Conservative SPDEs as fluctuating mean field limits of stochastic gradient descent

Vitalii Konarovskyi

Bielefeld University

GPSD 2023 — Essen

joint work with Benjamin Gess and Rishabh Gvalani

Table of Contents

1. Derivation of SPDE and application in Machine Learning

2. Well-posedness, superposition principle and connection with SGD dynamics

Supervised Learning

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

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

where $x_k \in \mathbb{R}^d$, $k \in \{1, \ldots, n\}$, are parameters which have to be found.
Example: $U(\theta, x) = c \cdot h(a \cdot \theta + b), \quad x = (a, b, c)$

Supervised Learning

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

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

  where $x_k \in \mathbb{R}^d$, $k \in \{1, \dots, n\}$, are parameters which have to be found.
  Example: $U(\theta, x) = c \cdot h(a \cdot \theta + b), \quad x = (a, b, c)$
- We measure the distance between f and f_n by the **generalization error**

$$\mathcal{L}[f_n] = \frac{1}{2}\mathbb{E}_m l(f(\theta), f_n(\theta)) = \frac{1}{2}\int_\Theta l(f(\theta), f_n(\theta))\mathrm{m}(d\theta),$$

where m is the distribution of θ_i.

Supervised Learning

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

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

 where $x_k \in \mathbb{R}^d$, $k \in \{1, \ldots, n\}$, are parameters which have to be found.
 Example: $U(\theta, x) = c \cdot h(a \cdot \theta + b), \quad x = (a, b, c)$
- We measure the distance between f and f_n by the **generalization error**

$$\mathcal{L}[f_n] = \frac{1}{2}\mathbb{E}_m |f(\theta) - f_n(\theta)|^2 = \frac{1}{2} \int_{\Theta} |f(\theta) - f_n(\theta)|^2 \mathrm{m}(d\theta),$$

 where m is the distribution of θ_i.

Stochastic Gradient Descent and (deterministic) PDE

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

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

where Δt is a **learning rate**, $t_i = i\Delta t$, $\{\theta_i, i \in \mathbb{N}\}$ are i.i.d. with distribution m,

Stochastic Gradient Descent and (deterministic) PDE

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

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

where Δt is a **learning rate**, $t_i = i\Delta t$, $\{\theta_i, i \in \mathbb{N}\}$ are i.i.d. with distribution m, $\nu^n_t = \frac{1}{n} \sum_{l=1}^{n} \delta_{x_l(t)}$.

Stochastic Gradient Descent and (deterministic) PDE

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

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

where Δt is a **learning rate**, $t_i = i\Delta t$, $\{\theta_i, i \in \mathbb{N}\}$ are i.i.d. with distribution m, $\nu^n_t = \frac{1}{n} \sum_{l=1}^n \delta_{x_l(t)}$.

Stochastic Gradient Descent and (deterministic) PDE

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

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

where Δt is a **learning rate**, $t_i = i\Delta t$, $\{\theta_i, i \in \mathbb{N}\}$ are i.i.d. with distribution m, $\nu^n_t = \frac{1}{n} \sum_{l=1}^n \delta_{x_l(t)}$.

If $x_k(0)$ are i.i.d. from μ_0, then

$$d(\nu^n_t, \mu_t) = O(n^{-1/2}) + O(\Delta t^{1/2}) = O(n^{-1/2}), \quad \text{for } \Delta t = \frac{1}{n},$$

where μ_t solves

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

with $V(x, \mu) = \mathbb{E}_{\mathrm{m}} V(x, \mu, \theta)$. [Mei, Montanarib, Nguyen, 2018]

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 $\Delta t \to 0$.

