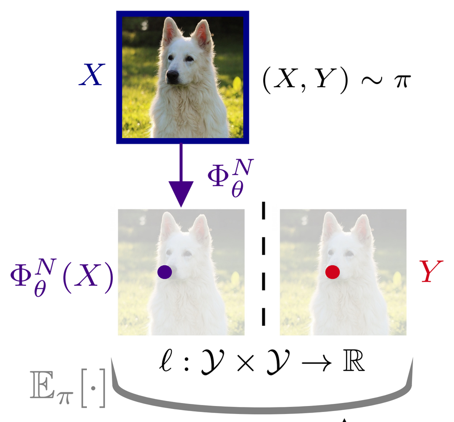
# Symmetries in Overparametrized Neural Networks: A Mean Field View

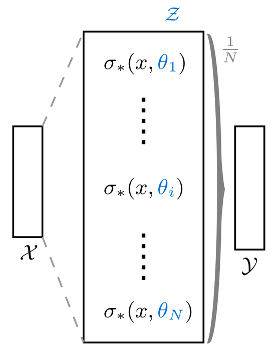
Javier Maass[1] Joaquín Fontbona[1]

[1]Center for Mathematical Modeling, University of Chile

## Problem Setting

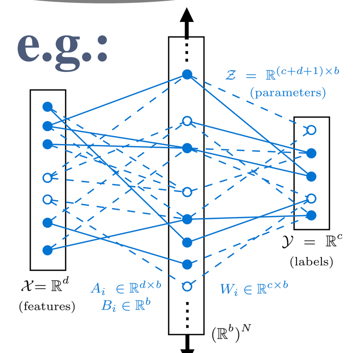
- $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ *feature, label, parameter* spaces.
- Data Distribution $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$.
- $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ **convex** loss function.
- $\Phi_\theta^N$ a *(shallow) neural network (NN).*
- Population risk: $R(\theta) = \mathbb{E}_\pi\left[\ell(\Phi_\theta^N(X), Y)\right]$
- Activation/Unit: $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$.



- Shallow NN model: $\Phi_\theta^N = \langle \sigma_*, \nu_\theta^N \rangle$; $\theta := (\theta_i)_{i=1}^N \in \mathcal{Z}^N$, $\nu_\theta^N := \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$.
- Shallow Model: $\Phi_\mu = \langle \sigma_*, \mu \rangle$; $\mu \in \mathcal{P}(\mathcal{Z})$. **Barron** space: $\mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{Z}))$.

**e.g.:**



We study $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ given by $R(\mu) := \mathbb{E}_\pi\left[\ell(\Phi_\mu(X), Y)\right]$ **(convex)**.

### Generalization in Learning: A Mean Field View

**SGD**: Initialize i.i.d. on $\mu_0 \in \mathcal{P}_2(\mathcal{Z})$ and iterate (for $\{(X_k, Y_k)\}_{k\in\mathbb{N}} \overset{i.i.d.}{\sim} \pi$):

$$\theta_i^{k+1} = \theta_i^k - s_k^N \nabla_z \sigma_*(X_k, \theta_i^k) \cdot \nabla_1 \ell(\Phi_{\theta^k}^N(X_k), Y_k) + s_k^N \tau \nabla r(\theta_i^k) + \sqrt{2\beta s_k^N}\,\xi_i^k.$$

*Step-size* $s_k^N = \varepsilon_N \varsigma(k\varepsilon_N)$; *Penalization* $r : \mathcal{Z} \to \mathbb{R}$; *Gaussian noise* $\xi_i^k \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathrm{Id}_\mathcal{Z})$, $\tau, \beta \geq 0$.

> **Thm.1** (MFL): $\left(\nu^N_{\theta^{\lfloor t/\varepsilon_N \rfloor}}\right)_{t\in[0,T]} \underset{N\to\infty}{\Longrightarrow} (\mu_t)_{t\in[0,T]}$ *in* $D_{\mathcal{P}(\mathcal{Z})}([0,T])$
> where $(\mu_t)_{t\geq 0}$ is the (unique) **WGF**$(R^{\tau,\beta})$ starting at $\mu_0$ [4, 7, 9, 10].

