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

Vitalii Konarovskyi

University of Hamburg

Conference on Mathematics of Machine Learning 2025

joint work with Benjamin Gess, Rishabh Gvalani and Sebastian Kassing

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

National Academy of Sciences of Ukraine
INSTITUTE OF MATHEMATICS

Supervised Learning and SGD Dynamics

Give some data $\{(\xi_i, \gamma_i),\ i \in I\}$, the goal is to predict a new label γ given a new input ξ.

Supervised Learning and SGD Dynamics

Give some data $\{(\xi_i, \gamma_i),\ i \in I\}$, the goal is to predict a new label γ given a new input ξ.

For simplicity assume that $\xi_i \sim P$ and $\gamma_i = f(\xi_i)$.

Supervised Learning and SGD Dynamics

Give some data $\{(\xi_i, \gamma_i),\ i \in I\}$, the goal is to predict a new label γ given a new input ξ.

For simplicity assume that $\xi_i \sim P$ and $\gamma_i = f(\xi_i)$.

The output $\gamma = f(\xi)$ is approximated by $f_z(\xi)$, where $z \in \mathbb{R}^d$ are parameters that have to be learned from

$$R(z) := \mathbb{E}_P\left[l\left(f(\xi), f_z(\xi)\right)\right] = \mathbb{E}_P\left[R(z, \xi)\right] \to \min$$

Supervised Learning and SGD Dynamics

Give some data $\{(\xi_i, \gamma_i),\ i \in I\}$, the goal is to predict a new label γ given a new input ξ.

For simplicity assume that $\xi_i \sim P$ and $\gamma_i = f(\xi_i)$.

The output $\gamma = f(\xi)$ is approximated by $f_z(\xi)$, where $z \in \mathbb{R}^d$ are parameters that have to be learned from

$$R(z) := \mathbb{E}_P\left[l\left(f(\xi), f_z(\xi)\right)\right] = \mathbb{E}_P\left[R(z, \xi)\right] \to \min$$

Stochastic Gradient Descent:

taking $z(0) \in \mathbb{R}^d$ define

$$z(t_{i+1}) = z(t_i) - \alpha \nabla_z R(z(t_{i+1}), \xi_i)$$

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

SGD and ODE

For $\alpha \to 0$ the SDG dynamics

$$z(t_{i+1}) = z(t_i) - \alpha \nabla_z R(z(t_{i+1}), \xi_i)$$

behaves as a solution to ODE

$$dZ_t = -\nabla_z R(Z_t)dt,$$

where

$$R(z) = \mathbb{E}_P R(z, \xi).$$

SGD and ODE

For $\alpha \to 0$ the SDG dynamics

$$z(t_{i+1}) = z(t_i) - \alpha \nabla_z R(z(t_{i+1}), \xi_i)$$

behaves as a solution to ODE

$$dZ_t = -\nabla_z R(Z_t) dt,$$

where

$$R(z) = \mathbb{E}_P R(z, \xi).$$

Theorem [see e.g. in Li, Tai, E '19, JMLR]

For f and R smooth enough one has

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

SGD and SDE

The ODE loses all information about the randomness in the SGD.

$$
\begin{aligned}
z(t_{i+1}) &= z(t_i) - \alpha \nabla R(z(t_i), \xi_i) \\
&= z(t_i) - \underbrace{\nabla \mathbb{E}_P R(\dots)}_{\nabla R(z(t_i))} \alpha + \underbrace{\sqrt{\alpha}}_{=\sqrt{\alpha}} \underbrace{\left(\nabla \mathbb{E}_P R(\dots) - \nabla R(z(t_i), \xi_i)\right)}_{=G(z(t_i), \xi_i)} \sqrt{\alpha}
\end{aligned}
$$

SGD and SDE

The ODE loses all information about the randomness in the SGD.

$$\begin{aligned} z(t_{i+1}) &= z(t_i) - \alpha \nabla R(z(t_i), \xi_i) \\ &= z(t_i) - \underbrace{\nabla \mathbb{E}_P R(\dots)}_{\nabla R(z(t_i))} \alpha + \underbrace{\sqrt{\alpha}}_{=\sqrt{\alpha}} \underbrace{\left(\nabla \mathbb{E}_P R(\dots) - \nabla R(z(t_i), \xi_i)\right)}_{=G(z(t_i),\xi_i)} \sqrt{\alpha} \end{aligned}$$

is the Euler scheme for the SDE

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

where $\Sigma(z) = \mathrm{Cov}\left(\nabla R(z, \xi), \nabla R(z, \xi)\right) = \mathbb{E}_P G(z, \xi) \otimes G(z, \xi)$.

SGD and SDE

The ODE loses all information about the randomness in the SGD.

$$\begin{aligned} z(t_{i+1}) &= z(t_i) - \alpha \nabla R(z(t_i), \xi_i) \\ &= z(t_i) - \underbrace{\nabla \mathbb{E}_P R(\dots)}_{\nabla R(z(t_i))} \alpha + \underbrace{\sqrt{\alpha}}_{=\sqrt{\alpha}} \underbrace{\left(\nabla \mathbb{E}_P R(\dots) - \nabla R(z(t_i), \xi_i)\right)}_{=G(z(t_i),\xi_i)} \sqrt{\alpha} \end{aligned}$$

is the Euler scheme for the SDE

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

where $\Sigma(z) = \mathrm{Cov}\left(\nabla R(z, \xi), \nabla R(z, \xi)\right) = \mathbb{E}_P G(z, \xi) \otimes G(z, \xi)$.

