Iterative Feature Matching: Toward Provable Domain Generalization with Logarithmic Environments

Yining Chen
Stanford University
cynnjjs@stanford.edu

Elan Rosenfeld
Carnegie Mellon University
elan@cmu.edu

Mark Sellke
Stanford University
msellke@stanford.edu

Tengyu Ma
Stanford University
tengyuma@stanford.edu

Andrej Risteski
Carnegie Mellon University
aristesk@andrew.cmu.edu

Abstract

Domain generalization aims at performing well on unseen test environments with data from a limited number of training environments. Despite a proliferation of proposed algorithms for this task, assessing their performance both theoretically and empirically is still very challenging. Distributional matching algorithms such as (Conditional) Domain Adversarial Networks [12, 28] are popular and enjoy empirical success, but they lack formal guarantees. Other approaches such as Invariant Risk Minimization (IRM) require a prohibitively large number of training environments—linear in the dimension of the spurious feature space d_s—even on simple data models like the one proposed by Rosenfeld et al. [37]. Under a variant of this model, we show that ERM and IRM can fail to find the optimal invariant predictor with $o(d_s)$ environments. We then present an iterative feature matching algorithm that is guaranteed with high probability to find the optimal invariant predictor after seeing only $O(\log d_s)$ environments. Our results provide the first theoretical justification for distribution-matching algorithms widely used in practice under a concrete nontrivial data model.

1 Introduction

Domain generalization aims at performing well on unseen environments using labeled data from a limited number of training environments [8]. In contrast to transfer learning or domain adaptation, domain generalization assumes that neither labeled nor unlabeled data from the test environments is available at training time. In the empirical literature, invariance of a "signal" feature distribution conditioned on the label, i.e. $P(\Phi(x) \mid y)$, is the underlying assumption in widely adopted algorithms such as Correlation Alignment (CORAL) [43, 42], Maximum Mean Discrepancy (MMD) [13, 25], and (Conditional) Domain Adversarial Networks [12, 28]. Distribution-matching algorithms regularize the distribution of intermediate representations of examples from different environments to be closer in some metric. They are among the top-performing methods for domain generalization benchmarks like PACS [24], VLCS [11], and DomainNet [34], but they lack formal guarantees. Previous empirical works usually characterize the performance of these algorithms using generalization bounds based on divergence between domain distributions [7, 31], but those bounds are vacuous for the typical benchmarks for domain generalization and thus cannot explain their success. For example, when different environments have disjoint supports (as is the case in the above-mentioned benchmarks) the $\mathcal{H}$-divergence [7] is 1 even for the linear hypothesis class. To obtain useful guarantees, it is necessary to study data models that encode structure reflective of settings of interest. Prior works attempting to theoretically characterize the performance of feature matching algorithms emphasize

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

lower bounds [49, 44]. In this work, we seek to give the first *positive* theoretical justification for feature matching algorithms.

Another common assumption in the literature is invariance of the label distribution conditioned on the signal features. Invariant Risk Minimization (IRM) [3] assumes $\mathbb{E}[y \mid \Phi(x)]$ is invariant, and follow-up works assume invariance of higher moments [46, 18, 30, 6]. However, empirical results for these algorithms are mixed: Gulrajani and Lopez-Paz [15], Aubin et al. [4] present experimental evidence that these methods do not consistently outperform ERM for either realistic or simple linear data models, when fairly evaluated.

Recent theoretical works [37, 19] also question the theoretical foundations of IRM and its variants, shedding light on their failure conditions. These works study specific data generative models; a common assumption is that, conditioned on the label, some invariant features have identical distribution for all environments, and other spurious features have varying distributions across environments. The goal of these methods is then to obtain an *invariant predictor*, i.e. a classifier which uses only the invariant features. These works also often assume each training environment contains infinite samples. Thus, **the central measure of domain generalization is the number of environments needed to recover an invariant predictor—we refer to this measure as the *environment complexity* of a learning algorithm.** Rosenfeld et al. [37] show even for a simple generative model and linear classifiers, the environment complexity of IRM—and other objectives based on the same principle of invariance—is at least as large as the dimension of the spurious latent features, d_s. Further results by Kamath et al. [19], Ahuja et al. [2] also point to a linear environment complexity. Although the models in these works are simple, they help elucidate why existing algorithms fail and can help inform better algorithmic design.

IRM's linear environment complexity is prohibitive for realistic applications. Domain generalization benchmarks have fewer training environments than spurious features dimensions—PACS, VLCS, DomainNet have 4-6 total environments but the stylistic variations between domains are likely high-dimensional. In this paper, we show that a variant of distributional matching algorithm (Algorithm 1) finds the optimal invariant predictor with sublinear (even logarithmic) environment complexity.

1.1 Our contributions

Conceptually, we propose a "smoothed covariance" extension of the data model in Rosenfeld et al. [37], which is rich enough to establish an environment complexity separation between IRM/ERM and feature matching algorithms. More precisely, we show that ERM and IRM can fail to find the optimal invariant predictor with fewer than $d_s + 1$ training environments (Theorems 4.2, 4.3), while a relatively simple and natural algorithm based on iterative feature matching (IFM) finds the optimal invariant predictor with $O(\log d_s)$ environments.

Moreover, our analysis highlights the value of iterative matching, much similar to matching at multiple layers in a deep neural network. IFM iteratively projects the features to a lower dimension while matching the label-conditioned feature distributions on a small, disjoint subset of the training environments. A projection inducing invariance in the non-invariant features across one subset of environments is unlikely to do so for a different subset, so each projection only removes (a constant fraction of) spurious feature dimensions with high probability. Intuitively, the iterative scheme prevents the different environments from "colluding" to create a solution which depends on spurious features. As a result, IFM recovers the optimal invariant predictor after $O(\log d_s)$ rounds and uses $O(1)$ environments per round, thus requiring $O(\log d_s)$ environments.

Our techniques for proving the upper and lower bounds may be of independent interest. The upper bound (Theorem 5.1) is proved by showing that any projection matrix that uses many spurious dimensions cannot match feature covariances in a large set of training environments. This is done via intricate matrix concentration bounds and decoupling inequalities [10].

We derive the exact lower bound for IRM (Theorem 4.3) using tools from differential topology. Each environment provides an ellipsoidal constraint on the solution, and we prove that there exists a non-trivial intersection of these constraints (besides the origin, which corresponds to the "intended" solution). Our key lemma shows that the total number of intersections between two manifolds of complementary dimensions $k, d-k$ is even when certain tranversality conditions hold, implying that the origin cannot be the only solution.

Finally, to corroborate the advantages of the proposed algorithm, we perform experiments on a Gaussian dataset and a semi-synthetic Noised MNIST dataset [23], where the background noise spuriously correlates with the label. Our results in Section 6 suggest that practitioners may benefit from feature matching algorithms when the distinguishing property of the signal feature is indeed conditional distributional invariance, and may get additional advantage via matching at multiple layers with diminishing dimensions, echoing existing empirical observations [27, 29].

2 Preliminaries

2.1 Domain generalization

In domain generalization, we are given a set of E training environments $\mathcal{E}_{tr}$ indexed by $e \in [E]$,[1] and a set of test environments $\mathcal{E}_{ts}$. For environment e we have n examples $\{(X_i^e, Y_i^e)\}_{i=1}^n$ drawn from the distribution P_e. In this work we study the infinite sample limit $n \to \infty$ so as to separate the effect of limited training environments from that of limited samples *per* environment, as is done in previous theoretical works [37, 19]. Let $\mathcal{X}$, $\mathcal{P}$, $\mathcal{Y}$ denote the space of inputs, intermediate features, and labels. For a featurizer $\Phi : \mathcal{X} \to \mathcal{P}$ and classifier $w : \mathcal{P} \to \mathcal{Y}$, their risk on environment e is denoted by $R^e_{\Phi,w} = \mathbb{E}_{(X,Y)\sim P_e}[l(w \circ \Phi(X), Y)]$ for any common loss function l. In this paper we focus on $\mathcal{Y} = \{\pm 1\}$, linear featurizers $\Phi(X) = UX$ for $U \in \mathbb{R}^{k\times d}$, and unit-norm predictors $\widehat{Y} = \text{sgn}\,(w^\top U x)$ where $w \in \mathbb{R}^k$ and $\|w^\top U\|_2 = 1$ for some feature dimension $k \leq d$ chosen by the algorithm. A predictor's 0-1 risk on environment e is denoted by $R^e_{U,w} = \Pr_{(X,Y)\sim P_e}[\text{sgn}\,(w^\top UX) \neq Y]$. We focus on unit-norm predictors because we evaluate on the 0-1 risk on test environments, which are invariant to the scaling of $w^\top U$ under our data model.

2.2 Baseline algorithms

We analyze the performance of our proposed method and compare it to two baseline algorithms, ERM and IRM. ERM learns a classifier that minimizes the average loss over all training environments, where l is any common training loss such as the logistic loss:

$$\min_{w\in\mathbb{S}^{d-1}} \frac{1}{E} \sum_{e\in[E]} \mathbb{E}_{(X,Y)\sim P_e}[l(w^\top X, Y)].$$

IRM learns a featurizer $\Phi(X) \in \mathbb{R}^k$ such that the optimal classifier on top of the featurizer is invariant across training environments. As we focus on linear classifier, it is equivalent to learning a linear transformation $U \in \mathbb{R}^{k\times d}$ such that it induces a classifier w that is optimal for all $e \in \mathbb{E}_{tr}$:

$$\begin{aligned} &\min_{U\in\mathbb{R}^{k\times d}, w\in\mathbb{R}^k, \|w^\top U\|_2=1} \frac{1}{E} \sum_{e\in[E]} \mathbb{E}_{(X,Y)\sim P_e} l((w^\top UX), Y) \\ &\text{s.t.} \quad w \in \arg\min_{w'\in\mathbb{R}^k} \mathbb{E}_{(X,Y)\sim P_e}[l((w'^\top UX), Y)], \forall e \in \mathcal{E}_{tr}. \end{aligned}$$

This is objective is *not* the same as feature distribution matching; IRM only tries to match the first moment. Observe that this constrained objective is intended to solve a minimax domain generalization problem, as opposed to ERM which is typically viewed as minimizing the risk in expectation.

3 Problem setup

We first recall the data model from Rosenfeld et al. [37]. We assume without loss of generality that the label Y is uniformly randomly drawn from $\{\pm 1\}$ (extension of our theorems to $Y = 1$ with probability $\eta \neq 0.5$ is straightforward). Latent variable Z consists of invariant features $Z_1 \in \mathbb{R}^r$ and spurious features $Z_2 \in \mathbb{R}^{d_s}$ where $d_s = d - r$. The number of spurious features can be much larger than the number of invariant features, i.e. $d_s \gg r$. The input $X \in \mathbb{R}^d$ is generated via a linear

[1] We define $[n] = \{1, \ldots, n\}$; $\mathbf{0}_{n\times m} \in \mathbb{R}^{n\times m}$ denotes an all-zero matrix; $\mathbb{S}^d$ is the unit sphere in $\mathbb{R}^{d+1}$; $\text{sgn}\,(c) \in \{\pm 1, 0\}$ is the sign of scalar $c \in \mathbb{R}$. $\dagger$ denotes the Moore-Penrose pseudo-inverse.

transformation of latent variable Z, i.e. $X = SZ$ for a matrix $S \in \mathbb{R}^{d \times d}$ such that its left r columns have rank r (so that there are r invariant dimensions).

For each training environment indexed by $e \in [E]$, the invariant features conditioned on Y are drawn from a Gaussian distribution with mean $Y \cdot \mu_1 \in \mathbb{R}^r$ and nonsingular covariance $\Sigma_1 \in \mathbb{R}^{r \times r}$. The spurious features conditioned on Y have mean $Y \cdot \mu_2^e \in \mathbb{R}^{d_e}$ and covariance $\Sigma_2^e \in \mathbb{R}^{d_e \times d_e}$ where μ_e's and Σ_e's vary across $e \in [E]$. The assumption of symmetric class center with respect to the origin can also be relaxed. Define $\mu^e = [\mu_1, \mu_2^e]$ and $\Sigma^e = [\Sigma_1, \mathbf{0}_{r \times d_s}; \mathbf{0}_{d_s \times r}, \Sigma_2^e]$. The overall data model for training environments is summarized below:

$$
\begin{aligned}
Y &\overset{iid}{\sim} \text{unif}\{\pm 1\} \\
Z_1 | Y &\sim N(Y \cdot \mu_1, \Sigma_1) \in \mathbb{R}^r \\
Z_2 | Y &\sim N(Y \cdot \mu_2^e, \Sigma_2^e) \in \mathbb{R}^{d_s} \\
Z &= [Z_1, Z_2] \in \mathbb{R}^d \\
X &= SZ.
\end{aligned}
$$

Since the goal of invariant feature learning is to learn a predictor that only uses the invariant features, one reasonable measure for domain generalization is a predictor's performance on test environments where the spurious features Z_2 are drawn from a different distribution—in particular, they are usually chosen adversarially. A classifier that predicts using the spurious features will perform badly on such test environments. When modeling the test environments, we consider the difficult scenario where there is one corresponding test environment for each training environment, whose parameters are the same except that the spurious means are flipped. Formally, for each environment $e \in \mathcal{E}_{tr}$ we construct a corresponding test environment $e' \in \mathcal{E}_{test}$ where

$$Z_2 \sim N(-Y \cdot \mu_2^e, \Sigma_2^e) \in \mathbb{R}^{d_s}.$$

In this setting where the observations X are a linear function of the latents Z, Rosenfeld et al. [37] assume that the covariances of spurious features are isotropic and vary only in magnitude:

Assumption 3.1 (Data model for spurious covariances in Rosenfeld et al. [37])**.**

$$\Sigma_2^e = \sigma_e^2 I_{d_s}, \textit{ where } \sigma_e \textit{ is a scalar for an environment indexed by } e.$$

We consider a generalized model where the covariances of spurious features for each environment is a generic random PSD matrix, instead of only random in scaling:

Assumption 3.2 (Data model for spurious covariances in this work)**.**

$$\Sigma_2^e \sim \overline{\Sigma_2^e} + G_e G_e^\top, \textit{ where } \overline{\Sigma_2^e} \in \mathbb{R}^{d_s \times d_s} \textit{ is arbitrary (and can be adversarial), and } [G_e]_{i,j} \overset{iid}{\sim} N(0,1) \textit{ for all } i, j \in [d_s]. \textit{ Furthermore,} \max_e \|\overline{\Sigma_2^e}\|_2^2 \leq D.$$

Assumption 3.2 allows the covariance matrix of the spurious features to be *almost worst-case*: it is an arbitrary matrix, plus a Gaussian perturbation. This is a common assumption in algorithmic complexity called *smoothed analysis* [41], as we allow the parameters to be a random smoothing of arbitrary parameters.