**Entropy-regularized risk:** $R^{\tau,\beta}(\mu) = R(\mu) + \tau\int r d\mu + \beta H_\lambda(\mu)$

**Wasserstein Gradient Flow (WGF)** for $R^{\tau,\beta}$: $(\mu_t)_{t\geq 0}$;

$$\partial_t \mu_t = \varsigma(t)\left[\mathrm{div}\left(\left(D_\mu R(\mu_t, \cdot) + \tau\nabla_\theta r\right)\mu_t\right) + \beta\Delta\mu_t\right],$$

with $D_\mu R$ the **intrinsic derivative** of $R$ [1, 2].

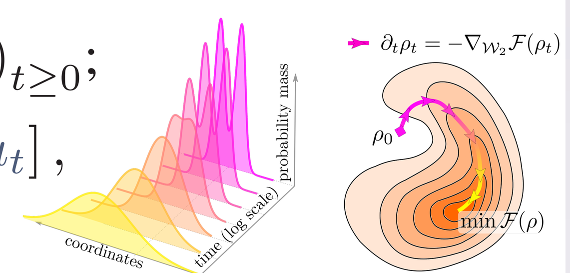


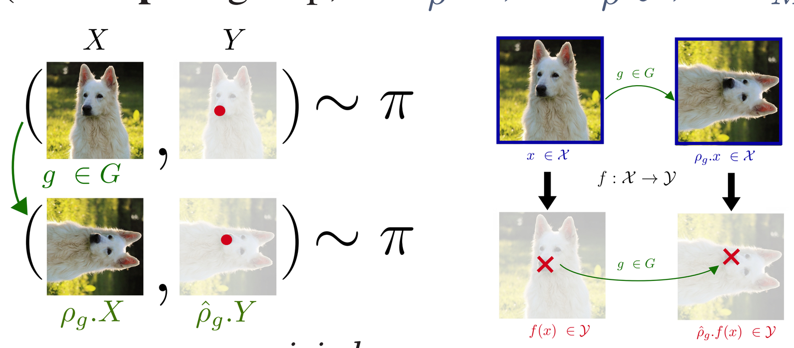

Image taken from [8]

When $\tau, \beta > 0$, flow **converges** to the global minimizer of $R^{\tau,\beta}$ (see e.g. [3, 6])

### Learning with Symmetries ($G$ **compact** group, $G \circlearrowright_\rho \mathcal{X}$, $G \circlearrowright_{\hat\rho} \mathcal{Y}$, $G \circlearrowright_M \mathcal{Z}$)

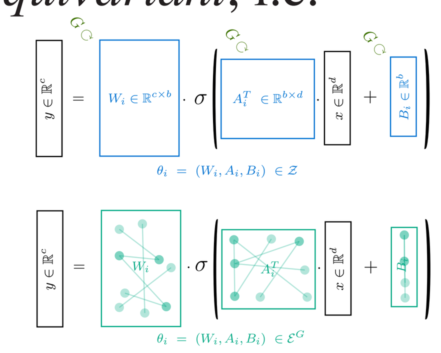
**Equivariant Data:**
$\forall g \in G,\ (\rho_g.X, \hat\rho_g.Y) \sim \pi$.

**Equivariant Function:**
$\forall g \in G,\ f(\rho_g.\cdot) = \hat\rho_g.f(\cdot)$




- **Data Augmentation (DA)**: Draw $\{g_k\}_{k\in\mathbb{N}} \overset{i.i.d.}{\sim} \lambda_G$ and do SGD with $\{(\rho_{g_k}.X_k, \hat\rho_{g_k}.Y_k)\}_{k\in\mathbb{N}}$. This optimizes the *symmetrized population risk*:
$$R^{DA}(\theta) := \mathbb{E}_\pi\left[\int_G \ell\left(\Phi_\theta^N(\rho_g.X), \hat\rho_g.Y\right) d\lambda_G(g)\right]$$
- **Feature Averaging (FA)**: Train the **symmetrized model**, via the **symmetrization operator**, $(\mathcal{Q}_G.f) := \int_G \hat\rho_{g^{-1}}.f(\rho_g.\cdot)d\lambda_G(g)$. This optimizes:
$$R^{FA}(\theta) := \mathbb{E}_\pi\left[\ell\left((\mathcal{Q}_G.\Phi_\theta^N)(X), Y\right)\right]$$
- **Equivariant Architectures (EA)**: For $\sigma_*$ *jointly equivariant*, i.e. $\forall g, x, z : \sigma_*(\rho_g.x, M_g.z) = \hat\rho_g\sigma_*(x,z)$; **EAs** are the fixed points: $\mathcal{E}^G := \{z : M_g.z = z,\ \forall g\}$.

