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

Vitalii Konarovskyi

Hamburg University

Gradient Flows, Large Deviation Theory, and Macroscopic Fluctuation Theory — Bielefeld 2024

joint work with Benjamin Gess and Sebastian Kassing

Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

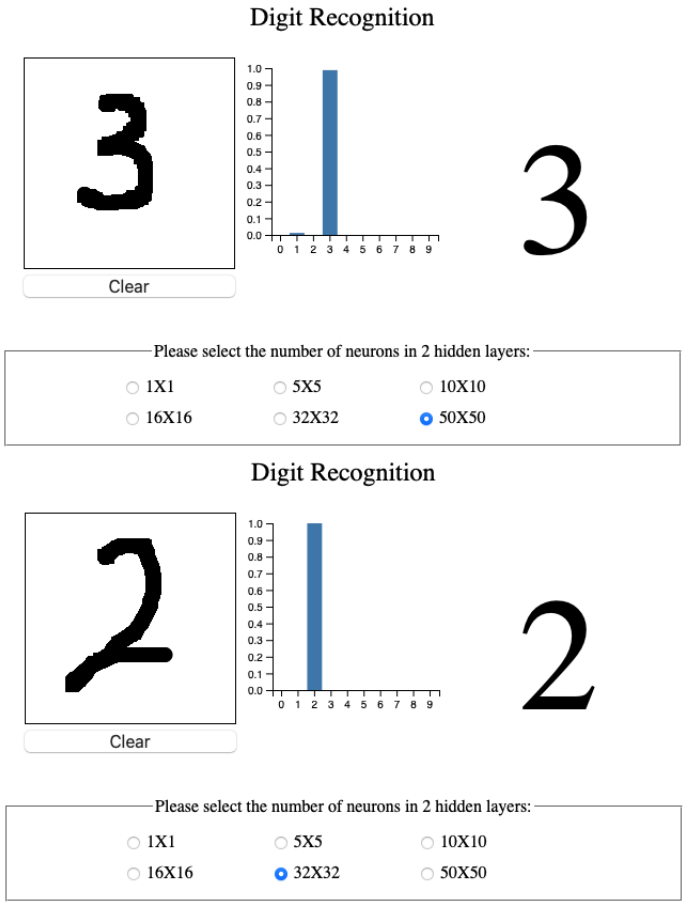
National Academy of Sciences of Ukraine

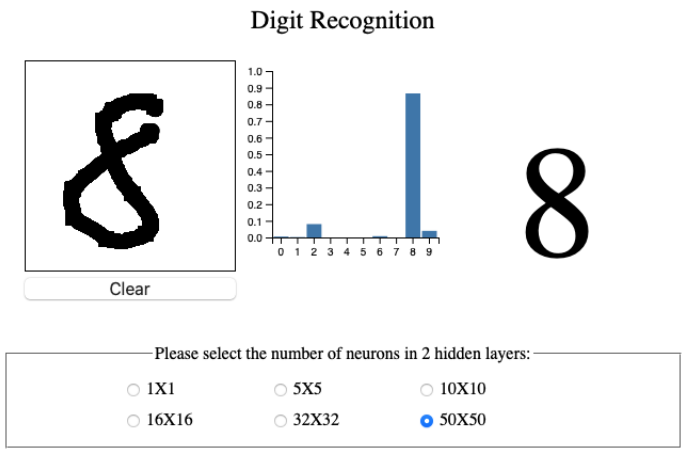
INSTITUTE OF MATHEMATICS

Table of Contents

1. Motivation: Stochastic Gradient Descent
2. Motivation: Overparameterized Stochastic Gradient Descent
3. Stochastic Modified Flow Driven by Inf.-dim Noise
4. Idea of Proof

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

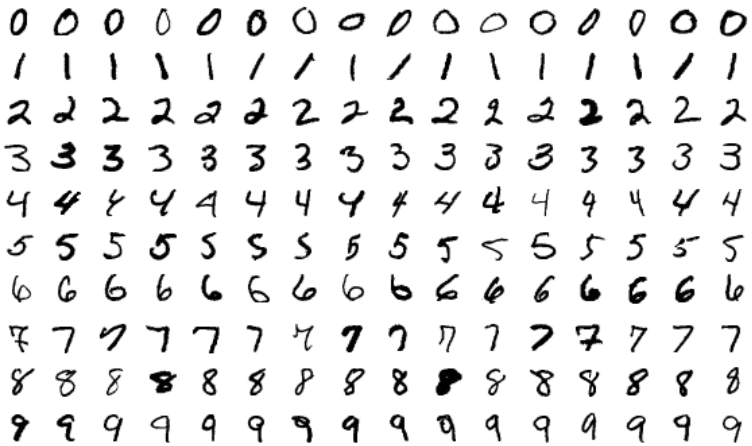
Simulation done by **Bohdan Tkachuk** Link: `http://54.72.31.237`
(student at Applied College of Yuriy Fedkovych Chernivtsi National University, Ukraine)