While our data model is simple, it has already been used as a sandbox to understand algorithms like IRM [37, 19] because it captures some important aspects of real-life data like latent variables and correlations between the labels and the spurious features.

In the next section, we show that baseline algorithms ERM and IRM still have $\Omega(d_s)$ environment complexity under assumption 3.2, whereas our iterative feature matching algorithm (Algorithm 1) requires only $O(\log d_s)$ training environments. Note that the environment complexity of our algorithm only depends logarithmically on the norm bound D.

4 Main results

Armed with assumption 3.2, we now present our main results. We begin by presenting our algorithm based on iterative feature matching. We then provide formal guarantees for its environment complexity in comparison to ERM and IRM.

Algorithm 1 Iterative Feature Matching (IFM) algorithm

Require: Invariant feature dimension r, total feature dimension d, number of training environments $E = |\mathcal{E}_{tr}|$, infinite samples $\{(X_i^e, Y_i^e)\}_{i=1}^{\infty} \sim P_e$ from each environment $e \in \mathcal{E}_{tr}$, an array of integers indicating the number of training environments to use at each iteration $E_1, E_2, \ldots$.

1: $r_0 \leftarrow d$, $t \leftarrow 0$.
2: **while** $r_t > r$ **do**
3: $\quad t \leftarrow t + 1$.
4: $\quad$ Uniformly randomly sample E_t training environments without replacement.
5: $\quad$ Binary search between 1 and r_{t-1} to find the maximum dimension r_t such that there exists orthonormal $U_t \in \mathbb{R}^{r_t \times r_{t-1}}$ and $C_t \in \mathbb{R}^{r_t \times r_t}$, where for all $e \in \mathcal{E}_t$,

$$\mathbb{E}_{(X,Y)\sim P_e}[U_t \ldots U_1 X X^\top U_1^\top \ldots U_t^\top | Y] = C_t. \tag{4.1}$$

6: Return a classifier on projected features that minimizes the average risk

$$\widehat{w} = \min_{w \in \mathbb{S}^{r-1}} \frac{1}{E} \sum_{e \in [E]} \mathbb{E}_{(X,Y)\sim P_e} l(w^\top U_t \ldots U_1 X, Y).$$

4.1 Iterative feature matching algorithm

We hope to recover the invariant features by imposing constraints which are satisfied by only those features (i.e., they are not satisfied by the spurious features). A natural idea is to match the feature means and covariances across $\mathcal{E}_{tr}$. Since μ_1, Σ_1 are constant, any orthonormal featurizer $U \in \mathbb{R}^{r \times d}$ such that US has only non-zero entries in the first r rows yields invariant means $US\mu^e$ and covariances $US\Sigma^e S^\top U^\top$. Thus we need E large enough such that any $U' \in \mathbb{R}^{r \times d}$ using spurious dimensions cannot match the means and covariances. Informally, for each $e \in [E]$ we get $r \times r$ equations from matching covariances $U\Sigma^e U^\top = C$, and we have $r \times d$ parameters to estimate in U. Rough parameter counting suggests that if we match covariances of all E environments jointly, we need at least $E > d/r$ environments to find a unique solution. Our key observation is that, due to the independence of randomness in Σ_2^e, we can split E environments into disjoint groups $\mathcal{E}_1, \ldots, \mathcal{E}_T$, and use $\mathcal{E}_t$ to train an orthonormal featurizer that shrinks the feature dimensions from r_{t-1} to r_t. Thus, in each round we shrink the dimension by a constant factor using a constant number of environments. The main theoretical challenge that remains is to show that in each iteration, with high probability, *only* spurious features are projected out.

This brings us to IFM (Algorithm 1), which proceeds in $T = O(\log d_s)$ rounds. Starting with an input dimension $r_0 = d$, each round we learn an orthonormal matrix U_t projecting features from r_{t-1} to r_t dimensions so that the feature covariances after projection match across a fresh set of training environments. In practice, for each choice of r_t, starting from random initialization, we perform SGD on objective $\min_{U_t \in \mathbb{R}^{r_t \times r_{t-1}}} \sum_{e,e' \in \mathcal{E}_t} \|U_t(\Sigma^e - \Sigma^{e'})U_t^T\|_F^2 + \lambda \|U_t^\top U_t - I\|_F^2$ until the objective is ϵ-close to 0. To ensure that all invariant dimensions are preserved, we always find the projection with the maximum possible dimension r_t that still matches the covariances (i.e. admits an ϵ-optimal solution U_t), until we are left with only r dimensions.

Among algorithms used in practice, IFM is most similar to CORAL [43]. We study IFM since it is more amenable to theoretical analysis. The differences between IFM and CORAL are: first, CORAL does not enforce that the featurizer is orthonormal; second, IFM learns to extract features in an unsupervised manner, whereas CORAL jointly minimizes the supervised loss and feature distribution discrepancy; third, IFM matches the feature distributions at multiple layers and uses a disjoint set of environments for each layer—this iterative process is necessary for the theoretical guarantees we provide—but CORAL matches only at the last layer. Despite these differences, our theoretical results serve as a justification for using feature matching algorithms in general, when the distinguishing attribute of signal vs. spurious features is that the former have invariant distributions across all environments. In section 6 we empirically show that adding the core features of IFM to CORAL can improve test accuracy.

The following theorem states that the environment complexity of IFM is logarithmic in the spurious feature dimension. A proof sketch is given in Section 5.

Theorem 4.1 (IFM upper bound)**.** *Under assumption 3.2, suppose at each round IFM uses* $|\mathcal{E}_t| = \tilde{\Omega}(1)$ [2] *training environments, with probability* $1 - \exp(-\Omega(d_s))$*, IFM terminates in* $O(\log d_s)$ *rounds and outputs* $\widehat{w} = w^*$*.*

As a remark, we note that there is in fact a simple algorithm that achieves $O(1)$ environment complexity under Assumption 3.2 (see Appendix D). However, this algorithm is extremely brittle and reliant on very specific aspects of the data model (such as Gaussianity) and it cannot be extended to other settings, whereas feature matching algorithms are regularly applied to general model architectures and real-world datasets. The goal of this paper is to provide theoretical justification for distribution matching algorithms and to investigate why they may outperform ERM and IRM, rather than to solve a specific data model.

4.2 ERM and IRM still have linear environment complexity

In the previous section, we showed that IFM has low environment complexity thanks to the additional structure we imposed in our model. However, it is possible that this additional structure also allows ERM and IRM to succeed. These next two results demonstrate that this is not the case.

ERM has low test accuracy In contrast to IFM, ERM still suffers from linear environment complexity under Assumption 3.2. The first theorem says there are hard instances where the ERM solution has worse-than-random performance on the test environments.

Theorem 4.2 (ERM lower bound)**.** *Suppose* $E \leq d_s$*, parameters* $\mu_1 \in \mathbb{R}^r$*,* $\Sigma_1 = \sigma_1^2 I_r$*,* $\mu_2^e \in \mathbb{R}^{d_s}$*,* $\overline{\Sigma_2^e} = \sigma_2^2 I_{d_s}$ *(recall* $\Sigma_2^e \sim \overline{\Sigma_2^e} + G_e G_e^\top$*). Then any unit-norm linear classifier which achieves accuracy* $\geq \Phi\left(\frac{2\|\mu_1\|}{\min(\sigma_1,\sigma_2)}\right)$ *on all training environments will suffer 0-1 error at least* $\frac{1}{2}$ *on every test environment with flipped spurious mean, where* Φ *is the standard Normal CDF.*

A complete proof of Theorem 4.2 is in Appendix B.3. Note that it is quite reasonable to assume that the ERM solution satisfies the accuracy condition. In particular, it is common to model the spurious features as having much greater magnitude than the invariant features, since they have much greater dimensionality. For example, with a unit-norm mean we would expect $\|\mu_1\|^2 \approx r/d$, $\|\mu_2^e\|^2 \approx d_s/d$. Then for $r \ll d$ and $\sigma_1, \sigma_2 = \omega(1/\sqrt{d})$ one can verify that $2\|\mu_1\|/\min(\sigma_1,\sigma_2) = o(\sqrt{r/d})$ is very close to 0, meaning the lower bound $\Phi\left(\frac{2\|\mu_1\|}{\min(\sigma_1,\sigma_2)}\right)$ is only slightly larger than $\frac{1}{2}$.

IRM fails to learn invariant features Our next theorem proves that even under Assumption 3.2, IRM is still not guaranteed to find an invariant predictor. We can show this by proving that when $E \leq d_s$, we can find a featurizer that only uses spurious dimensions, i.e., $u_s \in \mathbb{R}^{d_s}$, such that $u_s^\top \Sigma_2^e u_s = u_s^\top \mu_2^e$ for all $e \in \mathcal{E}_{tr}$. If so, the optimal predictor on top of features $u_s^\top Z_2$, $\widehat{w^e} = (u_s^\top \Sigma_2^e u_s)^{-1} u_s^\top \mu^e$ is invariant across all e, and can therefore be the preferred solution IRM when the spurious features have greater magnitude than the invariant features on the training environments.

Theorem 4.3 (IRM lower bound)**.** *Suppose* $E \leq d_s$*. If* $\mu_2^1, \ldots, \mu_2^E \in \mathbb{R}^{d_s}$ *are linearly independent, then there exists* $u_s \in \mathbb{R}^{d_s}$*,* $\|u_s\|_2 > 0$*, such that* $u_s^\top \Sigma_2^e u_s = u_s^\top \mu_2^e$ *for all* $e \in [E]$*.*

Proof Sketch. Observe that each environment provides an ellipsoidal constraint $E_e = \{u_s \in \mathbb{R}^{d_s} : u_s^\top \Sigma_2^e u_s - u_2^\top \mu_2^e = 0\}$. The origin is a trivial intersection. We prove the existence of a non-trivial intersection using tools from differential topology. The key lemma is that the total number of intersection points between two manifolds of complementary dimensions $k, d-k$ is even when certain tranversality conditions hold. Using these techniques, we show that $|\bigcap_e E_e| \geq 2$ for almost all matrices $\Sigma_2^1, \ldots, \Sigma_2^E$, as long as the means are linearly independent. $\square$

A complete proof of Theorem 4.3 is in Appendix B.4.

[2] $\tilde{\Omega}(\cdot)$ hides logarithmic factors in D, r, d_s.

5 Proof sketch for the main upper bound Theorem 4.1

To argue that IFM outputs a featurizer $U_1 \dots U_T$ that does not use the spurious features, we need to show that the right d_s columns of matrix $U_T \dots U_1 S$ are all-zero. The main lemma below says that this happens with high probability if we match $\tilde{\Omega}(1)$ environments at every iteration,

Lemma 5.1. *If for all $1 \le t \le T$, $|\mathcal{E}_t| = E_t = \Omega\left(c(\log(D/d_s) + \log d_s)\right)$, and $U_1, \dots, U_T$ are the orthonormal matrices returned by IFM, then with probability $1 - \exp(-\Omega(d_s))$, $r_t - r < (r_{t-1} - r + 1)/c$ for all t, and if we write $U_T \dots U_1 S = [A, B]$, where $B \in \mathbb{R}^{r \times d_s}$, then $B = \mathbf{0}_{r \times d_s}$.*

Theorem 4.1 follows from Lemma 5.1 as follows: We take $c = 2$ and the algorithm terminates in $T = O(\log d_s)$ rounds. Therefore with an environment complexity of $O(\log d_s)$, we learn a feature extractor $U = U_T \dots U_1$ that does not use any spurious dimensions. Since U is orthonormal, it must contain all signal dimensions. The predictor on top of this representation uses all and only signal dimensions, so with high probability, IFM outputs $\widehat{w} = w^*$.

The first step towards proving Lemma 5.1 is to show that with high probability, any *one-layer* featurizer $Q_1 \in \mathbb{R}^{k_1 \times d_s}$ that uses *only* spurious dimensions cannot match feature covariances from $\tilde{\Omega}(d_s/k_1)$ environments. If a featurizer $U_1 \in \mathbb{R}^{r_1 \times d}$ uses k_1 spurious dimensions, there is a corresponding rank-k_1 featurizer $Q_1 \in \mathbb{R}^{k_1 \times d_s}$ that uses *only* spurious dimensions. So Lemma 5.2 implies that any U_1 that matches covariances in $\mathcal{E}_1$ must use at most d_s/E_1 spurious dimensions. We will then apply this argument recursively until we have 0 spurious dimensions.

Lemma 5.2 (Informal version of Lemma B.3)**.** *For any integer $2 \le k_1 \le d_s/2$, when $|\mathcal{E}_1| = E_1 = \Omega\left(\frac{d_s - k_1}{k_1 - 1} \max\left\{1, \log\left(\frac{D}{(k_1-1)d_s}\right), \log\left(\frac{d_s}{k_1 - 1}\right)\right\}\right)$, with probability $1 - O(\exp(-d_s))$, no orthonormal $Q \in \mathbb{R}^{k_1 \times d_s}$ satisfies that for some constant $C_1 \in \mathbb{R}^{k_1 \times k_1}$.*

$$\forall e \in [E_1], \quad Q \Sigma_2^e Q^\top = C_1. \tag{5.1}$$

Section B.1 gives a proof sketch. The formal statement Lemma B.3 and its proof can be found in Appendix B.2.

