Symmetries in Overparametrized Neural Networks: A Mean-Field View

Javier Maass Martínez

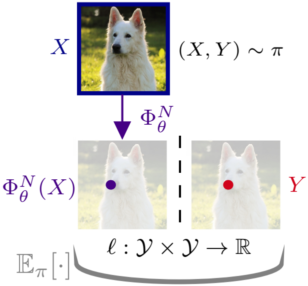
Joint work with Joaquín Fontbona

Center for Mathematical Modeling
University of Chile

Context

Our Setting

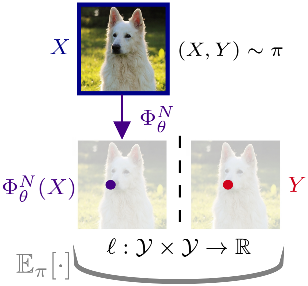
- $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ separable Hilbert spaces. (*features*, *labels*, *parameters* resp.).
- Data Distribution $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. (samples $(X, Y) \sim \pi$).
- $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ **convex** loss function.
- Φ_θ^N a *(shallow) neural network (NN)* of N units and parameters $\theta \in \mathcal{Z}^N$.



Dog image taken from [10]

Our Setting

- $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ separable Hilbert spaces. (*features*, *labels*, *parameters* resp.).
- Data Distribution $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. (samples $(X, Y) \sim \pi$).
- $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ **convex** loss function.
- Φ_θ^N a *(shallow) neural network (NN)* of N units and parameters $\theta \in \mathcal{Z}^N$.



Dog image taken from [10]

We want to minimize the population risk (generalization error):

$$R(\theta) = \mathbb{E}_\pi \left[\ell(\Phi_\theta^N(X), Y) \right]$$

Our Setting

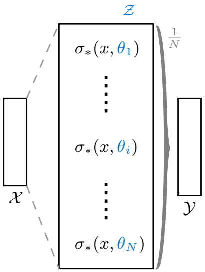
General Activation function (also called *unit*) $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$.

Def. Shallow NN models (general)

$\Phi_\theta^N : \mathcal{X} \to \mathcal{Y}$ with $\theta := (\theta_i)_{i=1}^N \in \mathcal{Z}^N$, is:

$$\forall x \in \mathcal{X},\ \Phi_\theta^N(x) := \frac{1}{N}\sum_{i=1}^{N} \sigma_*(x;\theta_i) = \langle \sigma_*(x;\cdot), \nu_\theta^N \rangle,$$

where $\nu_\theta^N := \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$. Simply put: $\Phi_\theta^N = \langle \sigma_*, \nu_\theta^N \rangle$.



Our Setting

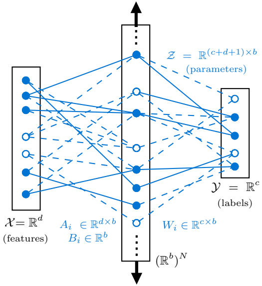
General Activation function (also called *unit*) $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$.

Example: Traditional 'shallow NN' unit

$\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}^c$, $\mathcal{Z} = \mathbb{R}^{c\times b} \times \mathbb{R}^{d\times b} \times \mathbb{R}^b$.
For $z = (W, A, B)$, $\sigma : \mathbb{R}^b \to \mathbb{R}^b$:

$$\sigma_*(x, z) := W\sigma(A^T x + B)$$

Our **general models** go far beyond this example !



Our Setting

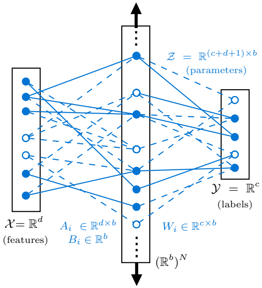
General Activation function (also called *unit*) $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$.

Example: Traditional 'shallow NN' unit

$\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}^c$, $\mathcal{Z} = \mathbb{R}^{c \times b} \times \mathbb{R}^{d \times b} \times \mathbb{R}^b$.
For $z = (W, A, B)$, $\sigma : \mathbb{R}^b \to \mathbb{R}^b$:

$$\sigma_*(x, z) := W\sigma(A^T x + B)$$

Our **general models** go far beyond this example !



Def. Shallow Models (general): $\Phi_\mu = \langle \sigma_*, \mu \rangle$ for $\mu \in \mathcal{P}(\mathcal{Z})$.
Barron space of such models: $\mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{Z}))$.

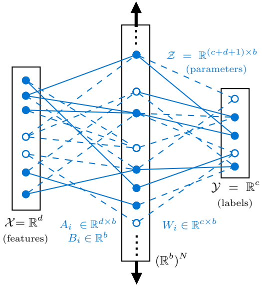
General Activation function (also called *unit*) $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$.

Example: Traditional 'shallow NN' unit

$\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}^c$, $\mathcal{Z} = \mathbb{R}^{c\times b} \times \mathbb{R}^{d\times b} \times \mathbb{R}^b$.
For $z = (W, A, B)$, $\sigma : \mathbb{R}^b \to \mathbb{R}^b$:

$$\sigma_*(x, z) := W\sigma(A^T x + B)$$

Our **general models** go far beyond this example !



Def. Shallow Models (general): $\Phi_\mu = \langle \sigma_*, \mu \rangle$ for $\mu \in \mathcal{P}(\mathcal{Z})$.
Barron space of such models: $\mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{Z}))$.

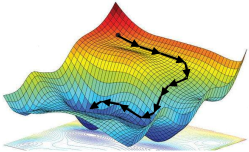
We study $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ given by $R(\mu) := \mathbb{E}_\pi \left[\ell(\Phi_\mu(X), Y) \right]$ (**convex**).

Generalization in Learning: A Mean-Field view

Approximate the optimization using (noisy) SGD ($\{(X_k, Y_k)\}_{k\in\mathbb{N}} \overset{i.i.d.}{\sim} \pi$).

