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

Vitalii Konarovskyi

Bielefeld University & Institute of Mathematics of NAS of Ukraine

Malliavin Calculus and its Applications – Kyiv

joint work with Benjamin Gess and Sebastian Kassing

Table of Contents

Supervised Machine Learning

$\{(\theta_i, \gamma_i),\ i \in I\}$ – training data set

θ_i – i.i.d. samples from P

$\gamma_i = f(\theta_i)$, where $f : \Theta \to \mathbb{R}^l$ has to be modeled

Supervised Machine Learning

$\{(\theta_i, \gamma_i),\ i \in I\}$ – training data set

θ_i – i.i.d. samples from P

$\gamma_i = f(\theta_i)$, where $f : \Theta \to \mathbb{R}^l$ has to be modeled

Idea: Model $f(\theta)$ by $U(\theta, z)$, where $z \in \mathbb{R}^d$ is a parameter that has to be found.

Ex.: $U(\theta, z) = c \cdot h(A\theta + b)$, $z = (c, A, b)$ – neural network approximation.

Supervised Machine Learning

$\{(\theta_i, \gamma_i),\ i \in I\}$ – training data set

θ_i – i.i.d. samples from P

$\gamma_i = f(\theta_i)$, where $f : \Theta \to \mathbb{R}^l$ has to be modeled

Idea: Model $f(\theta)$ by $U(\theta, z)$, where $z \in \mathbb{R}^d$ is a parameter that has to be found.

Ex.: $U(\theta, z) = c \cdot h(A\theta + b)$, $z = (c, A, b)$ – neural network approximation.

Consider a **loss function** $l : \mathbb{R}^2 \to \mathbb{R}_+$
(Ex. $l(a, b) = |a - b|^2$, $l(a, b) = |a - b|$, $l(a, b) = \mathbb{I}_{\{a \neq b\}}$) and define

$$R(z) = \mathbb{E}_P l(f(\theta), U(\theta, z)) \ - \ \textbf{generalization error}.$$

$$R \to \min$$

Stochastic gradient descent

Set

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

Stochastic gradient descent

Set

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

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

$$Z_{t_{n+1}} = Z_{t_n} - \eta \nabla \tilde{R}(Z_{t_n}, \theta_n)$$

for learning rate η, $t_n = \eta n$ and $\theta_n \sim P$ – i.i.d.

Stochastic Differential equation

$$Z_{t_{n+1}} = Z_{t_n} - \eta \nabla \tilde{R}(Z_{t_n}, \theta_n)$$

Stochastic Differential equation

$$
\begin{aligned}
Z_{t_{n+1}} &= Z_{t_n} - \eta \nabla \tilde{R}(Z_{t_n}, \theta_n) \\
&= Z_{t_n} - \nabla R(Z_{t_n})\eta + \sqrt{\eta} \underbrace{\left(\nabla R(Z_{t_n}) - \nabla \tilde{R}(Z_{t_n}, \theta_n) \right)}_{G(Z_{t_n}, \theta_n)} \sqrt{\eta}
\end{aligned}
$$

Stochastic Differential equation

$$
\begin{aligned}
Z_{t_{n+1}} &= Z_{t_n} - \eta \nabla \tilde{R}(Z_{t_n}, \theta_n) \\
&= Z_{t_n} - \nabla R(Z_{t_n})\eta + \sqrt{\eta} \underbrace{\left(\nabla R(Z_{t_n}) - \nabla \tilde{R}(Z_{t_n}, \theta_n)\right)}_{G(Z_{t_n}, \theta_n)} \sqrt{\eta}
\end{aligned}
$$

is the Euler scheme for the SDE

$$
dY_t = -\nabla R(Y_t)dt + \sqrt{\eta}\Sigma^{\frac{1}{2}}(Y_t)dW,
$$

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

Stochastic Differential equation

$$Z_{t_{n+1}} = Z_{t_n} - \eta\nabla\tilde{R}(Z_{t_n},\theta_n)$$
$$= Z_{t_n} - \nabla R(Z_{t_n})\eta + \sqrt{\eta}\underbrace{\left(\nabla R(Z_{t_n}) - \nabla\tilde{R}(Z_{t_n},\theta_n)\right)}_{G(Z_{t_n},\theta_n)}\sqrt{\eta}$$

is the Euler scheme for the SDE

$$dY_t = -\nabla R(Y_t)dt + \sqrt{\eta}\Sigma^{\frac{1}{2}}(Y_t)dW,$$

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

Theorem Li, Tai, E '19

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

$$\sup_{t_n\le T}|\mathbb{E}f(Z_{t_n}) - \mathbb{E}f(Y_{t_n})| \le C\eta.$$

Stochastic Differential equation

$$\begin{aligned} Z_{t_{n+1}} &= Z_{t_n} - \eta \nabla \tilde{R}(Z_{t_n}, \theta_n) \\ &= Z_{t_n} - \nabla R(Z_{t_n})\eta + \sqrt{\eta} \underbrace{\left(\nabla R(Z_{t_n}) - \nabla \tilde{R}(Z_{t_n}, \theta_n) \right)}_{G(Z_{t_n}, \theta_n)} \sqrt{\eta} \end{aligned}$$

is the Euler scheme for the SDE

$$dY_t = -\nabla R(Y_t)dt - \frac{\eta}{4} \nabla |\nabla R(Y_t)|^2 dt + \sqrt{\eta} \Sigma^{\frac{1}{2}}(Y_t) dW,$$

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

Theorem Li, Tai, E '19

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

$$\sup_{t_n \leq T} |\mathbb{E} f(Z_{t_n}) - \mathbb{E} f(Y_{t_n})| \leq C\eta^2.$$

Disadvantages of the SDE approximation

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

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

Disadvantages of the SDE approximation

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

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

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

Denoted the SGD dynamics started from z by $Z(z)$, i.e, for $Z_0(z) = z \in \mathbb{R}^d$ define

$$Z_{t_{n+1}}(z) = Z_{t_n}(z) - \eta \nabla \tilde{R}(Z_{t_n}(z), \theta_n)$$

for learning rate η, $t_n = \eta n$ and $\theta_n \sim P$ – i.i.d.