Supervised Learning

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

Supervised Learning

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

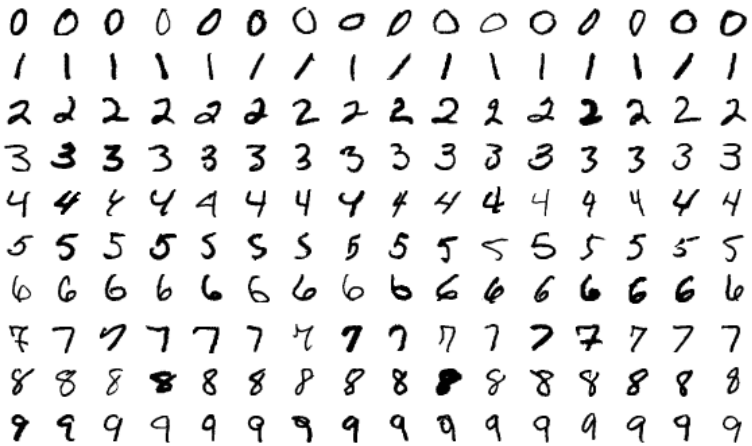


MNIST database

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

Supervised Learning

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

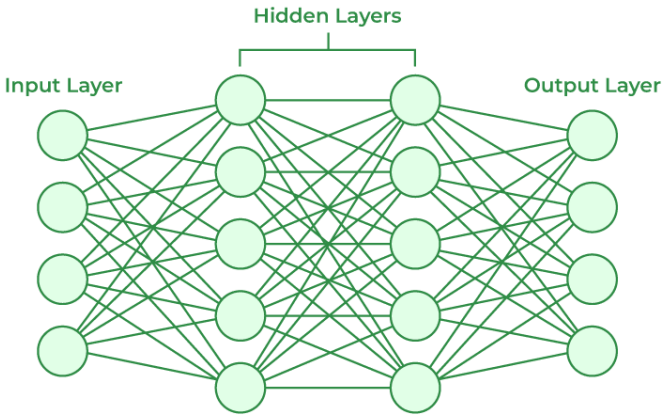


MNIST database

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

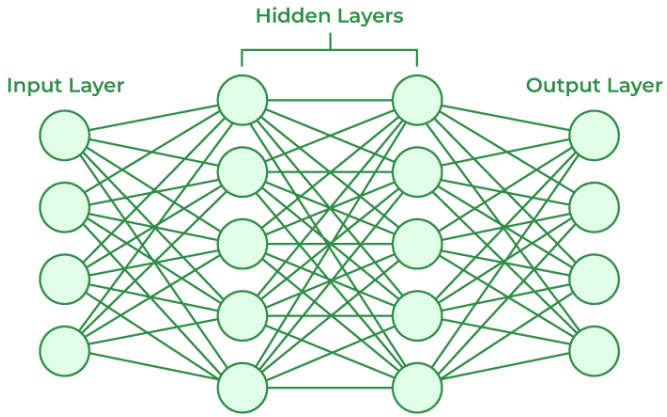
Assume that $\theta_i \sim P$ i.i.d. and $f(\theta_i) = \gamma_i$ (in general: γ_i may not be a deterministic function of θ)

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

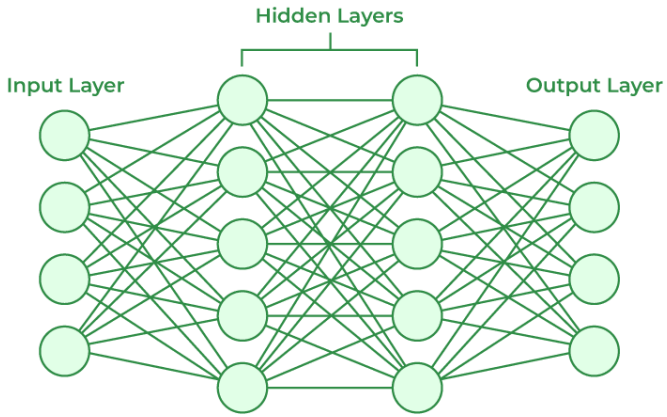


- $L \in \mathbb{N}$ – number of layers
- $d_1, \ldots, d_L$ – dimension of each layer
- σ – activation functions ($\sigma(x) = (1 + e^{-x})$, $\sigma(x) = \max(x, 0)$,...)

Motion of "signals" from layer to layer:

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

Neural Network



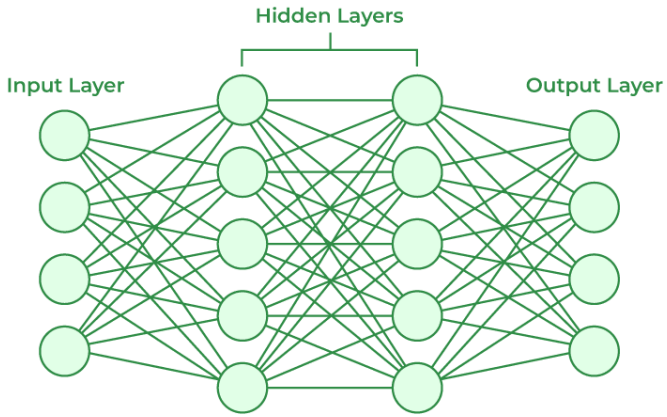
- $L \in \mathbb{N}$ – number of layers
- $d_1, \ldots, d_L$ – dimension of each layer
- σ – activation functions ($\sigma(x) = (1 + e^{-x})$, $\sigma(x) = \max(x, 0)$,...)

Motion of "signals" from layer to layer:

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

Output $\gamma = f(\theta)$ is approximated by $f(\theta; z)$ (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:

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

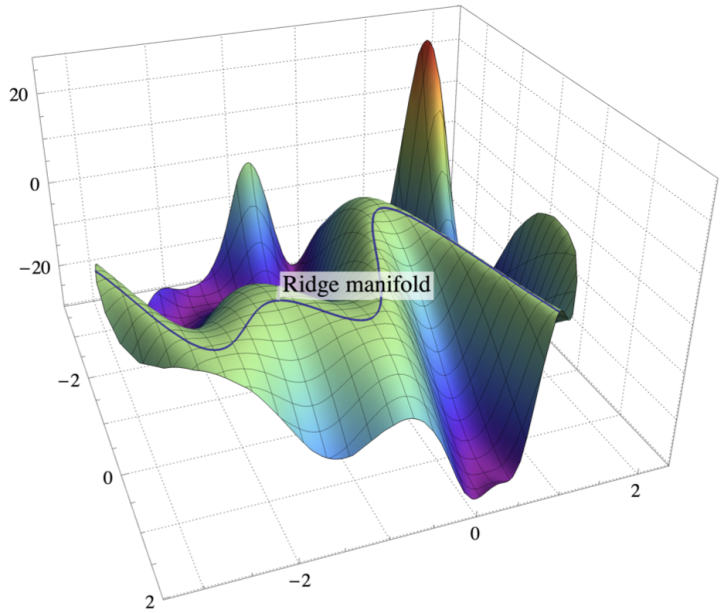
Output $\gamma = f(\theta)$ is approximated by $f(\theta; z)$ (in this example $z = (W_1, b_1, W_2, b_2, \ldots)$
We measure the distance between f and $f(\cdot; z)$ by the **risk function**

