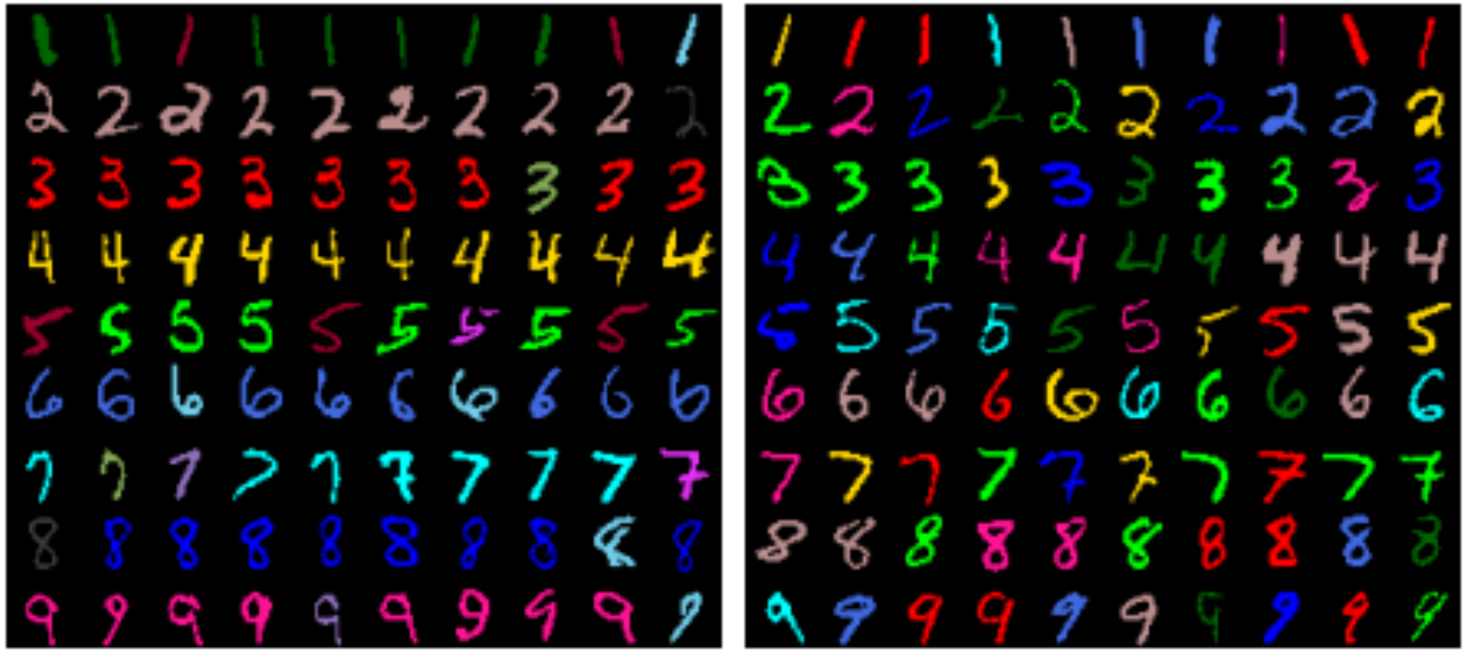
Intro of Out-of-distribution Generalization

Dinghuai Zhang

What’s “spurious” correlation?

“true label” and “spurious label”

What’s “spurious” correlation?



“true label” and “spurious label”

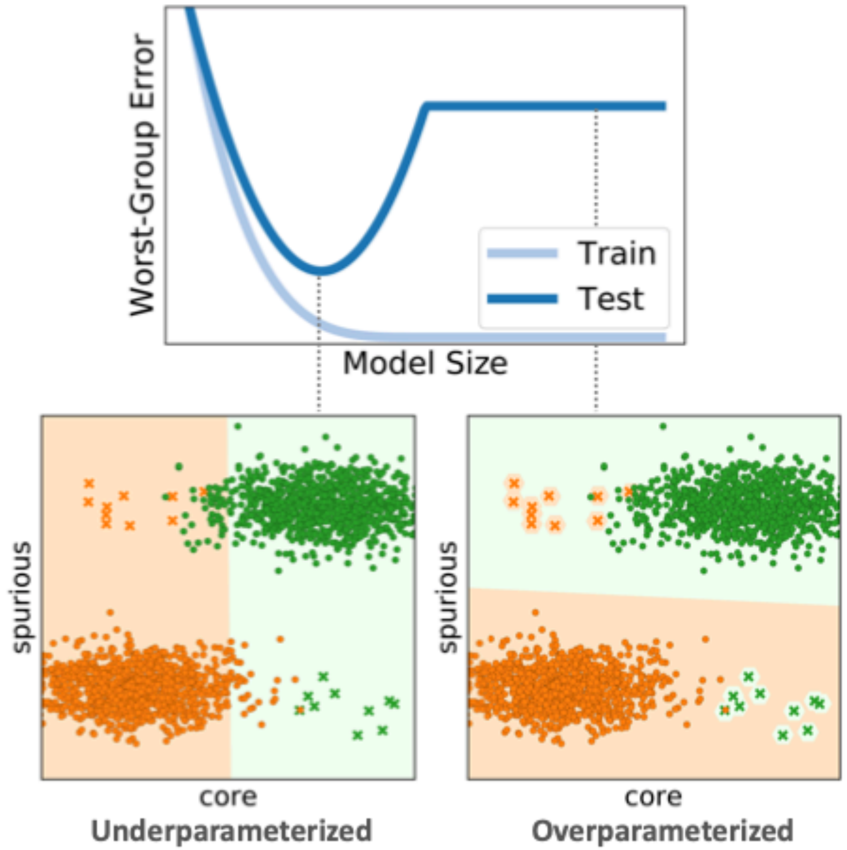
GroupDRO

$$\hat{\theta}_{\mathrm{DRO}} := \underset{\theta \in \Theta}{\arg\min} \left\{ \hat{\mathcal{R}}(\theta) := \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y) \sim \hat{P}_g} [\ell(\theta; (x, y))] \right\},$$

DISTRIBUTIONALLY ROBUST NEURAL NETWORKS FOR GROUP SHIFTS: ON THE IMPORTANCE OF REGULARIZATION FOR WORST-CASE GENERALIZATION, Shiori Sagawa et al. ICLR2020

			Average Accuracy		Worst-Group Accuracy	
			ERM	DRO	ERM	DRO
Standard Regularization	Waterbirds	Train	100.0	100.0	100.0	100.0
		Test	97.3	97.4	60.0	76.9
	CelebA	Train	100.0	100.0	99.9	100.0
		Test	94.8	94.7	41.1	41.1
	MultiNLI	Train	99.9	99.3	99.9	99.0
		Test	82.5	82.0	65.7	66.4
Strong ℓ_2 Penalty	Waterbirds	Train	97.6	99.1	35.7	97.5
		Test	95.7	96.6	21.3	84.6
	CelebA	Train	95.7	95.0	40.4	93.4
		Test	95.8	93.5	37.8	86.7