The next claim says that Lemma 5.2 can be applied iteratively, i.e. fixing a featurizer from previous iterations that uses k_{t-1} spurious dimensions, with high probability, any U_t that matches features from $\Omega\left(k_{t-1}/k_t\right)$ new environments uses at most k_t spurious dimensions.

Corollary 5.3 (Informal version of Corollary B.7)**.** *Suppose $2 \le k_t \le k_{t-1}/2 \le d_s/2$. When $|\mathcal{E}_t| = E_t = \Omega\left(\frac{k_{t-1} - k_t}{k_t - 1} \max\left\{1, \log\left(\frac{D}{(k_t-1)d_s}\right), \log\left(\frac{d_s}{k_t - 1}\right)\right\}\right)$, for fixed orthonormal $P \in \mathbb{R}^{k_{t-1} \times d_s}$, with probability $1 - O(\exp(-d_s))$, no orthonormal $Q \in \mathbb{R}^{k_t \times k_{t-1}}$ satisfies $\forall e \in [E_t]$, $QP\Sigma_2^e P^\top Q^\top = C_t$ for some constant $C_t \in \mathbb{R}^{r_t \times r_t}$.*

The formal statement Corollary B.7 and its proof can be found in Appendix B.2. Lemma 5.1 follows from iterative application of Corollary 5.3, as shown in Appendix B.2.

6 Experiments

In light of the differences between IFM and CORAL discussed in section 4.1, we test several questions inspired by our theory: (Q1) Do feature matching algorithms (IFM and CORAL) have much smaller environment complexity compared to ERM and IRM, with finite samples drawn from data models similar to our assumptions? (Q2) Can decoupling feature matching and supervised training of the classifier (IFM) improve over joint training (CORAL)? (Q3) For neural network featurizers, can matching feature distributions at multiple layers improve over matching at only the last layer (naive CORAL)? (Q4) Can matching disjoint sets of environments at each layer perform as well as matching all environments at all layers? (Q5) Is it important to shrink feature dimensions? We use two tasks to investigate those questions empirically. Appendix C contains additional details.

Gaussian dataset is a binary classification task that closely reflects our assumptions in section 3.

Noised MNIST is a 10-way semi-synthetic classification task modified from LeCun and Cortes [23]. Conditioned on the class, we add noise with identical mean but changing covariances across training environments. In test environments, the noise is uncorrelated with the label.

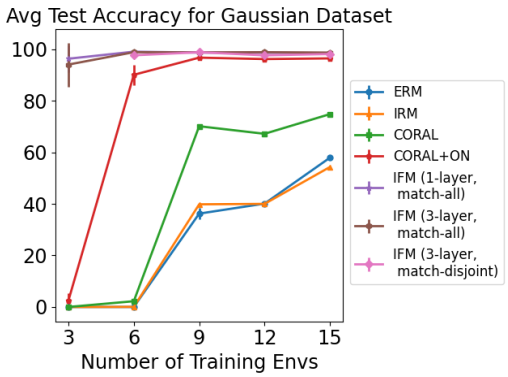

Figure 1: For Gaussian dataset, our algorithm IFM achieves highest test accuracy with the same number of training environments.

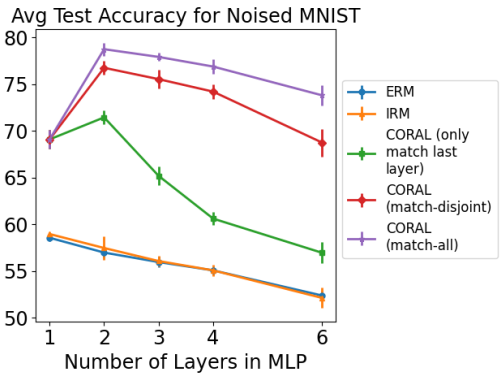

Figure 2: For Noised MNIST, matching feature distributions from multiple layers improves over naive CORAL across different architectures.

Algorithms and architectures. For Gaussian dataset we use linear predictors. **IRM** follows the implementation in Arjovsky et al. [3]; **CORAL** jointly minimizes average supervised loss on training environments and L_{coral}, which is the average of squared distances in conditonal feature means (in l_2 norm) and covariances (in Frobenius norm) between adjacent training environments; **CORAL+ON** adds orthonormal penalty loss $L_{on}(U) = \|UU^\top - I\|_F^2$ where U is the featurizer; **IFM** is our Algorithm 1, where for each layer U_t, the training objective is $L_t(U_t) = \lambda_1 L_{coral} + \lambda_2 L_{on}$. We test IFM with 1 vs. 3-layer featurizers, either matching all (**match-all**) or a disjoint set of training environments (**match-disjoint**) at each layer.

For Noised MNIST we use ReLU networks with 1-6 layers. Although our theory for IFM is only for the linear setting, in practice we can extend the idea to deep network by training the classifier head together with feature matching at different layers.**CORAL (match-all)** and **CORAL (match-disjoint)** in Figure 2 are natural extensions of IFM to nonlinear models. They match at all layers post-activation, using either all (**match-all**) or a disjoint subset of training environments (**match-disjoint**) per layer. Figure 2 shows that they outperform naive **CORAL**, which matches features only at the last layer. Thus, the core ideas of IFM may be used to improve performance of existing feature matching algorithms with nonlinear models and realistic datasets (e.g. DANN, MMD).

Results. (Q1) Figures 1 and 2 show that IFM and CORAL have much smaller environment complexity compared to ERM and IRM in both datasets. (Q2) In Gaussian dataset, IFM improves over CORAL. (Q3) For Noised MNIST, matching feature distributions at multiple layers (CORAL match-all, CORAL match-disjoint) improves over matching at only the last layer (CORAL). (Q4) In both datasets, matching disjoint sets of environments at each layer (IFM match-disjoint, CORAL match-disjoint) is almost as good as matching all environments at all layers (IFM match-all, CORAL match-all) while saving computation. (Q5) For Noised MNIST (Table 2 in Appendix C), shrinking feature dimensions is crucial for the advantage of feature matching at multiple layers, e.g. matching features at 3 layers with widths $[24, 24, 24]$ does not significantly improve over matching features at the last layer (CORAL). Overall, our results suggest that practitioners may benefit from feature matching algorithms when the data is similar to our assumed model, and may get additional advantage via matching at multiple layers with diminishing dimensions, echoing existing empirical works [27, 29].

7 Conclusion

This work presents the first domain generalization algorithm which provably recovers an invariant predictor with a number of environments that scales sub-linearly with the spurious feature dimension. Our results demonstrate that generalization which does not suffer from the "curse of dimensionality" is possible, and based on our theory we believe the use of an *iterative* approach is a key insight which could lead to additional positive results for out-of-distribution generalization. Notably, this work also represents the first theoretical justification for the empirical success of existing algorithms which use feature distribution matching. However, there remains much room for improvement: our results are

for a linear data model with fairly special assumptions. It would be interesting to analyze data models in which the observables are a non-linear function of the latent variables.

Acknowledgements

We would like to thank Colin Wei, Sang Michael Xie, Ananya Kumar, and Ciprian Manolescu for helpful discussions. YC is supported by Stanford Graduate Fellowship. MS is supported by NSF and Stanford Graduate Fellowships. TM acknowledges support of Google Faculty Award and NSF IIS 2045685. AR and ER acknowledge support from the NSF via IIS-2211907 and the CMU/PwC DT&I Center. AR acknowledges support of an Amazon Research Award.

References

[1] K. Ahuja, E. Caballero, D. Zhang, J.-C. Gagnon-Audet, Y. Bengio, I. Mitliagkas, and I. Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

[2] K. Ahuja, J. Wang, A. Dhurandhar, K. Shanmugam, and K. R. Varshney. Empirical or invariant risk minimization? a sample complexity perspective. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=jrA5GAccy_`.

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

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

[5] S. Beery, G. Van Horn, and P. Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[6] A. Bellot and M. van der Schaar. Generalization and invariances in the presence of unobserved confounding. *arXiv preprint arXiv:2007.10653*, 2020.

[7] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[8] G. Blanchard, G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL `https://proceedings.neurips.cc/paper/2011/file/b571ecea16a9824023ee1af16897a582-Paper.pdf`.

[9] M. Chevalley, C. Bunne, A. Krause, and S. Bauer. Invariant causal mechanisms through distribution matching. *arXiv preprint arXiv:2206.11646*, 2022.

[10] V. H. de la Peña and S. J. Montgomery-Smith. Decoupling inequalities for the tail probabilities of multivariate u-statistics. *The Annals of Probability*, pages 806–816, 1995.

[11] C. Fang, Y. Xu, and D. N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013.

[12] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[13] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(null):723–773, Mar. 2012. ISSN 1532-4435.

[14] V. Guillemin and A. Pollack. *Differential topology*, volume 370. American Mathematical Soc., 2010.

[15] I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

[16] A. Hyvarinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

[17] A. Hyvarinen, H. Sasaki, and R. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 859–868. PMLR, 16–18 Apr 2019.

[18] W. Jin, R. Barzilay, and T. Jaakkola. Domain extrapolation via regret minimization. *arXiv preprint arXiv:2006.03908*, 2020.

[19] P. Kamath, A. Tangella, D. Sutherland, and N. Srebro. Does invariant risk minimization capture invariance? In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 4069–4077. PMLR, 13–15 Apr 2021. URL http://proceedings.mlr.press/v130/kamath21a.html.

[20] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[21] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

[22] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 10 2000. doi: 10.1214/aos/1015957395. URL http://dx.doi.org/10.1214/aos/1015957395.

[23] Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

[24] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[25] H. Li, S. J. Pan, S. Wang, and A. C. Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018.

[26] Y. Li and Y. Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

[27] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 97–105, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/long15.html.

[28] M. Long, Z. CAO, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/ab88b15733f543179858600245108dd8-Paper.pdf.

[29] Z. Luo, Y. Zou, J. Hoffman, and L. F. Fei-Fei. Label efficient learning of transferable representations acrosss domains and tasks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/a8baa56554f96369ab93e4f3bb068c22-Paper.pdf.

[30] D. Mahajan, S. Tople, and A. Sharma. Domain generalization using causal matching. *arXiv preprint arXiv:2006.07500*, 2020.

[31] Y. Mansour, M. Mohri, and A. Rostamizadeh. Multiple source adaptation and the rényi divergence. *arXiv preprint arXiv:1205.2628*, 2012.

[32] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 10–18, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL `http://proceedings.mlr.press/v28/muandet13.html`.

[33] A. Pajor. Metric entropy of the grassmann manifold. *Convex Geometric Analysis*, 34:181–188, 1998.

[34] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019.

[35] A. Rame, C. Dancette, and M. Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 18347–18377. PMLR, 2022.

[36] A. Robey, G. J. Pappas, and H. Hassani. Model-based domain generalization. *Advances in Neural Information Processing Systems*, 34:20210–20229, 2021.

[37] E. Rosenfeld, P. Ravikumar, and A. Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=BbNIbVPJ-42`.

[38] M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(none):1 – 9, 2013. doi: 10.1214/ECP.v18-2865. URL `https://doi.org/10.1214/ECP.v18-2865`.

[39] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[40] D. Serre. *Matrices: Theory and Applications*. Graduate Texts in Mathematics. Springer, 2010. ISBN 9781441930101. URL `https://books.google.to/books?id=IYWLcgAACAAJ`.

[41] D. A. Spielman and S.-H. Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *J. ACM*, 51(3):385–463, May 2004. ISSN 0004-5411. doi: 10.1145/990308.990310. URL `https://doi.org/10.1145/990308.990310`.

[42] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[43] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2058–2065. AAAI Press, 2016.

[44] R. Tachet des Combes, H. Zhao, Y.-X. Wang, and G. J. Gordon. Domain adaptation with conditional distribution matching and generalized label shift. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19276–19289. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/dfbfa7ddcfffeb581f50edcf9a0204bb-Paper.pdf`.

[45] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing, Theory and Applications*, chapter 5, pages 210–268. Cambridge University Press, 2012.

[46] C. Xie, F. Chen, Y. Liu, and Z. Li. Risk variance penalization: From distributional robustness to causality. *arXiv preprint arXiv:2006.07544*, 2020.

[47] G. Zhang, H. Zhao, Y. Yu, and P. Poupart. Quantifying and improving transferability in domain generalization. *Advances in Neural Information Processing Systems*, 34:10957–10970, 2021.

[48] M. Zhang, H. Marklund, N. Dhawan, A. Gupta, S. Levine, and C. Finn. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34:23664–23678, 2021.

[49] H. Zhao, R. T. D. Combes, K. Zhang, and G. Gordon. On learning invariant representations for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7523–7532. PMLR, 09–15 Jun 2019. URL `http://proceedings.mlr.press/v97/zhao19a.html`.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1.1.
   (b) Did you describe the limitations of your work? [Yes] See Section 7 and Appendix E.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section E.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 4, Appendix B.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code attached in supplemental file.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 6, Appendix C.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See figures 1, 2.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix C.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cited LeCun and Cortes [23].
   (b) Did you mention the license of the assets? [Yes] See Appendix C.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Additional related works

The general problem of domain generalization has been studied since Blanchard et al. [8], Muandet et al. [32]. For a domain generalization problem, it is crucial to make carefully reasoned assumptions on what remains constant and what varies across environments, as different domain shift assumptions call for different algorithms. Beyond invariance of the conditional signal feature distribution $P(\Phi(x) \mid y)$ (shared by feature-matching algorithms) and invariance of the label distribution conditioned on the signal features $P(y \mid \Phi(x))$ (shared by IRM variants) already discussed in our paper, other types of assumptions include invariant gradients [35], learnable domain transformation models [36], the test domain being a convex combination of training domains [39], etc. Some works assume various causal data models and the corresponding algorithms are inspired by causal inference, including Krueger et al. [21], Chevalley et al. [9], Ahuja et al. [1]. Zhang et al. [47] seeks to quantify and improve the transferability among domains. A related line of work is test-time training [48] or meta-learning [26] where we are not limited to using a single model for all domains but can adapt at test time.

