A Additional Notation for the Appendix

To avoid overloading, we use ϕ, F for the standard Gaussian PDF and CDF respectively. We write $\mathcal{S}^c$ to denote the set complement of a set $\mathcal{S}$. We write $|\cdot|$ to denote entrywise absolute value.

B Proof of Proposition 4.1

Recall the IRM objective:

$$\begin{aligned} \min_{\Phi,\beta} \quad & \mathbb{E}_{(x,y)\sim p(x,y)}[-\log \sigma(y \cdot \hat{\beta}^T \Phi(x))] \\ \text{subject to} \quad & \frac{\partial}{\partial \hat{\beta}} \mathbb{E}_{(x,y)\sim p^e}[-\log \sigma(y \cdot \hat{\beta}^T \Phi(x))] = 0. \; \forall e \in \mathcal{E}. \end{aligned}$$

Concretely, we represent Φ as some parametrized function Φ_θ, over whose parameters θ we then optimize. The derivative of the negative log-likelihood for logistic regression with respect to the β coefficients is well known:

$$\frac{\partial}{\partial \hat{\beta}} \Big[-\log \sigma(y \cdot \hat{\beta}^T \Phi_\theta(x))\Big] = (\sigma(\hat{\beta}^T \Phi_\theta(x)) - \mathbf{1}\{y = 1\})\Phi_\theta(x).$$

Suppose we recover the true invariant features $\Phi_\theta(x) = \begin{bmatrix} z_c \\ \mathbf{0} \end{bmatrix}$ and coefficients $\hat{\beta} = \begin{bmatrix} \beta \\ \mathbf{0} \end{bmatrix}$ (in other words, we allow for the introduction of new features). Then the IRM constraint becomes:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \hat{\beta}} \mathbb{E}_{(x,y)\sim p^e}[-\log \sigma(y \cdot \hat{\beta}^T \Phi_\theta(x))] \\ &= \int_{\mathcal{Z}} p^e(z_c) \sum_{y \in \{\pm 1\}} p^e(y \mid z_c) \frac{\partial}{\partial \hat{\beta}} \left[-\log \sigma(y \cdot \beta^T z_c)\right] \, dz_c \\ &= \int_{\mathcal{Z}} p^e(z_c) \Phi_\theta(x) \Big[\sigma(\hat{\beta}^T z_c)(\sigma(\hat{\beta}^T z_c) - 1) + (1 - \sigma(\hat{\beta}^T z_c))\sigma(\hat{\beta}^T z_c)\Big] \, dz_c. \end{aligned}$$

Since $\hat{\beta}$ is constant across environments, this constraint is clearly satisfied for every environment, and is therefore also the minimizing $\hat{\beta}$ for the training data as a whole.

Considering now the derivative with respect to the featurizer Φ_θ:

$$\frac{\partial}{\partial \theta} \Big[-\log \sigma(y \cdot \hat{\beta}^T \Phi_\theta(x))\Big] = (\sigma(\hat{\beta}^T \Phi_\theta(x)) - \mathbf{1}\{y = 1\}) \frac{\partial}{\partial \theta} \hat{\beta}^T \Phi_\theta(x).$$

Then the derivative of the loss with respect to these parameters is

$$\int_{\mathcal{Z}} p^e(z_c) \left(\frac{\partial}{\partial \theta} \hat{\beta}^T \Phi_\theta(x)\right) \Big[\sigma(\beta^T z_c)(\sigma(\beta^T z_c) - 1) + (1 - \sigma(\beta^T z_c))\sigma(\beta^T z_c)\Big] \, dz_c = 0.$$

So, the optimal invariant predictor is a stationary point with respect to the feature map parameters as well.

C Results from Section 5

C.1 Proof of Theorem 5.1

We begin by formally stating the non-degeneracy condition. Consider any environmental mean μ_e, and suppose it can be written as a linear combination of the others means with coefficients α^e:

$$\mu_e = \sum_i \alpha_i^e \mu_i.$$

Then the environments are considered non-degenerate if the following inequality holds for any such set of coefficients:

$$\sum_i \alpha_i^e \neq 1, \tag{9}$$

and furthermore that the following ratio is different for at least two different environments a, b:

$$\exists \alpha^a, \alpha^b.\ \frac{\sigma_a^2 - \sum_i \alpha_i^a \sigma_i^2}{1 - \sum_i \alpha_i^a} \neq \frac{\sigma_b^2 - \sum_i \alpha_i^b \sigma_i^2}{1 - \sum_i \alpha_i^b}. \tag{10}$$

The first inequality says that none of the environmental means are are an affine combination of the others; in other words, they lie in *general linear position*, which is the same requirement as [1]. The other inequality is a similarly lax non-degeneracy requirement regarding the relative scale of the variances. It is clear that the set of environmental parameters that do not satisfy Equations 9 and 10 has measure zero under any absolutely continuous density, and similarly, if $E \leq d_e$ then the environmental means will be linearly independent almost surely.

We can now proceed with the proof, beginning with some helper lemmas:

Lemma C.1. *Suppose we observe E environments $\mathcal{E} = \{e_1, e_2, \ldots, e_E\}$, each with environmental mean of dimension $d_e \geq E$, such that all environmental means are linearly independent. Then there is a unique unit-norm vector p such that*

$$p^T \mu_e = \sigma_e^2 \tilde{\mu} \ \ \forall e \in \mathcal{E}, \tag{11}$$

where $\tilde{\mu}$ is the largest scalar which admits such a solution.