Then

$$(Z(z^1), \ldots, Z(z^m)) \not\approx (Y(z^1), \ldots, Y(z^m)).$$

Stochastic Modified Flow

Stochastic Gradient Descent:

$$Z_{t_{n+1}} = Z_{t_n} - \nabla R(Z_{t_n})\eta + \sqrt{\eta}\underbrace{\left(\nabla R(Z_{t_n}) - \nabla \tilde{R}(Z_{t_n}, \theta_n)\right)}_{G(Z_{t_n}, \theta_n)}\sqrt{\eta}.$$

Stochastic Modified flow:

$$dX_t = -\nabla R(X_t)dt - \frac{\eta}{4}\nabla|\nabla R(X_t)|^2 dt + \sqrt{\eta}\Sigma^{\frac{1}{2}}(X_t)dW.$$

Stochastic Modified Flow

Stochastic Gradient Descent:

$$Z_{t_{n+1}} = Z_{t_n} - \nabla R(Z_{t_n})\eta + \sqrt{\eta} \underbrace{\left(\nabla R(Z_{t_n}) - \nabla \tilde{R}(Z_{t_n}, \theta_n)\right)}_{G(Z_{t_n}, \theta_n)} \sqrt{\eta}.$$

Stochastic Modified flow:

$$dX_t = -\nabla R(X_t)dt - \frac{\eta}{4}\nabla|\nabla R(X_t)|^2 dt + \sqrt{\eta}\int_\Theta G(X_t, \theta)W(d\theta, dt),$$

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

Stochastic Modified Flow

Stochastic Gradient Descent:

$$Z_{t_{n+1}} = Z_{t_n} - \nabla R(Z_{t_n})\eta + \sqrt{\eta}\underbrace{\left(\nabla R(Z_{t_n}) - \nabla \tilde{R}(Z_{t_n}, \theta_n)\right)}_{G(Z_{t_n},\theta_n)}\sqrt{\eta}.$$

Stochastic Modified flow:

$$dX_t = -\nabla R(X_t)dt - \frac{\eta}{4}\nabla|\nabla R(X_t)|^2 dt + \sqrt{\eta}\int_\Theta G(X_t, \theta)W(d\theta, dt),$$

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

Theorem 1 Gess, Kassing, K. '23

Let $\tilde{R}(\cdot, \theta) \in \mathcal{C}_b^6$ for P-a.e. θ and $\int_\Theta \|\tilde{R}(\cdot, \theta)\|^2_{\mathcal{C}_b^6} P(d\theta) < \infty$. Then for all $f \in \mathcal{C}_b^4$ and $\Phi \in \mathcal{C}_b^4(\mathcal{P}_2)$

$$\sup_{t_n \le T}\left|\mathbb{E}f(Z_{t_n}(z^1), \dots, Z_{t_n}(z^m)) - \mathbb{E}f(X_{t_n}(z^1), \dots, X_{t_n}(z^m))\right| \le C\eta^2$$

and $$\sup_{t_n \le T}\left|\mathbb{E}\Phi(\mu \circ Z_{t_n}^{-1}) - \mathbb{E}\Phi(\mu \circ X_{t_n}^{-1})\right| \le C\eta^2.$$

Table of Contents

Supervised Learning in Overparametrized Regime

$\{(\theta_i, \gamma_i),\ i \in I\}$ – training data set

θ_i – i.i.d. samples from P

$\gamma_i = f(\theta_i)$, where $f : \Theta \to \mathbb{R}^l$ has to be modeled

Supervised Learning in Overparametrized Regime

$\{(\theta_i, \gamma_i),\ i \in I\}$ – training data set

θ_i – i.i.d. samples from P

$\gamma_i = f(\theta_i)$, where $f : \Theta \to \mathbb{R}^l$ has to be modeled

Idea: Model $f(\theta)$ by $U(\theta, z)$, $z \in \mathbb{R}^d$.

Ex.: $U(\theta, z) = c \cdot h(A\theta + b)$, $z = (c, A, b)$ – neural network approximation.

Supervised Learning in Overparametrized Regime

$\{(\theta_i, \gamma_i),\ i \in I\}$ – training data set

θ_i – i.i.d. samples from P

$\gamma_i = f(\theta_i)$, where $f : \Theta \to \mathbb{R}^l$ has to be modeled

Idea: Model $f(\theta)$ by

$$f_n(\theta, z) = \frac{1}{m} \sum_{k=1}^{m} U(\theta, z^k) = \langle U(\theta, \cdot), \nu \rangle,$$

where $z^k \in \mathbb{R}^d$ are parameters that has to be found, and $\nu = \frac{1}{m} \sum_{k=1}^{m} \delta_{z^k}$.

Ex.: $U(\theta, z) = c \cdot h(A\theta + b)$, $z = (c, A, b)$ – neural network approximation.

Supervised Learning in Overparametrized Regime

$\{(\theta_i, \gamma_i),\ i \in I\}$ – training data set

θ_i – i.i.d. samples from P

$\gamma_i = f(\theta_i)$, where $f : \Theta \to \mathbb{R}^l$ has to be modeled

Idea: Model $f(\theta)$ by

$$f_n(\theta, z) = \frac{1}{m} \sum_{k=1}^{m} U(\theta, z^k) = \langle U(\theta, \cdot), \nu \rangle,$$

where $z^k \in \mathbb{R}^d$ are parameters that has to be found, and $\nu = \frac{1}{m} \sum_{k=1}^{m} \delta_{z^k}$.