Overparameterization exacerbates spurious correlations

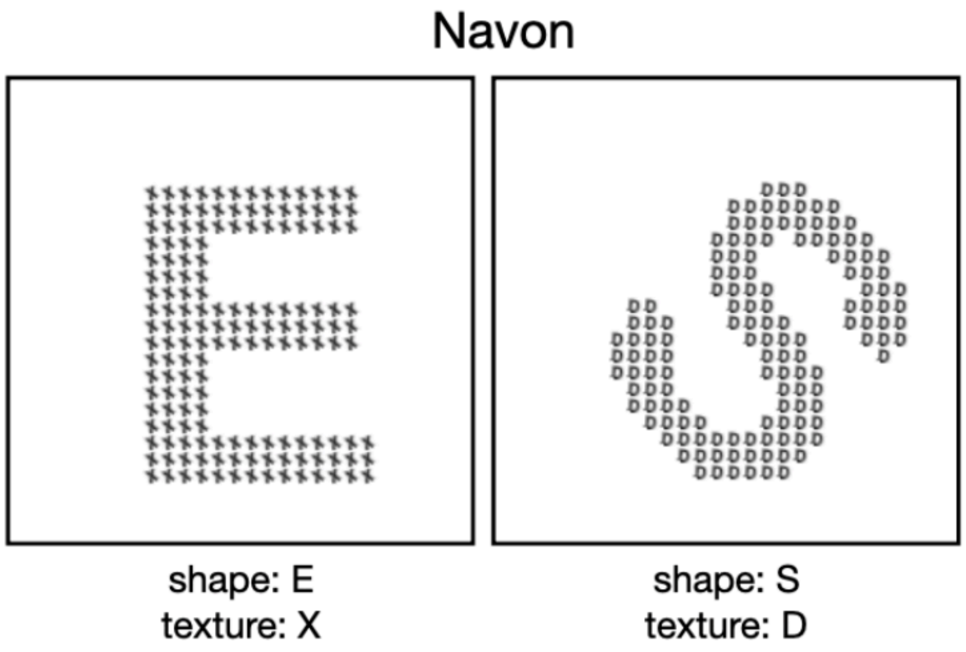


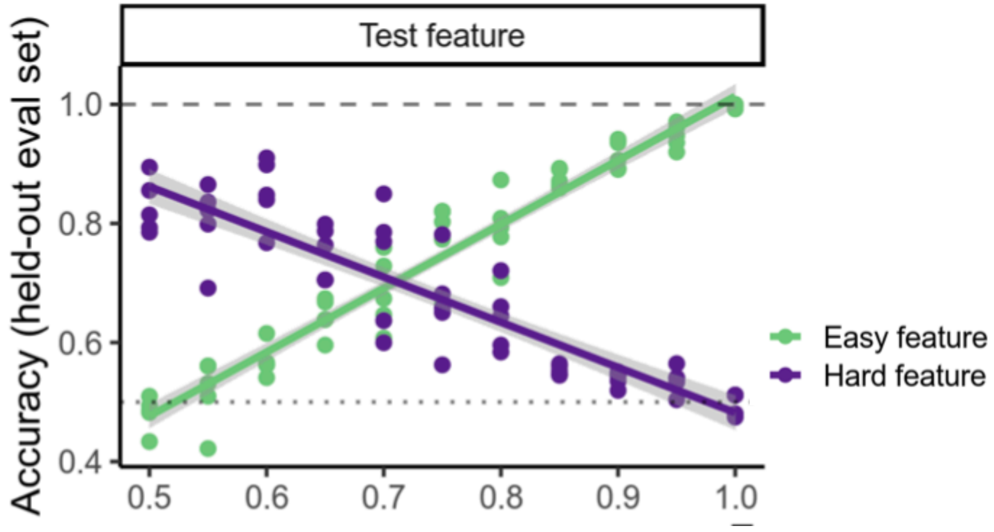
As model size grows, avg errors decrease, but worst group error increases

Reason: overparametrized models use spurious feature to classify

An investigation of why overparameterization exacerbates spurious correlations, Shiori Sagawa et al. ICML2020

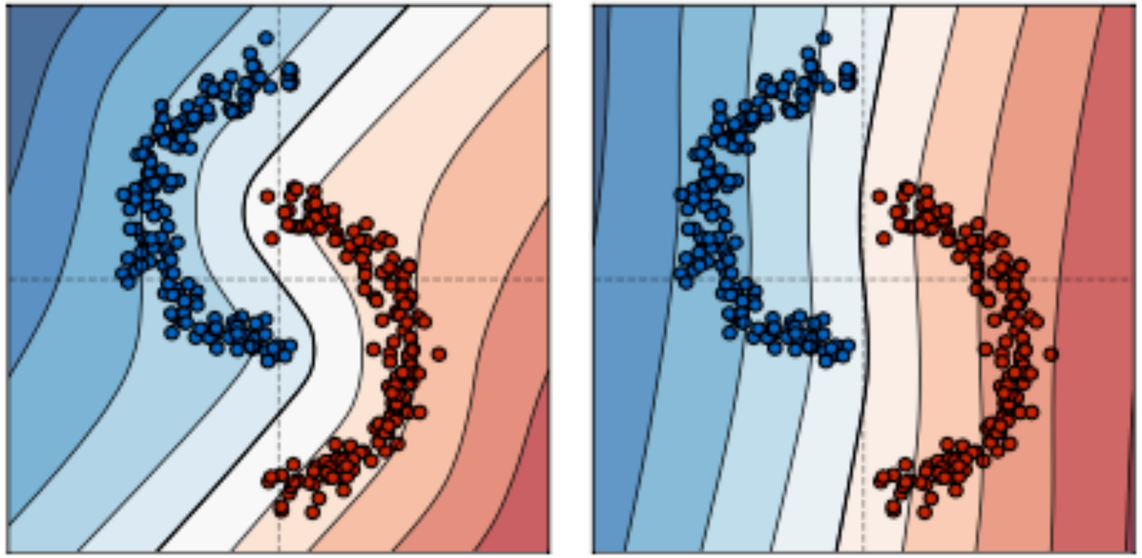
Another name: Simplicity Bias





What shapes feature representations? Exploring datasets, architectures, and training. NeurIPS2020.

Gradient Starvation

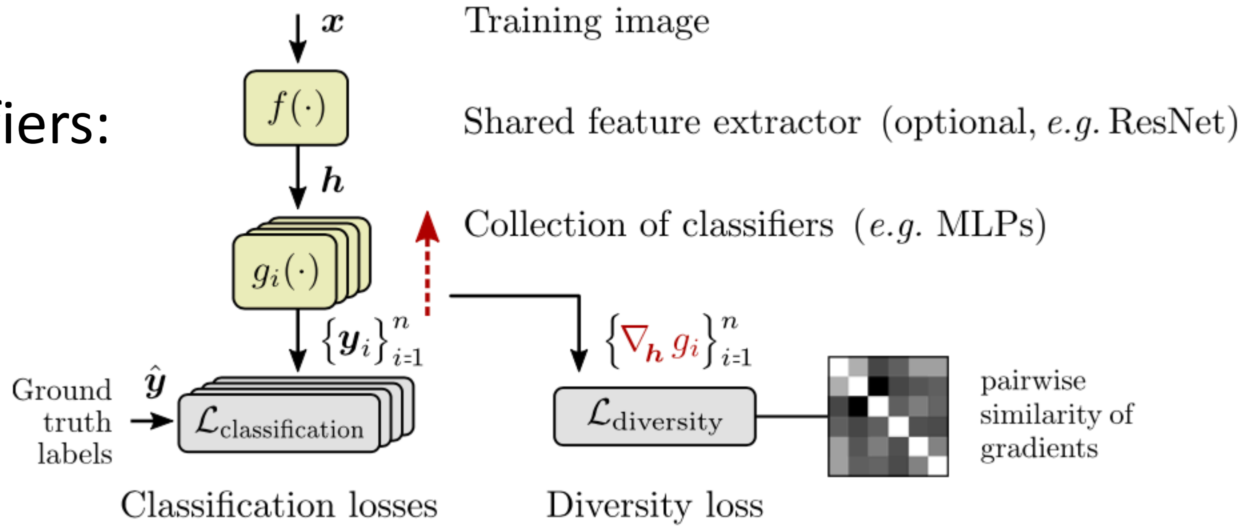


"overfitting" property of ERM

Gradient Starvation: A Learning Proclivity in Neural Networks Mohammad Pezeshki et al.