Theorem [Li, Tai, E '19, JMLR]

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

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

SGD and SDE

The ODE loses all information about the randomness in the SGD.

$$\begin{aligned} z(t_{i+1}) &= z(t_i) - \alpha \nabla R(z(t_i), \xi_i) \\ &= z(t_i) - \underbrace{\nabla \mathbb{E}_P R(\dots)}_{\nabla R(z(t_i))} \alpha + \underbrace{\sqrt{\alpha}}_{=\sqrt{\alpha}} \underbrace{\left(\nabla \mathbb{E}_P R(\dots) - \nabla R(z(t_i), \xi_i)\right)}_{=G(z(t_i),\xi_i)} \sqrt{\alpha} \end{aligned}$$

is the Euler scheme for the SDE

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

where $\Sigma(z) = \mathrm{Cov}\left(\nabla R(z,\xi), \nabla R(z,\xi)\right) = \mathbb{E}_P G(z,\xi) \otimes G(z,\xi)$.

Theorem [Li, Tai, E '19, JMLR]

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

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

Limitations of Modified SDE

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

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

Limitations of Modified SDE

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

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

The SDE does not catch n-point motion:

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

$$z_1(t_{i+1}) = z_1(t_i) - \alpha \nabla R(z_1(t_i), \xi_i)$$

$$\dots$$

$$z_n(t_{i+1}) = z_n(t_i) - \alpha \nabla R(z_n(t_i), \xi_i)$$

Limitations of Modified SDE

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

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

The SDE does not catch n-point motion:

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

$$z_1(t_{i+1}) = z_1(t_i) - \alpha \nabla R(z_1(t_i), \xi_i)$$
$$\dots$$
$$z_n(t_{i+1}) = z_n(t_i) - \alpha \nabla R(z_n(t_i), \xi_i)$$

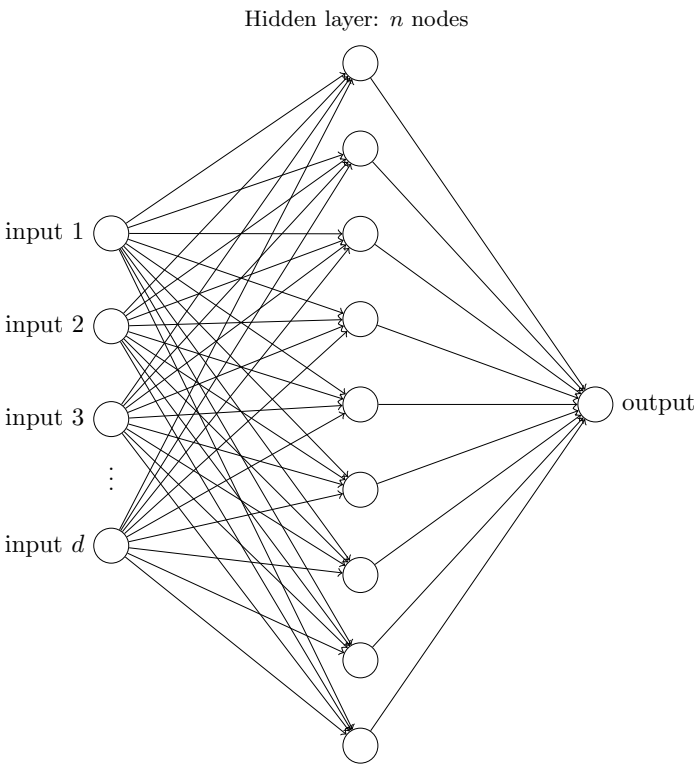
Then

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

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

$$\mathrm{Cov}\left(z_k, z_l\right) = \mathrm{Cov}\left(\nabla R(z_k, \xi), \nabla R(z_l, \xi)\right) \neq \Sigma^{\frac{1}{2}}\left(Z^k\right)\Sigma^{\frac{1}{2}}\left(Z^l\right) = \mathrm{Cov}\left(Z^k, Z^l\right)$$

Neural Network with One Hidden Layer



Network with a single hidden layer:

$$f_z(\xi) = \frac{1}{n}\sum_{k=1}^{n} \Phi_{z_k}(\xi)$$

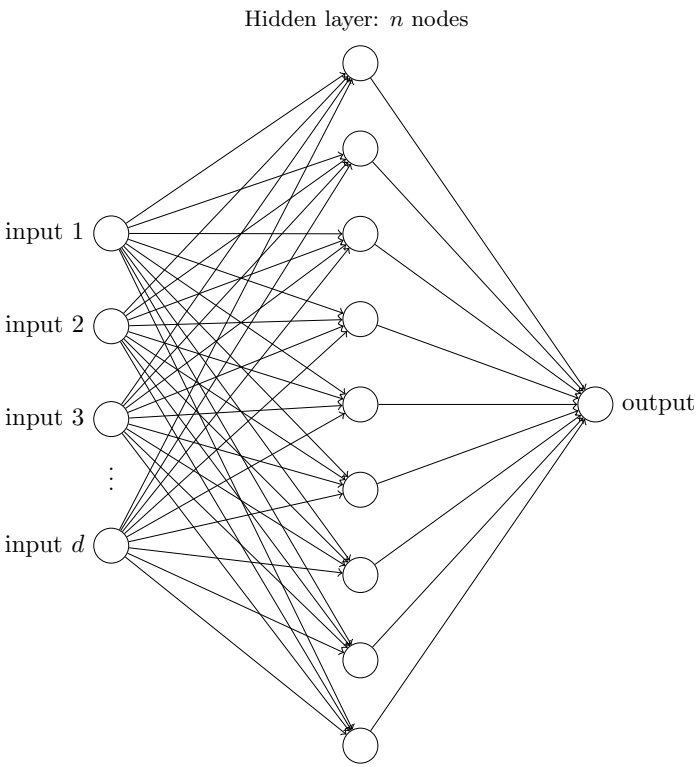
$$= \int \Phi_z(\xi)\mu^n(dz) = \langle \Phi(\xi), \mu^n \rangle ,$$

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

$$\mu^n = \frac{1}{n}\sum_{k=1}^{n} \delta_{z_k}$$

[Chizat, Bach, Mei, Nguye, Rotskoff, Sirignano, Vanden-Eijnden...]

