Equivariance Discovery by Learned Parameter-Sharing

Raymond A. Yeh† Yuan-Ting Hu Mark Hasegawa-Johnson Alexander G. Schwing

Toyota Technological Institute at Chicago† University of Illinois at Urbana-Champaign

Introduction

Motivation:

- Equivariance is an effective inductive bias
- Convolution networks achieve shift equivariance through parameter-sharing
- Can we discover equivariance from data?

Contribution:

- Equivariance discovery by learning how to share model parameters from data
- Analysis on the benefit of learning a parameter-sharing scheme
- Empirical results on recovering shift and permutation equivariance

Running Example

Convolution:

$$\mathbf{y}[k] = \sum_j \mathbf{x}[j+k]\theta[j]$$

and let $\mathbf{x} \in \mathbb{R}^3$ and $\mathbf{k} = [2,1]$, we equivalently write convolution as $\mathbf{y} = \mathbf{K}\mathbf{x}$, where

$$\mathbf{K} = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix}$$

Approach

Parameterizing Parameter-Sharing:

$$\text{Flatten}(\mathbf{K}) = \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \\ 2 \\ 1 \end{bmatrix} \quad \mathbf{A} = \begin{bmatrix} 0&0&1&0&0&0 \\ 0&1&0&0&0&0 \\ 1&0&0&0&0&0 \\ 1&0&0&0&0&0 \\ 0&0&1&0&0&0 \\ 0&1&0&0&0&0 \end{bmatrix} \quad \boldsymbol{\psi} = \begin{bmatrix} 0 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

- A selection matrix $\boldsymbol{A}$ to represent a sharing scheme
- $\boldsymbol{A}_{ij} \in \{0,1\}, \sum_j \boldsymbol{A}_{ij} = 1$
- $\boldsymbol{A}$ selects the model parameters from ψ

Learning parameter-sharing as a bi-level optimization:

$$\overbrace{\min_{\boldsymbol{A}} \mathcal{L}(\underbrace{\boldsymbol{A}\boldsymbol{\psi}^*(\boldsymbol{A})}_{\boldsymbol{\theta}}, \mathcal{V})}^{\text{upper-level task}} \text{ s.t. } \boldsymbol{\psi}^*(\boldsymbol{A}) = \overbrace{\arg\min_{\boldsymbol{\psi}} \mathcal{L}(\boldsymbol{A}\boldsymbol{\psi}, \mathcal{T})}^{\text{lower-level task}},$$

$$\boldsymbol{A} \in \{0,1\}^{K\times K}, \sum_j \boldsymbol{A}_{ij} = 1 \;\; \forall i.$$

Practical Considerations:

- Relax $\boldsymbol{A} \in [0,1]^{K\times K}$ and optimize using gradient based methods
- Penalty term 1: Entropy $H(\boldsymbol{A})$ encourages $\boldsymbol{A}_{ij}$ to be closer to 0 or 1
- Penalty term 2: Nuclear norm $\|\boldsymbol{A}\|_*$ encourages $\boldsymbol{A}$ to be low-rank

Analysis on Gaussian Data:

$$\mathbf{y} \sim \mathcal{N}(\boldsymbol{A}_{\text{gt}}\boldsymbol{\psi}_{\text{gt}}, \sigma^2\boldsymbol{I}). \tag{1}$$

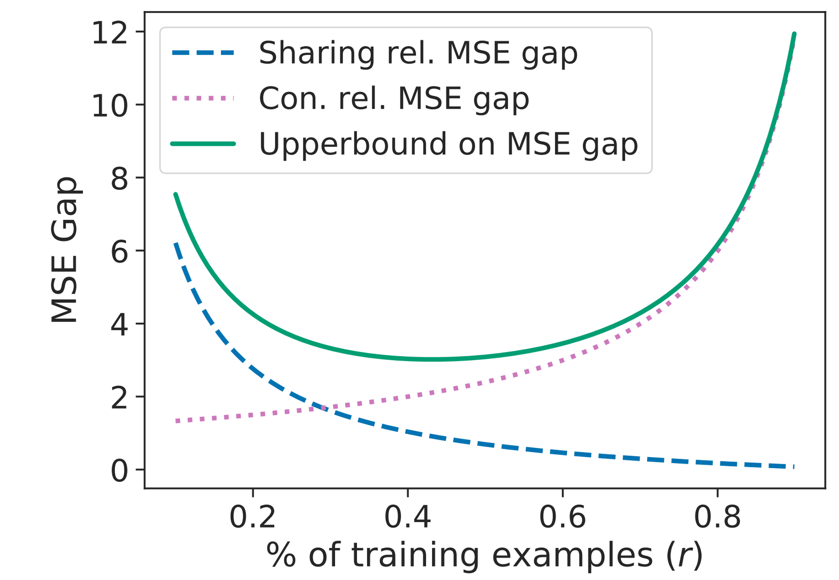
- *i.i.d.* Gaussian with shared means across dimensions
- $\boldsymbol{\theta}_{\text{gt}} = \boldsymbol{A}_{\text{gt}}\boldsymbol{\psi}_{\text{gt}}$: Ground-truth mean
- $\hat{\boldsymbol{\theta}}_{\texttt{val}} = \boldsymbol{A}_{\texttt{val}}\hat{\boldsymbol{\psi}}$: Estimated $\boldsymbol{\psi}$ with $\boldsymbol{A}_{\texttt{val}}$ from our approach
- $\hat{\boldsymbol{\theta}}_{\text{gt}} = \boldsymbol{A}_{\text{gt}}\hat{\boldsymbol{\psi}}$: Estimated $\boldsymbol{\psi}$ with $\boldsymbol{A}_{\text{gt}}$
- MSE Gap: $\text{MSE}(\hat{\boldsymbol{\theta}}_{\texttt{val}}) - \text{MSE}(\hat{\boldsymbol{\theta}}_{\text{gt}})$