- Initialize $(\theta_i^0)_{i=1}^N \overset{i.i.d.}{\sim} \mu_0 \in \mathcal{P}_2(\mathcal{Z})$.
- Iterate, for $k \in \mathbb{N}$, defining $\forall i \in \{1, \dots, N\}$:

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



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

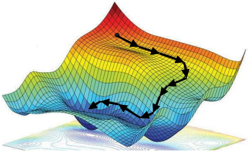
CMM Center for Mathematical Modeling

Generalization in Learning: A Mean-Field view

Approximate the optimization using (noisy) SGD $(\{(X_k, Y_k)\}_{k\in\mathbb{N}} \overset{i.i.d.}{\sim} \pi)$.

- Initialize $(\theta_i^0)_{i=1}^N \overset{i.i.d.}{\sim} \mu_0 \in \mathcal{P}_2(\mathcal{Z})$.
- Iterate, for $k \in \mathbb{N}$, defining $\forall i \in \{1, \dots, N\}$:

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



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

Theorem (Mean-Field limit; sketch) **(see [6, 14, 19, 20] and [4, 7, 8, 15, 21, 22])**

$$\left(\nu^N_{\theta^{\lfloor t/\varepsilon_N \rfloor}}\right)_{t\in[0,T]} \xrightarrow[N\to\infty]{} (\mu_t)_{t\in[0,T]} \quad \text{in } D_{\mathcal{P}(\mathcal{Z})}([0,T])$$

where $(\mu_t)_{t\geq 0}$ is given by the **unique WGF**$(R^{\tau,\beta})$ starting at μ_0.

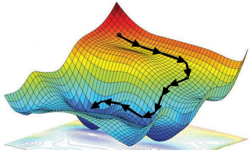
Generalization in Learning: A Mean-Field view

Approximate the optimization using (noisy) SGD ($\{(X_k, Y_k)\}_{k\in\mathbb{N}} \overset{i.i.d.}{\sim} \pi$).

- Initialize $(\theta_i^0)_{i=1}^N \overset{i.i.d.}{\sim} \mu_0 \in \mathcal{P}_2(\mathcal{Z})$.
- Iterate, for $k \in \mathbb{N}$, defining $\forall i \in \{1, \dots, N\}$:

$$\theta_i^{k+1} = \theta_i^k - s_k^N \nabla_z \sigma_*(X_k, \theta_i^k) \cdot \nabla_1 \ell(\Phi^N_{\theta^k}(X_k), Y_k)$$

$$+ s_k^N \tau \nabla r(\theta_i^k) + \sqrt{2\beta s_k^N} \xi_i^k.$$



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

Theorem (Mean-Field limit; sketch) **(see [6, 14, 19, 20] and [4, 7, 8, 15, 21, 22])**

$$\left(\nu^N_{\theta^{\lfloor t/\varepsilon_N \rfloor}}\right)_{t\in[0,T]} \xrightarrow[N\to\infty]{} (\mu_t)_{t\in[0,T]} \quad \text{in } D_{\mathcal{P}(\mathcal{Z})}([0,T])$$

where $(\mu_t)_{t\geq 0}$ is given by the **unique WGF**($R^{\tau,\beta}$) starting at μ_0.

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

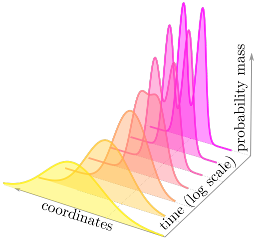
λ is the Lebesgue Measure on $\mathcal{Z}$, and H_λ the *Boltzmann entropy*.

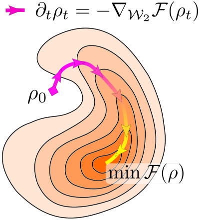
Wasserstein Gradient Flow (WGF) for $R^{\tau,\beta}$ (denoted WGF($R^{\tau,\beta}$))

It is (given an i.c. $\mu_0 \in \mathcal{P}_2(\mathcal{Z})$) the unique (weak) solution, $(\mu_t)_{t\geq 0}$, to:

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

with $D_\mu R : \mathcal{P}_2(\mathcal{Z}) \times \mathcal{Z} \to \mathcal{Z}$ the **intrinsic derivative** of R (see [1, 2, 12]).





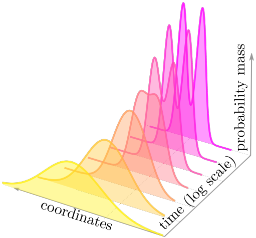
Generalization in Learning: A Mean-Field view

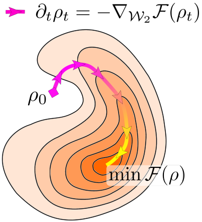
Wasserstein Gradient Flow (WGF) for $R^{\tau,\beta}$ (denoted **WGF**$(R^{\tau,\beta})$)

It is (given an i.c. $\mu_0 \in \mathcal{P}_2(\mathcal{Z})$) the unique (weak) solution, $(\mu_t)_{t\geq 0}$, to:

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

with $D_\mu R : \mathcal{P}_2(\mathcal{Z}) \times \mathcal{Z} \to \mathcal{Z}$ the **intrinsic derivative** of R (see [1, 2, 12]).





