# Approximate Equivariance and Generalization

Shubhendu Trivedi

Boston Symmetry Day

07 April 2023

Northeastern University

$$f(g(x)) = g(f(x))$$

# G-Equivariance

$$\begin{tikzcd} L(X_1) \arrow[r, "\mathbb{T}_g"] \arrow[d, "\phi"] & L(X_1) \arrow[d, "\phi"] \\ L(X_2) \arrow[r, "\mathbb{T}'_g"] & L(X_2) \end{tikzcd}$$

A note on the title
(let's see if we can get to displacement structures!).

# Approximate Equivariance

- Most relevant to us:

$$|f(\rho_{in}(g)x) - \rho_{out}(g)f(x)| < \varepsilon$$

# Approximate Equivariance

- Most relevant to us:

$$|f(\rho_{in}(g)x) - \rho_{out}(g)f(x)| < \varepsilon$$

- Real world symmetries are rarely exact

# Approximate Equivariance

- Most relevant to us:

$$|f(\rho_{in}(g)x) - \rho_{out}(g)f(x)| < \varepsilon$$

- Real world symmetries are rarely exact
- Not to be confused with partial symmetries.

# Approximate Equivariance

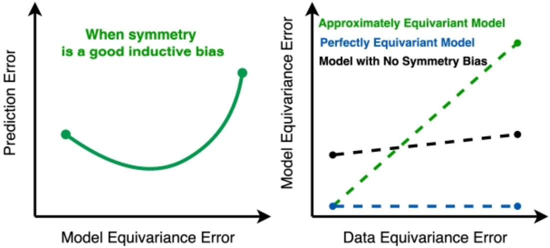



*Figure from: Wang, Walters, and Yu ICML 2022*

- There is a common intuition that there is a *sweet spot* for balancing between model and data equivariance that can lead to good generalization

# Approximate Equivariance

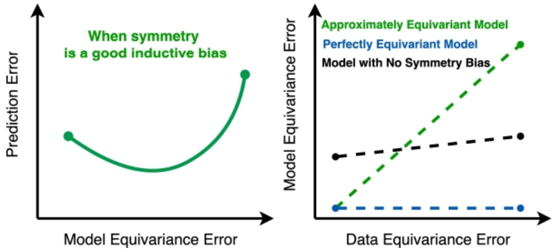



*Figure from: Wang, Walters, and Yu ICML 2022*

- There is a common intuition that there is a *sweet spot* for balancing between model and data equivariance that can lead to good generalization
- How do we show this is the case?

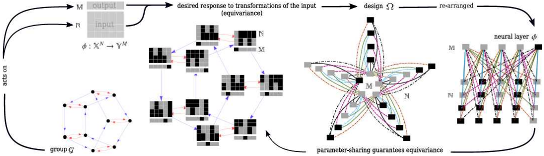
## Some Theoretical Developments

An Approximate Survey

# Some Axes of Theoretical Development

## 1 Architectural Characterizations

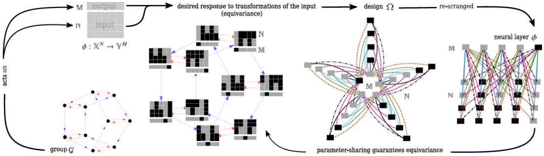
- Discrete groups: Equivariance and Parameter Sharing. Ravanbaksh, Schneider, Poczos, ICML 2017



# Some Axes of Theoretical Development

1 Architectural Characterizations

- Discrete groups: Equivariance and Parameter Sharing. Ravanbaksh, Schneider, Poczos, ICML 2017

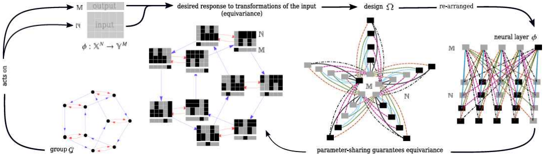



- Convolution $\iff$ Equivariance (Kondor & Trivedi); scalar fields, general compact groups, ICML 2018.

# Some Axes of Theoretical Development

1 Architectural Characterizations

- Discrete groups: Equivariance and Parameter Sharing. Ravanbaksh, Schneider, Poczos, ICML 2017




- Convolution $\iff$ Equivariance (Kondor & Trivedi); scalar fields, general compact groups, ICML 2018.
- Convolution $\iff$ Equivariance (Cohen, Geiger, & Weiler); steerable case, general compact groups, NeurIPS 2020

# Some Axes of Theoretical Development

1 Architectural Characterizations

- Convolution $\iff$ Equivariance (extensions): Aronsson, Olhsson, Persson et al.

# Some Axes of Theoretical Development

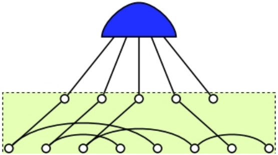
1 Architectural Characterizations

- Convolution $\iff$ Equivariance (extensions): Aronsson, Olhsson, Persson et al.
- Classification of equivariant networks; space of invariant networks. Agrawal and Ostrowski, 2022 NeurIPS, 2023.

# Some Axes of Theoretical Development

1 Architectural Characterizations

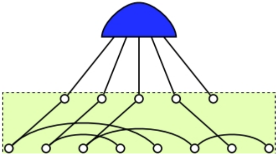
- Convolution $\iff$ Equivariance (extensions): Aronsson, Olhsson, Persson et al.
- Classification of equivariant networks; space of invariant networks. Agrawal and Ostrowski, 2022 NeurIPS, 2023.
- Characterization of Linear Layers:
  - $\mathbb{S}_n$, Maron, Ben-Hamu, Shamir, Lipman, ICLR 2019
  - $A_n$, $Sp(n)$, $O(n)$, $SO(n)$, Pearce-Crump 2023a, 2023b, 2023c.