Solution: Promote Diversity

Train multiple classifiers:



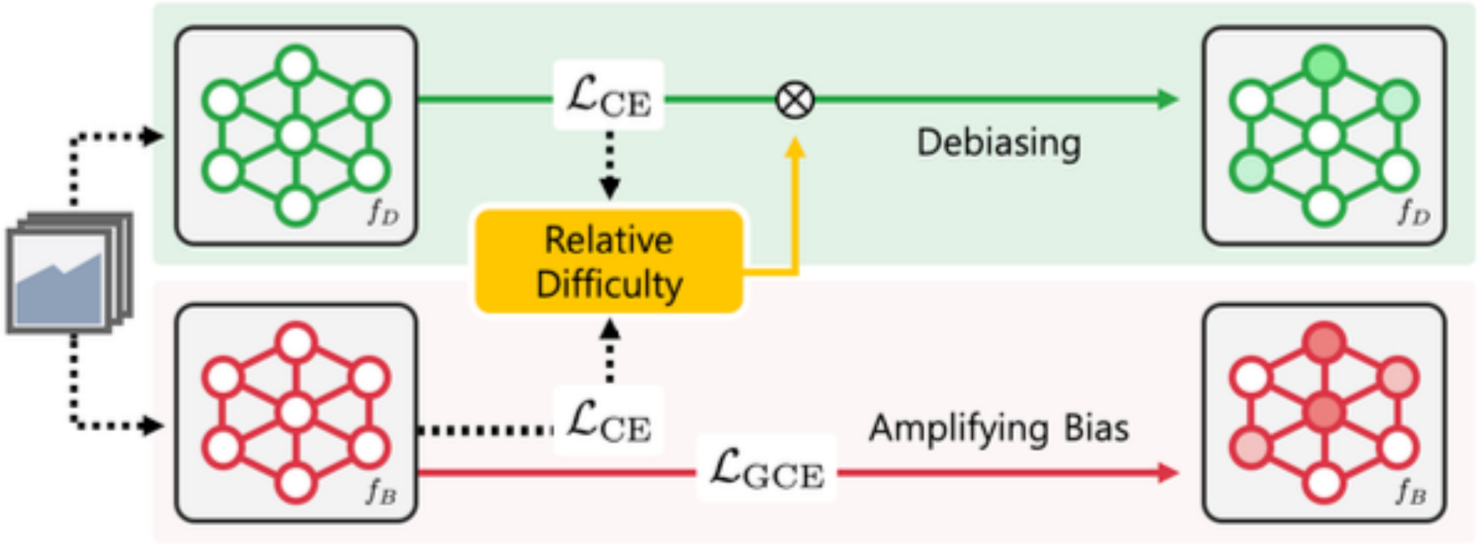
$$\min_{(\boldsymbol{\theta},\{\boldsymbol{\phi}_i\})} \Sigma_i^n \, \mathcal{R}(g_{\boldsymbol{\phi}_i} \circ f_{\boldsymbol{\theta}}) + \lambda \, \Sigma_{i \neq j} \Sigma_k^K \, \delta_{g_{\phi i}, g_{\phi j}}(\boldsymbol{h}^k) \qquad \delta_{g_{\phi_1}, g_{\phi_2}}(\boldsymbol{h}) = \nabla_{\boldsymbol{h}} \, g^\star_{\boldsymbol{\phi}_1}(\boldsymbol{h}) \,.\, \nabla_{\boldsymbol{h}} \, g^\star_{\boldsymbol{\phi}_2}(\boldsymbol{h})$$

With post model selection method

Evading the Simplicity Bias: Training a Diverse Set of Models Discovers Solutions with Superior OOD Generalization

Learning from Failure

- Setting: eg. No multiple domains
- 99% data: label & color has 1 to 1 corresponding
- 1% data: label & color has no corresponding

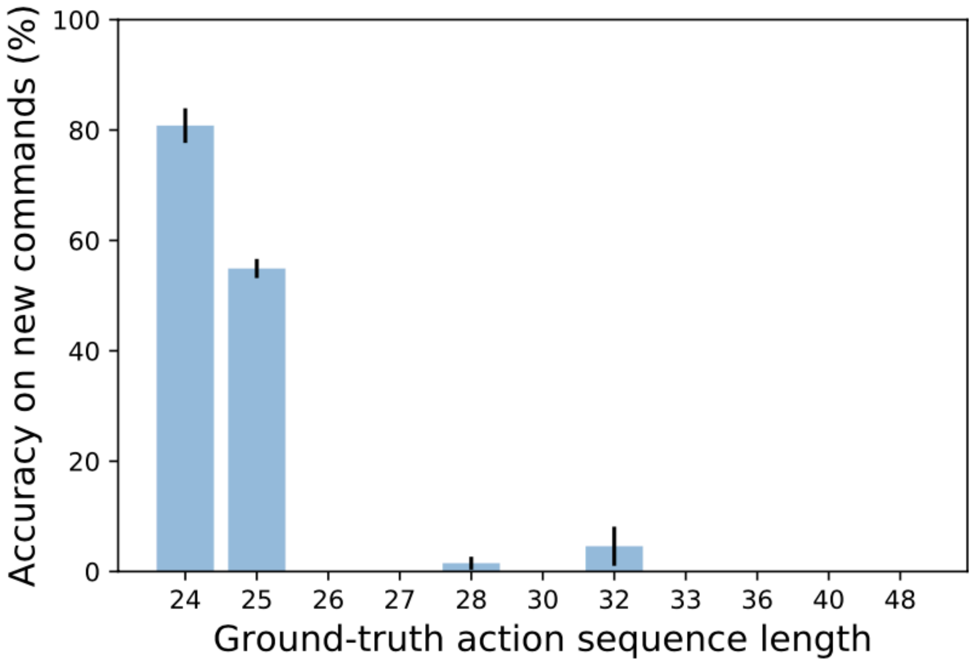


Learning from Failure: Training Debiased Classifier from Biased Classifier Junhyun Nam et al., NeurIPS2020