Relation to ICA Our data model bears some semblance to ICA since the invariant and spurious features are assumed to be independent and the goal is to disentangle the two sources. However, for identifiability in linear ICA, at most one source has to be non-Gaussian, which is different from our data model. The algorithms being analyzed here are also distinct from ICA methods such as minimization of mutual information or maximization of non-Gaussianity. A line of recent work on nonlinear ICA [16, 17, 20] proves identifiability results for deep latent variable models but they require additional conditions or auxiliary supervision.

B Additional proofs

B.1 Proof sketch of Lemma 5.2

This section gives a proof sketch of the main lemma 5.2. Our goal is to show that when $E = \Omega\left(d_s/k\right)$, with high probability, no orthonormal $Q \in \mathbb{R}^{k\times d_s}$ satisfies $Q\Delta_e Q^\top = 0$ for all e, where we define the difference between the covariances of spurious features in two adjacent environments $e, e+1$ as $\Delta_2^e = \Sigma_2^e - \Sigma_2^{e+1} = (\overline{\Sigma_2^e} - \overline{\Sigma_2^{e+1}}) + (G_e G_e^\top - G_{e+1}G_{e+1}^\top)$.

Step 1: Discretization To show this, we discretize over the space of orthonormal matrices Q, and show that for fixed Q, the probability that $Q\Delta_2^e Q^\top = 0$ for all e is small. We then union bound over the covering.

Step 2: Designing a testing statistic To show that for fixed Q, $Q\Delta_e Q^\top = \mathbf{0}_{k\times k}$ is unlikely, we focus on showing that the sum of squares of all entries in $Q\Delta_e Q^\top$ is bounded away from 0. More formally, let q_i be the i-th row of Q. Define $Z_{ije} = (q_i^\top \Delta_2^e q_j)^2$ so it is $\left[Q\Delta_e Q^\top\right]_{i,j}^2$.

Define the nonrandom part $A_{ije} = q_i^\top(\overline{\Sigma_2^e} - \overline{\Sigma_2^{e+1}})q_j$, and $A = \sum A_{ije}^2$, we write

$$\sum_e \|Q\Delta_e Q^\top\|_F^2 = \sum Z_{ije} = \sum (A_{ije} + q_i^\top G_e G_e^\top q_j - q_i^\top G_{e+1}G_{e+1}^\top q_j)^2.$$

We will eventually show that with high probability, $\sum Z_{ije} = \Omega\left(A + Ek^2 d_s\right)$.

Step 3: Decoupling One difficulty is that $\sum Z_{ije}$ is not a sum of independent random variables, because $q_i^\top G_e G_e^\top q_j$ and $q_i^\top G_e G_e^\top q_{j'}$ are dependent. The key insight is to use a decoupling tool from de la Peña and Montgomery-Smith [10].

In more details, define $V_{i,e} = G_e q_i$. For fixed Q, $V_{i,e} \sim \mathcal{N}(0, I_{d_s})$ and the ensemble $\{V_{i,e}\}_{i\in[k],e\in[E]}$ is independent. Therefore

$$q_i^\top G_e G_e^\top q_j - q_i^\top G_{e+1}G_{e+1}^\top q_j = V_{i,e}^\top V_{j,e} - V_{i,e+1}^\top V_{j,e+1}$$

For further simplification, we define $X_{i,e} = [V_{i,e}; V_{i,e+1}] \in \mathbb{R}^{2d_s}$, and $I^* = [I_{d_s}, \mathbf{0}; \mathbf{0}, -I_{d_s}]$, so

$$V_{i,e}^\top V_{j,e} - V_{i,e+1}^\top V_{j,e+1} = X_{i,e}^\top I^* X_{j,e}$$

Note that $X_{i,e}^\top I^* X_{j,e}$ and $X_{i,e}^\top I^* X_{j',e}$ are dependent for $j' \neq j, j' \neq i, i \neq j$. To resolve this, We apply Lemma B.4 to decouple them. Lemma B.4 says to show concentration of sum of Z_{ije} we just need to show concentration of the sum of some other random variables Z'_{ije}, where

$$Z'_{ije} = A_{ije}^2 + 2A_{ije}X_{i,e}^\top I^* Y_{j,e} + (X_{i,e}^\top I^* Y_{j,e})^2.$$

Here $Y_{i,e}$ and $X_{i,e}$ are identically distributed.

We then turn $\{Z'_{ije}\}$ into their identically distributed counterparts $\{Z''_{ije}\}$, where

$$Z''_{ije} = A_{ije}^2 + 2A_{ije}X_{i,e}^\top Y_{j,e} + (X_{i,e}^\top Y_{j,e})^2.$$

Step 4: High probability norm bounds for $\{Y_{i,e}\}$ We first consider the randomness in $\{Y_{i,e}\}$, and prove that with high probability $\{Y_{i,e}\}$ satisfies two norm bounds; we then show the concentration of $\sum Z''_{ije}$ conditioned on the event that $\{Y_{i,e}\}$ satisfies these bounds. We pack $\{Y_{i,e}\}$ as rows of a matrix $Y_e \in \mathbb{R}^{k\times 2d}$. We design event $\mathcal{E}_1$ so that conditioned on $\mathcal{E}_1$, for all $Q \in \mathcal{Q}$ and e, $\|Y_e\|_2 = O(\sqrt{d_s})$. Event $\mathcal{E}_2$ denotes the event that for all cover elements in $\tilde{Q}$, all e, $\|Y_e\|_F^2 = \Theta(Ekd_s)$.

Step 5: Conditioned on Y_e, proving that $\sum Z''_{ije}$ concentrates Once we fix Y_e, the random variables $\{P_{ei} = \sum_j Z''_{ije}\}$ are independent, so the concentration of their sum is immediate.

$$P_{ei} = \sum A_{ije}^2 + 2X_{i,e}^\top \left(\sum_{j\neq i} A_{ije}Y_{j,e}\right) + X_{i,e}^\top Y_e Y_e^\top X_{i,e} \tag{B.1}$$

We show concentration of the second and the third terms of equation (B.1) using Hoeffding's inequality and Hanson-Wright inequality. Conditioned on good events $\mathcal{E}_1$ and $\mathcal{E}_2$, for fixed Q,

$$\Pr\left[\sum_{e,i} P_{ei} \lesssim A + Ek^2 d_s \middle| \mathcal{E}_1, \mathcal{E}_2\right] \leq \exp\left(-\min\left\{\frac{(A+Ek^2d_s)^2}{Ad_s}, Ek^2\right\}\right).$$

In other words, with high probability all $Q \in \tilde{Q}$ satisfies $\sum_{e,i} P_{ei} = O\left(A + Ek^2 d_s\right)$.

We choose ϵ to make the discretization error smaller than $O(Ek^2d_s)$. Therefore, with high probability, for all $Q \in \mathcal{Q}$, $\sum_{eij} Z_{ije} = \Omega\left(A + Ek^2 d_s\right)$, i.e. the testing statistic is bounded away from zero.

B.2 Proof of Lemma 5.1

Proof of Lemma 5.1. We shall prove that for all $t < T$, if we write $U_t S = [A_t, B_t]$ where A_t, B_t are the left r and right d_s columns of $U_t S$, then $rank((I - P_{B_t})A_t) = r$ and $k_t = rank(B_t) = r_t - r < (r_{t-1} - r + 1)/c$ with probability $1 - O(t\exp(-d_s))$. Since $T = O(d_s)$, for $t = T-1$, the probability $1 - O(T\exp(-d_s)) = 1 - \exp(-\Omega(d_s))$.

Lemma B.1 in Appendix B.2 says that if $rank((I - P_{B_t})A_t) < r$, then we can construct orthonormal U'_t with higher dimension $r'_t > r_t$ that still matches the covariances for environments $\mathcal{E}_t$. Hence IFM always finds U_t with $rank((I - P_{B_t})A_t) = r$.

To show that the number of spurious dimension decreases, we prove by induction on t. For the base case $t = 1$, Lemma 5.2 says $r_1 - r \leq k_1 < (d_s - r + 1)/c$ with probability $1 - O(\exp(-d_s))$. For $t \geq 2$, suppose to the contrary that there is orthonormal $U_t \in \mathbb{R}^{r_t \times r_{t-1}}$ satisfying (4.1) such that $U_t \dots U_1 S = [A_t, B_t]$ where $B_t \in \mathbb{R}^{r_t \times d_s}$, and $rank(B_t) = k_t > (r_{t-1} - r + 1)/c$. By induction hypothesis, with probability $1 - O((t-1)\exp(-d_s))$, we can write $U_{t-1}\dots U_1 S = [A_{t-1}, B_{t-1}]$ where $B_{t-1} \in \mathbb{R}^{r_{t-1}\times d_s}$ has rank $k_{t-1} \geq r_{t-1} - r$. Below we condition on this event.

Writing B_{t-1} in terms of the compact SVD, we get $B_{t-1} = P_{t-1}\Lambda_{t-1}Q_{t-1}$, where $Q_{t-1} \in \mathbb{R}^{k_{t-1}\times d_s}$. Therefore

$$U_t[A_{t-1}, B_{t-1}]\begin{bmatrix}\Sigma_1 & 0\\ 0 & \Sigma_2^e\end{bmatrix}[A_{t-1}, B_{t-1}]^\top U_t^T = C_t$$
$$\implies U_t B_{t-1}\Sigma_2^e B_{t-1}^\top U_t^\top = C'_t.$$

Writing $U_t P_{t-1}\Lambda_{t-1}$ in terms of its compact SVD, we get $U_t P_{t-1}\Lambda_{t-1} = P_t \Lambda_t Q_t$ where Λ_t has k^* non-zero singular values and $Q_t \in \mathbb{R}^{k^* \times k_{t-1}}$. Therefore

$$B_t = U_t B_{t-1} = P_t \Lambda_t Q_t Q_{t-1}. \tag{B.2}$$

Note that $Q = Q_t Q_{t-1} \in \mathbb{R}^{k^* \times d_s}$ satisfies $QQ^\top = I$, since both Q_t and Q_{t-1} satisfies this. Therefore (B.2) forms an SVD decomposition of B_t, and due to the uniqueness of non-zero singular values up to permutation, we have $k^* = k_t$.

Therefore $Q_t \in \mathbb{R}^{k_t \times k_{t-1}}$ satisfies $\forall e \in \mathcal{E}_t, Q_t Q_{t-1} \Sigma_2^e Q_{t-1}^\top Q^\top = C_t''$.

Applying Corollary 5.3 with $P = Q_{t-1}$, $Q = Q_t$, with probability $1 - O(\exp(-d_s))$, no Q_t satisfies $\forall e \in [E]$, $Q_t Q_{t-1} \Sigma_2^e Q_{t-1}^\top Q_t^\top = C_t$ for some constant $C_t \in \mathbb{R}^{r_t \times r_t}$.

For the last iteration $t = T$, $E_T = 3$. We assume without loss of generality $rank(B_{T-1}) = k_{T-1} \in \{1, 2\}$, since we can always half the spurious dimensions $r_t - r \leq (r_{t-1} - r)/2$ until $r_{t-1} - r = 2$.

Lemma B.8 and Lemma B.9 in Appendix B.2 deal with the cases when $k_{T-1} = 2$ and $k_{T-1} = 1$, respectively. Suppose $rank(B_{T-1}) = 2$, its associated orthonormal matrix $Q_{T-1} \in \mathbb{R}^{2 \times d_s}$. Lemma B.8 says that with yields that, with probability 1, no vector on the unit circle $q_T \in \mathbb{S}^1$ satisfies $q_T^\top Q_{T-1}(\Sigma_2^e - \Sigma_2^{e+1}) Q_{T-1}^\top q_T = 0$ for $e \in \{1, 2\}$. Suppose $rank(B_{T-1}) = 1$, its associated unit-norm vector $q_{T-1} \in \mathbb{R}^{d_s}$. Lemma B.9 says that with probability 1, no non-zero scalar q_T satisfies $q_T^2 q_{T-1}^\top (\Sigma_2^1 - \Sigma_2^2) q_{T-1} = 0$. Combining the two cases, with probability $1 - O((T-1)\exp(-d_s))$, $rank(B_T) = k_T = 0$. □

The first lemma says IFM always finds U_t that uses all invariant dimensions.

