Computationally Efficient Methods for Invariant Feature Selection with Sparsity

Jane Du[1] **Arindam Banerjee**[1]

[1]Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign

Abstract

Invariant Risk Minimization (IRM) (Arjovsky et al., 2020) proposes an optimization scheme that uses causal features to improve generalization. However, in most realizations, it does not have an explicit feature selection strategy. Prior investigation (Rosenfeld et al., 2020; Zhang et al., 2023) reveals failure cases when searching for causal features, and in light of these concerns, recent work has demonstrated the promise of using sparsity (Zhou et al., 2022; Fan et al., 2024) in IRM, and we make two specific contributions on that theme. First, for the original sparse IRM formulation, we present the first correct non-asymptotic analysis of the effectiveness of sparsity for selecting invariant features. We show that sparse IRM with L_0 constraints can select invariant features and ignore spurious and random features. We show that sample complexity depends polynomially on the number of invariant features and otherwise logarithmically on the ambient dimensionality. Second, we present the first invariant feature recovery guarantees with a computationally-efficient implementation of such sparse IRM based on iterative hard-thresholding. Prior methods are limited to combinatorially searching over the space of all sparse models, but we present a different loss function. We show this new optimization implies recovery of invariant features under standard assumptions. We present empirical results on standard benchmark datasets to demonstrate the effectiveness and efficiency of the proposed sparse IRM models.

1 INTRODUCTION

While overparameterized deep neural networks (DNN) are ubiquitous in the modern landscape of machine learning, the risk of memorization or other shortcuts leading to poor out-of-distribution performance remains an issue. The baseline assumption that training and test data are drawn i.i.d. from the same distribution, necessary for Empirical Risk Minimization (ERM) (Vapnik, 1991) to provide generalization guarantees, is arguably not true in many modern settings but also challenging to work around.

One approach to out-of-distribution (OOD) generalization is Invariant Causal Prediction (ICP) (Peters et al., 2016), in which data is drawn from different training environments, but the parent "causes" of the label, or target variable, are unchanging and independent of the environment. In other words, given the set of causal features, the conditional distribution of the label must be identical across multiple training environments. A popular line of work that has been developed in recent years is Invariant Risk Minimization (IRM) (Arjovsky et al., 2020), which aims to find an invariant data representation that induces a classifier that performs uniformly across all environments, including unseen test environments.

To take an illustrative example, consider classifying cows and camels. (Beery et al., 2018; Arjovsky et al., 2020) show that a model may be fooled into learning the background (green pastures and yellow desert respectively) over actual identifying features, thereby misclassifying, e.g., a cow on beach sand. Many following works build on this paradigm to adapt it to a variety of experimental and theoretical frameworks (Ahuja et al., 2022, 2020; Lin et al.; Creager et al., 2021). However, others have identified when it is impossible to provide formal guarantees in the nonlinear and linear regimes (Rosenfeld et al., 2020), which can lead to poorer generalization than unconstrained ERM (Kamath et al., 2021). Prior works demonstrate a large train-test gap in a variety of models and domain generalization datasets (Lin et al., 2022; Zhou et al., 2022; Krueger et al., 2021; Gulrajani and Lopez-Paz, 2020).

Fan et al. (2024) addresses the statistical challenge of estimating a stable linear relationship across multiple envi-

ronments with a data model that relaxes restrictions on the heterogeneity of the environments, requiring only the conditional expectation of the response and not the joint distribution to remain invariant for invariant features, but is restricted to linear models and does not extend to overparameterized deep models. In contrast, Zhou et al. (2022) suggest that IRM fails to drop spurious features when paired with deep models. They propose a global sparsity constraint to further eliminate spurious features from feature representation, based on a probabilistic approach (Zhou et al., 2021) to the lottery ticket hypothesis (Frankle and Carbin, 2018).

Two key challenges arise from this line of work. First, the sample complexity result in (Zhou et al., 2022) does not correctly capture the non-asymptotic case, due to errors in the analysis which mix up empirical and population terms, and the result incurs an additional dependency on the ambient dimensionality when working with finite samples. Second, the existing methods for sparse IRM are computationally inefficient. They either require searching over subsets of features, (Fan et al., 2024) or probabilistically prune network weights, which is computationally slow (Zhou et al., 2022).

We address the first challenge by providing a correct result through a generalized information-theoretic analysis. With d_{inv} invariant features and d total features, we present an information theoretic analysis with L_0-norm constraint selecting d_{inv} features. We show that a variant of the IRM formulation will provably find the correct d_{inv} features, with sample complexity depending polynomially on d_{inv} and logarithmically on d. The analysis is effectively information theoretic, with no consideration for computational demands, and it implies working with all the $\binom{d}{d_{\text{inv}}}$ L_0-constrained problems, but showing that this will identify the correct invariant features. For the second concern, we focus on practical efficient algorithms based on projected gradient descent (PGD) based on L_1-norm constraints and iterative hard thresholding (IHT) (Blumensath and Davies, 2009; Jain et al., 2014), to avoid the combinatorial complexity of the L_0-constrained approach. Our approach is efficient and guaranteed to recover the invariant optimal predictors. To summarize, our work makes the following contributions:

Non-Asymptotic Theory. We present a non-asymptotic analysis of using sparsity to select invariant features on the proposed IdepRM penalty. Our results show that L_0 constrained estimation in IRM is able to find the correct invariant features under suitable assumptions. The sample complexity is $O(\text{poly}(d_{\text{inv}})\log(d))$, where d includes the undesirable features. Our model captures more realistic scenarios where spurious features vary in their correlation with the label, using novel *scale* parameters. Expressing complexity in terms of these parameters decouples it from dataset size (see Section 3.2).

Modularity. Prior work on sparse IRM necessarily trains deep neural networks with sparse subnetwork selection, i.e., to change the training procedure to get invariant features. In contrast, our approach, based on sparsity on the last layer of the neural network, can be directly applied to many different settings, including the myriad pretrained models, without the need to change their training. Further, the modularity in our approach, i.e., feature selection happening at the last layer, makes it flexible by allowing the ability to "hot-swap" different sparse estimators, e.g., based on IHT or PGD with convex relaxations.

Experiments. We present experimental results with different instances of our sparse IRM, demonstrating better performance than that of existing IRM methods, including Sparse IRM with subnetworks. We also show that our methods are computationally efficient.