# Some Axes of Theoretical Development

1 Architectural Characterizations

- Convolution $\iff$ Equivariance (extensions): Aronsson, Olhsson, Persson et al.
- Classification of equivariant networks; space of invariant networks. Agrawal and Ostrowski, 2022 NeurIPS, 2023.
- Characterization of Linear Layers:
  - $\mathbb{S}_n$, Maron, Ben-Hamu, Shamir, Lipman, ICLR 2019
  - $A_n$, $Sp(n)$, $O(n)$, $SO(n)$, Pearce-Crump 2023a, 2023b, 2023c.

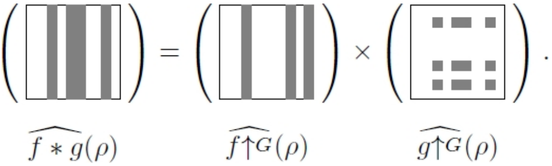


- Similar for Quantum group equivariant neural networks.

# Some Axes of Theoretical Development

2 Fourier picture

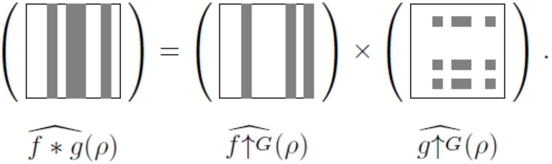
- For scalar fields, Kondor & Trivedi, 2018



# Some Axes of Theoretical Development

2 Fourier picture

- For scalar fields, Kondor & Trivedi, 2018




- For the steerable case, Xu, Lei, Dobriban, & Daniilidis, ICML 2022.

# Some Axes of Theoretical Development

3 Universality Results

# Some Axes of Theoretical Development

3 Universality Results

- Large body of work at this point: Yarotsky, 2018; Keriven & Peyre, 2019; Sannai *et al.*, 2019; Maron *et al.*, 2019; Segol & Lipman, 2020; Ravanbakhsh, 2020.

# Some Axes of Theoretical Development

3 Universality Results

- Large body of work at this point: Yarotsky, 2018; Keriven & Peyre, 2019; Sannai *et al.*, 2019; Maron *et al.*, 2019; Segol & Lipman, 2020; Ravanbakhsh, 2020.

4 More Universality Results

- Barron's analogue for equivariant networks, Lawrence 2022.

# Some Axes of Theoretical Development

4 Generalization/Sample Complexity

- Initial results go back to John-Shawe Taylor

# Some Axes of Theoretical Development

4 Generalization/Sample Complexity

- Initial results go back to John-Shawe Taylor 
- A whole body of work on group NNs (1989, 1991, 1993, 1995)
- Jeffrey Wood (1996)

# Some Axes of Theoretical Development

4 Generalization/Sample Complexity

- Initial results go back to John-Shawe Taylor
- A whole body of work on group NNs (1989, 1991, 1993, 1995)
- Jeffrey Wood (1996)
- Elsedy & Zaidi, ICML 2021: Strict generalization benefit for equivariant linear models. Generalization gap depends on the dimension of the space of anti-symmetric linear maps.

5 PAC-Bayesian Style Bounds:

# Some Axes of Theoretical Development

4 Generalization/Sample Complexity

- Initial results go back to John-Shawe Taylor 
- A whole body of work on group NNs (1989, 1991, 1993, 1995)
- Jeffrey Wood (1996)
- Elsedy & Zaidi, ICML 2021: Strict generalization benefit for equivariant linear models. Generalization gap depends on the dimension of the space of anti-symmetric linear maps.

5 PAC-Bayesian Style Bounds:

- Behboodi, Cesa, & Cohen, NeurIPS 2022.

# Some Axes of Theoretical Development

6 Augmentation:

## Some Axes of Theoretical Development

6 Augmentation:

- A theory for data augmentation, Chen, Dobriban, Lee, NeurIPS 2020, JMLR 2021; Lyle, van der Wilk, *et al.*, ICML 2020.

## Some Axes of Theoretical Development

6 Augmentation:

- A theory for data augmentation, Chen, Dobriban, Lee, NeurIPS 2020, JMLR 2021; Lyle, van der Wilk, *et al.*, ICML 2020.

7 Expressivity

# Some Axes of Theoretical Development

6 Augmentation:

- A theory for data augmentation, Chen, Dobriban, Lee, NeurIPS 2020, JMLR 2021; Lyle, van der Wilk, *et al.*, ICML 2020.

7 Expressivity

- Group invariant capacity, Farrell, Bordelon, Trivedi, Pehlevan, ICLR 2022.

# Some Axes of Theoretical Development

6 Augmentation:

- A theory for data augmentation, Chen, Dobriban, Lee, NeurIPS 2020, JMLR 2021; Lyle, van der Wilk, *et al.*, ICML 2020.

7 Expressivity

- Group invariant capacity, Farrell, Bordelon, Trivedi, Pehlevan, ICLR 2022.

8 Partial, approximate and Incorrect

# Some Axes of Theoretical Development

6 Augmentation:

- A theory for data augmentation, Chen, Dobriban, Lee, NeurIPS 2020, JMLR 2021; Lyle, van der Wilk, *et al.*, ICML 2020.

7 Expressivity

- Group invariant capacity, Farrell, Bordelon, Trivedi, Pehlevan, ICLR 2022.

8 Partial, approximate and Incorrect

- e.g. Wang, Zhu, Park, Platt, & Walters

Mircea Petrache
Pontificia Universidad Católica de Chile

# Goals for [no longer a] Vignette One

- Part 1: Sketch quantitative bounds for the common intuition that model respecting underlying symmetries afford better generalization

# Goals for [no longer a] Vignette One

- Part 1: Sketch quantitative bounds for the common intuition that model respecting underlying symmetries afford better generalization
- Part 2: Use above to tease out dependence on the optimal equivariance error due to model symmetries and the equivariance error due to data symmetries

# Goals for [no longer a] Vignette One

- Part 1: Sketch quantitative bounds for the common intuition that model respecting underlying symmetries afford better generalization

- Part 2: Use above to tease out dependence on the optimal equivariance error due to model symmetries and the equivariance error due to data symmetries

- Overall Goal: Theoretically understand the dependence of optimal model equivariance for data with pre-specified symmetries.

