A Technical Survey on Invariant Risk Minimization and Related Methods for Out-of-Distribution Generalization

Saksham Jain

19 March, Winter 2023

Abstract

Invariant Risk Minimization (IRM) has emerged as a promising technique for learning domain-invariant representations to enable out-of-distribution generalization in machine learning. In this technical survey, we review IRM and its extensions, as well as some related methods that aim to achieve similar goals. We provide background information on each method, identify the formal and informal promises made, discuss their successes and failures, and illuminate possible future directions in the field.

1 Introduction

Machine learning models often face challenges when it comes to generalizing to out-of-distribution (OOD) data. Invariant Risk Minimization (IRM) was introduced by Arjovsky et al. [2019] as a method to learn domain-invariant representations for OOD generalization. This survey aims to provide an overview of IRM and related methods, their successes and failures, and the state-of-the-art in the field. We first describe the problem statement and desired goals, followed by a discussion of the related literature.

2 Problem Statement

Causality-inspired literature on OOD generalization suggests that real-world datasets share causal mechanisms that remain constant across different environments. Some features pertain to these invariant mechanisms, while others are spurious. Learning invariant correlations across multiple training distributions from different environments is related to learning the causal structures underlying the data, which can enable OOD generalization. In this project, we will conduct a technical survey of methods inspired by and related to IRM, focusing on consolidating works demonstrating their successes and failures and identifying the state-of-the-art.

3 Desiderata

This survey aims to achieve the following goals:

- Develop background on IRM and subsequent approaches to domain-invariant representation learning for OOD generalization.
- Identify the formal and informal promises made by these methods, the settings/problems in which they fail and why, and potential future directions.

4 Overview

We briefly introduce the literature that will be discussed in this survey. IB-IRM [Ahuja et al., 2021] and V-REx [Krueger et al., 2021] build upon the initial IRM, while IGA [Koyama and Yamaguchi, 2020] and EIIL [Creager et al., 2021] follow related but tangential ideas. Works by Rosenfeld et al. [2020], and Aubin et al. [2021] contain further analyses of the ideas behind IRM and formulate challenging linear problems for invariance-learning approaches.

5 Invariant Risk Minimization (IRM)

IRM is based on the idea that real-world datasets share causal mechanisms that remain constant across different environments. The method seeks to learn a representation that captures these invariant causal mechanisms, allowing for better generalization to new and unseen environments. The intuition is that only the causes of the target are the ideal representation for prediction, and thus, we can find the 'invariant' predictor that achieves the lowest training risk, and is composed of a linear representation map and a representation-level classifier that is simultaneously optimal for all environments.

Given a set of environments indexed by $1, 2, \ldots, m$, the goal of IRM is to find a predictor that performs well across all environments by optimizing the empirical risk within each environment. IRM achieves this by minimizing the worst-case risk across environments, which can be formulated as the following optimization problem:

$$\min_{f:\mathcal{X}\to\mathcal{Y}} \max_{e\in[m]} \mathbb{E}_{(X,Y)\sim E_e}[L(f(X), Y)], \tag{1}$$

where the data {X,Y} hold their usual meaning, L is an appropriate loss function (e.g. MSE for regression, and 0-1 loss for classification), and f is the predictor learned by the model. In practice, IRM does this by solving:

$$\min_{f:\mathcal{X}\to\mathcal{Y}} \sum_{e\in[m]} R^e(1 \cdot f) + \lambda \cdot \mathbb{D}(w, f, e)\Big|_{w=1.0}, \tag{2}$$

where $R^e(\phi)$ is the risk of function ϕ under environment E_e, f is the invariant classifier, $\mathbb{D}(w, f, e)$ is a metric for how close w is to minimizing $R^e(w \cdot f)$ (e.g. $||\nabla_w R^e(w \cdot f)||^2$) when $w \overset{(set)}{=} 1.0$, and λ is a hyperparameter balancing the predictive power (the second term), and invariance $(1 \cdot f)$.

