Simetrías en Redes Neuronales Sobreparametrizadas: Una mirada de Campo Medio

Javier Maass Martínez

DIM, Universidad de Chile
Tesis de Magíster en Matemáticas Aplicadas
Memoria de Ingeniería Civil Matemática

25 de marzo de 2024

Objetivo

Entender, mediante un enfoque de Campo Medio, el rol de las *simetrías* (y su *aprovechamiento*) en el entrenamiento de Redes Neuronales Sobreparametrizadas.

Índice

Aprendizaje Supervisado con Redes Neuronales

Problema (genérico) de aprendizaje supervisado

Dados espacios $\mathcal{X}$ e $\mathcal{Y}$, y una ley $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$.

Suponemos que nos llegan datos de la forma:

$$\left(\overset{X}{\text{  }}, \overset{Y}{\textbf{Perro}} \right) \overset{\text{i.i.d.}}{\sim} \pi$$

Problema (genérico) de aprendizaje supervisado

Dados espacios $\mathcal{X}$ e $\mathcal{Y}$, y una ley $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$.
Suponemos que nos llegan datos de la forma:

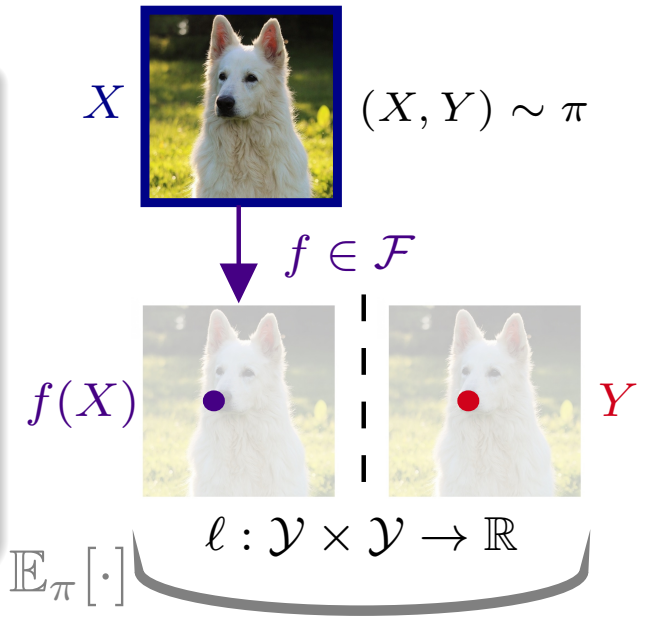
$$\left(\overset{X}{\text{  }}, \overset{Y}{\text{Perro}} \right) \overset{\text{i.i.d.}}{\sim} \pi$$

Problema de Aprendizaje Supervisado

- Conjunto Hipótesis: $\mathcal{F} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$
- Función de pérdida: $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$.
- Riesgo de Población: $R(f) = \mathbb{E}_\pi[\ell(f(X), Y)]$

Nos interesa resolver: $\min_{f \in \mathcal{F}} R(f)$.

i.e. encontrar un modelo en $\mathcal{F}$ que *generalice bien*.



Problema (genérico) de aprendizaje supervisado

Dados espacios $\mathcal{X}$ e $\mathcal{Y}$, y una ley $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$.
Suponemos que nos llegan datos de la forma:

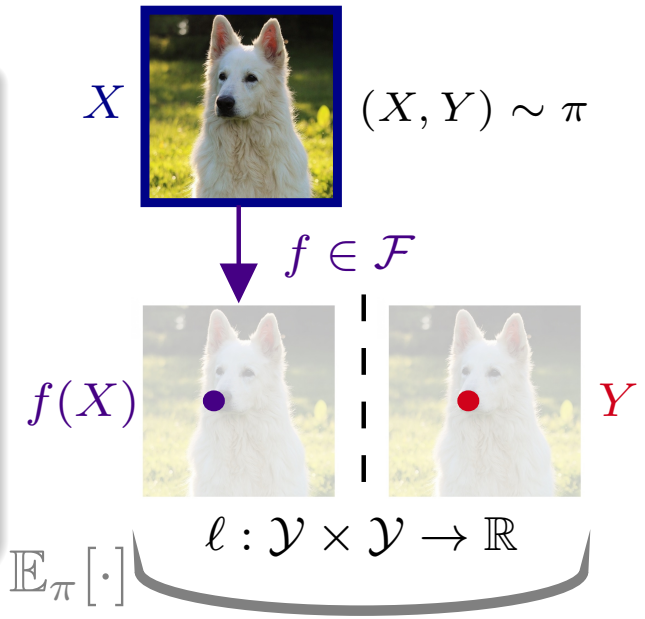
$$\left(\overset{X}{\text{  }}, \overset{Y}{\text{Perro}} \right) \overset{\text{i.i.d.}}{\sim} \pi$$

Problema de Aprendizaje Supervisado

- Conjunto Hipótesis: $\mathcal{F} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$
- Función de pérdida: $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$.
- Riesgo de Población: $R(f) = \mathbb{E}_\pi[\ell(f(X), Y)]$

Nos interesa resolver: $\min_{f \in \mathcal{F}} R(f)$.

i.e. encontrar un modelo en $\mathcal{F}$ que *generalice bien*.



En general, no tenemos acceso a π, tan solo a una muestra i.i.d. $S = (X_k, Y_k)_{k=1}^m$.

Problema (genérico) de aprendizaje supervisado

Dados espacios $\mathcal{X}$ e $\mathcal{Y}$, y una ley $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$.
Suponemos que nos llegan datos de la forma:

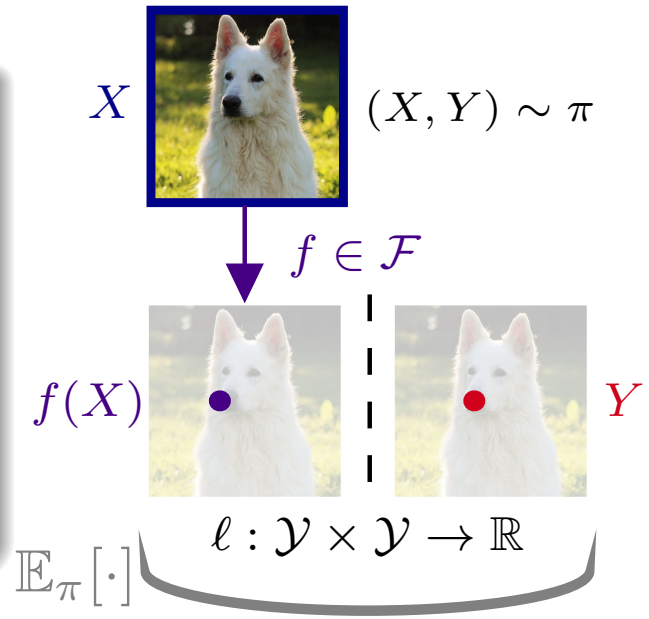
$$\left(\overset{X}{\text{  }}, \overset{Y}{\text{Perro}} \right) \overset{\text{i.i.d.}}{\sim} \pi$$

Problema de Aprendizaje Supervisado

- Conjunto Hipótesis: $\mathcal{F} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$
- Función de pérdida: $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$.
- Riesgo de Población: $R(f) = \mathbb{E}_\pi[\ell(f(X), Y)]$

Nos interesa resolver: $\min_{f \in \mathcal{F}} R(f)$.

i.e. encontrar un modelo en $\mathcal{F}$ que *generalice bien*.



En general, no tenemos acceso a π, tan solo a una muestra i.i.d. $S = (X_k, Y_k)_{k=1}^m$.

Aproximamos usando el *riesgo empírico* $\hat{R}_S(f) = \frac{1}{m} \sum_{k=1}^m \ell(f(X_k), Y_k)$

Problema (genérico) de aprendizaje supervisado

Dados espacios $\mathcal{X}$ e $\mathcal{Y}$, y una ley $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$.

Suponemos que nos llegan datos de la forma:

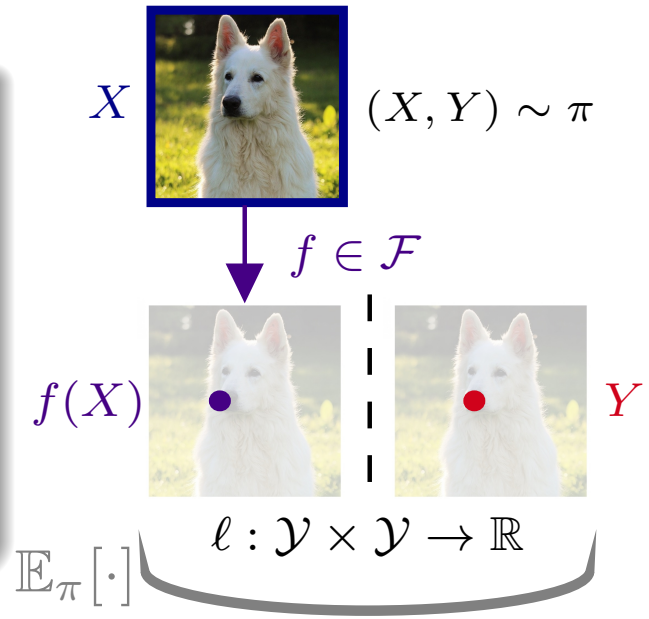
$$\left(\overset{X}{\text{  }}, \overset{Y}{\text{Perro}} \right) \overset{\text{i.i.d.}}{\sim} \pi$$

Problema de Aprendizaje Supervisado

- Conjunto Hipótesis: $\mathcal{F} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$
- Función de pérdida: $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$.
- Riesgo de Población: $R(f) = \mathbb{E}_\pi[\ell(f(X), Y)]$

Nos interesa resolver: $\min_{f \in \mathcal{F}} R(f)$.

i.e. encontrar un modelo en $\mathcal{F}$ que *generalice bien*.

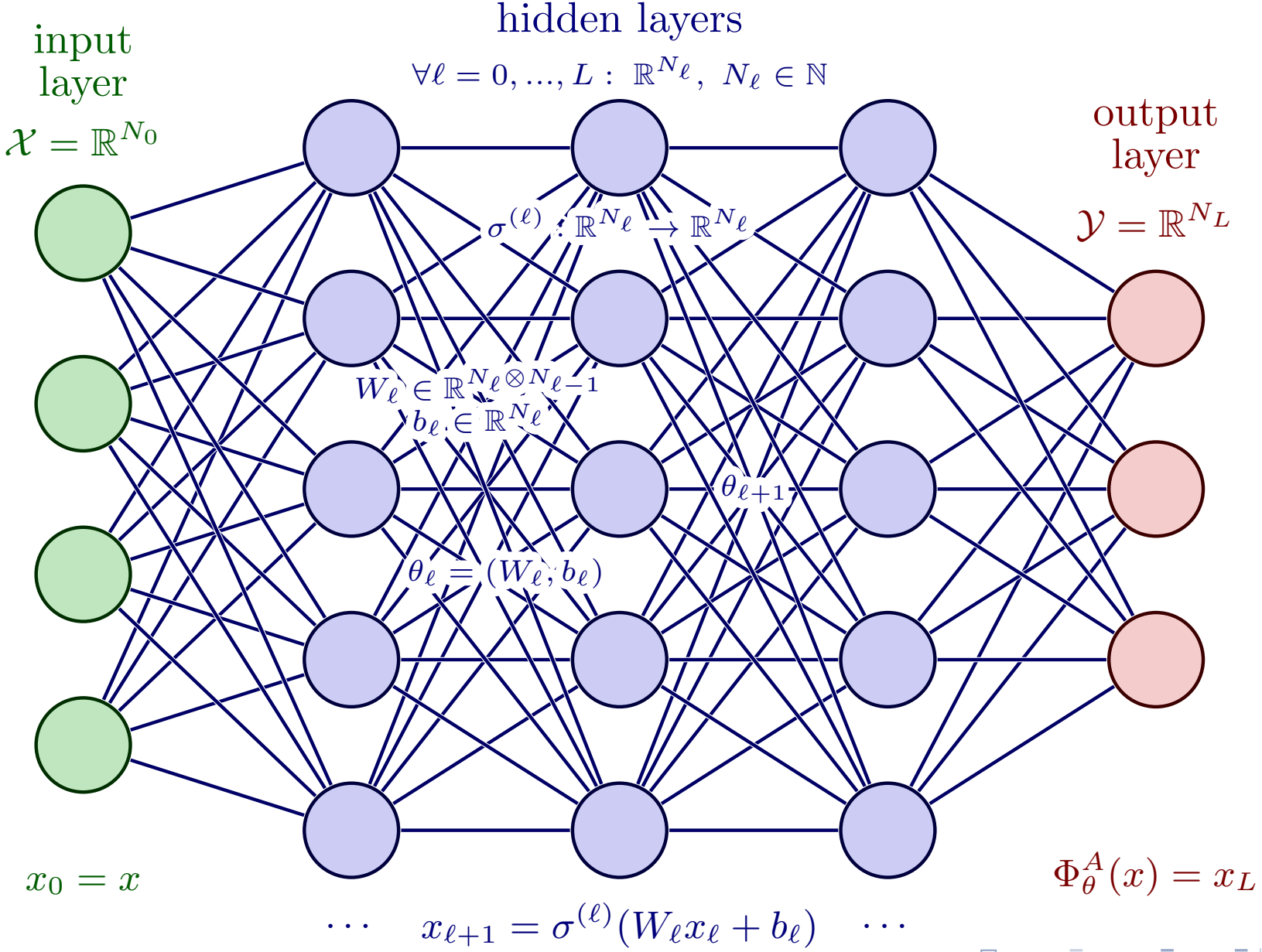


En general, no tenemos acceso a π, tan solo a una muestra i.i.d. $S = (X_k, Y_k)_{k=1}^m$.

Aproximamos usando el *riesgo empírico* $\hat{R}_S(f) = \frac{1}{m} \sum_{k=1}^m \ell(f(X_k), Y_k)$

Es importante que $\mathcal{F}$ sea suficientemente *robusto* para *aprender* sin *memorizar*.

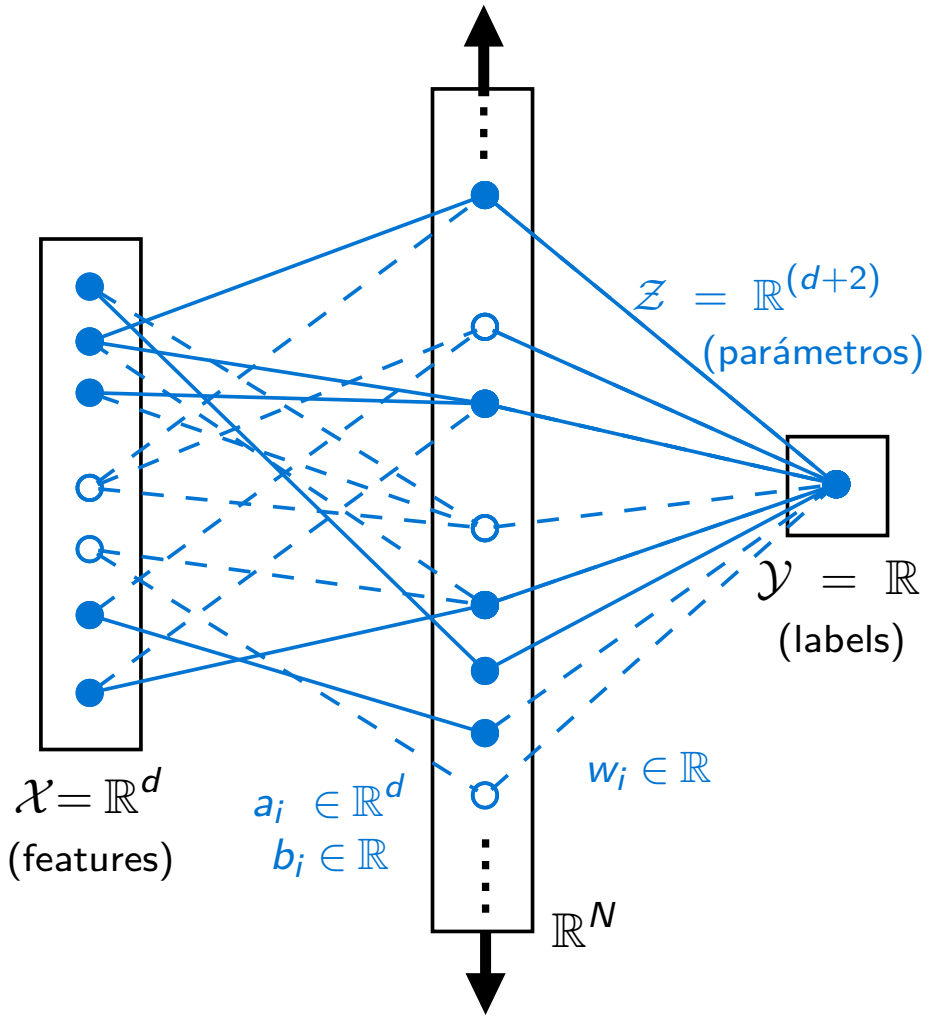
Red Neuronal Multicapa (Fully Connected)



Teoría Mean Field de Redes Neuronales

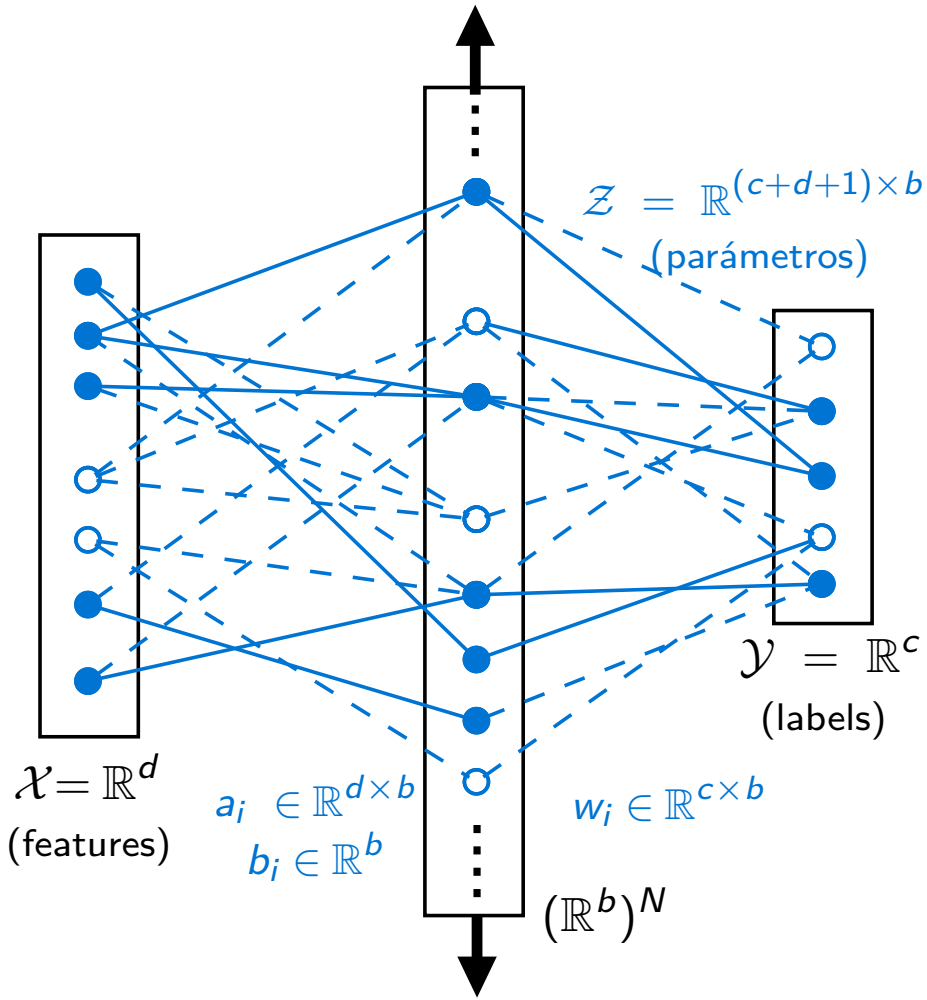
Introducido por Mei u. a. (2018), Sirignano und Spiliopoulos (2018), Rotskoff und Vanden-Eijnden (2022) y Chizat und Bach (2018); profundizado en: Mei u. a. (2019), Sirignano und Spiliopoulos (2019), Chen u. a. (2022b), Bortoli u. a. (2020), Descours u. a. (2023), Hu u. a. (2020), Chizat (2022), Chen u. a. (2022a) y Nitanda u. a. (2022).

Teoría Mean Field de *Shallow NNs*



Modelo de NN con 1 capa oculta: $\sigma_*(x;\theta_i) = w_i\sigma(a_i^T x + b_i)$, y $\theta_i = (w_i, a_i, b_i) \in \mathcal{Z}$.

Teoría Mean Field de *Shallow NNs*



Modelo de NN con 1 capa oculta: $\sigma_*(x;\theta_i) = w_i\sigma(a_i^T x + b_i)$, y $\theta_i = (w_i, a_i, b_i) \in \mathcal{Z}$.

Teoría Mean Field de *Shallow NNs*

Sea $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ la *activación/unidad*.

Sea $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ la *activación/unidad*.

Modelo de *Shallow* NN $(\mathcal{N}_{\sigma_*}(\mathcal{Z}))$

Para $N \in \mathbb{N}$ y $\theta = (\theta_i)_{i=1}^N \in (\mathcal{Z})^N$; es $\Phi_\theta^N : \mathcal{X} \to \mathcal{Y}$:

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

Permite describir diferentes *settings* (e.g. *RBF networks*).

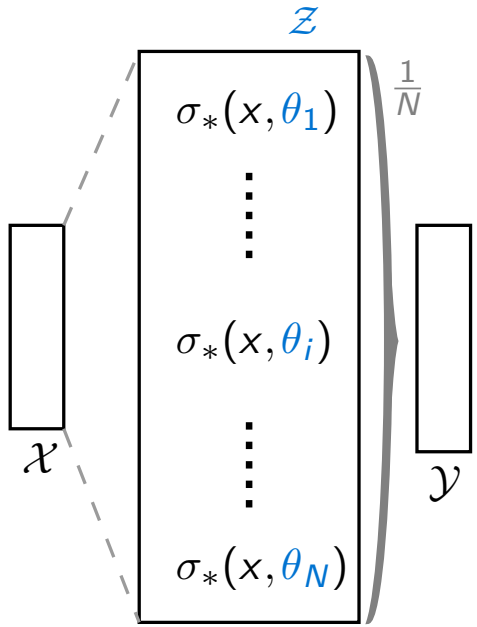
Sea $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ la *activación/unidad*.

Modelo de *Shallow* NN $(\mathcal{N}_{\sigma_*}(\mathcal{Z}))$

Para $N \in \mathbb{N}$ y $\theta = (\theta_i)_{i=1}^N \in (\mathcal{Z})^N$; es $\Phi_\theta^N : \mathcal{X} \to \mathcal{Y}$:

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

Permite describir diferentes *settings* (e.g. *RBF networks*).

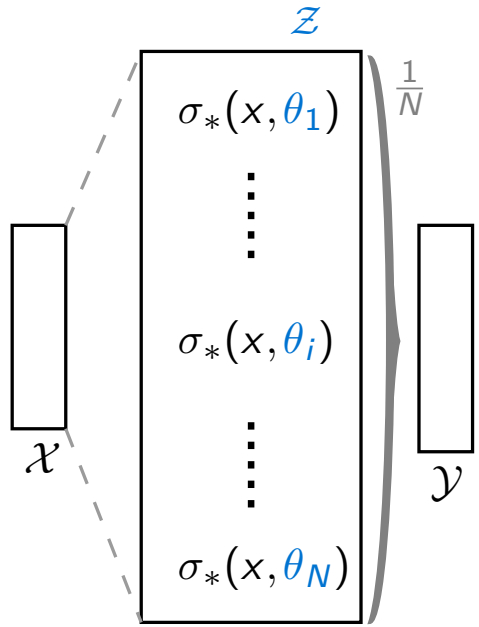


Sea $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ la *activación/unidad*.

Modelo de *Shallow* NN $(\mathcal{N}_{\sigma_*}(\mathcal{Z}))$

Para $N \in \mathbb{N}$ y $\theta = (\theta_i)_{i=1}^N \in (\mathcal{Z})^N$; es $\Phi_\theta^N : \mathcal{X} \to \mathcal{Y}$:

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



Permite describir diferentes *settings* (e.g. *RBF networks*).

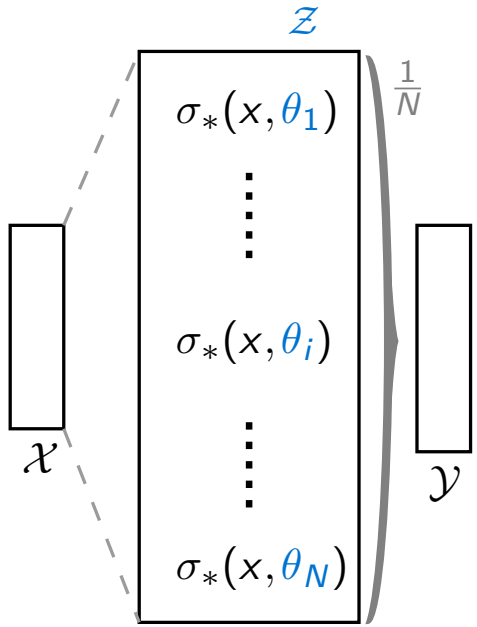
Equivalentemente, es una **integral** contra $\nu_\theta^N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_i}$: $\Phi_\theta^N(x) = \langle \sigma_*(x; \cdot), \nu_\theta^N \rangle$

Sea $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ la *activación/unidad*.