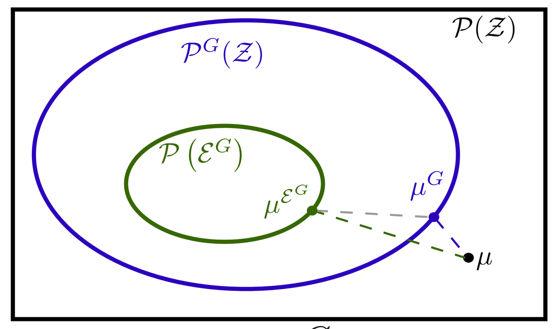
Optimizes: $R^{EA}(\theta) := \mathbb{E}_\pi\left[\ell\left(\Phi_\theta^{N,EA}(X), Y\right)\right]$;
$\Phi_\theta^{N,EA} := \langle \sigma_*, P_{\mathcal{E}^G}\#\nu_\theta^N \rangle$, $P_{\mathcal{E}^G}.z := \int_G M_g.z\, d\lambda_G(g)$.



## Main Results

### Optimizing Invariant Functionals

- **Weakly-Invariant (WI) measures**
$\mathcal{P}^G(\mathcal{Z}) := \{\mu : \forall g \in G,\ M_g\#\mu = \mu\}$
- **Strongly-Invariant (SI) measures**
$\mathcal{P}(\mathcal{E}^G) := \{\mu : \mu(\mathcal{E}^G) = 1\}$



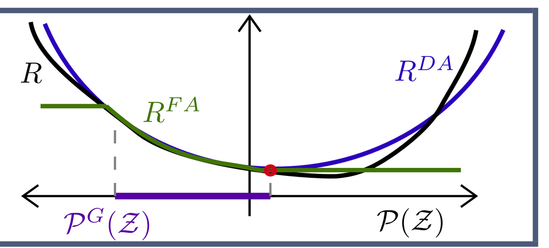

**Symmetrization**: $\mu^G := \int_G (M_g\#\mu)d\lambda_G$ and **Projection**: $\mu^{\mathcal{E}^G} := P_{\mathcal{E}^G}\#\mu$

**Assumption 1**: $\pi \in \mathcal{P}_2(\mathcal{X} \times \mathcal{Y})$; $\ell$ convex, invariant; $\sigma_*$ jointly equivariant
+ standard MF assumptions (regularity and boundedness).

**Proposition 1**: For $\Phi_\mu \in \mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{Z}))$, $(\mathcal{Q}_G\Phi_\mu) = \Phi_{\mu^G}$.

**Proposition 2**: $R^{DA}$, $R^{FA}$, $R^{EA}$ are **invariant** and can be written in terms of $R$ and the above operations. When $\pi$ is equivariant, $R$ is invariant too.

> **Thm.2** (Equivalence of **DA** and **FA**):
> $$\inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) = \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R^{DA}(\mu) = \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R^{FA}(\mu).$$




If $\pi$ is equivariant, using **DA**, **FA** or **no SL technique** makes no difference.

**Prop. 4:** For simple examples, with equivariant $\pi$, we can get:
$$\inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu) < \inf_{\nu\in\mathcal{P}(\mathcal{E}^G)} R(\nu)$$

**Prop. 5:** Quadratic $\ell$ + equivariant $\pi$ + $\mathcal{E}^G$ *universal on equivariant functions*:
$$\inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu) = \inf_{\nu\in\mathcal{P}(\mathcal{E}^G)} R(\nu) = R_*$$

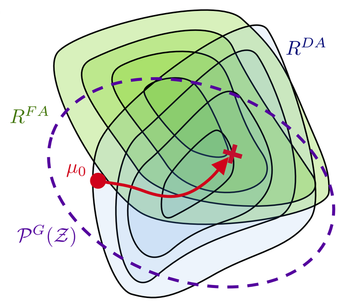
### Symmetries in the Training Dynamic

> **Thm.3** (Invariant **WGF**): Invariant $F : \mathcal{P}(\mathcal{Z}) \to \overline{\mathbb{R}}$, with well-defined **WGF**$(F)$, $(\mu_t)_{t\geq 0}$. If i.c. $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$, then: $\mu_t \in \mathcal{P}_2^G(\mathcal{Z}), \forall t \geq 0$.