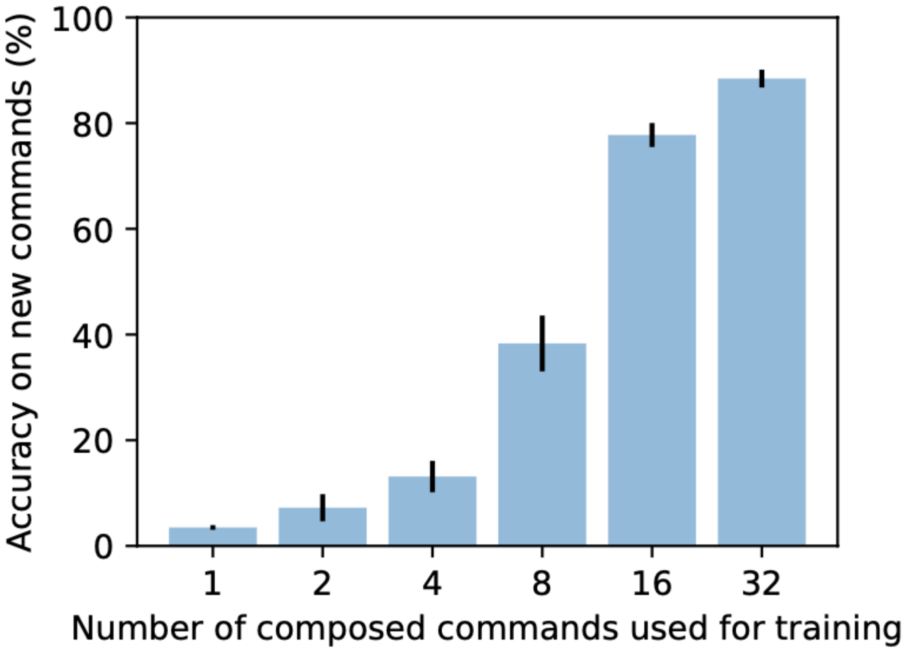
NLP & CogSci: Compositionality

- How RNN generalize systematically under distribution shift

Different sequence length



Different word combination methods



Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. ICML2018

NLP: Debias

MNLI synthetic:

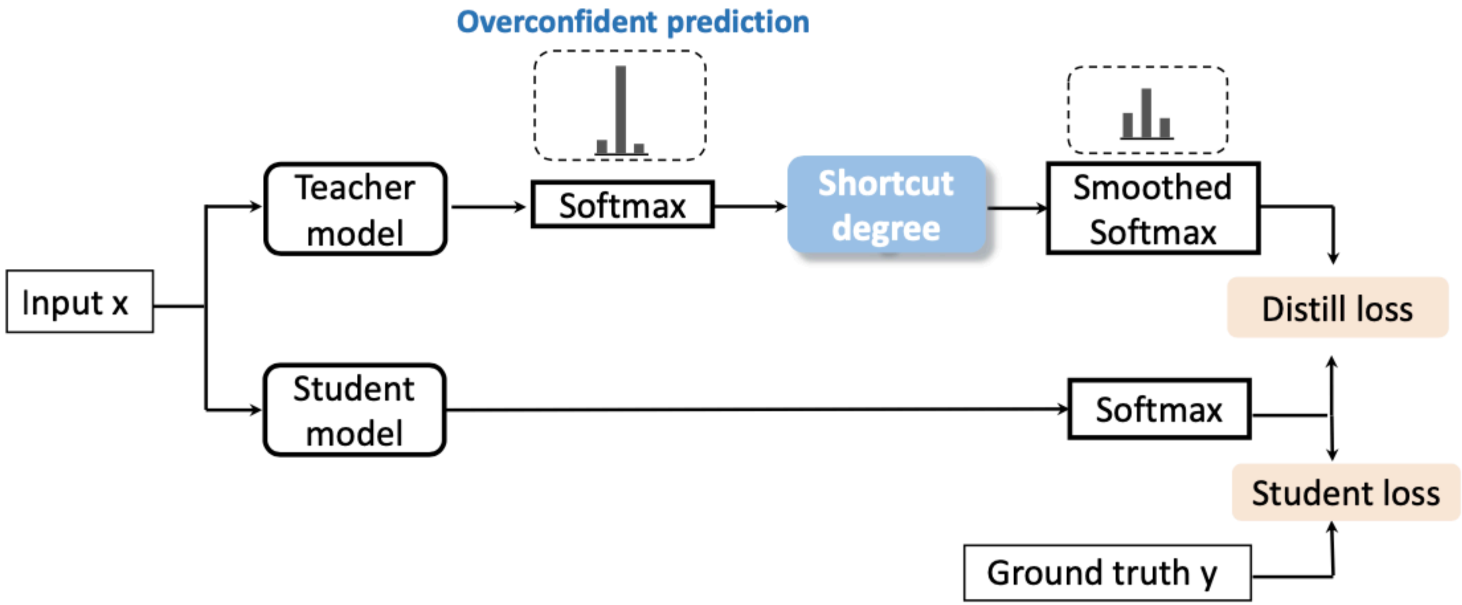
premise: What's truly striking, though, is that Jobs has never really let this idea go.

orig. hypo.: Jobs never held onto an idea for long.

biased: 0 Jobs never held onto an idea for long.

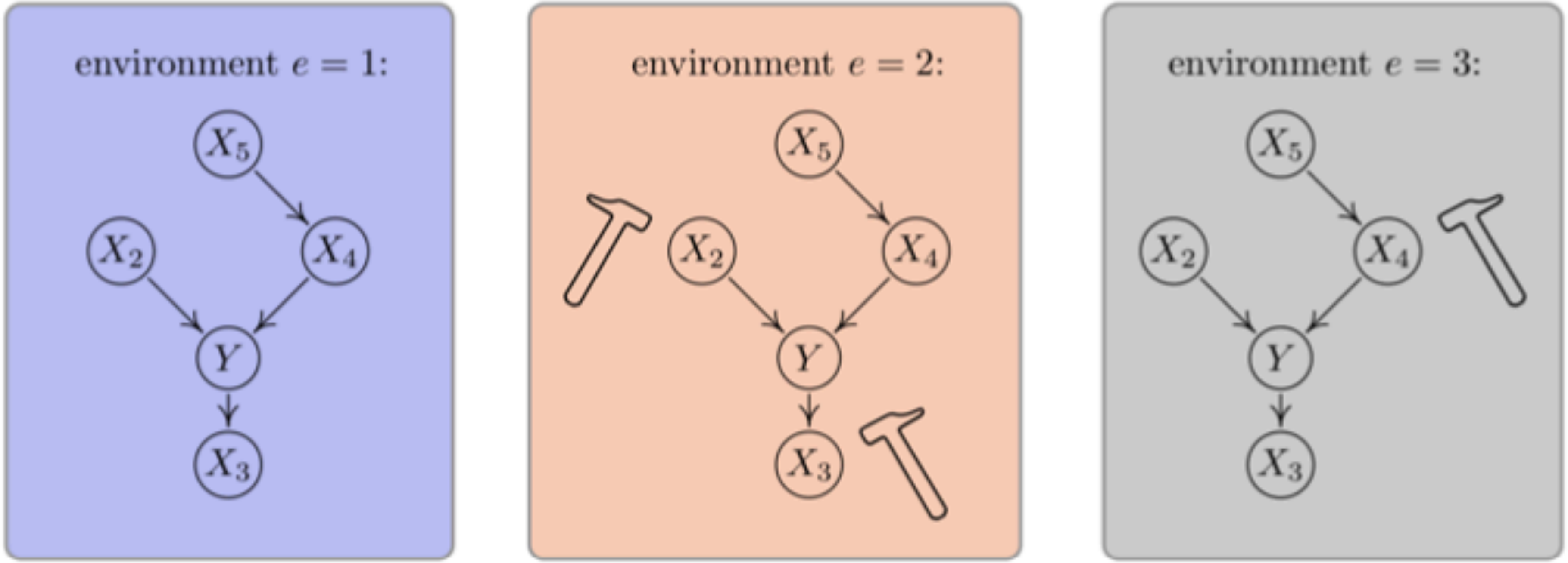
anti-biased: 1 Jobs never held onto an idea for long.