Modelo de *Shallow* NN $(\mathcal{N}_{\sigma_*}(\mathcal{Z}))$

Para $N \in \mathbb{N}$ y $\theta = (\theta_i)_{i=1}^N \in (\mathcal{Z})^N$; es $\Phi_\theta^N : \mathcal{X} \to \mathcal{Y}$:

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



Permite describir diferentes *settings* (e.g. *RBF networks*).

Equivalentemente, es una **integral** contra $\nu_\theta^N = \frac{1}{N} \sum_{i=1}^N \delta_{\theta_i}$: $\Phi_\theta^N(x) = \langle \sigma_*(x; \cdot), \nu_\theta^N \rangle$

Espacio de Barron: $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z})) = \{f : \mathcal{X} \to \mathcal{Y} \mid \exists \gamma \in \mathcal{M}^S(\mathcal{Z}),\ f = \langle \sigma_*, \gamma \rangle\}$

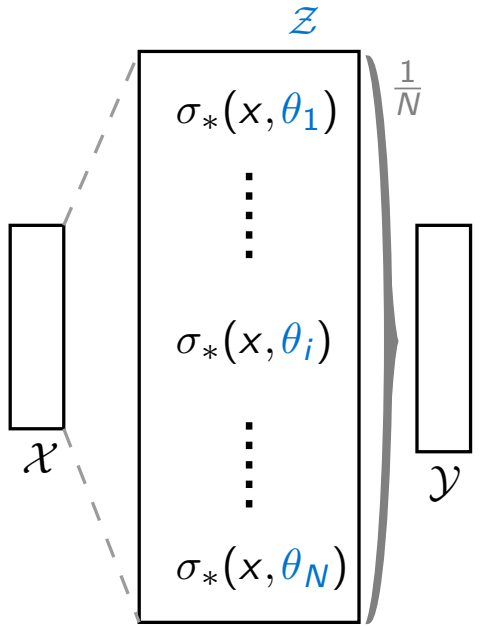
Teoría Mean Field de *Shallow NNs*

Sea $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ la *activación/unidad*.

Modelo de *Shallow* NN ($\mathcal{N}_{\sigma_*}(\mathcal{Z})$)

Para $N \in \mathbb{N}$ y $\theta = (\theta_i)_{i=1}^N \in (\mathcal{Z})^N$; es $\Phi_\theta^N : \mathcal{X} \to \mathcal{Y}$:

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



Permite describir diferentes *settings* (e.g. *RBF networks*).

Equivalentemente, es una **integral** contra $\nu_\theta^N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$: $\Phi_\theta^N(x) = \langle \sigma_*(x;\cdot), \nu_\theta^N \rangle$

Espacio de Barron: $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z})) = \{f : \mathcal{X} \to \mathcal{Y} \mid \exists \gamma \in \mathcal{M}^S(\mathcal{Z}),\ f = \langle \sigma_*, \gamma \rangle\}$

Teorema (Universalidad) **(C'89,H'89,B'93,RVE'18)**

Bajo **(C.T.)**, $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z}))$ es **subespacio lineal denso** de $L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})$.

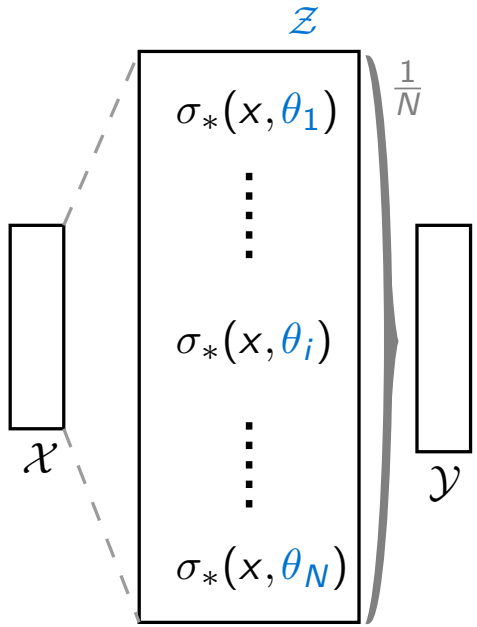
Teoría Mean Field de *Shallow NNs*

Sea $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ la *activación/unidad*.

Modelo de *Shallow* NN ($\mathcal{N}_{\sigma_*}(\mathcal{Z})$)

Para $N \in \mathbb{N}$ y $\theta = (\theta_i)_{i=1}^N \in (\mathcal{Z})^N$; es $\Phi_\theta^N : \mathcal{X} \to \mathcal{Y}$:

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



Permite describir diferentes *settings* (e.g. *RBF networks*).

Equivalentemente, es una **integral** contra $\nu_\theta^N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$: $\Phi_\theta^N(x) = \langle \sigma_*(x;\cdot), \nu_\theta^N \rangle$

Espacio de Barron: $\mathcal{F}_{\sigma_*}(\mathcal{M}^S(\mathcal{Z})) = \{f : \mathcal{X} \to \mathcal{Y} \mid \exists \gamma \in \mathcal{M}^S(\mathcal{Z}),\ f = \langle \sigma_*, \gamma \rangle\}$

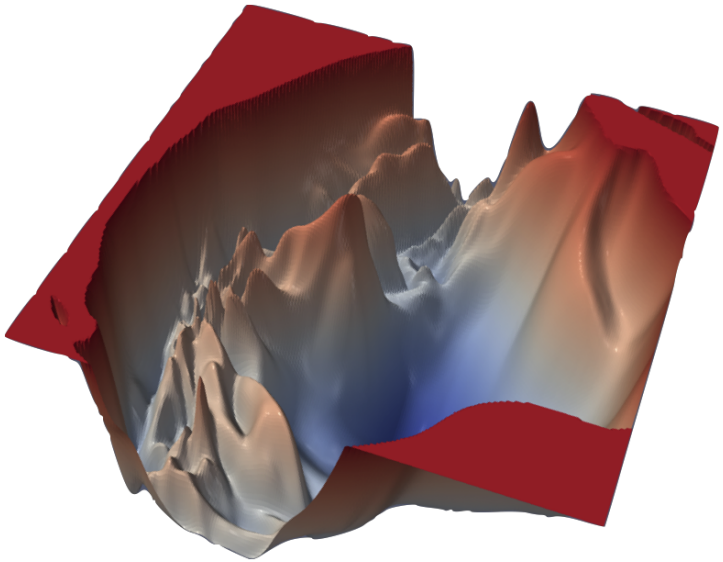
Corolario (Universalidad) **(C'89,H'89,B'93,RVE'18)**

[**(C.T.)** + $\pi_{\mathcal{X}}$ de soporte compacto] $\Rightarrow$ $\mathcal{N}_{\sigma_*}(\mathcal{Z})$ **denso** en $L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})$.

Teoría Mean Field de *Shallow NNs*

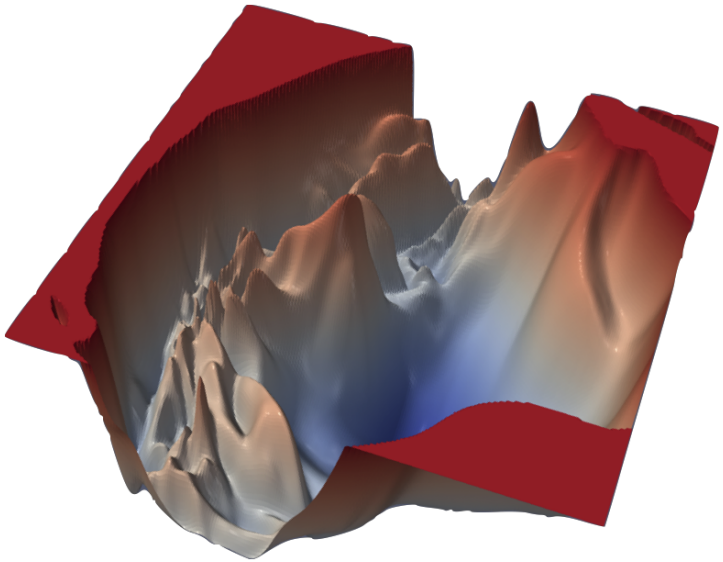
$\min_{\theta \in \mathcal{Z}^N} \mathbb{E}_\pi[\ell(\Phi_\theta^N(X), Y)]$

Altamente complejo y **no convexo**.



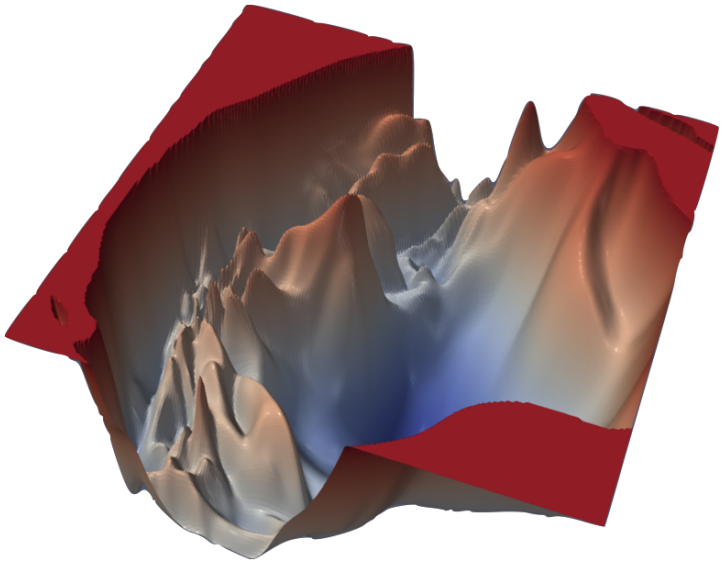
Teoría Mean Field de *Shallow NNs*

$\min_{\theta \in \mathcal{Z}^N} \mathbb{E}_\pi[\ell(\Phi_\theta^N(X), Y)]$
Altamente complejo y **no convexo**.



Recordando: $\Phi_\theta^N = \langle \sigma_*, \nu_\theta^N \rangle \in \mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{Z}))$

$\min_{\theta \in \mathcal{Z}^N} \mathbb{E}_\pi[\ell(\Phi_\theta^N(X), Y)]$
Altamente complejo y **no convexo**.



Recordando: $\Phi_\theta^N = \langle \sigma_*, \nu_\theta^N \rangle \in \mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{Z}))$

Convexificación del problema

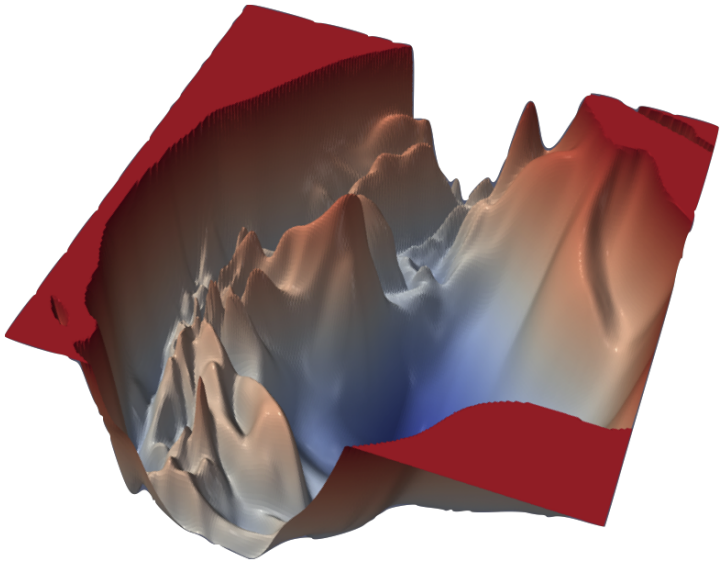
Podemos definir (para ℓ convexo):

$$R(\mu) := \mathbb{E}_\pi[\ell(\langle \sigma_*(X, \cdot), \mu \rangle, Y)]$$

con lo que $\min_{\mu \in \mathcal{P}(\mathcal{Z})} R(\mu)$ es **convexo**.

Teoría Mean Field de *Shallow NNs*

$\min_{\theta \in \mathcal{Z}^N} \mathbb{E}_\pi[\ell(\Phi_\theta^N(X), Y)]$
Altamente complejo y **no convexo**.



Recordando: $\Phi_\theta^N = \langle \sigma_*, \nu_\theta^N \rangle \in \mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{Z}))$

Convexificación del problema

Podemos definir (para ℓ convexo):

$$R(\mu) := \mathbb{E}_\pi[\ell(\langle \sigma_*(X, \cdot), \mu \rangle, Y)]$$

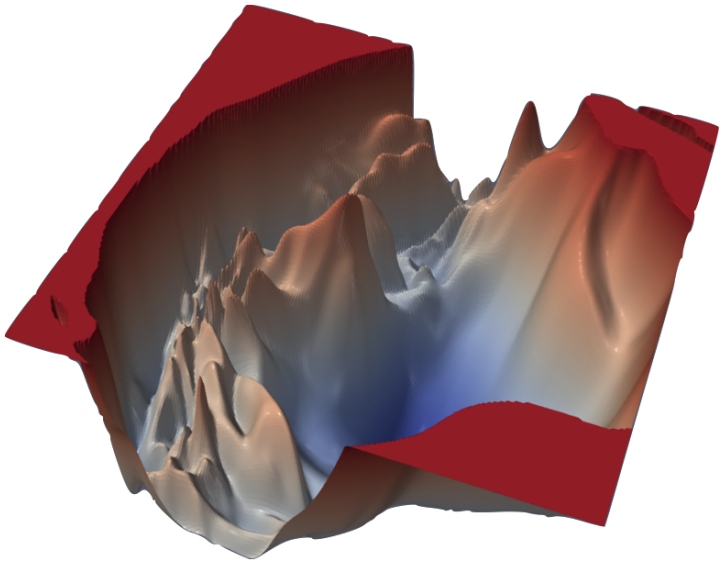
con lo que $\min_{\mu \in \mathcal{P}(\mathcal{Z})} R(\mu)$ es **convexo**.

Además, **(HRSS'19)**

$$\left| \inf_{\theta \in \mathcal{Z}^N} R\left(\nu_\theta^N\right) - \inf_{\mu \in \mathcal{P}_2(\mathcal{Z})} R(\mu) \right| \leq \frac{\text{cte}}{N}$$

Teoría Mean Field de *Shallow NNs*

$\min_{\theta \in \mathcal{Z}^N} \mathbb{E}_\pi[\ell(\Phi_\theta^N(X), Y)]$
Altamente complejo y **no convexo**.



Recordando: $\Phi_\theta^N = \langle \sigma_*, \nu_\theta^N \rangle \in \mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{Z}))$

Convexificación del problema

Podemos definir (para ℓ convexo):

$$R(\mu) := \mathbb{E}_\pi[\ell(\langle \sigma_*(X, \cdot), \mu \rangle, Y)]$$

con lo que $\min_{\mu \in \mathcal{P}(\mathcal{Z})} R(\mu)$ es **convexo**.

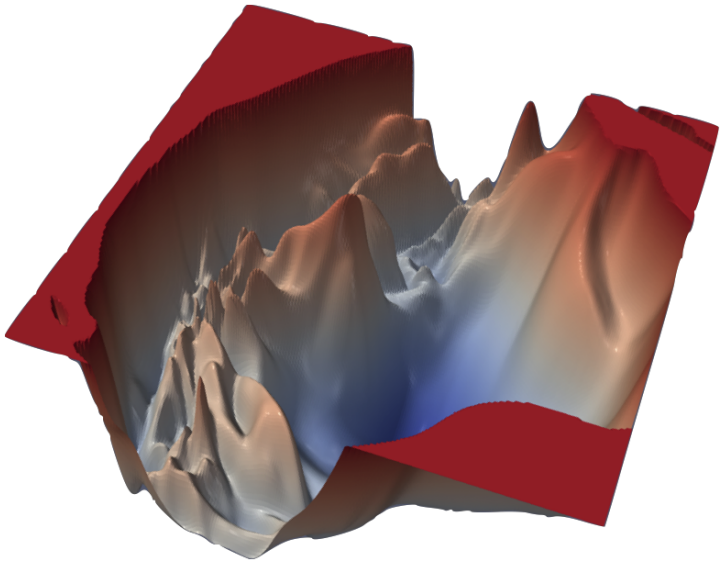
Además, **(HRSS'19)**

$$\left| \inf_{\theta \in \mathcal{Z}^N} R\left(\nu_\theta^N\right) - \inf_{\mu \in \mathcal{P}_2(\mathcal{Z})} R(\mu) \right| \leq \frac{\text{cte}}{N}$$

$$\left[\textbf{Universalidad} + \left(\ell(y, \hat{y}) = \|y - \hat{y}\|_{\mathcal{Y}}^2 \right) \right] \xrightarrow{\text{Lema 2}} \left[\inf_{\mu \in \mathcal{P}(\mathcal{Z})} R(\mu) = \inf_{f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})} R(f) =: R_* \right]$$

Teoría Mean Field de *Shallow NNs*

$\min_{\theta \in \mathcal{Z}^N} \mathbb{E}_\pi[\ell(\Phi_\theta^N(X), Y)]$
Altamente complejo y **no convexo**.



Recordando: $\Phi_\theta^N = \langle \sigma_*, \nu_\theta^N \rangle \in \mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{Z}))$

Convexificación del problema

Podemos definir (para ℓ convexo):

$$R(\mu) := \mathbb{E}_\pi[\ell(\langle \sigma_*(X, \cdot), \mu \rangle, Y)]$$

con lo que $\min_{\mu \in \mathcal{P}(\mathcal{Z})} R(\mu)$ es **convexo**.

Además, **(HRSS'19)**

$$\left| \inf_{\theta \in \mathcal{Z}^N} R\left(\nu_\theta^N\right) - \inf_{\mu \in \mathcal{P}_2(\mathcal{Z})} R(\mu) \right| \leq \frac{\text{cte}}{N}$$

$$\left[\textbf{Universalidad} + \left(\ell(y, \hat{y}) = \|y - \hat{y}\|_{\mathcal{Y}}^2 \right) \right] \xrightarrow{\text{Lema 2}} \left[\inf_{\mu \in \mathcal{P}(\mathcal{Z})} R(\mu) = \inf_{f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})} R(f) =: R_* \right]$$

¿Cómo se **encuentra** ese óptimo (en la práctica) con el entrenamiento de la NN?

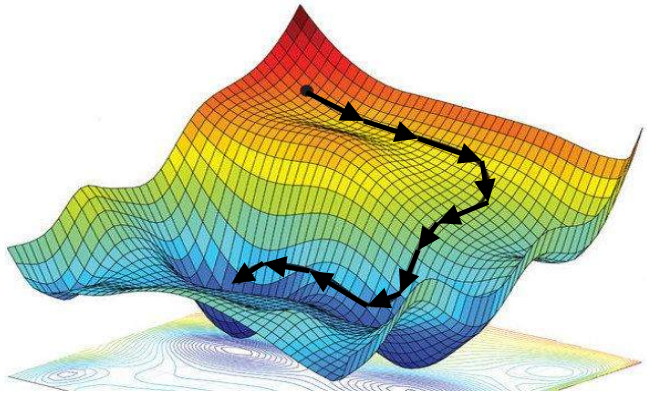
¡Sólo disponemos de muestras $(X_k, Y_k)_{k \in \mathbb{N}} \overset{i.i.d.}{\sim} \pi$!

Stochastic Gradient Descent (SGD)

- **Inicialización**: $(\theta_i^0)_{i=1}^N \overset{i.i.d.}{\sim} \mu_0 \in \mathcal{P}_2(\mathcal{Z})$
- $\forall k \in \mathbb{N}$, $\forall i \in \{1, \dots, N\}$ se itera:

$$\theta_i^{k+1} = \theta_i^k - s_k^N \partial_1 \ell(\Phi_\theta^N(X_k), Y_k) \nabla_{\theta_i}(\sigma_*(X_k; \theta_i^k))$$

Donde $s_k^N = \varepsilon_N \varsigma(k\varepsilon_N)$ es el *learning rate*, con $\varepsilon_N > 0$ y $\varsigma : \mathbb{R}_+ \to \mathbb{R}_+$ una función regular.

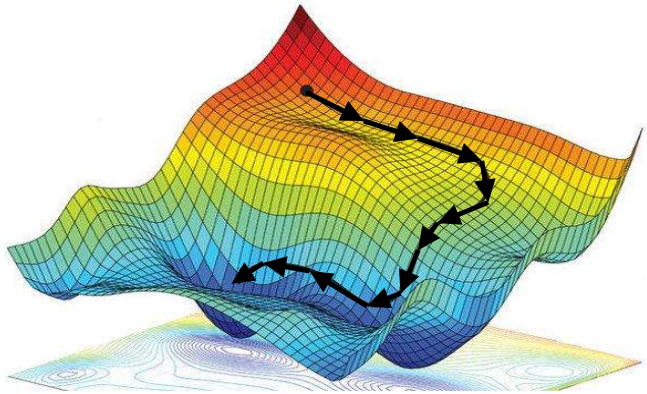


Stochastic Gradient Descent (SGD)

- **Inicialización**: $(\theta_i^0)_{i=1}^N \overset{i.i.d.}{\sim} \mu_0 \in \mathcal{P}_2(\mathcal{Z})$
- $\forall k \in \mathbb{N}$, $\forall i \in \{1, \dots, N\}$ se itera:

$$\theta_i^{k+1} = \theta_i^k - s_k^N \partial_1 \ell(\Phi_\theta^N(X_k), Y_k) \nabla_{\theta_i}(\sigma_*(X_k; \theta_i^k))$$

Donde $s_k^N = \varepsilon_N \varsigma(k\varepsilon_N)$ es el *learning rate*, con $\varepsilon_N > 0$ y $\varsigma : \mathbb{R}_+ \to \mathbb{R}_+$ una función regular.



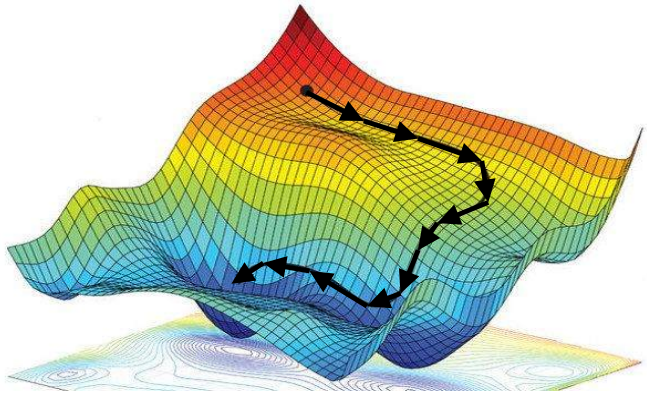
Se podría incluir también *minibatches*, una *penalización* e incluso *ruido gaussiano*.

Stochastic Gradient Descent (SGD)

- **Inicialización**: $(\theta_i^0)_{i=1}^N \overset{i.i.d.}{\sim} \mu_0 \in \mathcal{P}_2(\mathcal{Z})$
- $\forall k \in \mathbb{N}$, $\forall i \in \{1, \dots, N\}$ se itera:

$$\theta_i^{k+1} = \theta_i^k - s_k^N \partial_1 \ell(\Phi_\theta^N(X_k), Y_k) \nabla_{\theta_i}(\sigma_*(X_k; \theta_i^k))$$

Donde $s_k^N = \varepsilon_N \varsigma(k\varepsilon_N)$ es el *learning rate*, con $\varepsilon_N > 0$ y $\varsigma : \mathbb{R}_+ \to \mathbb{R}_+$ una función regular.



Se podría incluir también *minibatches*, una *penalización* e incluso *ruido gaussiano*.

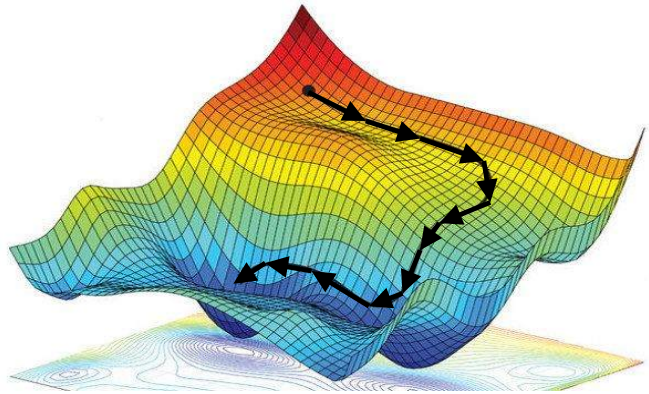
¿Cómo se relaciona con $R(\mu)$?

Stochastic Gradient Descent (SGD)

- **Inicialización**: $(\theta_i^0)_{i=1}^N \overset{i.i.d.}{\sim} \mu_0 \in \mathcal{P}_2(\mathcal{Z})$
- $\forall k \in \mathbb{N}$, $\forall i \in \{1, \dots, N\}$ se itera:

$$\theta_i^{k+1} = \theta_i^k - s_k^N \, \partial_1 \ell(\Phi_\theta^N(X_k), Y_k) \nabla_{\theta_i}(\sigma_*(X_k; \theta_i^k))$$

Donde $s_k^N = \varepsilon_N \varsigma(k\varepsilon_N)$ es el *learning rate*, con $\varepsilon_N > 0$ y $\varsigma : \mathbb{R}_+ \to \mathbb{R}_+$ una función regular.



Se podría incluir también *minibatches*, una *penalización* e incluso *ruido gaussiano*.