**Corollary 3**: For $R$ and $r$ invariant, under technical assumptions [4], if the i.c. of **WGF**$(R^{\tau,\beta})$ satisfies $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$, then: $\mu_t \in \mathcal{P}_2^G(\mathcal{Z})\ \forall t \geq 0$.

This applies to **freely-trained NN, even without SL-techniques**.

> **Thm.4**: If $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$: **WGF**$(R^{DA})$, **WGF**$(R^{FA})$ (and **WGF**$(R)$ if $R$ invariant), exactly coincide.

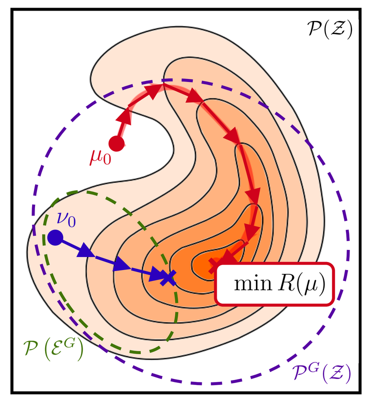



Similar results hold for $\mathcal{P}(\mathcal{E}^G)$; consider a SGD variant with **projected noise**:

$$\theta_i^{k+1} = \theta_i^k - s_k^N\left(\nabla_z\sigma_*(X_k, \theta_i^k) \cdot \nabla_1\ell(\Phi_{\theta^k}^N(X_k), Y_k) + \tau\nabla r(\theta_i^k)\right) + \sqrt{2\beta s_k^N}\,P_{\mathcal{E}^G}\xi_i^k.$$

It approximates the **WGF** of $R_{\mathcal{E}^G}^{\tau,\beta}(\mu) := R(\mu) + \tau\int r d\mu + \beta H_{\lambda_{\mathcal{E}^G}}(\mu^{\mathcal{E}^G})$.

> **Thm.5**: Invariant $R$ and $r$, under technical assumptions [5]; if i.c. of **WGF**$(R_{\mathcal{E}^G}^{\tau,\beta})$ is s.t. $\nu_0 \in \mathcal{P}_2(\mathcal{E}^G)$, then: $\nu_t \in \mathcal{P}_2(\mathcal{E}^G)\ \forall t \geq 0$.




If $\pi$ equivariant, parameters *stay* **SI**, despite facing **no explicit constraint, nor using any SL-technique.**

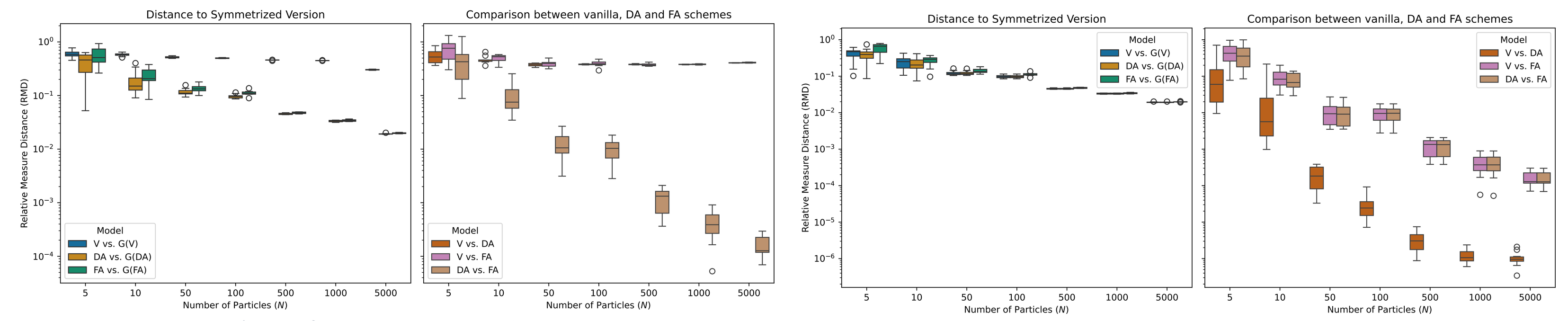
**Thm.5** holds for $R^{DA}$, $R^{FA}$ and $R^{EA}$, **even if $\pi$ is not equivariant**.