Ex.: $U(\theta, z) = c \cdot h(A\theta + b)$, $z = (c, A, b)$ – neural network approximation.

Consider the **square loss function** $l(a, b) = |a - b|^2$

$$R(z) = \mathbb{E}_P l |f(\theta) - f_n(\theta, z)|^2 \ - \ \textbf{generalization error}.$$

$$R \to \min$$

SGD in Overparametrized Regime

Stochastic Gradient Descent: $Z_0^k \sim \mu_0$ i.i.d.

$$Z_{t_{j+1}}^k = Z_{t_j}^k - \eta \nabla_{z^k} \left(\frac{1}{2} |f(\theta_j) - f_m(\theta_j, Z_{t_j})|^2 \right)$$

for learning rate η, $t_n = \eta n$, $\theta_n \sim P$ – i.i.d.

SGD in Overparametrized Regime

Stochastic Gradient Descent: $Z_0^k \sim \mu_0$ i.i.d.

$$Z_{t_{j+1}}^k = Z_{t_j}^k - \eta \nabla_{z^k} \left(\frac{1}{2} |f(\theta_j) - f_m(\theta_j, Z_{t_j})|^2 \right)$$
$$= Z_{t_j}^k + \eta \tilde{V}(Z_{t_j}^k, \nu_{t_j}, \theta_j)$$

for learning rate η, $t_n = \eta n$, $\theta_n \sim P$ – i.i.d. and

$$\nu_t = \frac{1}{m} \sum_{k=1}^{m} \delta_{Z_t^k}.$$

SGD in Overparametrized Regime

Stochastic Gradient Descent: $Z_0^k \sim \mu_0$ i.i.d.

$$Z_{t_{j+1}}^k = Z_{t_j}^k - \eta \nabla_{z^k} \left(\frac{1}{2} |f(\theta_j) - f_m(\theta_j, Z_{t_j})|^2 \right)$$
$$= Z_{t_j}^k + \eta \tilde{V}(Z_{t_j}^k, \nu_{t_j}, \theta_j)$$

for learning rate η, $t_n = \eta n$, $\theta_n \sim P$ – i.i.d. and

$$\nu_t = \frac{1}{m} \sum_{k=1}^{m} \delta_{Z_t^k}.$$

Distribution dependent SDE:

$$dX_t^k = V(X_t^k, \mu_t^m)dt + \sqrt{\eta} \int_\Theta G(X_t^k, \mu_t^m, \theta) W(d\theta, dt),$$
$$\mu_t^m = \frac{1}{m} \sum_{k=1}^{m} \delta_{X_t^k}, \quad V := \mathbb{E}_P \tilde{V}, \quad G := \tilde{V} - V.$$

See also [Rotskoff, Vanden-Eijnden, CPAM, '22; Gess, Gvalani, K. '22]

Distribution Dependent Stochastic Modified Flow

Stochastic Gradient Descent: $Z^k(0) \sim \mu_0$ i.i.d.

$$Z^k_{t_{j+1}} = Z^k_{t_j} + \eta \tilde{V}(Z^k_{t_j}, \nu_{t_j}, \theta_j)$$

for learning rate η, $t_n = \eta n$ and $\theta_n \sim P$ – i.i.d. $\nu_t = \frac{1}{m}\sum_{k=1}^m \delta_{Z^k_t}$.

Distribution Dependent Stochastic Modified Flow:

$$dX_t(x) = V(X_t(x), \mu_t)dt + \sqrt{\eta}\int_\Theta G(X_t(x), \mu_t, \theta)W(d\theta, dt),$$

$$X_0(x) = x, \quad \mu_t = \mu_0 \circ X_t^{-1},$$

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

[Dorogovtsev, Kotelenez, Pilipenko, Ostapenko, Weiß, Wang . . .]

Distribution Dependent Stochastic Modified Flow

Stochastic Gradient Descent: $Z^k(0) \sim \mu_0$ i.i.d.

$$Z^k_{t_{j+1}} = Z^k_{t_j} + \eta \tilde{V}(Z^k_{t_j}, \nu_{t_j}, \theta_j)$$

for learning rate η, $t_n = \eta n$ and $\theta_n \sim P$ – i.i.d. $\nu_t = \frac{1}{m}\sum_{k=1}^m \delta_{Z^k_t}$.

Distribution Dependent Stochastic Modified Flow:

$$dX_t(x) = V(X_t(x), \mu_t)dt - \frac{\eta}{4}\nabla|V(X_t(x), \mu_t)|^2 dt - \frac{\eta}{4}\langle D|V(X_t(x), \mu_t)|^2, \mu_t\rangle dt$$
$$+ \sqrt{\eta}\int_\Theta G(X_t(x), \mu_t, \theta)W(d\theta, dt),$$
$$X_0(x) = x, \quad \mu_t = \mu_0 \circ X_t^{-1},$$

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

[Dorogovtsev, Kotelenez, Pilipenko, Ostapenko, Weiß, Wang ...]

Distribution Dependent Stochastic Modified Flow

Stochastic Gradient Descent: $Z^k(0) \sim \mu_0$ i.i.d.

$$Z^k_{t_{j+1}} = Z^k_{t_j} + \eta \tilde{V}(Z^k_{t_j}, \nu_{t_j}, \theta_j)$$

for learning rate η, $t_n = \eta n$ and $\theta_n \sim P$ – i.i.d. $\nu_t = \frac{1}{m}\sum_{k=1}^m \delta_{Z^k_t}$.