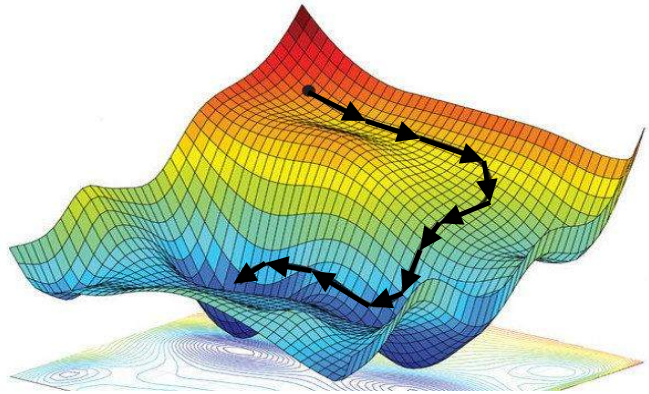
¿Cómo se relaciona con $R(\mu)$? $\to$ Se estudia la evolución de $\nu_k^N := \nu_{\theta^k}^N = \dfrac{1}{N} \displaystyle\sum_{i=1}^N \delta_{\theta_i^k}$

Stochastic Gradient Descent (SGD)

- **Inicialización**: $(\theta_i^0)_{i=1}^N \overset{i.i.d.}{\sim} \mu_0 \in \mathcal{P}_2(\mathcal{Z})$
- $\forall k \in \mathbb{N}$, $\forall i \in \{1, \dots, N\}$ se itera:

$$\theta_i^{k+1} = \theta_i^k - s_k^N \, \partial_1 \ell(\Phi_\theta^N(X_k), Y_k) \nabla_{\theta_i}(\sigma_*(X_k; \theta_i^k))$$

Donde $s_k^N = \varepsilon_N \varsigma(k\varepsilon_N)$ es el *learning rate*, con $\varepsilon_N > 0$ y $\varsigma : \mathbb{R}_+ \to \mathbb{R}_+$ una función regular.



Se podría incluir también *minibatches*, una *penalización* e incluso *ruido gaussiano*.

¿Cómo se relaciona con $R(\mu)$? $\to$ Se estudia la evolución de $\nu_k^N := \nu_{\theta^k}^N = \dfrac{1}{N} \displaystyle\sum_{i=1}^N \delta_{\theta_i^k}$

Teorema (Propagación de Caos) **(MMN'18,SS'18,CB'18,RVE'18)**

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

donde $(\mu_t)_{t\geq 0}$ satisface el **Flujo de Gradiente de Wasserstein** del riesgo R.

Flujo de Gradiente de Wasserstein (WGF) de un funcional $R : \mathcal{P}_2(\mathcal{Z}) \to \mathbb{R}$

Es cualquier trayectoria $(\mu_t)_{t\in[0,T[} \subseteq \mathcal{P}_2(\mathcal{Z})$ que (débilmente) satisface:

$$\partial_t \mu_t = \varsigma(t)\,\mathrm{div}_\theta\left(D_\mu R(\mu_t, \cdot)\mu_t\right)$$

Donde $D_\mu R(\mu, \theta) = \nabla_\theta \frac{\partial R}{\partial \mu}(\mu, \theta)$ es la *derivada intrínseca* de R

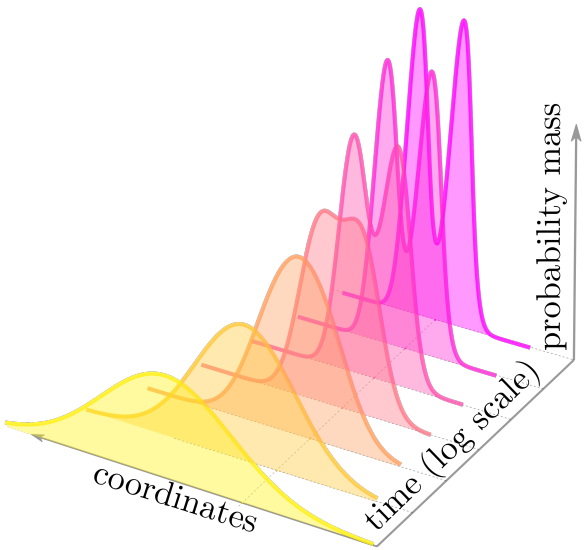
Límite Mean Field de *Shallow NNs*

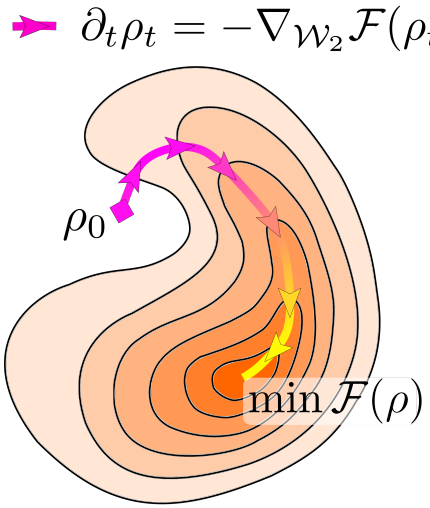
Flujo de Gradiente de Wasserstein (WGF) de un funcional $R : \mathcal{P}_2(\mathcal{Z}) \to \mathbb{R}$

Es cualquier trayectoria $(\mu_t)_{t\in[0,T[} \subseteq \mathcal{P}_2(\mathcal{Z})$ que (débilmente) satisface:

$$\partial_t \mu_t = \varsigma(t)\,\mathrm{div}_\theta\left(D_\mu R(\mu_t, \cdot)\mu_t\right)$$

Donde $D_\mu R(\mu, \theta) = \nabla_\theta \frac{\partial R}{\partial \mu}(\mu, \theta)$ es la *derivada intrínseca* de R



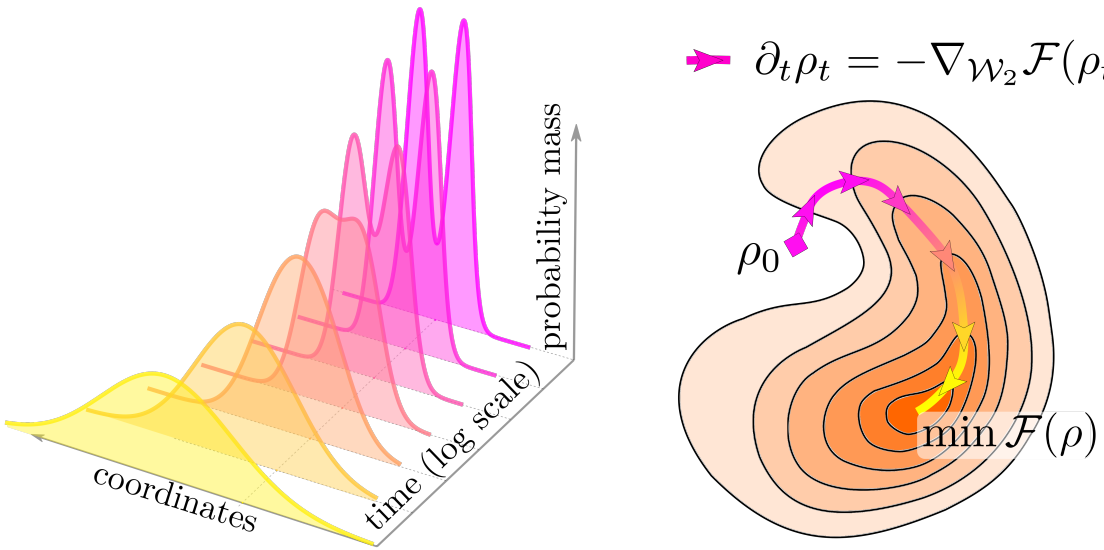


Flujo de Gradiente de Wasserstein (WGF) de un funcional $R : \mathcal{P}_2(\mathcal{Z}) \to \mathbb{R}$

Es cualquier trayectoria $(\mu_t)_{t\in[0,T[} \subseteq \mathcal{P}_2(\mathcal{Z})$ que (débilmente) satisface:

$$\partial_t \mu_t = \varsigma(t)\,\mathrm{div}_\theta\left(D_\mu R(\mu_t, \cdot)\mu_t\right)$$

Donde $D_\mu R(\mu, \theta) = \nabla_\theta \frac{\partial R}{\partial \mu}(\mu, \theta)$ es la *derivada intrínseca* de R



El WGF equivale a la SDE no lineal: $\mathrm{d}Z_t = -\varsigma(t) D_\mu R(\mu_t, Z_t)\mathrm{d}t$ con $\mu_t = \textbf{Ley}(Z_t)$. Esta es la llamada dinámica de **McKean-Vlasov**.

Teorema (Convergencia al Óptimo SIN ruido) **(CB'18)**

Sea $(\mu_t)_{t\geq 0}$ el WGF de R. Bajo **(C.T.)**, si $W_2(\mu_t, \mu_\infty) \xrightarrow{t\to\infty} 0$, entonces:

$$R(\mu_\infty) = \min_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu)$$

i.e. El entrenamiento converge al **óptimo global** del problema! (cuando converge)

Teorema (Convergencia al Óptimo SIN ruido) **(CB'18)**

Sea $(\mu_t)_{t\geq 0}$ el WGF de R. Bajo **(C.T.)**, si $W_2(\mu_t, \mu_\infty) \xrightarrow{t\to\infty} 0$, entonces:

$$R(\mu_\infty) = \min_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu)$$

i.e. El entrenamiento converge al **óptimo global** del problema! (cuando converge)

Riesgo Regularizado: $R^{\tau,\beta}(\mu) = R(\mu) + \tau\int r d\mu + \beta H_\lambda(\mu)$ $(\tau, \beta > 0,\ r : \mathcal{Z} \to \mathbb{R})$.

Noisy SGD: $\theta_i^{k+1} = \theta_i^k - s_k^N \left(\partial_1 \ell(\Phi_\theta^N(X_k), Y_k)\nabla_{\theta_i}(\sigma_*(X_k;\theta_i^k)) + \tau\nabla_{\theta_i} r(\theta_i^k)\right) + \sqrt{2\beta s_k^N}\xi_i^k$

Límite Mean Field de *Shallow NNs*

Teorema (Convergencia al Óptimo SIN ruido) **(CB'18)**

Sea $(\mu_t)_{t\geq 0}$ el WGF de R. Bajo **(C.T.)**, si $W_2(\mu_t, \mu_\infty) \xrightarrow{t\to\infty} 0$, entonces:

$$R(\mu_\infty) = \min_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu)$$

i.e. El entrenamiento converge al **óptimo global** del problema! (cuando converge)

Riesgo Regularizado: $R^{\tau,\beta}(\mu) = R(\mu) + \tau\int r d\mu + \beta H_\lambda(\mu)$ $(\tau,\beta>0,\ r:\mathcal{Z}\to\mathbb{R})$.

Noisy SGD: $\theta_i^{k+1} = \theta_i^k - s_k^N\left(\partial_1\ell(\Phi_\theta^N(X_k),Y_k)\nabla_{\theta_i}(\sigma_*(X_k;\theta_i^k)) + \tau\nabla_{\theta_i} r(\theta_i^k)\right) + \sqrt{2\beta s_k^N}\xi_i^k$

Teorema (Convergencia al Óptimo) **(MMN'18,HRSS'19,CRW'22)**

Sea $(\mu_t)_{t\geq 0}$ el WGF de $R^{\tau,\beta}$ y $\mu_*^{\tau,\beta}$ su (único) mínimo. Bajo **(C.T.)**:

$$W_2(\mu_t,\mu_*^{\tau,\beta}) \xrightarrow[t\to\infty]{} 0 \quad \left(\text{ y } H_\lambda(\mu_t||\mu_*^{\tau,\beta}) \xrightarrow[t\to\infty]{} 0\right)$$

Teorema (Convergencia al Óptimo SIN ruido) **(CB'18)**

Sea $(\mu_t)_{t\geq 0}$ el WGF de R. Bajo **(C.T.)**, si $W_2(\mu_t, \mu_\infty) \xrightarrow{t\to\infty} 0$, entonces:

$$R(\mu_\infty) = \min_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu)$$

i.e. El entrenamiento converge al **óptimo global** del problema! (cuando converge)

Riesgo Regularizado: $R^{\tau,\beta}(\mu) = R(\mu) + \tau\int r d\mu + \beta H_\lambda(\mu)$ $(\tau,\beta>0,\ r:\mathcal{Z}\to\mathbb{R})$.

Noisy SGD: $\theta_i^{k+1} = \theta_i^k - s_k^N\left(\partial_1\ell(\Phi_\theta^N(X_k), Y_k)\nabla_{\theta_i}(\sigma_*(X_k;\theta_i^k)) + \tau\nabla_{\theta_i} r(\theta_i^k)\right) + \sqrt{2\beta s_k^N}\xi_i^k$

Teorema (Convergencia al Óptimo) **(MMN'18,HRSS'19,CRW'22)**

Sea $(\mu_t)_{t\geq 0}$ el WGF de $R^{\tau,\beta}$ y $\mu_*^{\tau,\beta}$ su (único) mínimo. Bajo **(C.T.)**:

$$W_2(\mu_t, \mu_*^{\tau,\beta}) \xrightarrow[t\to\infty]{} 0 \quad \left(\text{ y } H_\lambda(\mu_t||\mu_*^{\tau,\beta}) \xrightarrow[t\to\infty]{} 0\right)$$

Prop. 10 (HRSS'19): $R_\nu^{\tau,\beta}$ Γ-converge a R con $\tau,\beta\downarrow 0$; y: $\overline{\lim_{\tau,\beta\to 0}} R(\mu_*^{\tau,\beta}) = \inf_{\mu\in\mathcal{P}_2(\mathcal{Z})} R(\mu)$

Diferencias con el caso de 1 capa (SS'19,AOY'19,NTP'20)

- Parámetros no son *estadísticamente independientes*.
- *Unidad básica* son los *caminos de pesos* en la red.
- **Dificultades**: *scalings distintos* entre capas *ocultas* y *extremas*.
  (Se han resuelto con *random features* y/o usando *learning rates* específicos).

Diferencias con el caso de 1 capa (SS'19,AOY'19,NTP'20)

- Parámetros no son *estadísticamente independientes*.
- *Unidad básica* son los *caminos de pesos* en la red.
- **Dificultades**: *scalings distintos* entre capas *ocultas* y *extremas*. (Se han resuelto con *random features* y/o usando *learning rates* específicos).

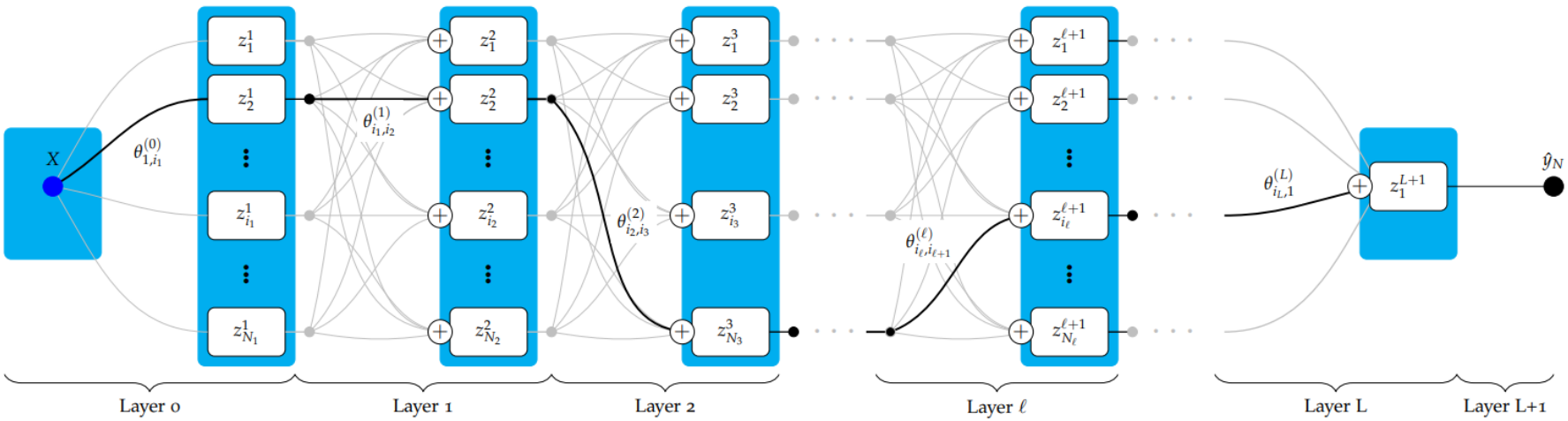


Figura: *Caminos de pesos* en una arquitectura de NN *fully-connected*

Diferencias con el caso de 1 capa (SS'19,AOY'19,NTP'20)

- Parámetros no son *estadísticamente independientes*.
- *Unidad básica* son los *caminos de pesos* en la red.
- **Dificultades**: *scalings distintos* entre capas *ocultas* y *extremas*. (Se han resuelto con *random features* y/o usando *learning rates* específicos).

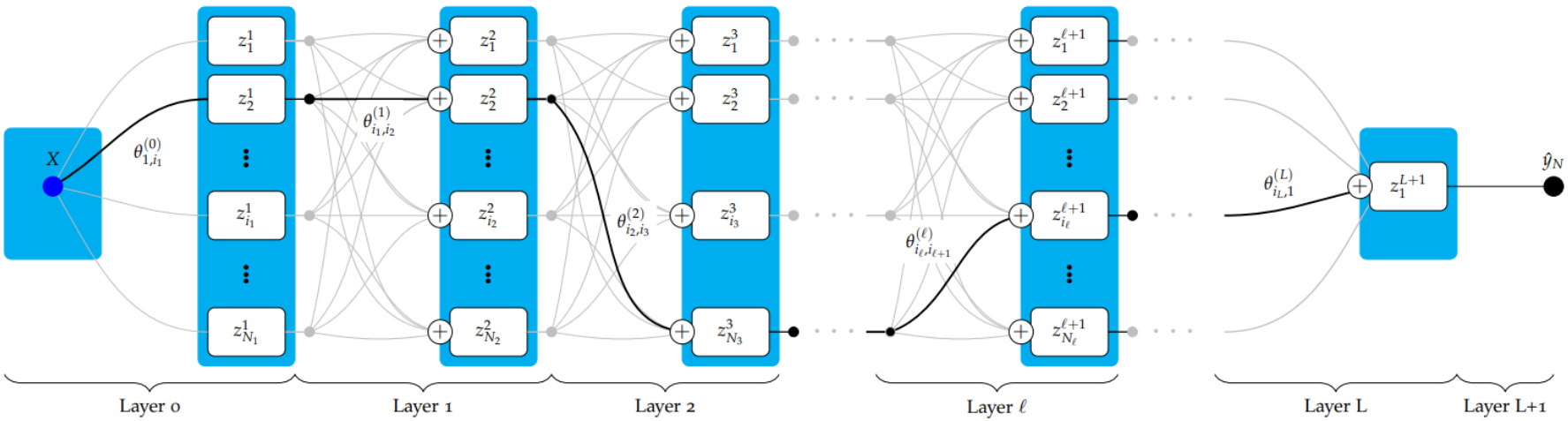


Figura: *Caminos de pesos* en una arquitectura de NN *fully-connected*

El estudio del caso multicapa se realizará como trabajo futuro.

Aprovechamiento de Simetrías con NNs

Contexto de *Geometric Deep Learning* (review en Bronstein u. a. (2021)); también influencias de Lyle u. a. (2020),Chen u. a. (2020), Elesedy und Zaidi (2021), Finzi u. a. (2021) y Flinth und Ohlsson (2023).

Datos Equivariantes

Digamos que hay un grupo G de *simetrías* que **actúa** sobre $\mathcal{X}, \mathcal{Y}$ y $\mathcal{Z}$.

Datos Equivariantes

Digamos que hay un grupo G de *simetrías* que **actúa** sobre $\mathcal{X}, \mathcal{Y}$ y $\mathcal{Z}$.

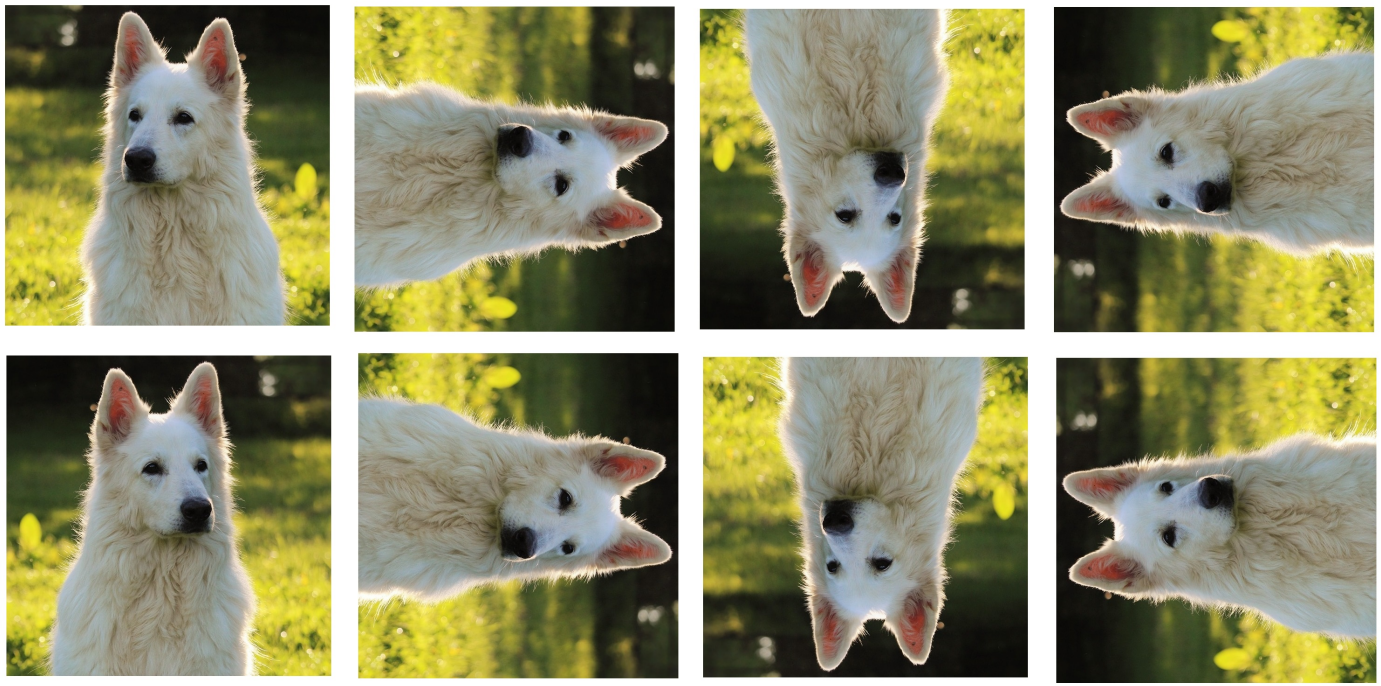


Figura: Imagen de Perro bajo la acción de $G = D_4$

Datos Equivariantes

Digamos que hay un grupo G de *simetrías* que **actúa** sobre $\mathcal{X}, \mathcal{Y}$ y $\mathcal{Z}$.

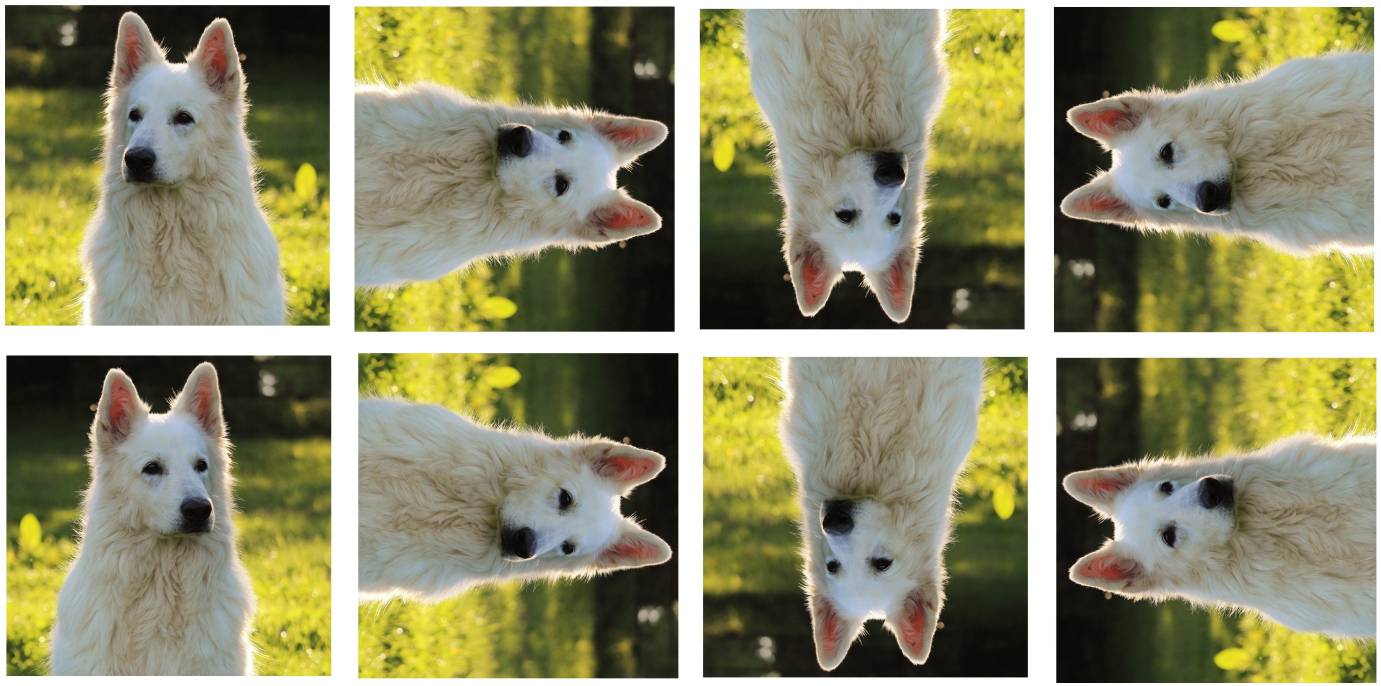


Figura: Imagen de Perro bajo la acción de $G = D_4$

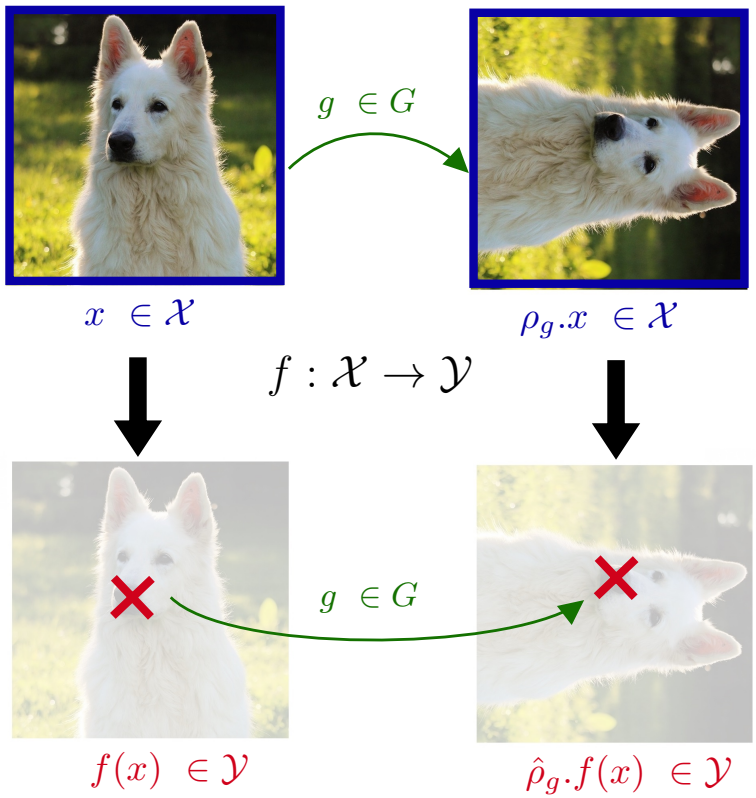
Consideraremos G **compacto** y que actúa via **representaciones ortogonales** $(G \circlearrowright \mathcal{X})$.