Lemma B.1. *Let $A_0 \in \mathbb{R}^{d \times r}, B_0 \in \mathbb{R}^{d \times d_s}$ be the left r and right d_s columns of S. Define projection matrix onto the column span of B, $P_{B_0} = B_0(B_0^\top B_0)^{-1} B_0^\top$. Suppose orthonormal $U_t \in \mathbb{R}^{r_t \times d}$ satisfies that $U_t S = [A_t, B_t]$ where $rank(U_t(I - P_{B_0})A_0) < r$, and for all $e \in \mathcal{E}$, $U S \Sigma^e S^\top U = C_t \in \mathbb{R}^{r_t \times r_t}$. Then there exist orthonormal $U_t' \in \mathbb{R}^{r_t' \times d}$ such that $r_t' > r_t$, $rank(U_t'(I - P_{B_0})A_0) > rank(U_t(I - P_{B_0})A_0)$, and for all $e \in \mathcal{E}$, $U_t' S \Sigma^e S^\top {U_t'}^\top = C_t' \in \mathbb{R}^{r_t' \times r_t'}$.*

Proof. We construct U_t' by adding one additional row u^+ to U_t. Denote the columns of $(I - P_{B_0})A_0$ as $a_1, \ldots, a_r \in \mathbb{R}^d$. Since $U_t(I - P_{B_0})A_0$ does not have full column rank, there is one column that can be written as linear combination of others. Assume without loss of generality that $U_t a_0 = \sum_{j=1}^r \alpha_j U_t a_j$, which implies that $U_t(a_0 - \sum_{j=1}^r \alpha_j a_j) = 0$. Since $(I - P_{B_0})A_0$ has full column rank r, $a^+ := a_0 - \sum_{j=1}^r \alpha_j a_j \neq 0$. Define $u^+ := a^+ / \|a^+\|_2$. Since $U_t a^+ = 0$, we have that $u_i^\top u^+ = 0$, for all existing rows of U_t ($i \in [r_t]$). Furthermore, since each a_j is orthogonal to the column space of B_0, ${u^+}^\top B_0 = 0$. Hence $U_t' = \begin{bmatrix} U_t \\ u^+ \end{bmatrix}$ is orthonormal, $r_t' = r_t + 1$, and $U_t' B_0 = \begin{bmatrix} U_t B_0 \\ \mathbf{0}_{1 \times d_s} \end{bmatrix}$ so

$$U_t' S \Sigma^e S^\top {U_t'}^\top = U_t' A_0 \Sigma_1 A_0^\top {U_t'}^\top + U_t' B_0 \Sigma_2^e B_0^\top {U_t'}^\top = U_t' A_0 \Sigma_1 A_0^\top {U_t'}^\top + \begin{bmatrix} U_t B_0 \Sigma_2^e B_0^\top U_t^\top & \mathbf{0}_{d_s \times 1} \\ \mathbf{0}_{1 \times d_s} & 0 \end{bmatrix}$$

which is constant for all $e \in \mathcal{E}$. □

We use the following lemma to discretize the space of orthonormal matrices $\mathcal{Q} = \{Q : QQ^\top = I_k, Q \in \mathbb{R}^{k \times d_s}\}$. For any $Q, Q' \in \mathcal{Q}$, we define the metric $\rho(Q, Q') = \|Q^\top Q - {Q'}^\top Q'\|_F$. We recall the following lemma about the existence of a cover of $\mathcal{Q}$ with respect to the metric ρ:

Lemma B.2 (Proposition 8 of Pajor [33])**.** *For $1 \leq k \leq d_s/2$, there exists absolute constant c_3 and covering $\tilde{\mathcal{Q}} \subset \mathcal{Q}$ such that for all $\epsilon > 0$, $|\tilde{\mathcal{Q}}| \leq (c_3\sqrt{k}/\epsilon)^{k(d_s - k)}$, and $\forall Q^* \in \mathcal{Q}$, $\exists Q \in \tilde{\mathcal{Q}}$ such that $\rho(Q, Q^*) \leq \epsilon$.*

For any odd integer $e < E$, define $\Delta_2^e = \Sigma_2^e - \Sigma_2^{e+1} = (\overline{\Sigma_2^e} - \overline{\Sigma_2^{e+1}}) + (G_e G_e^\top - G_{e+1} G_{e+1}^\top)$.

For any $Q \in \mathcal{Q}$, let q_i be the i-th row of Q, for $i \in [k]$. Let $Z_{ije} = (q_i^\top \Delta_2^e q_j)^2$. Define $A_{ije} = q_i^\top (\overline{\Sigma_2^e} - \overline{\Sigma_2^{e+1}}) q_j$, and $A = \sum_{\text{odd } e < E, i, j \in [k], i \neq j} A_{ije}^2$. The main lemma below shows that the sum of Z_{ije}'s are bounded away from 0.

Lemma B.3. *There exists constants $c_1, c_2, b_1, b_2 > 0$ such that for any integer $2 \le k \le d_s/2$, for all E satisfying*

$$b_1 \frac{d_s - k}{k-1} \max\left\{1, \log\left(\frac{D}{(k-1)d_s}\right), \log\left(\frac{d_s}{k-1}\right)\right\} < E < b_2 d_s,$$

where $\max_e \|\overline{\Sigma_2^e}\|_2^2 \le D$ for some constant D, with probability $1 - c_1 \exp(-d_s)$, for all $Q \in \mathcal{Q}$,

$$\sum_{\text{odd } e<E, i,j\in[k], i\neq j} Z_{ije} > c_2(A + Ek(k-1)d_s).$$

Proof. For any odd $e < E$ and $i \in [k]$, by definition

$$\sum_{j\neq i} Z_{ije} = \sum_{j\neq i} (A_{ije} + q_i^\top G_e G_e^\top q_j - q_i^\top G_{e+1} G_{e+1}^\top q_j)^2$$

Define $V_{i,e} = G_e q_i$ for $e \in [E], i \in [k]$. For fixed orthonormal Q, $V_{i,e} \sim \mathcal{N}(0, I_{d_s})$ and the ensemble $\{V_{i,e}\}_{i\in[k], e\in[E]}$'s is independent. Therefore

$$q_i^\top G_e G_e^\top q_j - q_i^\top G_{e+1} G_{e+1}^\top q_j = V_{i,e}^\top V_{j,e} - V_{i,e+1}^\top V_{j,e+1}$$

For further simplification, we define $W_{i,e} = [V_{i,e}; V_{i,e+1}] \in \mathbb{R}^{2d_s}$, and $I^* = [I_{d_s}, \mathbf{0}; \mathbf{0}, -I_{d_s}]$, so

$$V_{i,e}^\top V_{j,e} - V_{i,e+1}^\top V_{j,e+1} = W_{i,e}^\top I^* W_{j,e}$$

We use the following lemma to decouple the correlations between $W_{i,e}^\top I^* W_{j,e}$ and $W_{i,e}^\top I^* W_{j',e}$ for $j' \neq j, j' \neq i, i \neq j$:

Lemma B.4 (Theorem 1 of de la Peña and Montgomery-Smith [10])**.** *Suppose $\{X_i\}$ ($i \in [k]$) are independent random variables, X_i and Y_i have the same distribution. There exists some absolute constant c_4 such that*

$$\Pr\left[\left|\sum_{i,j\in[k], i\neq j} f(X_i, X_j)\right| \ge t\right] \le c_4 \Pr\left[\left|\sum_{i,j\in[k], i\neq j} f(X_i, Y_j)\right| \ge t/c_4\right].$$

We apply Lemma B.4 with $X_i = W_{i,e}$ and $f(X_i, X_j) = Z_{ije} - \mathbb{E}[Z_{ije}]$ to get

$$\Pr\left[\left|\sum_{i,j\in[k], i\neq j} Z_{ije} - \mathbb{E}[Z_{ije}]\right| \ge t\right] \le c_4 \Pr\left[\left|\sum_{i,j\in[k], i\neq j} Z'_{ije} - \mathbb{E}[Z'_{ije}]\right| \ge t/c_4\right].$$

where $Y_{i,e}$ and $X_{i,e}$ are identically distributed and

$$Z'_{ije} = A_{ije}^2 + 2A_{ije} X_{i,e}^\top I^* Y_{j,e} + (X_{i,e}^\top I^* Y_{j,e})^2.$$

Note that $\{Z'_{ije}\}$ and $\{Z''_{ije}\}$ are identically distributed, where

$$Z''_{ije} = A_{ije}^2 + 2A_{ije} X_{i,e}^\top Y_{j,e} + (X_{i,e}^\top Y_{j,e})^2.$$

Below we first consider the randomness in $\{Y_{i,e}\}$, and prove that with high probability $\{Y_{i,e}\}$ satisfies some good properties; we then show the concentration of $\sum_{i,j,e} Z''_{ije}$ conditioned on the event that $\{Y_{i,e}\}$ satisfies these properties.

First, for fixed Q, since $Y_{i,e} = [G_e v_i; G_{e+1} v_i] \sim \mathcal{N}(0, I_{2d_s})$, if we write $Y_e = [Y_{1,e}; \ldots; Y_{k,e}] \in \mathbb{R}^{k\times 2d_s}$, it is a random matrix with iid standard normal entries. We show that the $\|Y_e\|_F^2 = \Theta(kd_s)$ with high probability. The following lemma is a standard concentration bound for chi-squared variable:

Lemma B.5 (Corollary of Lemma 1 in Laurent and Massart [22])**.** *Suppose $Z_i \sim \mathcal{N}(0,1)$ for $i \in [n]$. For any $t > 0$,*

$$\Pr\left[\sum_{i=1}^n Z_i^2 \ge n + 2\sqrt{nt} + 2t\right] \le \exp(-t),$$

$$\Pr\left[\sum_{i=1}^n Z_i^2 \le n - 2\sqrt{nt}\right] \le \exp(-t).$$

Applying Lemma B.5 to $n = Ekd_s$ entries of $\{Y_e\}_{e=1}^E$ and setting $t = Ekd_s/16$ we get with probability $1 - 2\exp(-Ekd_s/16)$,

$$\frac{Ekd_s}{2} \leq \sum_e \|Y_e\|_F^2 \leq \frac{13Ekd_s}{8}. \tag{B.3}$$

Second, we show that with high probability over the randomness of G_e, $\|Y_e\|_2$ viewed as a function of Q satisfies $\|Y_e\|_2 = O(\sqrt{d_s})$ for all orthonormal Q. We use the following lemma to upper bound $\|G_e\|_2$:

Lemma B.6 (Corollary 5.35 of Vershynin [45]). *Suppose $G \in \mathbb{R}^{D\times d}$ and $[G]_{ij} \sim \mathcal{N}(0,1)$ for all $i \in [D], j \in [d]$. For every $t \geq 0$, with probability $1 - 2\exp(-t^2/2)$,*

$$\|G\|_2 \leq \sqrt{D} + \sqrt{d} + t$$

Applying Lemma B.6 with $G = [G_e; G_{e+1}]$, $D = 2d_s$, $d = d_s$, $t = \sqrt{d_s}$, we get with probability $1 - 2\exp(-d_s/2)$, $\|G\|_2 \leq (2+\sqrt{2})\sqrt{d_s}$, and therefore for all orthonormal $Q \in \mathbb{R}^{k\times d_s}$,

$$\|Y_e\|_2 = \|QG^\top\|_2 \leq \|Q\|_2\|G\|_2 \leq (2+\sqrt{2})\sqrt{d_s}. \tag{B.4}$$

For any odd $e < E$, $i \in [k]$, and fixed Y_e, we prove $P_{ei} = \sum_{j\neq i} Z''_{ije}$ concentrates. Once we fix Y_e, the $Er/2$ random variables $\{P_{ei}\}$ are independent, so the concentration of their sum is immediate. Let $Y_{-i,e}$ be Y_e without the i-th row,

$$P_{ei} = \sum_{j\neq i} Z''_{ije} = \sum_{j\neq i} A_{ije}^2 + 2X_{i,e}^\top \left(\sum_{j\neq i} A_{ije}Y_{j,e}\right) + X_{i,e}^\top Y_{-i,e}Y_{-i,e}^\top X_{i,e} \tag{B.5}$$

Define $B_{i,e} = Y_{-i,e}Y_{-i,e}^\top$. Let $a_{i,e} \in \mathbb{R}^{k-1}$ be the column vector consisting of A_{ije} for $j \neq i$.

Since $X_{i,e} \sim \mathcal{N}(0, I_{2d_s})$, $X_{i,e}^\top \left(\sum_{j\neq i} A_{ije}Y_{j,e}\right)$ is a Gaussian variable with mean 0 and variance $a_{i,e}^\top B_{i,e}a_{i,e} \leq \|a_{i,e}\|_2^2\|B_{i,e}\|_2$, so by Hoeffding's inequality, for all $t \geq 0$,

$$\Pr\left[2X_{i,e}^\top \left(\sum_{j\neq i} A_{ije}Y_{j,e}\right) > t \mid Y_e\right] \leq \exp\left(-\frac{t^2}{8\|a_{i,e}\|^2\|B_{i,e}\|_2}\right). \tag{B.6}$$

By Hanson-Wright Inequality (e.g. Theorem 1.1 of Rudelson and Vershynin [38]), there exists constant c_5 such that

$$\Pr\left[\mathbb{E}[X_{i,e}^\top B_{i,e}X_{i,e}] - X_{i,e}^\top B_{i,e}X_{i,e} > t \mid Y_e\right] \leq \exp\left(-c_5 \min\left\{\frac{t^2}{\|B_{i,e}\|_F^2}, \frac{t}{\|B_{i,e}\|_2}\right\}\right). \tag{B.7}$$

Combining equations (B.5), (B.6), (B.7), we get

$$\Pr\left[\mathbb{E}[P_{ei}] - P_{ei} > t \mid Y_e\right] \leq \exp\left(-\frac{t^2}{32\|a_{i,e}\|_2^2\|B_{i,e}\|_2}\right) + \exp\left(-c_5 \min\{\frac{t^2}{4\|B_{i,e}\|_F^2}, \frac{t}{2\|B_{i,e}\|_2}\}\right).$$

Summing over all $e \in [E]$ and $i \in [k]$ we get

$$\Pr\left[\mathbb{E}\left[\sum_{e,i} P_{ei}\right] - \sum_{e,i} P_{ei} > t \mid Y_1, \ldots, Y_E\right] \leq \exp\left(-\frac{t^2}{32\sum_{e,i}\|a_{i,e}\|_2^2\|B_{i,e}\|_2}\right)$$
$$+ \exp\left(-c_5 \min\left\{\frac{t^2}{4\sum_{e,i}\|B_{i,e}\|_F^2}, \frac{t}{2\max_{e,i}\|B_{i,e}\|_2}\right\}\right).$$

Note that $\mathbb{E}[X_{i,e}^\top BX_{i,e}] = \mathbb{E}\sum_{j\neq i}(X_{i,e}^\top Y_{j,e})^2 = \|Y_{-i,e}\|_F^2$ so

$$E\left[\sum_{e,i} P_{ei}\right] = \sum_{e,i}\|a_{i,e}\|_2^2 + \sum_{e,i}\|Y_{-i,e}\|_F^2 = A + (k-1)\sum_e \|Y_e\|_F^2.$$

Since $\|B_{i,e}\|_2 \leq \|Y_e\|_2^2$, $\|B_{i,e}\|_F^2 \leq \|Y_{-i,e}\|_F^2 \|Y_e\|_2^2$, taking $t = \frac{1}{2}E[\sum_{e,i} P_{ei}]$,

$$\Pr\left[\sum_{e,i} P_{ei} < \frac{1}{2}\left(A + (k-1)\sum_e \|Y_e\|_F^2\right) \mid Y_1, \ldots, Y_E\right] \leq \exp\left(-\frac{(A + (k-1)\sum_e \|Y_e\|_F^2)^2}{128\sum_{e,i}\|a_{i,e}\|_2^2 \|Y_e\|_2^2}\right)$$
$$+ \exp\left(-c_5 \min\left\{\frac{(k-1)^2(\sum_e \|Y_e\|_F^2)^2}{16(k-1)\sum_e \|Y_e\|_F^2 \|Y_e\|_2^2}, \frac{(k-1)\sum_e \|Y_e\|_F^2}{2\max_e \|Y_e\|_2^2}\right\}\right).$$

Let $\mathcal{E}_1$ denote the event that for all odd $e < E$, $[G_e; G_{e+1}] \in \mathbb{R}^{2d_s \times d_s}$ denote the matrix with $G_e, G_{e+1} \in \mathbb{R}^{d_s \times d_s}$ in its first and last d_s rows, respectively, we have

$$\|[G_e; G_{e+1}]\|_2 \leq (2+\sqrt{2})\sqrt{d_s}.$$

Due to equation (B.4) and the union bound, $\Pr[\mathcal{E}_1] \geq 1 - E\exp(-d_s/2)$. Conditioned on $\mathcal{E}_1$, for all $Q \in \mathcal{Q}$ and odd $e < E$,

$$\|Y_e\|_2 \leq (2+\sqrt{2})\sqrt{d_s}.$$

Let $\mathcal{E}_2$ denote the event that for all cover elements $Q \in \tilde{Q}$,

$$\frac{Ekd_s}{2} \leq \sum_e \|Y_e\|_F^2 \leq \frac{13Ekd_s}{8}.$$

Due to equation (B.3) and the union bound, $\Pr[\mathcal{E}_2] \geq 1 - 2|\tilde{Q}|\exp(-Ekd_s/16)$.

