Sparse Invariant Risk Minimization

Xiao Zhou*[1], Yong Lin*[1], Weizhong Zhang*[1], Tong Zhang[1,2]

[1]The Hong Kong University of Science and Technology [2]Google Research

June 27, 2022

Motivation and Challenges

Motivation: How can we address a basic and intractable contradiction between the model trainability and generalization ability in IRM?

Dilemma

- Large-sized or even overparameterized neural networks are important to make the model easy to train.
- Generalization ability of IRM is much easier to be demolished by overfitting caused by overparameterization.

Investigating the Effects of Overparameterization on IRM

Data Generation Process

Two training environments $\mathcal{E}_{tr} = \{e_1, e_2\}$. $\boldsymbol{x}^e$ the input feature of environment $e \in \mathcal{E}_{tr}$, concatenated from invariant feature $\boldsymbol{x}^e_{inv}$, spurious feature $\boldsymbol{x}^e_s$ and the random feature $\boldsymbol{x}^e_r$, i.e., $\boldsymbol{x}^e := [\boldsymbol{x}^e_{inv}, \boldsymbol{x}^e_s, \boldsymbol{x}^e_r] \in \mathbb{R}^d$. The data is generated as follows:

$$
\begin{aligned}
y^e &= \gamma^\top \boldsymbol{x}^e_{inv} + \epsilon_{inv}, \\
\boldsymbol{x}^e_s &= y^e \mathbf{1}^{\mathbf{s}} + \alpha^e \circ \epsilon_s \\
\boldsymbol{x}^e_r &= \epsilon_r,
\end{aligned}
$$

where $e \in \mathcal{E}_{tr}$, $\epsilon_{inv}, \epsilon_s$ and ϵ_r are independent random noise that follows sub-Gaussian distributions with zero mean and bounded variance.

Investigating the Effects of Overparameterization on IRM

Setting

We aim to learn a linear model to predict y based on $\boldsymbol{x}$:

$$f(\boldsymbol{x}; \boldsymbol{w}) = (\Phi \circ \boldsymbol{x})^\top \boldsymbol{v} + b, \tag{1}$$

where $\Phi \in \{0,1\}^{d_{inv}+d_s+d_r}$ is a binary vector to perform feature selection . $\boldsymbol{v} \in \mathbb{R}^{d_{inv}+d_s+d_r}$ is the parameter of the linear function on the top of Φ. We denote the $\hat{\mathcal{L}}(\Phi)$ as loss of a given Φ when $\boldsymbol{v}$ is solved optimally, $\hat{\mathcal{L}}(\Phi) := \min_{\boldsymbol{v}} \mathcal{L}(\boldsymbol{w})$.
The ideal feature selector is $\Phi_{inv} = [\mathbf{1}^{d_{inv}}, \mathbf{0}^{d_s+d_r}]$, merely selecting the invariant feature $\boldsymbol{x}_{inv}$ and discarding spurious features $\boldsymbol{x}_s$ and random features $\boldsymbol{x}_r$. IRM learns $\boldsymbol{w}$ by minimizing $\mathcal{L}(\boldsymbol{w})$, therefore, it can finally find the ideal feature selector Φ_{inv} if and only if the following condition holds

$$\hat{\mathcal{L}}(\Phi_{inv}) < \hat{\mathcal{L}}(\Phi), \forall \Phi \neq \Phi_{inv}. \tag{2}$$

Investigating the Effects of Overparameterization on IRM

Proposition

(Failure of IRM in Overparameterization Region). If $d_{inv} + d_s + d_r > n_{e_1} + n_{e_2}$, *then*

$$\hat{\mathcal{L}}(\Phi_{all}) = 0 \leq \hat{\mathcal{L}}(\Phi_{inv}), \tag{3}$$

where $\Phi_{all} = \mathbf{1}^{d_{inv}+d_s+d_r}$.

Corollary

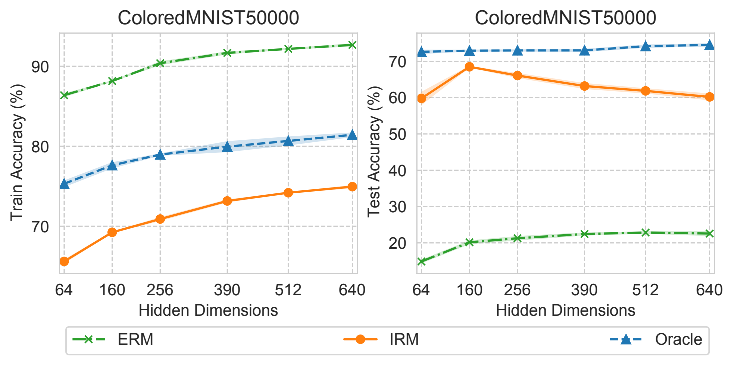
(Worse Case) If $d_s + d_r > n_{e_1} + n_{e_2}$, *then*

$$\hat{\mathcal{L}}(\Phi_{all}) = \hat{\mathcal{L}}(\Phi_{sr}) = 0 \leq \hat{\mathcal{L}}(\Phi_{inv}) \tag{4}$$

where $\Phi_{sr} = [\mathbf{0}^{d_{inv}}, \mathbf{1}^{d_s+d_r}]$.