Distribution Dependent Stochastic Modified Flow:

$$dX_t(x) = V(X_t(x), \mu_t)dt - \frac{\eta}{4}\nabla|V(X_t(x), \mu_t)|^2 dt - \frac{\eta}{4}\langle D|V(X_t(x), \mu_t)|^2, \mu_t\rangle dt$$
$$+ \sqrt{\eta}\int_\Theta G(X_t(x), \mu_t, \theta)W(d\theta, dt),$$
$$X_0(x) = x, \quad \mu_t = \mu_0 \circ X_t^{-1},$$

where is a cylindrical Wiener process on $L_2(\Theta, P)$.
[Dorogovtsev, Kotelenez, Pilipenko, Ostapenko, Weiß, Wang . . .]

Theorem 2 Gess, Kassing, K. '23

Let $\mu_0 \in \mathcal{P}_2$ and $\int_\Theta \left(\|U(\cdot,\theta)\|^2_{\mathcal{C}^6_b} + |f(\theta)|^2\right)\|U(\cdot,\theta)\|^2_{\mathcal{C}^6_b} P(d\theta) < \infty$. Then for every $\Phi \in \mathcal{C}^4_b(\mathcal{P}_2)$ and $m \geq 1/\eta^{2d}$

$$\sup_{t_n \leq T} |\mathbb{E}\Phi(\mu_{t_n}) - \mathbb{E}\Phi(\nu_{t_n})| \leq C\eta^2.$$

Table of Contents

Lion's Derivative

We say that a function $f : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}^k$ is L-differentiable at μ, if there exists $Df(\mu) \in L_2\left(\mathbb{R}^d \to \mathbb{R}^k \times \mathbb{R}^d, \mu\right)$ such that

$$f\left(\mu \circ (\mathrm{id} + h)^{-1}\right) - f(\mu) = \langle Df(\mu), h \rangle_\mu + o\left(\|h\|_\mu\right).$$

Ex. If $f(\mu) = g(\langle \varphi, \mu \rangle)$, then $Df(\mu, x) = g'(\langle \varphi, \mu \rangle)\nabla\varphi(x)$.

Definition

We write $f \in \mathcal{C}_b^1(\mathcal{P}_2)$ if f is L-differentiable at every point $\mu \in \mathcal{P}_2$ and its derivative at μ has μ-version $Df(\mu, x)$ which is jointly continuous and bounded.

Similarly, we can define the class $\mathcal{C}_b^m(\mathcal{P}_2)$.

Discrete and Continuous Dynamics

Fix measurable $V : R^d \times \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}^d$, $G : \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \to L_2\left(\Theta \to \mathbb{R}^d, P\right)$ with $\mathbb{E}_P G(\mu, x, \theta) = 0$.

Discrete and Continuous Dynamics

Fix measurable $V : R^d \times \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}^d$, $G : \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \to L_2\left(\Theta \to \mathbb{R}^d, P\right)$ with $\mathbb{E}_P G(\mu, x, \theta) = 0$. Define

$$Z_{t_{n+1}}(z) = Z_{t_n}(z) + \eta V(Z_{t_n}(z), \nu_{t_n}) + \eta G(Z_{t_n}(z), \nu_{t_n}, \theta_n),$$
$$Z_0(z) = z, \quad \nu_{t_n} = \mu_0 \circ Z_{t_n}^{-1}, \quad t_n = n\eta, \quad \theta_n \sim P - \text{i.i.d.}$$

and

$$dX_t(x) = \left[V(X_t(x), \mu_t) - \frac{\eta}{4}\nabla|V(X_t(x), \mu_t)|^2 - \frac{\eta}{4}\langle D|V(X_t(x), \mu_t)|^2, \mu_t\rangle\right] dt$$
$$+ \sqrt{\eta}\int_\Theta G(X_t(x), \mu_t, \theta) W(d\theta, dt)$$
$$X_0(x) = x, \quad \mu_t = \mu_0 \circ X_t^{-1}.$$

Main result

Theorem 3 Gess, Kassing, K. '23

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

$$\sup_{\mu_0 \in \mathcal{P}_2} \sup_{t_n \leq T} |\mathbb{E}\Phi(\nu_{t_n}) - \mathbb{E}\Phi(\mu_{t_n})| \leq C\eta^2.$$

Table of Contents

Interpolation of One-Step estimate

Set

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

and

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

Interpolation of One-Step estimate

Set

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

and

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

Then for $t_n = \eta n$

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

Interpolation of One-Step estimate

Set

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

and

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

Then for $t_n = \eta n$

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

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

Interpolation of One-Step estimate

Set

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

and

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

Then for $t_n = \eta n$

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

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

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

Interpolation of One-Step estimate

Set

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

and

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

Then for $t_n = \eta n$

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

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

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

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

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

Expansion of $\Psi(\nu_\eta)$

For fixed $\theta \in \Theta$ we consider

$$Z_\eta(\mu_0, z) := z + \eta V(z, \mu_0) + \eta G(z, \mu_0, \theta_1), \quad z \in \mathbb{R}^d,$$

as a random variable on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu_0)$.

Expansion of $\Psi(\nu_\eta)$

For fixed $\theta \in \Theta$ we consider

$$Z_\eta(\mu_0, z) := z + \eta V(z, \mu_0) + \eta G(z, \mu_0, \theta_1), \quad z \in \mathbb{R}^d,$$

as a random variable on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu_0)$.

Define

$$\xi_s(z) := (1 - s)z + sZ_\eta(z, \mu_0), \quad s \in [0, 1].$$

Expansion of $\Psi(\nu_\eta)$

For fixed $\theta \in \Theta$ we consider

$$Z_\eta(\mu_0, z) := z + \eta V(z, \mu_0) + \eta G(z, \mu_0, \theta_1), \quad z \in \mathbb{R}^d,$$

as a random variable on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu_0)$.

