begin{aligned} x_k(t_{i+1}) &= x_k(t_i) + V(x_k(t_i), \nu^n_{t_i}, \xi_i)\Delta t \\ &= x_k(t_i) + \underbrace{\mathbb{E}_\xi V(\dots)}_{=V(x_k(t_i),\nu^n_{t_i})}\Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}}\underbrace{(V(\dots) - \mathbb{E}_\xi V(\dots))}_{=G(x_k(t_i),\nu^n_{t_i},\xi_i)}\sqrt{\Delta t} \end{aligned}$$

is the Euler-Maruyama scheme for the SDE

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

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

SDE Driven by Inf-Dim Noise for SGD Dynamics

Stochastic gradient descent

$$\begin{aligned} x_k(t_{i+1}) &= x_k(t_i) + V(x_k(t_i), \nu^n_{t_i}, \xi_i)\Delta t \\ &= x_k(t_i) + \underbrace{\mathbb{E}_\xi V(\dots)}_{=V(x_k(t_i),\nu^n_{t_i})}\Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}}\underbrace{(V(\dots) - \mathbb{E}_\xi V(\dots))}_{=G(x_k(t_i),\nu^n_{t_i},\xi_i)}\sqrt{\Delta t} \end{aligned}$$

is the Euler-Maruyama scheme for the SDE

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

where $\mu^n_t = \frac{1}{n}\sum_{i=1}^n \delta_{X_i(t)}$, W – white noise on $L_2(\xi, P)$ (P is the distribution of ξ).

[Gess, Kassing, K. '23]

Stochastic Mean-Field Equation

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

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

Stochastic Mean-Field Equation

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

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

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

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

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

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

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

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

Stochastic Mean-Field Equation

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

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

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

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

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

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

Related Works

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

Well-posedness results for similar SPDEs:

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

Related Works

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

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

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

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

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

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

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

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

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

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

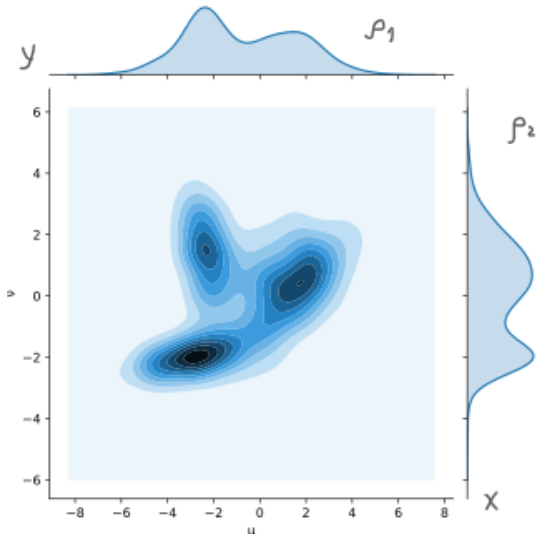
Wasserstein distance

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

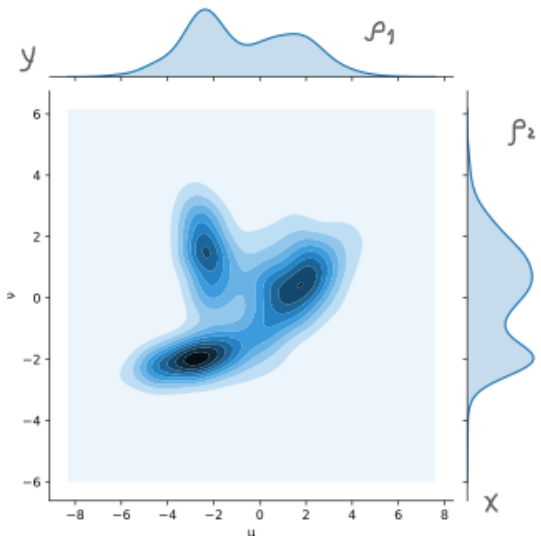
Let (E, d) be a Polish space, and for $p \geq 1$ $\mathcal{P}_p(E)$ be a space of all probability measures ρ on E with