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

Stochastic Gradient Descent

Set

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

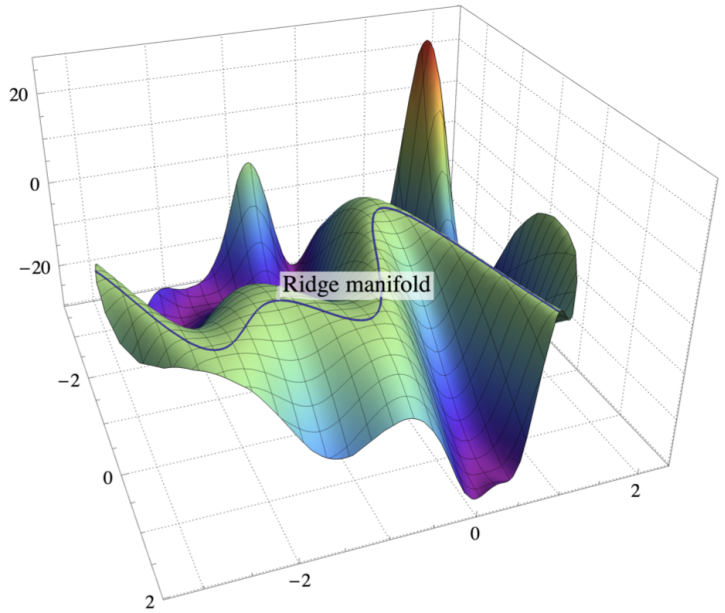


P. Mertikopoulos, N. Hallak, A. Kavis, V. Cevher '20

Stochastic Gradient Descent

Set

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



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

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

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

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

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

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.

$$
\begin{aligned}
z(t_{i+1}) &= z(t_i) - \nabla R(z(t_i), \theta_i)\Delta t \\
&= z(t_i) - \nabla \underbrace{\mathbb{E}_\theta R(\ldots)}_{R(z(t_i))} \Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}} \underbrace{\left(\nabla \mathbb{E}_\theta R(\ldots) - \nabla R(z(t_i), \theta_n)\right)}_{=G(z(t_i),\theta_i)} \sqrt{\Delta t}
\end{aligned}
$$

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.

$$
\begin{aligned}
z(t_{i+1}) &= z(t_i) - \nabla R(z(t_i), \theta_i)\Delta t \\
&= z(t_i) - \nabla \underbrace{\mathbb{E}_\theta R(\ldots)}_{R(z(t_i))} \Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}} \underbrace{(\nabla \mathbb{E}_\theta R(\ldots) - \nabla R(z(t_i), \theta_n))}_{=G(z(t_i),\theta_i)} \sqrt{\Delta t}
\end{aligned}
$$

is the Euler scheme for the SDE

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

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

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.

$$\begin{aligned} z(t_{i+1}) &= z(t_i) - \nabla R(z(t_i), \theta_i)\Delta t \\ &= z(t_i) - \nabla \underbrace{\mathbb{E}_\theta R(\ldots)}_{R(z(t_i))} \Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}} \underbrace{(\nabla \mathbb{E}_\theta R(\ldots) - \nabla R(z(t_i), \theta_n))}_{=G(z(t_i),\theta_i)} \sqrt{\Delta t} \end{aligned}$$

is the Euler scheme for the SDE

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

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

Theorem (Li, Tai, E '19, JMLR)

For f, R and $\Sigma^{\frac{1}{2}}$ smooth enough with bounded derivatives one has

$$\sup_{t_i \leq T} |\mathbb{E} f(z(t_i)) - \mathbb{E} f(Z_{t_i})| = O(\alpha).$$

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.

$$\begin{aligned} z(t_{i+1}) &= z(t_i) - \nabla R(z(t_i), \theta_i)\Delta t \\ &= z(t_i) - \nabla \underbrace{\mathbb{E}_\theta R(\ldots)}_{R(z(t_i))} \Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}} \underbrace{(\nabla \mathbb{E}_\theta R(\ldots) - \nabla R(z(t_i), \theta_n))}_{=G(z(t_i),\theta_i)} \sqrt{\Delta t} \end{aligned}$$

is the Euler scheme for the SDE

$$dZ_t = -\nabla R(Z_t)dt - \frac{\alpha}{4}\nabla|\nabla R(Z_t)|^2 dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(Z_t)dw_t,$$

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

Theorem (Li, Tai, E '19, JMLR)

For f, R and $\Sigma^{\frac{1}{2}}$ smooth enough with bounded derivatives one has

$$\sup_{t_i \leq T} |\mathbb{E} f(z(t_i)) - \mathbb{E} f(Z_{t_i})| = O(\alpha^2).$$

Some Limitation of Modified SDE

1. Limited regularity of $\Sigma^{\frac{1}{2}}$:

Ex. $\Sigma(z) = z^2 \quad \Longrightarrow \quad \Sigma^{\frac{1}{2}}(z) = |z|.$

Some Limitation of Modified SDE

1. Limited regularity of $\Sigma^{\frac{1}{2}}$:

Ex. $\Sigma(z) = z^2 \quad \Longrightarrow \quad \Sigma^{\frac{1}{2}}(z) = |z|$.

2. The SDE does not catch n-point motion:

Let $z_k(t_i)$ be the SGD dynamics started from $z_k(0)$

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

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

Then

$$(z_1(t_i), \ldots, z_n(t_i)) \not\approx (Z^1_{t_i}, \ldots, Z^n_{t_i}).$$

for solutions of the Modified SDE started from $Z^k_0 = z_k(0)$.

(e.g. important for a dynamical systems approach [Wu et al. '18; Sato et al. '22])

Table of Contents

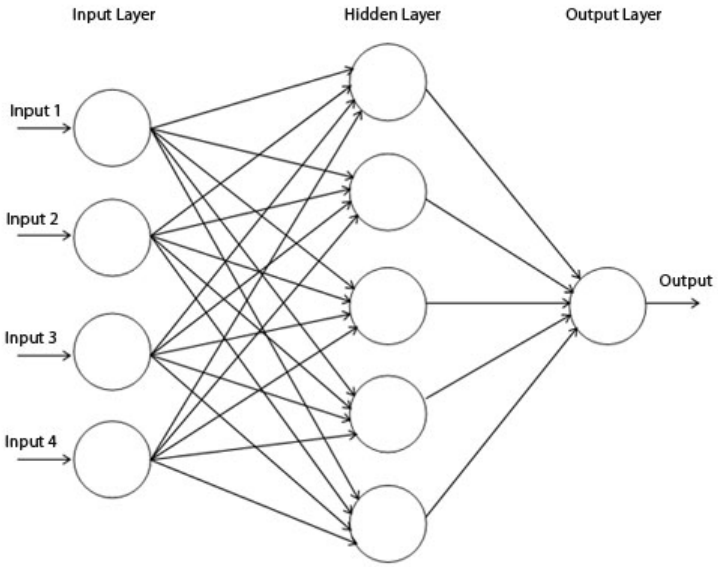
1 Motivation: Stochastic Gradient Descent