label: 0 (contradiction)



Towards Interpreting and Mitigating Shortcut Learning Behavior of NLU models, NAACL2021

How to get rid of "spurious" feature?
Or, how to do invariant learning



Causal inference using invariant prediction: identification and confidence intervals. Jonas Peters et al. JRSSB

Invariant Causal Prediction (ICP)

Assumption 1 (Invariant prediction) *There exists a vector of coefficients* $\gamma^* = (\gamma_1^*, \ldots, \gamma_p^*)^t$ *with support* $S^* := \{k : \gamma_k^* \neq 0\} \subseteq \{1, \ldots, p\}$ *that satisfies*

$$\begin{aligned} \textit{for all } e \in \mathcal{E}: \quad & X^e \textit{ has an arbitrary distribution and} \\ & Y^e = \mu + X^e \gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \textit{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}, \end{aligned} \tag{3}$$

where $\mu \in \mathbb{R}$ *is an intercept term,* ε^e *is random noise with mean zero, finite variance and the same distribution* F_ε *across all* $e \in \mathcal{E}$.

We will interchangeably use “domain” and “environment”.

Causal Transfer Learning

Algorithm 1: Subset search

Inputs: Sample $(\mathbf{x}_i^k, y_i^k)_{i=1}^{n_k}$ for tasks $k \in \{1, \ldots, D\}$, threshold δ for independence test.

Outputs: Estimated invariant subset $\hat{S}$.

```
Set S_acc = {}, MSE = {}.
for S ⊆ {1, ..., p} do
    linearly regress Y on X_S and compute the residuals R_{β^{CS(S)}} on a validation set.
    compute H = HSIC_b((R_{β^{CS(S)},i}, K_i)_{i=1}^n) and the corresponding p-value p* (or
      the p-value from an alternative test, e.g., Levene test.).
    if p* > δ then
        compute Ê_{P1,...,D}(β^{CS(S)}), the empirical estimate of E_{P1,...,D}(β^{CS(S)}) on a
          validation set.
        S_acc.add(S), MSE.add(Ê_{P1,...,D}(β^{CS(S)}))
    end
end
Select Ŝ according to RULE, see Section 3.4.
```

Invariant Models for Causal Transfer Learning Mateo Rojas-Carulla et al

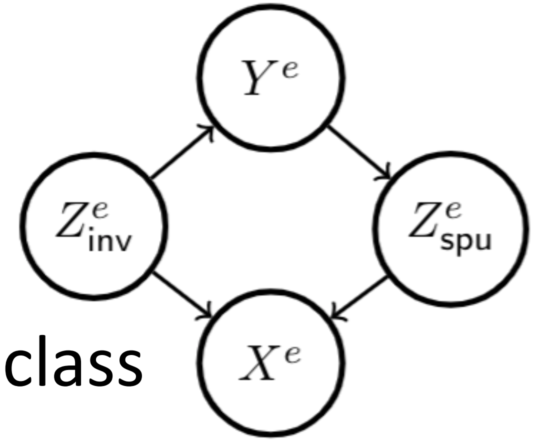
Invariant Risk Minimization

$$\begin{aligned} \min_{\substack{\Phi:\mathcal{X}\to\mathcal{H}\\ w:\mathcal{H}\to\mathcal{Y}}} \quad & \sum_{e\in\mathcal{E}_{\mathrm{tr}}} R^e(w\circ\Phi) \\ \text{subject to} \quad & w \in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w}\circ\Phi), \text{ for all } e\in\mathcal{E}_{\mathrm{tr}}. \end{aligned} \qquad \text{(IRM)}$$

Require the classifier to be simultaneously optimal for all environments!

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} R^e(\Phi) + \lambda \cdot \left\|\nabla_{w|w=1.0}\, R^e(w\cdot\Phi)\right\|^2, \qquad \text{(IRMv1)}$$

ColoredMNIST



- Binary classification: 0~4 as positive class, 5~9 as negative class
- Each image is either red or green
- Domain1 (train): In all positive images, 70% are red; in all negative images, 30% are red.

Algorithm	**Acc. train envs.**	**Acc. test env.**
ERM	87.4 ± 0.2	17.1 ± 0.6
IRM (ours)	70.8 ± 0.9	$\mathbf{66.9 \pm 2.5}$
Random guessing (hypothetical)	50	50
Optimal invariant model (hypothetical)	75	75
ERM, grayscale model (oracle)	73.5 ± 0.2	73.0 ± 0.4

Game theory formulation

$$\min_{\Phi \in \mathcal{H}_\Phi, w^{av} \in \mathcal{H}_w} \sum_{e \in \mathcal{E}_{tr}} R^e(w^{av} \circ \Phi)$$

$$\text{s.t. } w^e \in \arg\min_{\bar{w}^e \in \mathcal{H}_w} R^e\left(\frac{1}{|\mathcal{E}_{tr}|}\left[\bar{w}^e + \sum_{q \neq e} w^q\right] \circ \Phi\right),\ \forall e \in \mathcal{E}_{tr}$$

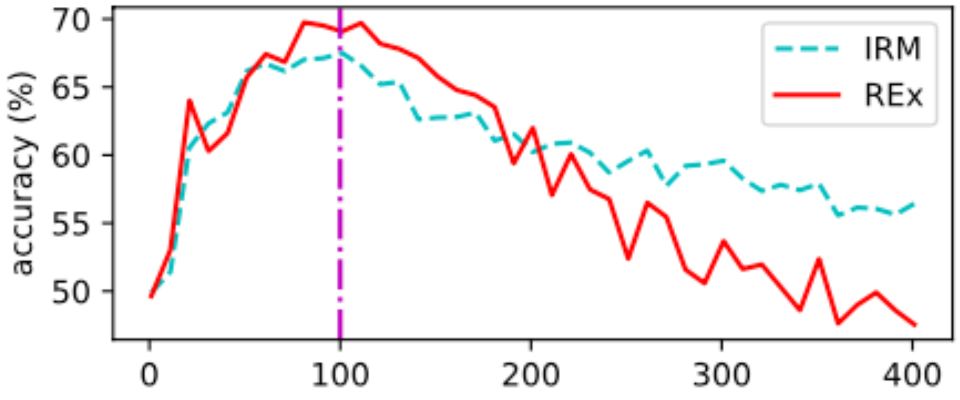
A game between many classifiers

$$\boxed{\begin{aligned} &\min_{\Phi \in \mathcal{H}_\Phi, w^{av}} \sum_{e \in \mathcal{E}_{tr}} R^e(w^{av} \circ \Phi) \\ &\text{s.t. } R^e\left(\frac{1}{|\mathcal{E}_{tr}|}\left[w^e + \sum_{q \neq e} w^q\right] \circ \Phi\right) \\ &\leq R^e\left(\frac{1}{|\mathcal{E}_{tr}|}\left[\bar{w}^e + \sum_{q \neq e} w^q\right] \circ \Phi\right)\ \forall \bar{w}^e \in \mathcal{H}_w\ \forall e \in \mathcal{E}_{tr} \qquad (3)\end{aligned}}$$