2 RELATED WORK

Invariant Risk Minimization Invariance as an indicator of causality was introduced by Peters et al. (2016), who outline the goal of seeking a subset of features that are causal for a target variable or label. The features are generated by structural equation models (SEMs), where interventions on different features create different environments. They suggest that with access to a sufficient number of independent environmental interventions, or environments, the invariant features can be recovered. The IRM paradigm (Arjovsky et al., 2020) applies this idea to learning causal features across a number of training domains. Because this optimization question is computationally intractable, they propose the IRMv1 variant, which uses the gradient norm as a constraining penalty for invariance.

A number of followup works propose variants that implement the paradigm, including IRM games (Ahuja et al., 2020), IRM with information bottlenecking (Ahuja et al., 2021a; Li et al.), risk extrapolation (Krueger et al., 2021), and learning spurious features without environment index (Tan et al., 2023). Theoretical works on the IRM paradigm largely analyze linear models (Arjovsky et al., 2020; Rosenfeld et al., 2020; Wang et al.), although analyses of nonlinear models for analysis exist to varying degrees of generality (Rosenfeld et al., 2020; Lai and Wang, 2024). Some of these works also highlight simple failure cases of IRM (Rosenfeld et al., 2020; Ahuja et al., 2021b). The data generation model introduced by (Arjovsky et al., 2020) has also been extended to overparameterized models (Zhou et al., 2022), or to cover different types of environmental variables (Kaur et al., 2022; Rosenfeld et al., 2020).

Domain Generalization IRM is closely related to other methods that tackle Domain Generalization (DG), which broadly targets good OOD generalization on unseen environments after training on more than one training domain. Similar lines of work include distributionally robust optimization (Sagawa et al., 2020; Volpi et al., 2018), which

aims to improve the overparameterized models over worst-case training loss on different data groups. Domain adaptation covers methods which also leverage information of the test domain to best capture distributional shift (Ben-David et al., 2006; Sun and Saenko, 2016; Ganin et al., 2016).

Sparse Representation Highly overparameterized DNNs are prevalent in modern machine learning, and many works have developed techniques to eliminate unnecessary weights or finding sparse representations (Han et al., 2016; Li et al., 2017; Hinton et al., 2015). A simple and popular technique is to use constrained L_0 norm, or its convex relaxation with LASSO, to enforce sparsity. Alternatively, projected gradient descent (PGD) methods are fast, efficient, and provably recover the optimal parameter with low estimation error (Loh and Wainwright, 2013; Negahban et al., 2009; Agarwal et al., 2010; Banerjee et al., 2015). Other paradigms explored to induce sparsity include probability-based methods for pruning (Louizos et al., 2017; Srinivas et al., 2017; Molchanov et al., 2017), which have shown empirical success in this regime as well. Finally, Zhou et al. (2022); Fan et al. (2024) provide evidence that combining sparsity with IRM can improve generalizability across domains, requiring knowledge of the number of sparse, invariant features, and combinatorially iterating through all feature subsets to find the causal subset.

3 PROBLEM SETTING

3.1 IRM SETTING

We denote a training set $\mathcal{D} = \{\mathcal{D}^e\}_{e\in\mathcal{E}}$ composed of environmental training datasets $\mathcal{D}^e := \{(\boldsymbol{x}_i^e, y_i)\}_{i=1}^{n_e}$, $\boldsymbol{x}_i^e \in \mathcal{X} \subseteq \mathbb{R}^p, y_i \in \mathcal{Y} \subseteq \mathbb{R}$. Each point is drawn i.i.d. from an environmental distribution $P^e(\boldsymbol{x}^e, y)$. Each environmental dataset $\mathcal{D}^e$ has n_e points for a total of $n = \sum_{e\in\mathcal{E}} n_e$ points in total. In the IRM paradigm outlined by Arjovsky et al. (2020), the goal is to find a predictor $f : \mathcal{X} \to \mathcal{Y}$, defined as $f(\boldsymbol{x}) = \boldsymbol{v}^\top \Phi(\boldsymbol{x})$, with a linear component $\boldsymbol{v} \in \mathbb{R}^d$ and a feature extractor $\Phi : \mathcal{X} \to \mathbb{R}^d$.

The mapping Φ is said to be invariant if there exists a $\boldsymbol{v}$ such that $f(\boldsymbol{x})$ is minimized across all environments simultaneously. Specifically, we define the population risk $\mathcal{R}^e(\boldsymbol{v}) = R^e(\boldsymbol{v}^\top\Phi(\boldsymbol{x}^e)) = \mathbb{E}^e[\ell(f(\boldsymbol{x}^e), y)]$ and empirical risk $\hat{\mathcal{R}}^e(\boldsymbol{v}) = \sum_{i=1}^{n_e} \ell(f(\boldsymbol{x}_i^e), y_i)$, per environment. The IRM formulation looks for the best $(\Phi, \boldsymbol{v})$ that minimizes the following constrained problem:

$$\begin{aligned} \min_{\substack{\Phi:\mathcal{X}\to\mathbb{R}^d \\ \boldsymbol{v}:\mathbb{R}^d\to\mathcal{Y}}} \quad & \sum_{e\in\mathcal{E}} R^e(\boldsymbol{v}^\top\Phi(\boldsymbol{x}^e)), \\ \text{subject to} \quad & \boldsymbol{v} \in \underset{\boldsymbol{v}^e:\mathbb{R}^d\to\mathcal{Y}}{\arg\min} R^e\left((\boldsymbol{v}^e)^\top\Phi(\boldsymbol{x}^e)\right) \quad \forall e \in \mathcal{E}. \end{aligned} \tag{1}$$

We consider a generative model in the style of previous lines of work in IRM (Rosenfeld et al., 2020; Ahuja et al., 2021a; Zhou et al., 2022) that explicitly has invariant, environmental ("non-invariant"), and random features. As coined by Ahuja et al. (2022), these are confounder, or anti-causal, models in which $P^{e_1}(y|\Phi(\boldsymbol{x}^e)) \neq P^{e_2}(y|\Phi(\boldsymbol{x}^e))$ if $\Phi(\boldsymbol{x}^e) = \boldsymbol{x}^e$.

