# Pareto Invariant Risk Minimization

Yongqiang Chen [1] Kaiwen Zhou [1] Yatao Bian [2] Binghui Xie [1] Kaili Ma [1] Yonggang Zhang [3] Han Yang [1]
Bo Han [3] James Cheng [1]

## Abstract

Despite the success of invariant risk minimization (IRM) in tackling the Out-of-Distribution generalization problem, IRM can compromise the optimality when applied in practice. The practical variants of IRM, e.g., IRMv1, have been shown to have significant gaps with IRM and thus could fail to capture the invariance even in simple problems. Moreover, the optimization procedure in IRMv1 involves two intrinsically conflicting objectives, and often requires careful tuning for the objective weights. To remedy the above issues, we reformulate IRM as a multi-objective optimization problem, and propose a new optimization scheme for IRM, called **PA**reto **I**nvariant **R**isk Minimization (`PAIR`). `PAIR` can adaptively adjust the optimization direction under the objective conflicts. Furthermore, we show `PAIR` can empower the practical IRM variants to overcome the barriers with the original IRM when with proper guidance. We conduct experiments with ColoredMNIST to confirm our theory and the effectiveness of `PAIR`.

## 1. Introduction

There are surging evidences showing that machine learning models using empirical risk minimization (ERM) (Vapnik, 1991) are prone to exploit shortcuts, or spurious features, and thus can fail to generalize to *out-of-distribution* (OOD) (Beery et al., 2018; DeGrave et al., 2021; Geirhos et al., 2020; Koh et al., 2021; Ji et al., 2022) data. To address the OOD failure, several strategies and algorithms were proposed (Peters et al., 2016; Rojas-Carulla et al., 2018; Bengio et al., 2020; Sagawa* et al., 2020; Koyama & Yamaguchi, 2020a; Krueger et al., 2021; Creager et al., 2021; Ahuja et al., 2021a; Shi et al., 2022; Rame et al., 2021; Chen et al., 2022; Zhang et al., 2022). Among them, the Invariant Risk Minimization (IRM) (Arjovsky et al., 2019) framework has attracted much attention recently. IRM assumes datasets are collected from multiple causally related environments, and tries to find an invariant data representation $\varphi$. When a predictor $w$ acting on $\varphi$ minimizes the risks in all of the environments simultaneously, $\varphi$ is expected to discard the spurious signals while keeping the (causally) invariant signals (Peters et al., 2016; Arjovsky et al., 2019). Although being powerful, IRM can lead to suboptimal solutions when applied in the practice (Kamath et al., 2021; Rosenfeld et al., 2021; Gulrajani & Lopez-Paz, 2021; Nagarajan et al., 2021; Aubin et al., 2021).

Specifically, to alleviate the difficulty in solving the challenging non-linear bi-level programming in the original IRM, Arjovsky et al. (2019) relax IRM into IRM$_\mathcal{S}$ by restricting the $w$ to be linear, and further propose a more practical soft-constrained variant, IRMv1. However, Kamath et al. (2021) show the relaxed variants can have huge gap with the original IRM. In a simple example following the setting of IRM (Arjovsky et al., 2019), both IRM$_\mathcal{S}$ and IRMv1 could fail to capture the desired invariance even with infinite amounts of samples and environments. Moreover, since the optimization of IRMv1 involves two intrinsically conflicting objectives, the objective weights require careful tuning, otherwise IRMv1 may even underperform the vanilla empirical risk minimization (ERM) algorithm (Vapnik, 1991; Gulrajani & Lopez-Paz, 2021; Zhang et al., 2022).

Aiming to bridge the gap between the practical variants and the original IRM, we propose a new perspective with multi-objective optimization (MOO) for understanding the different behaviors between IRM$_\mathcal{S}$, IRMv1 and IRM, and propose a new optimization scheme for IRM, termed as **PA**reto **I**nvariant **R**isk Minimization (`PAIR`). By reformulating IRM as a MOO problem, we find the failures of the practical variants are caused by using the improper objectives for optimization. Therefore, we propose to pair IRM$_\mathcal{S}$ and IRMv1 with additional guidance from other robust objectives. More concretely, we show that, when guided by REx (Krueger et al., 2021), `PAIR` avoids the failure cases of IRM$_\mathcal{S}$ and IRMv1, and is capable of locating the desired invariant predictors using gradient-based methods such as the multiple-gradient descent algorithm (Désidéri, 2012). Consequently, `PAIR` resolves the theoretical drawbacks in IRM$_\mathcal{S}$ and thus is more powerful than IRM$_\mathcal{S}$. We also conduct experiments on ColoredMNIST (Kamath et al., 2021)

[1] The Chinese University of Hong Kong [2] Tencent AI Lab [3] Hong Kong Baptist University.

*Preliminary work.*

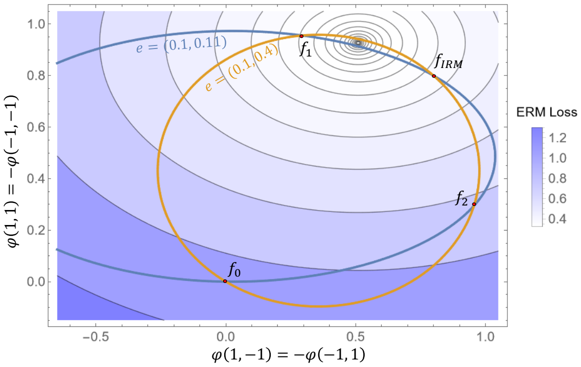


*Figure 1.* Theoretical failure case of IRM$_\mathcal{S}$ and IRMv1.