> **Thm.6**: If $\nu_0 \in \mathcal{P}_2(\mathcal{E}^G)$, **WGF** for $R^{DA}$, $R^{FA}$, $R^{EA}$ (& $R$ if invariant) coincide.

## Numerical Experiments: Teacher-Student Setting

For $\mathcal{X} = \mathcal{Y} = \mathbb{R}^2$, $\mathcal{Z} = \mathbb{R}^{2\times 2}$, we take $G = C_2$ acting naturally, and $\sigma_*(x, z) = \sigma(z \cdot x)$ with $\sigma$ pointwise sigmoidal.

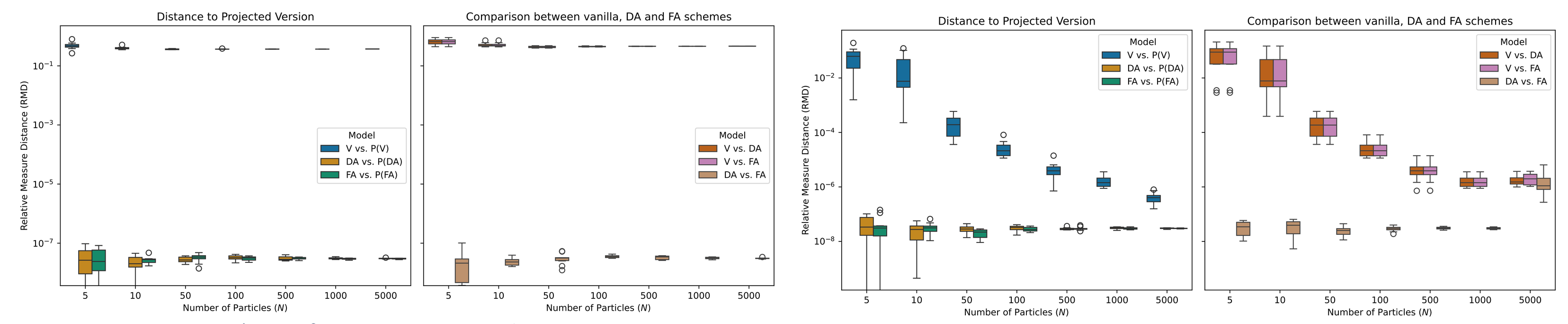
**WI-initialized students:**




Arbitrary teacher WI teacher

- If $f_*$ is **arbitrary**, **DA/FA** increasingly *stay* **WI** and approach each other
- If $f_*$ is **WI** (i.e. equivariant $\pi$), the same holds for **vanilla** training.