Proof. Let $v_1, v_2, \ldots, v_E$ be a set of basis vectors for span$\{\mu_1, \mu_2, \ldots, \mu_E\}$. Each mean can then be expressed as a combination of these basis vectors: $u_i = \sum_{j=1}^E \alpha_{ij} v_j$. Since the means are linearly independent, we can combine these coefficients into a single invertible matrix

$$U = \begin{bmatrix} \alpha_{11} & \alpha_{21} & \ldots & \alpha_{E1} \\ \alpha_{12} & \alpha_{22} & \ldots & \alpha_{E2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1E} & \alpha_{2E} & \ldots & \alpha_{EE} \end{bmatrix}.$$

We can then combine the constraints (11) as

$$U^T p_\alpha = \boldsymbol{\sigma} \triangleq \begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \vdots \\ \sigma_E^2 \end{bmatrix},$$

where p_α denotes our solution expressed in terms of the basis vectors $\{v_i\}_{i=1}^E$. This then has the solution

$$p_\alpha = U^{-T} \boldsymbol{\sigma}.$$

This defines the entire space of solutions, which consists of p_α plus any element of the remaining $(d_e - E)$-dimensional orthogonal subspace. However, we want p to be unit-norm—observe that the current vector solves Equation 11 with $\tilde{\mu} = 1$, which means that after normalizing we get $\tilde{\mu} = \frac{1}{\|p_\alpha\|_2}$. Adding any element of the orthogonal subspace would only increase the norm of p, decreasing $\tilde{\mu}$. Thus, the unique maximizing solution is

$$p_\alpha = \frac{U^{-T}\boldsymbol{\sigma}}{\|U^{-T}\boldsymbol{\sigma}\|_2}, \quad \text{with} \quad \tilde{\mu} = \frac{1}{\|U^{-T}\boldsymbol{\sigma}\|_2}.$$

Finally, p_α has to be rotated back into the original space by defining $p = \sum_{i=1}^E p_{\alpha i} v_i$. □

Lemma C.2. *Assume f is linear. Suppose we observe $E \leq d_e$ environments whose means are linearly independent. Then there exists a linear Φ with* $\text{rank}(\Phi) = d_c + d_e + 1 - E$ *whose output depends on the environmental features, yet the optimal classifier on top of Φ is invariant.*

Proof. We will begin with the case when $E = d_e$ and then show how to modify this construction for when $E < d_e$. Consider defining

$$\Phi = \begin{bmatrix} I & 0 \\ 0 & M \end{bmatrix} \circ f^{-1}$$

with

$$M = \begin{bmatrix} — & p^T & — \\ — & 0 & — \\ & \vdots & \\ — & 0 & — \end{bmatrix}.$$

Here, $p \in \mathbb{R}^{d_c}$ is defined as the unit-norm vector solution to

$$p^T \mu_e = \sigma_e^2 \tilde{\mu} \quad \forall e$$

such that $\tilde{\mu}$ is maximized—such a vector is guaranteed to exist by Lemma C.1. Thus we get $\Phi(x) = \begin{bmatrix} z_c \\ p^T z_e \end{bmatrix}$, which is of rank $d_c + 1$ as desired. Define $\tilde{z}_e = p^T z_e$, which means that $\tilde{z}_e \mid y \sim \mathcal{N}(y \cdot \sigma_e^2 \tilde{\mu}, \sigma_e^2)$. For each environment we have

$$\begin{aligned} p(y \mid z_c, \tilde{z}_e) &= \frac{p(z_c, \tilde{z}_e, y)}{p(z_c, \tilde{z}_e)} \\ &= \frac{\sigma(y \cdot \beta_c^T z_c) p(\tilde{z}_e \mid y \cdot \sigma_e^{e2} \tilde{\mu}, \sigma_e^2)}{[\sigma(y \cdot \beta_c^T z_c) p(\tilde{z}_e \mid y \cdot \sigma_e^{e2} \tilde{\mu}, \sigma_e^2) + \sigma(-y \cdot \beta_c^T z_c) p(\tilde{z}_e \mid -y \cdot \sigma_e^{e2} \tilde{\mu}, \sigma_e^2)]} \\ &= \frac{\sigma(y \cdot \beta_c^T z_c) \exp(y \cdot \tilde{z}_e \tilde{\mu})}{[\sigma(y \cdot \beta_c^T z_c) \exp(y \cdot \tilde{z}_e \tilde{\mu}) + \sigma(-y \cdot \beta_c^T z_c) \exp(-y \cdot \tilde{z}_e \tilde{\mu})]} \\ &= \frac{1}{1 + \exp(-y \cdot (\beta_c^T z_c + 2 \tilde{z}_e \tilde{\mu}))}. \end{aligned}$$

The log-odds of y is linear in these features, so the optimal classifier is

$$\hat{\beta} = \begin{bmatrix} \beta_c \\ 2\tilde{\mu} \end{bmatrix},$$

which is the same for all environments.

Now we show how to modify this construction for when $E < d_e$. If we remove one of the environmental means, Φ trivially maintains its feasibility. Note that since they are linearly independent, the mean which was removed has a component in a direction orthogonal to the remaining means. Call this component p', and consider redefining M as

$$M = \begin{bmatrix} — & p^T & — \\ — & p'^T & — \\ — & 0 & — \\ & \vdots & \\ — & 0 & — \end{bmatrix}.$$

The distribution of $\tilde{z}_e$ in each of the remaining dimensions is normal with mean 0, which means a corresponding coefficient of 0 is optimal for all environments. So the classifier $\hat{\beta}$ remains optimal for all environments, yet we've added another row to M which increases the dimensionality of its span, and therefore the rank of Φ, by 1. Working backwards, we can repeat this process for each additional mean, such that $\text{rank}(\Phi) = d_c + 1 + (d_e - E)$, as desired. Note that for $E = 1$ any Φ will be vacuously feasible. □

Lemma C.3. *Suppose we observe E environments $\mathcal{E} = \{e_1, e_2, \ldots, e_E\}$ whose parameters satisfy the non-degeneracy conditions (9, 10). Let $\Phi(x) = Az_c + Bz_e$ be any feature vector which is a linear function of the invariant and environmental features, and suppose the optimal $\hat{\beta}$ on top of Φ is invariant. If $E > d_e$, then $\hat{\beta} = 0$ or $B = 0$.*