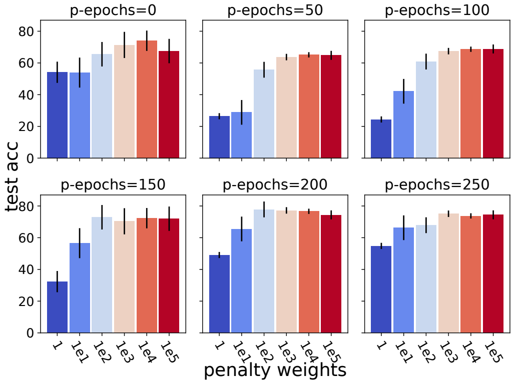


*Figure 2.* IRMv1 requires significant tuning efforts.

and its modified variants to verify the effectiveness of PAIR.

## 2. Drawbacks of IRM in Practice

We begin by introducing the basics of IRM and drawbacks of its practical variants following Kamath et al. (2021).

**Problem Setup.** The IRM (Arjovsky et al., 2019) framework typically considers a supervised learning setting based on the data $\mathcal{D} = \{\mathcal{D}^e\}_{e\in\mathcal{E}_{\text{all}}}$ collected from multiple causally related environments $\mathcal{E}_{\text{all}}$, where $\mathcal{D}^e = \{X_i^e, Y_i^e\}$ from environment $e \in \mathcal{E}_{\text{all}}$ are considered as drawn independently from an identical distribution $\mathbb{P}^e$. The goal of OOD generalization is to find a predictor $f : \mathcal{X} \to \mathcal{Y}$ that generalizes well to all (unseen) environments, i.e., to minimize $\max_{e\in\mathcal{E}_{\text{all}}} R^e(f)$ where $R^e$ is the empirical risk under environment $e$. IRM approaches the problem by finding an invariant representation $\varphi : \mathcal{X} \to \mathcal{Z}$, such that there exists a predictor $w : \mathcal{Z} \to \mathcal{Y}$ acting on $\varphi$ that is simultaneously optimal among $\mathcal{E}_{\text{all}}$, i.e., $w \in \arg\min_{\bar{w}:\mathcal{Z}\to\mathcal{Y}} \mathcal{L}_e(\bar{w}\circ\varphi),\ \forall e \in \mathcal{E}_{\text{all}}$. Hence, IRM leads to a challenging bi-level optimization problem (Arjovsky et al., 2019) as,

$$\begin{aligned} &\min_{w,\varphi} \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{L}_e(w\circ\varphi), \\ &\text{s.t. } w \in \mathop{\arg\min}_{\bar{w}:\mathcal{Z}\to\mathcal{Y}} \mathcal{L}_e(\bar{w}\circ\varphi),\ \forall e \in \mathcal{E}_{\text{tr}}. \end{aligned} \tag{1}$$

Let $\mathcal{I}(\mathcal{E}_{\text{tr}})$ denote the set of invariant predictors $w \circ \varphi$ by solving Eq. 1, given the training environments $\mathcal{E}_{\text{tr}}$, and $\mathcal{W}$ and $\varPhi$ be the functional spaces for $w$ and $\varphi$, respectively. Characterizing $\mathcal{I}(\mathcal{E}_{\text{tr}})$ is particularly difficult in practice, given the access only to finite samples from a small subset of environments. It is natural to introduce a restriction that $\mathcal{W}$ is the space of linear functions on $\mathcal{Z} = \mathbb{R}^d$ (Jacot et al., 2021). Furthermore, Arjovsky et al. (2019) argues that linear predictors actually do not provide additional representation power than *scalar* predictors, i.e., $d = 1, \mathcal{W} = \mathcal{S} = \mathbb{R}^1$. The scalar restriction on $\mathcal{W}$ elicits a practical variant IRM$_\mathcal{S}$ as

$$\begin{aligned} &\min_{\varphi} \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{L}_e(\varphi), \\ &\text{s.t. } \nabla_{w|w=1}\mathcal{L}_e(w\cdot\varphi) = 0,\ \forall e \in \mathcal{E}_{\text{tr}}. \end{aligned} \tag{2}$$

Let $\mathcal{I}_\mathcal{S}(\mathcal{E}_{\text{tr}})$ denote the set of invariant predictors elicited by IRM$_\mathcal{S}$. As proven by Kamath et al. (2021), given a convex and differentiable loss $\mathcal{L}_e$, we have $\mathcal{I}(\mathcal{E}_{\text{tr}}) \subseteq \mathcal{I}_\mathcal{S}(\mathcal{E}_{\text{tr}})$. Yet, Eq. 2 remains a constrained programming. Hence, Arjovsky et al. (2019) introduce a soft-constrained variant IRMv1 as

$$\min_{\varphi} \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{L}_e(\varphi) + \lambda|\nabla_{w|w=1}\mathcal{L}_e(w\cdot\varphi)|^2. \tag{3}$$