SGD and Martingale Problem (standard approach)

Stochastic gradient descent

$$
\begin{aligned}
x_k(t_{i+1}) &= x_k(t_i) + V(x_k(t_i), \nu^n_{t_i}, \theta_i)\Delta t \\
&= x_k(t_i) + V(x_k(t_i), \nu^n_{t_i})\Delta t + \sqrt{\Delta t}\left(V(x_k(t_i), \nu^n_{t_i}, \theta_i) - V(x_k(t_i), \nu^n_{t_i})\right)\sqrt{\Delta t}
\end{aligned}
$$

SGD and Martingale Problem (standard approach)

Stochastic gradient descent

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

SGD and Martingale Problem (standard approach)

Stochastic gradient descent

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

is the Euler-Maruyama scheme for the SDE

$$dx_k(t) = V(x_k(t), \mu^n_t)dt + \sqrt{\alpha} dB_k(t), \quad k \in \{1, \ldots, n\}$$

$$d[B_k, B_l]_t = A(x_k(t), x_l(t), \mu^n_t)dt,$$

where $\mu^n_t = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$ and $A(x, y, \mu) = \mathbb{E}_{\mathrm{m}} G(x, \mu, \theta) \otimes G(y, \mu, \theta)$.

SGD and Martingale Problem (standard approach)

Stochastic gradient descent

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

is the Euler-Maruyama scheme for the SDE

$$
dx_k(t) = V(x_k(t), \mu^n_t)dt + \sqrt{\alpha} dB_k(t), \quad k \in \{1, \dots, n\}
$$

$$
d[B_k, B_l]_t = A(x_k(t), x_l(t), \mu^n_t)dt,
$$

where $\mu^n_t = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$ and $A(x, y, \mu) = \mathbb{E}_{\mathrm{m}} G(x, \mu, \theta) \otimes G(y, \mu, \theta)$.

$$
d\mu^n_t = \frac{\alpha}{2}\nabla^2 : (A(\cdot, \mu^n_t)\mu^n_t)dt - \nabla \cdot (V(\cdot, \mu^n_t)\mu^n_t)dt + \nabla \cdot \sqrt{\alpha} dW^{\mathrm{cor}}(\cdot, t),
$$

with $[dW^{\mathrm{cor}}(x, t), dW^{\mathrm{cor}}(y, t)] = A(x, y, \mu^n_t)\mu^n_t(x)\mu^n_t(y)$.

[Rotskoff, Vanden-Eijnden, CPAM, 2022]

SGD and SPDE (new approach)

Stochastic gradient descent

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

where $\nu^n_t = \frac{1}{n}\sum_{l=1}^n \delta_{x_l(t)}$, $\alpha = \Delta t$, $G(x,\mu,\theta) = V(x,\mu,\theta) - V(x,\mu)$ and θ_i are i.i.d. with distribution m on Θ.

SGD and SPDE (new approach)

Stochastic gradient descent

$$x_k(t_{i+1}) = x_k(t_i) + V(x_k(t_i), \nu_{t_i}^n)\Delta t + \sqrt{\alpha} G(x_k(t_i), \nu_{t_i}^n, \theta_i)\sqrt{\Delta t},$$

where $\nu_t^n = \frac{1}{n}\sum_{l=1}^n \delta_{x_l(t)}$, $\alpha = \Delta t$, $G(x, \mu, \theta) = V(x, \mu, \theta) - V(x, \mu)$ and θ_i are i.i.d. with distribution m on Θ.

We take a cylindrical Wiener process W on $L_2(\Theta, \mathrm{m})$ and consider the equation

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

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

[Kotelenez '95, Dorogotsev, Wang '21]

(See [Gess, Kassing, K. '23] for further connection with SDG dynamics)

Stochastic Mean-Field Equation

Applying Itô 's formula to $\langle \varphi, \mu_t \rangle$, we come to the **Stochastic Mean-Field Equation** (SMFE):

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

Stochastic Mean-Field Equation

Applying Itô 's formula to $\langle \varphi, \mu_t \rangle$, we come to the **Stochastic Mean-Field Equation** (SMFE):

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

For comparison:

$$d\mu_t = \frac{\alpha}{2}\nabla^2 : (A(\cdot, \mu_t)\mu_t)dt - \nabla \cdot (V(\cdot, \mu_t)\mu_t)dt + \nabla \cdot \sqrt{\alpha} dW^{\mathrm{cor}}(\cdot, t),$$

with $[dW^{\mathrm{cor}}(x,t), dW^{\mathrm{cor}}(y,t)] = A(x, y, \mu_t)\mu_t(x)\mu_t(y)$ and $A = \mathbb{E}_m G \otimes G$.
[Rotskoff, Vanden-Eijnden, CPAM, 2022]