3.2 DATA GENERATION

IRM struggles to discover invariant data representations in the overparameterized regime, where the number of model parameters exceeds the size of the training set (Li and Liang, 2018; Allen-Zhu et al., 2019). Even in the simple linear model introduced by Zhou et al. (2022), unmodified IRM fails to recover the underlying invariant structure. Because the data representation $\Phi(\boldsymbol{x}^e)$ may not completely isolate the invariant features, we are interested in finding the subset of invariant features in the data representation.

We let $\boldsymbol{x}^e = \Phi(\boldsymbol{x}^e)$, and work directly with the representation. This reflects the interpretation that $\boldsymbol{x}^e$ is the output of the all-but-last layer of a deep neural network, which may have captured non-invariant features. Then, for a given sample $(\boldsymbol{x}^e, y)$ drawn from any environment $e \in \mathcal{E}$, write the feature vector as a concatenation of invariant, spurious, and random feature blocks, i.e., $(\boldsymbol{x}^e)^\top = [\boldsymbol{x}_{\text{inv}}^\top, (\boldsymbol{x}_s^e)^\top, \boldsymbol{x}_r^\top]$, for $\boldsymbol{x}^e \in \mathbb{R}^d$ and $\boldsymbol{x}_{\text{inv}} \in \mathbb{R}^{d_{\text{inv}}}, \boldsymbol{x}_s^e \in \mathbb{R}^{d_s}, \boldsymbol{x}_r \in \mathbb{R}^{d_r}$, where $d = d_{\text{inv}} + d_s + d_r$. Although the term "spurious" formally refers to features that are not caused by the label yet share a strong correlation with it (Rosenfeld et al., 2020), we adopt it as the common nomenclature for features that are caused by the label for clarity. We use the superscript e to denote a *dependency* on the environment to which the feature belongs; any variable (i.e., $\boldsymbol{x}_{\text{inv}}, \boldsymbol{x}_r$) that does not have the superscript indicates that it is independent of the environment. The dependencies are illustrated in Figure 1. We use $\odot$ to represent the Hadamard (element-wise) product between two vectors of the same length d, i.e., $(\boldsymbol{v} \odot \boldsymbol{w})_i = v_i w_i \; \forall i \in [d]$. The generative model is as follows:

$$\begin{aligned} y &= \gamma^\top \boldsymbol{x}_{\text{inv}} + \epsilon_{\text{inv}} \,, \\ \boldsymbol{x}_s^e &= y\boldsymbol{\zeta}_s + \boldsymbol{\alpha}^e \odot \boldsymbol{\epsilon}_s \,, \\ \boldsymbol{x}_r &= \boldsymbol{\zeta}_r \odot \boldsymbol{\epsilon}_r \,. \end{aligned} \tag{2}$$

As discussed in Zhou et al. (2022), the label y is generated from a fixed vector $\gamma \in \mathbb{R}^{d_{\text{inv}}}$, which is invariant across environments. Spurious features depend on both the labels as well as the environment; the variable $\boldsymbol{\alpha}^e \in \mathbb{R}^{d_s}$ controls the environment-dependent noise in each spurious feature and $\epsilon_{\text{inv}}, \boldsymbol{\epsilon}_s, \boldsymbol{\epsilon}_r$ are independent noise variables added to the system. We assume they are sub-Gaussian and centered, and we are interested in the regime in which d_s, d_r are very large. Additional scaling parameters $\boldsymbol{\zeta}_s \in \mathbb{R}^{d_s}$ and $\boldsymbol{\zeta}_r \in \mathbb{R}^{d_r}$. $\boldsymbol{\zeta}_s$ control the strength of the correlation between a spurious feature $[\boldsymbol{x}_s^e]_j$ for $j \in [d_s]$ and the label, y. Likewise, $\boldsymbol{\zeta}_r$ determines the scale of the random features. We also

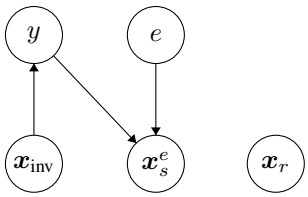

Figure 1: Causal relationship between observed variables in the generative model by Zhou et al. (2022). All variables are observed at training time, but the type of an individual feature $\boldsymbol{x}^e$ is unknown.

Table 1: List of variables for the generative model with invariant features. * indicates newly introduced variables. Column header Dim. is short for dimensionality.

Variable	L_2 **norm**	**Dim.**	**Definition**
γ	1	d_{inv}	Ground truth
ϵ_{inv}	-	1	Ground truth
$\boldsymbol{\zeta}_s$	c_s	d_s	*Label correlation
$\boldsymbol{\alpha}^e$	c_a	d_s	Spurious noise
$\boldsymbol{\epsilon}_s$	-	d_s	Sub-Gaussian noise
$\boldsymbol{\zeta}_r$	c_r	d_r	*Noise scale
$\boldsymbol{\epsilon}_r$	-	d_r	Sub-Gaussian noise

assume the basic noise random variables $\epsilon_{\text{inv}}, \boldsymbol{\epsilon}_s, \boldsymbol{\epsilon}_r$ are sub-Gaussian.

3.3 SELECTING INVARIANT FEATURES

Presented with a large feature vector dominated by spurious and random features, we want to find a model $f(\boldsymbol{x}^e) = \boldsymbol{v}^\top \boldsymbol{x}^e$ that is invariant across $\boldsymbol{x}^e$ drawn from different $e \in \mathcal{E}$ as in Equation (2). By the IRM paradigm (Arjovsky et al., 2020; Rosenfeld et al., 2020), this can only be achieved if $f(\boldsymbol{x}^e)$ depends only on invariant features, i.e., the support of $\boldsymbol{v}$ is a subset of the features $\boldsymbol{x}^e_{\text{inv}}$. Thus, we formulate the problem in terms of subsets of features.

Formally, let S be a subset of features, $S \in 2^d$, that represents the footprint for $\boldsymbol{v}$. We denote the set of all predictors that are only nonzero on S as $\text{Sp}(S)$. Then,

$$\text{Sp}(S) \coloneqq \{\boldsymbol{v} \in \mathbb{R}^d : \boldsymbol{v}_i = 0 \; \forall i \notin S\}, \tag{3}$$

and contains $\boldsymbol{v}$ that can take any value in features in S, and are 0 elsewhere.