Proof. Write $\Phi = [A|B]$ where $A \in \mathbb{R}^{d \times d_c}$, $B \in \mathbb{R}^{d \times d_e}$ and define

$$\begin{aligned} \bar{\mu}_e &= \Phi \begin{bmatrix} \mu_c \\ \mu_e \end{bmatrix} && = A\mu_c + B\mu_e, \\ \bar{\Sigma}_e &= \Phi \begin{bmatrix} \sigma_c^2 I_{d_c} & 0 \\ 0 & \sigma_e^2 I_{d_e} \end{bmatrix} \Phi^T && = \sigma_c^2 AA^T + \sigma_e^2 BB^T. \end{aligned}$$

Without loss of generality we assume $\bar{\Sigma}$ is invertible (if it is not, we can consider just the subspace in which it is—outside of this space, the features have no variance and therefore cannot carry information about the label). By Lemma F.2, the optimal coefficient for each environment is $2\bar{\Sigma}_e^{-1}\bar{\mu}_e$. In order for this vector to be invariant, it must be the same across environments; we write it as a constant $\hat{\beta}$. Suppose $\bar{\mu}_e = 0$ for some environment e—then the claim is trivially true with $\hat{\beta} = 0$. We therefore proceed under the assumption that $\bar{\mu}_e \neq 0 \; \forall e \in \mathcal{E}$.

With this fact, we have that $\forall e \in \mathcal{E}$,

$$\begin{aligned} & \hat{\beta} = 2(\sigma_c^2 AA^T + \sigma_e^2 BB^T)^{-1}(A\mu_c + B\mu_e) \\ \iff & (\sigma_c^2 AA^T + \sigma_e^2 BB^T)\hat{\beta} = 2A\mu_c + 2B\mu_e \\ \iff & \sigma_e^2 BB^T\hat{\beta} - 2B\mu_e = 2A\mu_c - \sigma_c^2 AA^T\hat{\beta}. \end{aligned} \tag{12}$$

Define the vector $\mathbf{v} = 2A\mu_c - \sigma_c^2 AA^T\hat{\beta}$. We will show that for any $\hat{\beta}, A$, with probability 1 only $B = 0$ can satisfy Equation 12 for every environment. If $E > d_e$, then there exists at least one environmental mean which can be written as a linear combination of the others. Without loss of generality, denote the parameters of this environment as $(\bar{\mu}, \bar{\sigma}^2)$ and write $\bar{\mu} = \sum_{i=1}^{d_e} \alpha_i \mu_i$. However, note that by assumption the means lie in general linear position, and therefore we actually have at least d_e sets of coefficients α for which this holds. Rearranging Equation 12, we get

$$\begin{aligned} \bar{\sigma}^2 BB^T\hat{\beta} - \mathbf{v} &= 2B\bar{\mu} \\ &= \sum_{i=1}^{d_e} \alpha_i 2B\mu_i \\ &= \sum_{i=1}^{d_e} \alpha_i \left[\sigma_i^2 BB^T\hat{\beta} - \mathbf{v}\right], \end{aligned}$$

and rearranging once more yields

$$\left(\bar{\sigma}^2 - \sum \alpha_i \sigma_i^2\right) BB^T\hat{\beta} = \left(1 - \sum \alpha_i\right)\mathbf{v}.$$

By assumption, $(1 - \sum \alpha_i)$ is non-zero. We can therefore rewrite this as

$$\hat{\alpha} BB^T\hat{\beta} = \mathbf{v},$$

where $\hat{\alpha} = \frac{\bar{\sigma}^2 - \sum \alpha_i \sigma_i^2}{1 - \sum \alpha_i}$ is a scalar. As the vectors $BB^T\hat{\beta}$ and $\mathbf{v}$ are both independent of the environment, this can only hold true if $\hat{\alpha}$ is fixed for all environments or if both $BB^T\hat{\beta}, \mathbf{v}$ are 0. The former is false by assumption, so the the latter must hold.

As a result, we see that Equation 12 reduces to

$$B\mu_e = 0 \quad \forall e \in \mathcal{E}.$$

As the span of the observed μ_e is all of $\mathbb{R}^{d_e}$, this is only possible if $B = 0$. □

We are now ready to prove the main claim. We restate the theorem here for convenience:

Theorem 5.1 (Linear case)**.** *Assume f is linear. Suppose we observe E training environments. Then the following hold:*

1. *Suppose $E > d_e$. Consider any linear featurizer Φ which is feasible under the IRM objective (4), with invariant optimal classifier $\hat{\beta} \neq 0$, and write $\Phi(f(z_c, z_e)) = Az_c + Bz_e$. Then under mild non-degeneracy conditions, it holds that $B = 0$. Consequently, $\hat{\beta}$ is the optimal classifier for all possible environments.*

2. *If $E \leq d_e$ and the environmental means μ_e are linearly independent, then there exists a linear Φ—where $\Phi(f(z_c, z_e)) = Az_c + Bz_e$ with rank$(B) = d_e + 1 - E$—which is feasible under the IRM objective. Further, both the logistic and 0-1 risks of this Φ and its corresponding optimal $\hat{\beta}$ are strictly lower than those of the optimal invariant predictor.*

Proof. 1. Since Φ, f are linear, we can write $\Phi(x) = Az_c + Bz_e$ for some matrices A, B. Assume the non-degeneracy conditions (9, 10) hold. By Lemma C.3, one of $B = 0$ or $\hat{\beta} = 0$ holds. Thus, $\Phi, \hat{\beta}$ uses only invariant features. Since the joint distribution $p^e(z_c, y)$ is invariant, this predictor has identical risk across all environments.

2. The existence of such a predictor is proven by Lemma C.2. It remains to show that the risk of this discriminator is lower than that of the optimal invariant predictor. Observe that these features are non-degenerate independent random variables with support over all of $\mathbb{R}$, and therefore by Lemma F.1, dropping the $\tilde{z}_e$ term and using

$$\Phi(x) = [z_c], \quad \hat{\beta} = \begin{bmatrix} \beta_c \\ \beta_0 \end{bmatrix}$$

results in strictly higher risk. The proof is completed by noting that this definition is precisely the optimal invariant predictor.

