Provably Strict Generalisation Benefit for Equivariant Models

Bryn Elesedy and Sheheryar Zaidi
University of Oxford

ICML 2021

What this paper is about

Background:

- Significant interest in symmetry in machine learning
- Improved generalisation is observed in practice
- Existing (worst-case) theoretical results do not show this

Contribution:

- Framework for analysing equivariant models and exact calculation of generalisation improvement

Notation

Input space $\mathcal{X}$, output space $\mathcal{Y} = \mathbb{R}^k$ with inner product $\langle \cdot, \cdot \rangle$

Compact group $\mathcal{G}$ with action ϕ on $\mathcal{X}$ and orthogonal representation ψ on $\mathcal{Y}$

Averaging operator for equivariance

$$\mathcal{Q}f(x) = \int_{\mathcal{G}} \psi(g^{-1}) f(\phi(g)x) \, \mathrm{d}\lambda(g)$$

where λ is the Haar measure on $\mathcal{G}$

Setting

Let μ be a $\mathcal{G}$-invariant distribution on $\mathcal{X}$

Consider

$$V = L^2(\mathcal{X}, \mu; \mathcal{Y})$$

which is the vector space of functions $f : \mathcal{X} \to \mathcal{Y}$ with inner product

$$\langle f_1, f_2 \rangle_\mu = \int_{\mathcal{X}} \langle f_1(x), f_2(x) \rangle \, \mathrm{d}\mu(x)$$

and norm $\|f\|_\mu = \sqrt{\langle f, f \rangle_\mu} < \infty$

Central Observations

Properties of $\mathcal{Q}$

1. Identification: $\mathcal{Q}f = f \iff f$ is $\mathcal{G}$-equivariant
2. Projection: $\mathcal{Q}$ is a projection
3. Decomposition: $f = \bar{f} + f^{\perp}$ where $\mathcal{Q}\bar{f} = \bar{f}$ and $\mathcal{Q}f^{\perp} = 0$
4. Self-Adjoint: $\langle \mathcal{Q}f_1, f_2 \rangle_\mu = \langle f_1, \mathcal{Q}f_2 \rangle_\mu$

Conclusion: *orthogonal decomposition*

$$V = S \oplus A$$

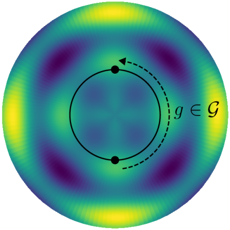
where $S = \{f \in V : f \text{ is } \mathcal{G}\text{-equivariant}\}$ and $A = \mathrm{null}(\mathcal{Q})$

Structure of Function Spaces: Example

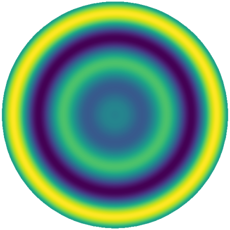
$X \sim \mathcal{N}(0, I_2)$ and $\mathcal{G} = \mathsf{SO}(2)$

$V = \{f : \mathbb{R}^2 \to \mathbb{R} \text{ with } \mathbb{E}[f(X)^2] < \infty\}$

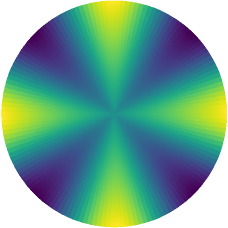
A picture for $f(r, \theta) = r \cos(r - 2\theta) \cos(r + 2\theta)$



$f \in V$



$\bar{f} \in S$



$f^{\perp} \in A$

Generalisation Benefit of Equivariance

Goal: Compare any predictor f to its equivariant version $\bar{f} = \mathcal{Q}f$

Setup:

- Task: $X \sim \mu$, $Y = f^*(X) + \xi$ with $\mathbb{E}[\xi] = 0$ and $\xi \perp\!\!\!\perp X$
- Equivariant target: $f^*(X) = \mathbb{E}[Y|X]$ is $\mathcal{G}$-equivariant

Result: Recall $f = \bar{f} + f^\perp$, the *generalisation gap* satisfies

$$\Delta(f, \bar{f}) \coloneqq \mathbb{E}[(f(X) - Y)^2] - \mathbb{E}[(\bar{f}(X) - Y)^2] = \|f^\perp\|_\mu^2$$

This is *strictly positive* if f is not equivariant!

Theorem: The Linear Case

Orthogonal representations ϕ on $\mathcal{X} = \mathbb{R}^d$ and ψ on $\mathcal{Y} = \mathbb{R}^k$

$X \sim \mathcal{N}(0, I)$ and $Y = h_\Theta(X) + \xi$ where $h_\Theta(x) = \Theta^\top x$ is equivariant and $\mathbb{E}[\xi] = 0$, $\mathrm{Cov}[\xi] = I$, $\xi \perp\!\!\!\perp X$

For a linear predictor f fit by least-squares on n i.i.d. examples:

- $n > d + 1$:
$$\mathbb{E}[\Delta(f, \bar{f})] = \frac{dk - (\chi_\phi | \chi_\psi)}{n - d - 1}$$
- $n \in [d-1, d+1]$: $\mathbb{E}[\Delta(f, \bar{f})] = \infty$
- $n < d - 1$:
$$\mathbb{E}[\Delta(f, \bar{f})] = \frac{n(dk - (\chi_\phi | \chi_\psi))}{d(d - n - 1)} + \mathcal{E}_\mathcal{G}(n, d)$$

The End

More in the paper: feature averaging, ideas for training NNs...

Bryn Elesedy

Sheheryar Zaidi