Define

$$\xi_s(z) := (1-s)z + sZ_\eta(z, \mu_0), \quad s \in [0,1].$$

Using Taylor's formula,

$$\Psi(\nu_\eta) = \Psi(\mu_0 \circ Z_\eta^{-1}(\mu_0, \cdot)) = \Psi(\operatorname{Law}\xi_1) = \Psi(\operatorname{Law}\xi_0) + \frac{d}{ds}\Psi(\operatorname{Law}\xi_s)|_{s=0} + \frac{1}{2}\frac{d^2}{ds^2}\Psi(\operatorname{Law}\xi_s)|_{s=0} + \frac{1}{2}\int_0^1 \frac{d^3}{ds^3}\Psi(\operatorname{Law}\xi_s)(1-s)^3 ds.$$

Chain rule

Lemma Ren, Wang '19, Wang '21

Let ξ_s, $s \geq 0$, be a family of square integrable random variables on $\mathbb{R}^k$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If

$$\xi_0' := \lim_{s \to 0+} \frac{\xi_s - \xi_0}{s}$$

exists in $L_2(\Omega \to \mathbb{R}^k, \mathbb{P})$, then $\forall f \in \mathcal{C}^1(\mathcal{P}_2)$ one has

$$\lim_{s \to 0+} \frac{f(\mathrm{Law}\, \xi_s) - f(\mathrm{Law}\, \xi_0)}{s} = \mathbb{E}\left[Df(\mathrm{Law}\, \xi_0, \xi_0) \cdot \xi_0' \right].$$

Expansion of $S\Psi(\mu_0)$

Recall

$$\xi_s(z) := (1-s)z + sZ_\eta(z,\mu_0), \quad s \in [0,1],$$

Expansion of $S\Psi(\mu_0)$

Recall

$$\xi_s(z) := (1-s)z + sZ_\eta(z,\mu_0), \quad s \in [0,1],$$

and note

$$\xi_0'(z) := Z_\eta(z,\mu_0) - z = \eta[V(z,\mu_0) + G(z,\mu_0,\theta)].$$

Expansion of $S\Psi(\mu_0)$

Recall

$$\xi_s(z) := (1-s)z + sZ_\eta(z,\mu_0), \quad s \in [0,1],$$

and note

$$\xi_0'(z) := Z_\eta(z,\mu_0) - z = \eta[V(z,\mu_0) + G(z,\mu_0,\theta)].$$

Then

$$\begin{aligned}\frac{d}{ds}\Psi(\mathrm{Law}\,\xi_s)|_{s=0} &= \mathbb{E}_{\mu_0}[Df(\mathrm{Law}\,\xi_0,\xi_0)\cdot\xi_0'] \\ &= \eta\int_{\mathbb{R}^d} D\Psi(\mu_0,z)\cdot[V(z,\mu_0)+G(z,\mu_0,\theta)]\,\mu(dz).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathcal{S}\Psi(\mu_0) = \mathbb{E}_P\Psi(\mathrm{Law}\,\xi_1) = \Psi(\mu_0) + \eta\int_{\mathbb{R}^d} D\Psi(z,\mu_0)\cdot V(z,\mu_0)\mu_0(dz) \\ + \eta^2(\dots) + \eta^3 R_1(\Psi,\mu_0),\end{aligned}$$

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

Expansion of $P_\eta \Psi(\mu_0)$

Recall that

$$dX_t(x) = \left[V(X_t(x), \mu_t) - \frac{\eta}{4} \nabla |V(X_t(x), \mu_t)|^2 - \frac{\eta}{4} \langle D|V(X_t(x), \mu_t)|^2, \mu_t \rangle \right] dt$$
$$+ \sqrt{\eta} \int_\Theta G(X_t(x), \mu_t, \theta) W(d\theta, dt)$$
$$X_0(x) = x, \quad \mu_t = \mu_0 \circ X_t^{-1}.$$

Then,

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

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

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

Expansion of $P_\eta\Psi(\mu_0)$

Recall that

$$dX_t(x) = \left[V(X_t(x),\mu_t) - \frac{\eta}{4}\nabla|V(X_t(x),\mu_t)|^2 - \frac{\eta}{4}\langle D|V(X_t(x),\mu_t)|^2,\mu_t\rangle\right]dt$$
$$+\sqrt{\eta}\int_\Theta G(X_t(x),\mu_t,\theta)W(d\theta,dt)$$
$$X_0(x) = x,\quad \mu_t = \mu_0\circ X_t^{-1}.$$

Then,

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

where $\mathcal{L} = \mathcal{L}_1 + \eta\mathcal{L}_2$ 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

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

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

End of Proof

For $t_n = \eta n \leq T$

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

End of Proof

For $t_n = \eta n \leq T$

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

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

End of Proof

For $t_n = \eta n \leq T$

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

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

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

End of Proof

For $t_n = \eta n \leq T$

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

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

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

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

Stochastic Modified Flow (again)

Stochastic Gradient Descent:

$$Z_{t_{n+1}} = Z_{t_n} - \nabla R(Z_{t_n})\eta + \sqrt{\eta}\,\underbrace{\left(\nabla R(Z_{t_n}) - \nabla \tilde{R}(Z_{t_n}, \theta_n)\right)}_{G(Z_{t_n}, \theta_n)}\sqrt{\eta}.$$

Stochastic Modified flow:

$$dX_t = -\nabla R(X_t)dt - \frac{\eta}{4}\nabla|\nabla R(X_t)|^2 dt + \sqrt{\eta}\int_\Theta G(X_t, \theta)W(d\theta, dt),$$