**SI-initialized students:**

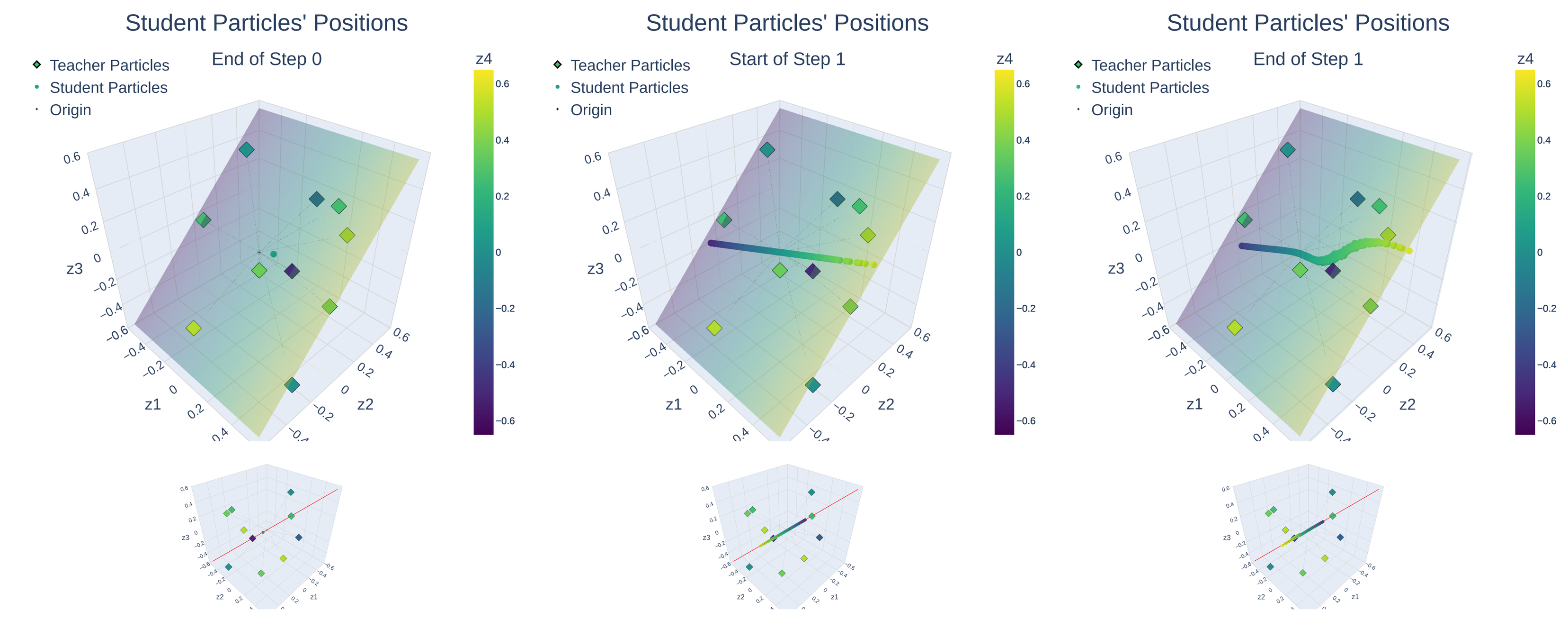



Arbitrary teacher WI teacher

- If $f_*$ is **arbitrary**, **vanilla** training escapes $\mathcal{E}^G$, regardless of $N$.
- If $f_*$ is **WI** , for large $N$, **vanilla** training stays **SI** and approaches **DA/FA**.

### Heuristic for Discovering Parameter-Sharing Schemes

Start with $E_0 = \{0\} \leq \mathcal{E}^G$ and iteratively: Train model initialized on $E_j$ and check if $\mathbf{dist}^2(\nu_T^N, P_{E_j}\#\nu_T^N) \leq \delta_j$ for $\delta_j > 0$. If not, expand $E_j$.




This should end on $E_* = \mathcal{E}^G$, which encodes *good* **SI** architectures.

## References

[1] P. Cardaliaguet. Notes on mean-field games (from P.-L.Lions lectures at Collège de France). 2013. URL: https://www.ceremade.dauphine.fr/cardaliaguet/MFG20130420.pdf. 2

[2] R. Carmona and F. Delarue. *Probabilistic Theory of Mean Field Games with Applications I: Mean Field FBSDEs, Control, and Games*. Probability Theory and Stochastic Modelling. Springer International Publishing, 2018. ISBN 9783319589206. URL https://books.google.cl/books?id=fZFODwAAQBAJ. 2

[3] F. Chen, Y. Lin, Z. Ren, and S. Wang. Uniform-in-time propagation of chaos for kinetic mean field langevin dynamics. *Electronic Journal of Probability*, 29:1–43, 2024. 2

[4] L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018. 2

[5] V. De Bortoli, A. Durmus, X. Fontaine, and U. Simsekli. Quantitative propagation of chaos for sgd in wide neural networks. *Advances in Neural Information Processing Systems*, 33:278–288, 2020. 2

[6] K. Hu, Z. Ren, D. Siska, and L. Szpruch. Mean-field langevin dynamics and energy landscape of neural networks. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 57, pages 2043–2065. Institut Henri Poincaré, 2021. 2

[7] S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. doi: 10.1073/pnas.1806579115. URL https://www.pnas.org/doi/abs/10.1073/pnas.1806579115. 2

[8] P. Mokrov, A. Korotin, L. Li, A. Genevay, J. Solomon, and E. Burnaev. Large-scale wasserstein gradient flows, 2021. 2

[9] G. Rotskoff and E. Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75 (9):1889–1935, jul 2022. doi: 10.1002/cpa.22074. URL https://doi.org/10.1002/cpa.22074. 2

[10] J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020. 2