□

C.2 Experiments for Theorem 5.1

To corroborate our theoretical findings, we run an experiment on data drawn from our model to see at what point IRM is able to recover a generalizing predictor. We generated data precisely according to our model in the linear setting, with $d_c = 3, d_e = 6$. The environmental means were drawn from a multivariate Gaussian prior; we randomly generated the invariant parameters and the parameters of the prior such that using the invariant features gave reasonable accuracy (71.9%) but the environmental features would allow for almost perfect accuracy on in-distribution test data (99.8%). Thus, the goal was to see if IRM could successfully learn a predictor which ignores meaningful covariates z_e, to the detriment of its training performance but to the benefit of OOD generalization. We chose equal class marginals ($\eta = 0.5$).

Figure C.1 shows the result of five runs of IRM, each with different environmental parameters but the same invariant parameters (the training data itself was redrawn for each run). We found that optimizing for the IRM objective was quite unstable, frequently collapsing to the ERM solution unless λ and the optimizer learning rate were carefully tuned. This echoes the results of [20] who found that tuning λ during training to specific values at precisely the right time is essential for good performance. To prevent collapse, we kept the same environmental prior and found a single setting for λ and the learning rate which resulted in reasonable performance across all five runs. At test time, we evaluated the trained predictors on additional, unseen environments whose parameters were drawn from the same prior. To simulate distribution shift, we evaluated the predictors on the same data but with the environmental means negated. Thus the correlations between the environmental features z_e and the label y were reversed.

Observe that the results closely track the expected outcome according to Theorem 5.1: up until $E = d_e$, IRM essentially matches ERM in performance both in-distribution and under distribution shift. As soon as we cross that threshold of observed environments, the predictor learned via IRM begins to perform drastically better under distribution shift, behaving more like the optimal invariant predictor. We did however observe that occasionally the invariant solution would be found after only $E = d_e = 6$ environments; we conjecture that this is because at this point the feasible-yet-not-invariant predictor with lower objective value presented in Theorem 5.1 is precisely a single point, as opposed to a multi-dimensional subspace, and therefore might be difficult for the optimizer to find.

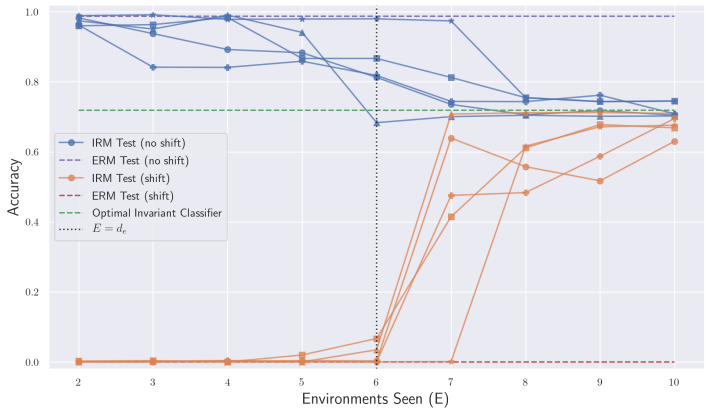

Figure C.1: Performance of predictors learned with IRM (5 different runs) and ERM (dashed lines) on test distributions where the correlation between environmental features and the label is consistent (no shift) or reversed (shift). The dashed green line is the performance of the optimal invariant predictor. Observe that up until $E = d_e$, IRM consistently returns a predictor with performance similar to ERM: good generalization without distribution shift, but catastrophic failure when the correlation is reversed. In contrast, once $E > d_e$, IRM is able to recover a $\Phi, \hat{\beta}$ with performance similar to that of the invariant optimal predictor.

C.3 Proof of Theorem 5.3

Theorem 5.3. *Suppose we observe $E \leq d_e$ environments, such that all environmental means are linearly independent. Then there exists a feasible $\Phi, \hat{\beta}$ which uses* only environmental features *and achieves lower 0-1 risk than the optimal invariant predictor on every environment e such that $\sigma_e\tilde{\mu} > \sigma_c^{-1}\|\mu_c\|_2$ and $2\sigma_e\tilde{\mu}\sigma_c^{-1}\|\mu_c\|_2 \geq |\beta_0|$.*

Proof. We consider the non-invariant predictor constructed as described in Lemma C.2, but dropping the invariant features and coefficients. By Lemma F.2, the optimal coefficients for the invariant and non-invariant predictors are

$$\hat{\beta}_{caus} = \begin{bmatrix} 2\sigma_c^{-2}\mu_c \\ \beta_0 \end{bmatrix} \qquad \text{and} \qquad \hat{\beta}_{non-caus} = \begin{bmatrix} 2\tilde{\mu} \\ \beta_0 \end{bmatrix},$$

respectively. Therefore, the 0-1 risk of the optimal invariant predictor is precisely

$$\begin{aligned} &\eta\mathbb{P}(2\sigma_c^{-2}\mu_c^T z_c + \beta_0 < 0) + (1-\eta)\mathbb{P}(-2\sigma_c^{-2}\mu_c^T z_c + \beta_0 > 0) \\ =&\eta F\left(-\sigma_c^{-1}\|\mu_c\|_2 - \frac{\beta_0\sigma_c}{2\|\mu_c\|_2}\right) + (1-\eta)F\left(-\sigma_c^{-1}\|\mu_c\|_2 + \frac{\beta_0\sigma_c}{2\|\mu_c\|_2}\right), \end{aligned}$$

where F is the Gaussian CDF. By the same reasoning, the 0-1 risk of the non-invariant predictor is

$$\eta F\left(-\sigma_e\tilde{\mu} - \frac{\beta_0}{2\sigma_e\tilde{\mu}}\right) + (1-\eta)F\left(-\sigma_e\tilde{\mu} + \frac{\beta_0}{2\sigma_e\tilde{\mu}}\right).$$

Define $\alpha = \sigma_c^{-1}\|\mu_c\|_2$ and $\gamma = \sigma_e\tilde{\mu}$. By monotonicity of the Gaussian CDF, the former risk is higher than the latter if

$$\alpha + \frac{\beta_0}{2\alpha} \leq \gamma + \frac{\beta_0}{2\gamma}, \tag{13}$$

$$\alpha - \frac{\beta_0}{2\alpha} < \gamma - \frac{\beta_0}{2\gamma}. \tag{14}$$

Without loss of generality, we will prove these inequalities for $\beta_0 \geq 0$; an identical argument proves it for $\beta_0 < 0$ but with the '$\leq$' and '$<$' swapped.