# Approximate Equivariance

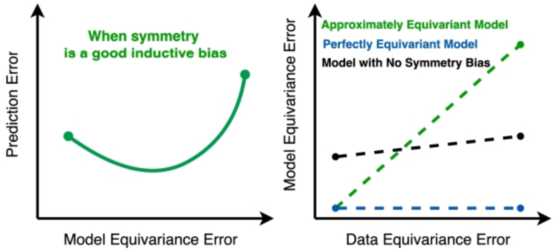
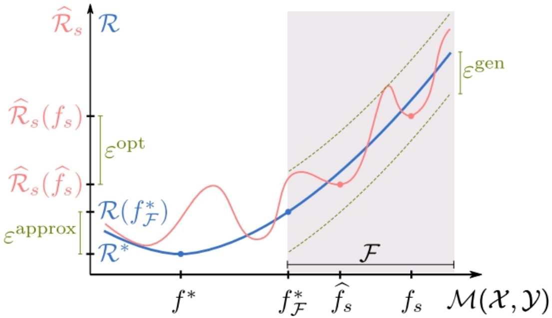



*Figure from: Wang, Walters, and Yu ICML 2022*

- There is a common intuition that there is a *sweet spot* for balancing between model and data equivariance that can lead to good generalization
- How do we show this is the case?

# How does the picture look like?




*Courtesy: J. Berner*

- Blue is the exemplary risk, red is the empirical risk [with respect to the projected space of measurable functions $\mathcal{M}(\mathcal{X}, \mathcal{Y})$)]
  - $\mathcal{R}^*$ is the so-called Bayes Risk.
  - Usual picture: Error: $\varepsilon^{\text{approx}} + \varepsilon^{\text{opt}} + \varepsilon^{\text{gen}}$

## Generalization Error

## Improved Generalization with Equivariance

# Let's Dive Straight In! – Initial Setup

- Consider a family of functions $\tilde{\mathcal{F}} \subset \{\tilde{f} : \mathcal{X} \to \mathcal{Y}\}$

# Let's Dive Straight In! – Initial Setup

- Consider a family of functions $\tilde{\mathcal{F}} \subset \{\tilde{f} : \mathcal{X} \to \mathcal{Y}\}$
- ... and a loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$

# Let's Dive Straight In! – Initial Setup

- Consider a family of functions $\tilde{\mathcal{F}} \subset \{\tilde{f} : \mathcal{X} \to \mathcal{Y}\}$
- ... and a loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$
- Fix a distribution for data $Z = (X, Y) \backsim \mathcal{D}$ and consider samples drawn i.i.d $Z_i = (X_i, Y_i)$

# Let's Dive Straight In! – Initial Setup

- Consider a family of functions $\tilde{\mathcal{F}} \subset \{\tilde{f} : \mathcal{X} \to \mathcal{Y}\}$
- ... and a loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$
- Fix a distribution for data $Z = (X, Y) \backsim \mathcal{D}$ and consider samples drawn i.i.d $Z_i = (X_i, Y_i)$
- Work with random functions $f(x, y) := \ell(\tilde{f}(x), y)$ and define:

$$\mathcal{F} := \{f : \tilde{f} \in \tilde{\mathcal{F}}\},$$

# Let's Dive Straight In! – Initial Setup

- Consider a family of functions $\tilde{\mathcal{F}} \subset \{\tilde{f} : \mathcal{X} \to \mathcal{Y}\}$
- ... and a loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$
- Fix a distribution for data $Z = (X, Y) \backsim \mathcal{D}$ and consider samples drawn i.i.d $Z_i = (X_i, Y_i)$
- Work with random functions $f(x, y) := \ell(\tilde{f}(x), y)$ and define:

$$\mathcal{F} := \{f : \tilde{f} \in \tilde{\mathcal{F}}\}, Pf := \mathbb{E}[f(X, Y)],$$

# Let's Dive Straight In! – Initial Setup

- Consider a family of functions $\tilde{\mathcal{F}} \subset \{\tilde{f} : \mathcal{X} \to \mathcal{Y}\}$
- ... and a loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$
- Fix a distribution for data $Z = (X, Y) \backsim \mathcal{D}$ and consider samples drawn i.i.d $Z_i = (X_i, Y_i)$
- Work with random functions $f(x, y) := \ell(\tilde{f}(x), y)$ and define:

$$\mathcal{F} := \{f : \tilde{f} \in \tilde{\mathcal{F}}\}, Pf := \mathbb{E}[f(X, Y)], \text{ and } P_n f := \frac{1}{n}\sum_{i=1}^{n} f(X_i, Y_i)$$

## Let's Dive Straight In! – Initial Setup

- Consider a family of functions $\tilde{\mathcal{F}} \subset \{\tilde{f} : \mathcal{X} \to \mathcal{Y}\}$
- ... and a loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$
- Fix a distribution for data $Z = (X, Y) \backsim \mathcal{D}$ and consider samples drawn i.i.d $Z_i = (X_i, Y_i)$
- Work with random functions $f(x, y) := \ell(\tilde{f}(x), y)$ and define:

$$\mathcal{F} := \{f : \tilde{f} \in \tilde{\mathcal{F}}\}, Pf := \mathbb{E}[f(X, Y)], \text{ and } P_n f := \frac{1}{n}\sum_{i=1}^{n} f(X_i, Y_i)$$

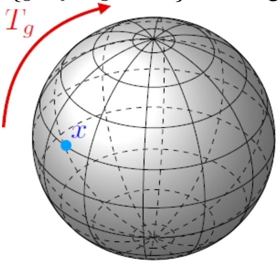
- Reminder: Usually care about $\sup_{f \in \mathcal{F}} Pf - P_n f$

# Initial Setup: Group Case

- Group $G$ acts over $\mathcal{X}$, $\mathcal{Y}$
- ... and transforms any $\tilde{f} \in \tilde{\mathcal{F}}$ into $g \cdot \tilde{f},$ so $x \mapsto g^{-1} \cdot \tilde{f}(g \cdot x)$