When $\tau, \beta > 0$, this flow **converges** to the (unique) global minimizer of $R^{\tau,\beta}$ (see [3, 5, 11, 17, 22])

Image taken from [16]

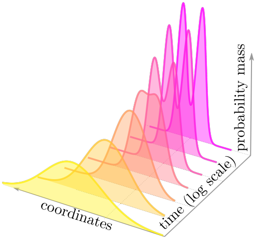
Generalization in Learning: A Mean-Field view

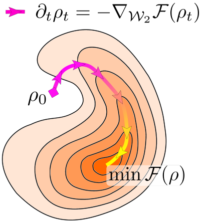
Wasserstein Gradient Flow (WGF) for $R^{\tau,\beta}$ (denoted **WGF**$(R^{\tau,\beta})$)

It is (given an i.c. $\mu_0 \in \mathcal{P}_2(\mathcal{Z})$) the unique (weak) solution, $(\mu_t)_{t\geq 0}$, to:

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

with $D_\mu R : \mathcal{P}_2(\mathcal{Z}) \times \mathcal{Z} \to \mathcal{Z}$ the **intrinsic derivative** of R (see [1, 2, 12]).





When $\tau, \beta > 0$, this flow **converges** to the (unique) global minimizer of $R^{\tau,\beta}$ (see [3, 5, 11, 17, 22])

Image taken from [16]

What if the data has some symmetries?

Learning with Symmetries

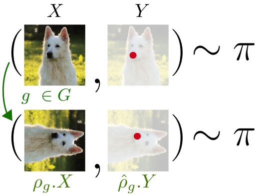
Let G **compact** group with Haar measure λ_G; $G \circlearrowright_{\rho} \mathcal{X}$, $G \circlearrowright_{\hat{\rho}} \mathcal{Y}$

Learning with Symmetries

Let G **compact** group with Haar measure λ_G; $G \circlearrowright_\rho \mathcal{X}$, $G \circlearrowright_{\hat{\rho}} \mathcal{Y}$

Equivariant Data: π s.t., if $(X, Y) \sim \pi$, then:

$$\forall g \in G,\ (\rho_g.X, \hat{\rho}_g.Y) \sim \pi.$$



Learning with Symmetries

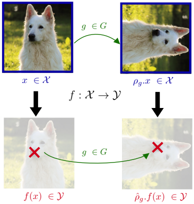
Let G **compact** group with Haar measure λ_G; $G \circlearrowright_\rho \mathcal{X}$, $G \circlearrowright_{\hat{\rho}} \mathcal{Y}$

Equivariant Data: π s.t., if $(X,Y) \sim \pi$, then:

$$\forall g \in G,\ (\rho_g.X, \hat{\rho}_g.Y) \sim \pi.$$

Equivariant Function: $f : \mathcal{X} \to \mathcal{Y}$ s.t. $\forall g \in G$:

$$f(\rho_g.x) = \hat{\rho}_g.f(x)\ \forall x \in \mathcal{X}$$



Learning with Symmetries

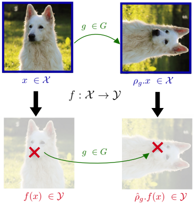
Let G **compact** group with Haar measure λ_G; $G \circlearrowright_\rho \mathcal{X}$, $G \circlearrowright_{\hat{\rho}} \mathcal{Y}$

Equivariant Data: π s.t., if $(X,Y) \sim \pi$, then:

$$\forall g \in G,\ (\rho_g.X, \hat{\rho}_g.Y) \sim \pi.$$

Equivariant Function: $f : \mathcal{X} \to \mathcal{Y}$ s.t. $\forall g \in G$:

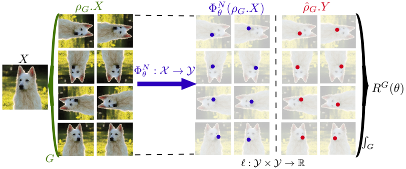
$$f(\rho_g.x) = \hat{\rho}_g.f(x)\ \forall x \in \mathcal{X}$$



Leveraging Symmetry: Data Augmentation (DA)

Draw $\{g_k\}_{k\in\mathbb{N}} \overset{i.i.d.}{\sim} \lambda_G$ and carry out SGD using $\{(\rho_{g_k}.X_k, \hat{\rho}_{g_k}.Y_k)\}_{k\in\mathbb{N}}$.
Aims at optimizing the *symmetrized population risk*:

$$R^{DA}(\theta) := \mathbb{E}_\pi \left[\int_G \ell\left(\Phi_\theta^N(\rho_g.X), \hat{\rho}_g.Y\right) d\lambda_G(g) \right]$$



Learning with Symmetries

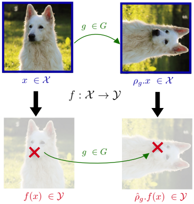
Let G **compact** group with Haar measure λ_G; $G \circlearrowright_\rho \mathcal{X}$, $G \circlearrowright_{\hat{\rho}} \mathcal{Y}$

Equivariant Data: π s.t., if $(X, Y) \sim \pi$, then:

$$\forall g \in G,\ (\rho_g.X, \hat{\rho}_g.Y) \sim \pi.$$

Equivariant Function: $f : \mathcal{X} \to \mathcal{Y}$ s.t. $\forall g \in G$:

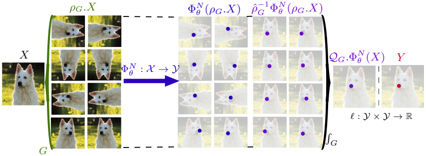
$$f(\rho_g.x) = \hat{\rho}_g.f(x)\ \ \forall x \in \mathcal{X}$$



Leveraging Symmetry: Feature Averaging (FA)

Training a **symmetrized model**, using the **symmetrization operator**, given by $(\mathcal{Q}_G.f)(x) := \int_G \hat{\rho}_{g^{-1}}.f(\rho_g.x)d\lambda_G(g)$. Aims at optimizing:

$$R^{FA}(\theta) := \mathbb{E}_\pi \left[\ell\left((\mathcal{Q}_G.\Phi_\theta^N)(X), Y\right)\right]$$



Leveraging Symmetry: Equivariant Architectures (EA)

Let $G \circlearrowright_M \mathcal{Z}$ and consider $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ *jointly equivariant*, namely:

$$\forall (g, x, z) \in G \times \mathcal{X} \times \mathcal{Z} : \ \sigma_*(\rho_g.x, M_g.z) = \hat{\rho}_g \sigma_*(x, z)$$

Leveraging Symmetry: Equivariant Architectures (EA)

Let $G \circlearrowright_M \mathcal{Z}$ and consider $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ *jointly equivariant*, namely:

$$\forall (g, x, z) \in G \times \mathcal{X} \times \mathcal{Z} : \sigma_*(\rho_g.x, M_g.z) = \hat{\rho}_g \sigma_*(x, z)$$

Fixed points: $\mathcal{E}^G := \{z \in \mathcal{Z} : \forall g \in G,\ M_g.z = z\}$, correspond exactly to **EAs** (e.g. CNNs, GNNs).