Datos Equivariantes

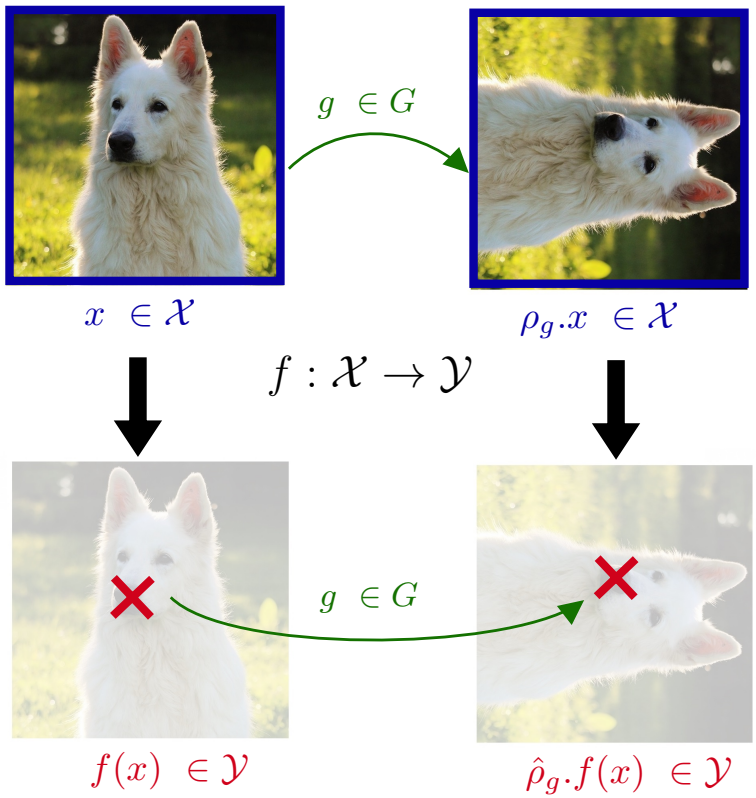
Sea $G \curvearrowright_{\rho} \mathcal{X}$, $G \curvearrowright_{\hat{\rho}} \mathcal{Y}$ y $G \curvearrowright_{M} \mathcal{Z}$.

Datos Equivariantes

Sea $G \curvearrowright_\rho \mathcal{X}$, $G \curvearrowright_{\hat{\rho}} \mathcal{Y}$ y $G \curvearrowright_M \mathcal{Z}$.



Sea $G \curvearrowright_\rho \mathcal{X}$, $G \curvearrowright_{\hat{\rho}} \mathcal{Y}$ y $G \curvearrowright_M \mathcal{Z}$.



Funciones G-Equivariantes

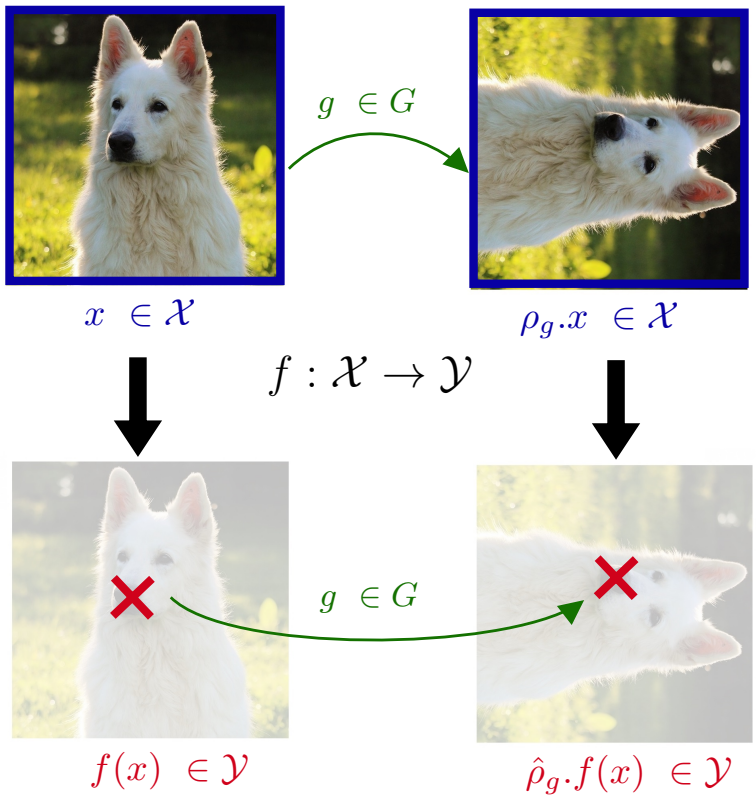
$f : \mathcal{X} \to \mathcal{Y}$ es $G-$**equivariante** si

$$\forall g \in G : \; f \circ \rho_g = \hat{\rho}_g \circ f$$

Si $\hat{\rho} \equiv id$, f se dice G-**invariante**.

Datos Equivariantes

Sea $G \curvearrowright_\rho \mathcal{X}$, $G \curvearrowright_{\hat{\rho}} \mathcal{Y}$ y $G \curvearrowright_M \mathcal{Z}$.



Funciones G-Equivariantes

$f : \mathcal{X} \to \mathcal{Y}$ es $G-$**equivariante** si

$$\forall g \in G : \; f \circ \rho_g = \hat{\rho}_g \circ f$$

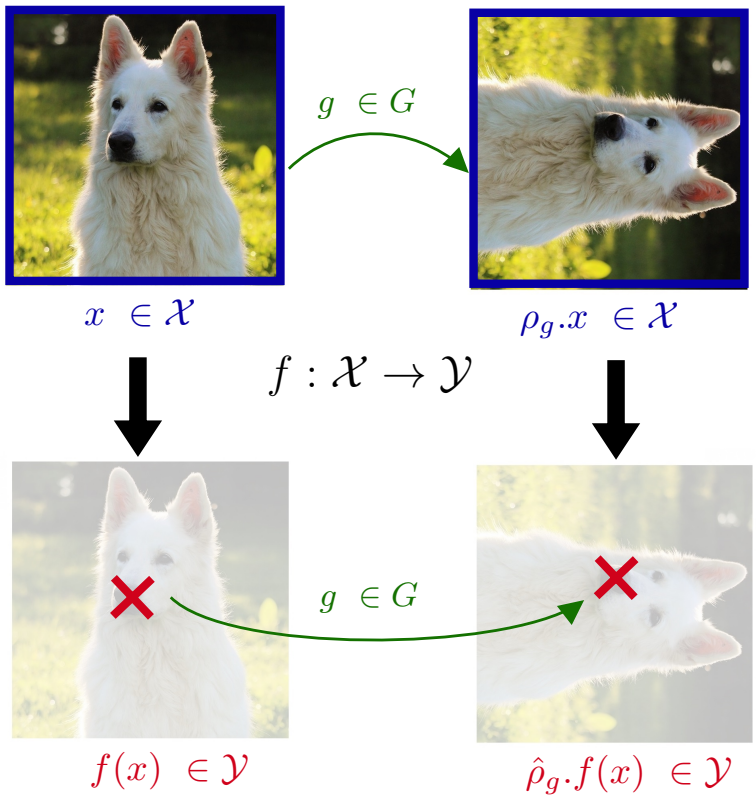
Si $\hat{\rho} \equiv id$, f se dice G-**invariante**.

Medidas G-invariantes, $\mathcal{M}^G(\mathcal{Z})$

$\mu \in \mathcal{M}(\mathcal{Z})$ tq: $\forall g \in G$, $M_g \# \mu = \mu$

Datos Equivariantes

Sea $G \curvearrowright_\rho \mathcal{X}$, $G \curvearrowright_{\hat{\rho}} \mathcal{Y}$ y $G \curvearrowright_M \mathcal{Z}$.



Funciones G-Equivariantes

$f : \mathcal{X} \to \mathcal{Y}$ es $G-$**equivariante** si

$$\forall g \in G : \; f \circ \rho_g = \hat{\rho}_g \circ f$$

Si $\hat{\rho} \equiv id$, f se dice G-**invariante**.

Medidas G-invariantes, $\mathcal{M}^G(\mathcal{Z})$

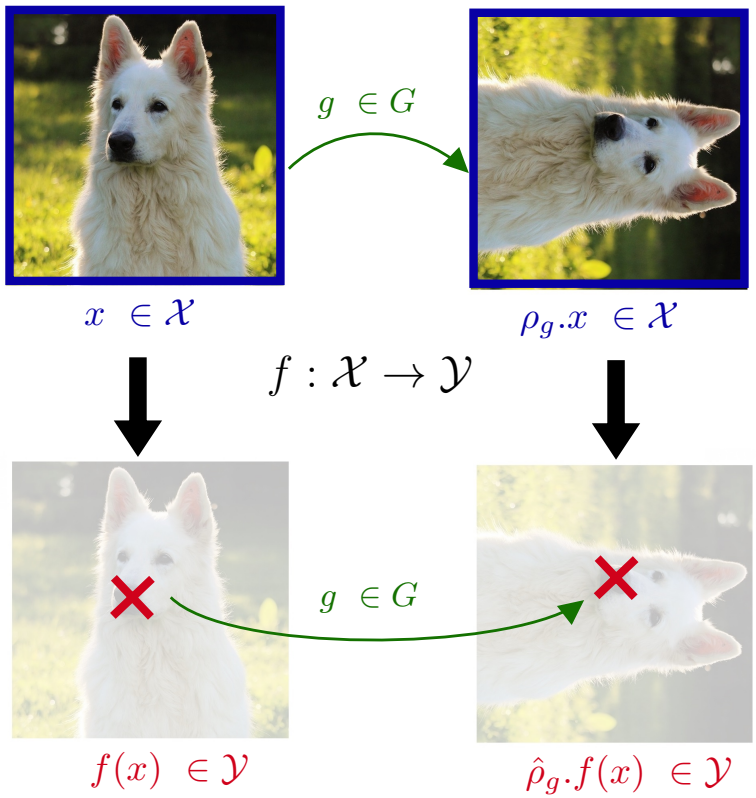
$\mu \in \mathcal{M}(\mathcal{Z})$ tq: $\forall g \in G$, $M_g \# \mu = \mu$

Datos *G-equivariantes*: $\pi \in \mathcal{P}^G(\mathcal{X} \times \mathcal{Y})$.

i.e. $\forall g \in G$, $(X, Y) \overset{(d)}{=} (\rho_g.X, \hat{\rho}_g.Y)$

Datos Equivariantes

Sea $G \curvearrowright_\rho \mathcal{X}$, $G \curvearrowright_{\hat{\rho}} \mathcal{Y}$ y $G \curvearrowright_M \mathcal{Z}$.



Funciones G-Equivariantes

$f : \mathcal{X} \to \mathcal{Y}$ es $G-$**equivariante** si

$$\forall g \in G : \; f \circ \rho_g = \hat{\rho}_g \circ f$$

Si $\hat{\rho} \equiv id$, f se dice G-**invariante**.

Medidas G-invariantes, $\mathcal{M}^G(\mathcal{Z})$

$\mu \in \mathcal{M}(\mathcal{Z})$ tq: $\forall g \in G$, $M_g \# \mu = \mu$

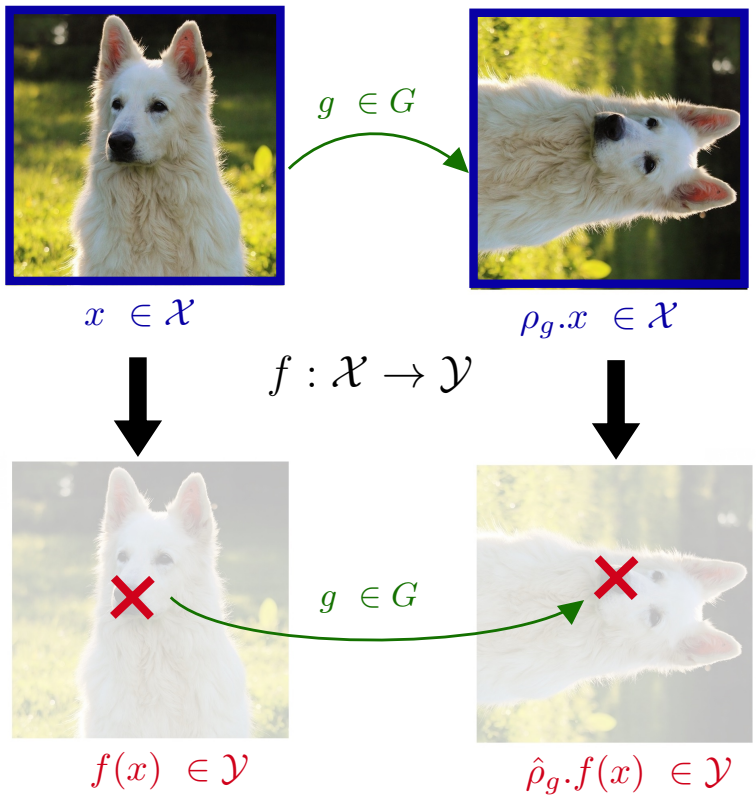
Datos *G-equivariantes*: $\pi \in \mathcal{P}^G(\mathcal{X} \times \mathcal{Y})$.

i.e. $\forall g \in G$, $(X, Y) \overset{(d)}{=} (\rho_g.X, \hat{\rho}_g.Y)$

Prop. 13: $\left[\pi \in \mathcal{P}_2^G(\mathcal{X} \times \mathcal{Y})\right] \Rightarrow [f^* := \mathbb{E}_\pi[Y|X = \cdot]$ G-equivariante $\pi_{\mathcal{X}}$-c.s.$]$

Datos Equivariantes

Sea $G \curvearrowright_\rho \mathcal{X}$, $G \curvearrowright_{\hat{\rho}} \mathcal{Y}$ y $G \curvearrowright_M \mathcal{Z}$.



Funciones G-Equivariantes

$f : \mathcal{X} \to \mathcal{Y}$ es $G-$**equivariante** si

$$\forall g \in G : \; f \circ \rho_g = \hat{\rho}_g \circ f$$

Si $\hat{\rho} \equiv id$, f se dice G-**invariante**.

Medidas G-invariantes, $\mathcal{M}^G(\mathcal{Z})$

$\mu \in \mathcal{M}(\mathcal{Z})$ tq: $\forall g \in G$, $M_g \# \mu = \mu$

Datos *G-equivariantes*: $\pi \in \mathcal{P}^G(\mathcal{X} \times \mathcal{Y})$.

i.e. $\forall g \in G$, $(X, Y) \overset{(d)}{=} (\rho_g.X, \hat{\rho}_g.Y)$

Prop. 13: $\left[\pi \in \mathcal{P}_2^G(\mathcal{X} \times \mathcal{Y})\right] \Rightarrow [f^* := \mathbb{E}_\pi[Y|X = \cdot]$ G-equivariante $\pi_{\mathcal{X}}$-c.s.$]$

¿Cómo aprovechamos esta simetría para construir mejores modelos de aprendizaje?

Data Augmentation (DA)

Se optimiza una versión *simetrizada* del riesgo de población:

$$R^G(f) = \int_G \mathbb{E}_\pi[\ell(f(\rho_g.X), \hat{\rho}_g.Y)]\mathrm{d}\lambda_G(g)$$

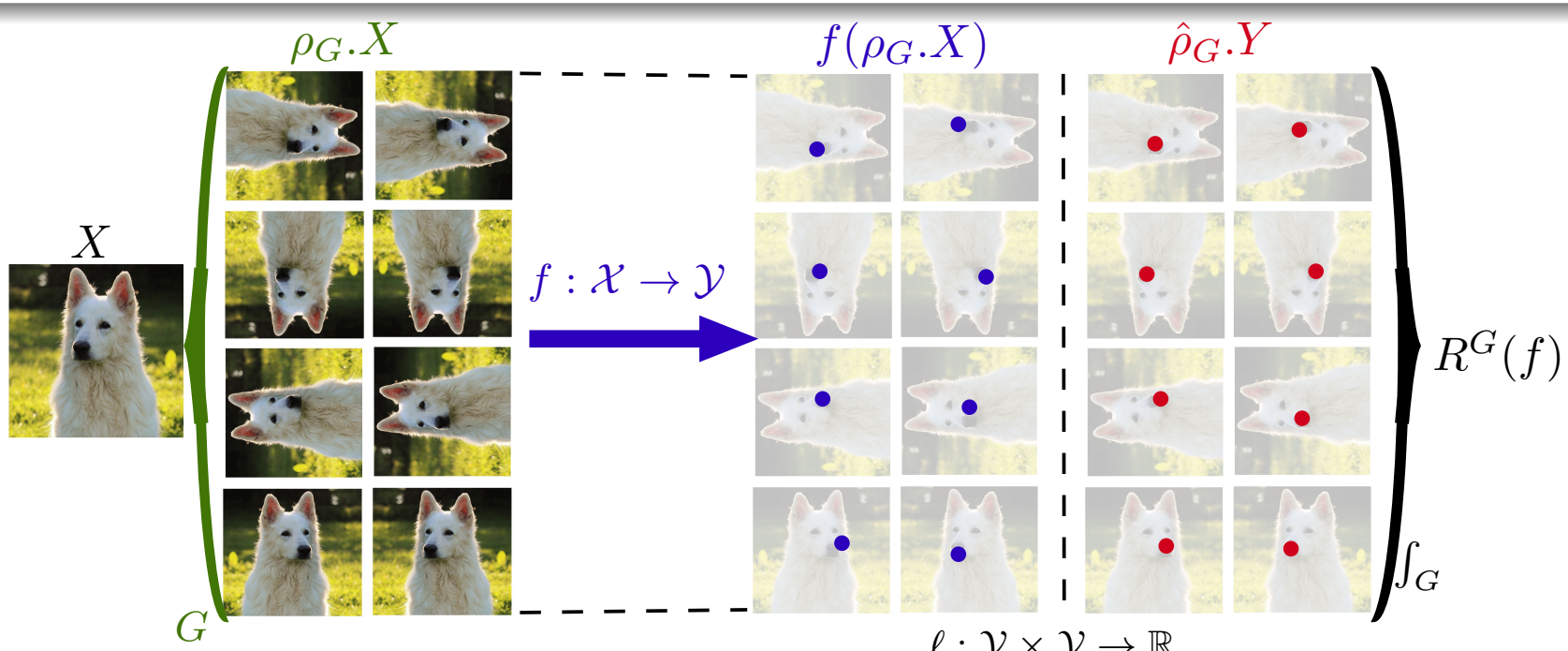
- *Promueve* un modelo final G-equivariante, pero **no** lo garantiza.
- *NO simplifica la red en cuestión* (ineficiente en parámetros).

Data Augmentation (DA)

Se optimiza una versión *simetrizada* del riesgo de población:

$$R^G(f) = \int_G \mathbb{E}_\pi[\ell(f(\rho_g.X), \hat{\rho}_g.Y)]\mathrm{d}\lambda_G(g)$$

- *Promueve* un modelo final G-equivariante, pero **no** lo garantiza.
- *NO simplifica la red en cuestión* (ineficiente en parámetros).



Operador Simetrización: Proyección ortogonal a las *funciones G-equivariantes.*

$\mathcal{Q} : L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}}) \to L^2_G(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$ dado por: $(\mathcal{Q}f)(x) = \int_G \hat{\rho}_g^{-1}.f(\rho_g.x)d\lambda_G(g)$

Operador Simetrización: Proyección ortogonal a las *funciones G-equivariantes.*

$\mathcal{Q} : L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}}) \to L^2_G(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$ dado por: $(\mathcal{Q}f)(x) = \int_G \hat{\rho}_g^{-1}.f(\rho_g.x)d\lambda_G(g)$

Feature Averaging (FA)

Se optimiza la *versión simetrizada* del modelo original: $f^{FA} = \mathcal{Q}f$.

Asegura un modelo *equivariante*, pero *no lo simplifica* y es *caro* de implementar.

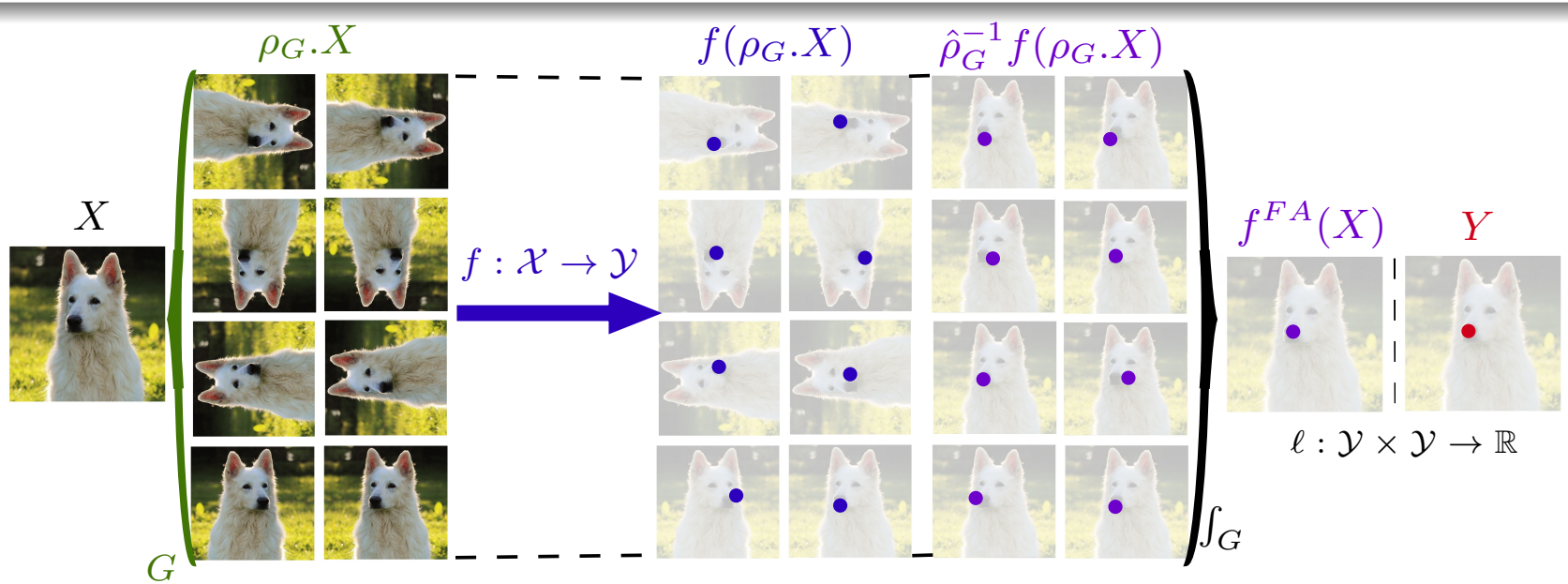
Operador Simetrización: Proyección ortogonal a las *funciones G-equivariantes.*

$\mathcal{Q} : L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}}) \to L^2_G(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$ dado por: $(\mathcal{Q}f)(x) = \int_G \hat{\rho}_g^{-1}.f(\rho_g.x)d\lambda_G(g)$

Feature Averaging (FA)

Se optimiza la *versión simetrizada* del modelo original: $f^{FA} = \mathcal{Q}f$.

Asegura un modelo *equivariante*, pero *no lo simplifica* y es *caro* de implementar.