Neural Network with One Hidden Layer



Network with a single hidden layer:

$$f_z(\xi) = \frac{1}{n}\sum_{k=1}^{n} \Phi_{z_k}(\xi)$$

$$= \int \Phi_z(\xi)\mu^n(dz) = \langle \Phi(\xi), \mu^n \rangle ,$$

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

$$\mu^n = \frac{1}{n}\sum_{k=1}^{n} \delta_{z_k}$$

[Chizat, Bach, Mei, Nguye, Rotskoff, Sirignano, Vanden-Eijnden...]

Risk Function: for $\mathbf{z} = (z_1, \ldots, z_n)$

$$\tilde{R}(\mathbf{z}) := \frac{1}{2}\mathbb{E}_P\left[\left(f(\xi) - f_{\mathbf{z}}(\xi)\right)|^2\right] =: \mathbb{E}\left[\tilde{R}(z_1, \ldots, z_n, \xi)\right] \to \min$$

SGD for Shallow Network

Taking $z_k(0) \sim \mu_0$ i.i.d. and $\xi_i \sim P$ i.i.d. we consider the SGD

$$\mathbf{z}(t_{i+1}) = \mathbf{z}(t_i) - \alpha \nabla_{\mathbf{z}} \tilde{R}\left(\mathbf{z}(t_i), \xi_i\right).$$

SGD for Shallow Network

Taking $z_k(0) \sim \mu_0$ i.i.d. and $\xi_i \sim P$ i.i.d. we consider the SGD

$$\mathbf{z}(t_{i+1}) = \mathbf{z}(t_i) - \alpha \nabla_{\mathbf{z}} \tilde{R}\left(\mathbf{z}(t_i), \xi_i\right).$$

A simple computation shows

$$\begin{aligned} z_1(t_{i+1}) &= z_1(t_i) - \alpha \nabla_{z_1} R\left(z_1(t_i), \mu^n_{t_i}, \xi_i\right) \\ &\dots \\ z_n(t_{i+1}) &= z_n(t_i) - \alpha \nabla_{z_n} R\left(z_n(t_i), \mu^n_{t_i}, \xi_i\right) \end{aligned}$$

SGD for Shallow Network

Taking $z_k(0) \sim \mu_0$ i.i.d. and $\xi_i \sim P$ i.i.d. we consider the SGD

$$\mathbf{z}(t_{i+1}) = \mathbf{z}(t_i) - \alpha \nabla_{\mathbf{z}} \tilde{R}\left(\mathbf{z}(t_i), \xi_i\right).$$

A simple computation shows

$$\begin{aligned} z_1(t_{i+1}) &= z_1(t_i) - \alpha \nabla_{z_1} R\left(z_1(t_i), \mu^n_{t_i}, \xi_i\right) \\ &\dots \\ z_n(t_{i+1}) &= z_n(t_i) - \alpha \nabla_{z_n} R\left(z_n(t_i), \mu^n_{t_i}, \xi_i\right) \end{aligned}$$

We need an SDE that would capture the dynamics of $\mathbf{z} = (z_1, \dots, z_n)$ or μ^n.

SGD for Shallow Network

Taking $z_k(0) \sim \mu_0$ i.i.d. and $\xi_i \sim P$ i.i.d. we consider the SGD

$$\mathbf{z}(t_{i+1}) = \mathbf{z}(t_i) - \alpha \nabla_{\mathbf{z}} \tilde{R}\left(\mathbf{z}(t_i), \xi_i\right).$$

A simple computation shows

$$\begin{aligned} z_1(t_{i+1}) &= z_1(t_i) - \alpha \nabla_{z_1} R\left(z_1(t_i), \mu^n_{t_i}, \xi_i\right) \\ &\dots \\ z_n(t_{i+1}) &= z_n(t_i) - \alpha \nabla_{z_n} R\left(z_n(t_i), \mu^n_{t_i}, \xi_i\right) \end{aligned}$$

We need an SDE that would capture the dynamics of $\mathbf{z} = (z_1, \dots, z_n)$ or μ^n.

Naively one can use:

$$d\mathbf{Z}_t = -\nabla \tilde{R}(\mathbf{Z}_t)dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(\mathbf{Z}_t)dw_t$$

SGD for Shallow Network

Taking $z_k(0) \sim \mu_0$ i.i.d. and $\xi_i \sim P$ i.i.d. we consider the SGD

$$\mathbf{z}(t_{i+1}) = \mathbf{z}(t_i) - \alpha \nabla_{\mathbf{z}} \tilde{R}\left(\mathbf{z}(t_i), \xi_i\right).$$

A simple computation shows

$$\begin{aligned} z_1(t_{i+1}) &= z_1(t_i) - \alpha \nabla_{z_1} R\left(z_1(t_i), \mu^n_{t_i}, \xi_i\right) \\ &\dots \\ z_n(t_{i+1}) &= z_n(t_i) - \alpha \nabla_{z_n} R\left(z_n(t_i), \mu^n_{t_i}, \xi_i\right) \end{aligned}$$

We need an SDE that would capture the dynamics of $\mathbf{z} = (z_1, \dots, z_n)$ or μ^n.

Naively one can use:

$$d\mathbf{Z}_t = -\nabla \tilde{R}(\mathbf{Z}_t)dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(\mathbf{Z}_t)dw_t$$

The coefficients depend on n!

Mean-Field ODE (Continuity Equation)

$$z_1(t_{i+1}) = z_1(t_i) - \alpha \nabla_{z_1} R\left(z_1(t_i), \mu^n_{t_i}, \xi_i\right)$$
$$\dots$$
$$z_n(t_{i+1}) = z_n(t_i) - \alpha \nabla_{z_n} R\left(z_n(t_i), \mu^n_{t_i}, \xi_i\right)$$

can be captured for $\alpha \to 0$ by

$$dZ^k_t = -\nabla_{Z^k} R\left(Z^k_t, \nu^n_t\right) dt,$$

with $\nu^n_t = \frac{1}{n}\sum_{k=1}^n \delta_{Z^k_t}$ and $R(z, \nu) = \mathbb{E}\left[R(z, \nu, \xi)\right]$.

