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

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

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 $e \in \mathcal{E}$, $US\Sigma^e S^\top U = C_t \in \mathbb{R}^{r_t \times r_t}$. 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 $\epsilon > 0$, $|\tilde{Q}| \leq (c_3\sqrt{k}/\epsilon)^{k(d_s - k)}$, and $\forall Q^* \in \mathcal{Q}$, $\exists Q \in \tilde{Q}$ 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,

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

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 $b_1, \ldots, b_d \in \mathbb{R}^d$ are linearly independent and $A_1, \ldots, A_d$ 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 $A_1, \ldots, A_d$, as long as $b_1, \ldots, b_d$ are linearly independent. Then we use a continuity argument to extend the result to all positive definite matrices $A_1, \ldots, A_d$.

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*

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

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

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

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 $b_1, \dots, b_d \in \mathbb{R}^d$ and let $A_1, \dots, A_d$ 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 $k = d-1, \dots, 1$ 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 $A_1, \dots, A_d$, consider a sequence of d-tuples $\left(A_1^{(k)}, \dots, A_d^{(k)}\right)_{k \geq 1}$ converging to $(A_1, \dots, A_d)$, 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 $g_{A,b}(x) = x^\top A x - b^\top x$ 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 $r = 3$, $d_s = 32$, $\mu_1 = \mathbf{1}_r$, $\Sigma_1 = I_r$, $\mu_2^e \sim \mathcal{N}(0, 10 I_{d_s})$, 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 $x_c \in \mathbb{R}^{28} \sim N(0, 2.5 I_{28})$. 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.