Suppose $\gamma > \alpha$ (the first condition). Then Equation 14 is immediate. Finally, for Equation 13, observe that

$$\begin{aligned} &\gamma + \frac{\beta_0}{2\gamma} \geq \alpha + \frac{\beta_0}{2\alpha} \\ \iff &\gamma - \alpha \geq \frac{\beta_0}{2\alpha} - \frac{\beta_0}{2\gamma} = \frac{(\gamma - \alpha)\beta_0}{2\gamma\alpha} \\ \iff &2\gamma\alpha \geq \beta_0, \end{aligned}$$

which is the second condition. □

C.4 Simulations of Magnitude of Environmental Features

As discussed in Section 5, analytically quantifying the solution $\tilde{\mu}$ to the equation in Lemma C.1 is difficult; instead, we present simulations to give a sense of how often these conditions would hold in practice.

For each choice of environmental dimension d_e, we generated a "base" correlation $b \sim \mathcal{N}(0, I_{d_e})$ as the mean of the prior over environmental means μ_e. Each of these μ_e was then drawn from $\mathcal{N}(b, 4I_{d_e})$—thus, while they all came from the same prior, the noise in the draw of each μ_e was significantly larger than the bias induced by the prior. We then solved for the precise value $\sigma_e\tilde{\mu}$, with the same variance σ_e^2 for all environments, chosen as a hyperparameter. The shaded area represents a 95% confidence interval over 20 runs.

The dotted lines are $\sqrt{d_c}$. If we imagine the invariant parameters are drawn from a standard Gaussian prior, then this is precisely $\mathbb{E}[\sigma_c^{-1}\|\mu_c\|_2]$. Thus, the point where $\sigma_e\tilde{\mu}$ crosses these dotted lines is approximately how many environments would need to be observed before the non-invariant predictor has higher risk than the optimal invariant predictor. We note that this value is quite large, on the order of $d_e - d_c$.

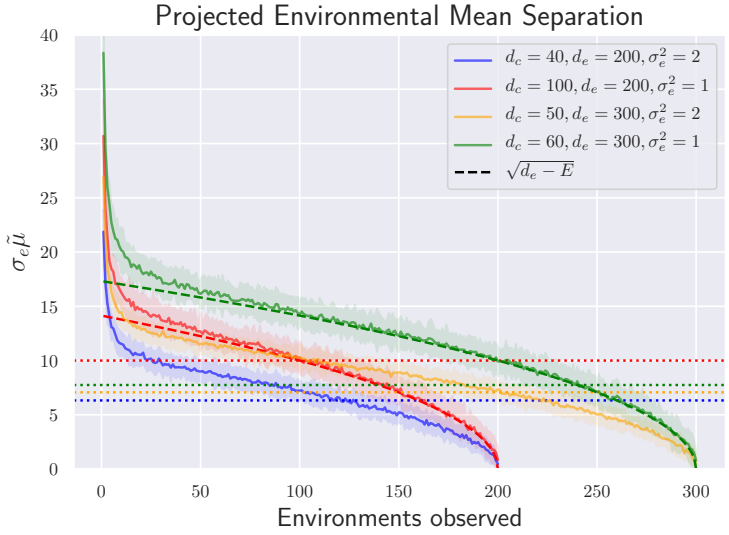

Figure C.2: Simulations to evaluate $\sigma_e\tilde{\mu}$ for varying ratios of $\frac{d_e}{d_c}$. When $\sigma_e^2 = 1$, the value closely tracks $\sqrt{d_e - E}$, giving a crossover point of $d_e - d_c$. These results imply the conditions of Theorem 5.3 are very likely to hold in the high-dimensional setting.

D Theorem 6.1 and Discussion

D.1 Proof of Theorem 6.1

We again begin with helper lemmas.

Our featurizer Φ is constructed to recover the environmental features only if they fall within a set $\mathcal{B}^c$. The following lemma shows that since only the environmental features contribute to the gradient penalty, the penalty can be bounded as a function of the measure and geometry of that set. This is used together with Lemmas F.3 and F.4 to bound the overall penalty of our constructed predictor.

Lemma D.1. *Suppose we observe environments $\mathcal{E} = \{e_1, e_2, \ldots\}$. Given a set $\mathcal{B} \subseteq \mathbb{R}^{d_e}$, consider the predictor defined by Equation 19. Then for any environment e, the penalty term of this predictor in Equation 5 is bounded as*

$$\|\nabla_{\hat{\beta}} \mathcal{R}^e(\Phi, \hat{\beta})\|_2^2 \leq \left\| \mathbb{P}(z_e \in \mathcal{B}^c) \mathbb{E}[|z_e| \mid z_e \in \mathcal{B}^c] \right\|_2^2.$$

Proof. We write out the precise form of the gradient for an environment e:

$$\nabla_{\hat{\beta}} \mathcal{R}^e(\Phi, \hat{\beta}) = \int_{\mathcal{Z}_c \times \mathcal{Z}_e} p^e(z_c, z_e) \left[\sigma(\hat{\beta}^T \Phi(f(z_c, z_e))) - p^e(y = 1 \mid z_c, z_e) \right] \Phi(f(z_c, z_e)) \, d(z_c, z_e).$$

Observe that since $z_c \perp\!\!\!\perp z_e \mid y$, the optimal invariant coefficients are unchanged, and therefore the gradient in the invariant dimensions is 0. We can split the gradient in the environmental dimensions into two integrals:

$$\begin{aligned} &\int_{\mathcal{Z}_c \times \mathcal{B}} p^e(z_c, z_e) \left[\sigma(\beta_c^T z_c + \beta_0) - p^e(y = 1 \mid z_c, z_e) \right] [0] \, d(z_c, z_e) \\ &\quad + \int_{\mathcal{Z}_c \times \mathcal{B}^c} p^e(z_c, z_e) \left[\sigma(\beta_c^T z_c + \beta_{e;\text{ERM}}^T z_e + \beta_0) - \sigma(\beta_c^T z_c + \beta_e^T z_e + \beta_0) \right] [z_e] \, d(z_c, z_e). \end{aligned}$$