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

Theorem 1 Gess, Kassing, K. '23

Let $\tilde{R}(\cdot, \theta) \in \mathcal{C}_b^6$ for P-a.e. θ and $\int_\Theta \|\tilde{R}(\cdot, \theta)\|^2_{\mathcal{C}_b^6} P(d\theta) < \infty$. Then for all $f \in \mathcal{C}_b^4$ and $\Phi \in \mathcal{C}_b^4(\mathcal{P}_2)$

$$\sup_{t_n \le T}\left|\mathbb{E}f(Z_{t_n}(z^1), \dots, Z_{t_n}(z^m)) - \mathbb{E}f(X_{t_n}(z^1), \dots, X_{t_n}(z^m))\right| \le C\eta^2$$

and $$\sup_{t_n \le T}\left|\mathbb{E}\Phi(\mu \circ Z_{t_n}^{-1}) - \mathbb{E}\Phi(\mu \circ X_{t_n}^{-1})\right| \le C\eta^2.$$

Prof of Theorem 1

Taking $\tilde{V}(x,\mu,\theta) = \tilde{R}(x,\theta)$ and $\mu_0 = \delta_{z^1}$, we get for $m = 1$

$$\sup_{t_n \le T} \left| \mathbb{E} f(Z_{t_n}(z^1)) - \mathbb{E} f(X_{t_n}(z^1)) \right| \le \sup_{t_n \le T} \left| \mathbb{E}\langle f, \mu_0 \circ Z_{t_n}^{-1} \rangle - \mathbb{E}\langle f, \mu_0 \circ X_{t_n}^{-1} \rangle \right| \le C\eta^2.$$

Prof of Theorem 1

Taking $\tilde{V}(x, \mu, \theta) = \tilde{R}(x, \theta)$ and $\mu_0 = \delta_{z^1}$, we get for $m = 1$

$$\sup_{t_n \leq T} \left| \mathbb{E} f(Z_{t_n}(z^1)) - \mathbb{E} f(X_{t_n}(z^1)) \right| \leq \sup_{t_n \leq T} \left| \mathbb{E} \langle f, \mu_0 \circ Z_{t_n}^{-1} \rangle - \mathbb{E} \langle f, \mu_0 \circ X_{t_n}^{-1} \rangle \right| \leq C\eta^2.$$

Define for $z = (z^1, \ldots, z^m) \in \mathbb{R}^{md}$

$$\tilde{R}^{ext}(z, \theta) := \tilde{R}(z^1, \theta) + \cdots + \tilde{R}(z^m, \theta)$$

and let $Z_{t_n}^{ext}$, X_t^{ext}, be defined as above for $\tilde{R}^{exp}$ instead of $\tilde{R}$.

Prof of Theorem 1

Taking $\tilde{V}(x,\mu,\theta) = \tilde{R}(x,\theta)$ and $\mu_0 = \delta_{z^1}$, we get for $m = 1$

$$\sup_{t_n \le T} \left| \mathbb{E} f(Z_{t_n}(z^1)) - \mathbb{E} f(X_{t_n}(z^1)) \right| \le \sup_{t_n \le T} \left| \mathbb{E}\langle f, \mu_0 \circ Z_{t_n}^{-1} \rangle - \mathbb{E}\langle f, \mu_0 \circ X_{t_n}^{-1} \rangle \right| \le C\eta^2.$$

Define for $z = (z^1, \ldots, z^m) \in \mathbb{R}^{md}$

$$\tilde{R}^{ext}(z,\theta) := \tilde{R}(z^1,\theta) + \cdots + \tilde{R}(z^m,\theta)$$

and let $Z_{t_n}^{ext}$, X_t^{ext}, be defined as above for $\tilde{R}^{exp}$ instead of $\tilde{R}$.

$$\nabla \tilde{R}^{ext}(z,\theta) = \left(\nabla_{z^i} \tilde{R}(z^i,\theta) \right)_{i \in [m]} \quad \Longrightarrow \quad Z_{t_n}^{ext}(z) = \left(Z_{t_n}(z^i) \right)_{i \in [m]}.$$

Prof of Theorem 1

Taking $\tilde{V}(x,\mu,\theta) = \tilde{R}(x,\theta)$ and $\mu_0 = \delta_{z^1}$, we get for $m = 1$

$$\sup_{t_n \le T} \left| \mathbb{E} f(Z_{t_n}(z^1)) - \mathbb{E} f(X_{t_n}(z^1)) \right| \le \sup_{t_n \le T} \left| \mathbb{E}\langle f, \mu_0 \circ Z_{t_n}^{-1} \rangle - \mathbb{E}\langle f, \mu_0 \circ X_{t_n}^{-1} \rangle \right| \le C\eta^2.$$

Define for $z = (z^1, \dots, z^m) \in \mathbb{R}^{md}$

$$\tilde{R}^{ext}(z,\theta) := \tilde{R}(z^1,\theta) + \cdots + \tilde{R}(z^m,\theta)$$

and let $Z_{t_n}^{ext}$, X_t^{ext}, be defined as above for $\tilde{R}^{exp}$ instead of $\tilde{R}$.

$$\nabla \tilde{R}^{ext}(z,\theta) = \left(\nabla_{z^i} \tilde{R}(z^i,\theta) \right)_{i \in [m]} \quad \Longrightarrow \quad Z_{t_n}^{ext}(z) = \left(Z_{t_n}(z^i) \right)_{i \in [m]}.$$

$$\left.\begin{array}{l} \nabla R^{ext}(z,\theta) = \left(\nabla_{z^i} R(z^i,\theta) \right)_i \\ \nabla |\nabla R^{ext}(z)|^2 = \left(\nabla_{z^i} |\nabla_{z^i} R(z^i)|^2 \right)_i \\ G^{ext}(z,\theta) = \left(G(z^i,\theta) \right)_i \end{array}\right\} \quad \Longrightarrow \quad X_t^{ext}(z) = \left(X_t(z^i) \right)_{i \in [m]}.$$