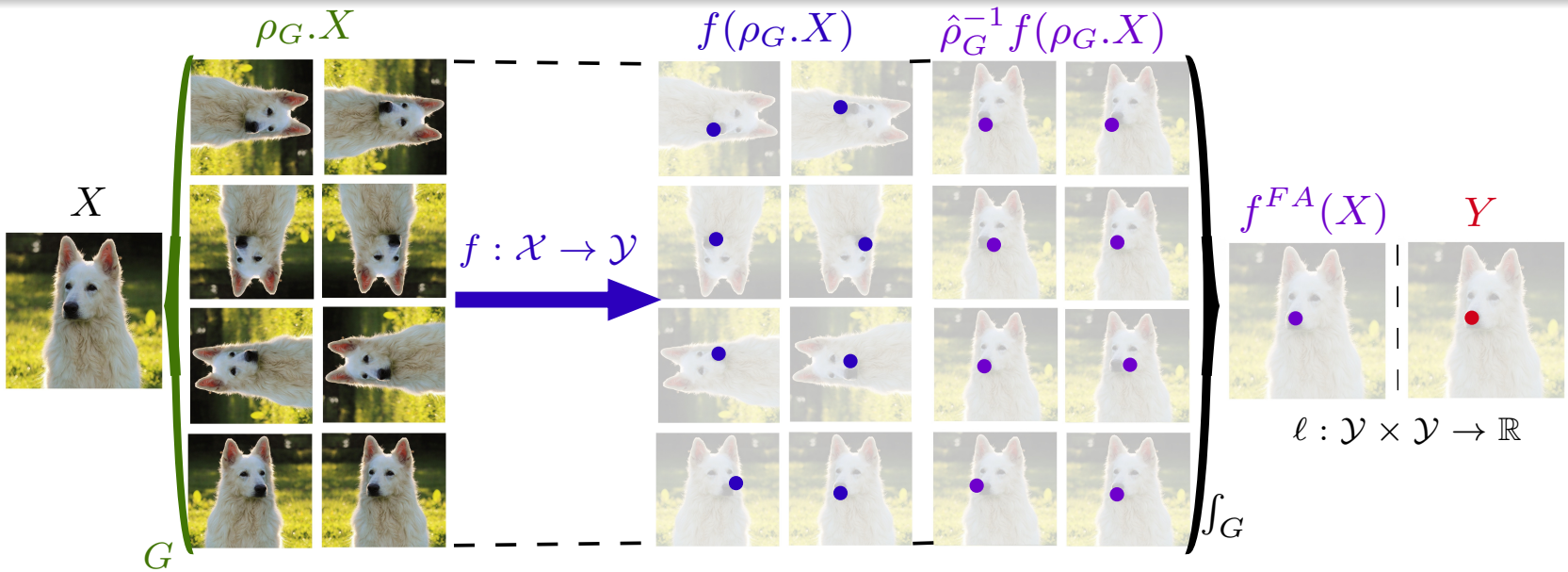
Operador Simetrización: Proyección ortogonal a las *funciones G-equivariantes.*

$\mathcal{Q} : L^2(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}}) \to L^2_G(\mathcal{X}, \mathcal{Y}; \pi_{\mathcal{X}})$ dado por: $(\mathcal{Q}f)(x) = \int_G \hat{\rho}_g^{-1}.f(\rho_g.x)d\lambda_G(g)$

Feature Averaging (FA)

Se optimiza la *versión simetrizada* del modelo original: $f^{FA} = \mathcal{Q}f$.

Asegura un modelo *equivariante*, pero *no lo simplifica* y es *caro* de implementar.



¿Cuál es preferible entre **DA** y **FA**? La literatura lo ha estudiado, con resultados mixtos.

Arquitecturas Equivariantes de NN (EA)

Se *diseña la arquitectura de la NN* para *aprovechar* las simetrías.

Si $\Phi_\theta^N = \sigma^{(L)} \circ A_L \circ \cdots \circ \sigma^{(1)} \circ A_1$, se asume que $\forall \ell \in [L]$, $G \circlearrowright_{\rho(\ell)} \mathcal{X}_\ell$, y:

- **Activaciones** $\sigma_\ell : \mathcal{X}_\ell \to \mathcal{X}_\ell$ G-equivariantes (e.g. cuando son *pointwise*).
- **Capas 'lineales'** $A_\ell : \mathcal{X}_{\ell-1} \to \mathcal{X}_\ell$ G-equivariantes. i.e. $\forall g,\ A_\ell = \rho_g^{(\ell)} . A_\ell . \rho_{g^{-1}}^{(\ell-1)}$.

Arquitecturas Equivariantes de NN (EA)

Se *diseña la arquitectura de la NN* para *aprovechar* las simetrías.

Si $\Phi_\theta^N = \sigma^{(L)} \circ A_L \circ \cdots \circ \sigma^{(1)} \circ A_1$, se asume que $\forall \ell \in [L]$, $G \circlearrowright_{\rho(\ell)} \mathcal{X}_\ell$, y:

- **Activaciones** $\sigma_\ell : \mathcal{X}_\ell \to \mathcal{X}_\ell$ G-equivariantes (e.g. cuando son *pointwise*).
- **Capas 'lineales'** $A_\ell : \mathcal{X}_{\ell-1} \to \mathcal{X}_\ell$ G-equivariantes. i.e. $\forall g,\ A_\ell = \rho_g^{(\ell)} . A_\ell . \rho_{g^{-1}}^{(\ell-1)}$.

Con $(\sigma^{(\ell)})_{\ell=1}^L$ fijo, basta conocer $\mathcal{E}^G := \Pi_{\ell=1}^L \mathrm{Hom}_G(\mathcal{X}_{\ell-1}, \mathcal{X}_\ell)$ para caracterizar las **EA**.

Arquitecturas Equivariantes de NN (EA)

Se *diseña la arquitectura de la NN* para *aprovechar* las simetrías.

Si $\Phi_\theta^N = \sigma^{(L)} \circ A_L \circ \cdots \circ \sigma^{(1)} \circ A_1$, se asume que $\forall \ell \in [L]$, $G \circlearrowright_{\rho(\ell)} \mathcal{X}_\ell$, y:

- **Activaciones** $\sigma_\ell : \mathcal{X}_\ell \to \mathcal{X}_\ell$ G-equivariantes (e.g. cuando son *pointwise*).
- **Capas 'lineales'** $A_\ell : \mathcal{X}_{\ell-1} \to \mathcal{X}_\ell$ G-equivariantes. i.e. $\forall g,\ A_\ell = \rho_g^{(\ell)} . A_\ell . \rho_{g^{-1}}^{(\ell-1)}$.

Con $(\sigma^{(\ell)})_{\ell=1}^L$ fijo, basta conocer $\mathcal{E}^G := \Pi_{\ell=1}^L \mathrm{Hom}_G(\mathcal{X}_{\ell-1}, \mathcal{X}_\ell)$ para caracterizar las **EA**.

$\mathcal{E}^G$ se caracteriza como las *convoluciones de grupo*; o como *matrices que comparten parámetros*.

(RSP'17,KT'18,CGW'18,FWGW'21,FO'23)

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

Arquitecturas Equivariantes de NN (EA)

Se *diseña la arquitectura de la NN* para *aprovechar* las simetrías.

Si $\Phi_\theta^N = \sigma^{(L)} \circ A_L \circ \cdots \circ \sigma^{(1)} \circ A_1$, se asume que $\forall \ell \in [L]$, $G \circlearrowright_{\rho(\ell)} \mathcal{X}_\ell$, y:

- **Activaciones** $\sigma_\ell : \mathcal{X}_\ell \to \mathcal{X}_\ell$ G-equivariantes (e.g. cuando son *pointwise*).
- **Capas 'lineales'** $A_\ell : \mathcal{X}_{\ell-1} \to \mathcal{X}_\ell$ G-equivariantes. i.e. $\forall g,\ A_\ell = \rho_g^{(\ell)} . A_\ell . \rho_{g^{-1}}^{(\ell-1)}$.

Con $(\sigma^{(\ell)})_{\ell=1}^L$ fijo, basta conocer $\mathcal{E}^G := \Pi_{\ell=1}^L \mathrm{Hom}_G(\mathcal{X}_{\ell-1}, \mathcal{X}_\ell)$ para caracterizar las **EA**.

$\mathcal{E}^G$ se caracteriza como las *convoluciones de grupo*; o como *matrices que comparten parámetros*.

(RSP'17,KT'18,CGW'18,FWGW'21,FO'23)

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

Ventaja: Modelo *simplificado* que es G-equivariante *por construcción*.

$\to$ Muy usado en la práctica! **CNNs**, **Transformers**, **GraphNNs**, **LieConv**, etc.

Arquitecturas Equivariantes de NN (EA)

Se *diseña la arquitectura de la NN* para *aprovechar* las simetrías.

Si $\Phi_\theta^N = \sigma^{(L)} \circ A_L \circ \cdots \circ \sigma^{(1)} \circ A_1$, se asume que $\forall \ell \in [L]$, $G \circlearrowright_{\rho(\ell)} \mathcal{X}_\ell$, y:

- **Activaciones** $\sigma_\ell : \mathcal{X}_\ell \to \mathcal{X}_\ell$ G-equivariantes (e.g. cuando son *pointwise*).
- **Capas 'lineales'** $A_\ell : \mathcal{X}_{\ell-1} \to \mathcal{X}_\ell$ G-equivariantes. i.e. $\forall g,\ A_\ell = \rho_g^{(\ell)} . A_\ell . \rho_{g^{-1}}^{(\ell-1)}$.

Con $(\sigma^{(\ell)})_{\ell=1}^L$ fijo, basta conocer $\mathcal{E}^G := \Pi_{\ell=1}^L \mathrm{Hom}_G(\mathcal{X}_{\ell-1}, \mathcal{X}_\ell)$ para caracterizar las **EA**.

$\mathcal{E}^G$ se caracteriza como las *convoluciones de grupo*; o como *matrices que comparten parámetros*.

(RSP'17,KT'18,CGW'18,FWGW'21,FO'23)

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

Desventaja: *Sobre-simplificar* puede hacernos perder la *universalidad*!

$\to$ Los modelos más usado en la práctica en general **sí** son universales.

Simetrías en modelos de *shallow NNs*

Consideremos $\mathcal{X}, \mathcal{Y}$ y $\mathcal{Z}$ espacios de Hilbert separables y G un grupo compacto tal que: $G \circlearrowright_{\rho} \mathcal{X}$, $G \circlearrowright_{M} \mathcal{Z}$ y $G \circlearrowright_{\hat{\rho}} \mathcal{Y}$. Estudiaremos modelos *shallow* dados por una función de *activación* $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$

Simetrías para *shallow NNs*

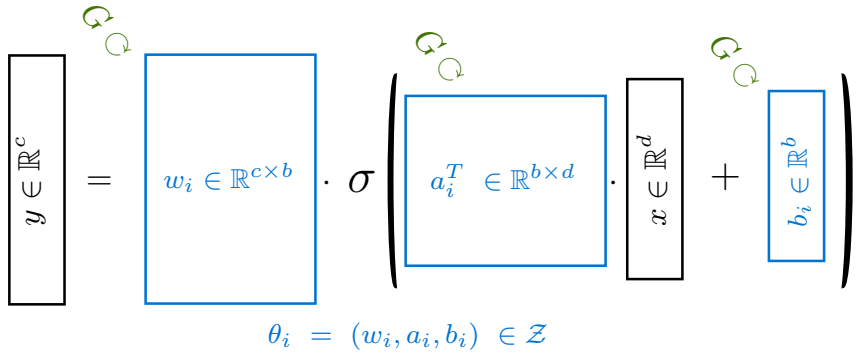
Sea $\mathcal{E}^G := \{\theta \in \mathcal{Z} \, : \, \forall g \in G, \, M_g.\theta = \theta\}$ el subespacio de *parámetros equivariantes*.

Simetrías para *shallow NNs*

Sea $\mathcal{E}^G := \{\theta \in \mathcal{Z} \,:\, \forall g \in G,\, M_g.\theta = \theta\}$ el subespacio de *parámetros equivariantes*.

En **NNs de 1 capa oculta**:

$$G \circlearrowright_M \mathcal{Z},\ \text{via: } M_g.\begin{pmatrix} w_i \\ a_i \end{pmatrix} = \begin{pmatrix} \hat{\rho}_g w_i \eta_g^T \\ \rho_g a_i \eta_g^T \end{pmatrix}$$

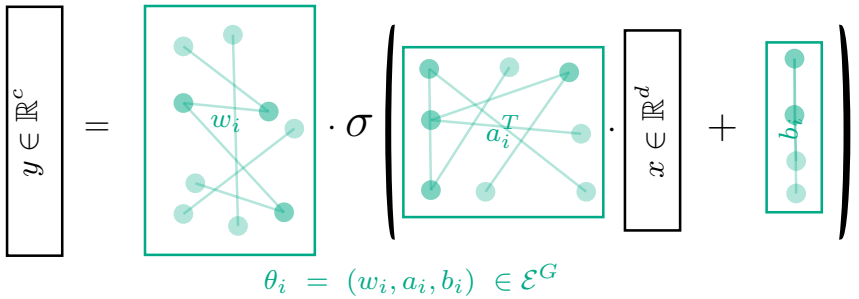


Simetrías para *shallow NNs*

Sea $\mathcal{E}^G := \{\theta \in \mathcal{Z} \,:\, \forall g \in G,\, M_g.\theta = \theta\}$ el subespacio de *parámetros equivariantes*.

En **NNs de 1 capa oculta**:

$$G \circlearrowright_M \mathcal{Z},\ \text{via: } M_g.\begin{pmatrix} w_i \\ a_i \end{pmatrix} = \begin{pmatrix} \hat{\rho}_g w_i \eta_g^T \\ \rho_g a_i \eta_g^T \end{pmatrix}$$

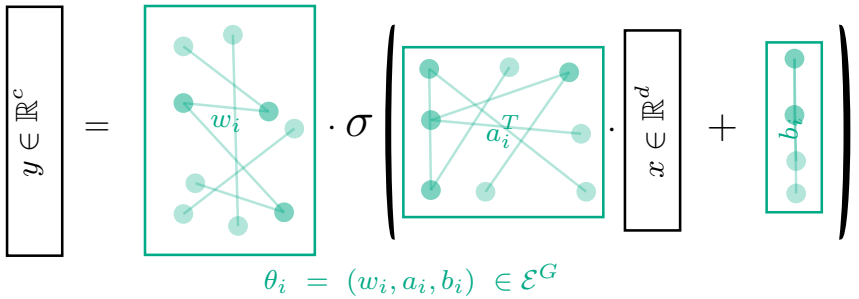


Simetrías para *shallow NNs*

Sea $\mathcal{E}^G := \{\theta \in \mathcal{Z} : \forall g \in G,\ M_g.\theta = \theta\}$ el subespacio de *parámetros equivariantes*.

En **NNs de 1 capa oculta**:

$$G \circlearrowright_M \mathcal{Z},\ \text{via: } M_g.\begin{pmatrix} w_i \\ a_i \end{pmatrix} = \begin{pmatrix} \hat{\rho}_g w_i \eta_g^T \\ \rho_g a_i \eta_g^T \end{pmatrix}$$



Modelos *shallow* G-equivariantes (definición)

Dado $\theta = (\theta_i)_{i=1}^N \in (\mathcal{Z})^N$; es un modelo shallow $\Phi_\theta^N = \frac{1}{N}\sum_{i=1}^N \sigma_*(\cdot, \theta_i)$ tal que:

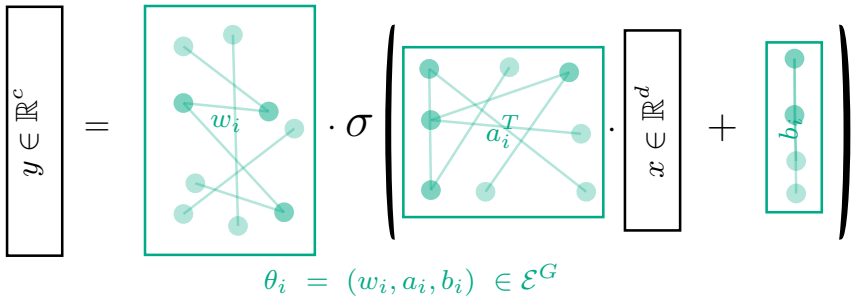
$$\forall i \in \{1, \ldots, N\},\ \theta_i \in \mathcal{E}^G \quad (\text{o, equivalentemente } \nu_\theta^N(\mathcal{E}^G) = 1)$$

Simetrías para *shallow NNs*

Sea $\mathcal{E}^G := \{\theta \in \mathcal{Z} : \forall g \in G,\ M_g.\theta = \theta\}$ el subespacio de *parámetros equivariantes*.

En **NNs de 1 capa oculta**:

$$G \circlearrowright_M \mathcal{Z},\ \text{via: } M_g.\begin{pmatrix} w_i \\ a_i \end{pmatrix} = \begin{pmatrix} \hat{\rho}_g w_i \eta_g^T \\ \rho_g a_i \eta_g^T \end{pmatrix}$$



Modelos *shallow* G-equivariantes (definición)

Dado $\theta = (\theta_i)_{i=1}^N \in (\mathcal{Z})^N$; es un modelo shallow $\Phi_\theta^N = \frac{1}{N}\sum_{i=1}^N \sigma_*(\cdot, \theta_i)$ tal que:

$$\forall i \in \{1, \ldots, N\},\ \theta_i \in \mathcal{E}^G \quad (\text{o, equivalentemente } \nu_\theta^N(\mathcal{E}^G) = 1)$$

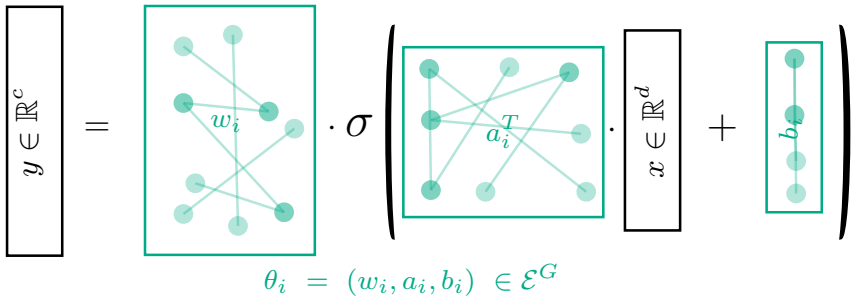
Más en general, el **modelo *shallow* G-equivariante** será: $\langle \sigma_*, \mu \rangle$ con $\mu(\mathcal{E}^G) = 1$.

Simetrías para *shallow NNs*

Sea $\mathcal{E}^G := \{\theta \in \mathcal{Z} : \forall g \in G,\ M_g.\theta = \theta\}$ el subespacio de *parámetros equivariantes*.

En **NNs de 1 capa oculta**:

$$G \circlearrowright_M \mathcal{Z},\ \text{via: } M_g.\begin{pmatrix} w_i \\ a_i \end{pmatrix} = \begin{pmatrix} \hat{\rho}_g w_i \eta_g^T \\ \rho_g a_i \eta_g^T \end{pmatrix}$$



Modelos *shallow* G-equivariantes (definición)

Dado $\theta = (\theta_i)_{i=1}^N \in (\mathcal{Z})^N$; es un modelo shallow $\Phi_\theta^N = \frac{1}{N}\sum_{i=1}^N \sigma_*(\cdot, \theta_i)$ tal que:

$$\forall i \in \{1, \ldots, N\},\ \theta_i \in \mathcal{E}^G \quad (\text{o, equivalentemente } \nu_\theta^N(\mathcal{E}^G) = 1)$$

Más en general, el **modelo *shallow* G-equivariante** será: $\langle \sigma_*, \mu \rangle$ con $\mu(\mathcal{E}^G) = 1$.

(Prop. 21) Si $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ es (*conjuntamente*) G-equivariante:

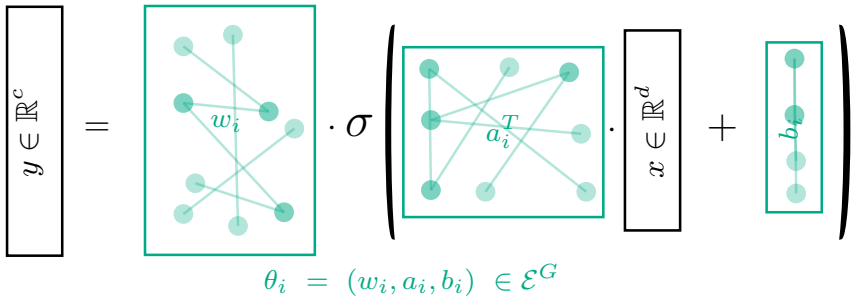
$$\text{Si } \theta \in (\mathcal{E}^G)^N,\ \text{entonces } \Phi_\theta^N : \mathcal{X} \to \mathcal{Y} \text{ es } G\text{-equivariante.}$$

Simetrías para *shallow NNs*

Sea $\mathcal{E}^G := \{\theta \in \mathcal{Z} \,:\, \forall g \in G,\ M_g.\theta = \theta\}$ el subespacio de *parámetros equivariantes*.

En **NNs de 1 capa oculta**:

$$G \circlearrowright_M \mathcal{Z},\ \text{via: } M_g.\begin{pmatrix} w_i \\ a_i \end{pmatrix} = \begin{pmatrix} \hat{\rho}_g w_i \eta_g^T \\ \rho_g a_i \eta_g^T \end{pmatrix}$$



Modelos *shallow* G-equivariantes (definición)

Dado $\theta = (\theta_i)_{i=1}^N \in (\mathcal{Z})^N$; es un modelo shallow $\Phi_\theta^N = \frac{1}{N}\sum_{i=1}^N \sigma_*(\cdot, \theta_i)$ tal que:

$$\forall i \in \{1, \ldots, N\},\ \theta_i \in \mathcal{E}^G \ \ \text{(o, equivalentemente } \nu_\theta^N(\mathcal{E}^G) = 1)$$

Más en general, el **modelo *shallow* G-equivariante** será: $\langle \sigma_*, \mu \rangle$ con $\mu(\mathcal{E}^G) = 1$.

(Prop. 21) Si $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ es (*conjuntamente*) G-equivariante:

$$\text{Si } \theta \in (\mathcal{E}^G)^N, \text{ entonces } \Phi_\theta^N : \mathcal{X} \to \mathcal{Y} \text{ es } G\text{-equivariante.}$$

La definición es *consistente* y *razonable*! $\to$ Supondremos σ_* (conj.) G-equivariante.

Medidas concentradas en $\mathcal{E}^G$: $\mathcal{P}(\mathcal{E}^G) := \{\nu \in \mathcal{P}(\mathcal{Z}) \,:\, \nu(\mathcal{E}^G) = 1\}$

Medidas concentradas en $\mathcal{E}^G$: $\mathcal{P}(\mathcal{E}^G) := \{\nu \in \mathcal{P}(\mathcal{Z}) \,:\, \nu(\mathcal{E}^G) = 1\}$

¿Es posible minimizar $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ G-invariante con medidas en $\mathcal{P}(\mathcal{E}^G)$?

(e.g. $R(\mu) = \mathbb{E}_\pi\left[\ell(\langle \sigma_*(X;\cdot), \mu\rangle, Y)\right]$ es G-invariante cuando ℓ, π son G-inv. y σ_* G-equiv.).

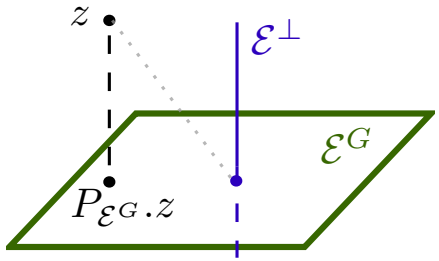
Medidas concentradas en $\mathcal{E}^G$: $\mathcal{P}(\mathcal{E}^G) := \{\nu \in \mathcal{P}(\mathcal{Z}) \,:\, \nu(\mathcal{E}^G) = 1\}$

¿Es posible minimizar $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ G-invariante con medidas en $\mathcal{P}(\mathcal{E}^G)$?

(e.g. $R(\mu) = \mathbb{E}_\pi\left[\ell(\langle \sigma_*(X;\cdot), \mu\rangle, Y)\right]$ es G-invariante cuando ℓ, π son G-inv. y σ_* G-equiv.).

Proyección Ortogonal a $\mathcal{E}^G$ (s.e.v. de $\mathcal{Z}$)

$$P_{\mathcal{E}^G} : \mathcal{Z} \to \mathcal{E}^G,\ P_{\mathcal{E}^G}(z) = \int_G M_g.z \, d\lambda_G(g)$$



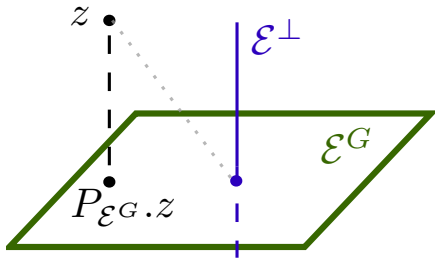
Medidas concentradas en $\mathcal{E}^G$: $\mathcal{P}(\mathcal{E}^G) := \{\nu \in \mathcal{P}(\mathcal{Z}) \,:\, \nu(\mathcal{E}^G) = 1\}$

¿Es posible minimizar $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ G-invariante con medidas en $\mathcal{P}(\mathcal{E}^G)$?

(e.g. $R(\mu) = \mathbb{E}_\pi\left[\ell(\langle \sigma_*(X;\cdot), \mu\rangle, Y)\right]$ es G-invariante cuando ℓ, π son G-inv. y σ_* G-equiv.).

Proyección Ortogonal a $\mathcal{E}^G$ (s.e.v. de $\mathcal{Z}$)

$$P_{\mathcal{E}^G} : \mathcal{Z} \to \mathcal{E}^G,\ P_{\mathcal{E}^G}(z) = \int_G M_g.z \, d\lambda_G(g)$$



(Prop. 25) $\mu \mapsto \mu^{\mathcal{E}^G} := P_{\mathcal{E}^G}\#\mu$ es una **proyección** sobre $(\mathcal{P}_p(\mathcal{E}^G), W_p)$.