We define the invariant footprint, i.e., the subset of invariant features corresponding to x_{inv}, as S_{inv}. Formally,

$$S_{\text{inv}} \coloneqq \{i \in [d] \mid \boldsymbol{x}_i^e \in \boldsymbol{x}_{\text{inv}}\}. \tag{4}$$

This is a small subset of all features if $d_{\text{inv}} \ll d$, and at training time, it is not known which of the available features are members of this set. We are then interested in seeking the **optimal invariant predictor**, as defined below.

Definition 1 (Optimal Invariant Predictor)**.** Let the optimal invariant predictor β^* be

$$\beta^* \coloneqq \operatorname*{argmin}_{\boldsymbol{v} \in \text{Sp}(S_{\text{inv}})} \sum_{e \in \mathcal{E}} \mathcal{R}^e(\boldsymbol{v}). \tag{5}$$

In other words, it is the best parameter that relies only on the invariant features $\boldsymbol{x}^e_{\text{inv}}$.

Two hurdles are evident: first, finding β^* requires prior knowledge of which features belong in S_{inv}, which we don't have. Thus, it is an information-theoretic target, i.e., without consideration for computational demands, since solving the outer problem of the best subset S. This involves searching over $\binom{d}{d_{\text{inv}}}$ subsets if we know d_{inv}; otherwise, the search space is $2^{d_{\text{inv}}}$. Further, we will be working with empirical loss whereas β^* is defined based on population loss.

Remark 1. In the problem setting defined by Equation (2), $\beta^* = [\gamma^\top, (\mathbf{0}^{d_s})^\top, (\mathbf{0}^{d_r})^\top]$, and is also a solution to Equation (1); this is easily shown, and details are provided in Proposition 14 in the appendix. □

We also use the subscript S and superscript e notation to represent environment and feature-restricted population optima,

$$\beta^e_S \coloneqq \operatorname*{argmin}_{\boldsymbol{v} \in \text{Sp}(S)} \mathcal{R}^e(\boldsymbol{v}), \quad \beta^*_S \coloneqq \operatorname*{argmin}_{\boldsymbol{v} \in \text{Sp}(S)} \sum_{e \in \mathcal{E}} \mathcal{R}^e(\boldsymbol{v}). \tag{6}$$

We extend this notation to the empirical minimizers,

$$\hat{\beta}^e_S \coloneqq \operatorname*{argmin}_{\boldsymbol{v} \in \text{Sp}(S)} \hat{\mathcal{R}}^e(\boldsymbol{v}), \quad \hat{\beta}_S \coloneqq \operatorname*{argmin}_{\boldsymbol{v} \in \text{Sp}(S)} \sum_{e \in \mathcal{E}} \hat{\mathcal{R}}^e(\boldsymbol{v}). \tag{7}$$

IRM introduces a penalty that penalizes non-invariant classifiers. It is generally formulated

$$\mathcal{L}(\boldsymbol{v}) \coloneqq \sum_{e \in \mathcal{E}_{tr}} \mathcal{R}^e(\boldsymbol{v}) + \rho \sum_{e \in \mathcal{E}} \mathcal{J}^e(\boldsymbol{v}), \tag{8}$$

with penalty weight $\rho > 0$ for some $\mathcal{J}^e : \mathbb{R}^d \to \mathbb{R}^+$ that captures a violation of invariance in Φ across environments.

For the analysis, we first adapt the IRM minimax penalty (Zhou et al., 2022), otherwise called the *loss difference* penalty, as a proxy for the constraint imposed by the original bi-level optimization formulation. With $\boldsymbol{v}_S \in \text{Sp}(S) \subseteq \mathbb{R}^d$, we have

$$\begin{aligned} \mathcal{L}(\boldsymbol{v}_S) \coloneqq & \sum_{e \in \mathcal{E}} \mathcal{R}^e(\boldsymbol{v}_S) \\ & + \rho \sum_{e \in \mathcal{E}} \max_{\boldsymbol{v}^e_S \in \text{Sp}(S)} \left[\mathcal{R}^e(\boldsymbol{v}_S) - \mathcal{R}^e\left(\boldsymbol{v}^e_S\right)\right]. \end{aligned} \tag{9}$$

If there exists some $\boldsymbol{v}_S$ which minimizes Equation (9), From this, the minimax loss can also be defined for a given subset of features $S \in 2^d$,

$$\mathcal{L}(S) \coloneqq \min_{\boldsymbol{v}_S \in \text{Sp}(S)} \mathcal{L}(\boldsymbol{v}_S) = \mathcal{L}(\beta_S^*). \tag{10}$$

However, computing this loss in practice, e.g., to use with gradient descent, would require solving an inner optimization problem in order to find the second term of the penalty, $\min_{\boldsymbol{v}^e \in \text{Sp}(S)} \mathcal{R}^e(\boldsymbol{v}^e)$. In practice, the penalty is often replaced with the gradient norm penalty introduced by (Arjovsky et al., 2020).

$$\mathcal{L}_{\text{IRMv1}}(\boldsymbol{v}) \coloneqq \sum_{e \in \mathcal{E}_{tr}} \mathcal{R}^e(\boldsymbol{v}) + \rho \sum_{e \in \mathcal{E}_{tr}} \|\nabla_{\boldsymbol{v}} \mathcal{R}^e(\boldsymbol{v})\|_2^2. \tag{11}$$

We show in Proposition 2 why this is an appropriate proxy for the minimax loss under reasonable assumptions, which are satisfied with linear least squares.