2 Motivation: Overparameterized Stochastic Gradient Descent

3 Stochastic Modified Flow Driven by Inf.-dim Noise

4 Idea of Proof

Neural network with one hidden layer



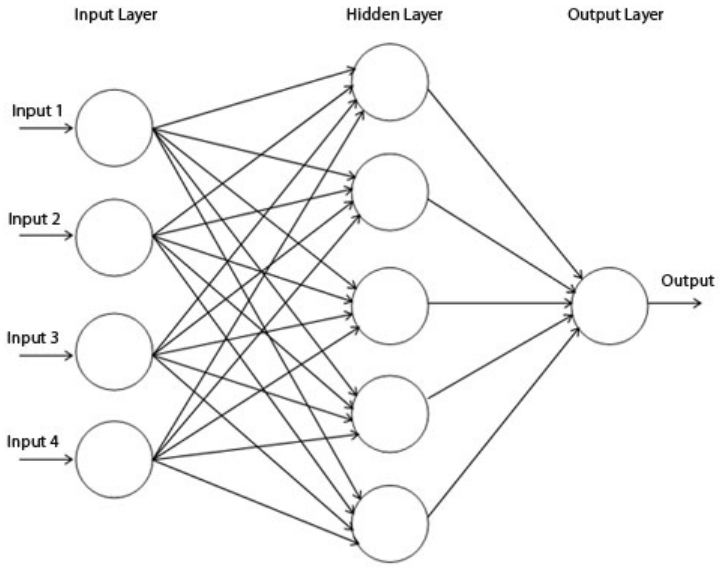
Network with a single hidden layer:

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

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:

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

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

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

where P 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**, $t_i = i\Delta t$, $\theta_i \sim P$ – 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

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

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

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

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

Classical SDE for Overparametrized SGD Dynamics

Stochastic gradient descent

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

Classical SDE for Overparametrized SGD Dynamics

Stochastic gradient descent

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

Classical SDE for Overparametrized SGD Dynamics

Stochastic gradient descent

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

Classical SDE for Overparametrized SGD Dynamics

Stochastic gradient descent

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

is the Euler-Maruyama scheme for the SDE

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

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

Classical SDE for Overparametrized SGD Dynamics

Stochastic gradient descent

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

is the Euler-Maruyama scheme for the SDE

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

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

$\rightsquigarrow$ $\Sigma^{\frac{1}{2}}$ is $dn \times dn$ matrix! — **Not good for $n \to \infty$**

Martingale Problem for Empirical distribution

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

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

Martingale Problem for Empirical distribution

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

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

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

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

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

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]

Martingale Problem for Empirical distribution

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

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

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]

The martingale problem does not depend on number of parameters (particles), but

- well-posedness is not clear;
- comparison with SGD is not clear

Disadvantages of Existing Models

1. SGD and Stochastic Modified Equation:

$$dZ_t = -\nabla R(Z_t)dt - \frac{\alpha}{4}\nabla|\nabla R(Z_t)|^2 dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(Z_t)dw_t$$

Disadvantages of Existing Models

1. SGD and Stochastic Modified Equation:

$$dZ_t = -\nabla R(Z_t)dt - \frac{\alpha}{4}\nabla|\nabla R(Z_t)|^2 dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(Z_t)dw_t$$

    1. Regularity of $\Sigma^{\frac{1}{2}}$

Disadvantages of Existing Models

1. SGD and Stochastic Modified Equation:

$$dZ_t = -\nabla R(Z_t)dt - \frac{\alpha}{4}\nabla|\nabla R(Z_t)|^2 dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(Z_t)dw_t$$

    1. Regularity of $\Sigma^{\frac{1}{2}}$
    2. non-comparable n-point motions

Disadvantages of Existing Models

1. SGD and Stochastic Modified Equation:

$$dZ_t = -\nabla R(Z_t)dt - \frac{\alpha}{4}\nabla|\nabla R(Z_t)|^2 dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(Z_t)dw_t$$

   1. Regularity of $\Sigma^{\frac{1}{2}}$
   2. non-comparable n-point motions

2. Overparametrized SGD and Measure Dependent SDE

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

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

Disadvantages of Existing Models

1. SGD and Stochastic Modified Equation:

$$dZ_t = -\nabla R(Z_t)dt - \frac{\alpha}{4}\nabla|\nabla R(Z_t)|^2 dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(Z_t)dw_t$$

   1. Regularity of $\Sigma^{\frac{1}{2}}$
   2. non-comparable n-point motions

2. Overparametrized SGD and Measure Dependent SDE

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

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

   1. Σ depends on n

Disadvantages of Existing Models

1. SGD and Stochastic Modified Equation:

$$dZ_t = -\nabla R(Z_t)dt - \frac{\alpha}{4}\nabla|\nabla R(Z_t)|^2 dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(Z_t)dw_t$$

   1. Regularity of $\Sigma^{\frac{1}{2}}$
   2. non-comparable n-point motions

2. Overparametrized SGD and Measure Dependent SDE

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

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

   1. Σ depends on n
   2. martingale problem for μ^n is not easy to be analyzed

Disadvantages of Existing Models

1. SGD and Stochastic Modified Equation:

$$dZ_t = -\nabla R(Z_t)dt - \frac{\alpha}{4}\nabla|\nabla R(Z_t)|^2 dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(Z_t)dw_t$$

   1. Regularity of $\Sigma^{\frac{1}{2}}$
   2. non-comparable n-point motions

2. Overparametrized SGD and Measure Dependent SDE

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

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

   1. Σ depends on n
   2. martingale problem for μ^n is not easy to be analyzed

Our Goal: Propose a new model (some stochastic flow), that would remove this disadvantages

Table of Contents

1 Motivation: Stochastic Gradient Descent

2 Motivation: Overparameterized Stochastic Gradient Descent

3 Stochastic Modified Flow Driven by Inf.-dim Noise

4 Idea of Proof

SDE Driven by Inf-Dim Noise for SGD Dynamics

Stochastic gradient descent

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

SDE Driven by Inf-Dim Noise for SGD Dynamics

Stochastic gradient descent

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

is the Euler-Maruyama scheme for the SDE

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

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

Stochastic Modified Flow and Martingale Problem

Distribution Dependent Stochastic Flow:

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

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

[Dorogovtsev, Kotelenez, Pilipenko, F-Y. Wang,...]

Stochastic Modified Flow and Martingale Problem

Distribution Dependent Stochastic Flow:

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

[Dorogovtsev, Kotelenez, Pilipenko, F-Y. Wang,...]