Since the features are 0 within $\mathcal{B}$, the first term reduces to 0. For the second term, note that $\forall\, x, y \in \mathbb{R},\ |\sigma(x) - \sigma(y)| \leq 1$, and therefore

$$|\nabla_{\hat{\beta}} \mathcal{R}^e(\Phi, \hat{\beta})| \leq \int_{\mathcal{Z}_c \times \mathcal{B}^c} p^e(z_c, z_e)[|z_e|] \, d(z_c, z_e).$$

We can marginalize out z_c, and noting that we want to bound the squared norm,

$$\begin{aligned} \|\nabla_{\hat{\beta}} \mathcal{R}^e(\Phi, \hat{\beta})\|_2^2 &\leq \left\| \int_{\mathcal{B}^c} p^e(z_e)[|z_e|] \, dz_e \right\|_2^2 \\ &= \left\| \mathbb{P}(z_e \in \mathcal{B}^c) \mathbb{E}[|z_e| \mid z_e \in \mathcal{B}^c] \right\|_2^2. \end{aligned}$$

□

This next lemma says that if the environmental mean of the test distribution is sufficiently separated from each of the training means, with high probability a sample from this distribution will fall outside of $\mathcal{B}_r$, and therefore $\Phi_\epsilon, \hat{\beta}$ will be equivalent to the ERM solution.

Lemma D.2. *For a set of E environments $\mathcal{E} = \{e_1, e_2, \ldots, e_E\}$ and any $\epsilon > 1$, construct $\mathcal{B}_r$ as in Equation 18 and define Φ_ϵ using $\mathcal{B}_r$ as in Equation 19. Suppose we now test on a new environment with parameters $(\mu_{E+1}, \sigma_{E+1}^2)$, and assume Equation 15 holds with parameter δ. Define $k = \min_{e \in \mathcal{E}} \frac{\sigma_e^2}{\sigma_{E+1}^2}$. Then with probability $\geq 1 - \frac{2E}{\sqrt{k\pi}\delta} \exp\{-k\delta^2\}$ over the draw of an observation from this new environment, we have*

$$\Phi_\epsilon(x) = f^{-1}(x) = \begin{bmatrix} z_c \\ z_e \end{bmatrix}.$$

Proof. By Equation 15 our new environmental mean is sufficiently far away from all the label-conditional means of the training environments. In particular, for any environment $e \in \mathcal{E}$ and any label $y \in \{\pm 1\}$, the ℓ_2 distance from that mean to μ_{E+1} is at least $(\sqrt{\epsilon} + \delta)\sigma_e\sqrt{d_e}$.

Recall that $\mathcal{B}_r$ is the union of balls $\pm B_e$, where B_e is the ball of ℓ_2 radius $\sqrt{\epsilon \sigma_e^2 d_e}$ centered at μ_e. For each environment e, consider constructing the halfspace which is perpendicular to the line connecting

μ_e and μ_{E+1} and tangent to B_e. This halfspace fully contains B_e, and therefore the measure of B_e is upper bounded by that of the halfspace.

By rotational invariance of the Gaussian distribution, we can rotate this halfspace into one dimension and the measure will not change. The center of the ball is $(\sqrt{\epsilon}+\delta)\sigma_e\sqrt{d_e}$ away from the mean μ_{E+1}, so accounting for its radius, the distance from the mean to the halfspace is $\delta\sigma_e\sqrt{d_e}$. The variance of the rotated distribution one dimension is $\sigma_{E+1}^2 d_e$, so the measure of this halfspace is upper bounded by

$$\begin{aligned} 1-\Phi\left(\frac{\delta\sigma_e\sqrt{d_e}}{\sqrt{\sigma_{E+1}^2 d_e}}\right) &\leq \Phi\left(-\sqrt{k}\delta\right) \\ &\leq \frac{1}{\sqrt{k\pi}\delta}\exp\{-k\delta^2\}, \end{aligned}$$

using results from [21]. There are $2E$ such balls comprising $\mathcal{B}_r$, which can be combined via union bound. □

With these two lemmas, we now state the full version of Theorem 6.1, with the main difference being that it allows for any environmental variance.

Theorem D.3 (Non-linear case, full)**.** *Suppose we observe E environments $\mathcal{E} = \{e_1, e_2, \ldots, e_E\}$. Then, for any $\epsilon > 1$, there exists a featurizer Φ_ϵ which, combined with the ERM-optimal classifier $\hat{\beta} = [\beta_c, \beta_{e;ERM}, \beta_0]^T$, satisfies the following properties, where we define $p_{\epsilon,d_e} := \exp\{-d_e \min((\epsilon-1), (\epsilon-1)^2)/8\}$:*

1. *Define $\sigma_{\max}^2 = \max_e \sigma_e^2$. Then the regularization term of $\Phi_\epsilon, \hat{\beta}$ is bounded as*
$$\frac{1}{E}\sum_{e\in\mathcal{E}} \|\nabla_{\hat{\beta}}\mathcal{R}^e(\Phi_\epsilon, \hat{\beta})\|_2^2 \in \mathcal{O}\left(p_{\epsilon,d_e}^2\left[\epsilon d_e \sigma_{\max}^4 \exp\{2\epsilon\sigma_{\max}^2\} + \overline{\|\mu\|_2^2}\right]\right).$$

2. *$\Phi_\epsilon, \hat{\beta}$ exactly matches the optimal invariant predictor on at least a $1 - p_{\epsilon,d_e}$ fraction of the training set. On the remaining inputs, it matches the ERM-optimal solution.*