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

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

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

$$= \inf\left\{\mathbb{E}d^p(\zeta_1, \zeta_2) : \quad \zeta_i \sim \rho_i\right\}$$



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

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

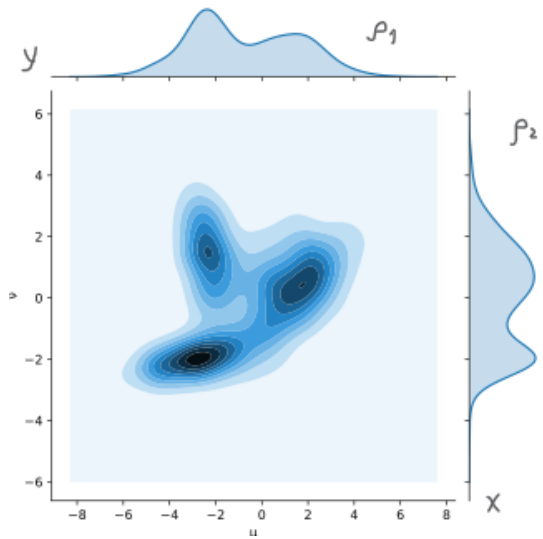
For $\rho_1, \rho_2 \in \mathcal{P}_p(E)$ we define the **Wasserstein distance** by

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

$$= \inf\left\{\mathbb{E}d^p(\zeta_1, \zeta_2) : \quad \zeta_i \sim \rho_i\right\}$$

Proposition

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



Wikipedia

Higher Order Approximation of SGD

Stochastic Mean-Field Equation:

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

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

Higher Order Approximation of SGD

Stochastic Mean-Field Equation:

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

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

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, $\eta_t^n := \sqrt{n}\left(\mu_t^n - \mu_t^0\right) \to \eta_t$ where η_t is a Gaussian process solving

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

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

Quantified Central Limit Theorem for SMFE

Theorem 2 (Gess, Gvalani, K. 2022)

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

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

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

Remark. [Sirignano, Spiliopoulos, '20]

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

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

CLT for SMFE + CLT for SGD $\Longrightarrow$ Higher Order Approx.

Note that

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

CLT for SMFE + CLT for SGD $\Longrightarrow$ 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 $\Longrightarrow$ Higher Order Approx.

Note that

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

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

CLT for SMFE + CLT for SGD $\Longrightarrow$ Higher Order Approx.

Note that

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

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

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

Table of Contents

Continuity Equation

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

Continuity Equation

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

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

where

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

[Ambrosio, Trevisan, Lions,. . .]

Continuity Equation

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

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

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:

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

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

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

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

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

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

$$X(u,0) = u, \quad u \in \mathbb{R}^d.$$

SDE with Interaction

SDE with interaction:

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

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

SDE with Interaction

SDE with interaction:

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

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

$X_t = X(\cdot,t)$ is a solution to the conditional McKean–Vlasov SDE

$$dX_t = V(X_t,\mathcal{L}_{X_t|W}) + \sqrt{\alpha}\int_\xi G(X_t,\mathcal{L}_{X_t|W},\xi)W(d\xi,dt), \quad \mathcal{L}_{X_0} = \mu_0$$

SDE with Interaction

SDE with interaction:

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

$X_t = X(\cdot,t)$ is a solution to the conditional McKean–Vlasov SDE

$$dX_t = V(X_t,\mathcal{L}_{X_t|W}) + \sqrt{\alpha}\int_\xi G(X_t,\mathcal{L}_{X_t|W},\xi)W(d\xi,dt), \quad \mathcal{L}_{X_0} = \mu_0$$

Theorem (Kotelenez '95, Dorogovtsev' 07, Wang '21)

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

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

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

Remark: We 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.

Remark: We 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

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

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:

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

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

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:

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

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

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

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

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

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, 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.

Transformed SPDE

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

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

we define

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

Transformed SPDE

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

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

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.

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

Reference

Gess, Gvalani, Konarovskyi,

Conservative SPDEs as fluctuating mean field limits of stochastic gradient descent

(arXiv:2207.05705)

Gess, Kassing, Konarovskyi,

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

Journal of Machine Learning Research 25 (2024) 1-27

Thank you!