5.1 Theoretical Guarantees and Informal Promises

The theoretical guarantees of IRM are related to its ability to learn domain-invariant representations that generalize well across diverse environments. By focusing on learning the causal structures underlying the data, IRM is expected to be less sensitive to the variations in the environment, such as changes in the distribution of spurious features.

Under the assumption that the causal mechanisms shared across environments are sufficient for making accurate predictions, IRM claims to be theoretically guaranteed to learn a predictor that generalizes well to new environments since the learned predictor relies on invariant causal mechanisms that remain constant across different environments, rather than on spurious correlations that may not generalize well. This approach provably generalizes OOD in linear regression tasks.

The informal promises of IRM are related to its ability to discover and exploit the causal structures underlying the data. By learning invariant representations that capture these causal

mechanisms, IRM is expected to enable better generalization to out-of-distribution samples and new environments. Moreover, IRM promises to be more robust to changes in the environment, as the method focuses on learning invariant causal mechanisms rather than spurious correlations.

5.2 Failure Conditions and Limitations

However, the theoretical guarantees and informal promises of IRM can be violated under certain conditions, as discussed by Rosenfeld et al. [2020] and Aubin et al. [2021]. Some of these conditions include:

- **Linear Classification:** A crucial failure case – when the invariant features capture the full information about the target labels – in the linear classification setting was identified by [Ahuja et al., 2021].
- **Significant changes in data distribution:** When the data distribution varies significantly between environments, the causal mechanisms shared across environments may not be sufficient for making accurate predictions. In this case, the performance of IRM can be suboptimal, as the method may not be able to capture the necessary information to generalize well to new environments.
- **Limited sample size:** When the sample size is limited, the empirical risk minimization principle underlying IRM can lead to overfitting, as the method may fit the noise in the data rather than the true underlying causal mechanisms.
- **Entangled causal mechanisms:** In some cases, the causal mechanisms underlying the data may be entangled with spurious features, making it challenging for IRM to learn invariant representations that separate the two. In such scenarios, IRM's performance may be compromised, as the method may not be able to effectively disentangle the causal mechanisms from the spurious features.
- **Nonlinear cases:** The objective function of IRM becomes non-convex in non-linear cases, making it more difficult to find the global minimum. This can result in convergence to local minima, leading to suboptimal solutions.
- **Hyperparamter sensitivity:** The IRM method is sensitive to hyperparameter choices, such as the learning rate, the gradient penalty weight, and the choice of the optimizer. In non-linear cases, it can be even more challenging to find the right hyperparamter settings.

6 V-REx

Krueger et al. [2021] introduced (REx), which is a method designed to improve the generalization of models to unseen environments by penalizing models that are overly specialized to the training environments.

The main idea behind risk extrapolation is to train a model by optimizing a regularized objective that encourages it to perform well not only on the training environments but also on extrapolated environments, which can be thought to be mixtures of training environments (i.e. affine combinations). This approach is based on the assumption that if a model can perform well on a range of environments created through extrapolation, it is more likely to generalize well to entirely new, unseen environments, where the magnitude of the mixing may be much larger. In essence, this

method works by reducing the average training risk while simultaneously increasing the similarity of the training risks.

The setup of the OOD problem is still the same as in 1. Krueger et. al first introduce Minimax-REx which performs robust optimization [Sagawa et al., 2020], over a perturbation set of affine combinations of training risks with lower-bounded coefficients, but they recommend using a simpler and more effective algorithm – V-REx, which, in practice, solves this optimization problem:

$$\min_{f:\mathcal{X}\to\mathcal{Y}} \sum_{e\in[m]} R^e(f) + \gamma \mathrm{Var}(\{R^1(f), \ldots, R^m(f)\}), \tag{3}$$

where the first term is from the ERM formulation, the second term encourages the risks across environments towards equality, and γ is the hyperparameter that balances these two objectives.