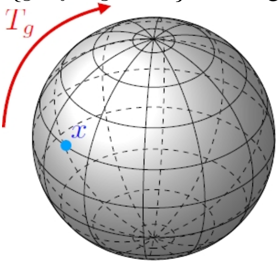
# Initial Setup: Group Case

- Group $G$ acts over $\mathcal{X}$, $\mathcal{Y}$
- ... and transforms any $\tilde{f} \in \tilde{\mathcal{F}}$ into $g \cdot \tilde{f}$, so $x \mapsto g^{-1} \cdot \tilde{f}(g \cdot x)$
- Get a new set $G \cdot f := \{g \cdot \tilde{f} : g \in G\}$ having *orbits* of a given $\tilde{f}$



# Initial Setup: Group Case

- Group $G$ acts over $\mathcal{X}$, $\mathcal{Y}$
- ... and transforms any $\tilde{f} \in \tilde{\mathcal{F}}$ into $g \cdot \tilde{f}$, so $x \mapsto g^{-1} \cdot \tilde{f}(g \cdot x)$
- Get a new set $G \cdot f := \{g \cdot \tilde{f} : g \in G\}$ having *orbits* of a given $\tilde{f}$

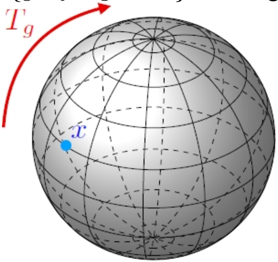



- Note: If all $\tilde{f} \in \tilde{\mathcal{F}}$ are invariant i.e. $G \cdot \tilde{f} = \{\tilde{f}\}$, then,

$$Pf = \mathbb{E}_g \mathbb{E}_Z[f(g \cdot Z)], P_n f = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_g f(g \cdot Z_i)$$

# Initial Setup: Group Case

- Group $G$ acts over $\mathcal{X}$, $\mathcal{Y}$
- ... and transforms any $\tilde{f} \in \tilde{\mathcal{F}}$ into $g \cdot \tilde{f}$, so $x \mapsto g^{-1} \cdot \tilde{f}(g \cdot x)$
- Get a new set $G \cdot f := \{g \cdot \tilde{f} : g \in G\}$ having *orbits* of a given $\tilde{f}$




- Note: If all $\tilde{f} \in \tilde{\mathcal{F}}$ are invariant i.e. $G \cdot \tilde{f} = \{\tilde{f}\}$, then,

$$Pf = \mathbb{E}_g \mathbb{E}_Z[f(g \cdot Z)], P_n f = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_g f(g \cdot Z_i)$$

- $g$ is a random variable, uniformly distributed over compact $G$
- More: Ulf Grenander, *Probabilities on Algebraic Structures* 1967

# Preliminary: Family of Invariant Functions

- We considered:

$$\mathcal{F} := \{f : \tilde{f} \in \tilde{\mathcal{F}}\}, Pf := \mathbb{E}[f(X, Y)], \text{ and } P_n f := \frac{1}{n}\sum_{i=1}^{n} f(X_i, Y_i)$$

- And:

$$Pf = \mathbb{E}_g \mathbb{E}_Z [f(g \cdot Z)], P_n f = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_g f(g \cdot Z_i)$$

# Preliminary: Family of Invariant Functions

- We considered:

$$\mathcal{F} := \{f : \tilde{f} \in \tilde{\mathcal{F}}\}, Pf := \mathbb{E}[f(X,Y)], \text{ and } P_n f := \frac{1}{n}\sum_{i=1}^{n} f(X_i, Y_i)$$

- And:

$$Pf = \mathbb{E}_g \mathbb{E}_Z[f(g \cdot Z)], P_n f = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_g f(g \cdot Z_i)$$

**Lemma**

*Assume $\tilde{\mathcal{F}}$ consists of $G$-equivariant functions, and $\mathcal{F}$ consists of $G$-invariant functions.*

# Preliminary: Family of Invariant Functions

- We considered:

$$\mathcal{F} := \{f : \tilde{f} \in \tilde{\mathcal{F}}\}, Pf := \mathbb{E}[f(X,Y)], \text{ and } P_n f := \frac{1}{n}\sum_{i=1}^{n} f(X_i, Y_i)$$

- And:

$$Pf = \mathbb{E}_g \mathbb{E}_Z [f(g \cdot Z)], P_n f = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_g f(g \cdot Z_i)$$

**Lemma**

*Assume $\tilde{\mathcal{F}}$ consists of $G$-equivariant functions, and $\mathcal{F}$ consists of $G$-invariant functions. Let $D^G := \frac{1}{|G|}\int_G g \cdot D dg$, then generalization errors for $\tilde{\mathcal{F}}$ for $D$ and $D^G$ are the same.*

# Invariant Functions

- PAC style bounds have been studied for invariant functions $\mathcal{F}$ in many previous works
  - Sannai *et al.* (2019); Sokolic *et al.* (2017); Zhu *et al.* (2021)

# Invariant Functions

- PAC style bounds have been studied for invariant functions $\mathcal{F}$ in many previous works
  - Sannai *et al.* (2019); Sokolic *et al.* (2017); Zhu *et al.* (2021)
- Equivalent to concentration bounds.
- Let's focus on bounds roughly of the type:

$$\mathbb{P}\left[\sup_{\mathcal{F}}(P - P_n)f \geq \mathcal{R}(\mathcal{F}_Z) + \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2 n}{2\|\mathcal{F}\|_\infty}\right),$$

# Usual Argument

- Let's focus on bounds roughly of the type:

$$\mathbb{P}\left[\sup_{\mathcal{F}}(P - P_n)f \geq \mathcal{R}(\mathcal{F}_Z) + \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2 n}{2\|\mathcal{F}\|_\infty}\right),$$

# Usual Argument

- Let's focus on bounds roughly of the type:

$$\mathbb{P}\left[\sup_{\mathcal{F}}(P - P_n)f \geq \mathcal{R}(\mathcal{F}_Z) + \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2 n}{2\|\mathcal{F}\|_\infty}\right),$$