Mean-Field ODE (Continuity Equation)

$$z_1(t_{i+1}) = z_1(t_i) - \alpha \nabla_{z_1} R\left(z_1(t_i), \mu^n_{t_i}, \xi_i\right)$$
$$\dots$$
$$z_n(t_{i+1}) = z_n(t_i) - \alpha \nabla_{z_n} R\left(z_n(t_i), \mu^n_{t_i}, \xi_i\right)$$

can be captured for $\alpha \to 0$ by

$$dZ^k_t = -\nabla_{Z^k} R\left(Z^k_t, \nu^n_t\right) dt,$$

with $\nu^n_t = \frac{1}{n}\sum_{k=1}^n \delta_{Z^k_t}$ and $R(z,\nu) = \mathbb{E}\left[R(z,\nu,\xi)\right]$.

One can write the equation for ν^n_t

$$d\nu_t = \nabla\left(\nabla_Z R(\cdot,\nu_t)\nu_t\right) dt$$
$$\nu_0 = \nu^n_0 = \frac{1}{n}\sum_{k=1}^n \delta_{z_k(0)}$$

Mean-Field ODE (Continuity Equation)

$$z_1(t_{i+1}) = z_1(t_i) - \alpha \nabla_{z_1} R\left(z_1(t_i), \mu^n_{t_i}, \xi_i\right)$$
$$\dots$$
$$z_n(t_{i+1}) = z_n(t_i) - \alpha \nabla_{z_n} R\left(z_n(t_i), \mu^n_{t_i}, \xi_i\right)$$

can be captured for $\alpha \to 0$ by

$$dZ^k_t = -\nabla_{Z^k} R\left(Z^k_t, \nu^n_t\right) dt,$$

with $\nu^n_t = \frac{1}{n}\sum_{k=1}^n \delta_{Z^k_t}$ and $R(z,\nu) = \mathbb{E}\left[R(z,\nu,\xi)\right]$.

One can write the equation for ν^n_t

$$d\nu_t = \nabla\left(\nabla_Z R(\cdot,\nu_t)\nu_t\right) dt \qquad \rightsquigarrow \qquad dZ_t(u) = -\nabla_Z R\left(Z_t(u),\nu_t\right) dt$$
$$\nu_0 = \nu^n_0 = \frac{1}{n}\sum_{k=1}^n \delta_{z_k(0)} \qquad\qquad Z_0(u) = u, \quad \nu_t = \nu^n_0 \circ Z_t^{-1}$$

Mean-Field ODE (Continuity Equation)

$$z_1(t_{i+1}) = z_1(t_i) - \alpha \nabla_{z_1} R\left(z_1(t_i), \mu^n_{t_i}, \xi_i\right)$$
$$\dots$$
$$z_n(t_{i+1}) = z_n(t_i) - \alpha \nabla_{z_n} R\left(z_n(t_i), \mu^n_{t_i}, \xi_i\right)$$

can be captured for $\alpha \to 0$ by

$$dZ^k_t = -\nabla_{Z^k} R\left(Z^k_t, \nu^n_t\right) dt,$$

with $\nu^n_t = \frac{1}{n}\sum_{k=1}^n \delta_{Z^k_t}$ and $R(z,\nu) = \mathbb{E}\left[R(z,\nu,\xi)\right]$.

One can write the equation for ν^n_t

$$d\nu_t = \nabla\left(\nabla_Z R(\cdot,\nu_t)\nu_t\right) dt \qquad \rightsquigarrow \qquad dZ_t(u) = -\nabla_Z R\left(Z_t(u),\nu_t\right) dt$$
$$\nu_0 = \nu^n_0 = \frac{1}{n}\sum_{k=1}^n \delta_{z_k(0)} \qquad\qquad Z_0(u) = u, \quad \nu_t = \nu^n_0 \circ Z_t^{-1}$$

Then $Z^k_t = Z_t(z_k(0))$ and $\nu^n_t = \nu_t$.

Martingale Problem for Empirical distribution

Recall

$$d\mathbf{Z}_t = -\nabla \tilde{R}(\mathbf{Z}_t)dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(\mathbf{Z}_t)dw_t$$

captures the motion of

$$z_k(t_{i+1}) = z_k(t_i) - \alpha \nabla_{z_k} R\left(z_k(t_i), \mu^n_{t_i}, \xi_i\right), \quad k = 1, \ldots, n.$$

Martingale Problem for Empirical distribution

Recall

$$d\mathbf{Z}_t = -\nabla \tilde{R}(\mathbf{Z}_t)dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(\mathbf{Z}_t)dw_t$$

captures the motion of

$$z_k(t_{i+1}) = z_k(t_i) - \alpha \nabla_{z_k} R\left(z_k(t_i), \mu^n_{t_i}, \xi_i\right), \quad k = 1, \dots, n.$$

Taking $\varphi \in C^2_c(\mathbb{R}^d)$, we get for the empirical measure $\nu^n_t = \frac{1}{n}\sum_{k=1}^n \delta_{Z^k_t}$

$$\langle \varphi, \nu^n_t \rangle = \langle \varphi, \nu^n_0 \rangle - \int_0^t \langle \nabla\varphi \cdot \nabla R(\cdot, \nu^n_s), \nu^n_s \rangle \, ds + \frac{\alpha}{2}\int_0^t \left\langle \nabla^2\varphi : A(\cdot, \nu^n_s), \nu^n_s \right\rangle ds + \text{Mart.},$$

with

$$\frac{d}{dt}\left[\text{Mart.}\right]_t = \alpha \int\int (\nabla\varphi(x) \otimes \nabla\varphi(y)) : \tilde{A}(x, y, \nu^n_t)\nu^n_t(dx)\nu^n_t(dy)$$

where $\tilde{A}(z, y, \mu) = \mathrm{Cov}\left(\nabla R(z, \mu, \xi), \nabla R(z, \mu, \xi)\right)$ and $A(z, \mu) = A(z, z, \mu)$