Remark 2. While the formulation in Equation (11) is a commonly used penalty in IRM optimization, and receives a detailed treatment in Fan et al. (2024), it does not absolve the need to search over all subsets $|S| \leq d_{\text{inv}}$ which may provide candidates for the invariant classifier. In fact, the variation across different subsets in Equation (10), under loss functions optimized over all of $\mathbb{R}^d$, prevents the direct application of LASSO and other convex relaxation techniques to Equation (10). As a result, both Zhou et al. (2022) and Fan et al. (2024) resort to gradient descent over the full space of $\mathbb{R}^d$. □

3.4 OPTIMIZATION

With the loss defined for the population case, we are ready to examine the minimax formulation for finite samples, the empirical counterpart to Equation (10):

$$\begin{aligned}\hat{\mathcal{L}}(\boldsymbol{v}_S) \coloneqq & \sum_{e \in \mathcal{E}} \hat{\mathcal{R}}^e(\boldsymbol{v}_S) \\ & + \rho \sum_{e \in \mathcal{E}} \max_{\boldsymbol{v}_S^e \in \text{Sp}(S)} \left[\hat{\mathcal{R}}^e(\boldsymbol{v}_S) - \hat{\mathcal{R}}^e\left(\boldsymbol{v}_S^e\right)\right].\end{aligned} \tag{12}$$

Again, we use the empirical minimizers defined in Equation (7) to indicate the loss incurred by the minimum of a given subset of features S:

$$\hat{\mathcal{L}}(S) = \min_{\boldsymbol{v}_S \in \text{Sp}(S)} \hat{\mathcal{L}}(\boldsymbol{v}_S) = \hat{\mathcal{L}}(\hat{\beta}_S). \tag{13}$$

This results in a two-step breakdown of the IRM problem. First, for any given subset $S \in 2^d$, we solve Equation (12), which can be solved by standard optimization methods. Second, we need to optimize over different subsets $S \in \mathcal{S} \subseteq 2^d$, to obtain the minimum over all subsets, which is a combinatorial problem over $\binom{d}{d_{\text{inv}}}$ subsets:

$$\hat{\mathcal{L}}(\bar{S}) = \min_{S \in \mathcal{S}} \hat{\mathcal{L}}(S). \tag{14}$$

In this setting, we first provide a sample complexity result (Theorem 1) when S_{inv} is optimal, i.e., the minimum number of samples $n \geq n_0$ needed per environment such that $\bar{S} = S_{\text{inv}}$. Thus, if $n \geq n_0$, running the combinatorial optimization will indeed find the correct set of invariant features.

Remark 3. Zhou et al. (2022) implement such optimization by searching for a robust, sparse data representation by applying ProbMask, a subnet-discovery alogrithm (Zhou et al., 2021) to the IRM problem. However, this approach is not *sparse feature selection*, but rather finding a *robust representation that is sparse*: the output of the representation is not necessarily sparse, and the last linear layer of their model is fully connected and dense. Both our analysis (Section 4) and experiments (Section 5) follow the line of *sparse feature selection* instead, by explicitly applying a sparsity constraint on the last layer. □

Remark 4. We also note that although many earlier works consider IRM for classification (Rosenfeld et al., 2020; Wang et al.), our regression model can be generalized to classification with conditional Bernoulli (or conditional multimodal) models. Further detail can be found in Appendix C.1. □

4 EFFICIENT SPARSE IRM

We want to show sample complexity bounds under which we can guarantee, with high probability, recovery of the invariant feature subset S_{inv} by minimizing Equation (12). In Section 4.3, we will examine both the use of the IRMv1 penalty, and the minimax penalty, the latter of which provides an additional result that demonstrates the optimality of the population parameter on even the empirical loss. We then provide an analysis of computationally efficient methods for maintaining sparsity. This involves fast projected gradient methods like Iterative Hard Thresholding, and we address this in terms of the gradient norm penalty Equation (11) which is more commonly used in practice.

4.1 THEORETICAL RESULTS

We first establish that, although IRM methods aim to eliminate spurious features already, that it fails in the overparameterized regime, motivating the need for sparsity-constrained IRM methods.

Proposition 1. *IRM fails in the overparameterized setting. We assume that* $d > n_{tot} = \sum_{e \in \mathcal{E}} n_e \geq d_{\text{inv}}$,

$$\hat{\mathcal{L}}(S_{\text{inv}}) \geq \hat{\mathcal{L}}(S), |S| > n_{tot} \tag{15}$$

Proof. Note $\min_{|S| > d_{\text{inv}}} \hat{\mathcal{L}}(S) = 0$ in the linear setting. Indeed, the set S_{inv} belongs to the set of footprints with cardinality $|S| > n_{\text{tot}}$, so $\min_{|S| > n_{\text{tot}}} \hat{\mathcal{L}}(S)$ is necessarily a lower bound. □

Empirically, the IRM paradigm alone struggles to eliminate the spurious features $\boldsymbol{x}_s$ and random features $\boldsymbol{x}_r$, which together constitutes the majority of features input to the linear classifier. Then, the natural starting point is the formulation of $\hat{\mathcal{L}}(\boldsymbol{v})$ as a IRM minimax loss function from Equation (12) with an explicit L_0 constraint,

$$\min_{\boldsymbol{v}\in\mathbb{R}^d} \hat{\mathcal{L}}(\boldsymbol{v}) \text{ s.t. } \|\boldsymbol{v}\|_0 \leq d_{\text{inv}}. \tag{16}$$

In this setting, we provide a guarantee of invariant feature recovery with finite samples on the minimax penalty.

Theorem 1 (Informal: sample complexity of optimizing Eqn. 12)**.** *Assume at least n samples per environment $e \in \mathcal{E}$, for a total of $N = |\mathcal{E}|n$ across the whole training set. If*

$$n \geq O\left(\text{poly}(d_{\text{inv}}) \log\left(\frac{|\mathcal{E}|d}{\delta}\right)\right),$$

together with assumptions in Appendix A.2, with probability at least $(1-\delta)$, the following holds:

$$\hat{\mathcal{L}}(S_{\text{inv}}) < \hat{\mathcal{L}}(S), \quad \forall |S| \leq d_{\text{inv}}, \quad S \neq S_{\text{inv}}\ , \tag{17}$$

Remark 5. The formal statement and a more detailed treatment of the constants in the sample complexity are provided in Appendix B.

Theorem 1 provides a sample complexity under which we guarantee that the resulting model depends on exactly the invariant features S_{inv}. With the definitions in Equation (7), we see that it is equivalent to the statement $\hat{\mathcal{L}}(\hat{\beta}_{\text{inv}}) < \hat{\mathcal{L}}(\hat{\beta}_S)$ for all $|S| \leq d_{\text{inv}}$. Informally, this implies that a parameter using any non-invariant features incurs a large enough penalty that it will have higher loss than $\hat{\mathcal{L}}(\hat{\beta}_{\text{inv}})$. Our result applies to $|\mathcal{E}|$ environments, noting that the minimum number of samples per environment scales with $\log(|\mathcal{E}|d/\delta)$, logarithmic in both the number of environments and the ambient dimensionality. In practice, this is easy to satisfy and is reflected in standard benchmarks Colored MNIST (Arjovsky et al., 2020), ColoredObject (Lin et al., 2014; Zhou et al., 2022), and MNISTCIFAR (Shah et al., 2020).