  - $\|\mathcal{F}\|_\infty := \sup_{f \in \mathcal{F}} \|f\|_\infty$

# Usual Argument

- Let's focus on bounds roughly of the type:

$$\mathbb{P}\left[\sup_{\mathcal{F}}(P - P_n)f \geq \mathcal{R}(\mathcal{F}_Z) + \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2 n}{2\|\mathcal{F}\|_\infty}\right),$$

  - $\|\mathcal{F}\|_\infty := \sup_{f\in\mathcal{F}} \|f\|_\infty = \sup_{f\in\mathcal{F}} \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |f(x,y)|$

# Usual Argument

- Let's focus on bounds roughly of the type:

$$\mathbb{P}\left[\sup_{\mathcal{F}}(P - P_n)f \geq \mathcal{R}(\mathcal{F}_Z) + \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2 n}{2\|\mathcal{F}\|_\infty}\right),$$

  - $\|\mathcal{F}\|_\infty := \sup_{f \in \mathcal{F}} \|f\|_\infty = \sup_{f \in \mathcal{F}} \sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} |f(x,y)|$
  - ... and the Rademacher complexity:

$$\mathcal{R}(\mathcal{F}_Z) := \mathbb{E}_\sigma \sup_{\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i)$$

# Usual Argument

- Let's focus on bounds roughly of the type:

$$\mathbb{P}\left[\sup_{\mathcal{F}}(P - P_n)f \geq \mathcal{R}(\mathcal{F}_Z) + \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2 n}{2\|\mathcal{F}\|_\infty}\right),$$

- $\|\mathcal{F}\|_\infty := \sup_{f\in\mathcal{F}} \|f\|_\infty = \sup_{f\in\mathcal{F}} \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |f(x,y)|$
- ... and the Rademacher complexity:

$$\mathcal{R}(\mathcal{F}_Z) := \mathbb{E}_\sigma \sup_{\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)$$

- $Z_i = (X_i, Y_i)$, and $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ are the so-called Rademacher random variables, distributed uniformly over $\{-1, 1\}^n$

# Usual Argument Continued..

- Want to bound $\mathcal{R}(\mathcal{F}_Z)$

# Usual Argument Continued..

- Want to bound $\mathcal{R}(\mathcal{F}_Z)$
- One path: Use the Dudley entropy integral

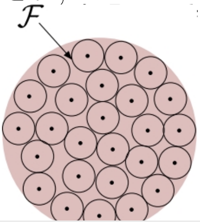
$$\mathcal{R}(\mathcal{F}_Z) \leq \inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{n}}\int_{\alpha}^{\infty}\sqrt{\ln \mathcal{N}(\mathcal{F}, t, \|\cdot\|_\infty)}dt\right)$$

# Usual Argument Continued..

- Want to bound $\mathcal{R}(\mathcal{F}_Z)$
- One path: Use the Dudley entropy integral

$$\mathcal{R}(\mathcal{F}_Z) \leq \inf_{\alpha>0} \left( 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{\infty} \sqrt{\ln \mathcal{N}(\mathcal{F}, t, \|\cdot\|_\infty)} dt \right)$$

- Use this covering number: $\mathcal{N}(\mathcal{F}, t, \|\cdot\|_\infty) :=$
$\min \left\{ k : \exists \{f_1, \ldots, f_k\} \subset \mathcal{F}, \max \; \min \; \|f_j - f\|_\infty \leq t \right\}$



# What happens with $G$-Equivariance?

- Consider the case where we have exact equivariance: $D = g \cdot D$ i.e. the equivariance errors $e_1^{eq}(D) = e_\infty^{eq}(D) = 0$

# What happens with $G$-Equivariance?

- Consider the case where we have exact equivariance: $D = g \cdot D$ i.e. the equivariance errors $e_1^{eq}(D) = e_\infty^{eq}(D) = 0$
- There are two equivalent ideas that can help improve generic bounds of the type discussed

# What happens with $G$-Equivariance? – Idea 1

1. Replace $\mathcal{F}$ by a set of $G$-orbit representatives $\mathcal{F}^0$
   - Probabilities and expectations don't change

# What happens with $G$-Equivariance? – Idea 1

1 Replace $\mathcal{F}$ by a set of $G$-orbit representatives $\mathcal{F}^0$

- Probabilities and expectations don't change
- Need to bound $\mathcal{R}(\mathcal{F}^0_Z)$

# What happens with $G$-Equivariance? – Idea 1

1 Replace $\mathcal{F}$ by a set of $G$-orbit representatives $\mathcal{F}^0$

- Probabilities and expectations don't change
- Need to bound $\mathcal{R}(\mathcal{F}^0_Z)$
- Need to analyze the smaller covering number $\mathcal{N}(\mathcal{F}^0, t, \|\cdot\|_\infty)$

# What happens with $G$-Equivariance? – Idea 2

2 Avoid the issue of choice of orbit representatives by considering orbits traced by our functions.

# What happens with $G$-Equivariance? – Idea 2

2 Avoid the issue of choice of orbit representatives by considering orbits traced by our functions.

- Instead of $f_1, \ldots, f_k$, take their orbits $G \cdot f_1, \ldots, G \cdot f_k$

# What happens with $G$-Equivariance? – Idea 2

2 Avoid the issue of choice of orbit representatives by considering orbits traced by our functions.

- Instead of $f_1, \dots, f_k$, take their orbits $G \cdot f_1, \dots, G \cdot f_k$
- Need to analyze the covering number $\mathcal{N}^G(\mathcal{F}, t, \|\cdot\|_\infty) :=$ $\min\left\{k : \exists\{f_1, \dots, f_k\} \subset \mathcal{F}, \max_{f\in\mathcal{F}} \min_{1\leq j\leq k}\left(\min_{g\in G}\|g\cdot f_j - f\|_\infty\right)\right\}$

# Improvement with $G$-Equivariance

- Improvement on bounds compared to the non-equivariant case is controlled by the following quantity:

$$\frac{\mathcal{N}^G(\mathcal{F}, t, \|\cdot\|_\infty)}{\mathcal{N}(\mathcal{F}, t, \|\cdot\|_\infty)}$$