[Rotskoff, Vanden-Eijnden, CPAM, 2022]

Martingale Problem for Empirical distribution

Recall

$$d\mathbf{Z}_t = -\nabla \tilde{R}(\mathbf{Z}_t)dt + \sqrt{\alpha}\Sigma^{\frac{1}{2}}(\mathbf{Z}_t)dw_t$$

captures the motion of

$$z_k(t_{i+1}) = z_k(t_i) - \alpha \nabla_{z_k} R\left(z_k(t_i), \mu^n_{t_i}, \xi_i\right), \quad k = 1, \dots, n.$$

Taking $\varphi \in C^2_c(\mathbb{R}^d)$, we get for the empirical measure $\nu^n_t = \frac{1}{n}\sum_{k=1}^n \delta_{Z^k_t}$

$$\langle \varphi, \nu^n_t \rangle = \langle \varphi, \nu^n_0 \rangle - \int_0^t \langle \nabla\varphi \cdot \nabla R(\cdot, \nu^n_s), \nu^n_s \rangle \, ds + \frac{\alpha}{2}\int_0^t \left\langle \nabla^2\varphi : A(\cdot, \nu^n_s), \nu^n_s \right\rangle ds + \text{Mart.},$$

with

$$\frac{d}{dt}\left[\text{Mart.}\right]_t = \alpha \int\int (\nabla\varphi(x) \otimes \nabla\varphi(y)) : \tilde{A}(x, y, \nu^n_t)\nu^n_t(dx)\nu^n_t(dy)$$

where $\tilde{A}(z, y, \mu) = \mathrm{Cov}\left(\nabla R(z, \mu, \xi), \nabla R(z, \mu, \xi)\right)$ and $A(z, \mu) = A(z, z, \mu)$

[Rotskoff, Vanden-Eijnden, CPAM, 2022]

$\rightsquigarrow$ Complicated to work

SDE with Inf-Dim Noise

We repeat the derivation of the equation for SGD with new idea:

$$
\begin{aligned}
z_k(t_{i+1}) &= z_k(t_i) - \alpha \nabla R(z_k(t_i), \mu^n_{t_i}, \xi_i) \\
&= z_k(t_i) - \alpha \underbrace{\mathbb{E}\nabla R(\dots)}_{=\nabla R(z_k(t_i), \mu^n_{t_i})} + \sqrt{\alpha} \underbrace{(\mathbb{E}\nabla R(\dots) - \nabla R(\dots))}_{=G(z_k(t_i), \mu^n_{t_i}, \xi_i)} \sqrt{\alpha}
\end{aligned}
$$

SDE with Inf-Dim Noise

We repeat the derivation of the equation for SGD with new idea:

$$
\begin{aligned}
z_k(t_{i+1}) &= z_k(t_i) - \alpha \nabla R(z_k(t_i), \mu^n_{t_i}, \xi_i) \\
&= z_k(t_i) - \alpha \underbrace{\mathbb{E}\nabla R(\dots)}_{=\nabla R(z_k(t_i), \mu^n_{t_i})} + \sqrt{\alpha} \underbrace{(\mathbb{E}\nabla R(\dots) - \nabla R(\dots))}_{=G(z_k(t_i), \mu^n_{t_i}, \xi_i)} \sqrt{\alpha}
\end{aligned}
$$

Note that $\mathrm{Cov}\,(z_k, z_l) = \mathbb{E}\,[G(z_k, \mu, \xi) \otimes G(z_l, \mu, \xi)]$

SDE with Inf-Dim Noise

We repeat the derivation of the equation for SGD with new idea:

$$\begin{aligned} z_k(t_{i+1}) &= z_k(t_i) - \alpha \nabla R(z_k(t_i), \mu^n_{t_i}, \xi_i) \\ &= z_k(t_i) - \alpha \underbrace{\mathbb{E}\nabla R(\dots)}_{=\nabla R(z_k(t_i),\mu^n_{t_i})} + \sqrt{\alpha} \underbrace{(\mathbb{E}\nabla R(\dots) - \nabla R(\dots))}_{=G(z_k(t_i),\mu^n_{t_i},\xi_i)} \sqrt{\alpha} \end{aligned}$$

Note that $\mathrm{Cov}\,(z_k, z_l) = \mathbb{E}\,[G(z_k, \mu, \xi) \otimes G(z_l, \mu, \xi)]$

We get

$$dZ^k_t = -\nabla R(Z^k_t, \nu^n_t)dt + \sqrt{\alpha} \underbrace{\int_\Xi G(Z^k_t, \nu^n_t, \xi) W(d\xi, dt)}_{=\sum_l (G, e_l)_{L_2} dw_l(t)}, \quad k = 1, \dots, n,$$

where $\mu^n_t = \frac{1}{n}\sum_{i=1}^n \delta_{Z^k}$, and $W(\xi, t) = \sum_l e_l(\xi) w_l(t)$ is white noise on $L_2(\Xi, P)$

($P = \mathrm{Law}\xi$, Ξ is the state space for ξ) .