Conditioned on $\mathcal{E}_1$ and $\mathcal{E}_2$, for fixed $Q \in \tilde{Q}$, there exists constants c_6, c_7 such that

$$\Pr\left[\sum_{e,i} P_{ei} < \frac{1}{2}A + \frac{1}{4}Ek(k-1)d_s\right] \leq \exp\left(-c_6\frac{(A + Ek(k-1)d_s)^2}{Ad_s}\right)$$
$$+ \exp\left(-c_7 \min\left\{\frac{(k-1)^2E^2k^2d_s^2}{Ek(k-1)d_s^2}, \frac{Ek(k-1)d_s}{d_s}\right\}\right),$$

which implies there exists constants c_8 such that

$$\Pr\left[\sum_{e,i} P_{ei} < \frac{1}{2}A + \frac{1}{4}Ek(k-1)d_s\right] \leq \exp\left(-c_8 \min\left\{\frac{(A + Ek(k-1)d_s)^2}{Ad_s}, Ek(k-1)\right\}\right).$$

Note that we always have $\frac{(A+Ek(k-1)d_s)^2}{Ad_s} \geq Ek(k-1)$. To see this, for $A > Ek(k-1)d_s$, $\frac{(A+Ek(k-1)d_s)^2}{Ad_s} > \frac{A}{d_s} > Ek(k-1)$. For $A \leq Ek(k-1)d_s$, $\frac{(A+Ek(k-1)d_s)^2}{Ad_s} \geq \frac{(Ek(k-1)d_s)^2}{Ek(k-1)d_s^2} = Ek(k-1)$.

In other words, with probability $1-\delta$, where

$$\delta = E\exp(-d_s/2) + 2|\tilde{Q}|\exp(-Ekd_s/16) + |\tilde{Q}|\exp(-c_8Ek(k-1)),$$

all $Q \in \tilde{Q}$ satisfies $\sum_{e,i} P_{ei} \geq \frac{1}{4}(A + Ek(k-1)d_s)$. Combined with Lemma B.4, with probability $1 - c_9\delta$, all $Q \in \tilde{Q}$ satisfies $\sum_{e,i,j} Z_{ije} < c_{10}(A + Ek(k-1)d_s)$ for some constants c_9, c_{10}.

For any $Q^* \in \mathcal{Q}$, let Q be the element in the cover closest to it, so that $\rho(Q, Q^*) = \|Q^\top Q - Q^{*\top}Q^*\|_F \leq \epsilon$. Let q_i^* be the i-th row of Q^*, and $Z_{ije}^* = (q_i^{*\top}\Delta_2^e q_j^*)$. Then

$$\begin{aligned}
\sum_{eij} Z_{ije}^* &= \sum_e \|Q^*\Delta_2^e Q^{*\top}\|_F^2 \\
&= \sum_e \|\Delta_2^e Q^{*\top}Q^*\|_F^2 \\
&\geq \frac{1}{2}\sum_e \|\Delta_2^e Q^\top Q\|_F^2 - \|\Delta_2^e\left(Q^\top Q - Q^{*\top}Q^*\right)\|_F^2 \\
&\geq \frac{1}{2}\sum_{eij} Z_{ije} - \|\Delta_2^e\|_2^2 \rho(Q, Q^*)^2.
\end{aligned}$$

Since $\|\Delta_2^e\|_2^2 \leq 2\|\overline{\Sigma_2^e}\|_2^2 + 2\|G_e G_e^\top\|_2^2$, and conditioned on $\mathcal{E}_1$, $\|G_e G_e^\top\|_2^2 \leq c_{11} d_s^2$ for all e, if $\max_e \|\overline{\Sigma_2^e}\|_2^2 \leq D$ for some constant D, we have with probability $1-\delta$,

$$\sum_{eij} Z_{ije}^* \geq \frac{c_{10}}{2}(A + Ek(k-1)d_s) - 2E(D + c_{11}d_s^2)\epsilon^2. \tag{B.8}$$

We choose $\epsilon^2 < \frac{c_{10}k(k-1)d_s}{8(D+c_{11}d_s^2)}$ so that $2E(D + c_{11}d_s^2)\epsilon^2 < \frac{c_{10}}{4}Ek(k-1)d_s$.

With this choice of ϵ, by equation (B.8) we have

$$\sum_{eij} Z_{ije}^* \geq \frac{c_{10}}{4}(A + Ek(k-1)d_s).$$

By Lemma B.2, $\log(|\tilde{Q}|) \leq k(d_s - k)\log(c_3\sqrt{k}/\epsilon) \leq c_{12}k(d_s-k)\log\left(\frac{D}{(k-1)d_s} + \frac{d_s}{k-1}\right)$.

Therefore there exists $b_1, b_2 > 0$ such that for E satisfying

$$b_2 d_s > E > b_1 \frac{d_s - k}{k-1} \max\left\{1, \log\left(\frac{D}{(k-1)d_s}\right), \log\left(\frac{d_s}{k-1}\right)\right\},$$

we have

$$\begin{aligned}
\delta &\leq \exp(-d_s/2 + \log(b_2 d_s)) + 2\exp\left(c_{12}k(d_s-k)\log\left(\frac{D}{(k-1)d_s} + \frac{d_s}{k-1}\right) - Ekd_s/16\right) \\
&\quad + \exp\left(c_{12}k(d_s-k)\log\left(\frac{D}{(k-1)d_s} + \frac{d_s}{k-1}\right) - c_8 Ek(k-1)\right) \\
&\leq c_1 \exp(-d_s)
\end{aligned}$$

for some constant c_1. Therefore with probability $1 - c_1\exp(-d_s)$, for all $Q^* \in \mathcal{Q}$, and $c_2 = c_{10}/4$,

$$\sum_{eij} Z_{ije}^* \geq c_2(A + Ek(k-1)d_s).$$

□

Corollary B.7 (Corollary of Lemma B.3)**.** *Suppose $2 \leq k \leq r/2 \leq d_s/2$. Let $\mathcal{P} = \{P \in \mathbb{R}^{r\times d_s} : PP^\top = I_r\}$, $\mathcal{Q} = \{Q \in \mathbb{R}^{k\times r} : QQ^\top = I_k\}$. For fixed $P \in \mathcal{P}$, there exists constants $c_1, c_2, b_1, b_2 > 0$ such that for all E satisfying*

$$b_1 \frac{r-k}{k-1} \max\left\{1, \log\left(\frac{D}{(k-1)d_s}\right), \log\left(\frac{d_s}{k-1}\right)\right\} < E < b_2 d_s,$$

where $\max_e \|\overline{\Sigma_2^e}\|_2^2 \leq D$ for some constant D, with probability $1 - c_1\exp(-d_s)$, for all $Q \in \mathcal{Q}$,

$$\sum_{\text{odd } e < E} \|QP\Delta_2^e P^\top Q^\top\|_F^2 > c_2 Ek(k-1)d_s.$$

Proof. The proof mostly follows that of Lemma B.3, with a few modifications below. We discretize over $\mathcal{Q}$ and get a ϵ-covering $\tilde{Q}$ of size $(c_3\sqrt{k}/\epsilon)^{r(r-k)}$.

For any $Q \in \mathcal{Q}$, let v_i be the i-th row of QP and define Z_{ije}, A_{ije} accordingly. For any $Q^* \in \mathcal{Q}$, let Q be its cover element, so $\rho(Q, Q^*) = \|Q^\top Q - Q^{*\top}Q^*\|_F \leq \epsilon$. Let q_i^* be the i-th row of Q^*P, and $Z_{ije}^* = (q_i^{*\top}\Delta_2^e q_j^*)$. Then

$$\begin{aligned}
\sum_{eij} Z_{ije}^* &= \sum_e \|Q^*P\Delta_2^e P^\top Q^{*\top}\|_F^2 \\
&= \sum_e \|P\Delta_2^e P^\top Q^{*\top}Q^*\|_F^2 \\
&\geq \frac{1}{2}\sum_e \|P\Delta_2^e P^\top Q^\top Q\|_F^2 - \|P\Delta_2^e P^\top\left(Q^\top Q - Q^{*\top}Q^*\right)\|_F^2 \\
&\geq \frac{1}{2}\sum_{eij} Z_{ije} - \|P\Delta_2^e P^\top\|_2^2 \rho(Q, Q^*)^2 \\
&\geq \frac{1}{2}\sum_{eij} Z_{ije} - \|\Delta_2^e\|_2^2 \rho(Q, Q^*)^2
\end{aligned}$$

Thus with the same choice of ϵ as Lemma B.3, $\log(|\tilde{Q}|) \leq k(r-k)\log(c_3\sqrt{k}/\epsilon) \leq c_{12}k(r-k)\log\left(\frac{D}{(k-1)d_s} + \frac{d_s}{k-1}\right)$. The rest of the argument is identical. □

Lemma B.8. *Let $\mathcal{P} = \{P \in \mathbb{R}^{2\times d_s} : PP^\top = I_2\}$. Suppose $\Sigma_2 = \overline{\Sigma_2^1} - \overline{\Sigma_2^2} + G_1G_1^\top - G_2G_2^\top$ and $\Sigma_2' = \overline{\Sigma_2^1} - \overline{\Sigma_2^3} + G_1G_1^\top - G_3G_3^\top$, where $G_e \in \mathbb{R}^{d_s\times d_s}$ and $[G_e]_{ij} \sim \mathcal{N}(0,1)$ for all $e \in [3]$, $i,j \in [d_s]$. For fixed $P \in \mathcal{P}$, with probability 1, no vector $q \in \mathbb{R}^2$ satisfies $\|q\|_2 = 1$ and*

$$q^\top \Sigma_2 q = 0, \quad q^\top \Sigma_2' q = 0.$$

Proof. For any fixed G_1, G_2, consider the system of quadratic equations over two variables,

$$\{q^\top \Sigma_2 q = 0, \|q\|_2 = 1\}.$$

With probability 1, it has at most 4 real solutions. Conditioned on G_1, G_2, consider the third quadratic equation where the randomness is in G_3.

$$\{q^\top \Sigma_2' q = 0\}.$$

With probability 1, any fixed solution from the first system does not satisfy this. □

The following lemma is trivial so proof is omitted:

Lemma B.9. *Suppose $p \in \mathbb{R}^{d_s}$ and $\|p\|_2 = 1$. Suppose $\Sigma_2 = \overline{\Sigma_2^1} - \overline{\Sigma_2^2} + G_1G_1^\top - G_2G_2^\top$, where $G_e \in \mathbb{R}^{d_s\times d_s}$ and $[G_e]_{ij} \sim \mathcal{N}(0,1)$ for $e \in [2]$, $i,j \in [d_s]$. With probability 1, no scalar $q \neq 0$ satisfies*

$$q^2 p^\top \Sigma_2 p = 0.$$

B.3 Proof of Theorem 4.2

Proof. Denote the unit-norm classifier β. For any environment with mean (μ_1, μ_2^i) and covariance Σ_1, Σ_2^i, the accuracy of β can be written

$$\begin{aligned}
\mathbb{E}[\mathbf{1}(\text{sgn}(\beta^\top x) = y] &= p(y=1)p(\beta^\top x \geq 0 \mid y=1) + p(y=-1)p(\beta^\top x < 0 \mid y=-1) \\
&= \frac{1}{2}\left[1 - \Phi\left(-\frac{\beta_1^\top \mu_1 + \beta_2^\top \mu_2^i}{\sqrt{\beta_1^\top \Sigma_1 \beta_1 + \beta_2^\top \Sigma_2^i \beta_2}}\right)\right] + \frac{1}{2}\Phi\left(\frac{\beta_1^\top \mu_1 + \beta_2^\top \mu_2^i}{\sqrt{\beta_1^\top \Sigma_1 \beta_1 + \beta_2^\top \Sigma_2^i \beta_2}}\right) \\
&= \Phi\left(\frac{\beta_1^\top \mu_1 + \beta_2^\top \mu_2^i}{\sqrt{\beta_1^\top \Sigma_1 \beta_1 + \beta_2^\top \Sigma_2^i \beta_2}}\right),
\end{aligned}$$

where Φ is the standard normal CDF. Observe that Φ is monotone and that $\sigma_2^2 I \preceq \Sigma_2^i$. Therefore, a training accuracy of at least γ on each environment implies that for each environment,

$$\begin{aligned}
\gamma &\leq \Phi\left(\frac{\beta_1^\top \mu_1 + \beta_2^\top \mu_2^i}{\sqrt{\beta_1^\top \Sigma_1 \beta_1 + \beta_2^\top \Sigma_2^i \beta_2}}\right) \\
&\leq \Phi\left(\frac{\beta_1^\top \mu_1 + \beta_2^\top \mu_2^i}{\sqrt{\sigma_1^2\|\beta_1\|^2 + \sigma_2^2\|\beta_2\|^2}}\right).
\end{aligned}$$

