Propagation of Chaos for Mean-Field Langevin Dynamics and its Application to Model Ensemble

Atsushi Nitanda[1,2], Anzelle Lee[3,1], Damian Tan[2,1], Mizuki Sakaguchi[4], Taiji Suzuki[5,6]

1 Centre for Frontier AI Research (CFAR) 2 Nanyang Technological University Singapore 3 NUS National University of Singapore 4 Kyutech Kyushu Institute of Technology 5 UTokyo 6 AIP

Overview

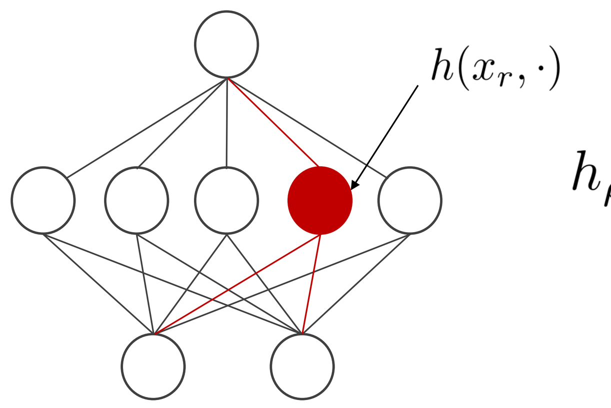
Mean-Field Neural Network:
2-layer NN formalized as an average w.r.t. neurons, which has the global convergence and feature learning properties.

Mean-Field Langevin dynamics:
Noisy gradient descent for mean-field neural networks.

New propagation of chaos result (PoC):
Convergence of mean-field Langevin dynamics in finite-particle setup (noisy GD) with improved particle approximation error.

Model Ensemble
Establish PoC-based ensemble method with nontrivial model approximation errors.

Two-layer Neural Network in Mean-Field Regime

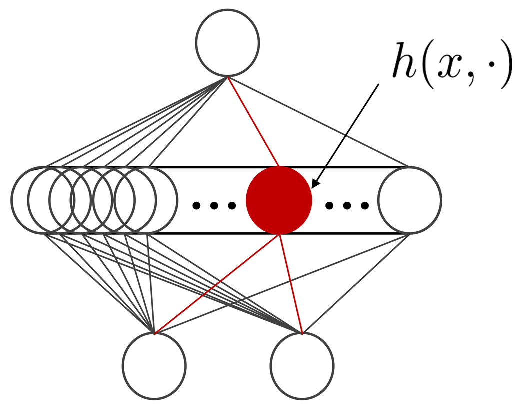


$$h_{\rho_{\mathbf{x}}}(z)=\frac{1}{N}\sum_{r=1}^{N}h(x_r,z)\quad \rho_{\mathbf{x}}=\frac{1}{N}\sum_{r=1}^{N}\delta_{x_r}$$

$\mathbf{x}=(x_r)_{r=1}^N$

$(x_r\sim\mu)$

$N\to\infty$ (mean-field limit)



$$h_\mu(z)=\mathbb{E}_{X\sim\mu}[h(X,z)]=\int h(x,z)\mu(\mathrm{d}x)$$

Linear w.r.t. μ

For loss function ℓ consider L_2-regularized loss:

$$F(\mu)=\mathbb{E}_{(Z,Y)}[\ell(h_\mu(Z),Y)]+\lambda'\mathbb{E}_\mu[\|X\|_2^2].$$

Noisy gradient descent for N-particle setting:

$$\mathrm{d}\mathbf{X}_t=-N\nabla_{\mathbf{X}_t}F(\rho_{\mathbf{X}_t})\mathrm{d}t+\sqrt{2\lambda}\mathrm{d}\mathbf{W}_t$$

where $\mathbf{W}_t=(W_t^1,\ldots,W_t^N),\ \mathbf{X}_t=(X_t^1,\ldots,X_t^N).$

Understand the optimization and approximation efficiency.

Mean-Field Langevin Dynamics

Noisy GD is a Langevin dynamics on $\mathbb{R}^{Nd}$ to solve

$$\min_{\mu^{(N)}\in\mathcal{P}_2(\mathbb{R}^{Nd})}\left\{\mathcal{L}^{(N)}(\mu^{(N)})=N\mathbb{E}_{\mathbf{X}\sim\mu^{(N)}}[F(\mu_{\mathbf{X}})]+\lambda\mathrm{Ent}(\mu^{(N)})\right\}$$

Model	**Optimization**
$h_{\mu_{\mathbf{x}}}(z)=\frac{1}{N}\sum_{r=1}^{N}h(x_r,z)$	Noisy gradient descent
$N\to\infty$ (mean-field limit)	$N\to\infty$ (mean-field limit)
$h_\mu(z)=\mathbb{E}_{X\sim\mu}[h(X,z)]$	**Mean-field Langevin dynamics**