**Theoretical Failure of Practical IRM Variants.** Although the practical variants seem promising, Kamath et al. (2021) show there exists huge gaps between the variants and the original IRM such that both IRM$_\mathcal{S}$ and IRMv1 can fail to capture the desired invariance, even being given the *population loss* and *infinite* amount of training environments. The failure case, called two-bit environment (Kamath et al., 2021), follows the setup of ColoredMNIST in IRM (Arjovsky et al., 2019), and defines environments with two parameters $\alpha_e, \beta_e \in [0, 1]$. Each $\mathcal{D}_e$ is defined as

$$Y := \text{Rad}(0.5), X_1 := Y\cdot\text{Rad}(\alpha_e), X_2 := Y\cdot\text{Rad}(\beta_e), \tag{4}$$

where $\text{Rad}(\sigma)$ is a random variable taking value $-1$ with probability $\sigma$ and $+1$ with probability $1 - \sigma$. We denote a environment $e$ with $(\alpha_e, \beta_e)$ for simplicity. The setup in IRM can then be denoted as $\mathcal{E}_\alpha = \{(\alpha, \beta_e) : 0 < \beta_e < 1\}$ where $X_1$ is the invariant feature as $\alpha$ is fixed for different $e$. Although IRM$_\mathcal{S}$ and IRMv1 is shown to be successful in $\mathcal{E}_{\text{tr}} := \{(0.25, 0.1), (0.25, 0.2)\}$, Kamath et al. (2021) show that the set of "invariant predictors" produced by IRM$_\mathcal{S}$ and IRMv1 is broader than our intuitive sense. For example, when given $\mathcal{E}_{\text{tr}} := \{(0.1, 0.11), (0.1, 0.4)\}$, the solutions satisfying the constraint in IRM$_\mathcal{S}$ are those intersected points in Fig. 1 (The ellipsoids are the constraints). Although $f_0, f_1, f_2, f_{\text{IRM}} \in \mathcal{I}_\mathcal{S}(\mathcal{E}_{\text{tr}})$, both IRM$_\mathcal{S}$ and IRMv1 prefer $f_1$ instead of $f_{\text{IRM}}$ (the predictor elicited by the original IRM), as $f_1$ has the smallest ERM loss. In fact, Kamath

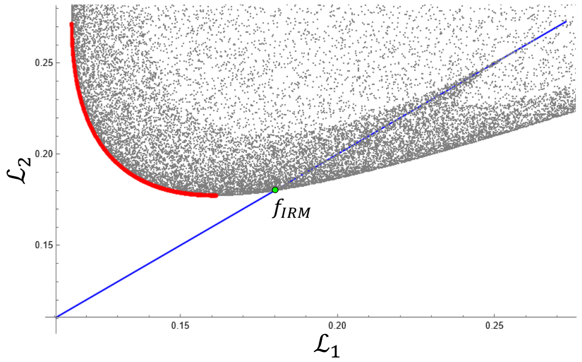


*Figure 3.* Approximated Pareto front.

et al. (2021) prove that, the failure can happen on a wide range of environments with $\alpha < 0.1464$ and $\alpha > 0.8356$, even being given *infinite* number of additional environments, when using the MSE loss. It follows that $\mathcal{I}(\mathcal{E}_{\text{tr}}) \subsetneq \mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\text{tr}})$, demonstrating the significant gap between the practical variants and the original IRM.

**Practical Drawback of Practical IRM Variants.** In addition to the theoretical gap, the optimization of IRMv1 is also difficult due to the conflicts between the IRM penalty and empirical risk in Eq. 3. It often requires significant efforts for choosing proper hyperparameters such as pretraining epochs and objective penalty weights, i.e., $\lambda$. Otherwise, IRMv1 may not enforce the constraint in IRM$_{\mathcal{S}}$, hence will lead to unsatisfactory performance, shown as Fig. 2.

## 3. Pareto Optimization for IRM

As shown that both IRM$_{\mathcal{S}}$ and IRMv1 have difficulties in handling the trade-off between ERM loss and IRM penalty, we introduce a new learning scheme for understanding the differences between the variants and the original IRM based on Multi-Objective optimization (MOO).

**IRM as Multi-Objective Optimization.** MOO typically considers solving $m$ objectives, with $\{\mathcal{L}_i\}_{i=1}^m$ loss functions, i.e., $\min_\theta \boldsymbol{L}(\theta) := (\mathcal{L}_1(\theta), ..., \mathcal{L}_m(\theta))^T$ (Kaisa, 1999). A solution $\theta$ dominates another $\bar{\theta}$ if $\mathcal{L}_i(\theta) \leq \mathcal{L}_i(\bar{\theta})$ for all $i$ and $\boldsymbol{L}(\theta) \neq \boldsymbol{L}(\bar{\theta})$. A solution $\theta^*$ is called **Pareto optimal** if there exists no other solution that dominates $\theta^*$. The set of Pareto optimal solutions is called Pareto set and its image is called **Pareto front**, denoted as $\mathcal{P}$. In practice, it is usual that we cannot find a global optimal solution for all objectives, hence Pareto optimal solutions are of particular value. The paradigm proposed in multiple-gradient descent algorithm (MGDA) (Désidéri, 2012) can efficiently find the Pareto optimal solutions, and has been widely applied in multi-task learning (Sener & Koltun, 2018; Lin et al., 2019; Ma et al., 2020; Mahapatra & Rajan, 2020).