6.1 Theoretical Guarantees and Informal Promises

Kreuger et. al prove that under certain assumptions [Peters et al., 2016] and under interventional shift with at least three distinct environments, if the REx condition is satisfied for the structural equation model, the algorithm recovers the invariant predictor that is optimal across all environments.

They also prove that a predictor that satisfies REx over all interventions that do not change the mechanism of the target (i.e., equality of risk across domains under homoskedasticity), uses the invariant predictor as the predictive distribution over the input features.

REx promises better generalization performance on out-of-distribution data compared to standard Empirical Risk Minimization (ERM) by minimizing the in-distribution risk while simultaneously equalizing extrapolated risks during optimization. The paper provides empirical evidence showing that the V-REx algorithm can outperform ERM and IRM on synthetic and real-world datasets under both covariate (of course, under the implicit support overlap assumption) and interventional shifts in terms of out-of-distribution generalization

The risk extrapolation condition, i.e. reduced training risks and closeness of the risks under all (extrapolated) environments promises an increase in OOD generalization, empirically verified on various tasks including classification, reinforcement learning, etc. Further, it is easy to enforces by simultaneously minimizing the variance of the per-environment risks in the regularization term of the overall objective.

6.2 Failure Conditions and Limitations

This algorithm also shares some of the limitations discussed in Section 5.2 such as the optimization issues caused by hyperparameter sensitivity, especially since $\gamma \in (0, \infty)$.

Another critical failure condition is either the violation of the implicit support overlap assumption, or a significant covariate shift for classification tasks. While REx claims good performance under covariate shift (and indeed outperforms IRM in several cases), this is by no means general, especially with increase in higher dimensions and/or number of environments.

7 Information Bottleneck IRM (IB-IRM)

IB-IRM Ahuja et al. [2021] is an extension of the original Invariant Risk Minimization (IRM) that incorporates the information bottleneck principle (i.e. learn a representation that, while preserving all the relevant information about the target, compresses the input as much as possible) with the invariance principle as formulated by IRM.

IB-IRM thus adds a regularization term to the IRM objective that minimizes the mutual information between the learned representations and the input features. The intuition is that the representation map that focuses on the invariant features is the one that has the least entropy and also achieves the lowest error. I.e., IB-IRM finds the optimal predictor by picking the predictors with the least entropy (IB part) among all the highly predictive invariant predictors (IRM part). The setup of the OOD problem is still the same as in 1. But IB-IRM, in practice, solves this optimization problem:

$$\min_{f:\mathcal{X}\to\mathcal{Y}} \sum_{e\in[m]} R^e(1\cdot f) + \lambda \cdot \mathbb{D}(w, f, e)\Big|_{w=1.0} + \beta \mathrm{Var}(f), \tag{4}$$

where the first two terms are as described in 2, and the third term includes a hyperparamter β and variance of the predictor. This is a surrogate for the unconditional (of environment) entropy $h(f)$, since for all conitnous random variables, the Gaussian has the highest entropy, and the Gaussian entropy increases with its variance.

7.1 Theoretical Guarantees and Informal Promises

The authors reexamine and verify the theory behind the invariance principle in IRM, that allows it to generalize OOD in the linear regression setting. In the linear classification setting, the authors prove that the support overlap assumption on distribution shifts (i.e. the train and test distribution must share the same support despite being different) for invariant features is crucial for OOD generalization. The intuition behind this is: if the support of the invariant features are strictly separable by the true (labelling) hyperplane, then another valid hyperplane can be found that could also have generated the data. This means that the data from the region where these two hyperplanes disagree cannot be classified by *any* algorithm in the linear classification setting. However, this assumption is not needed in the linear regression setting, a fact that is corroborated by the provable results in [Arjovsky et al., 2019], where even in the absence of this assumption, OOD generalization is guaranteed.

Another important result from the paper is that even under the above overlap assumptions on the distribution shift, the invariance principle [Arjovsky et al., 2019] by itself is not enough. Ahuja et al. [2021] show that in the classification setting in which the invariant features *fully* capture the label information, ERM and IRM provably fail to generalize. However, by combining the invaraince principle with the information bottleneck (IB) constraint, OOD generalization becomes possible for both cases – in the presence of fully informative invariant features, and in their absence.