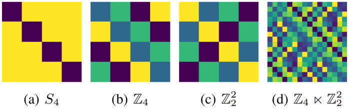
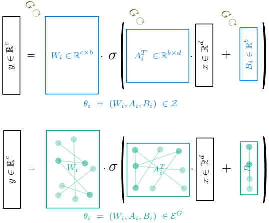


Image taken from [9]



Leveraging Symmetry: Equivariant Architectures (EA)

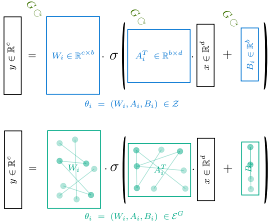
Let $G \circlearrowright_M \mathcal{Z}$ and consider $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ *jointly equivariant*, namely:

$$\forall (g,x,z) \in G \times \mathcal{X} \times \mathcal{Z} : \ \sigma_*(\rho_g.x, M_g.z) = \hat{\rho}_g \sigma_*(x,z)$$

Fixed points: $\mathcal{E}^G := \{z \in \mathcal{Z} \ : \ \forall g \in G, \ M_g.z = z\}$, correspond exactly to **EA**s (e.g. CNNs, GNNs).

(a) S_4 (b) $\mathbb{Z}_4$ (c) $\mathbb{Z}_2^2$ (d) $\mathbb{Z}_4 \ltimes \mathbb{Z}_2^2$

Image taken from [9]



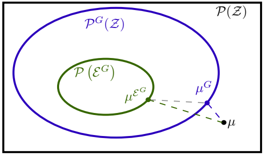
EA aims at minimizing $R^{EA}(\theta) := \mathbb{E}_\pi \left[\ell \left(\Phi_\theta^{N,EA}(X), Y \right) \right]$, with $\Phi_\theta^{N,EA} := \langle \sigma_*, P_{\mathcal{E}^G} \# \nu_\theta^N \rangle$ and $P_{\mathcal{E}^G}.z := \int_G M_g.z \, d\lambda_G(g)$ **orthogonal projection** on $\mathcal{E}^G$.

Main Results

Two Relevant Notions of Symmetry

Subspaces of $\mathcal{P}(\mathcal{Z})$ and modifications of $\mu \in \mathcal{P}(\mathcal{Z})$

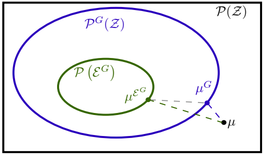
- **Weakly-Invariant (WI) measures**
  $\mathcal{P}^G(\mathcal{Z}) := \{\mu \in \mathcal{P}(\mathcal{Z}) \,:\, \forall g \in G,\; M_g \# \mu = \mu\}$
- **Strongly-Invariant (SI) measures**
  $\mathcal{P}(\mathcal{E}^G) := \{\mu \in \mathcal{P}(\mathcal{Z}) \,:\, \mu(\mathcal{E}^G) = 1\}$



Two Relevant Notions of Symmetry

Subspaces of $\mathcal{P}(\mathcal{Z})$ and modifications of $\mu \in \mathcal{P}(\mathcal{Z})$

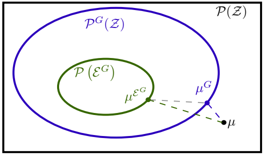
- **Symmetrized** version: $\mu^G := \int_G (M_g \# \mu) d\lambda_G$.
- **Projected** version: $\mu^{\mathcal{E}^G} := P_{\mathcal{E}^G} \# \mu$



Two Relevant Notions of Symmetry

Subspaces of $\mathcal{P}(\mathcal{Z})$ and modifications of $\mu \in \mathcal{P}(\mathcal{Z})$

- **Symmetrized** version: $\mu^G := \int_G (M_g \# \mu) d\lambda_G$.
- **Projected** version: $\mu^{\mathcal{E}^G} := P_{\mathcal{E}^G} \# \mu$

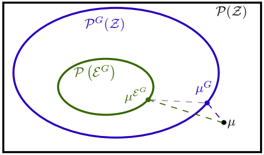


Assumption 1: $\pi \in \mathcal{P}_2(\mathcal{X} \times \mathcal{Y})$; ℓ convex, invariant; σ_* jointly equivariant
$+$ standard assumptions from MF theory (regularity and boundedness).

Two Relevant Notions of Symmetry

Subspaces of $\mathcal{P}(\mathcal{Z})$ and modifications of $\mu \in \mathcal{P}(\mathcal{Z})$

- **Symmetrized** version: $\mu^G := \int_G (M_g \# \mu) d\lambda_G$.
- **Projected** version: $\mu^{\mathcal{E}^G} := P_{\mathcal{E}^G} \# \mu$



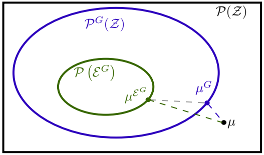
Assumption 1: $\pi \in \mathcal{P}_2(\mathcal{X} \times \mathcal{Y})$; ℓ convex, invariant; σ_* jointly equivariant
$+$ standard assumptions from MF theory (regularity and boundedness).