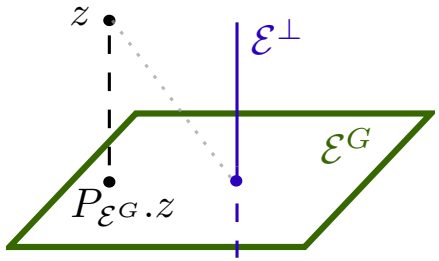
Medidas concentradas en $\mathcal{E}^G$: $\mathcal{P}(\mathcal{E}^G) := \{\nu \in \mathcal{P}(\mathcal{Z}) \,:\, \nu(\mathcal{E}^G) = 1\}$

¿Es posible minimizar $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ G-invariante con medidas en $\mathcal{P}(\mathcal{E}^G)$?

(e.g. $R(\mu) = \mathbb{E}_\pi\left[\ell(\langle \sigma_*(X;\cdot), \mu\rangle, Y)\right]$ es G-invariante cuando ℓ, π son G-inv. y σ_* G-equiv.).

Proyección Ortogonal a $\mathcal{E}^G$ (s.e.v. de $\mathcal{Z}$)

$$P_{\mathcal{E}^G} : \mathcal{Z} \to \mathcal{E}^G,\ P_{\mathcal{E}^G}(z) = \int_G M_g.z \, d\lambda_G(g)$$



(Prop. 25) $\mu \mapsto \mu^{\mathcal{E}^G} := P_{\mathcal{E}^G}\#\mu$ es una **proyección** sobre $(\mathcal{P}_p(\mathcal{E}^G), W_p)$.

Pero... Nada garantiza que $R(\mu^{\mathcal{E}^G}) \leq R(\mu)$. ¿Bajo qué condiciones sí se tiene esto?

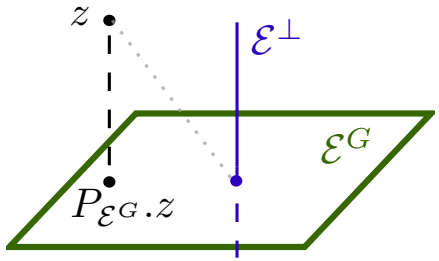
Medidas concentradas en $\mathcal{E}^G$: $\mathcal{P}(\mathcal{E}^G) := \{\nu \in \mathcal{P}(\mathcal{Z}) \,:\, \nu(\mathcal{E}^G) = 1\}$

¿Es posible minimizar $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ G-invariante con medidas en $\mathcal{P}(\mathcal{E}^G)$?

(e.g. $R(\mu) = \mathbb{E}_\pi\left[\ell(\langle \sigma_*(X;\cdot), \mu\rangle, Y)\right]$ es G-invariante cuando ℓ, π son G-inv. y σ_* G-equiv.).

Proyección Ortogonal a $\mathcal{E}^G$ (s.e.v. de $\mathcal{Z}$)

$$P_{\mathcal{E}^G} : \mathcal{Z} \to \mathcal{E}^G,\ P_{\mathcal{E}^G}(z) = \int_G M_g.z \, d\lambda_G(g)$$



(Prop. 25) $\mu \mapsto \mu^{\mathcal{E}^G} := P_{\mathcal{E}^G}\#\mu$ es una **proyección** sobre $(\mathcal{P}_p(\mathcal{E}^G), W_p)$.

Pero... Nada garantiza que $R(\mu^{\mathcal{E}^G}) \leq R(\mu)$. ¿Bajo qué condiciones sí se tiene esto?

Contraejemplo (Prop. 35). Incluso con G finito y $\mathrm{supp}(\pi)$ compacto...

En general: $\min_{\mu \in \mathcal{P}(\mathcal{Z})} R(\mu) < \min_{\nu \in \mathcal{P}(\mathcal{E}^G)} R(\nu)$

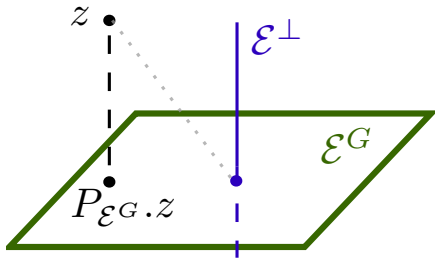
Medidas concentradas en $\mathcal{E}^G$: $\mathcal{P}(\mathcal{E}^G) := \{\nu \in \mathcal{P}(\mathcal{Z}) \,:\, \nu(\mathcal{E}^G) = 1\}$

¿Es posible minimizar $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ G-invariante con medidas en $\mathcal{P}(\mathcal{E}^G)$?

(e.g. $R(\mu) = \mathbb{E}_\pi\left[\ell(\langle \sigma_*(X;\cdot), \mu\rangle, Y)\right]$ es G-invariante cuando ℓ, π son G-inv. y σ_* G-equiv.).

Proyección Ortogonal a $\mathcal{E}^G$ (s.e.v. de $\mathcal{Z}$)

$$P_{\mathcal{E}^G} : \mathcal{Z} \to \mathcal{E}^G,\ P_{\mathcal{E}^G}(z) = \int_G M_g.z \, d\lambda_G(g)$$



(Prop. 25) $\mu \mapsto \mu^{\mathcal{E}^G} := P_{\mathcal{E}^G}\#\mu$ es una **proyección** sobre $(\mathcal{P}_p(\mathcal{E}^G), W_p)$.

Pero... Nada garantiza que $R(\mu^{\mathcal{E}^G}) \leq R(\mu)$. ¿Bajo qué condiciones sí se tiene esto?

Proposición 36 (Univ. Equivariante). Sean $\pi \in \mathcal{P}_2^G(\mathcal{X} \times \mathcal{Y})$, y ℓ cuadrática.

$$\left[\mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{E}^G)) \text{ universal en } L^2_G(\mathcal{X}, \mathcal{Y}; \pi|_{\mathcal{X}})\right] \Rightarrow \left[\inf_{\nu \in \mathcal{P}(\mathcal{E}^G)} R(\nu) = R_*\right]$$

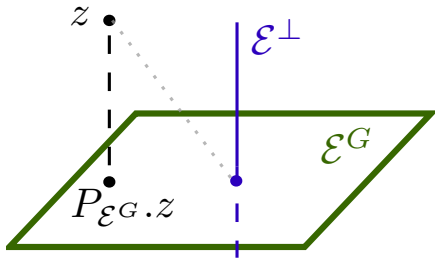
Medidas concentradas en $\mathcal{E}^G$: $\mathcal{P}(\mathcal{E}^G) := \{\nu \in \mathcal{P}(\mathcal{Z}) \,:\, \nu(\mathcal{E}^G) = 1\}$

¿Es posible minimizar $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ G-invariante con medidas en $\mathcal{P}(\mathcal{E}^G)$?

(e.g. $R(\mu) = \mathbb{E}_\pi\left[\ell(\langle \sigma_*(X;\cdot), \mu\rangle, Y)\right]$ es G-invariante cuando ℓ, π son G-inv. y σ_* G-equiv.).

Proyección Ortogonal a $\mathcal{E}^G$ (s.e.v. de $\mathcal{Z}$)

$$P_{\mathcal{E}^G} : \mathcal{Z} \to \mathcal{E}^G,\ P_{\mathcal{E}^G}(z) = \int_G M_g.z \, d\lambda_G(g)$$



(Prop. 25) $\mu \mapsto \mu^{\mathcal{E}^G} := P_{\mathcal{E}^G}\#\mu$ es una **proyección** sobre $(\mathcal{P}_p(\mathcal{E}^G), W_p)$.

Pero... Nada garantiza que $R(\mu^{\mathcal{E}^G}) \leq R(\mu)$. ¿Bajo qué condiciones sí se tiene esto?

Proposición 36 (Univ. Equivariante). Sean $\pi \in \mathcal{P}_2^G(\mathcal{X} \times \mathcal{Y})$, y ℓ cuadrática.

$$\left[\mathcal{F}_{\sigma_*}(\mathcal{P}(\mathcal{E}^G)) \text{ universal en } L^2_G(\mathcal{X}, \mathcal{Y}; \pi|_{\mathcal{X}})\right] \Rightarrow \left[\inf_{\nu \in \mathcal{P}(\mathcal{E}^G)} R(\nu) = R_*\right]$$

¿Dónde es *natural* esperar soluciones sabiendo que π es G-invariante?

Medidas G-invariantes:

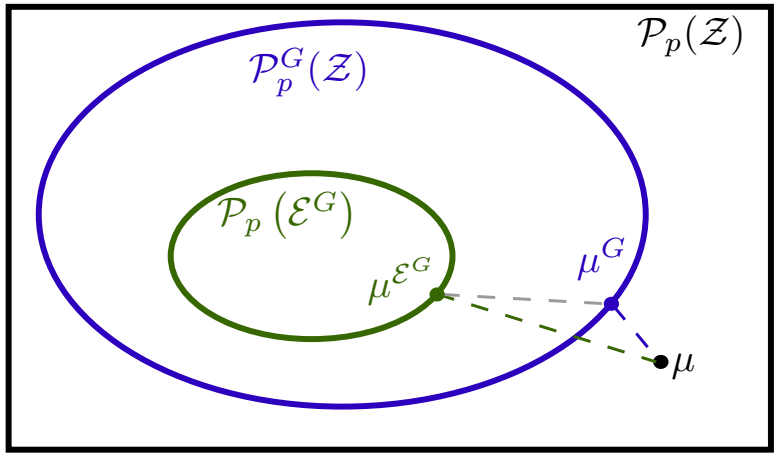
$$\mathcal{P}^G(\mathcal{Z}) := \{\mu \in \mathcal{P}(\mathcal{Z}) \,:\, \forall g \in G,\, M_g \# \mu = \mu\}$$

Simetrización: $\mu \mapsto \mu^G := \int_G (M_g \# \mu) d\lambda_G(g)$

Medidas G-invariantes:

$$\mathcal{P}^G(\mathcal{Z}) := \{\mu \in \mathcal{P}(\mathcal{Z}) \,:\, \forall g \in G,\, M_g \# \mu = \mu\}$$

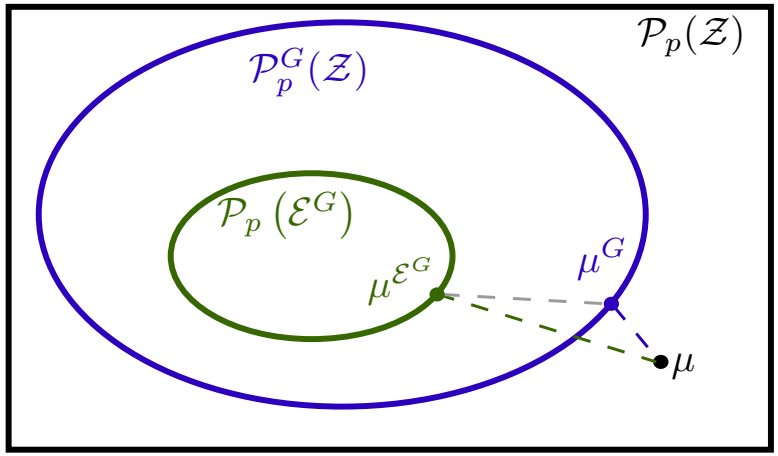
Simetrización: $\mu \mapsto \mu^G := \int_G (M_g \# \mu) d\lambda_G(g)$



Medidas G-invariantes:

$$\mathcal{P}^G(\mathcal{Z}) := \{\mu \in \mathcal{P}(\mathcal{Z}) \,:\, \forall g \in G,\, M_g \# \mu = \mu\}$$

Simetrización: $\mu \mapsto \mu^G := \int_G (M_g \# \mu) d\lambda_G(g)$

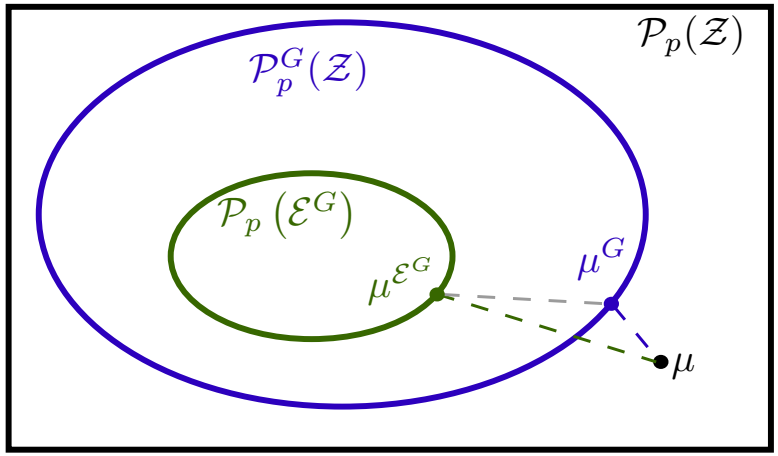


(Lema 13 y Prop. 26) estudian las propiedades de $\mathcal{P}^G(\mathcal{Z})$ y $\mathcal{P}(\mathcal{E}^G)$ (y su relación).

Medidas G-invariantes:

$$\mathcal{P}^G(\mathcal{Z}) := \{\mu \in \mathcal{P}(\mathcal{Z}) \,:\, \forall g \in G,\, M_g \# \mu = \mu\}$$

Simetrización: $\mu \mapsto \mu^G := \int_G (M_g \# \mu) d\lambda_G(g)$



(Lema 13 y Prop. 26) estudian las propiedades de $\mathcal{P}^G(\mathcal{Z})$ y $\mathcal{P}(\mathcal{E}^G)$ (y su relación).

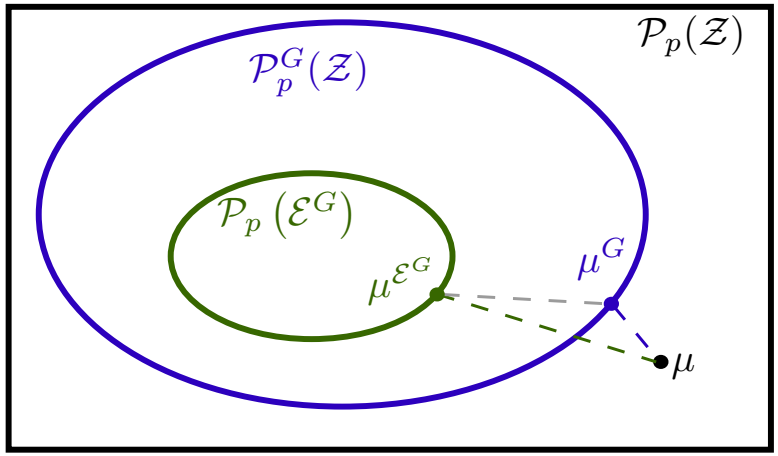
Proposición 32 (Desigualdad de Jensen). Sea $\{\mu_s\}_{s\in S} \subseteq \mathcal{P}(\mathcal{Z})$ y $\lambda \in \mathcal{P}(S)$.

$$\Big[\, R \textbf{ convexa y } \mathcal{C}^1 \Big] \Rightarrow \left[R\left(\int_S \mu_s d\lambda(s) \right) \leq \int_S R(\mu_s) d\lambda(s) \right]$$

Medidas G-invariantes:

$$\mathcal{P}^G(\mathcal{Z}) := \{\mu \in \mathcal{P}(\mathcal{Z}) : \forall g \in G,\ M_g \# \mu = \mu\}$$

Simetrización: $\mu \mapsto \mu^G := \int_G (M_g \# \mu) d\lambda_G(g)$



(Lema 13 y Prop. 26) estudian las propiedades de $\mathcal{P}^G(\mathcal{Z})$ y $\mathcal{P}(\mathcal{E}^G)$ (y su relación).

Proposición 32 (Desigualdad de Jensen). Sea $\{\mu_s\}_{s\in S} \subseteq \mathcal{P}(\mathcal{Z})$ y $\lambda \in \mathcal{P}(S)$.

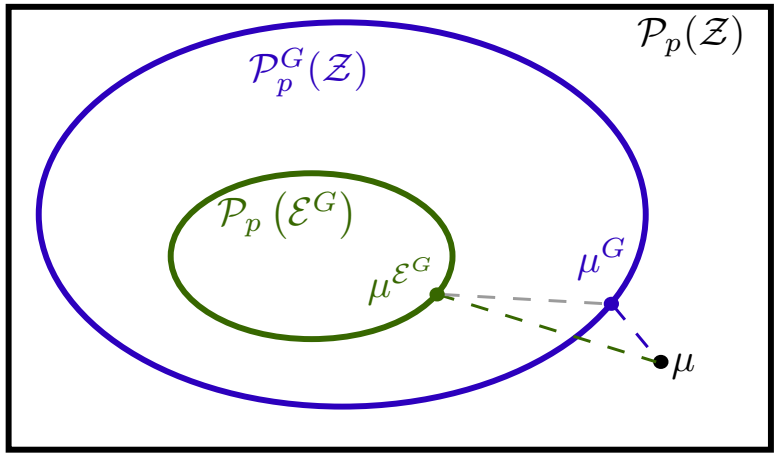
$$\Big[\ R \textbf{ convexa y } \mathcal{C}^1\Big] \Rightarrow \left[R\left(\int_S \mu_s d\lambda(s)\right) \leq \int_S R(\mu_s) d\lambda(s)\right]$$

En particular, si R es G-invariante, $R(\mu^G) \leq R(\mu)$.

Medidas G-invariantes:

$$\mathcal{P}^G(\mathcal{Z}) := \{\mu \in \mathcal{P}(\mathcal{Z}) : \forall g \in G,\ M_g \# \mu = \mu\}$$

Simetrización: $\mu \mapsto \mu^G := \int_G (M_g \# \mu) d\lambda_G(g)$



(Lema 13 y Prop. 26) estudian las propiedades de $\mathcal{P}^G(\mathcal{Z})$ y $\mathcal{P}(\mathcal{E}^G)$ (y su relación).

Proposición 32 (Desigualdad de Jensen). Sea $\{\mu_s\}_{s\in S} \subseteq \mathcal{P}(\mathcal{Z})$ y $\lambda \in \mathcal{P}(S)$.

$$\Big[\ R \textbf{ convexa y } \mathcal{C}^1\Big] \Rightarrow \left[R\left(\int_S \mu_s d\lambda(s)\right) \leq \int_S R(\mu_s) d\lambda(s)\right]$$

En particular, si R es G-invariante, $R(\mu^G) \leq R(\mu)$.

(Cor. 8) $\Big[\ R$ **convexa**, $\mathcal{C}^1$ y G-invariante$\Big] \Rightarrow \left[\inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) = \inf_{\mu\in\mathcal{P}(\mathcal{Z})} R(\mu)\right]$

Estudio de Medidas Simétricas

Sea que la proyección $p : \mathcal{Z} \to G\backslash\mathcal{Z}$ **admite una sección** $s : G\backslash\mathcal{Z} \to \mathcal{Z}$ **medible**.

Sea que la proyección $p : \mathcal{Z} \to G\backslash\mathcal{Z}$ **admite una sección** $s : G\backslash\mathcal{Z} \to \mathcal{Z}$ **medible**.

Proposición 28 (Teo. de Descomposición Ergódica; $\mathcal{P}^G(\mathcal{Z}) \cong \mathcal{P}(G\backslash\mathcal{Z})$)

$\Psi : \mathcal{P}^G(\mathcal{Z}) \to \mathcal{P}(G\backslash\mathcal{Z})$ dado por $\Psi(\mu) = p\#\mu$ es **biyección bimedible**

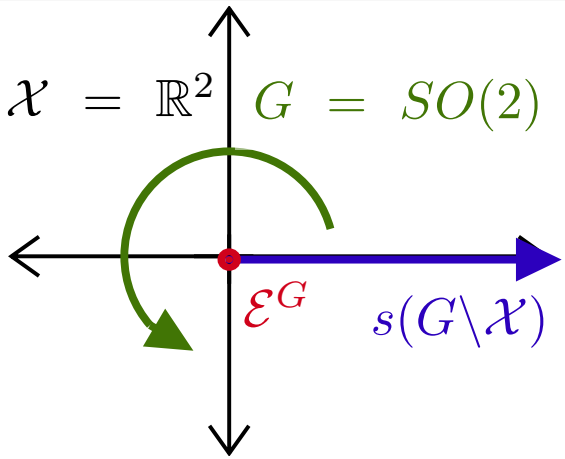
Su inversa está dada por $\Psi^{-1}(\overline{\mu}) = (s\#\overline{\mu})^G$ (Teo. Desc. Ergódica).

Sea que la proyección $p : \mathcal{Z} \to G\backslash\mathcal{Z}$ **admite una sección** $s : G\backslash\mathcal{Z} \to \mathcal{Z}$ **medible**.

Proposición 28 (Teo. de Descomposición Ergódica; $\mathcal{P}^G(\mathcal{Z}) \cong \mathcal{P}(G\backslash\mathcal{Z})$)

$\Psi : \mathcal{P}^G(\mathcal{Z}) \to \mathcal{P}(G\backslash\mathcal{Z})$ dado por $\Psi(\mu) = p\#\mu$ es **biyección bimedible**

Su inversa está dada por $\Psi^{-1}(\overline{\mu}) = (s\#\overline{\mu})^G$ (Teo. Desc. Ergódica).

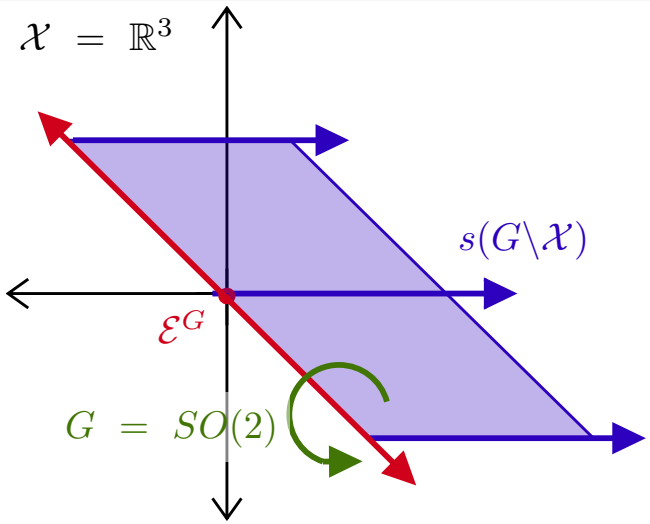


Sea que la proyección $p : \mathcal{Z} \to G\backslash\mathcal{Z}$ **admite una sección** $s : G\backslash\mathcal{Z} \to \mathcal{Z}$ **medible**.

Proposición 28 (Teo. de Descomposición Ergódica; $\mathcal{P}^G(\mathcal{Z}) \cong \mathcal{P}(G\backslash\mathcal{Z})$)

$\Psi : \mathcal{P}^G(\mathcal{Z}) \to \mathcal{P}(G\backslash\mathcal{Z})$ dado por $\Psi(\mu) = p\#\mu$ es **biyección bimedible**

Su inversa está dada por $\Psi^{-1}(\overline{\mu}) = (s\#\overline{\mu})^G$ (Teo. Desc. Ergódica).

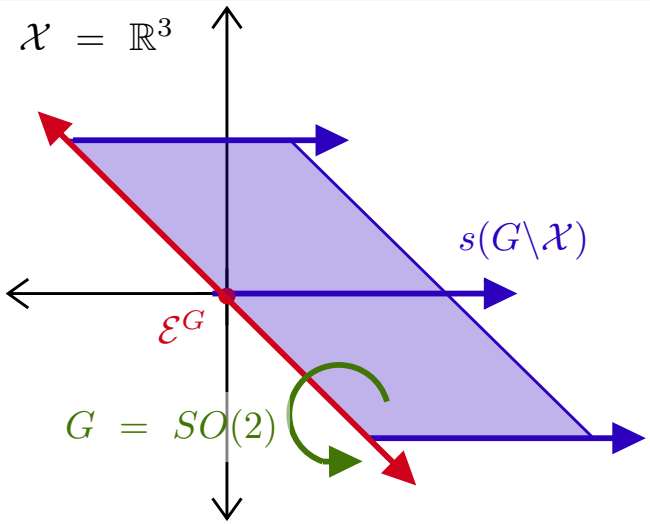


Sea que la proyección $p : \mathcal{Z} \to G\backslash\mathcal{Z}$ **admite una sección** $s : G\backslash\mathcal{Z} \to \mathcal{Z}$ **medible**.

Proposición 28 (Teo. de Descomposición Ergódica; $\mathcal{P}^G(\mathcal{Z}) \cong \mathcal{P}(G\backslash\mathcal{Z})$)

$\Psi : \mathcal{P}^G(\mathcal{Z}) \to \mathcal{P}(G\backslash\mathcal{Z})$ dado por $\Psi(\mu) = p\#\mu$ es **biyección bimedible**

Su inversa está dada por $\Psi^{-1}(\overline{\mu}) = (s\#\overline{\mu})^G$ (Teo. Desc. Ergódica).



Proposición 34 (Reducción a $G\backslash\mathcal{Z}$)

$$\inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) = \inf_{\overline{\mu}\in\mathcal{P}(G\backslash\mathcal{Z})} R((s\#\overline{\mu})^G)$$

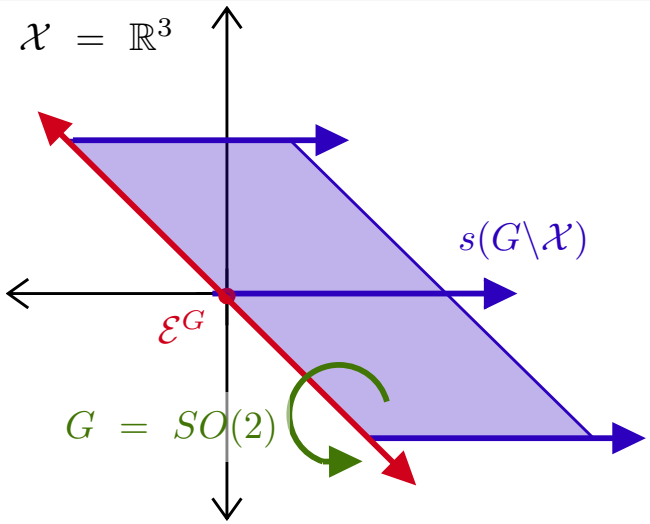
Basta buscar una medida sobre $G\backslash\mathcal{Z}$

Sea que la proyección $p : \mathcal{Z} \to G\backslash\mathcal{Z}$ **admite una sección** $s : G\backslash\mathcal{Z} \to \mathcal{Z}$ **medible**.