Interestingly, another result from the paper is that in the linear classification setting, support overlap is not a necessary condition for spurious features, but approaches such as IRM may fail in its absence.

Finally, the theoretical analyses also guarantee that even under covariate shift and the presence of a common labeling function, support overlap for invariant features is both necessary and *sufficient* in many cases to achieve OOD generalization.

7.2 Failure Conditions and Limitations

Since the first two terms of the objective function are from IRM, the guarantees and promises of IB-IRM can be violated under certain conditions. These conditions include a high degree of covariate shift, limited sample size (this is actually quite stark - the method requires a large number of samples (or from a large number of distinct environments) in order to generalize well OOD, especially when invariant and spurious features are more closely entangled). However, IB-IRM is expected to be

more robust under these conditions due to the information bottleneck constraint acting when the invariance principle by itself fails. E.g., it is possible for it to generalize OOD in classification under *both* the fully informative invariant features setting and the partial one, whereas it is impossible for IRM.

8 Inter-environmental Gradient Alignment (IGA)

Koyama and Yamaguchi [2020] show that the out-of-distribution (OOD) generalization problem can be formulated as an invariance problem across different environments under certain conditions. The goal is to learn a predictor function that remains invariant across these environments and is also expressive enough to capture the underlying structure of the data (hence, a maximal invariant predictor). The causal structure underlying the task is assumed to be shared across environments.

The authors propose the Inter-environmental Gradient Alignment (IGA) algorithm to learn this maximal invariant predictor by minimizing the inter-environmental gradient discrepancy – the differences in gradients of the predictor's risk across environments (i.e. aligning the inter-environmental gradients).

The algorithm first computes the risk gradients (w.r.t the weights) for each environment, then computes the inter-environmental gradient discrepancy, and minimizes this simultaneously with the average training risk. The predictor is updated using gradient-based optimization methods (e.g., stochastic gradient descent) by taking a step towards minimizing both components of the objective. The optimization problem can be written as:

$$\min_{\theta} \sum_{e\in[m]} R^e(f_\theta) + \alpha \sum_{e\in[m]} \left|\left| \nabla_\theta R^e(f_\theta) - \frac{1}{m} \sum_{e\in[m]} \nabla_\theta R^e(f_\theta) \right|\right|^2, \tag{5}$$

where the first term is the average training risk and the second term is the inter-environmental gradient discrepancy. This formulation is claimed to agree with the general formulation of the IRM objective (in contrast to the practical IRM objective (2) for categorical target). α is the hyperparameter that balances the expressiveness and invariance of the predictor.

8.1 Theoretical Guarantees and Informal Promises

The paper provably shows that under the controllability condition (intuitively, that for an invariant representation, there exists an environment in which the target distribution is the same regardless of if conditioned on the representation or the input), the invariant predictor solves the OOD problem (1).

They also guarantee that under certain assumptions (i.e. the above condition is satisfied for at least one invariant representation, and that all other invariant representations can be written as a function of a common one), then the invariance problem can be reformulated as an InfoMax [Linsker, 1988] objective with the invariance constraint, whose solution gives the maximal invariant predictor (MIP, that agrees with the predictor based on the common invariant representation) which solves the OOD problem.

Finally, they show that the MIP problem can be further reparametrized to achieve a solution using the IGA algorithm, whose objective actually upper bounds the OOD objective (1) and so agrees with the intuition behind all the invariant learning algorithms we have seen so far.

They informally promise better (than IRM) OOD generalization, and empirically demonstrate the effectiveness of the IEGA algorithm through a series of experiments on synthetic and real-world

datasets. Their theoretical results also claim that IRM-based techniques show promise for nonlinear problems.