IRM Games. ICML2020

REx

$$\mathcal{R}_{\text{V-REx}} \doteq \beta \operatorname{Var}(\{\mathcal{R}_1, ..., \mathcal{R}_m\}) + \sum_{e=1}^{m} \mathcal{R}_e$$

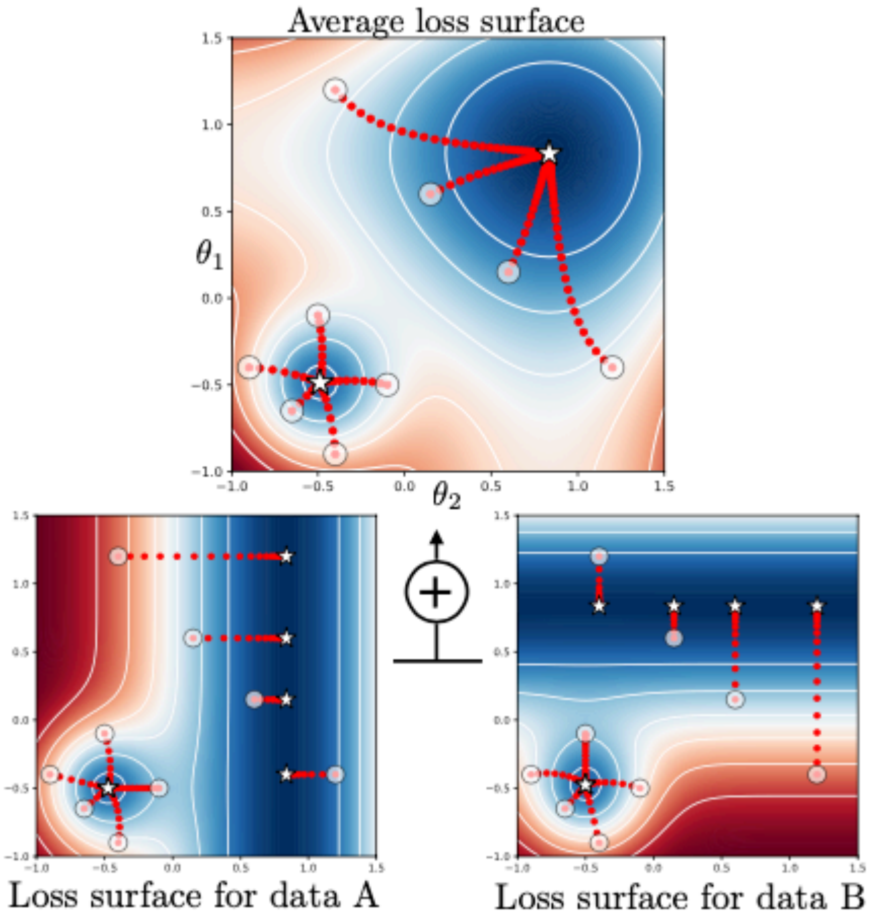


Feature selection effects

Theorem 1. *Given a Linear SEM, $X_i \leftarrow \sum_{j \neq i} \beta_{(i,j)} X_j + \varepsilon_i$, with $Y \doteq X_0$, and a predictor $f_\beta(X) \doteq \sum_{j:j>0} \beta_j X_j + \varepsilon_j$ that satisfies REx (with mean-squared error) over a perturbation set of domains that contains 3 distinct do() interventions for each $X_i : i > 0$. Then $\beta_j = \beta_{0,j}, \forall j$.*

Out-of-Distribution Generalization via Risk Extrapolation (REx), David Krueger et al. ICML2021 oral

Learning explanations that are hard to vary



$$\mathcal{C}^{\epsilon}(\theta^{*}) := - \max_{(e,e')\in\mathcal{E}^2} \max_{\theta\in N^{\epsilon}_{e,\theta^*}} |\mathcal{L}_{e'}(\theta) - \mathcal{L}_{e}(\theta)|.$$

Find the solution where the local geometry is invariant
(2 order information)

Learning explanations that are hard to vary Giambattista Parascandolo et al. ICLR2021

“and mask”

a *threshold* $\tau \in [0, 1]$

$$[m_\tau]_j = \mathbb{1}\left[\tau d \le \left|\sum_e \mathrm{sign}([\nabla \mathcal{L}_e]_j)\right|\right]$$

$$m_t(\theta) \odot \nabla \mathcal{L}(\theta)$$

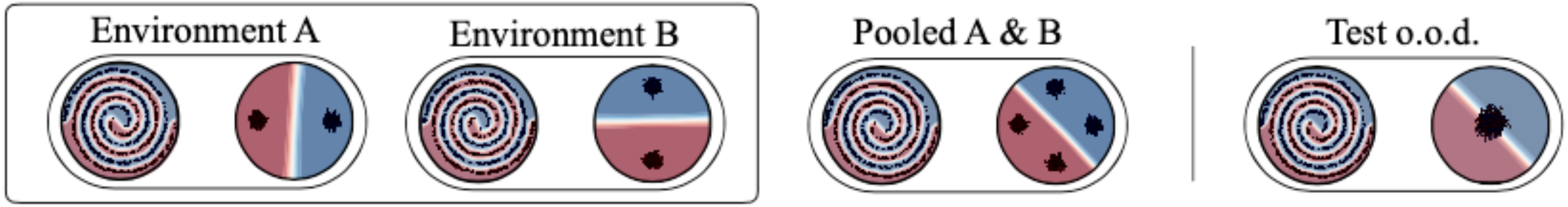


Figure 5: A 4-dimensional instantiation of the synthetic memorization dataset for visualization. Every example is a dot in both circles, and it can be classified by finding either of the "oracle" decision boundaries shown.

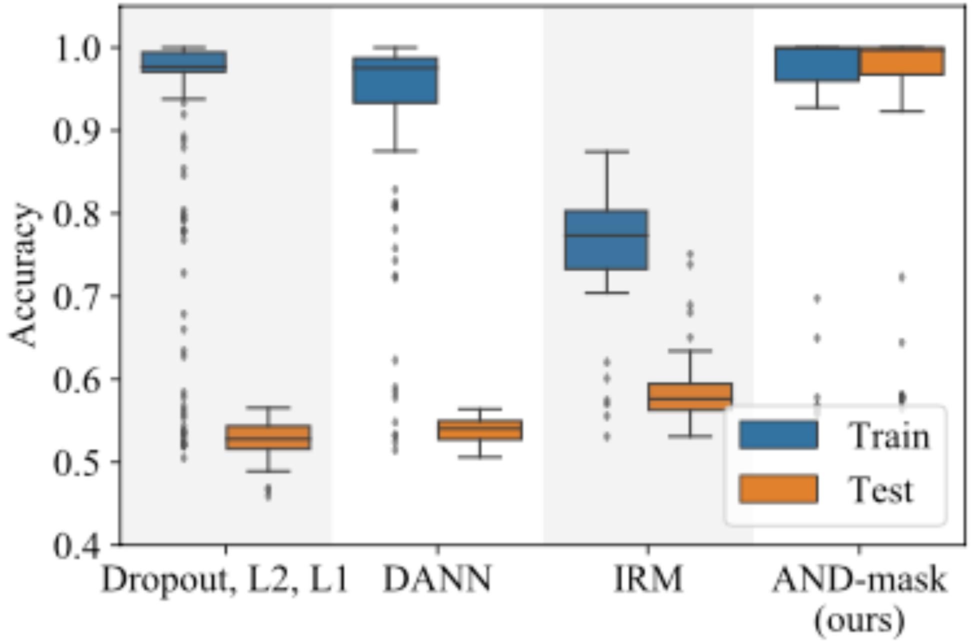


Figure 6: Results on the synthetic dataset.

