What is Missing in IRM Training and Evaluation? Challenges and Solutions

Yihua Zhang[1], Pranay Sharma[2], Parikshit Ram[3], Mingyi Hong[4], Kush Varshney[3], Sijia Liu[1, 3]

[1]Michigan State University, [2]Carnegie Mellon University, [3]IBM Research, [4]University of Minnesota, Twin City

Code

Paper

ICLR

What is Missing in IRM Training and Evaluation?

Invariant Risk Minimization

- to acquire **environment-agnostic** data representations
- to avoid learning **spurious correlations** in the data

Problem Setup[1]

➢ IRM is formulated as a bi-level optimization (BLO) problem:

$$\min_{\theta} \sum_{e\in\mathcal{E}_{tr}} \ell^{(e)}(\mathrm{w}^*(\theta) \circ \theta);\ s.t.\ \ \mathrm{w}^*(\theta) \in \min_{\mathrm{w}} \ell^{(e)}(\mathrm{w}\circ\theta), \forall e \in \mathcal{E}_{tr}$$

➢ IRMv1 simplifies the BLO to a single-level problem:

$$\min_{\theta} \sum_{e\in\mathcal{E}_{tr}} [\ell^{(e)}(\theta) + \gamma \left\| \nabla_{w|w=1.0} \ell^{(e)}(\mathrm{w}\circ\theta) \right\|_2^2]$$

❖ Challenge I: Large-Batch Training

➢ Large-batch optimization causes **suboptimal** IRM training.

➢ **Our proposal: Small-batch training** is effective versus a zoo of large-batch optimization enhancements.

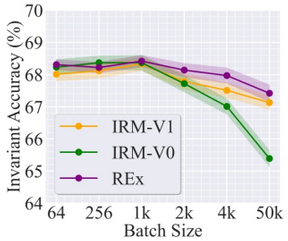
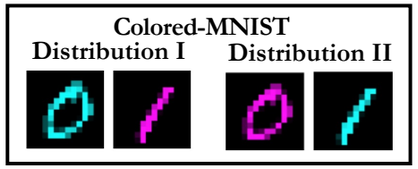


Figure 1. The performance of three IRM methods (IRMV1, IRMV0, and REX) vs. batch-size under Colored-MNIST.

Method	Original	LSGD	LALR	SAM	**Small**
IRMV1	67.13	67.31	67.44	67.79	**68.33**
IRMV0	65.39	66.42	66.76	66.99	**68.37**
IRM-GAME	65.69	65.82	65.47	66.23	**67.73**
REX	67.42	67.53	67.59	67.82	**68.42**
BIRM	67.93	67.99	68.21	68.32	**68.71**
SPARSEIRM	67.72	67.85	67.99	68.13	**68.81**
FISHR	67.88	67.82	67.93	68.11	**68.69**
Average	67.02	67.25	67.34	67.63	**68.44**

Table 1. Accuracy of IRM methods on Colored-MNIST using the original large-batch implementation (Original), the large-batch optimization-integrated implementations (LSGD/LALR/SAM), and the small-batch training recipe.

❖ Challenge II: Multi-Environment Invariance Evaluation



Data Environment (β)

β Dist. I+ $(1-\beta)$ Dist. II

Training Env. I: $\beta = 0.1$

Training Env. II: $\beta = 0.2$

➢ The evaluation metric adopted by all existing work focuses on a **single** test environment.

➢ IRM performance is sensitive to test environment choices. Single-environment evaluation leads to a false sense of invariance.

➢ **Our proposal**: evaluation across multiple environments. Good method should achieve high Avg. Acc. and low Acc. Gap.

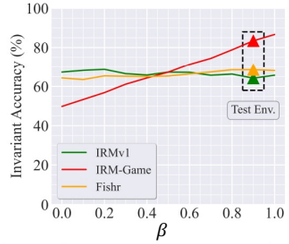


Figure 2. Performance comparison of different IRM methods under diverse test environments. Existing methods only evaluate at $\beta = 0.9$.

❖ Challenge III: IRM-Game[2] with Invariant Predictor

➢ IRM-Game assigns each environment an individual classifier $\mathrm{w}^{(e)}$. The output relies on the **ensemble** of individual predictors.

➢ **Our proposal: BLOC-IRM** (**BLO** with Consensus **IRM**):

$$\underset{\theta}{minimize} \sum_{e\in\mathcal{E}_{tr}} [\ell^{(e)}(\mathrm{w}^*(\theta)\circ\theta) + \gamma \left\| \nabla_w \ell^{(e)}(\mathrm{w}^*(\theta)\circ\theta) \right\|_2^2]$$

$$\text{subject to } \ \mathrm{w}^{(e)}(\theta) \in \underset{\mathrm{w}^{(e)}}{\operatorname{argmin}}\ \ell^{(e)}(\mathrm{w}^{(e)}\circ\theta), \forall e \in \mathcal{E}_{tr};\ \mathrm{w}^*(\theta) = \frac{1}{N}\sum_{e\in\mathcal{E}_{tr}} \mathrm{w}^{(e)}(\theta)$$

- The lower level ensures (I) per-environment risk minimization and (II) an environment-invariant predictor.
- The upper level minimizes the ERM loss and the **regularization term** penalizes the **lower-level stationarity**.
- This problem can be solved using an ordinary BLO solver.