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

Two Relevant Notions of Symmetry

Subspaces of $\mathcal{P}(\mathcal{Z})$ and modifications of $\mu \in \mathcal{P}(\mathcal{Z})$

- **Symmetrized** version: $\mu^G := \int_G (M_g \# \mu) d\lambda_G$.
- **Projected** version: $\mu^{\mathcal{E}^G} := P_{\mathcal{E}^G} \# \mu$



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

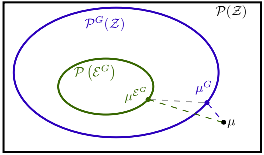
Proposition 1: For $\Phi_\mu \in \mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{Z}))$, $(\mathcal{Q}_G \Phi_\mu) = \Phi_{\mu^G}$.

We **lift** R^{DA}, R^{FA} and R^{EA} to $\mathcal{P}(\mathcal{Z})$ (analogous to R).

Two Relevant Notions of Symmetry

Subspaces of $\mathcal{P}(\mathcal{Z})$ and modifications of $\mu \in \mathcal{P}(\mathcal{Z})$

- **Symmetrized** version: $\mu^G := \int_G (M_g \# \mu) d\lambda_G$.
- **Projected** version: $\mu^{\mathcal{E}^G} := P_{\mathcal{E}^G} \# \mu$



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

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

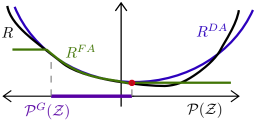
We **lift** R^{DA}, R^{FA} and R^{EA} to $\mathcal{P}(\mathcal{Z})$ (analogous to R).

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

Invariant Functionals and their Optima

Theorem 2 (Equivalence of **DA** and **FA**):

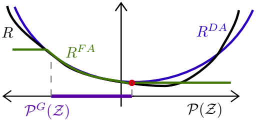
$$\inf_{\mu \in \mathcal{P}^G(\mathcal{Z})} R(\mu) = \inf_{\mu \in \mathcal{P}(\mathcal{Z})} R^{DA}(\mu) = \inf_{\mu \in \mathcal{P}(\mathcal{Z})} R^{FA}(\mu).$$



Invariant Functionals and their Optima

Theorem 2 (Equivalence of **DA** and **FA**):

$$\inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) = \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R^{DA}(\mu) = \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R^{FA}(\mu).$$



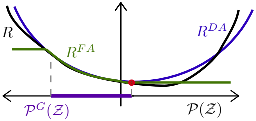
Corollary 1 (quadratic ℓ, invariant $\pi_{\mathcal{X}}$). For $f_* = \mathbb{E}_\pi[Y|X=\cdot]$ and $\tilde{R}_* \geq 0$:

$$\inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) = \tilde{R}_* + \inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} \|\Phi_\mu - \mathcal{Q}_G.f_*\|^2_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})}$$

Invariant Functionals and their Optima

Theorem 2 (Equivalence of **DA** and **FA**):

$$\inf_{\mu \in \mathcal{P}^G(\mathcal{Z})} R(\mu) = \inf_{\mu \in \mathcal{P}(\mathcal{Z})} R^{DA}(\mu) = \inf_{\mu \in \mathcal{P}(\mathcal{Z})} R^{FA}(\mu).$$



Corollary 1 (quadratic ℓ, invariant $\pi_{\mathcal{X}}$). For $f_* = \mathbb{E}_\pi[Y|X = \cdot]$ and $\tilde{R}_* \geq 0$:

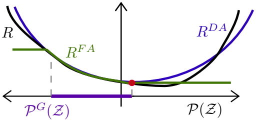
$$\inf_{\mu \in \mathcal{P}^G(\mathcal{Z})} R(\mu) = \tilde{R}_* + \inf_{\mu \in \mathcal{P}^G(\mathcal{Z})} \|\Phi_\mu - \mathcal{Q}_G.f_*\|^2_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})}$$

When $\pi \in \mathcal{P}^G(\mathcal{X} \times \mathcal{Y})$, using **DA**, **FA** or **no SL technique** makes no difference.

Invariant Functionals and their Optima

Theorem 2 (Equivalence of **DA** and **FA**):

$$\inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) = \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R^{DA}(\mu) = \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R^{FA}(\mu).$$



Corollary 1 (quadratic ℓ, invariant $\pi_{\mathcal{X}}$). For $f_* = \mathbb{E}_\pi[Y|X=\cdot]$ and $\tilde{R}_* \geq 0$:

$$\inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) = \tilde{R}_* + \inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} \|\Phi_\mu - \mathcal{Q}_G.f_*\|^2_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})}$$

When $\pi \in \mathcal{P}^G(\mathcal{X}\times\mathcal{Y})$, using **DA**, **FA** or **no SL technique** makes no difference.

On the other hand, regarding **EA**:

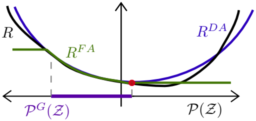
Proposition 4: For really simple examples, with equivariant π, we can get:

$$\inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu) < \inf_{\nu\in\mathcal{P}(\mathcal{E}^G)} R(\nu)$$

Invariant Functionals and their Optima

Theorem 2 (Equivalence of **DA** and **FA**):

$$\inf_{\mu \in \mathcal{P}^G(\mathcal{Z})} R(\mu) = \inf_{\mu \in \mathcal{P}(\mathcal{Z})} R^{DA}(\mu) = \inf_{\mu \in \mathcal{P}(\mathcal{Z})} R^{FA}(\mu).$$



Corollary 1 (quadratic ℓ, invariant $\pi_{\mathcal{X}}$). For $f_* = \mathbb{E}_\pi[Y|X = \cdot]$ and $\tilde{R}_* \geq 0$:

$$\inf_{\mu \in \mathcal{P}^G(\mathcal{Z})} R(\mu) = \tilde{R}_* + \inf_{\mu \in \mathcal{P}^G(\mathcal{Z})} \|\Phi_\mu - \mathcal{Q}_G . f_*\|^2_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})}$$

When $\pi \in \mathcal{P}^G(\mathcal{X} \times \mathcal{Y})$, using **DA**, **FA** or **no SL technique** makes no difference.