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 Modified Flow and Martingale Problem

Distribution Dependent Stochastic Flow:

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

[Dorogovtsev, Kotelenez, Pilipenko, F-Y. Wang,...]

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

Well-posedness, comparison with SGD obtained in [Gess, Gvalani, K. '22]

Stochastic Modified Flow and Martingale Problem

Distribution Dependent Stochastic Flow:

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

[Dorogovtsev, Kotelenez, Pilipenko, F-Y. Wang,...]

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

Well-posedness, comparison with SGD obtained in [Gess, Gvalani, K. '22]

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

Distribution Dependent Stochastic Modified Flow

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

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

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

Distribution Dependent Stochastic Modified Flow

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

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

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

Distribution Dependent Stochastic Flow:

$$\begin{aligned} dX(u,t) &= V(X(u,t), \mu_t)dt \\ &\quad + \sqrt{\alpha}\int_\Theta G(X(u,t), \mu_t, \theta)W(d\theta, dt), \\ X(u,0) &= u, \quad \mu_t = \mu_0 \circ X_t^{-1}, \end{aligned}$$

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

Distribution Dependent Stochastic Modified Flow

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

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

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

Distribution Dependent Stochastic Flow:

$$\begin{aligned} dX(u,t) &= V(X(u,t), \mu_t)dt \\ &\quad + \sqrt{\alpha}\int_\Theta G(X(u,t), \mu_t, \theta)W(d\theta, dt), \\ X(u,0) &= u, \quad \mu_t = \mu_0 \circ X_t^{-1}, \end{aligned}$$

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

Theorem 3 (Gess, Kassing, K. '24, JMLR)

Let $\mu_0 \in \mathcal{P}_2$ and V, G be regular enough. Then for every $\Phi \in \mathcal{C}^4_b(\mathcal{P}_2)$

$$\sup_{t_i \le T} |\mathbb{E}\Phi(\mu_{t_i}) - \mathbb{E}\Phi(\nu_{t_i})| \le C\alpha + C\sqrt{\mathbb{E}\mathcal{W}_2^2(\mu_0, \nu_0^n)}.$$

Distribution Dependent Stochastic Modified Flow

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

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

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

Distribution Dependent Stochastic Modified Flow:

$$dX(u,t) = V(X(u,t), \mu_t)dt - \frac{\alpha}{4}\nabla|V(X(u,t), \mu_t)|^2 dt - \frac{\alpha}{4}\langle D|V(X(u,t), \mu_t)|^2, \mu_t\rangle dt$$
$$+ \sqrt{\alpha}\int_\Theta G(X(u,t), \mu_t, \theta)W(d\theta, dt),$$
$$X(u,0) = u, \quad \mu_t = \mu_0 \circ X_t^{-1},$$

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

Theorem 3 (Gess, Kassing, K. '24, JMLR)

Let $\mu_0 \in \mathcal{P}_2$ and V, G be regular enough. Then for every $\Phi \in \mathcal{C}_b^4(\mathcal{P}_2)$

$$\sup_{t_i \le T} |\mathbb{E}\Phi(\mu_{t_i}) - \mathbb{E}\Phi(\nu_{t_i})| \le C\alpha + C\sqrt{\mathbb{E}\mathcal{W}_2^2(\mu_0, \nu_0^n)}.$$

Distribution Dependent Stochastic Modified Flow

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

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

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

Distribution Dependent Stochastic Modified Flow:

$$dX(u,t) = V(X(u,t), \mu_t)dt - \frac{\alpha}{4}\nabla|V(X(u,t), \mu_t)|^2 dt - \frac{\alpha}{4}\langle D|V(X(u,t), \mu_t)|^2, \mu_t\rangle dt$$
$$+ \sqrt{\alpha}\int_\Theta G(X(u,t), \mu_t, \theta)W(d\theta, dt),$$
$$X(u,0) = u, \quad \mu_t = \mu_0 \circ X_t^{-1},$$

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

Theorem 3 (Gess, Kassing, K. '24, JMLR)

Let $\mu_0 \in \mathcal{P}_2$ and V, G be regular enough. Then for every $\Phi \in \mathcal{C}^4_b(\mathcal{P}_2)$

$$\sup_{t_i \le T} |\mathbb{E}\Phi(\mu_{t_i}) - \mathbb{E}\Phi(\nu_{t_i})| \le C\alpha^2 + C\sqrt{\mathbb{E}\mathcal{W}_2^2(\mu_0, \nu_0^n)}.$$

Corollary: n-point motion for SGD

Assume that $V(x,\nu,\theta) = -\nabla R(x,\theta)$, then

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

describes n-**point motion of SGD.**
Consider the Distribution Dependent Stochastic Modified Flow:

$$\begin{aligned} dX(u,t) &= V(X(u,t),\mu_t)dt - \frac{\alpha}{4}\nabla|V(X(u,t),\mu_t)|^2 dt - \frac{\alpha}{4}\langle D|V(X(u,t),\mu_t)|^2,\mu_t\rangle dt \\ &\quad + \sqrt{\alpha}\int_\Theta G(X(u,t),\mu_t,\theta)W(d\theta,dt), \\ X(u,0) &= u, \quad \mu_t = \mu_0 \circ X_t^{-1}, \end{aligned}$$

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

Corollary: n-point motion for SGD

Assume that $V(x,\nu,\theta) = -\nabla R(x,\theta)$, then

$$x_k(t_{i+1}) = x_k(t_i) - \nabla R(x_k(t_i),\theta_i)\Delta t, \quad k \in \{1,\dots,n\},$$

describes n-**point motion of SGD.**
Consider the Distribution Dependent Stochastic Modified Flow:

$$\begin{aligned} dX(u,t) &= V(X(u,t),\mu_t)dt - \frac{\alpha}{4}\nabla|V(X(u,t),\mu_t)|^2 dt - \frac{\alpha}{4}\langle D|V(X(u,t),\mu_t)|^2,\mu_t\rangle dt \\ &\quad + \sqrt{\alpha}\int_\Theta G(X(u,t),\mu_t,\theta)W(d\theta,dt), \\ X(u,0) &= u, \quad \mu_t = \mu_0 \circ X_t^{-1}, \end{aligned}$$

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

Corollary: n-point motion for SGD

Assume that $V(x,\nu,\theta) = -\nabla R(x,\theta)$, then

$$x_k(t_{i+1}) = x_k(t_i) - \nabla R(x_k(t_i),\theta_i)\Delta t, \quad k \in \{1,\dots,n\},$$

describes n-**point motion of SGD.**
Consider the ~~Distribution Dependent~~ Stochastic Modified Flow:

$$\begin{aligned} dX(u,t) &= -\nabla R(X(u,t))dt - \frac{\alpha}{4}\nabla|\nabla R(X(u,t))|^2 dt \\ &\quad + \sqrt{\alpha}\int_\Theta G(X(u,t),\theta)W(d\theta,dt), \\ X(u,0) &= u, \end{aligned}$$

where is a cylindrical Wiener process on $L_2(\Theta,P)$,$G(x,\theta) = \nabla R(x) - \nabla R(x,\theta)$.