Further, for any test distribution with environmental parameters $(\mu_{E+1}, \sigma_{E+1}^2)$, suppose the environmental mean μ_{E+1} is sufficiently far from the training means:

$$\forall e \in \mathcal{E}, \min_{y\in\{\pm 1\}} \|\mu_{E+1} - y\cdot\mu_e\|_2 \geq (\sqrt{\epsilon}+\delta)\sigma_e\sqrt{d_e} \tag{15}$$

for some $\delta > 0$. Define the constants:

$$\begin{aligned} k &= \min_{e\in\mathcal{E}} \frac{\sigma_e^2}{\sigma_{E+1}^2} \\ q &= \frac{2E}{\sqrt{k\pi}\delta}\exp\{-k\delta^2\}. \end{aligned}$$

Then the following holds:

3. *$\Phi_\epsilon, \hat{\beta}$ is equivalent to the ERM-optimal predictor on at least a $1 - q$ fraction of the test distribution.*

4. *Under Assumption 1, suppose it holds that*
$$\mu_{E+1} = -\sum_{e\in\mathcal{E}} \alpha_e \mu_e \tag{16}$$
for some set of coefficients $\{\alpha_e\}_{e\in\mathcal{E}}$. Then for any $c \in \mathbb{R}$, so long as
$$\sum_{e\in\mathcal{E}} \alpha_e \frac{\|\mu_e\|_2^2}{\sigma_e^2} \geq \frac{\|\mu_c\|_2^2/\sigma_c^2 + |\beta_0|/2 + c\sigma_{ERM}}{1-\gamma}, \tag{17}$$
the 0-1 risk of $\Phi_\epsilon, \hat{\beta}$ is lower bounded by $F(2c) - q$.

Proof. Define $r = \sqrt{\epsilon\sigma_e^2 d_e}$ and construct $\mathcal{B}_r \subset \mathbb{R}^{d_e}$ as

$$\mathcal{B}_r = \left[\bigcup_{e\in\mathcal{E}} B_r(\mu_e)\right] \cup \left[\bigcup_{e\in\mathcal{E}} B_r(-\mu_e)\right], \tag{18}$$

where $B_r(\alpha)$ is the ball of ℓ_2-norm radius r centered at α. Further construct Φ_ϵ using $\mathcal{B}_r$ as follows:

$$\Phi_\epsilon(x) = \begin{cases} \begin{bmatrix} z_c \\ 0 \end{bmatrix}, & z_e \in \mathcal{B}_r \\ \begin{bmatrix} z_c \\ z_e \end{bmatrix}, & z_e \in \mathcal{B}_r^c, \end{cases} \qquad \text{and} \qquad \hat{\beta} = \begin{bmatrix} \beta_c \\ \hat{\beta}_e \\ \beta_0 \end{bmatrix}. \tag{19}$$

Without loss of generality, fix an environment e.

1. By Lemma D.1, the squared gradient norm is upper bounded by

$$\|\nabla_{\hat{\beta}}\mathcal{R}^e(\Phi_\epsilon, \hat{\beta})\|_2^2 \leq \left\|\mathbb{P}(z_e \in \mathcal{B}_r^c)\mathbb{E}[|z_e| \mid z_e \in \mathcal{B}_r^c]\right\|_2^2. \tag{20}$$

   Define $B_e := B_r(\mu_e)$, and observe that $\mathcal{B}_r^c \subseteq B_e^c$. Since $|z_e|$ is non-negative,

$$\mathbb{P}(z_e \in \mathcal{B}_r^c)\mathbb{E}[|z_e| \mid z_e \in \mathcal{B}_r^c] \leq \mathbb{P}(z_e \in B_e^c)\mathbb{E}[|z_e| \mid z_e \in B_e^c]$$

   (this inequality is element-wise). Plugging this into Equation 20 yields

$$\begin{aligned}\|\nabla_{\hat{\beta}}\mathcal{R}^e(\Phi_\epsilon, \hat{\beta})\|_2^2 &\leq \left\|\mathbb{P}(z_e \in B_e^c)\mathbb{E}[|z_e| \mid z_e \in B_e^c]\right\|_2^2 \\ &= [\mathbb{P}(z_e \in B_e^c)]^2 \left\|\mathbb{E}[|z_e| \mid z_e \in B_e^c]\right\|_2^2.\end{aligned}$$

   Define $p = \mathbb{P}(z_e \in \mathcal{B}_r^c) \leq \mathbb{P}(z_e \in B_e^c)$. By Lemma F.3,

$$p \leq p_{\epsilon,d_e} = e^{-d_e \min((\epsilon-1),(\epsilon-1)^2)/8}.$$

   Combining Lemmas F.4 and F.5 gives

$$\begin{aligned}\left\|\mathbb{E}[|z_e| \mid z_e \in B_e^c]\right\|_2^2 &\leq 2d_e\left[\sigma\frac{\phi(r/\sqrt{d_e})}{F(-r/\sqrt{d})}\right]^2 + 2\|\mu_e\|_2^2 \\ &\leq d_e\sigma_e^2 \exp\left\{2\epsilon(\sigma_e^2 - 1/2)\right\}\left[\epsilon\sigma_e^2 + 1\right] + 2\|\mu_e\|_2^2.\end{aligned}$$

   Putting these two bounds together, we have

$$\|\nabla_{\hat{\beta}}\mathcal{R}^e(\Phi_\epsilon, \hat{\beta})\|_2^2 \in \mathcal{O}\left(p_{\epsilon,d_e}^2\left[\epsilon d_e\sigma_{\max}^4 \exp\{2\epsilon\sigma_{\max}^2\} + \|\mu_e\|_2^2\right]\right),$$

   and averaging this value across environments gives the result.

2. $\Phi_\epsilon, \hat{\beta}$ is equal to the optimal invariant predictor on $\mathcal{B}_r$ and the ERM solution on $\mathcal{B}_r^c$. The claim then follows from Lemma F.3.