8.2 Failure Conditions and Limitations

Since the first term of the objective is analogous to the ERM objective, this method also suffers from some of the limitations described in Section 5.2. Although, since the theory for gradient descent is well-established, the actual optimization of IGA is expected to be nicer than IRM and its extensions, even in nonlinear problems. Apart from that, the theory for this method for OOD generalization breaks down when:

- **Controllability condition is not satisfied:** As the paper shows, the controllability condition is necessary for the formulation of the OOD problem as an invariance problem.
- **Other conditions for the MIP problem are unsatisfied:** Even if the controllability condition is satisfied it may not be appropriate to reformulate the invariance problem as an InfoMax objective.

9 Environment Inference for Invariant Learning (EIIL)

Creager et al. [2021] introduced EIIL, a method that touches upon a valid criticism of most invariant-learning methods. The invariant learning methods we have surveyed so far work in the presence of environment labels, which are either known a priori or handcrafted e.g. using perturbation.

Creager et al. propose an algorithm that trains an invariant predictor in the absence of environment labels. It does so by splitting the training into two phases - the environment inference (EI) phase for environment assignment for each sample, and then the invariant learning (IL) phase which outputs the invariant predictor. It frames the invariant constraint in approaches such as IRM and REx as an *environment*-invariant constraint (EIC) since, for invariance, the learned representation must elicit the same conditional distribution over the targets for all environments. Then, in the EI phase, the environment assignment that maximally violates the EIC is found, by optimizing:

$$\mathbf{p}* = \arg\max_{\mathbf{p}} \left|\left| \nabla_{w|w=1.0} \left(\frac{1}{\sum_{i'} \mathbf{p}_{i'}(e)} \sum_{i} \mathbf{p}_i(e) L(w \cdot \tilde{f}(X_i), Y_i) \right) \right|\right|^2, \tag{6}$$

which is a soft relaxation of the regularization term in 2. $\tilde{f}$ is a reference model (i.e. a suboptimal predictor, e.g. using only ERM), $\mathbf{p}_i(e)$ is the probability of assignment of the i^{th} sample to the e^{th} environment, and L is the appropriate loss function. Then for the IL step, the per-environment risk is:

$$\tilde{R}^e(f, \mathbf{p}*) = \frac{1}{\sum_{i'} \mathbf{p}*_{i'}(e)} \sum_{i} \mathbf{p}*_i(e) L(f(X_i), Y_i), \tag{7}$$

and we can find the invariant predictor by using the objective function of any of the above invariant learning algorithms, replacing all instances of $R^e(f)$ with $\tilde{R}^e(f, \mathbf{p}*)$, and minimizing over f.

9.1 Theoretical Guarantees and Informal Promises

While there are no guarantees on OOD generalization in general, the method provably shows in the binary classification setting (support overlap 7.1 is an implicit assumption, although this goes unremarked in the paper) that the reference model that learns solely from the spurious features maximally violates the EIC and finds the environment partitions with the greatest contrast for the invariant learning step. However, when spurious features are not known a priori, the ERM-based reference model is not guaranteed to maximally violate EIC.

The approach promises that the method enables learning an invariant predictor without requiring explicit environment labels during training. The authors provide empirical evidence on synthetic and real-world datasets, demonstrating that the proposed method can indeed learn invariant predictors that generalize well to out-of-distribution data in various settings.

Other informal promises include empirical evidence for several claims. E.g., even if the reference model focuses on a mix of invariant and spurious features, EIIL may still find environment partitions that are effective for invariant learning. Moreover, even when handcrafted environment labels are known, the EI phase usually helps find better environment assignments for invariant learning.

9.2 Failure Conditions and Limitations

Since the IL step of this approach uses the objective from IRM, limitations such as hyperparamter sensitivity, and difficulties in optimization as described in Section 5.2 still hold (although it is empirically shown to be more robust to changes in the distribution). However, due to the EI step, there are other failure conditions:

- **Poor environment partitions:** EIIL is highly dependent on the reference model for its performance. Therefore, if the reference model fails to maximally violate EIC, (e.g. when spurious and invariant features are highly entangled), then the overall performance may suffer - since these environments are then passed down to the IL step.
- **Choice of reference model:** If spurious features are not known a priori, the reference model must be recovered from the data, and in non-simple settings, it becomes difficult to heuristically design. And as in the previous point, ERM-based models may not always give the best (or even good) environment partitions.