Corollary: n-point motion for SGD

Assume that $V(x, \nu, \theta) = -\nabla R(x, \theta)$, then

$$x_k(t_{i+1}) = x_k(t_i) - \nabla R(x_k(t_i), \theta_i)\Delta t, \quad k \in \{1, \dots, n\},$$

describes n-**point motion of SGD.**
Consider the ~~Distribution Dependent~~ Stochastic Modified Flow:

$$\begin{aligned} dX(u,t) &= -\nabla R(X(u,t))dt - \frac{\alpha}{4}\nabla|\nabla R(X(u,t))|^2 dt \\ &\quad + \sqrt{\alpha}\int_\Theta G(X(u,t),\theta)W(d\theta, dt), \\ X(u,0) &= u, \end{aligned}$$

where is a cylindrical Wiener process on $L_2(\Theta, P)$, $G(x,\theta) = \nabla R(x) - \nabla R(x,\theta)$.

Corollary (Gess, Kassing, K. '24, JMLR)

Define $X_k(t) := X(x_k(0), t)$, $k \in [n]$. Then for every $f \in C_b^4(\mathbb{R}^{dn})$

$$\sup_{t_i \le T} |\mathbb{E}f(x_1(t_i), \dots, x_n(t_i)) - \mathbb{E}f(X_1(t_i), \dots, X_n(t_i))| \le C\alpha^2.$$

Stoch. Modified Flow vs Stoch. Modified Equation

Stochastic Modified Flow:

$$dX(u,t) = -\nabla R(X(u,t))dt - \frac{\alpha}{4}\nabla|\nabla R(X(u,t))|^2 dt$$
$$+ \sqrt{\alpha}\int_{\Theta} G(X(u,t),\theta)W(d\theta,dt),$$
$$X(u,0) = u,$$

where is a cylindrical Wiener process on $L_2(\Theta, P)$, $G(x,\theta) = \nabla R(x) - \nabla R(x,\theta)$.

Stoch. Modified Flow vs Stoch. Modified Equation

Stochastic Modified Flow:

$$dX(t) = -\nabla R(X(t))dt - \frac{\alpha}{4}\nabla|\nabla R(X(t))|^2 dt$$
$$+ \sqrt{\alpha}\int_{\Theta} G(X(t),\theta)W(d\theta, dt),$$

where is a cylindrical Wiener process on $L_2(\Theta, P)$, $G(x,\theta) = \nabla R(x) - \nabla R(x,\theta)$.

Stoch. Modified Flow vs Stoch. Modified Equation

Stochastic Modified Flow:

$$dX(t) = -\nabla R(X(t))dt - \frac{\alpha}{4}\nabla|\nabla R(X(t))|^2 dt$$
$$+ \sqrt{\alpha}\int_{\Theta} G(X(t),\theta)W(d\theta, dt),$$

where is a cylindrical Wiener process on $L_2(\Theta, P)$, $G(x,\theta) = \nabla R(x) - \nabla R(x,\theta)$.

Stochastic Modified Equation

$$dX_t = -\nabla R(X_t)dt - \frac{\alpha}{4}\nabla|\nabla R(X_t)|^2 dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(X_t)dw,$$

where $\Sigma(x) = \mathbb{E}_\theta G(x,\theta) \otimes G(x,\theta)$.

Stoch. Modified Flow vs Stoch. Modified Equation

Stochastic Modified Flow:

$$dX(t) = -\nabla R(X(t))dt - \frac{\alpha}{4}\nabla|\nabla R(X(t))|^2 dt$$
$$+ \sqrt{\alpha}\int_{\Theta} G(X(t),\theta)W(d\theta, dt),$$

where is a cylindrical Wiener process on $L_2(\Theta, P)$, $G(x,\theta) = \nabla R(x) - \nabla R(x,\theta)$.

Stochastic Modified Equation

$$dX_t = -\nabla R(X_t)dt - \frac{\alpha}{4}\nabla|\nabla R(X_t)|^2 dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(X_t)dw,$$

where $\Sigma(x) = \mathbb{E}_\theta G(x,\theta) \otimes G(x,\theta)$.

1. SMF discribes and SME have the same martingale problem;

Stoch. Modified Flow vs Stoch. Modified Equation

Stochastic Modified Flow:

$$dX(t) = -\nabla R(X(t))dt - \frac{\alpha}{4}\nabla|\nabla R(X(t))|^2 dt$$
$$+ \sqrt{\alpha}\int_{\Theta} G(X(t),\theta)W(d\theta, dt),$$

where is a cylindrical Wiener process on $L_2(\Theta, P)$, $G(x,\theta) = \nabla R(x) - \nabla R(x,\theta)$.

Stochastic Modified Equation

$$dX_t = -\nabla R(X_t)dt - \frac{\alpha}{4}\nabla|\nabla R(X_t)|^2 dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(X_t)dw,$$

where $\Sigma(x) = \mathbb{E}_\theta G(x,\theta) \otimes G(x,\theta)$.

1. SMF discribes and SME have the same martingale problem;
2. SMF describes n-point motion of SGD, SME – doesn't;

Stoch. Modified Flow vs Stoch. Modified Equation

Stochastic Modified Flow:

$$dX(t) = -\nabla R(X(t))dt - \frac{\alpha}{4}\nabla|\nabla R(X(t))|^2 dt$$
$$+ \sqrt{\alpha}\int_{\Theta} G(X(t),\theta)W(d\theta, dt),$$

where is a cylindrical Wiener process on $L_2(\Theta, P)$, $G(x,\theta) = \nabla R(x) - \nabla R(x,\theta)$.

Stochastic Modified Equation

$$dX_t = -\nabla R(X_t)dt - \frac{\alpha}{4}\nabla|\nabla R(X_t)|^2 dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(X_t)dw,$$

where $\Sigma(x) = \mathbb{E}_\theta G(x,\theta) \otimes G(x,\theta)$.