# Improvement with $G$-Equivariance

- Improvement on bounds compared to the non-equivariant case is controlled by the following quantity:

$$\frac{\mathcal{N}^G(\mathcal{F}, t, \|\cdot\|_\infty)}{\mathcal{N}(\mathcal{F}, t, \|\cdot\|_\infty)}$$

- Rough strategy to bound for equi-Lipschitz functions:
  - Discretize domain and co-domain for all $f$

# Improvement with $G$-Equivariance

- Improvement on bounds compared to the non-equivariant case is controlled by the following quantity:

$$\frac{\mathcal{N}^G(\mathcal{F}, t, \|\cdot\|_\infty)}{\mathcal{N}(\mathcal{F}, t, \|\cdot\|_\infty)}$$

- Rough strategy to bound for equi-Lipschitz functions:
  - Discretize domain and co-domain for all $f$
  - Reduce covering number problem for $\mathcal{X} \times \mathcal{Y}$ to independent problems

# Improvement with $G$-Equivariance

- Improvement on bounds compared to the non-equivariant case is controlled by the following quantity:

$$\frac{\mathcal{N}^G(\mathcal{F}, t, \|\cdot\|_\infty)}{\mathcal{N}(\mathcal{F}, t, \|\cdot\|_\infty)}$$

- Rough strategy to bound for equi-Lipschitz functions:
  - Discretize domain and co-domain for all $f$
  - Reduce covering number problem for $\mathcal{X} \times \mathcal{Y}$ to independent problems
  - Each problem deals with orbit representatives $\mathcal{X}_0$ and covering $G$

# Improvement with $G$-Equivariance

- Improvement on bounds compared to the non-equivariant case is controlled by the following quantity:

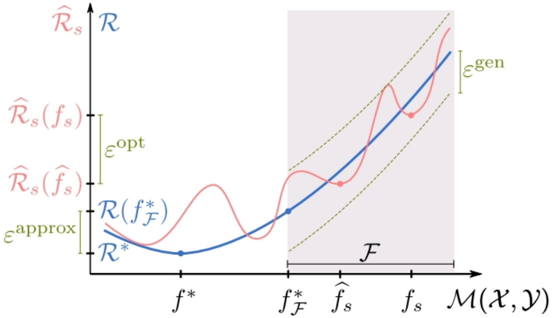
$$\frac{\mathcal{N}^G(\mathcal{F}, t, \|\cdot\|_\infty)}{\mathcal{N}(\mathcal{F}, t, \|\cdot\|_\infty)}$$

- Rough strategy to bound for equi-Lipschitz functions:
  - Discretize domain and co-domain for all $f$
  - Reduce covering number problem for $\mathcal{X} \times \mathcal{Y}$ to independent problems
  - Each problem deals with orbit representatives $\mathcal{X}_0$ and covering $G$
- Has appeared in some works in different contexts (Chen *et al.*, 2023).

# Improvement with $G$-Equivariance

- Can get bounds for $G$-equivariance... but then what?

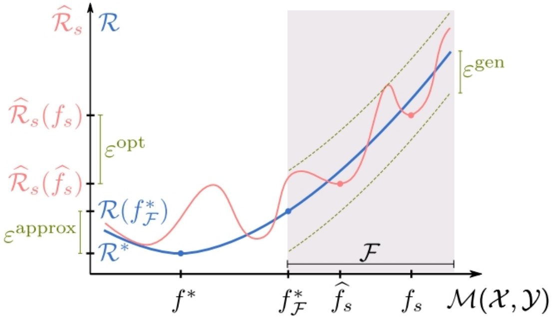
# Only Part of the Story




*Courtesy: J. Berner*

- Blue is the exemplary risk, red is the empirical risk
  - $\mathcal{R}^*$ is the so-called Bayes Risk.
  - Usual picture: Error: $\varepsilon^{\text{approx}} + \varepsilon^{\text{opt}} + \varepsilon^{\text{gen}}$

# Only Part of the Story




*Courtesy: J. Berner*

- Blue is the exemplary risk, red is the empirical risk
  - $\mathcal{R}^*$ is the so-called Bayes Risk.
  - Usual picture: Error: $\varepsilon^{\text{approx}} + \varepsilon^{\text{opt}} + \varepsilon^{\text{gen}}$
  - Have only considered the generalization error!

## Approximation Error

### Model Equivariance versus Data Equivariance

# What about the Approximation Error?

- Suppose we have data with approximate symmetries and model equivariance does not exactly correspond to data equivariance

# What about the Approximation Error?

- Suppose we have data with approximate symmetries and model equivariance does not exactly correspond to data equivariance
- Want to get a quantitative estimate (explicit formulae) of the approximation error lower bound due to the mismatch

# Sketch of Idea:Setup

- Fix $\tilde{\mathcal{F}}$ to be the set of $G$-equivariant functions.

## Sketch of Idea:Setup

- Fix $\tilde{\mathcal{F}}$ to be the set of $G$-equivariant functions.
- Let's first compare to measurable functions $\mathcal{M}(\mathcal{X}, \mathcal{Y})$.

# Sketch of Idea:Setup

- Fix $\tilde{\mathcal{F}}$ to be the set of $G$-equivariant functions.
- Let's first compare to measurable functions $\mathcal{M}(\mathcal{X}, \mathcal{Y})$.
- Assume non-degeneracy on the loss $f(Z) = \ell(\tilde{f}(X), Y)$ and set

$$\mathcal{M}' := \{F(x, y) = \ell(m(x), y), m \in \mathcal{M}\}$$

## Sketch of Idea:Setup

- Fix $\tilde{\mathcal{F}}$ to be the set of $G$-equivariant functions.
- Let's first compare to measurable functions $\mathcal{M}(\mathcal{X}, \mathcal{Y})$.
- Assume non-degeneracy on the loss $f(Z) = \ell(\tilde{f}(X), Y)$ and set

$$\mathcal{M}' := \{F(x, y) = \ell(m(x), y), m \in \mathcal{M}\}$$

- Want to get a quantitative estimate (explicit formulae) of the approximation error lower bound due to the mismatch