For brevity, moving forward we will denote $\psi := \sqrt{\sigma_1^2\|\beta_1\|^2 + \sigma_2^2\|\beta_2\|^2}$. Applying the inverse CDF (which is also monotone) and rearranging, we have

$$\beta_2^\top \mu_2^i \geq \psi\Phi^{-1}(\gamma) - \beta_1^\top \mu_1,$$

which implies

$$\beta_1^\top \mu_1 - \beta_2^\top \mu_2^i \leq 2\beta_1^\top \mu_1 - \psi\Phi^{-1}(\gamma).$$

If $\gamma \geq \Phi\left(\frac{2\|\mu_1\|}{\min(\sigma_1,\sigma_2)}\right) \geq \Phi\left(\frac{2\beta_1^\top \mu_1}{\psi}\right)$ then we have $\beta_1^\top \mu_1 - \beta_2^\top \mu_2^i \leq 0$ for all environments and therefore the classifier has accuracy $< \frac{1}{2}$ on all test environments. □

B.4 Proof of Theorem 4.3

Definition B.10. *For a positive definite matrix $A \in Mat_{d\times d}(\mathbb{R})$ and vector $b \in \mathbb{R}^d$, the associated ellipsoid $E_{A,b} \subseteq \mathbb{R}^d$ is given by*

$$E_{A,b} = \{x \in \mathbb{R}^d : x^\top Ax - b^\top x = 0\}.$$

Observe that the origin is contained in any such ellipsoid $E_{A,b}$. Therefore, any collection of ellipsoids E_{A_i,b_i} has the origin as a trivial point in its intersection. Our main result ensures the existence of another (non-trivial) intersection of any d such ellipses whenever the vectors b_i are linearly independent.

Theorem B.11. *If $b_1, \dots, b_d \in \mathbb{R}^d$ are linearly independent and $A_1, \dots, A_d$ are positive-definite matrices, then*

$$\left|\bigcap_{i=1}^{d} E_{A_i,b_i}\right| \geq 2.$$

To prove this result we use technical tools from differential topology. The most central tool, Proposition B.15, ensures that the total number of intersection points between two manifolds of complementary dimensions $k, d-k$ is even when certain generic tranversality conditions hold. Using these techniques, we show that $\left|\bigcap_{i=1}^{d} E_{A_i,b_i}\right| \geq 2$ for almost all matrices $A_1, \dots, A_d$, as long as $b_1, \dots, b_d$ are linearly independent. Then we use a continuity argument to extend the result to all positive definite matrices $A_1, \dots, A_d$.

Throughout we say a function is *smooth* to mean it is infinitely differentiable, i.e. C^∞. All manifolds considered are smooth, i.e. they have a smooth structure. When $F(x, y)$ has two arguments we denote by F_x the function $F_x(y) = F(x, y)$ of y given by fixing x, and similarly define F_y. If $x \in X$ is a point in the smooth manifold X, we denote by $T_x(X)$ its *tangent space*, which is intuitively the vector space of all tangent vectors to X at x. The derivative of a smooth map $f : X \to Y$ at $x \in X$ is a linear map $df_x : T_x(X) \to T_{f(x)}(Y)$.

Definition B.12. *[14, Chapter 1.5]*

Let X, Y, Z be smooth manifolds (without boundary) such that $Z \subseteq Y$. The smooth map $f : X \to Y$ is tranverse *to Z if for each $x \in X$ with $f(x) \in Z$, it holds that*

$$\textit{Image}(df_x) + T_{f(x)}(Z) = T_{f(x)}(Y).$$

If $X, Z \subseteq Y$ are both submanifolds of Y, we say they are transverse if the inclusion $\iota_X : X \hookrightarrow Y$ is transverse to Z. Equivalently, this means that for any $x \in X \cap Z$,

$$T_x(X) + T_x(Z) = T_x(Y).$$

Roughly speaking, smooth two manifolds X, Z are transversal if all intersection points are "typical". For example, if $\dim(X) + \dim(Z) < \dim(Y)$, then X, Z being transverse is equivalent to their intersection being empty. This corresponds to the intuition that their total dimension is too small for them to generically intersect. If $\dim(X) + \dim(Z) = \dim(Y)$, transversality rules out "unstable" intersections such as a line tangent to a circle.

Proposition B.13. *[14, Chapter 1.5]*

The intersection $W = X \cap Z$ of two transversal submanifolds $X, Z \subseteq Y$ is itself a submanifold of Y, and $\dim(W) = \dim(X) + \dim(Z) - \dim(Y)$.

Proposition B.14. *[14, Chapter 2.3]*

Suppose that $F : X \times S \to Y$ is a smooth map of manifolds, and let Z be a sub-manifold of Y. If F is transversal to Z, then for almost every $s \in S$, the map $f_s = F(\cdot, s) : X \to Y$ is also transversal to Z.

Proposition B.15. *[14, Chapter 2.4, Exercise 5]*

Suppose the smooth, compact manifolds $X, Y \subseteq \mathbb{R}^d$ are transversal, and that $\dim(X)+\dim(Y) = d$. Then $|X \cap Y|$ is finite and even.

Remark B.16. *Proposition B.15 follows from the methods of [14, Chapter 2.4], which shows that the parity of $|X \cap Y|$ is invariant under homotopy as long as transversality is enforced. One simply argues that by a homotopy $X \to X', Y \to Y'$, we can arrange that $|X' \cap Y'| = 0$ by translating X far away and invoking compactness.*

Lemma B.17. *The tangent space $T_0 E_{A,b}$ is exactly the orthogonal complement $b^\perp$.*

Proof. Since $E_{A,b}$ is an ellipsoid, it is a smooth manifold of dimension $d-1$. If $\gamma : [0,1] \to E_{A,b}$ is a smooth curve with $\gamma(0) = 0$, then we claim $\langle b, \gamma'(t) \rangle = 0$. This suffices to prove the desired result since $\gamma'(t)$ can be any vector in $T_0 E_{A,b}$. Indeed, differentiating the equation for $E_{A,b}$ gives

$$\begin{aligned} 0 &= 2\frac{d}{dt}\langle 0, A\gamma(t)\rangle \\ &= \frac{d}{dt}\langle \gamma(t), A\gamma(t)\rangle|_{t=0} \\ &= \frac{d}{dt}\langle b, \gamma(t)\rangle|_{t=0} \\ &= \langle b, \gamma'(t)\rangle|_{t=0}. \end{aligned}$$

□

Set $\mathcal{A}^\circ$ to be the set of all $d \times d$ strictly positive-definite matrices with distinct eigenvalues. Note that $\mathcal{A}^\circ$ is open in the space of all positive definite matrices, and its complement has Lebesgue measure 0. Denote by $\mathbb{S}^{d-1} \subseteq \mathbb{R}^d$ the unit sphere so that $(c_1, \ldots, c_d) \in \mathbb{S}^{d-1}$ if and only if $\sum_{i=1}^d c_i^2 = 1$.

Proposition B.18. *[40, Theorem 5.3]*

For any $A_0 \in \mathcal{A}^\circ$, there is an open neighborhood $U_{A_0} \subseteq \mathcal{A}^\circ$ of A_0 such that the eigenvalues $\lambda_1(A) > \cdots > \lambda_d(A)$ and associated orthonormal eigenvectors $v_1, \ldots, v_d$ can be chosen to depend smoothly on the entries of $A \in U_{A_0}$.

We remark that is it impossible to make a *globally* smooth choice of the eigenvectors and eigenvalues as above. This is because of problems caused by higher multiplicity eigenvalues, and also by the need to choose a sign for the eigenvectors.

Lemma B.19. *For $A \in \mathcal{A}^\circ$ and non-zero $b \in \mathbb{R}^d$, let $\lambda_1 > \cdots > \lambda_d$ be the eigenvalues of A, with associated orthonormal eigenvectors $v_1, \ldots, v_d$. Then $x \in E_{A,b}$ if and only if $x = x_0 + x_1$ where $x_0 = \frac{A^{-1}b}{2}$ and*

$$x_1 = \frac{\sqrt{b^\top A^{-1} b}}{2} \sum_{i=1}^d \frac{c_i v_i}{\sqrt{\lambda_i}}$$

for $(c_1, \ldots, c_d) \in \mathbb{S}^{d-1}$.

Proof. Writing $x = x_0 + x_1$, we derive

$$\begin{aligned} x_1^\top A x_1 + x_1^\top b + \frac{b^\top A^{-1} b}{4} &= x_1^\top A x_1 + 2x_1^\top A x_0 + x_0 A x_0 \\ &= x^\top A x \\ &= b^\top (x_1 + x_0) \\ &= b^\top x_1 + \frac{b^\top A^{-1} b}{2}. \end{aligned} \tag{B.9}$$

Since we used the condition $x \in E_{A,b}$ only in reaching line (B.9), the initial and final expressions are equal if and only if $x \in E_{A,b}$. It follows that $x = x_0 + x_1 \in E_{A,b}$ if and only if

$$x_1^\top A x_1 = \frac{b^\top A^{-1} b}{4}.$$

This easily leads to the parametrization given and concludes the proof. □

Lemma B.20. *Let $M^k \subseteq \mathbb{R}^d$ be a compact manifold of dimension $k \geq 1$ passing through the origin, and such that $T_0(M^k) \subsetneq b^\perp$. Then for all but a measure-zero set of positive-definite matrices A, the ellipsoid $E_{A,b}$ is transversal to M^k.*

Proof of Lemma B.20. Fixing $A_0 \in \mathcal{A}^\circ$, Proposotion B.18 ensures the existence of an open neighborhood $U_{A_0} \subseteq \mathcal{A}^\circ$ of A_0 on which the eigenvalues $\lambda_1(A) > \lambda_2(A) > \cdots > \lambda_d(A)$ and associated orthonormal eigenvectors $v_1(A), \ldots, v_d(A)$ are defined smoothly on all $A \in U_{A_0}$. Define $F : U_A \times \mathbb{S}^{d-1} \to \mathbb{R}^d$ by:

$$F(A, (c_1, \ldots, c_d)) = \frac{A^{-1}b}{2} + \frac{\sqrt{b^\top A^{-1} b}}{2} \sum_{i=1}^{d} \frac{c_i v_i(A)}{\sqrt{\lambda_i(A)}}.$$

Lemma B.19 implies that for each fixed A we obtain a diffeomorphism $F_A : \mathbb{S}^{n-1} \to E_{A,b}$. Moreover, F is smooth by construction. We claim that F and M^k are transversal. To check this, we must verify that for any $z = F(A, c) \in M^k$, it holds that

$$\text{Image}\left(dF \circ T_{F^{-1}(z)}(U_{A_0} \times \mathbb{S}^{N-1})\right) + T_z(M^k) = \mathbb{R}^d.$$

First, recall that fixing $A = A_0$, the map $F_{A_0} : \mathbb{S}^{n-1} \to E_{A_0,b}$ is a diffeomorphism. Therefore

$$\text{Image}\left(dF \circ T_{F^{-1}(z)}(U_{A_0} \times \mathbb{S}^{N-1})\right)$$

contains the tangent space $T_z(E_{A,b}) = b^\perp$ of $E_{A,b}$ at z. When $z = 0$ is the origin, the assumption $T_0(M^k) \subsetneq b^\perp$ implies

$$\dim\left(\text{Image}\left(dF \circ T_{F^{-1}(z)}(U_{A_0} \times \mathbb{S}^{N-1})\right) + T_z(M^k)\right) \geq \dim(b^\perp) + 1 = d$$

and the claim follows. Supposing for the remainder of the proof that $z \neq 0$ is not the zero vector, we claim that in fact

$$\text{Image}\left(dF \circ T_{F^{-1}(z)}(U_{A_0} \times \mathbb{S}^{N-1})\right) + T_z(M^k) = \mathbb{R}^d,$$

i.e. the tangent space of M^k is unnecessary. Indeed fixing $c \in \mathbb{S}^{N-1}$, we may vary $A \in U_A$ along the path $\gamma_A(t) = \frac{A}{t}$ for $t \in (1 - \varepsilon, 1 + \varepsilon)$. It is not difficult to see directly that

$$F(tA, c) = tF(A, c).$$

Therefore differentiating F along γ gives

$$\frac{d}{dt} F(\gamma_A(t), (c_1, \ldots, c_d))|_{t=1} = F(A, c).$$

This means $z \in \text{Image}\left(dF \circ T_{F^{-1}(z)}(U_{A_0} \times \mathbb{S}^{N-1})\right) + T_z(M^k)$. Because $E_{A,b}$ is strictly convex and passes through the origin, it follows that the tangent hyperplane to $E_{A,b}$ at z does not pass through the origin, hence $z \notin T_z(E_{A,b})$. We have establish that $\text{Image}\left(dF \circ T_{F^{-1}(z)}(U_{A_0} \times \mathbb{S}^{N-1})\right) + T_z(M^k)$ contains both $T_z(E_{A,b})$ and $z \notin T_z(E_{A,b})$. Since $\dim\left(T_z(E_{A,b})\right) = d - 1$ it follows that $\text{Image}\left(dF \circ T_{F^{-1}(z)}(U_{A_0} \times \mathbb{S}^{N-1})\right) + T_z(M^k) = \mathbb{R}^d$ for $z \neq 0$ as claimed. This shows the desired transversality for almost all $A \in U_{A_0}$.