1. SMF discribes and SME have the same martingale problem;
2. SMF describes n-point motion of SGD, SME – doesn't;
3. SMF avoids the irregularity of $\sqrt{\Sigma}$, e.g. $\Sigma(x) = x^2$.

Table of Contents

1 Motivation: Stochastic Gradient Descent

2 Motivation: Overparameterized Stochastic Gradient Descent

3 Stochastic Modified Flow Driven by Inf.-dim Noise

4 Idea of Proof

Flow structure of overparameterized SGD

The SGD

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

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

Flow structure of overparameterized SGD

The SGD

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

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

$$x(u, t_{i+1}) = x(u, t_i) + V(x(u, t_i), \nu_{t_i}, \theta_i)\Delta t,$$
$$x(u, 0) = u, \quad \nu_{t_i} = \nu_0^{-1} \circ x(\cdot, t_i)$$

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

Interpolation of One-Step estimate

Set $(t_1 = \Delta t = \alpha)$

$$\mathcal{S}\Psi(\mu_0) := \mathbb{E}_P \Psi(\nu_{t_1}) = \mathbb{E}_P \Psi(\mu_0 \circ x(\cdot, t_1))^{-1})$$

and

$$\mathcal{T}_t \Psi(\mu_0) := \mathbb{E}_P \Psi(\mu_t) = \mathbb{E}_P \Psi(\mu_0 \circ X(\cdot, t)^{-1}).$$

Interpolation of One-Step estimate

Set $(t_1 = \Delta t = \alpha)$

$$\mathcal{S}\Psi(\mu_0) := \mathbb{E}_P \Psi(\nu_{t_1}) = \mathbb{E}_P \Psi(\mu_0 \circ x(\cdot, t_1))^{-1})$$

and

$$\mathcal{T}_t \Psi(\mu_0) := \mathbb{E}_P \Psi(\mu_t) = \mathbb{E}_P \Psi(\mu_0 \circ X(\cdot, t)^{-1}).$$

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

$$\mathbb{E}\Phi(\mu_0 \circ x(\cdot, t_n)^{-1}) - \mathbb{E}\Phi(\mu_0 \circ X_{t_n}^{-1}) = \mathbb{E}\Phi(\nu_{t_n}) - \mathbb{E}\Phi(\mu_{t_n}) = \mathcal{S}^n \Phi(\mu_0) - \mathcal{T}_{t_n} \Phi(\mu_0)$$

Interpolation of One-Step estimate

Set $(t_1 = \Delta t = \alpha)$

$$\mathcal{S}\Psi(\mu_0) := \mathbb{E}_P \Psi(\nu_{t_1}) = \mathbb{E}_P \Psi(\mu_0 \circ x(\cdot, t_1))^{-1})$$

and

$$\mathcal{T}_t \Psi(\mu_0) := \mathbb{E}_P \Psi(\mu_t) = \mathbb{E}_P \Psi(\mu_0 \circ X(\cdot, t)^{-1}).$$

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

$$\mathbb{E}\Phi(\mu_0 \circ x(\cdot, t_n)^{-1}) - \mathbb{E}\Phi(\mu_0 \circ X_{t_n}^{-1}) = \mathbb{E}\Phi(\nu_{t_n}) - \mathbb{E}\Phi(\mu_{t_n}) = \mathcal{S}^n\Phi(\mu_0) - \mathcal{T}_{t_n}\Phi(\mu_0)$$

$$= \sum_{i=0}^{n-1} \left(\mathcal{S}^{n-i}\mathcal{T}_{t_i}\Phi(\mu_0) - \mathcal{S}^{n-i-1}\mathcal{T}_{t_{i+1}}\Phi(\mu_0) \right)$$

Interpolation of One-Step estimate

Set $(t_1 = \Delta t = \alpha)$

$$\mathcal{S}\Psi(\mu_0) := \mathbb{E}_P \Psi(\nu_{t_1}) = \mathbb{E}_P \Psi(\mu_0 \circ x(\cdot, t_1))^{-1})$$

and

$$\mathcal{T}_t \Psi(\mu_0) := \mathbb{E}_P \Psi(\mu_t) = \mathbb{E}_P \Psi(\mu_0 \circ X(\cdot, t)^{-1}).$$

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

$$\mathbb{E}\Phi(\mu_0 \circ x(\cdot, t_n)^{-1}) - \mathbb{E}\Phi(\mu_0 \circ X_{t_n}^{-1}) = \mathbb{E}\Phi(\nu_{t_n}) - \mathbb{E}\Phi(\mu_{t_n}) = \mathcal{S}^n \Phi(\mu_0) - \mathcal{T}_{t_n} \Phi(\mu_0)$$

$$= \sum_{i=0}^{n-1} \left(\mathcal{S}^{n-i} \mathcal{T}_{t_i} \Phi(\mu_0) - \mathcal{S}^{n-i-1} \mathcal{T}_{t_{i+1}} \Phi(\mu_0) \right)$$

$$= \sum_{i=0}^{n-1} \mathcal{S}^{n-i-1} \left(\mathcal{S}\mathcal{T}_{t_i} \Phi(\mu_0) - \mathcal{T}_\alpha \underbrace{\mathcal{T}_{t_i} \Phi(\mu_0)}_{=: U(t_i, \mu_0)} \right).$$

Interpolation of One-Step estimate

Set $(t_1 = \Delta t = \alpha)$

$$\mathcal{S}\Psi(\mu_0) := \mathbb{E}_P \Psi(\nu_{t_1}) = \mathbb{E}_P \Psi(\mu_0 \circ x(\cdot, t_1))^{-1})$$

and

$$\mathcal{T}_t \Psi(\mu_0) := \mathbb{E}_P \Psi(\mu_t) = \mathbb{E}_P \Psi(\mu_0 \circ X(\cdot, t)^{-1}).$$

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

$$\begin{aligned}
\mathbb{E}\Phi(\mu_0 \circ x(\cdot, t_n)^{-1}) - \mathbb{E}\Phi(\mu_0 \circ X_{t_n}^{-1}) &= \mathbb{E}\Phi(\nu_{t_n}) - \mathbb{E}\Phi(\mu_{t_n}) = \mathcal{S}^n \Phi(\mu_0) - \mathcal{T}_{t_n} \Phi(\mu_0) \\
&= \sum_{i=0}^{n-1} \left(\mathcal{S}^{n-i} \mathcal{T}_{t_i} \Phi(\mu_0) - \mathcal{S}^{n-i-1} \mathcal{T}_{t_{i+1}} \Phi(\mu_0) \right) \\
&= \sum_{i=0}^{n-1} \mathcal{S}^{n-i-1} \left(\mathcal{S}\mathcal{T}_{t_i} \Phi(\mu_0) - \mathcal{T}_\alpha \underbrace{\mathcal{T}_{t_i} \Phi(\mu_0)}_{=: U(t_i, \mu_0)} \right).
\end{aligned}$$