Stochastic Mean-Field Equation

Applying Itô 's formula to $\langle \varphi, \mu_t \rangle$, we come to the **Stochastic Mean-Field Equation** (SMFE):

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

For comparison:

$$d\mu_t = \frac{\alpha}{2}\nabla^2 : (A(\cdot,\mu_t)\mu_t)dt - \nabla \cdot (V(\cdot,\mu_t)\mu_t)dt + \nabla \cdot \sqrt{\alpha} dW^{\mathrm{cor}}(\cdot, t),$$

with $[dW^{\mathrm{cor}}(x,t), dW^{\mathrm{cor}}(y,t)] = A(x,y,\mu_t)\mu_t(x)\mu_t(y)$ and $A = \mathbb{E}_m G \otimes G$.
[Rotskoff, Vanden-Eijnden, CPAM, 2022]

$\rightsquigarrow$ Both solutions satisfy the same martingale problem!

Table of Contents

1. Derivation of SPDE and application in Machine Learning

2. Well-posedness, superposition principle and connection with SGD dynamics

Related Works to SMFE

$$d\mu_t = \frac{1}{2}\nabla^2 : (A(\cdot, \mu_t)\mu_t)\, dt - \nabla \cdot (V(\cdot, \mu_t)\mu_t)\, dt - \nabla \cdot \int_\Theta (G(\cdot, \mu_t, \theta)\mu_t)\, W(d\theta, dt),$$

Well-posedness results for similar SPDEs:

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

Related Works to SMFE

$$d\mu_t = \frac{1}{2}\nabla^2 : (A(\cdot,\mu_t)\mu_t)\,dt - \nabla\cdot(V(\cdot,\mu_t)\mu_t)\,dt - \nabla\cdot\int_\Theta (G(\cdot,\mu_t,\theta)\mu_t)\,W(d\theta,dt),$$

Well-posedness results for similar SPDEs:

- **Continuity equation in the fluid dynamics and optimal transportation** [Ambrosio, Trevisan, Crippa. . .]. There $A = G = 0$.
- **Stochastic nonlinear Fokker-Planck equation** [Coghi, Gess '19]. The covariance A has more general structure but the noise is finite-dimensional.

Related Works to SMFE

$$d\mu_t = \frac{1}{2}\nabla^2 : (A(\cdot,\mu_t)\mu_t)\,dt - \nabla\cdot(V(\cdot,\mu_t)\mu_t)\,dt - \nabla\cdot\int_\Theta (G(\cdot,\mu_t,\theta)\mu_t)\,W(d\theta,dt),$$

Well-posedness results for similar SPDEs:

- **Continuity equation in the fluid dynamics and optimal transportation** [Ambrosio, Trevisan, Crippa. . .]. There $A = G = 0$.
- **Stochastic nonlinear Fokker-Planck equation** [Coghi, Gess '19]. The covariance A has more general structure but the noise is finite-dimensional.
- **Particle representations for a class of nonlinear SPDEs** [Kurtz, Xiong '99]. The equation has more general form but the initial condition μ_0 must have an L_2-density w.r.t. the Lebesgue measure.

Well-posedness of SMFE

Theorem (Gess, Gvalani, K. 2022)

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

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

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

$$dX(u,t) = V(X(u,t),\mu_t)dt + \sqrt{\alpha}\int_\Theta G(X(u,t),\mu_t,\theta)W(d\theta,dt), \quad X(u,0) = u.$$

Convergence to deterministic PDE

Theorem (Gess, Gvalani, K. 2022)