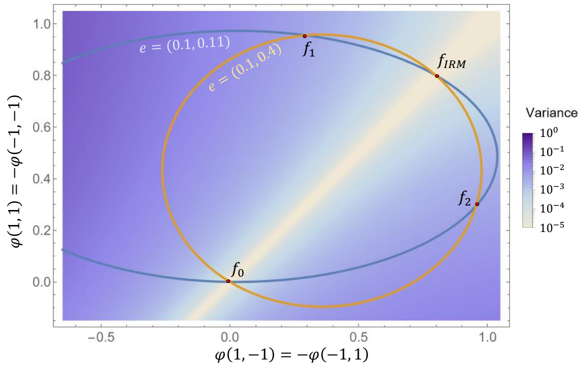


*Figure 4.* Visualization of variance distribution.

To understand the behavior of IRM$_{\mathcal{S}}$ and IRMv1, it is natural to introduce the following objectives as

$$\min_{\varphi} (\mathcal{L}_1, ..., \mathcal{L}_{|\mathcal{E}_{\text{tr}}|}, \mathcal{L}_{\text{IRM}})^T, \quad (5)$$

where we denote $\sum_e |\nabla_{w|w=1} \mathcal{L}_e(w \cdot \varphi)|^2$ as $\mathcal{L}_{\text{IRM}}$ for short. We then visualize the Pareto front with w.r.t. $\mathcal{L}_1, \mathcal{L}_2$ in Fig. 3, using the failure case in Fig. 1. Let $\mathcal{P}(\mathcal{L}_1(\theta), ..., \mathcal{L}_m(\theta))$ denote the set of Pareto optimal solutions w.r.t. the objectives $(\mathcal{L}_1(\theta), ..., \mathcal{L}_m(\theta))$. At first, we can find that $f_{\text{IRM}} \notin \mathcal{P}(\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_{\text{IRM}})$. In other words, any weighted ERM cannot find $f_{\text{IRM}}$ hence not capture the invariance. Moreover, combining Fig. 1, we can also find that $f_{\text{IRM}} \notin \mathcal{P}(\mathcal{E}_{\text{tr}})$ w.r.t. $(\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_{\text{IRM}})$, either, since $f_{\text{IRM}}$ is dominated by $f_1$. Therefore, the failures of IRM$_{\mathcal{S}}$ and IRMv1 can be understood as using improper objectives whose Pareto front does not contain the desired solution. Thus, choosing proper objectives is of great importance. The ideal objectives should constitute a Pareto front that contains the desired solutions.

**Pareto Robust Risk Minimization for IRM.** In pursuit of proper Pareto objectives, we resort to Distributionally Robust Optimization (Namkoong & Duchi, 2016) that aims to minimize the maximal error for any distribution in a convex hull formed under mixtures with positive weights of training distributions, or $\{\mathcal{D}_e\}_{e \in \mathcal{E}_{\text{tr}}}$ in our case. Bottou et al. (2019) argue that learning the invariance in IRM can extrapolate outside the convex hull by finding stationary points of $\sum_{e \in \mathcal{E}_{\text{tr}}} \lambda_e \mathcal{L}_e$ for some (possibly negative) $\{\lambda_e\}_{e \in \mathcal{E}_{\text{tr}}} \subseteq \mathbb{R}$. However, IRM$_{\mathcal{S}}$ and IRMv1 can fail to learn the invariant representations, weakening the extrapolation power. Nevertheless, we can introduce additional objectives to directly improve the extrapolation outside the convex hull. In particular, we exemplify the idea with REx (Krueger et al., 2021) that minimizes the exact maximal error on distributions outside the convex hull of $\{\mathcal{D}_e\}_{e \in \mathcal{E}_{\text{tr}}}$ to a certain distance, i.e., $\min \max_{e \in \mathcal{E}_{\text{tr}}} \lambda_e \mathcal{L}_e$ for $\{\lambda_e\}_{e \in \mathcal{E}_{\text{tr}}} \geq -\beta$ and $\sum_{e \in \mathcal{E}_{\text{tr}}} \lambda_e = 1$, where $\beta \geq 0$ is a hyperparamter to indicate the extrapolation distance. Krueger et al. (2021) further translates the objective into the variance of $\{\mathcal{L}_e\}_{e \in \mathcal{E}_{\text{tr}}}$. In Fig. 4, we visualize the variance of the failure case in Fig. 1 with respect to $\mathcal{L}_1, \mathcal{L}_2$ implemented in MSE. It can be found

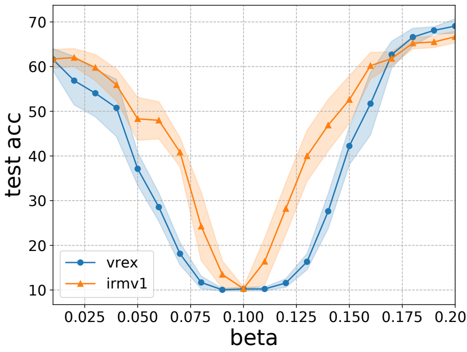


*Figure 5.* Drawbacks of Eq. 6 in practice.

that, $f_{\text{IRM}}$ lies in low variance region. Similarly, in Fig. 3, the zero variance solutions (shown in blue line) points out the underlying $f_{\text{IRM}}$ beyond the Pareto front. In other words, incorporating $\text{var}(\{\mathcal{L}_e\}_{e\in\mathcal{E}_{\text{tr}}})$ into objectives can relocate $f_{\text{IRM}}$ onto the Pareto front, which can be formulated as

$$\min_{\varphi}(\mathcal{L}_1, ..., \mathcal{L}_{|\mathcal{E}_{\text{tr}}|}, \mathcal{L}_{\text{IRM}}, \text{var}(\{\mathcal{L}_e\}_{e\in\mathcal{E}_{\text{tr}}}))^T. \qquad (6)$$

By resolving a large class of failure cases of $\text{IRM}_{\mathcal{S}}$, Eq. 6 is more powerful than $\text{IRM}_{\mathcal{S}}$ in learning the invariance.