To extend the transversality to all of $\mathcal{A}^\circ_{M^k}$, we use the fact that $\mathcal{A}^\circ_{M^k}$ is σ-compact, i.e. is the union of countably many compact sets. In fact, any open subset of $\mathbb{R}^d$ is σ-compact. As a consequence, $\mathcal{A}^\circ_{M^k}$ is contained the union of countably many of open neighborhoods U_{A_0} as constructed above. Since the set of matrices A inside each U_{A_0} violating the transversality statement has measure 0, we conclude by countable additivity that the set of $A \in \mathcal{A}^\circ_{M^k}$ violating transversality has measure 0 as well. This concludes the proof. □

Lemma B.21. *Fix linearly independent vectors $b_1, \ldots, b_d \in \mathbb{R}^d$ and let $A_1, \ldots, A_d$ be positive-definite matrices sampled independently from probability distributions on $\mathbb{R}^{\binom{d+1}{2}}$ which are absolutely continuous with respect to Lebesgue measure (i.e. which have a density). Then*

$$\left| \bigcap_{i=1}^{d} E_{A_i, b_i} \right| \geq 2$$

holds almost surely.

Proof. We proceed iteratively. For $k = d-1, \ldots, 1$ set

$$M^k = E_{A_1, b_1} \cap \ldots \cap E_{A_{d-k}, b_{d-k}}.$$

We show by induction that M^k is almost surely a smooth compact manifold of dimension k. The base case $k = d-1$ is obvious, and for smaller k, we have

$$M^k = M^{k+1} \cap E_{A,b}.$$

Lemma B.20 combined with Lemma B.13 now implies that M^k is a smooth compact manifold of dimension k almost surely, completing the inductive step.

Finally Proposition B.15 implies that assuming M^1 and E_{A_d, b_d} are transverse (which holds with probability 1), the number of intersection points $|M^1 \cap E_{A_d, b_d}|$ is finite and even. Of course $|M^1 \cap E_{A_d, b_d}| = \left| \cap_{i=1}^{d} E_{A_i, b_i} \right|$. Since $\cap_{i=1}^{d} E_{A_i, b_i}$ trivially contains the origin, it must also contain another point. This completes the proof. □

Proof of Theorem B.11. Given $A_1, \ldots, A_d$, consider a sequence of d-tuples $\left(A_1^{(k)}, \ldots, A_d^{(k)}\right)_{k \geq 1}$ converging to $(A_1, \ldots, A_d)$, i.e. satisfying

$$\lim_{k \to \infty} A_i^{(k)} = A_i$$

for each $i \in [d]$. Moreover assume that $\left| \bigcap_{i \in [d]} E_{A_i^{(k)}, b_i} \right| \geq 2$ for each k; such a sequence certainly exists by Lemma B.21. We also assume that the estimates

$$\ell \leq \lambda_d(A_i^{(k)}) \leq \lambda_1(A_i^{(k)}) \leq L \tag{B.10}$$

hold for some positive constants ℓ, L where λ_d, λ_1 are the minimum and maximum eigenvalues. This last assumption is without loss of generality by restricting the values of k to $k \geq k_0$ for suitably large k_0. For each k, choose a non-zero point

$$x_k \in \bigcap_{i \in [d]} E_{A_i^{(k)}, b_i} \backslash \{0\}.$$

Such points exist because $|\bigcap_{i \in [d]} E_{A_i^{(k)}, b_i}| \geq 2$. We claim the norms $|x_k|$ are bounded away from infinity, bounded away from zero, and that any sub-sequential limit x_* satisfies

$$x_* \in \bigcap_{i \in [d]} E_{A_i, b_i}.$$

It follows from the above claims that at least one sub-sequential limit x_* exists (using the Bolzano-Weierstrass theorem) and that $|x_*| \neq 0$. Therefore the above claims suffice to finish the proof, and we now turn to their individual proofs.

First, since $x_k^\top A_i^{(k)} x_k \geq \lambda_d(A_i^{(k)})|x_k|^2 \geq \ell|x_k|^2$ and $|b_i^\top x_k| \leq |b_i^\top| \cdot |x_k|$, it follows that $|x_k| \leq \frac{|b_1|}{\ell}$ for all k, so in particular these norms are bounded above. Next we show the values $|x_k|$ are bounded away from 0. Suppose for sake of contradiction that $|x_{a_j}| \to 0$ along some subsequence $(a_j)_{j\geq 1}$. Then

$$\langle b_i, x_{a_j} \rangle = x_{a_j}^\top A_i^{(a_j)} x_{a_j} \leq L|x_{a_j}|^2 = o(|x_{a_j}|).$$

Defining the rescaled unit vectors $\widehat{x}_{a_j} = \frac{x_{a_j}}{|x_{a_j}|}$, it follows that

$$\lim_{j\to\infty} \langle b_i, \widehat{x}_{a_j} \rangle = 0$$

for each i. As the $\widehat{x}_{a_j}$ are unit vectors, the Bolzano-Weierstrass theorem guarantees existence of a subsequential limit $\widehat{x}_*$ which is also a unit vector. It follows $\langle b_i, \widehat{x}_* \rangle = 0$ for all $i \in [d]$. However because the vectors b_i are linearly independent, this implies $|\widehat{x}_*| = 0$ which is a contradiction. We conclude that $|x_k|$ is bounded away from 0.

Finally we show that any subsequential limit satisfies $x_* \in E_{A,b}$. With b fixed, observe that the functions $g_{A,b}(x) = x^\top Ax - b^\top x$ are uniformly Lipschitz for A obeying the eigenvalue bound (B.10) and $|x| \leq \frac{|b_1|}{\ell}$. It follows that

$$\lim_{k\to\infty} g_{A_i^{(k)},b_i}(x_*) = \lim_{k\to\infty} g_{A_i^{(k)},b_i}(x_k) = 0.$$

Having established the three claims we conclude the proof of Theorem B.11. □

C Additional experimental details

Gaussian dataset is a binary classification task that closely reflects our assumptions in section 3. We take $r = 3$, $d_s = 32$, $\mu_1 = \mathbf{1}_r$, $\Sigma_1 = I_r$, $\mu_2^e \sim \mathcal{N}(0, 10I_{d_s})$, and $\Sigma_2^e = G_e G_e^\top$. We use $1k$ samples per environment and vary the number of training / test environments from $E = 3$ to $E = 15$.

Noised MNIST is a 10-way semi-synthetic classification task modified from LeCun and Cortes [23] to test generalization of our theory to multi-class classification and different neural network architectures. The construction is inspired by the situation where certain background features spuriously correlate with labels ("most cows appear in grass and most camels appear in sand") [5, 3, 4], but the covariance of the background features changes across environments. Concretely, we divide the 60k images into $E = 12$ groups. Each group is further divided into a training and a test environment with ratio 9:1. We add an additional row of noise (28 pixels) to the original grayscale digits of dimension 28×28. In training environments, the added noise is the spurious feature that, conditioned on the label, has identical mean but changing covariances across environments. In test environments, the noise is uncorrelated with the label.

For Noised MNIST dataset, for each class $c \in \{0, \ldots, 9\}$, we first generative a class signature $x_c \in \mathbb{R}^{28} \sim N(0, 2.5I_{28})$. For each of the $E = 12$ groups, we generate a training spurious covariance $\Sigma_2^e = G_e G_e^\top$ and a test spurious covariance $\Sigma_2^{e\prime} = G_e' {G_e'}^\top$. The noise code for digit c in training environment e is drawn from $\mathcal{N}(x_c, \Sigma_2^e)$. In test environment, the noise is drawn from $\mathcal{N}(x_{c'}, \Sigma_2^{e\prime})$ for random label $c' \sim \text{unif}\{0, \ldots, 9\}$).

We use SGD optimizer for both datasets. The hyperparameters are the coefficients for coral penalty, orthonormal penalty, and irm penalty $\lambda_{coral}, \lambda_{on}, \lambda_{irm}$, and learning rate lr. For each algorithm in Figures 1 and 2, we select penalization strengths from $\{0.1, 1, 10, 100\}$ and lr from $\{0.1, 0.01, 0.001, 0.0001\}$ that achieves highest average test accuracy within 500 epochs (for Gaussian dataset) and 400 epochs (for Noised MNIST). Gaussian dataset has batch size 100 and Noised MNIST has batch size 1000 from each training environment.

The average test accuracies for each algorithm with error bars are shown in Figures 1 and 2. We fix the datasets and use different random seeds for algorithmic randomness. Error bar indicates mean and standard deviation across 5 runs.

Table 1: MLP network architectures for Noised MNIST

Number of layers	1	2	3	4	6
Layer widths	24	96, 24	128, 50, 24	192, 96, 48, 24	400, 300, 200, 100, 50, 24

Table 2: Matching features at 3 layers with identical widths does not have significant advantage over matching only at the last layer (CORAL).

Layer widths	24	128, 50, 24	24, 24, 24
ERM	58.6 ± 0.4	56.0 ± 0.6	62.1 ± 0.6
IRM	59.0 ± 0.2	56.1 ± 0.6	62.3 ± 1.0
CORAL (only match last layer)	69.1 ± 1.0	65.2 ± 1.0	67.2 ± 0.4
CORAL (match-disjoint)	69.1 ± 1.0	75.5 ± 1.0	70.6 ± 0.9
CORAL (match-all)	69.1 ± 1.0	77.9 ± 0.4	70.4 ± 0.9

The MLP architecture in Figure 2 is in Table 1:

To answer (Q5), we compare performances of algorithms on a 3-layer MLP that does not shrink feature dimensions (right column) with those on a 3-layer MLP that does (middle column) and a 1-layer MLP (left column) in Table 2. Results show that without shrinking feature dimensions, matching at multiple layers does not improve over naive CORAL on a smaller architecture.

No run in any of our experiments take more than 10 minutes on a single GPU. MNIST dataset [23] is made available under the terms of the Creative Commons Attribution-Share Alike 3.0 license.

D A simple algorithm achieves O(1) environment complexity under Assumption 3.2

Intuitively, subtracting the label-conditional covariances of any two environments yields the subspace of spurious coordinates (the column subspace of $B \in \mathbb{R}^{d \times d_s}$, the right d_s columns of S). Once we obtain the projection matrix onto this subspace denoted as P_B, we can transform all observations (X_i^e, Y_i^e) to $(X_i^{e\prime}, Y_i^e)$ where $X_i^{e\prime} = (I - P_B)X_i^e$ is the projection of X_i^e onto the orthogonal subspace of B. The transformed inputs have no signal in any spurious dimension, so the optimal classifier on $(X_i^{e\prime}, Y_i^e)_{i=1}^{\infty}$ from any environment e is the invariant predictor w^*.

This is formalized as Algorithm 2, and the following theorem provides formal guarantees for the environment complexity of this algorithm:

Theorem D.1. *Under Assumption 3.2, Algorithm 2 satisfies $\widehat{w} = w^*$.*

Proof. Define A, B as the left r and right d_s columns of S. The optimal output is characterized by

$$\begin{bmatrix} A^\top \\ B^\top \end{bmatrix} w^* = \begin{bmatrix} \Sigma_1^{-1}\mu_1 \\ 0 \end{bmatrix} \iff A^\top w^* = \Sigma_1^{-1}\mu_1, B^\top w^* = 0 \iff \Sigma_1 A^\top w^* = \mu_1, P_B w^* = 0.$$

The algorithmic output $\widehat{w}$ satisfies

$$\begin{aligned} (I - P_B)A\Sigma_1 A^\top (I - P_B)\widehat{w} &= (I - P_B)A\mu_1, P_B\widehat{w} = 0 \\ \implies (I - P_B)A\Sigma_1 A^\top \widehat{w} &= (I - P_B)A\mu_1, P_B\widehat{w} = 0 \end{aligned}$$

Multiplying the first equation on the RHS by its pseudo-inverse, we get:

$$\begin{aligned} (A^\top(I - P_B)A)^{-1}A^\top(I - P_B)A\Sigma_1 A^\top \widehat{w} &= (A^\top(I - P_B)A)^{-1}A^\top(I - P_B)A\mu_1 \\ \implies \Sigma_1 A^\top \widehat{w} &= \mu_1. \end{aligned}$$

Therefore $\widehat{w} = w^*$. □

E Limitations and potential negative impact

We study a restricted data model and linear hypothesis class, so our results may not apply to realistic datasets with finite samples, or other hypotheses families. Our work aims to improve robustness

Algorithm 2 A simple algorithm under Assumption 3.2

Require: Invariant feature dimension r, spurious feature dimensions d_s, 2 training environments with infinite samples $\{(X_i^e, Y_i^e)\}_{i=1}^{\infty} \sim P_e$, $\{(X_i^{e'}, Y_i^{e'})\}_{i=1}^{\infty} \sim P_{e'}$.

1: Subtract the covariances of class 1 examples between the two environments
$B \leftarrow Cov_e[X|Y=1] - Cov_{e'}[X|Y=1]$.

2: Perform SVD on $B = Q\Gamma Q^\top$ to get orthonormal $Q \in \mathbb{R}^{d \times d_s}$ and diagonal $\Gamma \in \mathbb{R}^{d_s \times d_s}$.

3: Project the mean of class 1 examples $\mathbb{E}[X|Y=1]$ to the orthogonal subspace of B,
$\mu' = (I - QQ^\top)\mathbb{E}[X|Y=1]$.

4: Project the covariance of class 1 examples $\Sigma' = (I - QQ^\top)Cov_e[X|Y=1](I - QQ^\top)$.

5: Return classifier $\widehat{w} = {\Sigma'}^\dagger \mu'$.

of machine learning models; however, we only provide a sufficient set of conditions for certain algorithm to be robust to distributional shifts; the set of conditions may not be necessary, and may not be satisfied in the real world. Misapplying the proposed algorithm to realistic datasets may lead to negative impacts.