Conservative SPDEs as Fluctuating Mean Field Limits of Stochastic Gradient Descent

Vitalii Konarovskyi

Bielefeld University

International Conference of Young Mathematicians — Kyiv

joint work with Benjamin Gess and Rishabh Gvalani

UNIVERSITÄT BIELEFELD

National Academy of Sciences of Ukraine
INSTITUTE OF MATHEMATICS

Supervised Learning

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

Supervised Learning

- Having a large sets of data $\{(\theta_i, \gamma_i),\ i \in I\}$, $\theta_i \sim \vartheta$ 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 $\{(\theta_i, \gamma_i),\ i \in I\}$, $\theta_i \sim \vartheta$ 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**

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

where ϑ is the distribution of θ_i.

Stochastic gradient descent

Let $x_k(0) \sim \mu_0$ – i.i.d.

Stochastic gradient descent

Let $x_k(0) \sim \mu_0$ – i.i.d.

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

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

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

Stochastic gradient descent

Let $x_k(0) \sim \mu_0$ – i.i.d.

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

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

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

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

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

where Δt – **learning rate**, $t_i = i\Delta t$, $\theta_i \sim \vartheta$ – 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, $t_i = i\Delta t$, $\theta_i \sim \vartheta$ – 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, $t_i = i\Delta t$, $\theta_i \sim \vartheta$ – 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, $t_i = i\Delta t$, $\theta_i \sim \vartheta$ – 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. and $\Delta t = \frac{1}{n}$, then

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

where μ_t solves

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

[Mei, Montanari, Nguyen '18]

Convergence to deterministic SPDE

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

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

where μ_t solves

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

[Mei, Montanari, Nguyen '18]

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

Main Goal

Problem. After passing to the limit the equation

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

loses the information about the fluctuations of the SGD dynamics

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i),\nu^n_{t_i},\theta_i)\Delta t, \quad \nu^n_t = \frac{1}{n}\sum_{l=1}^{n}\delta_{x_l(t)}.$$

Main Goal

Problem. After passing to the limit the equation

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

loses the information about the fluctuations of the SGD dynamics

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i), \nu^n_{t_i}, \theta_i)\Delta t, \quad \nu^n_t = \frac{1}{n}\sum_{l=1}^{n} \delta_{x_l(t)}.$$

Goal: Propose a **stochastic** PDE which would capture the fluctuations of the SGD dynamics. Then, probably, its solutions would better approximate the SGD dynamics as $n \to \infty$ and $\Delta t \to 0$.

Classical SDE for SGD Dynamics

Stochastic gradient descent

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

Classical SDE for SGD Dynamics

Stochastic gradient descent

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

Classical SDE for 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 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

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

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

Classical SDE for SGD Dynamics

Stochastic gradient descent

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

is the Euler-Maruyama scheme for the SDE

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

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

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

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 $L_2(\Theta, \vartheta)$.

[Gess, Kassing, K. '23]

Stochastic Mean-Field Equation

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

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

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

Stochastic Mean-Field Equation

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

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

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

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

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

Stochastic Mean-Field Equation

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

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

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

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

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

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

Stochastic Mean-Field Equation

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

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

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

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

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

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

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

Higher Order Approximation of SGD

Stochastic Mean-Field Equation:

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

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

Higher Order Approximation of SGD

Stochastic Mean-Field Equation:

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

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

Theorem 1 **(Gess, Gvalani, K. 2022)**

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

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

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

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

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

Quantified Central Limit Theorem for SMFE

Theorem 2 (Gess, Gvalani, K. 2022)

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

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

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

Quantified Central Limit Theorem for SMFE

Theorem 2 **(Gess, Gvalani, K. 2022)**

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

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

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

Remark. **[Sirignano, Spiliopoulos, '20]**

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

$$\tilde{\eta}^n \to \eta.$$

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

Note that

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

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

Note that

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

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

Note that

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

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

Continuity Equation

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

Continuity Equation

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

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

where

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

[Ambrosio, Trevisan, Lions,. . .]

Continuity Equation

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

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

where

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

[Ambrosio, Trevisan, Lions,...]

The Stochastic Mean-Field Equation was derived from:

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

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

Well-Posedness of SMFE

Theorem 3 (Gess, Gvalani, K. 2022)

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

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

has a unique solution. Moreover, μ_t is a superposition solution, i.e.,

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

where X solves

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

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

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
(arXiv:2302.07125)

Thank you!