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

Vitalii Konarovskyi

Hamburg University and Institute of Mathematics of NASU

Kyiv-Leipzig Seminar — 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

Supervised Learning

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

Supervised Learning

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

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

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

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

Supervised Learning

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

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

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

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

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

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

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

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

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

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

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

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

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

Continuous Dynamics of Parameters

Recall that $x_k(0) \sim \mu_0$ – i.i.d., Δt – learning rate, $t_i = i\Delta t$, $\theta_i \sim P$ – 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 $\nu^n_t = \frac{1}{n}\sum_{k=1}^n \delta_{x_k(t)}$.

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, $t_i = i\Delta t$, $\theta_i \sim P$ – 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 $\nu^n_t = \frac{1}{n}\sum_{k=1}^n \delta_{x_k(t)}$.

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., then

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

where μ_t solves

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

[Mei, Montanari, Nguyen '18]

Convergence to deterministic SPDE

If $x_k(0) \sim \mu_0$ – i.i.d., then

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

where μ_t solves

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

[Mei, Montanari, Nguyen '18]

$\Longrightarrow$ The mean behavior of the SGD dynamics can then be analysed by considering μ_t.

Convergence to deterministic SPDE

If $x_k(0) \sim \mu_0$ – i.i.d., then

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

where μ_t solves

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

[Mei, Montanari, Nguyen '18]

$\Longrightarrow$ The mean behavior of the SGD dynamics can then be analysed by considering μ_t.

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.

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).

SDE Driven by Inf-Dim Noise for SGD Dynamics

Stochastic gradient descent

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

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).

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

SDE Driven by Inf-Dim Noise for SGD Dynamics

Stochastic gradient descent

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

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).

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

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

Well-Posedness of SMFE

Theorem 1 (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

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

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

Table of Contents

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 2 **(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 $p \in [1, 2)$

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

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 2 (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 $p \in [1, 2)$

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

$$\rightsquigarrow O(n^{-1}), \quad \text{if quintified CLT for SGD holds.}$$

Table of Contents

Stochastic Modified Equation and SGD

$$x(t_{n+1}) = x(t_n) - \nabla R(x(t_n), \theta_n)\Delta t$$

Stochastic Modified Equation and SGD

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

Stochastic Modified Equation and SGD

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

is the Euler scheme for the SDE

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

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

Stochastic Modified Equation and SGD

$$x(t_{n+1}) = x(t_n) - \nabla R(x(t_n), \theta_n)\Delta t$$
$$= x(t_n) - \nabla \underbrace{\mathbb{E}_\theta R(\dots)}_{R(x(t_n))} \Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}} \underbrace{(\nabla \mathbb{E}_\theta R(\dots) - \nabla R(x(t_n), \theta_n))}_{=G(x(t_n),\theta_n)} \sqrt{\Delta t}$$

is the Euler scheme for the SDE

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

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

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

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

$$\sup_{t_i \le T} |\mathbb{E} f(x_{t_i}) - \mathbb{E} f(X_{t_i})| = O(\alpha).$$

Stochastic Modified Equation and SGD

$$x(t_{n+1}) = x(t_n) - \nabla R(x(t_n), \theta_n)\Delta t$$
$$= x(t_n) - \nabla \underbrace{\mathbb{E}_\theta R(\dots)}_{R(x(t_n))} \Delta t + \underbrace{\sqrt{\Delta t}}_{=\sqrt{\alpha}} \underbrace{(\nabla \mathbb{E}_\theta R(\dots) - \nabla R(x(t_n), \theta_n))}_{=G(x(t_n),\theta_n)} \sqrt{\Delta t}$$

is the Euler scheme for the SDE

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

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

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

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

$$\sup_{t_i \le T} |\mathbb{E}f(x_{t_i}) - \mathbb{E}f(X_{t_i})| = O(\alpha^2).$$

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:

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

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 $\Phi \in \mathcal{C}^4_b(\mathcal{P}_2)$

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

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 $\Phi \in \mathcal{C}^4_b(\mathcal{P}_2)$

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

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:

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

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

Corollary: n-point motion for SGD

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

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

describes ***n*-point motion of SGD.**
Consider the 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 $L_2(\Theta,P)$,

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:

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

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

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:

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

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

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

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

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

Stoch. Modified Flow vs Stoch. Modified Equation

Stochastic Modified Flow:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $t_n = \alpha n \leq T$

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

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 $t_n = \alpha n \le T$

$$\begin{aligned}\sup_{\mu_0 \in \mathcal{P}} \left| \mathbb{E}\Phi(\mu_0 \circ Z_{t_n}^{-1}) - \mathbb{E}\Phi(\mu_0 \circ X_{t_n}^{-1}) \right| &\le \sum_{i=0}^{n-1} \sup_{\mu_0 \in \mathcal{P}_2} \left| \mathcal{S}U(t_i, \mu_0) - P_\alpha U(t_i, \mu_0) \right| \\ &\le \sum_{i=0}^{n-1} \sup_{\mu_0 \in \mathcal{P}_2} \alpha^3 \left| R_1(U(t_i, \mu_0), \mu_0) - R_2(U(t_i, \mu_0), \mu_0) \right| \\ &\le \alpha^3 n C \|U\|_{\mathcal{C}_b^{0,4}([0,T] \times \mathcal{P}_2)} \le C_1 T \alpha^2.\end{aligned}$$

Comparison of Generators and End of Proof

For $t_n = \alpha n \leq T$

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

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

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

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!