Proposición 28 (Teo. de Descomposición Ergódica; $\mathcal{P}^G(\mathcal{Z}) \cong \mathcal{P}(G\backslash\mathcal{Z})$)

$\Psi : \mathcal{P}^G(\mathcal{Z}) \to \mathcal{P}(G\backslash\mathcal{Z})$ dado por $\Psi(\mu) = p\#\mu$ es **biyección bimedible**

Su inversa está dada por $\Psi^{-1}(\overline{\mu}) = (s\#\overline{\mu})^G$ (Teo. Desc. Ergódica).



Proposición 34 (Reducción a $G\backslash\mathcal{Z}$)

$$\inf_{\mu\in\mathcal{P}^G(\mathcal{Z})} R(\mu) = \inf_{\overline{\mu}\in\mathcal{P}(G\backslash\mathcal{Z})} R((s\#\overline{\mu})^G)$$

Basta buscar una medida sobre $G\backslash\mathcal{Z}$

El problema se reduce a un espacio más pequeño! (pero NO lo suficiente)

(Prop. 38) $\left[\forall \nu \in \mathcal{P}^{G}(\mathcal{Z}), \langle \sigma_*, \nu \rangle \text{ es } G\text{-equiv.}\right]$ y $\left[\forall \mu,\ \mathcal{Q}(\langle \sigma_*, \mu \rangle) = \langle \sigma_*, \mu^{G} \rangle\right]$

(Prop. 38) $\left[\forall \nu \in \mathcal{P}^G(\mathcal{Z}), \langle \sigma_*, \nu \rangle \text{ es } G\text{-equiv.}\right]$ y $\left[\forall \mu,\ \mathcal{Q}(\langle \sigma_*, \mu \rangle) = \langle \sigma_*, \mu^G \rangle\right]$

Al *simetrizar modelos shallow* aparece *naturalmente* $\mathcal{P}^G(\mathcal{Z})$ (y *no* $\mathcal{P}(\mathcal{E}^G)$).

(Prop. 38) $\left[\forall \nu \in \mathcal{P}^G(\mathcal{Z}), \langle \sigma_*, \nu \rangle \text{ es } G\text{-equiv.}\right]$ y $\left[\forall \mu,\ \mathcal{Q}(\langle \sigma_*, \mu \rangle) = \langle \sigma_*, \mu^G \rangle\right]$

Al *simetrizar modelos shallow* aparece *naturalmente* $\mathcal{P}^G(\mathcal{Z})$ (y *no* $\mathcal{P}(\mathcal{E}^G)$).

Proposición 40 (DA, FA y EA en el contexto MF). Sea $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$.

$$R^G(\mu) = \int_G R(M_g \# \mu) d\lambda_G(g),\ \ R^{FA}(\mu) = R(\mu^G) \text{ y } R^{EA}(\mu) = R(\mu^{\mathcal{E}^G})$$

Son todos G-invariantes.

(Prop. 38) $\left[\forall \nu \in \mathcal{P}^G(\mathcal{Z}), \langle \sigma_*, \nu \rangle \text{ es } G\text{-equiv.}\right]$ y $\left[\forall \mu,\ \mathcal{Q}(\langle \sigma_*, \mu \rangle) = \langle \sigma_*, \mu^G \rangle\right]$

Al *simetrizar modelos shallow* aparece *naturalmente* $\mathcal{P}^G(\mathcal{Z})$ (y *no* $\mathcal{P}(\mathcal{E}^G)$).

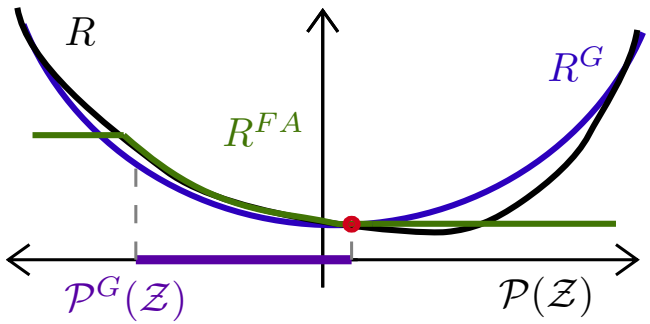
Proposición 40 (**DA**, **FA** y **EA** en el contexto MF). Sea $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$.

$$R^G(\mu) = \int_G R(M_g \# \mu) d\lambda_G(g),\ R^{FA}(\mu) = R(\mu^G) \text{ y } R^{EA}(\mu) = R(\mu^{\mathcal{E}^G})$$

Son todos G-invariantes.

Proposición 42 (Minimización de **DA** y **FA**)

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



(Prop. 38) $\left[\forall \nu \in \mathcal{P}^G(\mathcal{Z}), \langle \sigma_*, \nu \rangle \text{ es } G\text{-equiv.}\right]$ y $\left[\forall \mu,\ \mathcal{Q}(\langle \sigma_*, \mu \rangle) = \langle \sigma_*, \mu^G \rangle\right]$

Al *simetrizar modelos shallow* aparece *naturalmente* $\mathcal{P}^G(\mathcal{Z})$ (y *no* $\mathcal{P}(\mathcal{E}^G)$).

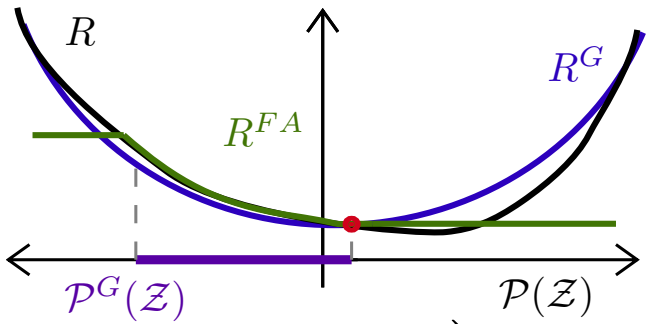
Proposición 40 (**DA**, **FA** y **EA** en el contexto MF). Sea $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$.

$$R^G(\mu) = \int_G R(M_g \# \mu) d\lambda_G(g),\ R^{FA}(\mu) = R(\mu^G) \text{ y } R^{EA}(\mu) = R(\mu^{\mathcal{E}^G})$$

Son todos G-invariantes.

Proposición 42 (Minimización de **DA** y **FA**)

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



Por su parte, $\inf_{\mu \in \mathcal{P}(\mathcal{Z})} R^{EA}(\mu) = \inf_{\mu \in \mathcal{P}(\mathcal{E}^G)} R(\mu) \left(> \inf_{\mu \in \mathcal{P}^G(\mathcal{Z})} R(\mu) \text{ en general...} \right)$

Resultados Extra (Opcional)

Los siguientes resultados son **extensiones** de resultados conocidos en la literatura. Hablan de los límites de *asumir que* π *es G-invariante*.

¿Qué pasa si asumo simetría con respecto al grupo incorrecto?

Para $\pi \in \mathcal{P}_2(\mathcal{X} \times \mathcal{Y})$ y *loss cuadrática*, se define el **gap de simetrización**:

$$\Delta(f, \mathcal{Q}_G f) := R(f) - R(\mathcal{Q}f) = \mathbb{E}_\pi[\|Y - f(X)\|_{\mathcal{Y}}^2] - \mathbb{E}_\pi[\|Y - (\mathcal{Q}_G f)(X)\|_{\mathcal{Y}}^2]$$

Generalización de resultados de **(EZ'21,HLV'23)**

Lema 10: Si $\pi_{\mathcal{X}} \in \mathcal{P}^G(\mathcal{X})$, pero π sólo es H-invariante para $H \leq G$:

$$\Delta(f, \mathcal{Q}_G f) = -2\langle f^*, f_G^\perp \rangle_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})} + \|f_G^\perp\|_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})}^2$$

Si además $\pi \in \mathcal{P}^G(\mathcal{X} \times \mathcal{Y})$, entonces $\Delta(f, \mathcal{Q}_G f) = \|f_G^\perp\|_{L^2(\mathcal{X},\mathcal{Y};\pi_{\mathcal{X}})}^2 \geq 0$

¿Qué tan cerca está R de ser *simétrico*? (aplicación de **(CDH'19)**)

$$\textbf{(Prop. 44)} \left[\begin{array}{c} \ell \text{ y } \sigma_* \\ \text{Lipschitz} \end{array} \right] \Rightarrow \left[\sup_{\mu \in \mathcal{P}(\mathcal{Z})} |R(\mu) - R^G(\mu)| \leq C \int_G W_1(g\#\pi, \pi) d\lambda_G(g) \right]$$

Explorar nociones de simetría *más generales* queda para trabajo futuro.

Simetrías en la Dinámica de Entrenamiento

Recordemos que tenemos $\mathcal{X}, \mathcal{Y}$ y $\mathcal{Z}$ Hilbert separables; G compacto tal que: $G \circlearrowright_\rho \mathcal{X}$, $G \circlearrowright_M \mathcal{Z}$ y $G \circlearrowright_{\hat{\rho}} \mathcal{Y}$; y la *activación* $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$

Recordemos que la **dinámica MF de entrenamiento** (DD) está dada por:

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

Dinámica de Entrenamiento

Recordemos que la **dinámica MF de entrenamiento** (DD) está dada por:

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

Teorema (WGF es G-invariante). Sea $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ convexo y $\mathcal{C}^1$.

Sea R G-invariante, con WGF bien definido y única solución débil $(\mu_t)_{t\geq 0}$.

Si la c.i. cumple $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$ entonces: c.t.p. $\forall t \geq 0$, $\mu_t \in \mathcal{P}_2^G(\mathcal{Z})$

Dinámica de Entrenamiento

Recordemos que la **dinámica MF de entrenamiento** (DD) está dada por:

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

Teorema (WGF es G-invariante). Sea $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ convexo y $\mathcal{C}^1$.

Sea R G-invariante, con WGF bien definido y única solución débil $(\mu_t)_{t\geq 0}$.

Si la c.i. cumple $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$ entonces: c.t.p. $\forall t \geq 0$, $\mu_t \in \mathcal{P}_2^G(\mathcal{Z})$

Este resultado general es aplicable en todos los casos de dinámica de entrenamiento que se encuentran en la literatura (con y sin regularización, diferentes *learning rates*, etc.).

Dinámica de Entrenamiento

Recordemos que la **dinámica MF de entrenamiento** (DD) está dada por:

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

Teorema (WGF es G-invariante). Sea $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ convexo y $\mathcal{C}^1$.

Sea R G-invariante, con WGF bien definido y única solución débil $(\mu_t)_{t\geq 0}$.

Si la c.i. cumple $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$ entonces: c.t.p. $\forall t \geq 0$, $\mu_t \in \mathcal{P}_2^G(\mathcal{Z})$

Este resultado general es aplicable en todos los casos de dinámica de entrenamiento que se encuentran en la literatura (con y sin regularización, diferentes *learning rates*, etc.).

Corolario (Caso *regularizado*). Sean R y r G-invariantes y $\tau, \beta > 0$.

- $R^{\tau,\beta}(\mu) = R(\mu) + \tau \int r d\mu + \beta H_\lambda(\mu)$ es G-invariante.
- Si $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$ y u_t densidad de $\mu_t \Rightarrow$ c.t.p. $\forall t \geq 0$, $u_t = \int_G u_t(M_g \cdot) d\lambda_G(g)$

Dinámica de Entrenamiento

Recordemos que la **dinámica MF de entrenamiento** (DD) está dada por:

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

Teorema (WGF es G-invariante). Sea $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ convexo y $\mathcal{C}^1$.

Sea R G-invariante, con WGF bien definido y única solución débil $(\mu_t)_{t\geq 0}$.

Si la c.i. cumple $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$ entonces: c.t.p. $\forall t \geq 0$, $\mu_t \in \mathcal{P}_2^G(\mathcal{Z})$

Este resultado general es aplicable en todos los casos de dinámica de entrenamiento que se encuentran en la literatura (con y sin regularización, diferentes *learning rates*, etc.).

Corolario (Convergencia Global). Sea R G-invariante y $(\mu_t)_{t\geq 0}$ su WGF.

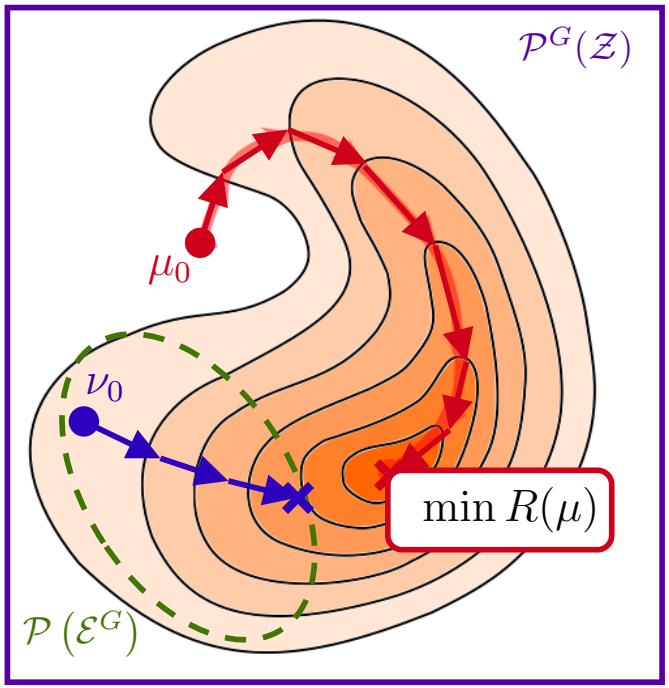
$$\left[\mu_0 \in \mathcal{P}^G(\mathcal{Z}) \text{ y } \mu_t \xrightarrow[t\to\infty]{W_2} \mu_\infty\right] \Rightarrow \left[\mu_\infty \in \mathcal{P}^G(\mathcal{Z}) \text{ y es } \textbf{óptimo}\right]$$

Teorema (WGF respeta a $\mathcal{E}^G$). Sean R y r G-invariantes. Sea $\beta = 0$.
Sea $(\mu_t)_{t\geq 0}$ la (única) solución de la SDE de **McKean-Vlasov** de $R^{\tau,0}$.

Si la c.i. cumple $\mu_0 \in \mathcal{P}_2(\mathcal{E}^G)$ entonces: c.t.p. $\forall t \geq 0$, $\mu_t \in \mathcal{P}_2(\mathcal{E}^G)$

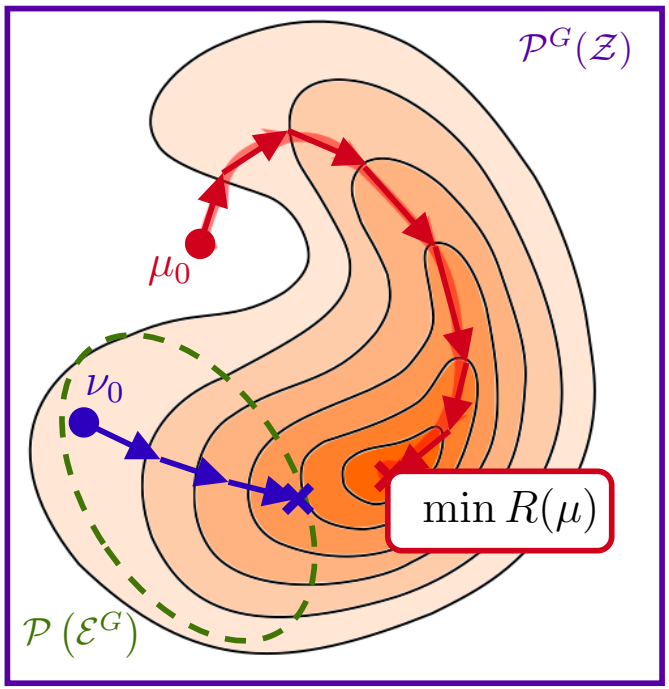
Teorema (WGF respeta a $\mathcal{E}^G$). Sean R y r G-invariantes. Sea $\beta = 0$.
Sea $(\mu_t)_{t\geq 0}$ la (única) solución de la SDE de **McKean-Vlasov** de $R^{\tau,0}$.

Si la c.i. cumple $\mu_0 \in \mathcal{P}_2(\mathcal{E}^G)$ entonces: c.t.p. $\forall t \geq 0$, $\mu_t \in \mathcal{P}_2(\mathcal{E}^G)$



Teorema (WGF respeta a $\mathcal{E}^G$). Sean R y r G-invariantes. Sea $\beta = 0$. Sea $(\mu_t)_{t\geq 0}$ la (única) solución de la SDE de **McKean-Vlasov** de $R^{\tau,0}$.

Si la c.i. cumple $\mu_0 \in \mathcal{P}_2(\mathcal{E}^G)$ entonces: c.t.p. $\forall t \geq 0$, $\mu_t \in \mathcal{P}_2(\mathcal{E}^G)$



Cor. Si $\tilde{\mathcal{Z}} = \mathcal{E}^G$ cumple **(C.T.)**, el WGF de $R|^{\tau,\beta}_{\mathcal{P}(\mathcal{E}^G)}$ satisface **convergencia global**.

Dinámica bajo Aprovechamiento de Simetrías

¿Cómo se ve el WGF de $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ cuando usamos **DA**, **FA** y **EA**?

¿Cómo se ve el WGF de $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ cuando usamos **DA**, **FA** y **EA**?

Corolario (Dinámicas **DA** y **FA**). Sea $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ convexo y $\mathcal{C}^1$.

Cuando los WGF de R^{FA} y R^{G} están bien definidos:

Si ambos inician en $\mu_0 \in \mathcal{P}_2^{\mathcal{G}}(\mathcal{Z})$, entonces **coinciden** $\forall t \geq 0$.

Dinámica bajo Aprovechamiento de Simetrías

¿Cómo se ve el WGF de $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ cuando usamos **DA**, **FA** y **EA**?

Corolario (Dinámicas **DA** y **FA**). Sea $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ convexo y $\mathcal{C}^1$.

Cuando los WGF de R^{FA} y R^G están bien definidos:

Si ambos inician en $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$, entonces **coinciden** $\forall t \geq 0$.

i.e. Si R es G invariante, su dinámica coincide con la de R^{FA}.

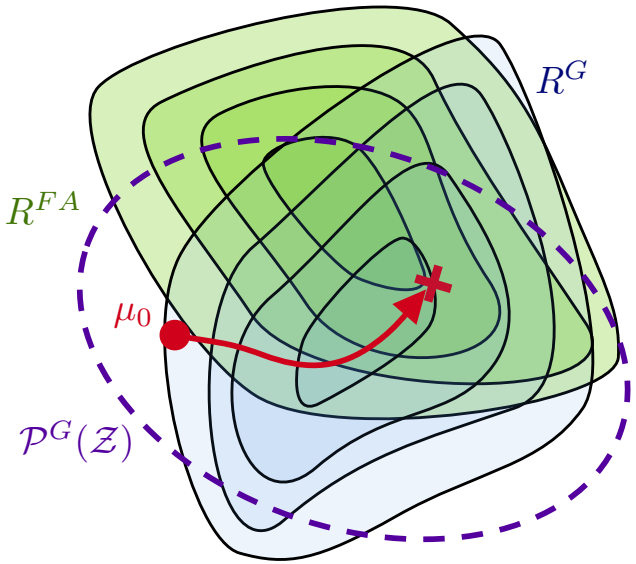
¿Cómo se ve el WGF de $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ cuando usamos **DA**, **FA** y **EA**?

Corolario (Dinámicas **DA** y **FA**). Sea $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ convexo y $\mathcal{C}^1$.

Cuando los WGF de R^{FA} y R^G están bien definidos:

Si ambos inician en $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$, entonces **coinciden** $\forall t \geq 0$.

i.e. Si R es G invariante, su dinámica coincide con la de R^{FA}.



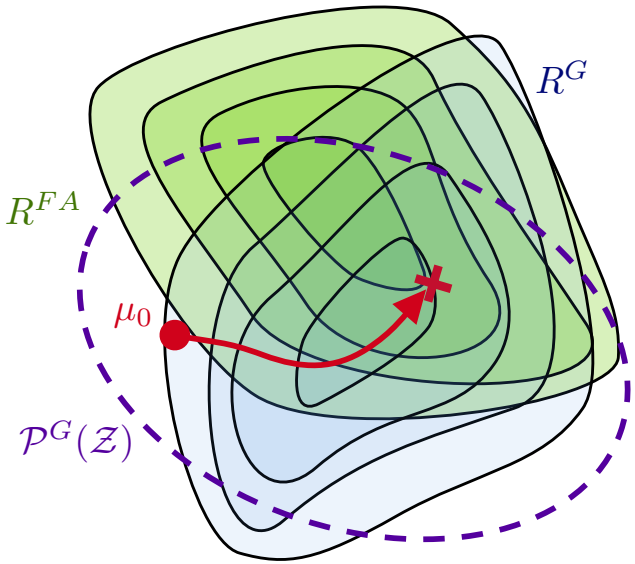
¿Cómo se ve el WGF de $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ cuando usamos **DA**, **FA** y **EA**?

Corolario (Dinámicas **DA** y **FA**). Sea $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ convexo y $\mathcal{C}^1$.

Cuando los WGF de R^{FA} y R^G están bien definidos:

Si ambos inician en $\mu_0 \in \mathcal{P}_2^G(\mathcal{Z})$, entonces **coinciden** $\forall t \geq 0$.

i.e. Si R es G invariante, su dinámica coincide con la de R^{FA}.



¿Cómo se compara el WGF de R^{EA} con el de $R|_{\mathcal{P}(\mathcal{E}^G)}$? ¿Existe algún resultado de coincidencia de dinámicas?

Conclusiones y Trabajo Futuro

Logros Principales

- Entender y describir en amplia **generalidad** tanto el límite MF de NNs como las técnicas de aprovechamiento de simetrías más comunes.

Logros Principales

- Entender y describir en amplia **generalidad** tanto el límite MF de NNs como las técnicas de aprovechamiento de simetrías más comunes.
- Demostrar *extensiones* de resultados conocidos para **unificar** el setting.

Logros Principales

- Entender y describir en amplia **generalidad** tanto el límite MF de NNs como las técnicas de aprovechamiento de simetrías más comunes.
- Demostrar *extensiones* de resultados conocidos para **unificar** el setting.
- Generalizar *razonablemente* las ideas de **EA**, **DA** y **FA** al contexto MF.

Logros Principales

- Entender y describir en amplia **generalidad** tanto el límite MF de NNs como las técnicas de aprovechamiento de simetrías más comunes.
- Demostrar *extensiones* de resultados conocidos para **unificar** el setting.
- Generalizar *razonablemente* las ideas de **EA**, **DA** y **FA** al contexto MF.
- Describir precisamente $\mathcal{P}^G(\mathcal{Z})$, $\mathcal{P}(\mathcal{E}^G)$ y sus propiedades principales.

Logros Principales

- Entender y describir en amplia **generalidad** tanto el límite MF de NNs como las técnicas de aprovechamiento de simetrías más comunes.
- Demostrar *extensiones* de resultados conocidos para **unificar** el setting.
- Generalizar *razonablemente* las ideas de **EA**, **DA** y **FA** al contexto MF.
- Describir precisamente $\mathcal{P}^G(\mathcal{Z})$, $\mathcal{P}(\mathcal{E}^G)$ y sus propiedades principales.
- Comprender de buena manera los funcionales G-invariantes: cómo deben ser sus **óptimos** y sus **WGF**. ***"La G-invarianza se preserva"***

Logros Principales

- Entender y describir en amplia **generalidad** tanto el límite MF de NNs como las técnicas de aprovechamiento de simetrías más comunes.
- Demostrar *extensiones* de resultados conocidos para **unificar** el setting.
- Generalizar *razonablemente* las ideas de **EA**, **DA** y **FA** al contexto MF.
- Describir precisamente $\mathcal{P}^G(\mathcal{Z})$, $\mathcal{P}(\mathcal{E}^G)$ y sus propiedades principales.
- Comprender de buena manera los funcionales G-invariantes: cómo deben ser sus **óptimos** y sus **WGF**. ***"La G-invarianza se preserva"***
- Entender las limitantes de $\mathcal{P}(\mathcal{E}^G)$; y la estrecha relación entre **DA** y **FA**.

Logros Principales

- Entender y describir en amplia **generalidad** tanto el límite MF de NNs como las técnicas de aprovechamiento de simetrías más comunes.
- Demostrar *extensiones* de resultados conocidos para **unificar** el setting.
- Generalizar *razonablemente* las ideas de **EA**, **DA** y **FA** al contexto MF.
- Describir precisamente $\mathcal{P}^G(\mathcal{Z})$, $\mathcal{P}(\mathcal{E}^G)$ y sus propiedades principales.
- Comprender de buena manera los funcionales G-invariantes: cómo deben ser sus **óptimos** y sus **WGF**. ***"La G-invarianza se preserva"***
- Entender las limitantes de $\mathcal{P}(\mathcal{E}^G)$; y la estrecha relación entre **DA** y **FA**.
- **Explicitar interrogantes** que, en un inicio, parecían *etéreas*.

Ventajas de la *reducción de dimensión/aprovechamiento de simetría*

- ¿Qué se gana *cuantitativamente* al *reducir* el problema a $\mathcal{E}^G$ o a $G\backslash\mathcal{Z}$?
- ¿Qué se gana por usar DA, FA o EA al entrenar? ¿Cuál es preferible?
- ¿Bajo qué condiciones sobre $G \circlearrowright \mathcal{Z}$ y $R^{\tau,\beta}$ existe un óptimo en $\mathcal{P}(\mathcal{E}^G)$? ¿Cómo se extiende *canónicamente* a todo $\mathcal{Z}$ (o al menos $G\backslash\mathcal{Z}$)?
- Si $\mathcal{E}^G$ es *universal*: ¿Se parecen los óptimos *restringidos* a los *normales*?