The next result shows that the empirical loss $\hat{\mathcal{L}}$ is also able to differentiate between the invariant optimal predictor β^*_{inv} from the population optimizers on non-invariant footprints $S \neq S_{\text{inv}}$, which we show in Theorem 2. This unusual connection between empirical loss and population minimizer is a consequence of the structure of the IRM penalty in Equation (10), and we are able to achieve this result with only mildly higher sample complexity: a multiplicative factor $O(\text{poly}(d_{\text{inv}}))$ more than the sample complexity in Theorem 1.

Theorem 2 (Sample complexity for sparse IRM with population optima)**.** *For population minimizers as defined in Equation* (6)*, and n samples per environment $e \in \mathcal{E}$, for a total of $N = |\mathcal{E}|n$ across the whole training set, we have*

$$\hat{\mathcal{L}}(\beta^*) < \hat{\mathcal{L}}(\beta^*_S), \quad |S| \leq d_{\text{inv}}, S \neq S_{\text{inv}}, \tag{18}$$

if $n > O\left(\text{poly}(d_{\text{inv}}) \log\left(\frac{d\cdot|\mathcal{E}|}{\delta}\right)\right)$ with constants specified in Appendix B.3.

Details that characterize this further are found in the proofs of Theorem 1 and Theorem 2 in Appendix B.

Remark 6. If we assume $\boldsymbol{\zeta}_s = \mathbf{1}^{d_s}$ and $\boldsymbol{\zeta}_r = \mathbf{1}^{d_r}$, we get the original linear model by Zhou et al. (2022). However, this will yield sample complexity and estimation error bounds which are dimension-dependent, i.e., dependent on d_{inv}, d_s, and d_r. To motivate variable $\boldsymbol{\zeta}_s$ as an example, consider that for d_s features, the size of the data $\|\boldsymbol{x}^e\|_2$ is $O(\sqrt{d_s})$ when $\boldsymbol{\zeta}_s = \mathbf{1}$. If we instead let the scaling parameter $\boldsymbol{\zeta}_s$ be changed, we allow different spurious features to correlate differently with labels. In addition to being a substantially more realistic assumption on the data, it allows us to create scale-dependent bounds. Then, the scale may be as low as $O(\frac{d_{\text{inv}}}{d_s+d_r})$ when instead generating the data with a fixed $\|\boldsymbol{\zeta}^e_s\|_2^2$. Corollary 9 compares this case.

4.2 PROOF SKETCHES

To prove Theorem 1, our analysis follows an approach similar to Zhou et al. (2022), but avoids the several errors in that analysis required to show Equation (12). Our approach is sketched in this section with full details in Appendix B.

Proof sketch. First, we break down the minimax penalty, defined in Equation (9), into a sum of three error components. In other words, $\mathcal{J}(\hat{\beta}_S) = \xi_a(S) + \xi_b(S) + \xi_c(S)$. We let $c_1, c_2, c_3 > 0$ be positive constants, S_{spu} be the set of spurious features, and $\alpha_i^2 = \frac{1}{|\mathcal{E}|}(\alpha_i^e)^2$ be the average value of the α_i scaling for a spurious feature i across environments.

$$\xi_a(S) = \sum_{e\in\mathcal{E}} \left[\hat{\mathcal{R}}^e(\hat{\beta}_S) - \mathcal{R}^e\left(\beta^*_S\right)\right] \leq c_1\sqrt{\frac{\log(\frac{1}{\delta})}{|\mathcal{E}|n}}, \tag{19}$$

$$\xi_b(S) = \sum_{e\in\mathcal{E}} \left[\hat{\mathcal{R}}^e(\hat{\beta}^e_S) - \mathcal{R}^e\left(\beta^e_S\right)\right] \leq c_2\sqrt{\frac{\log(\frac{|\mathcal{E}|}{\delta})}{n}}. \tag{20}$$

These two intermediate quantities $\xi_a(S)$ and $\xi_b(S)$ bound similar gaps, but $\xi_a(S)$ works with the across-environment minimizers $\hat{\beta}_S$ and β^*_S, and $\xi_b(S)$ bounds the environment-specific $\hat{\beta}^e_S$ and β^e_S. Both sum the gap across all environments, and the generalization-style bound is tighter for $\xi_a(S)$'s single classifier and greater sample complexity.

$$\begin{aligned}\xi_c(S) &= \sum_{e\in\mathcal{E}} [\mathcal{R}^e(\beta^*_S) - \mathcal{R}^e\left(\beta^e_S\right)] \\ &\geq c_3 \min_{i\in S_{\text{spu}}} |\alpha_i^2 - (\alpha_i^e)^2|.\end{aligned} \tag{21}$$

Note that Equation (19) and Equation (20) are not a result of directly applying Hoeffding's inequality for sub-Gaussian random variables, as the different errors are not

independent. Instead, we apply triangle inequality and $\hat{\mathcal{R}}^e(\hat{\beta}_S) - \hat{\mathcal{R}}^e(\beta^*) < 0$, by by definition of $\hat{\beta}_S$. We may then apply Hoeffding's inequality on the errors incurred on β^*. Thus, $\xi_a(S), \xi_b(S)$ decrease with sample complexity.