Distribution Dependent Stochastic Flow

Distribution Dependent Stochastic Flow:

$$dZ_t(u) = -\nabla R(Z_t(u), \nu_t)dt + \sqrt{\alpha} \int_{\Xi} G(Z_t(u), \nu_t, \xi) W(d\xi, dt)$$

$$Z_0(u) = u, \quad \nu_t = \nu_0 \circ Z_t^{-1}$$

[Dorogovtsev, Kotelenez, Pilipenko, F-Y. Wang,...] see also McKean–Vlasov equation with common noise (no independent noise)

Distribution Dependent Stochastic Flow

Distribution Dependent Stochastic Flow:

$$dZ_t(u) = -\nabla R(Z_t(u), \nu_t)dt + \sqrt{\alpha}\int_{\Xi} G(Z_t(u), \nu_t, \xi)W(d\xi, dt)$$

$$Z_0(u) = u, \quad \nu_t = \nu_0 \circ Z_t^{-1}$$

[Dorogovtsev, Kotelenez, Pilipenko, F-Y. Wang,...] see also McKean–Vlasov equation with common noise (no independent noise)

Using Itô 's formula, we come to the

Stochastic Mean-Field Equation:

$$d\nu_t = \nabla \cdot (\nabla R(\cdot, \nu_t)\nu_t)dt + \frac{\alpha}{2}\nabla^2 : (A(\cdot, \nu_t)\nu_t)dt$$

$$+\sqrt{\alpha}\nabla \cdot \int_{\Xi} G(\cdot, \nu_t, \xi)\mu_t \, W(d\xi, dt)$$

where $A(z, \mu) = \mathbb{E}G(z, \mu, \xi) \otimes G(z, \mu, \xi)$.

Distribution Dependent Stochastic Flow

Distribution Dependent Stochastic Flow:

$$dZ_t(u) = -\nabla R(Z_t(u), \nu_t)dt + \sqrt{\alpha}\int_{\Xi} G(Z_t(u), \nu_t, \xi)W(d\xi, dt)$$

$$Z_0(u) = u, \quad \nu_t = \nu_0 \circ Z_t^{-1}$$

[Dorogovtsev, Kotelenez, Pilipenko, F-Y. Wang,...] see also McKean–Vlasov equation with common noise (no independent noise)

Using Itô 's formula, we come to the

Stochastic Mean-Field Equation:

$$d\nu_t = \nabla \cdot (\nabla R(\cdot, \nu_t)\nu_t)dt + \frac{\alpha}{2}\nabla^2 : (A(\cdot, \nu_t)\nu_t)dt$$

$$+\sqrt{\alpha}\nabla \cdot \int_{\Xi} G(\cdot, \nu_t, \xi)\mu_t \, W(d\xi, dt)$$

where $A(z, \mu) = \mathbb{E}G(z, \mu, \xi) \otimes G(z, \mu, \xi)$.

Well-posedness is obtained in [Gess, Gvalani, K. '25, PTRF]

Distribution Dependent Stochastic Flow

Distribution Dependent Stochastic Flow:

$$dZ_t(u) = -\nabla R(Z_t(u), \nu_t)dt + \sqrt{\alpha}\int_{\Xi} G(Z_t(u), \nu_t, \xi)W(d\xi, dt)$$

$$Z_0(u) = u, \quad \nu_t = \nu_0 \circ Z_t^{-1}$$

[Dorogovtsev, Kotelenez, Pilipenko, F-Y. Wang,...] see also McKean–Vlasov equation with common noise (no independent noise)

Using Itô 's formula, we come to the

Stochastic Mean-Field Equation:

$$d\nu_t = \nabla \cdot (\nabla R(\cdot, \nu_t)\nu_t)dt + \frac{\alpha}{2}\nabla^2 : (A(\cdot, \nu_t)\nu_t)dt$$

$$+\sqrt{\alpha}\nabla \cdot \int_{\Xi} G(\cdot, \nu_t, \xi)\mu_t \, W(d\xi, dt)$$

where $A(z, \mu) = \mathbb{E}G(z, \mu, \xi) \otimes G(z, \mu, \xi)$.

Well-posedness is obtained in [Gess, Gvalani, K. '25, PTRF]

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

Connection with SGD

Recall that $z_k(0) \sim \mu_0$ – i.i.d., α – learning rate, $t_i = \alpha i$, $\xi_i \sim P$ – i.i.d.

$$z_k(t_{i+1}) = z_k(t_i) - \alpha \nabla R(z_k(t_i), \mu^n_{t_i}, \xi_i), \quad k = 1, \dots, n$$

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

Connection with SGD

Recall that $z_k(0) \sim \mu_0$ – i.i.d., α – learning rate, $t_i = \alpha i$, $\xi_i \sim P$ – i.i.d.

$$z_k(t_{i+1}) = z_k(t_i) - \alpha \nabla R(z_k(t_i), \mu^n_{t_i}, \xi_i), \quad k = 1, \dots, n$$

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

Distribution Dependent Stochastic Flow:

$$dZ_t(u) = -\nabla R(Z_t(u), \nu_t) dtred - \frac{\alpha}{4} \nabla |\nabla R(Z_t(u), \nu_t)|^2 dt - \frac{\alpha}{4} \langle D |\nabla R(Z_t(u), \nu_t)|^2, \nu_t \rangle dt$$

$$+ \sqrt{\alpha} \int_\Theta G(Z_t(u), \nu_t, \xi) W(d\xi, dt),$$

$$Z_0(u) = u, \quad \nu_t = \mu_0 \circ Z_t^{-1},$$

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

Connection with SGD

Recall that $z_k(0) \sim \mu_0$ – i.i.d., α – learning rate, $t_i = \alpha i$, $\xi_i \sim P$ – i.i.d.

$$z_k(t_{i+1}) = z_k(t_i) - \alpha \nabla R(z_k(t_i), \mu^n_{t_i}, \xi_i), \quad k = 1, \dots, n$$

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

Distribution Dependent Stochastic Flow:

$$dZ_t(u) = -\nabla R(Z_t(u), \nu_t) dtred - \frac{\alpha}{4} \nabla |\nabla R(Z_t(u), \nu_t)|^2 dt - \frac{\alpha}{4} \langle D|\nabla R(Z_t(u), \nu_t)|^2, \nu_t \rangle dt$$

$$+ \sqrt{\alpha} \int_\Theta G(Z_t(u), \nu_t, \xi) W(d\xi, dt),$$

$$Z_0(u) = u, \quad \nu_t = \mu_0 \circ Z_t^{-1},$$

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

Theorem [Gess, Kassing, K. '24, JMLR]