On the other hand, regarding **EA**:

Proposition 5: For quadratic ℓ and equivariant π, if $\mathcal{E}^G$ is *universal on equivariant functions* (see e.g. [13, 18, 23, 24]), then:

$$\inf_{\mu \in \mathcal{P}(\mathcal{Z})} R(\mu) = \inf_{\nu \in \mathcal{P}(\mathcal{E}^G)} R(\nu) = R_*$$

Symmetries in the shallow NN training dynamics

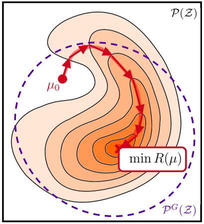
Theorem 3 (Invariant **WGF**s): For invariant $F : \mathcal{P}(\mathcal{Z}) \to \overline{\mathbb{R}}$ with well-defined **WGF**(F) of unique (weak) solution $(\mu_t)_{t\geq 0}$:

$$\text{If i.c. } \mu_0 \in \mathcal{P}_2^G(\mathcal{Z}), \text{ then: } \mu_t \in \mathcal{P}_2^G(\mathcal{Z}) \, \forall t \geq 0.$$

Theorem 3 (Invariant **WGF**s): For invariant $F : \mathcal{P}(\mathcal{Z}) \to \overline{\mathbb{R}}$ with well-defined **WGF**(F) of unique (weak) solution $(\mu_t)_{t\geq 0}$:

$$\text{If i.c. } \mu_0 \in \mathcal{P}_2^G(\mathcal{Z}), \text{ then: } \mu_t \in \mathcal{P}_2^G(\mathcal{Z})\ \forall t \geq 0.$$

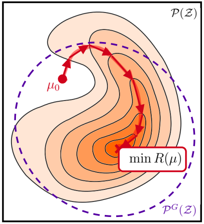
Corollary 3: For R and r invariant, under technical assumptions [6], if i.c. of **WGF**$(R^{\tau,\beta})$ satisfies $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$, then: $\mu_t \in \mathcal{P}_2^G(\mathcal{Z})\ \forall t \geq 0$.



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

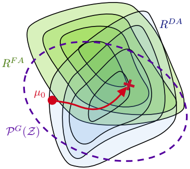
Theorem 3 (Invariant **WGF**s): For invariant $F : \mathcal{P}(\mathcal{Z}) \to \overline{\mathbb{R}}$ with well-defined **WGF**(F) of unique (weak) solution $(\mu_t)_{t\geq 0}$:

$$\text{If i.c. } \mu_0 \in \mathcal{P}_2^G(\mathcal{Z}), \text{ then: } \mu_t \in \mathcal{P}_2^G(\mathcal{Z})\ \forall t \geq 0.$$



Corollary 3: For R and r invariant, under technical assumptions [6], if 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, without SL-techniques**.



Theorem 4: Also, if $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$, then: **WGF**($R^{DA}$), **WGF**($R^{FA}$) (and **WGF**($R$) if R invariant), are equal.

Training with **DA**, **FA** or **no SL-technique** is the same.

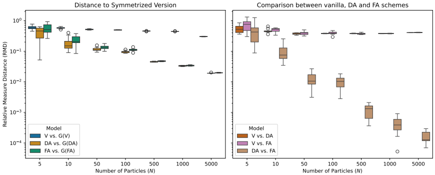
Symmetries in the shallow NN training dynamics

Numerical Validation of our Results: **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 σ pointwise sigmoidal.

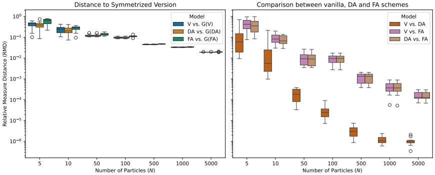
Symmetries in the shallow NN training dynamics

Numerical Validation of our Results: **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 σ pointwise sigmoidal.

WI-initialized students:



Arbitrary teacher



WI teacher

- If f_* is **arbitrary**, as N grows **DA/FA** increasingly *stay* **WI** and approach each other (see **Cor.3 & Thm.4**).
- If f_* is **WI**, the same holds for **vanilla** training (see **Cor.3 & Thm.4**).

Symmetries in the shallow NN training dynamics

Similar results hold for $\mathcal{P}(\mathcal{E}^G)$; consider a variant of SGD 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 $R_{\mathcal{E}^G}^{\tau,\beta}(\mu) := R(\mu) + \tau \int r d\mu + \beta H_{\lambda_{\mathcal{E}^G}}(\mu^{\mathcal{E}^G})$.

Symmetries in the shallow NN training dynamics

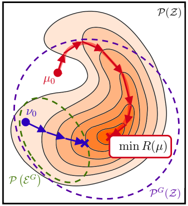
Similar results hold for $\mathcal{P}(\mathcal{E}^G)$; consider a variant of SGD with **projected noise**:

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

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

Theorem 5: For R and r are invariant, under technical assumptions [7]: if i.c. of **WGF**$(R^{\tau,\beta}_{\mathcal{E}^G})$ satisfies $\nu_0 \in \mathcal{P}_2(\mathcal{E}^G)$, then: $\nu_t \in \mathcal{P}_2(\mathcal{E}^G)\ \forall t \geq 0$.