**Drawbacks of Robust Minimization in Practice.** After showing REx (Krueger et al., 2021) can help avoiding the failure cases of $\text{IRM}_{\mathcal{S}}$, a natural question is that, does $\mathcal{L}_{\text{IRM}}$ remain necessary? We show the answer is "Yes". In Fig. 5, we use a modified example of $\mathcal{E}_{\text{tr}} = \{(0.25, 0.1), (0.25, \beta)\}$ with ColoredMNIST (Arjovsky et al., 2019), where we change the variance between two environments through different $\beta$. It can be found that, as the variance between two environments getting closer, the performance of REx (Krueger et al., 2021) (implemented as vrex) drops more sharply than IRMv1 as well as IRMv1 plus an additional regularization of vrex (denoted as irmx). The main reason is that, as the variation of spurious signals in two environments tend to be smaller, the gradient signal of $\text{var}(\{\mathcal{L}_e\}_{e\in\mathcal{E}_{\text{tr}}})$ tends to vanish, while the signals from $\mathcal{L}_{\text{IRM}}$ maintains. This issue can be more serious in stochastic gradient descent where the estimates of the variance of $\{\mathcal{L}_e\}_{e\in\mathcal{E}_{\text{tr}}}$ in minibatches tends to be noisy, leading to weaker signals.

**Pareto Invariant Risk Minimization.** To summarize, by formulating IRM as a MOO problem in Eq. 5, we find that the failures of $\text{IRM}_{\mathcal{S}}$ and IRMv1 are caused by using improper objectives for optimization. Nevertheless, we can introduce additional guidance from more robust objectives such REx (Krueger et al., 2021) to narrow the gap between $\text{IRM}_{\mathcal{S}}$ and IRM by resolving the failure cases of $\text{IRM}_{\mathcal{S}}$. Moreover, leaving $\mathcal{L}_{\text{IRM}}$ in the objectives also keeps better gradient signals in practice where the signals from REx can vanish. Thus, we arrive the following MOO formulation, called **PA**reto **I**nvariant **R**isk Minimization (`PAIR`),

$$\min_{\varphi}(\sum_e \mathcal{L}_{|\mathcal{E}_{\text{tr}}|}, \mathcal{L}_{\text{IRM}}, \text{var}(\{\mathcal{L}_e\}_{e\in\mathcal{E}_{\text{tr}}}))^T, \qquad (7)$$

*Table 1.* Accuracy (percent) on different modified ColoredMNIST

| Method | CMNIST-10 | CMNIST-25 | Avg |
|---|---|---|---|
| ERM | $73.3 \pm 0.9$ | $17.1 \pm 0.9$ | 45.2 |
| IRMv1 | $76.8 \pm 3.2$ | $67.3 \pm 1.9$ | 72.1 |
| V-REx | $82.9 \pm 1.3$ | $\mathbf{68.6 \pm 0.7}$ | 75.8 |
| IRMx | $81.6 \pm 2.0$ | $65.8 \pm 2.9$ | 73.7 |
| **`PAIR` (ours)** | $\mathbf{85.2} \pm 0.9$ | $68.4 \pm 1.1$ | **76.8** |
| Grayscale oracle | $86.5 \pm 0.3$ | $72.2 \pm 0.2$ | 79.4 |
| Optimum | 90 | 75 | 82.5 |
| Chance | 50 | 50 | 50 |

where we merge $(\mathcal{L}_1, ..., \mathcal{L}_{|\mathcal{E}_{\text{all}}|})$ into the summation for reducing the search dimensions over the Pareto front. To guarantee the achievability, we can leverage the off-the-shelf MGDA algorithms. The reason is that, in practice, $\mathcal{L}_{\text{IRM}}$ and $\text{var}(\{\mathcal{L}_e\}_{e\in\mathcal{E}_{\text{tr}}})$ will often converge to some $\epsilon \geq 0$ due to the potential noises of invariant features (Ahuja et al., 2021b; Kamath et al., 2021). Using scalarization scheme to combine the objectives with scalars can lead to sub-optimal solutions when the desired solutions lie in the concave part of the Pareto front (Boyd & Vandenberghe, 2014; Lin et al., 2019). Therefore, we use EPO solver (Mahapatra & Rajan, 2020) to find descent directions at each step. By specifying high preferences for $\mathcal{L}_{\text{IRM}}$ and $\text{var}(\{\mathcal{L}_e\}_{e\in\mathcal{E}_{\text{tr}}})$ to EPO, we can obtain the desired solutions in practice.