Mean-field Langevin dynamics for infinite-particle setting:
(Mean-field limit: $N\to\infty$)

$$\mathrm{d}X_t=-\nabla\frac{\delta F(\mu_t)}{\delta\mu}(X_t)\mathrm{d}t+\sqrt{2\lambda}\mathrm{d}W_t,$$

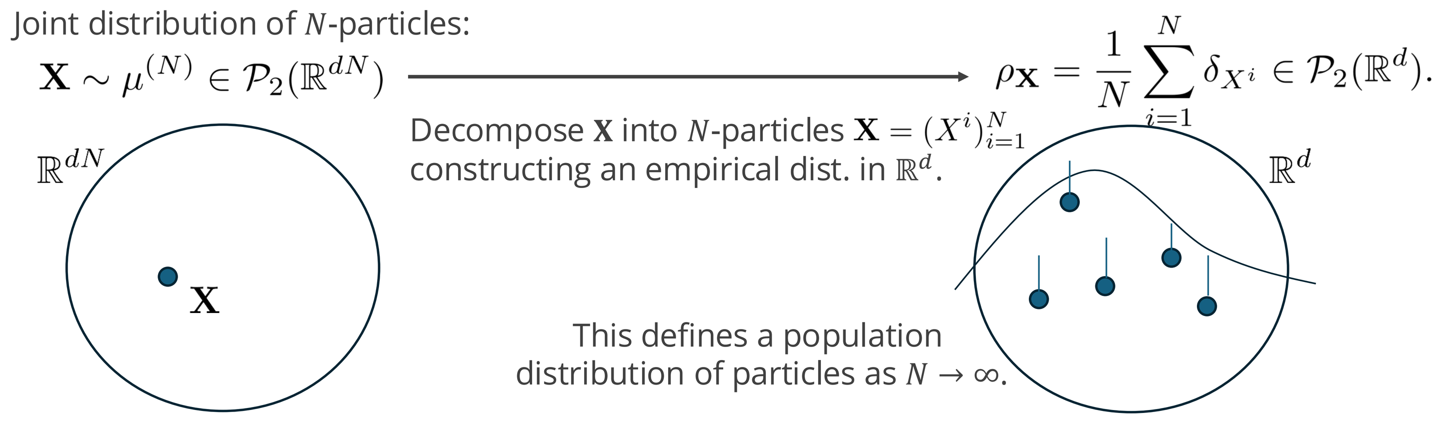
[Mei, Montanari & Nguyen (2018)], [Hu, Ren, Siska, & Szpruch (2021)]

where $\mu_t=\mathrm{Law}(X_t)$.

MFLD solve the following problem: [Nitanda, Wu, & Suzuki (2022)], [Chizat (2022)]

$$\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)}\{\mathcal{L}(\mu)=F(\mu)+\lambda\mathrm{Ent}(\mu)\}$$

Hence, we expect $\mu_*^{(N)}\to\mu_*^{\otimes N},\quad \frac{1}{N}\mathcal{L}^{(N)}(\mu_*^{(N)})\to\mathcal{L}(\mu_*)\quad(N\to\infty)$



[Chen, Ren, & Wang (2024)], [Suzuki, Nitanda, & Wu (2023)]

Question: bound on the following error (opt. + approx. errors)

$$0\le\frac{1}{N}\mathcal{L}^{(N)}(\mu_t^{(N)})-\mathcal{L}(\mu_*)\le\ ?$$

Minimum in the mean-field limit

(Improved) Propagation of Chaos

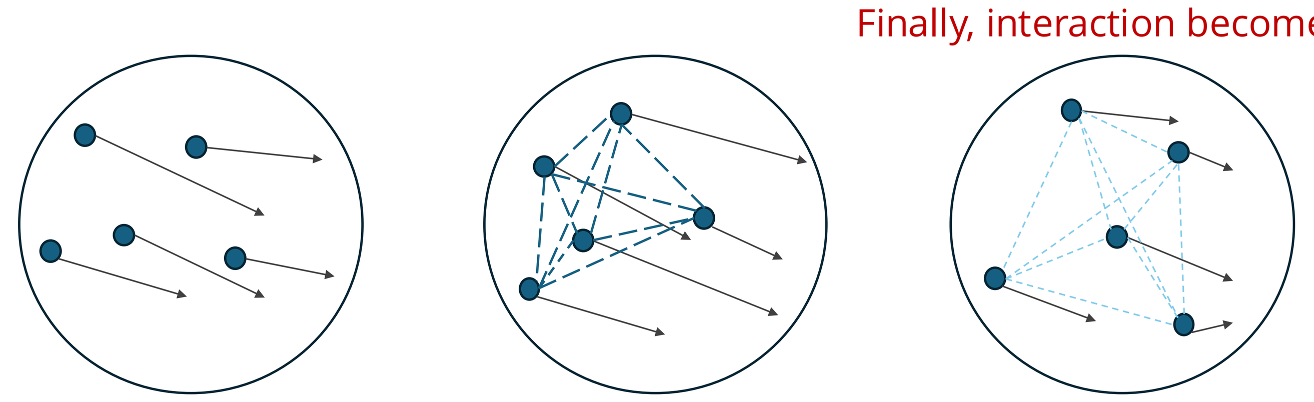
Theorem (PoC)

$$\frac{\lambda}{N}\mathrm{KL}(\mu_t^{(N)}\|\mu_*^{\otimes N})\le\frac{1}{N}\mathcal{L}^{(N)}(\mu_t^{(N)})-\mathcal{L}(\mu_*)\le\frac{B}{N}+\exp(-2\alpha\lambda t)\Delta_0^{(N)}.$$