Preguntas Abiertas

Ventajas de la *reducción de dimensión/aprovechamiento de simetría*

- ¿Qué se gana *cuantitativamente* al *reducir* el problema a $\mathcal{E}^G$ o a $G\backslash\mathcal{Z}$?
- ¿Qué se gana por usar DA, FA o EA al entrenar? ¿Cuál es preferible?
- ¿Bajo qué condiciones sobre $G \circlearrowright \mathcal{Z}$ y $R^{\tau,\beta}$ existe un óptimo en $\mathcal{P}(\mathcal{E}^G)$? ¿Cómo se extiende *canónicamente* a todo $\mathcal{Z}$ (o al menos $G\backslash\mathcal{Z}$)?
- Si $\mathcal{E}^G$ es *universal*: ¿Se parecen los óptimos *restringidos* a los *normales*?

Resultados Experimentales

- ¿Podemos comprobar nuestros resultados empíricamente?
- ¿Qué *insights prácticos* podemos obtener desde los resultados teóricos?

Preguntas Abiertas

Ventajas de la *reducción de dimensión/aprovechamiento de simetría*

- ¿Qué se gana *cuantitativamente* al *reducir* el problema a $\mathcal{E}^G$ o a $G\backslash\mathcal{Z}$?
- ¿Qué se gana por usar DA, FA o EA al entrenar? ¿Cuál es preferible?
- ¿Bajo qué condiciones sobre $G \circlearrowright \mathcal{Z}$ y $R^{\tau,\beta}$ existe un óptimo en $\mathcal{P}(\mathcal{E}^G)$? ¿Cómo se extiende *canónicamente* a todo $\mathcal{Z}$ (o al menos $G\backslash\mathcal{Z}$)?
- Si $\mathcal{E}^G$ es *universal*: ¿Se parecen los óptimos *restringidos* a los *normales*?

Resultados Experimentales

- ¿Podemos comprobar nuestros resultados empíricamente?
- ¿Qué *insights prácticos* podemos obtener desde los resultados teóricos?

Teoría Mean Field de Redes Neuronales

- ¿Es la universalidad compatible con que se *alcance el ínfimo*?
- ¿Qué arquitecturas *interesantes* pueden modelarse en el setting *shallow*? ¿Qué tan factible es *tener universalidad* con ellos?
- ¿Estudio de simetrías en el CLT? ¿Extensión al caso Multicapa?

Referencias Bibliográficas I

[Bortoli u. a. 2020] Bortoli, Valentin D. ; Durmus, Alain ; Fontaine, Xavier ; Simsekli, Umut: *Quantitative Propagation of Chaos for SGD in Wide Neural Networks*. 2020

[Bronstein u. a. 2021] Bronstein, Michael M. ; Bruna, Joan ; Cohen, Taco ; Veličković, Petar: *Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges*. 2021. – URL `https://arxiv.org/abs/2104.13478`

[Chen u. a. 2022a] Chen, Fan ; Ren, Zhenjie ; Wang, Songbo: *Uniform-in-Time Propagation of Chaos for Mean Field Langevin Dynamics*. 2022. – URL `https://arxiv.org/abs/2212.03050`

[Chen u. a. 2020] Chen, Shuxiao ; Dobriban, Edgar ; Lee, Jane H.: *A Group-Theoretic Framework for Data Augmentation*. 2020

[Chen u. a. 2022b] Chen, Zhengdao ; Rotskoff, Grant M. ; Bruna, Joan ; Vanden-Eijnden, Eric: *A Dynamical Central Limit Theorem for Shallow Neural Networks*. 2022

[Chizat und Bach 2018] Chizat, Lenaic ; Bach, Francis: *On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport*. 2018

Referencias Bibliográficas II

[Chizat 2022] Chizat, Lénaïc: *Mean-Field Langevin Dynamics: Exponential Convergence and Annealing*. 2022

[Descours u. a. 2023] Descours, Arnaud ; Guillin, Arnaud ; Michel, Manon ; Nectoux, Boris: *Law of large numbers and central limit theorem for wide two-layer neural networks: the mini-batch and noisy case*. 2023

[Elesedy und Zaidi 2021] Elesedy, Bryn ; Zaidi, Sheheryar: *Provably Strict Generalisation Benefit for Equivariant Models*. 2021

[Finzi u. a. 2021] Finzi, Marc ; Welling, Max ; Wilson, Andrew G.: *A Practical Method for Constructing Equivariant Multilayer Perceptrons for Arbitrary Matrix Groups*. 2021

[Flinth und Ohlsson 2023] Flinth, Axel ; Ohlsson, Fredrik: *Optimization Dynamics of Equivariant and Augmented Neural Networks*. 2023

[Hu u. a. 2020] Hu, Kaitong ; Ren, Zhenjie ; Siska, David ; Szpruch, Lukasz: *Mean-Field Langevin Dynamics and Energy Landscape of Neural Networks*. 2020

[Lyle u. a. 2020] Lyle, Clare ; Wilk, Mark van der ; Kwiatkowska, Marta ; Gal, Yarin ; Bloem-Reddy, Benjamin: *On the Benefits of Invariance in Neural Networks*. 2020

Referencias Bibliográficas III

[Mei u. a. 2019] Mei, Song ; Misiakiewicz, Theodor ; Montanari, Andrea: *Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit*. 2019

[Mei u. a. 2018] Mei, Song ; Montanari, Andrea ; Nguyen, Phan-Minh: A mean field view of the landscape of two-layer neural networks. In: *Proceedings of the National Academy of Sciences* 115 (2018), Nr. 33, S. E7665–E7671. – URL `https://www.pnas.org/doi/abs/10.1073/pnas.1806579115`

[Nitanda u. a. 2022] Nitanda, Atsushi ; Wu, Denny ; Suzuki, Taiji: *Convex Analysis of the Mean Field Langevin Dynamics*. 2022

[Rotskoff und Vanden-Eijnden 2022] Rotskoff, Grant ; Vanden-Eijnden, Eric: Trainability and Accuracy of Artificial Neural Networks: An Interacting Particle System Approach. In: *Communications on Pure and Applied Mathematics* 75 (2022), jul, Nr. 9, S. 1889–1935. – URL `https://doi.org/10.1002%2Fcpa.22074`

[Sirignano und Spiliopoulos 2018] Sirignano, Justin ; Spiliopoulos, Konstantinos: *Mean Field Analysis of Neural Networks: A Law of Large Numbers*. 2018. – URL `https://arxiv.org/abs/1805.01053`

[Sirignano und Spiliopoulos 2019] Sirignano, Justin ; Spiliopoulos, Konstantinos: *Mean Field Analysis of Neural Networks: A Central Limit Theorem*. 2019

¡Gracias por su atención!

Simetrías en Redes Neuronales Sobreparametrizadas: Una mirada de Campo Medio

Javier Maass Martínez

DIM, Universidad de Chile
Tesis de Magíster en Matemáticas Aplicadas
Memoria de Ingeniería Civil Matemática

25 de marzo de 2024

Propiedades Extra

Linear Functional Derivative

Es una función: $\frac{\partial R}{\partial \mu} : \mathcal{P}(\mathcal{Z}) \times \mathcal{Z} \to \mathbb{R}$ tq:

$$\forall \mu, \nu \in \mathcal{P}(\mathcal{Z}), \quad \lim_{h \to 0} \frac{R((1-h)\mu + h\nu) - R(\mu)}{h} = \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu, \theta) d(\nu - \mu)(\theta)$$

y que cumple: $$\int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu, \theta) d\mu(\theta) = 0$$

A $R' : \mu \in \mathcal{P}(\mathcal{Z}) \mapsto \frac{\partial R}{\partial \mu}(\mu, \cdot)$ se le conoce como la *primera variación* de R en μ.

Derivada Intrínseca

Si $\frac{\partial R}{\partial \mu} : \mathcal{P}(\mathcal{Z}) \times \mathcal{Z} \to \mathbb{R}$ existe y es **diferenciable** en su segundo argumento; la *derivada intrínseca* se define como:

$$D_\mu R(\mu, \theta) = \nabla_\theta \left(\frac{\partial R}{\partial \mu}(\mu, \theta) \right)$$

Propiedades Extra

En el caso del problema de Aprendizaje:

$$\frac{\partial R}{\partial \mu}(\mu,\theta) = \mathbb{E}_\pi\left[\langle \nabla_1 \ell(\langle \sigma_*(X;\cdot),\mu\rangle, Y), \sigma_*(X;\theta)\rangle_{\mathcal{Y}}\right] + (\text{cte no dep. de } z)$$

$$D_\mu R(\mu,\theta) = \mathbb{E}_\pi\left[\nabla_\theta \sigma_*(X;\theta).\nabla_1 \ell(\langle \sigma_*(X;\cdot),\mu\rangle, Y)\right]$$

Ser $\mathcal{C}^1$

$R : \mathcal{P}_p(\mathcal{Z}) \to \mathbb{R}$ se dice de clase $\mathcal{C}^1$ si $\frac{\partial R}{\partial \mu}(\mu,\cdot)$ está bien definida y es acotada para cada $\mu \in \mathcal{P}_p(\mathcal{Z})$; y además$(\mu, z) \in \mathcal{P}_p(\mathcal{Z}) \times \mathcal{Z} \mapsto \frac{\partial R}{\partial \mu}(\mu, z)$ es **continua**.

Lema Subgradiente

Si $R : \mathcal{P}_p(\mathcal{Z}) \to \mathbb{R}$ es **convexa** y de clase $\mathcal{C}^1$. Entonces $\forall \mu, \mu' \in P_p(\mathcal{Z})$:

$$R(\mu') - R(\mu) \geq \int_{\mathcal{Z}} \frac{\partial R}{\partial \mu}(\mu, z) d(\mu' - \mu)(z)$$

Propagación de Caos (versión LGN)

$\forall t > 0$, $(\theta_1^{\lfloor \frac{t}{\varepsilon_N} \rfloor}, \ldots, \theta_N^{\lfloor \frac{t}{\varepsilon_N} \rfloor})$ es μ_t-**caótico**. i.e. $\forall j \in \mathbb{N}^*$, $\textbf{Ley}(\theta_1^{\lfloor Nt \rfloor}, \ldots, \theta_j^{\lfloor Nt \rfloor}) \xrightarrow[N\to\infty]{} (\mu_t)^{\otimes j}$

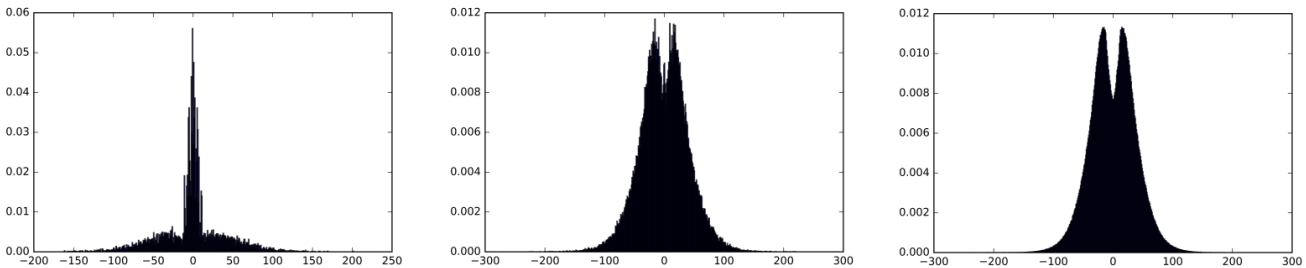


Figura: Distribución de parámetros en el límite de ancho infinito de NNs

Ejemplo clásico: SGD con regularización

Para la iteración de forma (con $\tau, \beta > 0$, $B \in \mathbb{N}^*$, $\xi_i^k \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathrm{Id}_{\mathcal{Z}})$ y $r : \mathcal{Z} \to \mathbb{R}$):

$$\theta_i^{k+1} = \theta_i^k - s_k^N \left(\frac{1}{B} \sum_{j=1}^{B} \partial_1 \ell(\Phi_\theta^N(X_j^k), Y_j^k) \nabla_{\theta_i}(\sigma_*(X_j^k; \theta_i^k)) + \tau \nabla_{\theta_i} r(\theta_i^k) \right) + \sqrt{2\beta s_k^N} \xi_i^k$$

El límite MF corresponde al WGF de: $R^{\tau \cdot \beta}(\mu) := R(\mu) + \tau \int r d\mu + \beta H_\lambda(\mu)$.

i.e. $\partial_t \mu_t = \varsigma(t) [\mathrm{div}((D_\mu R(\mu_t, \cdot) + \tau \nabla_\theta r) \mu_t) + \beta \Delta \mu_t]$ o, equivalentemente:

$$dZ_t = \varsigma(t) \left[-(D_\mu R(\mu_t, Z_t) + \tau \nabla_\theta r(Z_t))\, dt + \sqrt{2\beta} dB_t \right] \text{ con } \mu_t = \textbf{Ley}(Z_t) \text{ y } (B_t)_{t \geq 0} \text{ MB}$$

Teorema (Disipación de Energía y Convergencia) **(HRSS'19,CRW'22)**

Sea $(\mu_t)_{t\geq 0}$ el WGF de $R^{\tau,\beta}$. Bajo **(C.T.)**, se tiene que $\forall t > 0$:

$$\frac{d}{dt}(R^{\tau,\beta}(\mu_t)) = -\varsigma(t)\int_{\mathcal{Z}} \left| D_\mu R(\mu_t, z) + \tau \nabla r(z) + \beta \frac{\nabla u_t}{u_t}(z) \right|^2 d\mu_t(z)$$

Además, $W_2(\mu_t, \mu_*^{\tau,\beta}) \xrightarrow[t\to\infty]{} 0$, donde $\mu_*^{\tau,\beta}$ es el único mínimo de $R^{\tau,\beta}$.

Propiedades Extra

Proposición 24

Dado $\mu \in \mathcal{P}_p(\mathcal{Z})$, la función $\nu \in \mathcal{P}_p(\mathcal{Z}) \mapsto W_p(\nu, \mu)$ es *convexa* y *continua*.
Si $\mathcal{Z} = \mathbb{R}^D$, $\mu \ll \lambda$ y $p > 1$, es de hecho *estrictamente convexa*.
Si $G \circlearrowright_M \mathcal{Z}$ *ortogonalmente*, la función $W_p : \mathcal{P}_p(\mathcal{Z}) \times \mathcal{P}_p(\mathcal{Z}) \to \mathbb{R}$ es (conjuntamente) G-invariante.

Proposición 25

Si $\mathcal{E}$ s.e.v. cerrado de $\mathcal{Z}$ con proyección ortogonal $P_{\mathcal{E}}$ y $\mu \in \mathcal{P}_p(\mathcal{Z})$; entonces:

1. $\mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$ y $\mathcal{P}_p^{G}(\mathcal{Z})$ son subespacios *cerrados* y *convexos* de $\mathcal{P}_p(\mathcal{Z})$.
2. $\mu^{\mathcal{E}} \in \mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$ y $\mu^{G} \in \mathcal{P}_p^{G}(\mathcal{Z})$
3. $\mu^{\mathcal{E}}$ es *una proyección* de μ en $\mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$; en el sentido de que minimiza $W_p(\mu, \cdot)$ sobre $\mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$. Si $\mathcal{Z} = \mathbb{R}^D$, $\mu \ll \lambda$ y $p > 1$, entonces es *la* única tal proyección.
4. $\mu \in \mathcal{P}_p^{G}(\mathcal{Z}) \iff \mu = \mu^{G}$ and $\mu \in \mathcal{P}_p^{\mathcal{E}}(\mathcal{Z}) \iff \mu = \mu^{\mathcal{E}}$

Propiedades Extra

Lema 13

Se sabe que: $\mathcal{P}^{\mathcal{E}^G}(\mathcal{Z}) \subseteq \mathcal{P}^G(\mathcal{Z})$ y $\forall \mu \in \mathcal{P}(\mathcal{Z})$, $\mu^{\mathcal{E}^G} = (\mu^G)^{\mathcal{E}^G} = (\mu^{\mathcal{E}^G})^G$.

Proposición 26

Si $\mathcal{Z} = \mathbb{R}^D$ y $\mu \in \mathcal{P}(\mathcal{Z})$ tq $\mu \lll \lambda$ (con densidad $u : \mathcal{Z} \to \mathbb{R}_+$):

- $\mu^{\mathcal{E}^G}$ tiene densidad c/r a $\lambda_{\mathcal{E}^G} := P_{\mathcal{E}^G} \# \lambda$ (restringido a $\mathcal{E}^G$)).
- $\mu^G \in \mathcal{P}^G(\mathcal{Z})$ tiene densidad $u^G := \int_G u \circ M_g d\lambda_G(g)$ c/r a λ. En particular: $\mu \in \mathcal{P}^G(\mathcal{Z}) \iff$ *u es G-invariante* $(\lambda - c.s.)$

Propiedades Extra

$\mathcal{X}, \mathcal{Y}$ y $\mathcal{Z}$ Hilbert y G grupo lcsH tal que $G \curvearrowright_{\chi} \mathcal{X}$, $G \curvearrowright_{\tilde{\chi}} \mathcal{Z}$, $G \curvearrowright_{\check{\chi}} \mathcal{Y}$.

Proposición (Derivada de Funciones Equivariantes)

Sea $f : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ G-equivariante y Fréchet-diferenciable en su primer argumento; entonces:

$$\forall g \in G,\ \forall x \in \mathcal{X},\ \forall z \in \mathcal{Z},\ \ D_x f(\chi_g.x, \tilde{\chi}_g.z) = \check{\chi}_g.D_x f(x,z)\chi_g^{-1}$$

Proposición (Integral de Funciones Equivariantes)

Sea $\mu \in \mathcal{P}(\mathcal{Z})$ y sea $f : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ G-equivariante y Bochner integrable en su segundo argumento c/r a μ; entonces:

$$\forall x \in \mathcal{X},\ \forall g \in G,\ \check{\chi}_g \langle f(x;\cdot), \mu \rangle = \langle f(\chi_g x;\cdot), \tilde{\chi}_g \# \mu \rangle$$

Por la cerradura de los espacios, si $(\mu_t)_{t\geq 0} \subseteq \mathcal{P}_p(\mathcal{Z})$ es un flujo de medidas tal que $W_p(\mu_t, \mu_*) \xrightarrow[t\to\infty]{} 0$ para algún $\mu_* \in \mathcal{P}_p(\mathcal{Z})$, entonces:

- Cuando $(\mu_t)_{t\geq 0} \subseteq \mathcal{P}_p^G(\mathcal{Z})$, $\mu_* \in \mathcal{P}_p^G(\mathcal{Z})$
- Cuando $(\mu_t)_{t\geq 0} \subseteq \mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$, $\mu_* \in \mathcal{P}_p^{\mathcal{E}}(\mathcal{Z})$

Propiedades Extra

Con $\beta > 0$, la *entropía obliga a tener densidad c/r a* $\lambda \to$ NO se cumple lo mismo.

Esto se *resuelve* **proyectando** el ruido del entrenamiento a $\mathcal{E}^G$:

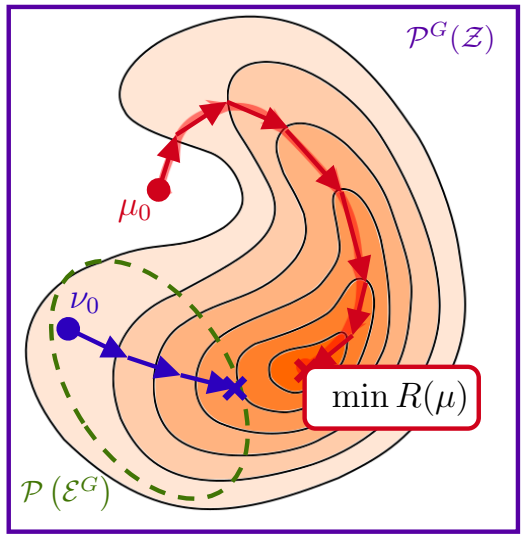
$$\theta_i^{k+1} = \theta_i^k - s_k^N \left(\partial_1 \ell(\Phi_\theta^N(X_k), Y_k) \nabla_{\theta_i}(\sigma_*(X_k; \theta_i^k)) + \tau \nabla_{\theta_i} r(\theta_i^k) \right) + \sqrt{2\tau s_k^N} P_{\mathcal{E}^G} \xi_i^k$$

Teo. La DD *proyectada* respeta a $\mathcal{E}^G$:

$$\left[\mu_0 \in \mathcal{P}_2(\mathcal{E}^G) \Rightarrow (\mu_t)_{t \geq 0} \subseteq \mathcal{P}_2(\mathcal{E}^G) \right]$$

¡Trampa! La DD está siendo **forzada** a seguir en $\mathcal{E}^G$.

Cor. Si $\tilde{\mathcal{Z}} = \mathcal{E}^G$ cumple **(C.T.)**, el WGF de $R|^{\tau,\beta}_{\mathcal{P}(\mathcal{E}^G)}$ satisface **convergencia global**.



Propiedades Extra

Con $\beta > 0$, la *entropía obliga a tener densidad c/r a* $\lambda \rightarrow$ NO se cumple lo mismo.

Esto se *resuelve* **proyectando** el ruido del entrenamiento a $\mathcal{E}^G$:

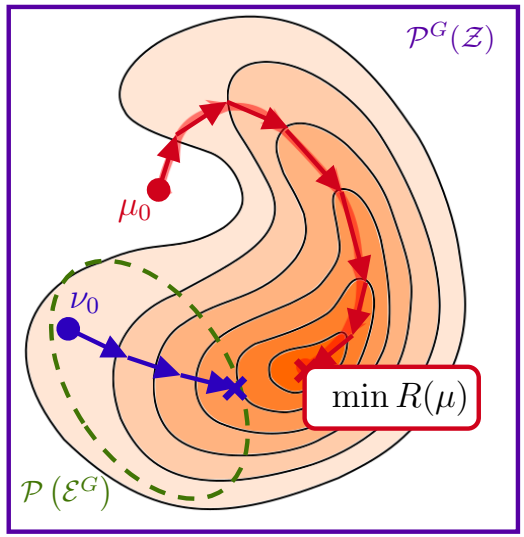
$$\mathrm{d}Z_t = -\left[D_\mu\left(R(\mu_t, Z_t)\right) + \tau \nabla_\theta r(Z_t)\right]\mathrm{d}t + \sqrt{2\tau}\, P_{\mathcal{E}^G}\, dB_t$$

Teo. La DD *proyectada* respeta a $\mathcal{E}^G$:

$$\left[\mu_0 \in \mathcal{P}_2(\mathcal{E}^G) \Rightarrow (\mu_t)_{t\geq 0} \subseteq \mathcal{P}_2(\mathcal{E}^G)\right]$$

¡Trampa! La DD está siendo **forzada** a seguir en $\mathcal{E}^G$.

Cor. Si $\tilde{\mathcal{Z}} = \mathcal{E}^G$ cumple **(C.T.)**, el WGF de $R|^{\tau,\beta}_{\mathcal{P}(\mathcal{E}^G)}$ satisface **convergencia global**.



Hipótesis NN

Standard Assumptions on *shallow NNs*

1. $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}^c$ y $\mathcal{Z} = \mathbb{R}^{c\times b} \times \mathbb{R}^{d\times b} \times \mathbb{R}^b$.
2. $\sigma_* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ es de la forma:

$$\forall \theta = (W, A, B) \in \mathcal{Z};\ \forall x \in \mathcal{X},\ \sigma_*(x;\theta) = \varphi(W)\sigma(A^T x + B)$$

   con $\sigma : \mathbb{R} \to \mathbb{R}$ aplicada *pointwise* y $\varphi : \mathbb{R} \to [-M, M]$ una función de *truncación*, con $M < +\infty$. σ y φ son al menos $\mathcal{C}^1$.
3. $\pi \in \mathcal{P}_2(\mathcal{X} \times \mathcal{Y})$
4. $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ convexa, suave, y cumple $\ell \geq 0$.
5. Sea $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ dado por $\forall \mu \in \mathcal{P}(\mathcal{Z})$, $R(\mu) = \mathbb{E}_\pi[\ell(\langle \sigma_*(X, \cdot), \mu \rangle, Y)]$.

Resultado

Si además:

- $\forall \theta \in \mathcal{Z}$, $\sigma_*(\cdot, \theta) \in L^2(\pi|_{\mathcal{X}})$ y
  $\exists C > 0$, $\forall \theta \in \mathcal{Z}$, $\|\sigma_*(\cdot, \theta)\|_{L^2(\pi|_{\mathcal{X}})} \leq C(1 + |\theta|^2)$. Entonces
  $R : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ es **convexo**, $\mathcal{C}^1$.
- Si $\ell(y, \hat{y}) = \|y - \hat{y}\|_2^2$ (cuadrática), $r(\theta) = \|\theta\|^2$ (cuadrática), y σ y φ tienen derivadas acotadas hasta el orden 4. Entonces:
  - $D_\mu^2 R^\tau$ tiene **2-norm** acotada, haciendo $D_\mu R^\tau(\cdot, z)$ W_1-Lipschitz de constante $M_{mm}^{R^\tau} = \left(\|\varphi'\|_\infty + \|\varphi\|_\infty \|\sigma'\|_\infty (1 + \int |x|^2 \pi_{\mathcal{X}}(dx)\right)^{1/2}$.
  - Por acotamiento de σ y φ, se tienen las cotas de Lipschitz para R.

  Por acotamiento de φ y que $\pi \in \mathcal{P}_2(\mathcal{X} \times \mathcal{Y})$, hay Uniform LSI con $\vartheta = \tau \exp\left(-2(E_\pi[\|Y\|] + \|\varphi\|_\infty)\|\varphi\|_\infty\right)$.