## 4. Experiments

We modify the ColoredMNIST dataset (Arjovsky et al., 2019) to verify our findings and the effectiveness of `PAIR`. Specifically, as the label flipping probability and colored probability in ColoredMNIST correspond to $\alpha_e$ and $\beta_e$ in two-bit environment (Eq. 4), respectively, we conduct experiments with two variants of ColoredMNIST, i.e., **CMNIST-10**: $\{(0.1, 0.25), (0.1, 0.2)\}$, and **CMNIST-25**: $\{(0.25, 0.10), (0.25, 0.20)\}$.We use the hyperparameter and optimization settings as in IRM (Arjovsky et al., 2019), and compare `PAIR` with ERM (Vapnik, 1991), IRMv1 (Arjovsky et al., 2019), V-REx (Krueger et al., 2021). We also construct a strong baseline IRMx where the variance penalty and $\mathcal{L}_{\text{IRM}}$ are simply added up with the same objective weights. Experimental results are shown in Table 1. It can be found that `PAIR` can significantly improve over IRMv1 to effectively capture the invariance across all environment settings, while the simple combination IRMx cannot yield satisfactory performance, confirming the drawbacks of linear scalarization (Boyd & Vandenberghe, 2014). We provide more results and discussions in Appendix C.2.

## 5. Conclusion

In this work, we provided a new understanding of the differences between $\text{IRM}_{\mathcal{S}}$, IRMv1 and IRM using MOO. We revealed that using the improper optimization objectives in $\text{IRM}_{\mathcal{S}}$ and IRMv1 is the main reason for their failures. Thus, we proposed `PAIR`, that offers new capacity for incorporating additional guidance from more robust objectives. We showed that, with additional guidance of REx, PAIR can effectively relocate and approach the invariant predictors onto the Pareto front. We hope PAIR can shed some light on understanding and improving the trade-off between ERM objective and OOD objective during the optimization. More discussions are deferred to Appendix A.

## References

Ahuja, K., Caballero, E., Zhang, D., Gagnon-Audet, J.-C., Bengio, Y., Mitliagkas, I., and Rish, I. Invariance principle meets information bottleneck for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, 2021a.

Ahuja, K., Wang, J., Dhurandhar, A., Shanmugam, K., and Varshney, K. R. Empirical or invariant risk minimization? a sample complexity perspective. In *International Conference on Learning Representations*, 2021b.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint*, arXiv:1907.02893, 2019.

Aubin, B., Słowik, A., Arjovsky, M., Bottou, L., and Lopez-Paz, D. Linear unit-tests for invariance discovery. *arXiv preprint arXiv:2102.10867*, 2021.

Beery, S., Horn, G. V., and Perona, P. Recognition in terra incognita. In *Computer Vision European Conference, Part XVI*, volume 11220, pp. 472–489, 2018.

Bengio, Y., Deleu, T., Rahaman, N., Ke, N. R., Lachapelle, S., Bilaniuk, O., Goyal, A., and Pal, C. J. A meta-transfer objective for learning to disentangle causal mechanisms. In *International Conference on Learning Representations*, 2020.

Bottou, L., Arjovsky, M., Gulrajani, I., and Lopez-Paz, D. Learning representations using causal invariance. Keynote in International Conference on Learning Representations, 2019. URL https://leon.bottou.org/talks/invariances.

Boyd, S. P. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2014.

Chen, Y., Zhang, Y., Yang, H., Ma, K., Xie, B., Liu, T., Han, B., and Cheng, J. Invariance principle meets out-of-distribution generalization on graphs. *arXiv preprint arXiv:2202.05441*, 2022.

Creager, E., Jacobsen, J., and Zemel, R. S. Environment inference for invariant learning. In *International Conference on Machine Learning*, volume 139, pp. 2189–2200, 2021.

DeGrave, A. J., Janizek, J. D., and Lee, S. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Désidéri, J.-A. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5):313–318, 2012.

Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R. S., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2 (11):665–673, 2020.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: convergence and generalization in neural networks. In *Annual ACM SIGACT Symposium on Theory of Computing*, pp. 6, 2021.

Ji, Y., Zhang, L., Wu, J., Wu, B., Huang, L.-K., Xu, T., Rong, Y., Li, L., Ren, J., Xue, D., Lai, H., Xu, S., Feng, J., Liu, W., Luo, P., Zhou, S., Huang, J., Zhao, P., and Bian, Y. DrugOOD: Out-of-Distribution (OOD) Dataset Curator and Benchmark for AI-aided Drug Discovery – A Focus on Affinity Prediction Problems with Noise Annotations. *arXiv preprint*, arXiv:2201.09637, 2022.

Kaisa, M. *Nonlinear Multiobjective Optimization*. Springer, 1999.

Kamath, P., Tangella, A., Sutherland, D., and Srebro, N. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pp. 4069–4077, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning,*, volume 139, pp. 5637–5664, 2021.

Koyama, M. and Yamaguchi, S. Out-of-distribution generalization with maximal invariant predictor. *arXiv preprint arXiv:2008.01883*, 2020a.

Koyama, M. and Yamaguchi, S. Out-of-distribution generalization with maximal invariant predictor. *arXiv preprint*, arXiv:2008.01883, 2020b.

Krueger, D., Caballero, E., Jacobsen, J., Zhang, A., Binas, J., Zhang, D., Priol, R. L., and Courville, A. C. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, volume 139, pp. 5815–5826, 2021.

Lin, X., Zhen, H., Li, Z., Zhang, Q., and Kwong, S. Pareto multi-task learning. In *Advances in Neural Information Processing Systems*, pp. 12037–12047, 2019.