If π equivariant, parameters *stay* **SI**, despite there being **no explicit constraint on them**, **nor any SL-technique** being used.



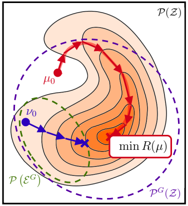
Symmetries in the shallow NN training dynamics

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

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

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

Theorem 5: For R and r are invariant, under technical assumptions [7]: if i.c. of **WGF**$(R^{\tau,\beta}_{\mathcal{E}^G})$ satisfies $\nu_0 \in \mathcal{P}_2(\mathcal{E}^G)$, then: $\nu_t \in \mathcal{P}_2(\mathcal{E}^G)\ \forall t \geq 0$.



If π equivariant, parameters *stay* **SI**, despite there being **no explicit constraint on them**, **nor any SL-technique** being used.

This holds for R^{DA}, R^{FA} and R^{EA} in the role of R, **even if π is not equivariant**.

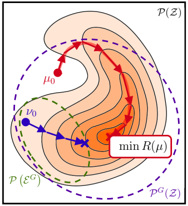
Symmetries in the shallow NN training dynamics

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

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

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

Theorem 5: For R and r are invariant, under technical assumptions [7]: if i.c. of **WGF**($R^{\tau,\beta}_{\mathcal{E}^G}$) satisfies $\nu_0 \in \mathcal{P}_2(\mathcal{E}^G)$, then: $\nu_t \in \mathcal{P}_2(\mathcal{E}^G)\ \forall t \geq 0$.



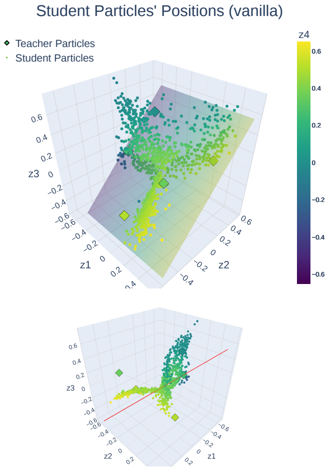
If π equivariant, parameters *stay* **SI**, despite there being **no explicit constraint on them**, **nor any SL-technique** being used.

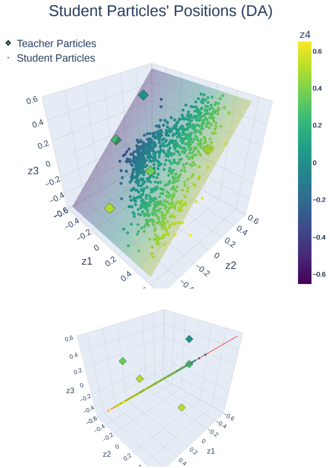
This holds for R^{DA}, R^{FA} and R^{EA} in the role of R, **even if** π **is not equivariant**.

Theorem 6: Also, if $\nu_0 \in \mathcal{P}_2(\mathcal{E}^G)$, then **WGF**($R^{DA}$), **WGF**($R^{FA}$), **WGF**($R^{EA}$) (and **WGF**($R$) if R invariant) all coincide.

Back to our **Numerical Experiments**:

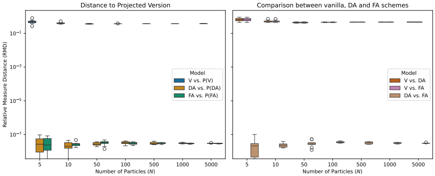
Example of optimization under an **arbitrary** teacher:



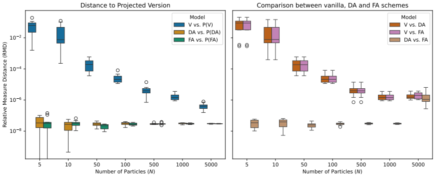


Symmetries in the shallow NN training dynamics

SI-initialized students:



Arbitrary teacher



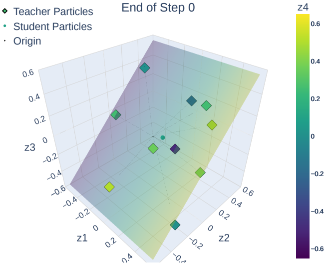
WI teacher

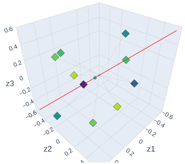
- If f_* is **arbitrary**, **vanilla** training escapes $\mathcal{E}^G$, regardless of N.
- **DA**/**FA** stay **SI** regardless of the teacher and of N (see **Thm.5**).
- If f_* is **WI** (i.e. equivariant π), for large N, **vanilla** training remains **SI** and approaches **DA/FA** (see **Thms.5 & 6**).

Architecture Discovery Heuristic

Finding *good parameter-sharing* schemes for **EA**s:

- Initialize $E_0 = \{0\} \leq \mathcal{E}^G$ and, for $j = 0, 1, \dots$:
  - Train model initialized at $\nu^N_{\theta_0} \in \mathcal{P}(E_j)$ for N_e epochs.
  - Check if $\mathbf{dist}^2(\nu^N_{N_e}, P_{E_j} \# \nu^N_{N_e}) \leq \delta_j$ for threshold $\delta_j > 0$.
  - If not, expand: $E_{j+1} := E_j \oplus v_{E_j}$, with $v_{E_j} = \frac{1}{N}\sum_{i=1}^N (\theta_i^{N_e} - P_{E_j}.\theta_i^{N_e})$.
- Finish with a space $E_* = \mathcal{E}^G$ which encodes *good* **SI** architectures.