# Sketch of Idea: First Steps

- Fix random variables $(X, Y) \backsim D$, denote $\mathcal{X}_0$ as G-orbit representatives in $\mathcal{X}$

$$X = g \cdot \bar{X}, \text{ with } g \in G, \bar{X} \in \mathcal{X}_0$$

# Sketch of Idea: First Steps

- Fix random variables $(X, Y) \backsim D$, denote $\mathcal{X}_0$ as G-orbit representatives in $\mathcal{X}$

$$X = g \cdot \bar{X}, \text{ with } g \in G, \bar{X} \in \mathcal{X}_0$$

- $g$ and $\bar{X}$ are now random variables $\implies$ distributions of $g \cdot \bar{X}, Y \simeq (X, Y)$

# Sketch of Idea: First Steps

- Fix random variables $(X, Y) \backsim D$, denote $\mathcal{X}_0$ as G-orbit representatives in $\mathcal{X}$

$$X = g \cdot \bar{X}, \text{ with } g \in G, \bar{X} \in \mathcal{X}_0$$

- $g$ and $\bar{X}$ are now random variables $\implies$ distributions of $g \cdot \bar{X}, Y \simeq (X, Y)$
- Distributions of the three objects can be obtained by suitable projections:
  - Of $\bar{X}$, denoted $D_{\bar{X}} = \pi_{X_0} \pi_X D$
  - Of $g$, denoted $D_g = D_{g|\bar{X}} D_{\bar{X}}$
  - Of $Y$, denoted $D_Y = D_{Y|g,\bar{X}} D_{g|\bar{X} D_{\bar{X}}}$

## Approximation Error: Data Equivariance

- To get approximation error while working with all measurable $f : \mathcal{X} \to \mathcal{Y}$ set $f(x) = y$, where:

# Approximation Error: Data Equivariance

- To get approximation error while working with all measurable $f : \mathcal{X} \to \mathcal{Y}$ set $f(x) = y$, where:

$$y \in \arg\min \mathbb{E}_{Y|g=g_x, \bar{X}=\bar{x}} \, \ell(y, Y)$$

# Approximation Error: Data Equivariance

- To get approximation error while working with all measurable $f : \mathcal{X} \to \mathcal{Y}$ set $f(x) = y$, where:

$$y \in \arg\min \mathbb{E}_{Y|g=g_x, \bar{X}=\bar{x}} \, \ell(y, Y)$$

- Assumes $x = g_x \bar{x}$ is the unique expression of of $x$ if $g_x \in G$ acts on $\bar{x} \in \mathcal{X}_0$

# Approximation Error: Data Equivariance

- To get approximation error while working with all measurable $f : \mathcal{X} \to \mathcal{Y}$ set $f(x) = y$, where:

$$y \in \arg\min \mathbb{E}_{Y|g=g_x,\bar{X}=\bar{x}}\ \ell(y, Y)$$

  - Assumes $x = g_x\bar{x}$ is the unique expression of of $x$ if $g_x \in G$ acts on $\bar{x} \in \mathcal{X}_0$

- Replace $x$ by $X$, and obtain:

$$AppErr_{equi}(D, \ell) = \mathbb{E}_{\bar{X}} \min_y \mathbb{E}_{g|\bar{X}} \mathbb{E}_{Y|g,\bar{X}} \ell(y, Y)$$

## Approximation Error: Model Equivariance

- Getting approximation error for equivariant measurable functions $f : \mathcal{X} \to \mathcal{Y}$, corresponds to optimizing over $f|_{\mathcal{X}_0}$

# Approximation Error: Model Equivariance

- Getting approximation error for equivariant measurable functions $f : \mathcal{X} \to \mathcal{Y}$, corresponds to optimizing over $f|_{\mathcal{X}_0}$

$$y \in \arg\min \mathbb{E}_{g|\bar{X}=\bar{x}} \mathbb{E}_{Y|g,\bar{X}=\bar{x}} \; \ell(y, Y)$$

# Approximation Error: Model Equivariance

- Getting approximation error for equivariant measurable functions $f : \mathcal{X} \to \mathcal{Y}$, corresponds to optimizing over $f|_{\mathcal{X}_0}$

$$y \in \arg\min \mathbb{E}_{g|\bar{X}=\bar{x}} \mathbb{E}_{Y|g,\bar{X}=\bar{x}} \, \ell(y, Y)$$

- Replace $\bar{x}$ by $\bar{X}$, and we get the error:

# Approximation Error: Model Equivariance

- Getting approximation error for equivariant measurable functions $f : \mathcal{X} \to \mathcal{Y}$, corresponds to optimizing over $f|_{\mathcal{X}_0}$

$$y \in \arg\min \mathbb{E}_{g|\bar{X}=\bar{x}} \mathbb{E}_{Y|g,\bar{X}=\bar{x}} \ \ell(y, Y)$$

- Replace $\bar{x}$ by $\bar{X}$, and we get the error:

$$AppErr(D, \ell) = \mathbb{E}_{\bar{X}} \mathbb{E}_{g|\bar{X}} \min_y \mathbb{E}_{Y|g,\bar{X}} \ \ell(y, Y)$$

# Approximation Gap: Model versus Data Equivariance

- The error terms that we obtained:

$$AppErr(D, \ell) = \mathbb{E}_{\bar{X}} \mathbb{E}_{g|\bar{X}} \min_{y} \mathbb{E}_{Y|g,\bar{X}} \, \ell(y, Y)$$

# Approximation Gap: Model versus Data Equivariance

- The error terms that we obtained:

$$AppErr(D, \ell) = \mathbb{E}_{\bar{X}} \mathbb{E}_{g|\bar{X}} \min_{y} \mathbb{E}_{Y|g,\bar{X}} \, \ell(y, Y)$$

$$AppErr_{equi}(D, \ell) = \mathbb{E}_{\bar{X}} \min_{y} \mathbb{E}_{g|\bar{X}} \mathbb{E}_{Y|g,\bar{X}} \ell(y, Y)$$