Lv, F., Liang, J., Gong, K., Li, S., Liu, C. H., Li, H., Liu, D., and Wang, G. Pareto domain adaptation. In *Advances in Neural Information Processing Systems*, 2021.

Ma, P., Du, T., and Matusik, W. Efficient continuous pareto exploration in multi-task learning. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 6522–6531, 2020.

Mahapatra, D. and Rajan, V. Multi-task learning with user preferences: Gradient descent with controlled ascent in pareto optimization. In *International Conference on Machine Learning*, pp. 6597–6607, 2020.

Nagarajan, V., Andreassen, A., and Neyshabur, B. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*, 2021.

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems*, pp. 2208–2216, 2016.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Rame, A., Dancette, C., and Cord, M. Fishr: Invariant gradient variances for out-of-distribution generalization. *arXiv preprint arXiv:2109.02934*, 2021.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

Rosenfeld, E., Ravikumar, P. K., and Risteski, A. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.

Sagawa*, S., Koh*, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

Sener, O. and Koltun, V. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, pp. 525–536, 2018.

Shi, Y., Seely, J., Torr, P., N, S., Hannun, A., Usunier, N., and Synnaeve, G. Gradient matching for domain generalization. In *International Conference on Learning Representations*, 2022.

Sutskever, I., Martens, J., Dahl, G. E., and Hinton, G. E. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pp. 1139–1147, 2013.

Vapnik, V. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, pp. 831–838, 1991.

Zhai, R., Dan, C., Kolter, J. Z., and Ravikumar, P. Understanding why generalized reweighting does not improve over ERM. *arXiv preprint arXiv:2201.12293*, 2022.

Zhang, J., Lopez-Paz, D., and Bottou, L. Rich feature construction for the optimization-generalization dilemma. *arXiv preprint arXiv:2203.15516*, 2022.

Zhou, K., Jin, Y., Ding, Q., and Cheng, J. Amortized nesterov's momentum: A robust momentum and its application to deep learning. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 211–220, 2020.

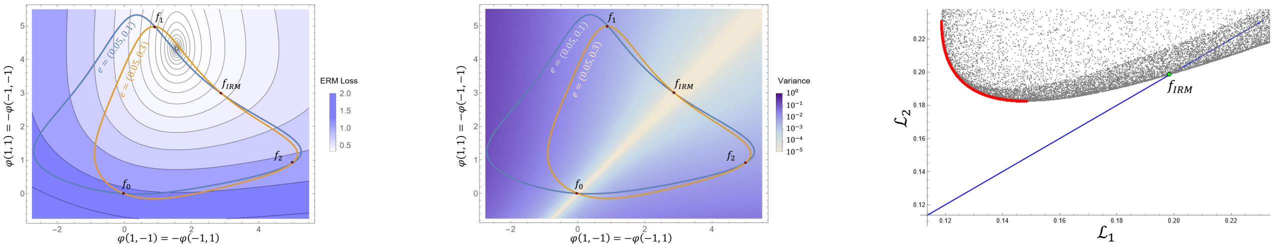
# A. Related Work

**Optimization Dilemma in OOD Algorithms.** Several recent works also notice the optimization dilemma in OOD algorithms, specifically, the trade-off between discovering the statistical correlations (i.e., ERM) and preventing the usage of spurious correlations (e.g., IRM). Empirically, Gulrajani & Lopez-Paz (2021) find that, with careful hyperparameter tuning and evaluation setting, many OOD algorithms cannot outperform ERM in domain generalization, demonstrating the difficulties in finding the desirable OOD solutions. Sagawa* et al. (2020); Zhai et al. (2022) find that, regularization on ERM, or sacrificing ERM performance, is usually needed for achieving satisfactory OOD performance, which aligns with our findings through Pareto front as Fig. 3 and Fig. 6(c). Zhang et al. (2022) find that, the performance of OOD algorithms largely relies on choosing a proper pretraining epochs which aligns with our findings in Fig. 2, hence propose to construct a ready-to-use features for stable OOD generalization performance. Orthogonal to Zhang et al. (2022), we focus on developing better optimization scheme for OOD algorithms, including choosing the proper objectives and the achievability of the invariant predictors.

**MOO for Multi-Task Learning.** As it is usual that we cannot find a global optimal solution for all objectives in practice, hence Pareto optimal solutions are of particular value. The multiple-gradient descent algorithm (MGDA) is one of the commonly used approaches to efficiently find the Pareto optimal solutions (Désidéri, 2012) but limited to low-dimensional data. Sener & Koltun (2018) then resolve the issue and apply MGDA to high-dimensional multi-task learning scenarios, where the objective conflicts may degenerate the performance when using linear scalarization. As pure MGDA cannot find a Pareto optimal solution specified by certain objective preferences, Lin et al. (2019); Ma et al. (2020) propose efficient methods to explore the Pareto set. Mahapatra & Rajan (2020) propose EPO to find the exact Pareto optimal solution with the specified objective preferences. However, they are unable to solve OOD generalization, when without a suitable Pareto front. Besides, Lv et al. (2021) propose ParetoDA to use guidance of validation loss based on the data that has the identical distribution to test distribution, to trade-off the conflicts in domain adaption objectives. However, there can be multiple test domains and the data are usually unavailable in OOD generalization, limiting the adoption of ParetoDA.