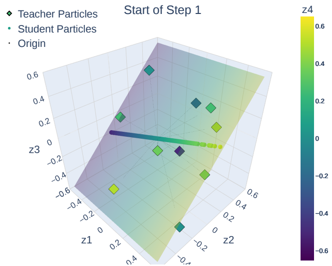


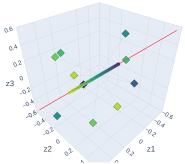


Architecture Discovery Heuristic

Finding *good parameter-sharing* schemes for **EA**s:

- Initialize $E_0 = \{0\} \leq \mathcal{E}^G$ and, for $j = 0, 1, \ldots$:
  - Train model initialized at $\nu^N_{\theta_0} \in \mathcal{P}(E_j)$ for N_e epochs.
  - Check if $\mathbf{dist}^2(\nu^N_{N_e}, P_{E_j}\#\nu^N_{N_e}) \leq \delta_j$ for threshold $\delta_j > 0$.
  - If not, expand: $E_{j+1} := E_j \oplus v_{E_j}$, with $v_{E_j} = \frac{1}{N}\sum_{i=1}^N(\theta_i^{N_e} - P_{E_j}.\theta_i^{N_e})$.
- Finish with a space $E_* = \mathcal{E}^G$ which encodes *good* **SI** architectures.

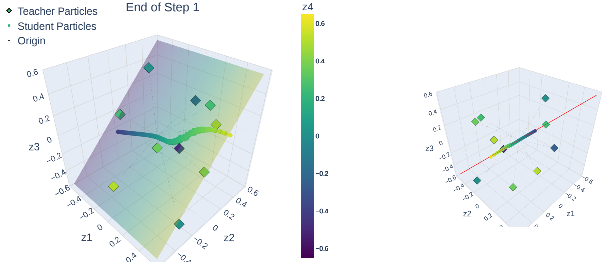




Architecture Discovery Heuristic

Finding *good parameter-sharing* schemes for **EA**s:

- Initialize $E_0 = \{0\} \leq \mathcal{E}^G$ and, for $j = 0, 1, \ldots$:
  - Train model initialized at $\nu_{\theta_0}^N \in \mathcal{P}(E_j)$ for N_e epochs.
  - Check if $\mathbf{dist}^2(\nu_{N_e}^N, P_{E_j} \# \nu_{N_e}^N) \leq \delta_j$ for threshold $\delta_j > 0$.
  - If not, expand: $E_{j+1} := E_j \oplus v_{E_j}$, with $v_{E_j} = \frac{1}{N} \sum_{i=1}^{N} (\theta_i^{N_e} - P_{E_j}.\theta_i^{N_e})$.
- Finish with a space $E_* = \mathcal{E}^G$ which encodes *good* **SI** architectures.



Conclusions and Future Directions

Conclusions and Future Directions

Conclusions

- SL techniques (**DA/FA/EA**) can be expressed in **MF** terms.
- Symmetries are *respected* in the **MFL**, even in a quite strong sense.
- **DA/FA** become equivalent in the **MFL** (and to **vanilla** if π equiv.).
- Numerical validation of results and possible heuristic for **EA** design.

Conclusions and Future Directions

Conclusions

- SL techniques (**DA/FA/EA**) can be expressed in **MF** terms.
- Symmetries are *respected* in the **MFL**, even in a quite strong sense.
- **DA/FA** become equivalent in the **MFL** (and to **vanilla** if π equiv.).
- Numerical validation of results and possible heuristic for **EA** design.

Future Directions

- Quantifying convergence rates to the **MFL** when using SL techniques.
- Extending our *shallow models* analysis to more complex architectures.
- Provide theoretical guarantees for our **EA**-discovery heuristic
- Larger scale experimental validation (*real* datasets, other settings).

Thank you for your attention!

Symmetries in Overparametrized Neural Networks: A Mean-Field View

Javier Maass Martínez

Joint work with Joaquín Fontbona

Center for Mathematical Modeling
University of Chile

CMM
Center for
Mathematical
Modeling

Relevant References I

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

[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.

[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.

[4] Z. Chen, G. Rotskoff, J. Bruna, and E. Vanden-Eijnden. A dynamical central limit theorem for shallow neural networks. *Advances in Neural Information Processing Systems*, 33:22217–22230, 2020.

[5] L. Chizat. Mean-field langevin dynamics : Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=BDqzLH1gEm.

Relevant References II

[6] 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.

[7] 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.

[8] A. Descours, A. Guillin, M. Michel, and B. Nectoux. Law of large numbers and central limit theorem for wide two-layer neural networks: the mini-batch and noisy case. *arXiv preprint arXiv:2207.12734*, 2022.

[9] M. Finzi, M. Welling, and A. G. Wilson. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. In *International conference on machine learning*, pages 3318–3328. PMLR, 2021.

[10] https://www.facebook.com/dogsplanetcom/. Pastor Suizo: Todo sobre esta raza - DogsPlanet.com — dogsplanet.com. `https://www.dogsplanet.com/es/razas-de-perros/pastor-blanco-suizo/`. [Accessed 05-11-2024].

Relevant References III

[11] 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.

[12] P. L. Lions. *Cours au College de France*. 2008.

[13] H. Maron, E. Fetaya, N. Segol, and Y. Lipman. On the universality of invariant networks. In *International conference on machine learning*, pages 4363–4371. PMLR, 2019.

[14] 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`.

[15] S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

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

Relevant References IV

[17] A. Nitanda, D. Wu, and T. Suzuki. Convex analysis of the mean field langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 9741–9757. PMLR, 2022.

[18] S. Ravanbakhsh. Universal equivariant multilayer perceptrons. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7996–8006. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/ravanbakhsh20a.html.

[19] 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%2Fcpa.22074.

[20] 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.

[21] J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

Relevant References V

[22] T. Suzuki, D. Wu, and A. Nitanda. Convergence of mean-field langevin dynamics: time-space discretization, stochastic gradient, and variance reduction. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[23] D. Yarotsky. Universal approximations of invariant maps by neural networks. *Constructive Approximation*, 55(1):407–474, 2022.

[24] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.