# Approximation Gap: Model versus Data Equivariance

- The error terms that we obtained:

$$AppErr(D,\ell) = \mathbb{E}_{\bar{X}}\mathbb{E}_{g|\bar{X}} \min_{y} \mathbb{E}_{Y|g,\bar{X}}\ \ell(y,Y)$$

$$AppErr_{equi}(D,\ell) = \mathbb{E}_{\bar{X}} \min_{y} \mathbb{E}_{g|\bar{X}}\mathbb{E}_{Y|g,\bar{X}}\ell(y,Y)$$

- Clearly: $AppErr(D,\ell) \leq AppErr_{equi}(D,\ell)$

# Approximation Gap: Model versus Data Equivariance

- The error terms that we obtained:

$$AppErr(D, \ell) = \mathbb{E}_{\bar{X}} \mathbb{E}_{g|\bar{X}} \min_{y} \mathbb{E}_{Y|g,\bar{X}} \; \ell(y, Y)$$

$$AppErr_{equi}(D, \ell) = \mathbb{E}_{\bar{X}} \min_{y} \mathbb{E}_{g|\bar{X}} \mathbb{E}_{Y|g,\bar{X}} \ell(y, Y)$$

- Clearly: $AppErr(D, \ell) \leq AppErr_{equi}(D, \ell)$
- How much is the gap?

# Approximation Gap: Model versus Data Equivariance

- We had $AppErr(D, \ell) \leq AppErr_{equi}(D, \ell)$

# Approximation Gap: Model versus Data Equivariance

- We had $AppErr(D, \ell) \leq AppErr_{equi}(D, \ell)$
- Leading to the following expression for the approximation gap:

$$AppGap(D, \ell) = \mathbb{E}_{\bar{X}} \min_{y \in \mathcal{Y}} \mathbb{E}_{g|\bar{X}} \left( L_{\bar{X}}(g, y) - L_{\bar{X}}(g, y_g^*) \right)$$

# Approximation Gap: Model versus Data Equivariance

- We had $AppErr(D, \ell) \leq AppErr_{equi}(D, \ell)$
- Leading to the following expression for the approximation gap:

$$AppGap(D, \ell) = \mathbb{E}_{\bar{X}} \min_{y \in \mathcal{Y}} \mathbb{E}_{g|\bar{X}} \left( L_{\bar{X}}(g, y) - L_{\bar{X}}(g, y_g^*) \right)$$

- Where $L_{\bar{X}}(g, y) := \mathbb{E}_{Y|g,\bar{X}} \ell(y, Y)$

# Approximation Gap: Model versus Data Equivariance

- We had $AppErr(D, \ell) \leq AppErr_{equi}(D, \ell)$
- Leading to the following expression for the approximation gap:

$$AppGap(D, \ell) = \mathbb{E}_{\bar{X}} \min_{y \in \mathcal{Y}} \mathbb{E}_{g|\bar{X}} \left( L_{\bar{X}}(g, y) - L_{\bar{X}}(g, y_g^*) \right)$$

- Where $L_{\bar{X}}(g, y) := \mathbb{E}_{Y|g,\bar{X}} \ell(y, Y)$
- .. and $y^* \in \arg\min_{y} L_{\bar{X}}(g, y)$

# Approximation Gap

- With some technical conditions on the loss, we will have (simplified):

# Approximation Gap

- With some technical conditions on the loss, we will have (simplified):

$$AppGap(D, C_L\| \cdot \|^2) \geq C_L \mathbb{E}_{\bar{X}} \text{ Var }_{g|\bar{X}} \left[ y_g^* \right]$$

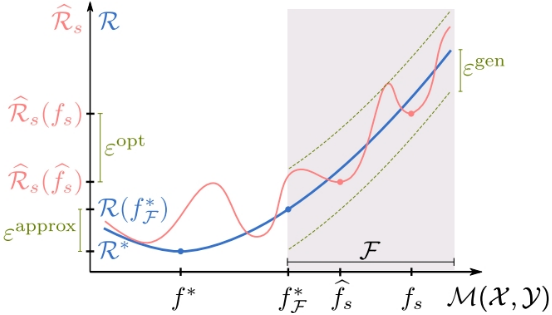
# Approximation Gap

- With some technical conditions on the loss, we will have (simplified):

$$AppGap(D, C_L \| \cdot \|^2) \geq C_L \mathbb{E}_{\bar{X}} \text{ Var }_{g|\bar{X}} \left[ y_g^* \right]$$

- Can obtain explicit formulae in many cases, also reasonably easy to compute

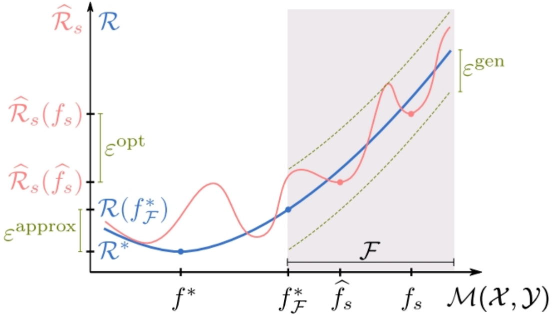
# Approximation Gap




*Courtesy: Wang, Walters, and Yu*

- The *sweet spot* for better generalization occurs when the generalization and approximation error plots cross.

# Approximation Gap

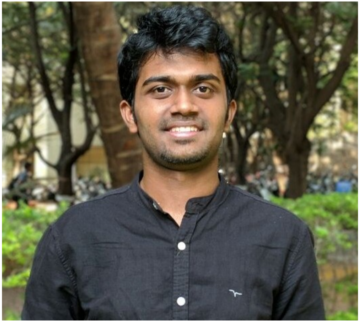



*Courtesy: Wang, Walters, and Yu*

- The *sweet spot* for better generalization occurs when the generalization and approximation error plots cross.
- Note: Skipped handling optimization error. But the picture remains similar with added caveats.

Learning Partial Symmetries

Exploiting Displacement Structure



Ashwin Samudre
U. British Columbia

Brian D. Nord
MIT Physics/Fermilab