Prof of Theorem 1

Taking $\tilde{V}(x,\mu,\theta) = \tilde{R}(x,\theta)$ and $\mu_0 = \delta_{z^1}$, we get for $m = 1$

$$\sup_{t_n \le T} \left| \mathbb{E} f(Z_{t_n}(z^1)) - \mathbb{E} f(X_{t_n}(z^1)) \right| \le \sup_{t_n \le T} \left| \mathbb{E}\langle f, \mu_0 \circ Z_{t_n}^{-1} \rangle - \mathbb{E}\langle f, \mu_0 \circ X_{t_n}^{-1} \rangle \right| \le C\eta^2.$$

Define for $z = (z^1, \dots, z^m) \in \mathbb{R}^{md}$

$$\tilde{R}^{ext}(z,\theta) := \tilde{R}(z^1,\theta) + \dots + \tilde{R}(z^m,\theta)$$

and let $Z_{t_n}^{ext}$, X_t^{ext}, be defined as above for $\tilde{R}^{exp}$ instead of $\tilde{R}$.

$$\nabla \tilde{R}^{ext}(z,\theta) = \left(\nabla_{z^i} \tilde{R}(z^i,\theta) \right)_{i \in [m]} \quad \Longrightarrow \quad Z_{t_n}^{ext}(z) = \left(Z_{t_n}(z^i) \right)_{i \in [m]}.$$

$$\left.\begin{array}{l} \nabla R^{ext}(z,\theta) = \left(\nabla_{z^i} R(z^i,\theta) \right)_i \\ \nabla |\nabla R^{ext}(z)|^2 = \left(\nabla_{z^i} |\nabla_{z^i} R(z^i)|^2 \right)_i \\ G^{ext}(z,\theta) = \left(G(z^i,\theta) \right)_i \end{array} \right\} \quad \Longrightarrow \quad X_t^{ext}(z) = \left(X_t(z^i) \right)_{i \in [m]}.$$

Remark: This trick does not work for the diffusion term $\sqrt{\eta}\Sigma^{\frac{1}{2}}(X_t)dW_t$ with $\Sigma = \mathbb{E}_p G(z,\theta) \otimes G(z,\theta)$.

Overparametrized case

Stochastic Gradient Descent: $Z_0^k \sim \mu_0$ i.i.d.

$$Z_{t_{j+1}}^k = Z_{t_j}^k + \eta \tilde{V}(Z_{t_j}^k, \nu_{t_j}, \theta_j)$$

for learning rate η, $t_n = \eta n$ and $\theta_n \sim P$ – i.i.d. $\nu_t = \frac{1}{m}\sum_{k=1}^m \delta_{Z_t^k}$.

Distribution Dependent Stochastic Modified Flow:

$$dX_t(x) = V(X_t(x), \mu_t)dt - \frac{\eta}{4}\nabla |V(X_t(x), \mu_t)|^2 dt - \frac{\eta}{4}\langle D|V(X_t(x), \mu_t)|^2, \mu_t\rangle dt$$
$$+ \sqrt{\eta}\int_\Theta G(X_t(x), \mu_t, \theta) W(d\theta, dt),$$
$$X_0(x) = x, \quad \mu_t = \mu_0 \circ X_t^{-1}$$

where is a cylindrical Wiener process on $L_2(\Theta, P)$.
[Dorogovtsev, Kotelenez, Pilipenko, Ostapenko, Weiß, Wang ...]

Theorem 2 Gess, Kassing, K. '23

Let $\mu_0 \in \mathcal{P}_2$ and $\int_\Theta \left(\|U(\cdot,\theta)\|^2_{\mathcal{C}_b^6} + |f(\theta)|^2 \right) \|U(\cdot,\theta)\|^2_{\mathcal{C}_b^6} P(d\theta) < \infty$. Then for every $\Phi \in \mathcal{C}_b^4(\mathcal{P}_2)$ and $m \geq 1/\eta^{2d}$

$$\sup_{t_n \leq T} |\mathbb{E}\Phi(\mu_{t_n}) - \mathbb{E}\Phi(\nu_{t_n})| \leq C\eta^2.$$

Proof of Theorem 2

Let μ_t^m be a solution to the Distribution Dependent Stochastic Modified Flow started from $\nu_0 = \frac{1}{m}\sum_{k=1}^m \delta_{Z_0^k}$.

Using the Lipshitz continuity of solutions to DDSMF
see e.g. [Dorogovtsev '07; Gess, Gvalani, K. '22],

$$\sup_{t_n \leq T} |\mathbb{E}\Phi(\mu_{t_n}) - \mathbb{E}\Phi(\nu_{t_n})|$$

Proof of Theorem 2

Let μ_t^m be a solution to the Distribution Dependent Stochastic Modified Flow started from $\nu_0 = \frac{1}{m}\sum_{k=1}^m \delta_{Z_0^k}$.

Using the Lipshitz continuity of solutions to DDSMF
see e.g. [Dorogovtsev '07; Gess, Gvalani, K. '22],

$$\sup_{t_n \leq T} |\mathbb{E}\Phi(\mu_{t_n}) - \mathbb{E}\Phi(\nu_{t_n})| \leq \sup_{t_n \leq T} |\mathbb{E}\Phi(\mu_{t_n}) - \mathbb{E}\Phi(\mu_{t_n}^m)| + \sup_{t_n \leq T} |\mathbb{E}\Phi(\mu_{t_n}^m) - \mathbb{E}\Phi(\nu_{t_n})|$$

Proof of Theorem 2

Let μ_t^m be a solution to the Distribution Dependent Stochastic Modified Flow started from $\nu_0 = \frac{1}{m}\sum_{k=1}^m \delta_{Z_0^k}$.