CIFAR10 random label

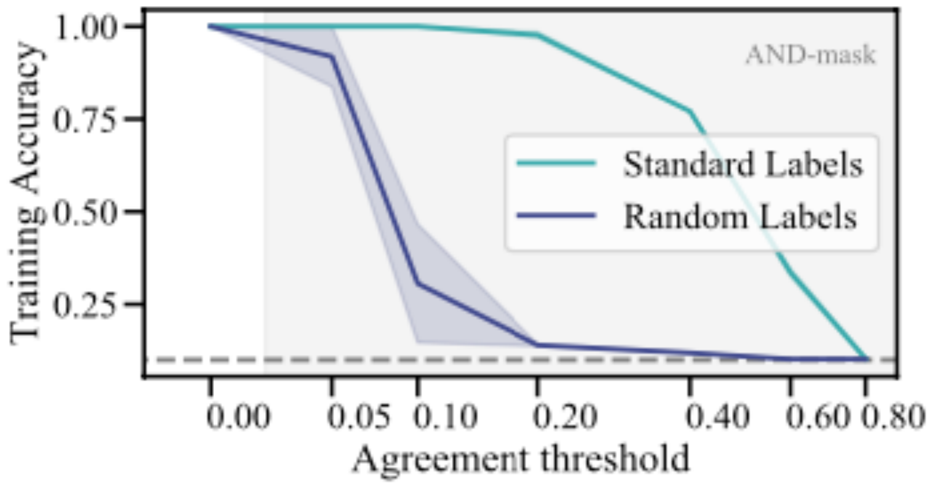


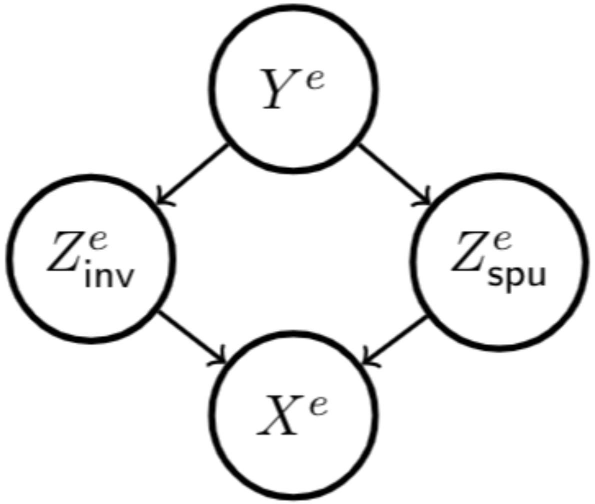
Figure 8: As the AND-mask threshold increases, memorization on CIFAR-10 with random labels is quickly hindered.

Risks of IRM

$$y = \begin{cases} 1, & \text{w.p. } \eta \\ -1, & \text{otherwise.} \end{cases}$$

$$z_c \sim \mathcal{N}(y \cdot \mu_c, \sigma_c^2 I), \qquad z_e \sim \mathcal{N}(y \cdot \mu_e, \sigma_e^2 I)$$

$$x = f(z_c, z_e).$$



$$\mathcal{R}^e(\Phi, \hat{\beta}) := \mathbb{E}_{(x,y)\sim p^e}\left[\ell(\sigma(\hat{\beta}^T \Phi(x)), y)\right]$$

$$\min_{\Phi, \hat{\beta}} \quad \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left[\mathcal{R}^e(\Phi, \hat{\beta}) + \lambda \|\nabla_{\hat{\beta}} \mathcal{R}^e(\Phi, \hat{\beta})\|_2^2\right]$$

Risks of IRM, ICLR2021

Proposition 4.1. *Suppose the observed data are generated according to Equations 1-3. Then recovering the (parametrized) invariant classifier $\Phi(x) = [z_c]$ and $\hat{\beta} = [\beta_c, \beta_0]$ is a stationary point for Equation 4.*

Theorem 5.1 (Linear case)**.** *Assume f is linear. Suppose we observe E training environments. Then the following hold:*

1. *Suppose $E > d_e$. Under mild non-degeneracy conditions, any linear featurizer Φ with an invariant optimal regression vector $\hat{\beta}$ uses only invariant features, and it therefore has identical risk on all possible environments.*

2. *If $E \leq d_e$ and the environmental means μ_e are linearly independent, then there exists a linear Φ with $\mathrm{rank}(\Phi) = d_c + d_e + 1 - E$ whose output depends on the environmental features, yet the optimal classifier on top of Φ is invariant. Further, both the logistic and 0-1 risks of this Φ and its corresponding $\hat{\beta}$ are strictly lower than those of the invariant classifier.*

Equivalence to fairness

- Env index can be seen as sensitive attributes $\mathbb{E}[y|\Phi(x) = h, e]$

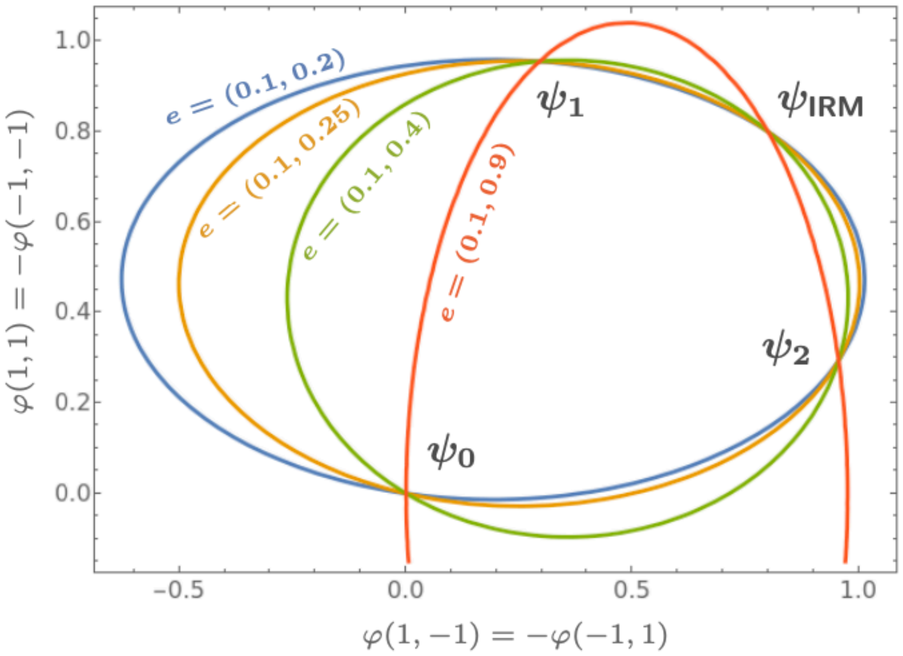
- Settings without domain label:

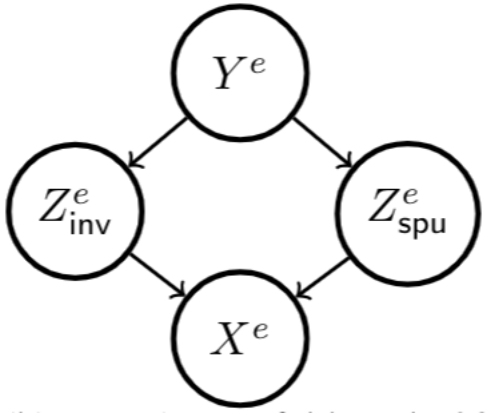
1. Input *reference model* $\tilde{\Phi}$;
2. Fix $\Phi \leftarrow \tilde{\Phi}$ and fully optimize the inner loop of (EIIL) to infer environments $\tilde{\mathbf{q}}_i(e) = \tilde{q}(e|x_i, y_i)$;
3. Fix $\mathbf{q} \leftarrow \tilde{\mathbf{q}}$ and fully optimize the outer loop to yield the new model Φ.