Since $\sup_{\mu_0 \in \mathcal{P}_2} |\mathcal{S}\Psi(\mu_0)| \leq \sup_{\mu_0 \in \mathcal{P}_2} |\Psi(\mu_0)|$,

$$\sup_{\mu_0 \in \mathcal{P}} \left| \mathbb{E}\Phi(\mu_0 \circ x(\cdot, t_n)^{-1}) - \mathbb{E}\Phi(\mu_0 \circ X(\cdot, t_n)^{-1}) \right| \leq \sum_{i=0}^{n-1} \sup_{\mu_0 \in \mathcal{P}_2} |\mathcal{S}U(t_i, \mu_0) - \mathcal{T}_\alpha U(t_i, \mu_0)|.$$

Expansions of $S\Psi(\mu_0)$ and $P_\alpha\Psi(\mu_0)$

Expansion in Taylor's series w.r.t $\alpha = \Delta t$

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

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

Expansions of $S\Psi(\mu_0)$ and $P_\alpha\Psi(\mu_0)$

Expansion in Taylor's series w.r.t $\alpha = \Delta t$

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

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

$$P_\alpha\Psi(\mu_0) = \Psi(\mu_0) + \int_0^\alpha \mathcal{L}P_s\Psi(\mu_0)ds,$$

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

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

Expansions of $S\Psi(\mu_0)$ and $P_\alpha\Psi(\mu_0)$

Expansion in Taylor's series w.r.t $\alpha = \Delta t$

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

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

$$P_\alpha\Psi(\mu_0) = \Psi(\mu_0) + \int_0^\alpha \mathcal{L}P_s\Psi(\mu_0)ds,$$

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

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

Iterating the equality above, one gets

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

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

Comparison of Generators and End of Proof

For $t_n = \alpha n \leq T$

$$\sup_{\mu_0 \in \mathcal{P}} \left| \mathbb{E}\Phi(\mu_0 \circ Z_{t_n}^{-1}) - \mathbb{E}\Phi(\mu_0 \circ X_{t_n}^{-1}) \right| \leq \sum_{i=0}^{n-1} \sup_{\mu_0 \in \mathcal{P}_2} |\mathcal{S}U(t_i, \mu_0) - P_\alpha U(t_i, \mu_0)|$$

Comparison of Generators and End of Proof

For $t_n = \alpha n \leq T$

$$\sup_{\mu_0 \in \mathcal{P}} \left| \mathbb{E}\Phi(\mu_0 \circ Z_{t_n}^{-1}) - \mathbb{E}\Phi(\mu_0 \circ X_{t_n}^{-1}) \right| \leq \sum_{i=0}^{n-1} \sup_{\mu_0 \in \mathcal{P}_2} |\mathcal{S}U(t_i, \mu_0) - P_\alpha U(t_i, \mu_0)|$$

$$\leq \sum_{i=0}^{n-1} \sup_{\mu_0 \in \mathcal{P}_2} \alpha^3 \, |R_1(U(t_i, \mu_0), \mu_0) - R_2(U(t_i, \mu_0), \mu_0)|$$

Comparison of Generators and End of Proof

For $t_n = \alpha n \leq T$

$$\sup_{\mu_0 \in \mathcal{P}} \left| \mathbb{E}\Phi(\mu_0 \circ Z_{t_n}^{-1}) - \mathbb{E}\Phi(\mu_0 \circ X_{t_n}^{-1}) \right| \leq \sum_{i=0}^{n-1} \sup_{\mu_0 \in \mathcal{P}_2} |\mathcal{S}U(t_i, \mu_0) - P_\alpha U(t_i, \mu_0)|$$

$$\leq \sum_{i=0}^{n-1} \sup_{\mu_0 \in \mathcal{P}_2} \alpha^3 \, |R_1(U(t_i, \mu_0), \mu_0) - R_2(U(t_i, \mu_0), \mu_0)|$$

$$\leq \alpha^3 n C \|U\|_{C_b^{0,4}([0,T] \times \mathcal{P}_2)} \leq C_1 T \alpha^2.$$

Comparison of Generators and End of Proof

For $t_n = \alpha n \leq T$

$$\sup_{\mu_0 \in \mathcal{P}} \left| \mathbb{E}\Phi(\mu_0 \circ Z_{t_n}^{-1}) - \mathbb{E}\Phi(\mu_0 \circ X_{t_n}^{-1}) \right| \leq \sum_{i=0}^{n-1} \sup_{\mu_0 \in \mathcal{P}_2} |\mathcal{S}U(t_i, \mu_0) - P_\alpha U(t_i, \mu_0)|$$

$$\leq \sum_{i=0}^{n-1} \sup_{\mu_0 \in \mathcal{P}_2} \alpha^3 \, |R_1(U(t_i, \mu_0), \mu_0) - R_2(U(t_i, \mu_0), \mu_0)|$$

$$\leq \alpha^3 n C \| U \|_{\mathcal{C}_b^{0,4}([0,T] \times \mathcal{P}_2)} \leq C_1 T \alpha^2.$$

Proposition [Feng-Yu Wang, J. Evol. Equ., '21]

Let $V \in \mathcal{C}_b^{5,5}(\mathbb{R}^d \times \mathcal{P}_2)$, $G(\cdot, \cdot, \theta) \in \mathcal{C}_b^{4,4}(\mathbb{R}^d \times \mathcal{P}_2)$ P-a.s. Then for every $\Phi \in \mathcal{C}_b^4(\mathcal{P}_2)$ the function $U(t, \mu_0) = \mathbb{E}\Phi(\mu_t)$ is a unique solution to the equation

$$\partial_t U(t, \mu_0) = \mathcal{L}_t U(t, \mu_0),$$
$$U(0, \mu_0) = \Phi(\mu_0).$$

Moreover, $U \in \mathcal{C}_b^{0,4}([0, T] \times \mathcal{P}_2)$ and $\partial_t U \in \mathcal{C}([0, T] \times \mathcal{P}_2)$.

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!