Let $\mu_0 \in \mathcal{P}_2$ and $\nabla R(z, \nu, \xi)$ be regular enough in z, ν. Then for every $\Phi \in \mathrm{C}^4_b(\mathcal{P}_2)$

$$\sup_{t_i \le T} |\mathbb{E}\Phi(\nu_{t_i}) - \mathbb{E}\Phi(\mu^n_{t_i})| \le C\alpha red^2 + C\sqrt{\mathbb{E}W_2^2(\mu_0, \mu^n_0)}.$$

Connection with SGD

Recall that $z_k(0) \sim \mu_0$ – i.i.d., α – learning rate, $t_i = \alpha i$, $\xi_i \sim P$ – i.i.d.

$$z_k(t_{i+1}) = z_k(t_i) - \alpha \nabla R(z_k(t_i), \mu^n_{t_i}, \xi_i), \quad k = 1, \dots, n$$

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

Distribution Dependent Stochastic Modified Flow:

$$dZ_t(u) = -\nabla R(Z_t(u), \nu_t)dt - \frac{\alpha}{4} \nabla |\nabla R(Z_t(u), \nu_t)|^2 dt - \frac{\alpha}{4} \langle D|\nabla R(Z_t(u), \nu_t)|^2, \nu_t \rangle dt$$

$$+ \sqrt{\alpha} \int_\Theta G(Z_t(u), \nu_t, \xi) W(d\xi, dt),$$

$$Z_0(u) = u, \quad \nu_t = \mu_0 \circ Z_t^{-1},$$

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

Theorem [Gess, Kassing, K. '24, JMLR]

Let $\mu_0 \in \mathcal{P}_2$ and $\nabla R(z, \nu, \xi)$ be regular enough in z, ν. Then for every $\Phi \in \mathrm{C}^4_b(\mathcal{P}_2)$

$$\sup_{t_i \le T} |\mathbb{E}\Phi(\nu_{t_i}) - \mathbb{E}\Phi(\mu^n_{t_i})| \le C\alpha^2 + C\sqrt{\mathbb{E}W_2^2(\mu_0, \mu^n_0)}.$$

n-point motion for SGD

Recall SGD:

$$z_k(t_{i+1}) = z_k(t_i) - \alpha \nabla R(z_k(t_i), \xi_i), \quad k = 1, \dots, n,$$

started from distinct initializations $z_k(0)$, $k = 1, \dots, n$.

n-point motion for SGD

Recall SGD:

$$z_k(t_{i+1}) = z_k(t_i) - \alpha \nabla R(z_k(t_i), \xi_i), \quad k = 1, \dots, n,$$

started from distinct initializations $z_k(0)$, $k = 1, \dots, n$.

Consider Stochastic Modified Flow:

$$\begin{aligned} dZ_t(u) &= -\nabla R(Z_t(u))dt - \frac{\alpha}{4}\nabla |\nabla R(Z_t(u))|^2 dt \\ &\quad + \sqrt{\alpha} \int_{\Xi} \left(\nabla R(Z_t(u)) - \nabla R(Z_t(u), \xi)\right) W(d\xi, dt), \\ Z_t(u) &= u, \end{aligned}$$

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

n-point motion for SGD

Recall SGD:

$$z_k(t_{i+1}) = z_k(t_i) - \alpha \nabla R(z_k(t_i), \xi_i), \quad k = 1, \dots, n,$$

started from distinct initializations $z_k(0)$, $k = 1, \dots, n$.

Consider Stochastic Modified Flow:

$$\begin{aligned} dZ_t(u) &= -\nabla R(Z_t(u))dt - \frac{\alpha}{4}\nabla |\nabla R(Z_t(u))|^2 dt \\ &\quad + \sqrt{\alpha} \int_{\Xi} \left(\nabla R(Z_t(u)) - \nabla R(Z_t(u), \xi)\right) W(d\xi, dt), \\ Z_t(u) &= u, \end{aligned}$$

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

Theorem [Gess, Kassing, K. '24, JMLR]

Define $Z_t^k := Z_t(x_k(0))$, $k = 1, \dots, n$. Then for every $f \in C_b^4(\mathbb{R}^{dn})$

$$\sup_{t_i \le T} \left| \mathbb{E} f\left(z_1(t_i), \dots, z_n(t_i)\right) - \mathbb{E} f\left(Z_{t_i}^1, \dots, Z_{t_i}^n\right) \right| \le C\alpha^2.$$

Weak Topology vs. Strong Topology

Observe that for a 1-d Brownian motion w the error between $\sqrt{\alpha}w_t$ and 0 can is different in different topologies

$$W_1\left(\mathrm{Law}\left(\sqrt{\alpha}w_t\right), \delta_0\right) = \mathbb{E}\left[\left|\sqrt{\alpha}w_t - 0\right|\right] = O\left(\sqrt{\alpha}\right)$$

Weak Topology vs. Strong Topology

Observe that for a 1-d Brownian motion w the error between $\sqrt{\alpha}w_t$ and 0 can is different in different topologies

$$W_1\left(\mathrm{Law}\left(\sqrt{\alpha}w_t\right), \delta_0\right) = \mathbb{E}\left[\left|\sqrt{\alpha}w_t - 0\right|\right] = O\left(\sqrt{\alpha}\right)$$

but

$$\mathbb{E}\left[f\left(\sqrt{\alpha}w_t\right)\right] - \mathbb{E}\left[f(0)\right] = \mathbb{E}\left[f'(0)\sqrt{\alpha}w_t + \frac{1}{2}f''(\theta)\alpha w_t^2\right] = O\left(\|f''\|\alpha\right)$$

Comparison in Wasserstein (strong) topology

SDE for overparametrized shallow neural network with $z_k(0) \sim \mu_0$ i.i.d.