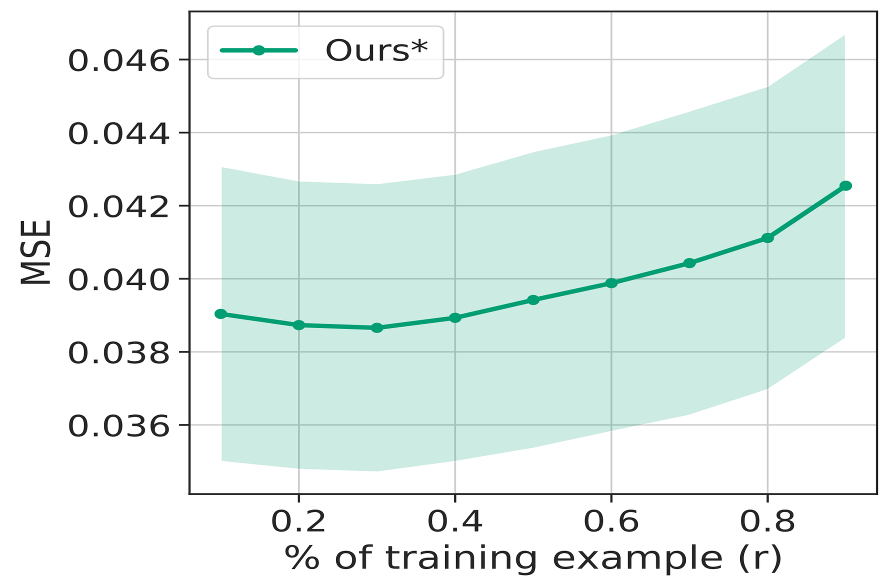
Results

Claim: Given data following Eq. (1), with probability $1-\alpha$ and $\alpha < \exp\frac{-K}{10}$, the MSE gap is upper bounded by

$$\sigma^2\left(\underbrace{\frac{1-r}{r|\mathcal{D}|}(\text{rk}(\boldsymbol{A}_{\text{gt}})-1)}_{\text{sharing rel. MSE gap}} - \underbrace{\frac{40\ln(\alpha)}{(1-r)|\mathcal{D}|}}_{\text{con. rel. MSE gap}}\right),$$

where $r = \frac{|\mathcal{T}|}{|\mathcal{D}|}$ denotes the ratio between the size of training and overall dataset.

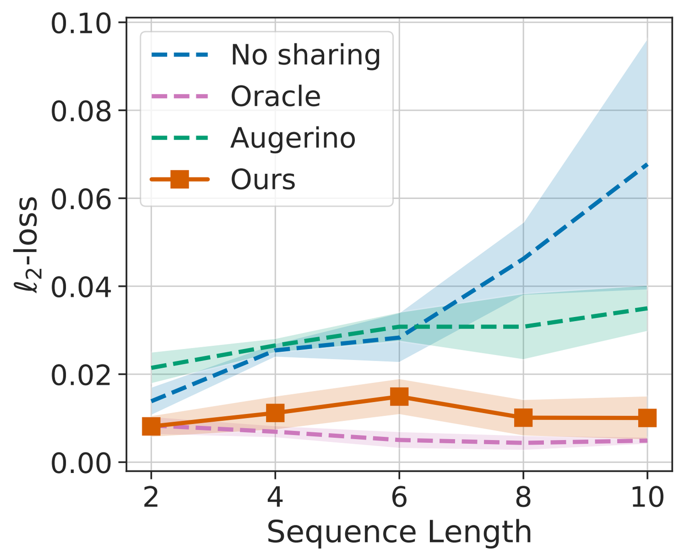


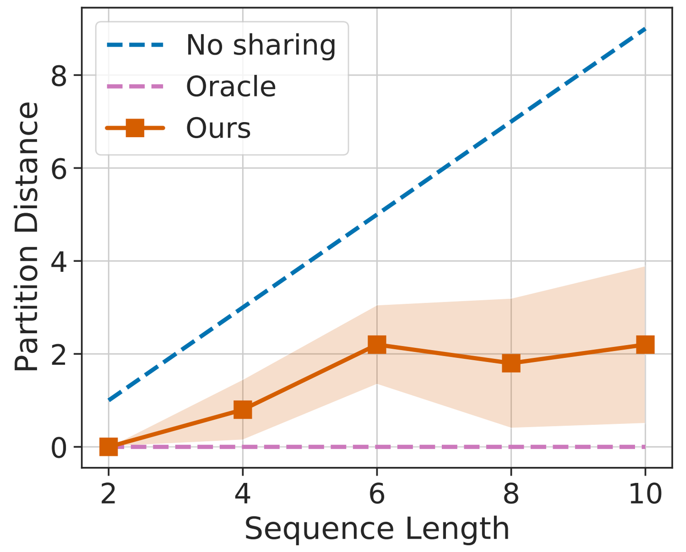


Experiments (Sum of Numbers):

The task is to regress to the sum of a sequence of numbers provided in text format.

- *E.g.*, given the input (âĂIJone,âĂİ âĂIJfiveâĂİ) the model should output 6
- ℓ_2-loss between the predicted an ground-truth
- Partition distance between the recovered $\boldsymbol{A}_{\texttt{val}}$ and $\boldsymbol{A}_{\text{gt}}$
- Partition distance measures the number of assignments that must be changed for one sharing scheme to be identical to the other.





Please find additional experiments in the paper