Let $\mu^{n,\frac{1}{n}}$ be superposition solutions to the SMFE ($\alpha = \frac{1}{n}$)

$$d\mu_t = \frac{1}{2n}\nabla^2 : (A(\cdot,\mu_t)\mu_t)\,dt - \nabla\cdot(V(\cdot,\mu_t)\mu_t)\,dt \\ - \frac{1}{\sqrt{n}}\nabla\cdot\int_\Theta G(\cdot,\mu_t,\theta)\mu_t W(d\theta,dt),$$

started from $\mu_0^{n,\frac{1}{n}} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ with $x_i \sim \mu_0$ i.i.d. Then

$$\mathbb{E}\sup_{t\in[0,T]} \mathcal{W}_2^2(\mu_t^{n,\frac{1}{n}},\mu_t^0) \leq Cn^{-1},$$

and $d\mu_t^0 = -\nabla\left(V(\cdot,\mu_t^0)\mu_t^0\right)dt$.

Quantified CLT for SMFE

Since $\mu_t^{n,\frac{1}{n}} = \mu_t^0 + O(n^{-1/2})$, we consider

$$\eta_t^n = \sqrt{n}\left(\mu^{n,\frac{1}{n}} - \mu^0\right).$$

Theorem (Gess, Gvalani, K. 2022)

There exists the Gaussian fluctuation field η, which is a solution to the linear SPDE

$$\begin{aligned} d\eta_t = -\nabla \cdot \left(V(\cdot,\mu_t^0)\eta_t + \langle \tilde{V}(x,\cdot),\eta_t\rangle \mu_t^0(dx)\right) dt \\ - \nabla \cdot \int_\Theta G(\cdot,\mu_t^0,\theta)\mu_t^0 W(d\theta,dt). \end{aligned}$$

Moreover,

$$\mathbb{E} \sup_{t\in[0,T]} \|\eta_t^n - \eta_t\|_{H^{-J}}^2 \leq Cn^{-1}.$$

Higher order approximation of the SGD dynamics

The quantified CLT gives us that

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

Higher order approximation of the SGD dynamics

The quantified CLT gives us that

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

The empirical distribution of SGD with n parameters and learning rate $\alpha = \frac{1}{n}$ satisfies

$$\nu_t^{n,\frac{1}{n}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i(\lfloor nt \rfloor)} = \mu_t^0 + n^{-1/2}\eta_t + o(n^{-1/2})$$

[Sirignano, Spiliopoulos, SPA, 2020]

Higher order approximation of the SGD dynamics

The quantified CLT gives us that

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

The empirical distribution of SGD with n parameters and learning rate $\alpha = \frac{1}{n}$ satisfies

$$\nu_t^{n,\frac{1}{n}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i(\lfloor nt \rfloor)} = \mu_t^0 + n^{-1/2}\eta_t + o(n^{-1/2})$$

[Sirignano, Spiliopoulos, SPA, 2020]

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

Higher order approximation of the SGD dynamics

Theorem (Gess, Gvalani, K. 2022)

Let $\mu^{n,\frac{1}{n}}$ be a superposition solution to the SMFE with learning rate $\alpha = \frac{1}{n}$ started from $\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$. Let also $\nu^{n,\frac{1}{n}}$ be the empirical process associated to the SGD dynamics with $\alpha = \frac{1}{n}$. Then

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

for all $p \in [1, 2)$.

Higher order approximation of the SGD dynamics

Theorem (Gess, Gvalani, K. 2022)

Let $\mu^{n,\frac{1}{n}}$ be a superposition solution to the SMFE with learning rate $\alpha = \frac{1}{n}$ started from $\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$. Let also $\nu^{n,\frac{1}{n}}$ be the empirical process associated to the SGD dynamics with $\alpha = \frac{1}{n}$. Then

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

for all $p \in [1,2)$.

Remark. The SMFE

$$d\mu_t = \frac{1}{2}\nabla^2 : (A(\cdot,\mu_t)\mu_t)\,dt - \nabla\cdot(V(\cdot,\mu_t)\mu_t)\,dt - \nabla\cdot\int_\Theta G(\cdot,\mu_t,\theta)\mu_t W(d\theta,dt)$$

captures the fluctuations of the SGD dynamics. Therefore, it gives a better approximation of the SGD dynamics than

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

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!