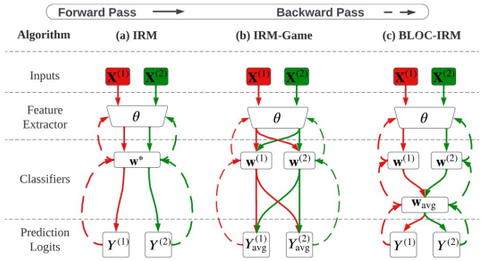


Figure 3. (Algorithm Comparison) Schematic overview of BLOC-IRM over two training environments (red and green), and its comparison to IRM and IRM-Game.

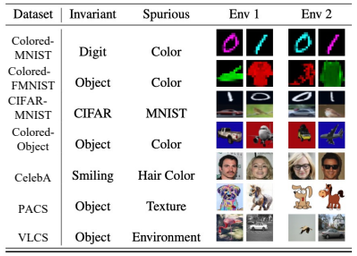


Figure 4. (Dataset Setups) An overview of the 'Invariant' and 'Spurious' features in different datasets used in this work.

Experiment Results

Algorithm	COLORED-OBJECT		CIFAR-MNIST		CELEBA		VLCS		PACS	
	Avg Acc (%)	Acc Gap	Avg Acc (%)	Acc Gap	Avg Acc (%)	Acc Gap	Avg Acc (%)	Acc Gap	Avg Acc (%)	Acc Gap
ERM	41.11±1.44	86.43±2.89	40.39±1.32	85.53±2.33	72.38±0.29	10.73±0.36	63.23±0.23	12.39±0.35	69.95±0.35	14.32±0.75
IRMV1	64.42±0.21	4.18±0.29	61.49±0.29	7.17±0.33	72.49±0.38	10.15±0.27	62.72±0.29	12.74±0.27	68.93±0.33	14.99±0.51
IRMV0	62.39±0.25	5.36±0.31	60.14±0.18	8.83±0.39	72.42±0.35	10.43±0.38	62.59±0.32	12.99±0.36	68.72±0.29	15.29±0.71
IRM-GAME	62.88±0.34	5.59±0.28	60.44±0.31	6.72±0.41	72.18±0.44	12.32±0.41	62.31±0.38	13.37±0.62	68.12±0.22	15.77±0.66
REX	63.37±0.35	5.42±0.31	62.32±0.24	5.55±0.32	72.34±0.26	10.31±0.23	63.19±0.31	12.87±0.31	69.43±0.34	15.31±0.67
BIRM	65.11±0.27	**3.31**±0.22	62.99±0.35	5.23±0.36	72.93±0.28	9.92±0.33	63.33±0.40	12.13±0.23	69.34±0.25	15.76±0.49
SPARSEIRM	64.97±0.39	3.97±0.25	62.16±0.29	**4.14**±0.31	72.42±0.33	9.79±0.21	62.86±0.26	12.79±0.35	69.52±0.39	15.81±0.82
FISHR	64.07±0.23	4.41±0.29	61.79±0.25	5.55±0.21	72.89±0.25	9.42±0.32	63.44±0.37	11.93±0.42	70.21±0.22	**14.52**±0.43
BLOC-IRM	**65.97**±0.33	4.10±0.36	**63.69**±0.32	4.89±0.36	**73.35**±0.32	**8.79**±0.21	**63.62**±0.35	**11.55**±0.32	**70.31**±0.21	14.73±0.65

Table 2. (Main Results) IRM performance comparison between BLOC-IRM and other baselines.

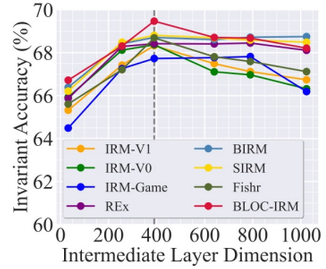


Figure 5. Ablation study on model size on the dataset Colored-MNIST.

Environment Metrics (%)	$p_{tr} \in \{0.1, 0.15\}$		$p_{tr} \in \{0.1, 0.15, 0.2\}$	
	Avg Acc	Acc Gap	Avg Acc	Acc Gap
OPTIMUM	75.00	0.00	75.00	0.00
GRAYSCALE ERM	73.82±0.11	0.37±0.05	73.97±0.14	0.29±0.08
	49.21±0.79	91.88±3.31	49.03±0.93	92.17±3.04
IRMV1	67.36±0.31	2.77±0.15	67.11±0.34	2.42±0.12
IRMV0	67.01±0.42	2.85±0.18	66.71±0.42	2.36±0.19
IRM-GAME	66.39±0.72	4.47±0.61	65.93±0.53	4.25±0.84
REX	66.82±0.44	2.59±0.11	67.14±0.38	2.16±0.13
BIRM	67.35±0.39	2.65±0.10	68.05±0.43	**1.99**±0.07
SPARSEIRM	67.12±0.53	2.33±0.18	67.72±0.41	2.11±0.19
FISHR	67.22±0.43	2.44±0.15	67.32±0.39	2.59±0.15
BLO-IRM	**68.72**±0.41	**2.19**±0.15	**68.89**±0.31	2.39±0.09

Table 3. Ablation study on different training environments on the dataset Colored-MNIST.

Reference [1] Martin Arjovsky et al., Invariant risk minimization. [2] Kartik Ahuja, et al. Invariant risk minimization games.

Contact: {zhan1908@msu.edu}