10 Collated Results

Table 1 shows the collated experimental results for the discussed methods. The results for the linear unit tests are collated from Aubin et al. [2021], and Ahuja et al. [2021] (for IB-IRM). The results for C-MNIST are collated from the reported results in each of the papers corresponding to each method.

11 Challenges and Future Directions

While the methods discussed above show promising results in learning invariant representations and enabling OOD generalization, they are not without limitations. Some challenges and future directions in this area include:

- Addressing the limitations when the data distribution varies significantly between environments, which may lead to suboptimal performance.

	ERM	IRMv1	IGA	IB-IRM	V-Rex	EIIL	Oracle
C-MNIST (Test accuracy)	16.84 ± 0.82	59.2 ± 0.011	62.0 ± 0.015	67.67 ± 1.78	68.7 ± 0.9	68.4 ± 2.7	-
Example 1 (MSE error)	13.36 ± 1.49	11.15 ± 0.71	17.47 ± 2.16	11.68 ± 0.90	*	*	10.42 ± 0.16
Example 1s (MSE error)	13.33 ± 1.49	11.06 ± 0.68	17.68 ± 3.31	11.74 ± 1.03	*	*	10.45 ± 0.19
Example 2 (Classification error)	0.42 ± 0.01	0.45 ± 0.00	0.45 ± 0.00	0.00 ± 0.00	*	*	0.00 ± 0.00
Example 2s (Classification error)	0.45 ± 0.01	0.45 ± 0.01	0.45 ± 0.01	0.06 ± 0.12	*	*	0.00 ± 0.00
Example 3 (Classification error)	0.48 ± 0.09	0.49 ± 0.07	0.47 ± 0.10	0.48 ± 0.07	*	*	0.00 ± 0.00
Example 3s (Classification error)	0.48 ± 0.07	0.49 ± 0.07	0.48 ± 0.09	0.49 ± 0.07	*	*	0.00 ± 0.00

Table 1: Shows experimental performance on the Color-MNIST experiment [Arjovsky et al., 2019], and the linear unit tests designed by Aubin et al. [2021]. The performance metrics are noted in parentheses beside the name of the experiment. * reflects the fact that these results are unavailable and that these experiments need to be run in the future.

- Uncovering and formalizing the implicit assumptions and informal promises. It seems like many underlying conditions may be, in fact, equivalent.
- Investigating the trade-off between invariance and discriminative power, as methods that focus solely on invariance may sacrifice their ability to discriminate between different target classes.
- Developing more efficient and scalable optimization techniques for learning invariant representations, as current methods may be computationally demanding, especially for large-scale datasets.
- Exploring the integration of causal reasoning with other machine learning paradigms, such as unsupervised learning, reinforcement learning, and transfer learning, to enhance generalization performance across a broader range of tasks and scenarios.

12 Conclusion

In this technical survey, we have reviewed IRM and related methods that aim to learn domain-invariant representations for out-of-distribution generalization. We provided background information on each method, identified the formal and informal promises made, discussed their successes and failures, and highlighted possible future directions in the field. As the demand for robust and generalizable machine learning models increases, the development and refinement of these methods will continue to play a crucial role in advancing the state-of-the-art in machine learning.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019. URL `https://arxiv.org/abs/1907.02893`.

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

Masanori Koyama and Shoichiro Yamaguchi. When is invariance useful in an out-of-distribution generalization problem? *arXiv preprint arXiv:2008.01883*, 2020.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Benjamin Aubin, Agnieszka Słowik, Martin Arjovsky, Leon Bottou, and David Lopez-Paz. Linear unit-tests for invariance discovery. *arXiv preprint arXiv:2102.10867*, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020.

R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988. doi: 10.1109/2.36.