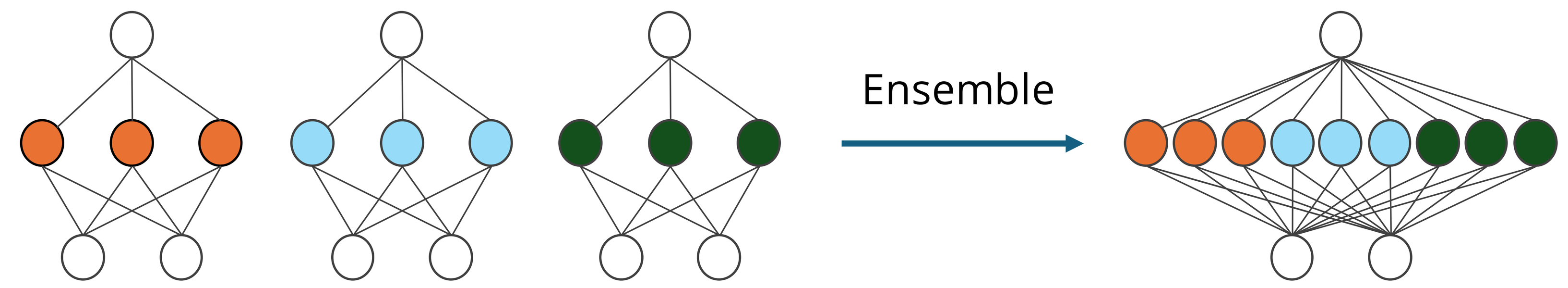
Extension of [Nitanda (2024)]

This implies POC: $\frac{1}{N}\mathrm{KL}(\mu_t^{(N)}\|\mu_*^{\otimes N})\to 0\quad(t,N\to\infty)$.

Finally, interaction becomes weak



Model Ensemble



Theorem (Approximation Error)

$$\mathbb{E}_{\{\mathbf{X}_j\}_{j=1}^M}\left[\left(\frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N}h(X_j^i,z)-\mathbb{E}_{X\sim\mu_*}[h(X,z)]\right)^2\right]$$
$$\le\frac{4R^2}{MN}+\frac{8R^2}{M}\sqrt{\frac{\mathrm{KL}(\mu^{(N)}\|\mu_*^{\otimes N})}{N}}+\frac{2R^2\mathrm{KL}(\mu^{(N)}\|\mu_*^{\otimes N})}{N}.$$

Upper bound after the training: $\frac{4R^2}{MN}+\frac{8R^2}{M}\sqrt{\frac{B}{\lambda N}}+\frac{2BR^2}{\lambda N}$.

Application (LoRA for LMs): $x\mapsto Wx=W_{\mathrm{pre}}x+\gamma BAx$.

Low-rank matrices

Model	Method	SIQA	PIQA	WinoGrande	OBQA	ARC-c	ARC-e	BoolQ	HellaSwag	Ave.
Llama2 7B	LoRA (best)	79.48	82.43	81.77	80.60	67.75	80.47	70.37	86.67	78.69
	PoC merge	81.17	84.60	85.16	86.60	72.53	86.62	72.45	92.79	82.74
Llama3 8B	LoRA (best)	81.22	89.50	86.74	86.00	79.86	90.53	72.91	95.34	85.26
	PoC merge	82.04	89.39	89.27	89.20	83.28	92.30	76.33	96.58	87.30

References
[Mei, Montanari & Nguyen (2018)] A mean field view of the landscape of two-layer neural networks. PNAS, 2018.
[Hu, Ren, Siska, & Szpruch (2019)] Mean-field Langevin dynamics and energy landscape of neural networks. AIHP, 2021.
[Nitanda, Wu, & Suzuki (2022)] Convex analysis of the mean field Langevin dynamics. AISTATS, 2022.
[Chizat (2022)] Mean-field Langevin dynamics: Exponential convergence and annealing. TMLR, 2022.
[Chen, Ren, and Wang] Uniform-in-time propagation of chaos for kinetic mean field Langevin dynamics. EJP, 2024.
[Suzuki, Nitanda, and Wu] Uniform-in-time Propagation of Chaos for the Mean Field Gradient Langevin Dynamics. ICLR, 2023.
[Nitanda] Improved Particle Approximation Error for Mean Field Neural Networks. NeurIPS, 2024.