Using the Lipshitz continuity of solutions to DDSMF
see e.g. [Dorogovtsev '07; Gess, Gvalani, K. '22],

$$\begin{aligned} &\sup_{t_n \le T} |\mathbb{E}\Phi(\mu_{t_n}) - \mathbb{E}\Phi(\nu_{t_n})| \\ &\qquad \le \sup_{t_n \le T} |\mathbb{E}\Phi(\mu_{t_n}) - \mathbb{E}\Phi(\mu_{t_n}^m)| + \sup_{t_n \le T} |\mathbb{E}\Phi(\mu_{t_n}^m) - \mathbb{E}\Phi(\nu_{t_n})| \\ &\qquad \le \|\Phi\|_{\mathcal{C}_b^1} \sup_{t_n \le T} \mathbb{E}\mathcal{W}_2(\mu_{t_n}, \mu_{t_n}^m) + C\eta^2 \end{aligned}$$

Proof of Theorem 2

Let μ_t^m be a solution to the Distribution Dependent Stochastic Modified Flow started from $\nu_0 = \frac{1}{m}\sum_{k=1}^m \delta_{Z_0^k}$.

Using the Lipshitz continuity of solutions to DDSMF
see e.g. [Dorogovtsev '07; Gess, Gvalani, K. '22],

$$
\begin{aligned}
&\sup_{t_n \le T} |\mathbb{E}\Phi(\mu_{t_n}) - \mathbb{E}\Phi(\nu_{t_n})| \\
&\qquad \le \sup_{t_n \le T} |\mathbb{E}\Phi(\mu_{t_n}) - \mathbb{E}\Phi(\mu_{t_n}^m)| + \sup_{t_n \le T} |\mathbb{E}\Phi(\mu_{t_n}^m) - \mathbb{E}\Phi(\nu_{t_n})| \\
&\qquad \le \|\Phi\|_{\mathcal{C}_b^1} \sup_{t_n \le T} \mathbb{E}\mathcal{W}_2(\mu_{t_n}, \mu_{t_n}^m) + C\eta^2 \\
&\qquad \le \|\Phi\|_{\mathcal{C}_b^1} \sup_{t_n \le T} \left(\mathbb{E}\mathcal{W}_2^2(\mu_{t_n}, \mu_{t_n}^m)\right)^{\frac{1}{2}} + C\eta^2
\end{aligned}
$$

Proof of Theorem 2

Let μ_t^m be a solution to the Distribution Dependent Stochastic Modified Flow started from $\nu_0 = \frac{1}{m}\sum_{k=1}^m \delta_{Z_0^k}$.

Using the Lipshitz continuity of solutions to DDSMF
see e.g. [Dorogovtsev '07; Gess, Gvalani, K. '22],

$$
\begin{aligned}
&\sup_{t_n\le T} |\mathbb{E}\Phi(\mu_{t_n}) - \mathbb{E}\Phi(\nu_{t_n})| \\
&\qquad \le \sup_{t_n\le T} |\mathbb{E}\Phi(\mu_{t_n}) - \mathbb{E}\Phi(\mu_{t_n}^m)| + \sup_{t_n\le T} |\mathbb{E}\Phi(\mu_{t_n}^m) - \mathbb{E}\Phi(\nu_{t_n})| \\
&\qquad \le \|\Phi\|_{\mathcal{C}_b^1} \sup_{t_n\le T} \mathbb{E}\mathcal{W}_2(\mu_{t_n}, \mu_{t_n}^m) + C\eta^2 \\
&\qquad \le \|\Phi\|_{\mathcal{C}_b^1} \sup_{t_n\le T} \left(\mathbb{E}\mathcal{W}_2^2(\mu_{t_n}, \mu_{t_n}^m)\right)^{\frac{1}{2}} + C\eta^2 \\
&\qquad \le C\left(\mathbb{E}\mathcal{W}_2^2(\mu_0, \mu_0^m)\right)^{\frac{1}{2}} + C\eta^2
\end{aligned}
$$

Proof of Theorem 2

Let μ^m_t be a solution to the Distribution Dependent Stochastic Modified Flow started from $\nu_0 = \frac{1}{m}\sum_{k=1}^m \delta_{Z_0^k}$.

Using the Lipshitz continuity of solutions to DDSMF see e.g. [Dorogovtsev '07; Gess, Gvalani, K. '22],

$$\begin{aligned}
&\sup_{t_n\le T} |\mathbb{E}\Phi(\mu_{t_n}) - \mathbb{E}\Phi(\nu_{t_n})| \\
&\qquad \le \sup_{t_n\le T} |\mathbb{E}\Phi(\mu_{t_n}) - \mathbb{E}\Phi(\mu^m_{t_n})| + \sup_{t_n\le T} |\mathbb{E}\Phi(\mu^m_{t_n}) - \mathbb{E}\Phi(\nu_{t_n})| \\
&\qquad \le \|\Phi\|_{\mathcal{C}^1_b} \sup_{t_n\le T} \mathbb{E}\mathcal{W}_2(\mu_{t_n}, \mu^m_{t_n}) + C\eta^2 \\
&\qquad \le \|\Phi\|_{\mathcal{C}^1_b} \sup_{t_n\le T} \left(\mathbb{E}\mathcal{W}_2^2(\mu_{t_n}, \mu^m_{t_n})\right)^{\frac{1}{2}} + C\eta^2 \\
&\qquad \le C\left(\mathbb{E}\mathcal{W}_2^2(\mu_0, \mu^m_0)\right)^{\frac{1}{2}} + C\eta^2
\end{aligned}$$

The control of $\mathbb{E}\mathcal{W}_2^2(\mu_0, \mu^m_0)$ follows from the quantified Law of Large Numbers from [Fournier, Guillin, On the rate of convergence in Wasserstein distance of the empirical measure, PTRF, '15].

Reference

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

Gess, Gvalani, Konarovskyi,
Conservative SPDEs as Fluctuating Mean-Field Limits of Stochastic Gradient Descent
(arXiv:2207.05705)

Thank you!