Equation (21), can be computed directly; its lower bound can be derived under reasonable assumptions of the environmental parameter $\boldsymbol{\alpha}^e$, detailed in Appendix A.2. Intuitively, the more α_i varies across environments, the better the bound. With these quantities, we compute the samples required n to have any non-invariant footprint $S \neq S_{\text{inv}}$ elicit a higher loss $\hat{\mathcal{L}}(S)$, provided that $|S| \leq d_{\text{inv}}$. Critically, $\xi_c(S_{\text{inv}}) = 0$, and $\xi_c(S_{\text{inv}}) > O$ for all $S \neq S_{\text{inv}}$. The full proof can be found in Appendix B. □

Remark 7. The quantity $\xi_c(S)$ is positive as long as there exist environments $e_1, e_2 \in \mathcal{E}$ and some spurious feature such that $i \in S$, where $\alpha_i^{e_1} \neq \alpha_i^{e_2}$. Intuitively, $\xi_c(S)$ captures the difference between environmental distributions, when only accessing features in S. As a result, $\xi_c(S_{\text{inv}}) = 0$, since S_{inv} contains only features that remain invariant across environments. Then, it is possible to lower bound $\xi_c(S)$ for $S \neq S_{\text{inv}}$ by leveraging environmental separation of the underlying distributions. It benefits from more "widely-ranging" values of $\boldsymbol{\alpha}^e$. In this way, it links back to previous works like (Ahuja et al., 2021a), which impose requirements on differences in environment to present general sample complexity results. □

The proof of Theorem 2 follows the same structure as that of Theorem 1, with an additional $\text{poly}(d_{\text{inv}})$ term incurred by $|\hat{\mathcal{R}}(\beta_S^*) - \hat{\mathcal{R}}(\hat{\beta}_S^*)|$ and $|\hat{\mathcal{R}}(\beta_S^e) - \hat{\mathcal{R}}(\hat{\beta}_S^e)|$. The full proof can be found in Appendix B.3.

Remark 8. Zhou et al. (2022) provides a bound for $\xi_b(\Phi) = \xi_b(S)$ which requires $\hat{\mathcal{R}}^e(\hat{\beta}_S^e) - \hat{\mathcal{R}}^e(\beta_S^e) \leq |\hat{\mathcal{R}}^e(\hat{\beta}_S^e) - \hat{\mathcal{R}}^e(\beta_S^e)| = \|\hat{\beta}_S^e - \beta_S^e\|_{\hat{\Sigma}^e}$. The equality is formally stated in Lemma 1 of Zhou et al. (2021) but is in general untrue for arbitrary feature subset S in the non-asymptotic setting. As a result, their final claim that this term is $O(n^{-0.5})$ is incorrect as well; they are missing an important term that arises from the misspecified model. We provide a corrected analysis in Lemma 5 in our appendix. □

Under the generative model introduced in Equation (2), it is impossible for ERM and Sparse ERM to recover the invariant features only in the asymptotic case; see Appendix B.4. In this setting, both IRM and IRM with sparsity constraints can recover the optimal invariant predictor. For the non-asymptotic case, we provide sample complexity bounds for Sparse IRM that leverage the invariant feature dimensionality. The result follows in Theorem 1, and the full proof is found in Appendix B.

4.3 EFFICIENT ALGORITHMS

The loss formulation in Theorem 1 uses a L_0 constraint, which is not computationally practical. We refer to the the rich line of work proving sharp convergence rates and bounds on estimation error under constraints for regression problems (Negahban et al., 2009; Agarwal et al., 2010; Banerjee et al., 2015). We later leverage a subset of these works (Loh and Wainwright, 2013; Jain et al., 2014) which show the same guarantees for methods in the family of Projected Gradient Descent (PGD) or Iterative Hard Thresholding (IHT) algorithms, which provide bounds for high-dimensional statistical settings. We apply IHT to solve Equation (12) and show that the sparse invariant feature recovery is possible with these fast, space-efficient methods.

Algorithm 1 Sparse IRM with Iterative Hard-Thresholding

1: **Input:** target nonzero features $d_{\text{inv}} < d$, $\mathcal{D} = \{\mathcal{D}^e\}_{e \in \mathcal{E}}$ and $\mathcal{D}^e := \{(\boldsymbol{x}_i^e, y_i)\}_{i=1}^{n_e}$.
2: Initialize weights $\boldsymbol{v}$.
3: **for** training iteration $t = 1, 2, \cdots, T$ **do**
4: $\quad \boldsymbol{v}^{t+1} \leftarrow \text{proj}_s(\boldsymbol{v}^t - \eta \nabla_{\boldsymbol{v}} \hat{\mathcal{L}}(\boldsymbol{v}^t))$
5: $\quad t = t + 1$
6: **end for**

Let $s \in \mathbb{N}$ be the sparsity level. Then, the *hard thresholding projection operator* $\text{proj}_s : \mathbb{R}^d \to \mathbb{R}^d$ is defined as:

$$\text{proj}_s(\boldsymbol{v}) := \arg\min_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \|\boldsymbol{v} - \boldsymbol{u}\|_2^2 \;\middle|\; \|\boldsymbol{u}\|_0 \leq s \right\}, \quad (22)$$

where $\|u\|_0$ denotes the number of nonzero entries in u. Algorithm 1 directly projects the gradient descent update onto the non-convex feasible set. Previous works (Jain et al., 2014) have shown that, despite the non-convexity of the problem, tight, minimax lower bounds can be achieved on the learned parameter, and we use constants from Theorem 3 in Jain et al. (2014) for sparse linear regression. Details are included in Appendix C.1.

Theorem 3 (Sparse IRM with IHT)**.** *Assume n samples per training environment, for $n > Q\left(\text{poly}(d_{\text{inv}}) \log(d) \log\left(\frac{|E|}{\delta}\right)\right)$. Together with assumptions in Appendix A.2, using the IRMv1 penalty as defined in Equation* (11)*, Algorithm 1 returns a parameter $\tilde{\beta} \in \mathbb{R}^d$ With s chosen to be $O(d_{\text{inv}})$, we have with probability at least $1 - \delta$, a bound on the estimation error,*

$$\|\tilde{\beta} - \beta_{\text{inv}}^*\|_2 = O\left(\lambda_{\max}^2 \sqrt{\frac{d_{\text{inv}} \log d}{n}} + \frac{\sigma_{\text{inv}}}{\kappa_s}\right). \quad (23)$$

The full proof and definitions for constants $Q, \sigma_{\text{inv}}, and \kappa_s$ are provided in Appendix C.1. Because we do not know $\|\beta^*\|_0 = d_{\text{inv}}$ beforehand, we discuss tuning s as a hyperparameter in Section 5. Overall, both methods provide guarantees of low estimation error in high probability, while

being fast and having low memory cost, scaling to much larger models and datasets.