$$z_k(t_{i+1}) = z_k(t_i) - \alpha \nabla_{z_k} R\left(z_k(t_i), \mu^n_{t_i}, \xi_i\right), \quad k = 1, \dots, n.$$

Consider a solution to continuity equation

$$d\nu_t = \nabla \left(\nabla R(\cdot, \nu_t)\nu_t\right) dtred + \frac{\alpha}{2}\nabla^2 : (A(\cdot, \nu_t)\nu_t)dt + \sqrt{\alpha}\int_\Xi G(Z_t(u), \nu_t, \xi) W(d\xi, dt)$$

with initial particle distributions μ_0, i.e. $\nu_t = \mu_0 \circ Z_t^{-1}$, where

$$dZ_t(u) = -\nabla R\left(Z_t(u), \nu_t\right) dtred + \sqrt{\alpha}\int_\Xi G(Z_t(u), \nu_t, \xi) W(d\xi, dt)$$

$$Z_0(u) = u.$$

Comparison in Wasserstein (strong) topology

SDE for overparametrized shallow neural network with $z_k(0) \sim \mu_0$ i.i.d.

$$z_k(t_{i+1}) = z_k(t_i) - \alpha \nabla_{z_k} R\left(z_k(t_i), \mu^n_{t_i}, \xi_i\right), \quad k = 1, \dots, n.$$

Consider a solution to continuity equation

$$d\nu_t = \nabla\left(\nabla R(\cdot, \nu_t)\nu_t\right) dtred + \frac{\alpha}{2}\nabla^2 : (A(\cdot, \nu_t)\nu_t)dt + \sqrt{\alpha}\int_\Xi G(Z_t(u), \nu_t, \xi) W(d\xi, dt)$$

with initial particle distributions μ_0, i.e. $\nu_t = \mu_0 \circ Z_t^{-1}$, where

$$dZ_t(u) = -\nabla R\left(Z_t(u), \nu_t\right) dtred + \sqrt{\alpha}\int_\Xi G(Z_t(u), \nu_t, \xi) W(d\xi, dt)$$

$$Z_0(u) = u.$$

Theorem [Mei, Montanari, Nguyen '18, PNAS]

For R quite regular smooth enough and $n \sim \frac{1}{\alpha}$ one has

$$\sup_{t_i \le T} W_1\left(\mathrm{Law}\mu^n_{t_i}, \mathrm{Law}\nu_{t_i}\right) = redoO\left(\sqrt{\alpha}\right).$$

Comparison in Wasserstein (strong) topology

SDE for overparametrized shallow neural network with $z_k(0) \sim \mu_0$ i.i.d.

$$z_k(t_{i+1}) = z_k(t_i) - \alpha \nabla_{z_k} R\left(z_k(t_i), \mu^n_{t_i}, \xi_i\right), \quad k = 1, \dots, n.$$

Consider a solution to continuity equation

$$d\nu_t = \nabla\left(\nabla R(\cdot, \nu_t)\nu_t\right) dt + \frac{\alpha}{2}\nabla^2 : (A(\cdot, \nu_t)\nu_t) dt + \sqrt{\alpha} \int_\Xi G(Z_t(u), \nu_t, \xi) W(d\xi, dt)$$

with initial particle distributions μ_0, i.e. $\nu_t = \mu_0 \circ Z_t^{-1}$, where

$$dZ_t(u) = -\nabla R\left(Z_t(u), \nu_t\right) dt + \sqrt{\alpha} \int_\Xi G(Z_t(u), \nu_t, \xi) W(d\xi, dt)$$

$$Z_0(u) = u.$$

Theorem [Gess, Gvalani, K. '25, PTRF]

For R quite regular smooth enough and $n \sim \frac{1}{\alpha}$ one has

$$\sup_{t_i \le T} W_1\left(\mathrm{Law}\mu^n_{t_i}, \mathrm{Law}\nu_{t_i}\right) = o\left(\sqrt{\alpha}\right).$$

Comparison in Wasserstein (strong) topology

SDE for overparametrized shallow neural network with $z_k(0) \sim \mu_0$ i.i.d.

$$z_k(t_{i+1}) = z_k(t_i) - \alpha \nabla_{z_k} R\left(z_k(t_i), \mu^n_{t_i}, \xi_i\right), \quad k = 1, \dots, n.$$

Consider a solution to continuity equation

$$d\nu_t = \nabla\left(\nabla R(\cdot, \nu_t)\nu_t\right) dt + \frac{\alpha}{2}\nabla^2 : (A(\cdot, \nu_t)\nu_t) dt + \sqrt{\alpha} \int_\Xi G(Z_t(u), \nu_t, \xi) W(d\xi, dt)$$

with initial particle distributions μ_0, i.e. $\nu_t = \mu_0 \circ Z_t^{-1}$, where

$$dZ_t(u) = -\nabla R\left(Z_t(u), \nu_t\right) dt + \sqrt{\alpha} \int_\Xi G(Z_t(u), \nu_t, \xi) W(d\xi, dt)$$

$$Z_0(u) = u.$$

Theorem [Gess, Gvalani, K. '25, PTRF]

For R quite regular smooth enough and $n \sim \frac{1}{\alpha}$ one has

$$\sup_{t_i \le T} W_1\left(\mathrm{Law}\mu^n_{t_i}, \mathrm{Law}\nu_{t_i}\right) = o\left(\sqrt{\alpha}\right).$$

Proof is based on CLT for SGD [Sirignano, Spiliopoulos '19, SPA] + CLT for the SPDE

References

[1] Gess, Kassing, Konarovskyi,
Stochastic Modified Flows, Mean-Field Limits and Dynamics of Stochastic Gradient Descent
Journal of Machine Learning Research 25 (2024)

[2] Gess, Gvalani, Konarovskyi,
Conservative SPDEs as Fluctuating Mean Field Limits of Stochastic Gradient Descent
Probability Theory and Related Fields 192 (2025)

Thank you!