Exchanging Lessons Between Algorithmic Fairness and Domain Generalization, Elliot Creager et al.

Does IRM Capture Invariance?

$$\begin{aligned} Y &\leftarrow \mathrm{Rad}(0.5), \\ X_1 &\leftarrow Y \cdot \mathrm{Rad}(\alpha_e), \qquad \text{(Two-Bit-Envs)} \\ X_2 &\leftarrow Y \cdot \mathrm{Rad}(\beta_e), \end{aligned}$$





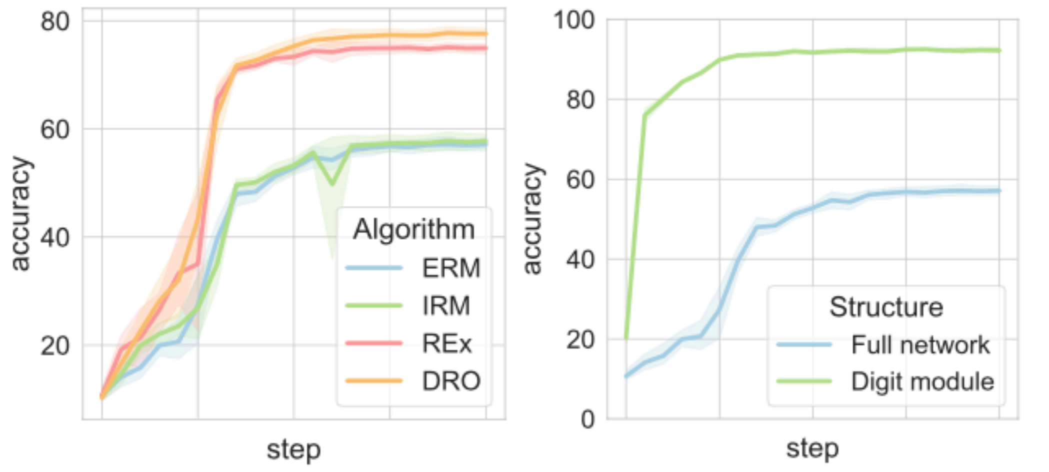
Observation 2. *Under Setting A, a representation $\varphi : \mathcal{X} \to \mathcal{Z}$ is invariant over $\mathcal{E}$ if and only if for all $e_1, e_2 \in \mathcal{E}$, it holds that*

$$\mathbb{E}_{\mathcal{D}_{e_1}}[Y \mid \varphi(X) = z] = \mathbb{E}_{\mathcal{D}_{e_2}}[Y \mid \varphi(X) = z]$$

Does Invariant Risk Minimization Capture Invariance? AISTATS2021 oral

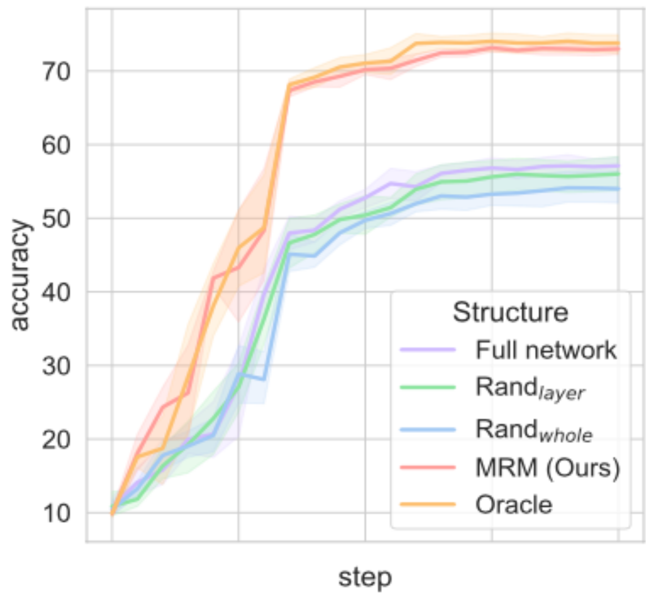
Invariant subnetwork property

Invariant subnetwork exists in normally trained large network:



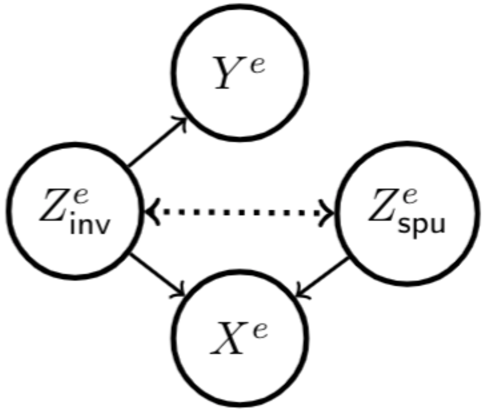
(a) Accuracy of baselines. (b) Accuracy of module in ERM.

Design algorithm to utilize this property:



Can Subnetwork Structure Be the Key to Out-of-Distribution Generalization? ICML2021

Information bottleneck (IB) principle



$$\min_{w\in\mathbb{R}^{k\times r},\Phi\in\mathbb{R}^{r\times d}} \frac{1}{|\mathcal{E}_{tr}|}\sum_{e\in\mathcal{E}_{tr}} h^e\big(w\cdot\Phi\big)$$

$$\text{s.t. } \frac{1}{|\mathcal{E}_{tr}|}\sum_{e\in\mathcal{E}_{tr}} R^e\big(w\cdot\Phi\big) \leq r^{\mathsf{th}},\ w \in \arg\min_{\tilde{w}\in\mathcal{H}_w} R^e(\tilde{w}\cdot\Phi), \forall e\in\mathcal{E}_{tr}$$

Theorem 4. ***IB-IRM and IB-ERM vs IRM and ERM***

- ***Fully informative invariant features.*** *Suppose that the data $\forall e \in \mathcal{E}_{all}$ follows Assumption 2. Assume that the invariant features are separable, bounded, and satisfy support overlap (Assumptions 3,5 and 7 hold). Also, $\forall e \in \mathcal{E}_{tr}$ $Z^e_{\mathsf{spu}} \leftarrow AZ^e_{\mathsf{inv}} + W^e$, where W^e is continuous, bounded, zero mean noise. Every solution of IB-IRM (equation* (6)*, ℓ is* 0*-*1 *loss, $r^{\mathsf{th}} = q$), and IB-ERM solves OOD generalization (equation* (1)*) but ERM and IRM fail.*

Invariance principle meets information bottleneck for out-of-distribution generalization. Submitted.

- More papers at https://sites.google.com/site/irinarish/ood_generalization
- Thank you very much!