# B. More Details on Figures in Sections 2 and 3

We plot Fig. 1 and Fig. 4 based on the Mathematica code provided by Kamath et al. (2021), where we focus on the odd predictors due to the symmetry in two-bit environments, i.e., predictors satisfying $\varphi(1,-1) = -\varphi(-1,1)$ and $\varphi(1,1) = -\varphi(-1,-1)$. Besides, Fig. 1, Fig. 4 and Fig. 3 are implemented in MSE loss. Their Logistic loss counterparts are given as Fig. 6.



(a) Failure case under Logistic loss.

(b) Variance distribution under Logistic loss.

(c) Pareto Front under Logistic loss.

*Figure 6.* Counterparts of Fig. 1, Fig. 4 and Fig. 3 implemented in Logistic loss.

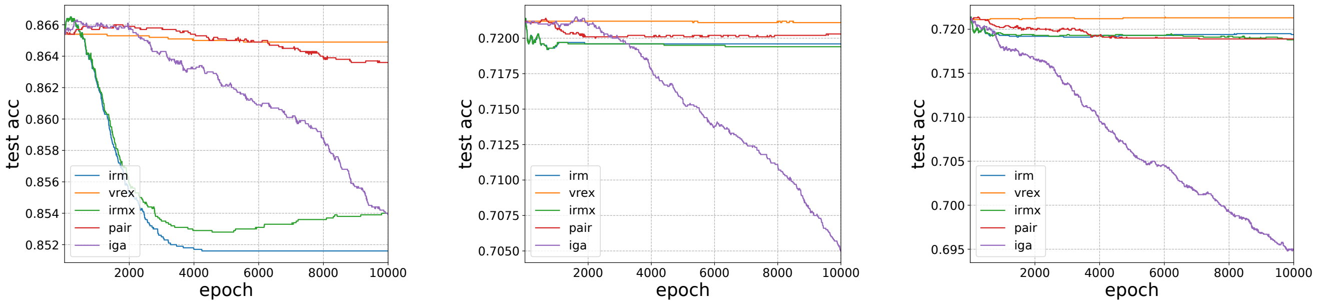
# C. More Details on Experiments

## C.1. Experimental Settings

We follow the evaluation settings as IRM (Arjovsky et al., 2019) and the test-domain selection as DomainBed (Gulrajani & Lopez-Paz, 2021) when conducting the experiments. Specifically, we use a 4-Layer MLP with a hidden dimension of 256. By default, we use Adam (Kingma & Ba, 2015) optimizer with a learning rate of $1e-3$ and a weight decay of $1e-3$ to train the model with 500 epochs and select the last epoch as the output model for each hyperparamter setting. We choose the final model as the one that maximizes the accuracy on the validation that share the same distribution as test domain. We then do grid search for the other hyperparameters. For pretraining epochs, we search from $\{0, 50, 100, 150, 200, 250\}$. For OOD penalty, we search from $\{1e1, 1e2, 1e3, 1e4, 1e5\}$. Besides, we will refresh the history in Adam optimizer when the pretraining finishes, as suggested by Gulrajani & Lopez-Paz (2021). While for PAIR, we use SGD with a momentum of $0.9$ (Sutskever et al., 2013) after pretraining to avoid the interference of Adam to the gradient direction and convergence of EPO (Mahapatra & Rajan, 2020) solver. Moreover, we empirically find that SGD requires larger learning rate (we use $0.1$) for approaching the direction. This is because of the design in EPO solver that it first fits to the preference direction then does the "pure" gradient descent, while the intrinsically conflicting directions pointed by the objectives can make the loss surface more steep. We will leave in-depth understanding of the above phenomenon and more sophisticated optimizer design in more complex tasks and network architectures to future works (Zhou et al., 2020).

### C.2. More Experimental Results and Discussions

We also conduct experiments with "perfect" initializations for different methods, to check whether the OOD constraints can enforce the invariance, following Zhang et al. (2022). Besides the OOD methods used in the paper, we also include another OOD method IGA (Koyama & Yamaguchi, 2020a) to give a more comprehensive overview of their performances with "perfect" initialization. All methods are initialized with a ERM model learned on gray-scale CMNIST data which is expected to learn to use digit shapes in the image to make predictions. The learning rate is $1e-3$ and the penalty weight is $1e5$. Different from Zhang et al. (2022), we use SGD to optimize the models, as Adam would generate larger step sizes when the gradients continue to be within a small range under the "perfect" initialization. Results are shown as in Fig. 7.



(a) "Perfect" initialization on CMNIST-10.
(b) "Perfect" initialization on CMNIST-11.
(c) "Perfect" initialization on CMNIST-25.

*Figure 7.* OOD performances with "Perfect" initializations.

It can be found that, in CMNIST-10, IRM, IRMx and IGA cannot enforce the invariance while V-REx and PAIR maintain the invariance, which is consistent to our previous findings. Moreover, IGA fails to maintain the invariance in CMNIST-11 and CMNIST-25, demonstrating the relatively low robustness of IGA objective. Besides, V-REx consistently maintain the invariance even in CMNIST-11, for the reason that the gradient signals of variance in "perfect" initialization tend to vanish. In contrast, PAIR improve over both IRM and IRMx to maintain the invariance, confirming the effectiveness of PAIR.