5 EXPERIMENTS

Algorithms: We compare our approach, IRM with iterative hard thresholding (IRMv1 + IHT), with relevant baselines ERM, sparse ERM, the oracle, and IRM-based methods. For IRM-based methods, we use IRMv1 (Arjovsky et al., 2020), and we provide Proposition 2 to prove it is an acceptable proxy for the minimax formulation in Equation (12). In order, ERM is the standard training loop on the mixture of all environments; and sparse ERM adds IHT (Jain et al., 2014). The oracle trains ERM with spurious features zeroed, upper bounding accuracies for other methods. For the IRM-based methods, we compare with the original IRMv1 (Arjovsky et al., 2020), and IRMv1 with ProbMask (IRMv1+PM) (Zhou et al., 2022, 2021). When comparing sparsity-based methods, we fix the target density of the feature representation to be same across methods.

Datasets: We use common invariant representation learning benchmarks, ColoredMNIST (2-CMNIST) is the original binary dataset introduced in Arjovsky et al. (2020), and FullColoredMNIST (10-CMNIST) (Ahmed et al., 2021) is also generated from MNIST, with two environments, 10 labels and 10 colors. MNISTCIFAR concatenates MNIST digits and CIFAR-10 images (Shah et al., 2020). The oracle baseline is constructed per dataset and only has the designated invariant features: the grayscale MNIST for 2- and 10-CMNIST, and the CIFAR image for MNISTCIFAR. Parameters for the dataset configurations, including label noise and environmental correlation, are in Appendix F.

Hyperparameter selection: Because we do not know d_{inv} at train time, it is common to treat s in algorithm 1 as a hyperparameter as in e.g. (Wainwright, 2019). Specifically, we take a uniform grid search per dataset. We find also that accuracy is not affected significantly by small perturbations in s, which is demonstrated by data from additional experiments on MNISTCIFAR in Table 4.

Evaluation metrics: Top-1 test accuracy is compared for the three tasks. For ResNet-18 on MNISTCIFAR, we also provide training time results, and the relative timing in comparison to standard ERM.

Discussion: We observe that IRM with IHT can match or exceed the performance of competing methods, including IRM with ProbMask sparsity, for larger models and datasets. Sparse ERM, IRMv1+PM, and IRMv1+IHT were computed with 88% weight density in Table 2; this corresponds to 12% of the weights zeroed out by sparsificaiton methods. The L_1 norms of the layer also reflect the sparsification. ProbMask incurs a noticeable computational overhead – an additional 23% over IRMv1. IHT only adds a 4% cost. We expect time savings to scale up with larger models. Additionally, we provide results for a MLP with two hidden layers of dimension 390, the median configuration of the model used by (Zhou et al., 2022) on these datasets.

6 CONCLUSIONS

In this paper, we provide a non-asymptotic analysis of IRM with sparsity constraints. First, we generalize the data model, relaxing the data model to allow for varying correlation between spurious features and the label. Next, we provide the non-asymptotic results for sparse IRM, including a refinement and correction of previous work in sparse IRM, including theoretical guarantees for L_1- and L_0-constrained IRM, resulting in a sparse representation that selects invariant features. Finally, we demonstrate that these methods can be computed in a fast and efficient matter using projected gradient descent-based methods, and we provide experimental results that demonstrate improved test accuracy and time savings on domain generalization datasets.

7 ACKNOWLEDGEMENTS AND DISCLOSURE OF FUNDING

The work was supported by the National Science Foundation (NSF) through awards IIS 21-31335, OAC 21-30835, DBI 20-21898, as well as a C3.ai research award.

References

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/7cce53cf90577442771720a370c3c723-Paper.pdf.

Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron C. Courville. Systematic generalisation with group invariant predictions. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=b9PoimzZFJ.

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 145–155. PMLR, 13–18 Jul

Table 2: Top-1 test accuracy of ResNet-18 with timings on MNISTCIFAR. Our method, IRMv1+IHT, bolded, has negligible overhead time cost and the overall best test accuracy.

Method	Test Accuracy	Train Time (s)	% time/ERM	L_1 norm of last layer
Oracle	77.85 ± 0.14	36.38 ± 0.26	99%	19.72 ± 3.88
ERM	44.93 ± 0.49	36.65 ± 0.25	- %	25.05 ± 2.32
Sparse ERM	44.82 ± 0.42	37.31 ± 0.37	102%	18.31 ± 2.46
IRMv1	52.86 ± 0.53	36.51 ± 0.17	100%	18.65 ± 0.93
IRMv1+PM	57.30 ± 0.45	44.98 ± 1.02	123%	21.59 ± 1.02
IRMv1+IHT	$\mathbf{62.44 \pm 0.96}$	$\mathbf{38.03 \pm 0.51}$	**104%**	$\mathbf{9.10 \pm 1.78}$

Table 3: Top-1 train and test accuracy of MLP390.

	10-CMNIST Accuracy (%)	
Method	Train	Test
Oracle	73.06 ± 0.21	71.36 ± 0.44
ERM	90.00 ± 0.29	28.32 ± 0.10
Sparse ERM	87.17 ± 1.16	29.15 ± 2.14
IRMv1	70.77 ± 0.27	58.88 ± 0.14
IRMv1+PM	$\mathbf{92.20 \pm 0.10}$	$\mathbf{65.16 \pm 0.09}$
IRMv1+IHT	$\mathbf{80.83 \pm 0.10}$	$\mathbf{63.03 \pm 0.51}$

2020. URL https://proceedings.mlr.press/v119/ahuja20a.html.

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. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3438–3450. Curran Associates, Inc., 2021a.

Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R Varshney. Empirical or invariant risk minimization? a sample complexity perspective. *ICLR*, 2021b.

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, 2022.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/allen-zhu19a.html.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

Arindam Banerjee, Sheng Chen, Farideh Fazayeli, and Vidyashankar Sivakumar. Estimation with norm regularization, 2015.

O E Barndorff-Neils. *Information and exponential families*. Wiley Series in Probability and Statistics. John Wiley & Sons, Nashville, TN, April 2014.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, page 472–489, Berlin, Heidelberg, 2018. Springer-Verlag. ISBN 978-3-030-01269-4. doi: 10.1007/978-3-030-01270-0_28. URL https://doi.org/10.1007/978-3-030-01270-0_28.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/b1b0432ceafb0ce714426e9114852ac7-Paper.pdf.

Thomas Blumensath and Mike E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2009.04.002. URL https://www.sciencedirect.com/science/article/pii/S1063520309000384.