Empirical Verification

As the hidden dimension increases, the training and testing accuracy of ERM and Oracle increases steadily. However, the testing accuracy of IRM decreases while its training accuracy increases.

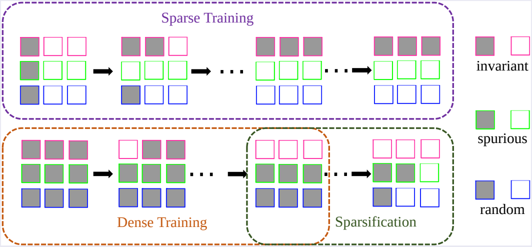


Our Method

SparseIRM

$$\min_{\boldsymbol{w},\boldsymbol{s}} \ \mathbb{E}_{p(\boldsymbol{m}|\boldsymbol{s})} \ \mathcal{L}(\{\boldsymbol{v}, \boldsymbol{m} \circ \Phi\}) \tag{5}$$
$$s.t. \ \boldsymbol{w} \in \mathbb{R}^{d_w}, \boldsymbol{s} \in \mathcal{S} := \{\boldsymbol{s} \in [0,1]^{d_\Phi} : \mathbf{1}^\top \boldsymbol{s} \leq K\}.$$

FlowChart



Understanding the Benefits of SparseIRM through Theoretical Analysis

Theorem

Under assumptions specified in Appendix B.6.2, assume $n_{e_1} = n_{e_1} = n$*, if* $n > Q_1 + Q_2 \ln(d/\delta)$ *and choosing* $K = d_{inv}$*, then with probability at least* $1 - \delta$ *the following inequality holds:*

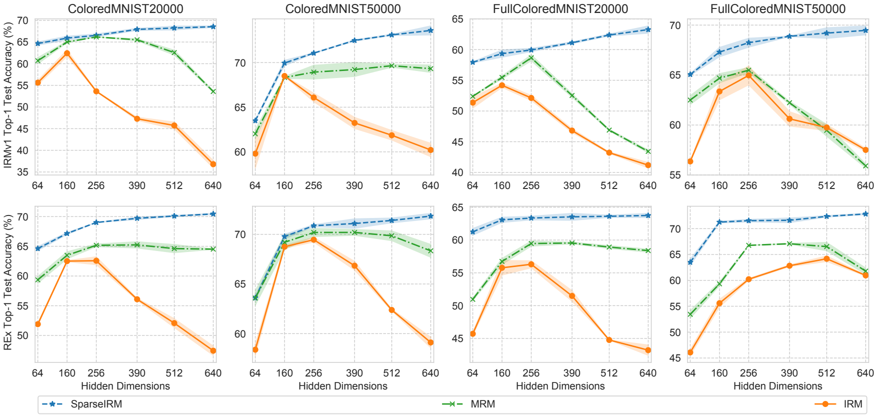
$$\hat{\mathcal{L}}(\Phi_{inv}) < \hat{\mathcal{L}}(\Phi), \forall\ \Phi \neq \Phi_{inv} \text{ and } \|\Phi\|_1 \leq K, \tag{6}$$

where Q_1 *and* Q_2 *are constants specified in the appendix.*

Theorem 1 indicates that in the linear case, SparseIRM can provably find the invariant features as long as the number of the data samples is larger than a logarithmic term of spurious and random features.

The sparsity constraint limits the number of features to be selected, in which way any combinations of spurious or random features not exceeding the constraint will only lead to a larger loss.

Experiments: ColoredMNIST



Experiments: ColoredObject/CIFARMNIST and Ablation Studies

Table 3. Comparison of Top-1 Test Accuracy on ResNet-18 on ColoredObject and CIFARMNIST.

Dataset		ColoredObject	CIFARMNIST
Oracle		87.9 ± 0.3	83.7 ± 1.5
ERM		51.6 ± 0.5	39.5 ± 0.4
SparseERM		54.4 ± 0.4	40.1 ± 0.8
BayesianIRM		78.1 ± 0.6	59.3 ± 0.8
IRMv1	IRM	72.5 ± 2.3	51.3 ± 3.0
	MRM	58.4 ± 0.9	56.7 ± 2.3
	SparseIRM	$\mathbf{87.4 \pm 0.6}$	$\mathbf{63.9 \pm 0.4}$
REx	IRM	73.8 ± 1.3	50.1 ± 2.2
	MRM	55.7 ± 2.9	52.6 ± 1.5
	SparseIRM	$\mathbf{80.3 \pm 1.1}$	$\mathbf{62.7 \pm 0.6}$

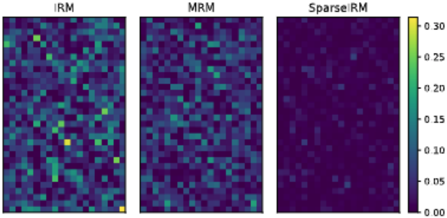


Figure 4. Comparison of absolute value of difference of feature representations by flipping spurious features. The dimension of feature representation 640 and we reshape it into 32×20 matrix for better visualization.