3. This follows directly from Lemma D.2.

4. With Equation 16, we have that

$$\begin{aligned}\beta_{e;\text{ERM}}^T\mu_{E+1} &= -\sum_{e\in\mathcal{E}} \alpha_e \beta_{e;\text{ERM}}^T\mu_e \\ &\leq -2(1-\gamma)\sum_{e\in\mathcal{E}} \alpha_e \frac{\|\mu_e\|_2^2}{\sigma_e^2} \\ &\leq -2(1-\gamma)\frac{\|\mu_c\|_2^2/\sigma_c^2 + |\beta_0|/2 + c\sigma_{\text{ERM}}}{1-\gamma} \\ &= -(2\|\mu_c\|_2^2/\sigma_c^2 + |\beta_0| + 2c\sigma_{\text{ERM}}).\end{aligned}$$

where we have applied Assumption 1 in the first inequality and Equation 17 in the second. Consider the full set of features $\Phi_\epsilon(x) = f^{-1}(x)$, and without loss of generality assume $y = 1$. The label-conditional distribution of the resulting logit is

$$\beta_c^T z_c + \beta_{e;\text{ERM}}^T z_e + \beta_0 \sim \mathcal{N}\left(\beta_c^T \mu_c + \beta_{e;\text{ERM}}^T \mu_{E+1} + \beta_0, \sigma_{\text{ERM}}^2\right).$$

Therefore, the 0-1 risk is equal to the probability that this logit is negative. This is precisely

$$\begin{aligned} F\left(-\frac{\beta_c^T \mu_c + \beta_{e;\text{ERM}}^T \mu_{E+1} + \beta_0}{\sigma_{\text{ERM}}}\right) &\geq F\left(\frac{(2\|\mu_c\|_2^2/\sigma_c^2 + |\beta_0| + 2c\sigma_{\text{ERM}}) - 2\|\mu_c\|_2^2/\sigma_c^2 - |\beta_0|}{\sigma_{\text{ERM}}}\right) \\ &= F\left(\frac{2c\sigma_{\text{ERM}}}{\sigma_{\text{ERM}}}\right) \\ &= F(2c). \end{aligned}$$

Observe that by the previous part, $\Phi_\epsilon \neq f^{-1}$ on at most a q fraction of observations, so the risk of our predictor $\Phi_\epsilon, \hat{\beta}$ can differ from that of $f^{-1}, \hat{\beta}$ by at most q. Therefore our predictor's risk is lower bounded by $F(2c) - q$. In particular, choosing $c = 1$ recovers the statement in the main body.

□

D.2 Discussion of Conditions and Assumption

To see just how often we can expect the conditions for Theorem D.3 to hold, we can do a Fermi approximation based on the expectations of each of the terms. A reasonable prior for the environmental means is a multivariate Gaussian $\mathcal{N}(m, \Sigma)$. We might expect them to be very concentrated (with $\text{Tr}(\Sigma)$ small), or perhaps to have a strong bias (with $\|m\|_2^2 \gg \text{Tr}(\Sigma)$). For simplicity we treat the variances σ_c^2, σ_e^2 as constants. Then the expected separation between any two means from this distribution is

$$\mathbb{E}[\|\mu_1 - \mu_2\|_2] = \mathbb{E}_{x\sim\mathcal{N}(0,2\Sigma)}[\|x\|_2] \approx \sqrt{2\,\text{Tr}(\Sigma)}.$$

In high dimensions this value will tightly concentrate around the mean, which is in $\mathcal{O}(\sqrt{d_e})$. On the other hand, even a slight deviation from this separation, to $\Omega(\sqrt{d_e \log E})$, means $\delta \in \Omega(\sqrt{\log E})$, which implies $q \in \mathcal{O}(1/E)$; this is plenty small to ensure worse-than-random error on the test distribution.

Now we turn our attention to the second condition (17). The expected squared norm of each mean is d_e, and in high dimensions we expect them to be reasonably orthogonal (as a rough approximation; this is technically not true with a non-centered Gaussian). Then so long as $\sum_i \alpha_i \in \Omega(1)$, the left-hand side of Equation 17 is approximately d_e. On the other hand, treating γ as a constant, the right-hand side is close to $d_c + \sqrt{d_c + d_e} \in \mathcal{O}(d_c + \sqrt{d_e})$. Thus, Equation 17 is quite likely to hold for any mean μ_{E+1} with the same scale as the training environments but with reversed correlations—again, this is *exactly the situation* where IRM hopes to outperform ERM, and we have shown that it does not.

We can also do a quick analysis of Assumption 1 under this prior: the ERM-optimal non-invariant coefficient will be approximately $2m/\sigma_e^2$ with high probability, meaning $\hat{\beta}^T \mu \approx 2\|m\|_2^2/\sigma_e^2$ for every environment. Thus, this vector will be γ-close to optimal with $\gamma \approx 1$ for every environment with high probability.

E Extensions to Alternative Objectives

E.1 Extensions for the Linear Case

Observe that the constraint of Equation 4 is strictly stronger than that of [4]; when the former is satisfied, the penalty term of the latter is necessarily 0. It is thus trivial to extend all results in the

Section 5 to this objective. As another example, consider the risk-variance-penalized objective of [20]:

$$\min_{\Phi,\hat{\beta}} \quad \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \mathcal{R}^e(\Phi,\hat{\beta}) + \lambda \mathrm{Var}_{e\in\mathcal{E}}\left(\mathcal{R}^e(\Phi,\hat{\beta})\right), \tag{21}$$

It is simple to extend Theorem 5.1 under an additional assumption:

Corollary E.1 (Extension to Theorem 5.1)**.** *Assume f is linear. Suppose we observe $E \leq d_e$ environments with linearly independent means and identical variance σ_e^2. Consider minimizing empirical risk subject to a penalty on the risk variance (21). Then there exists a $\Phi, \hat{\beta}$ dependent on the non-invariant features which achieves a lower objective value than the optimal invariant predictor for* any *choice of regularization parameter $\lambda \in [0, \infty]$.*

Proof. Consider the featurizer Φ constructed in Lemma C.2. If the environmental variance is constant, then the label-conditional distribution of the environmental features,

$$z_e \mid y \sim \mathcal{N}(y \cdot \tilde{\mu}\sigma_e^2, \sigma_e^2),$$

is also invariant. This implies that the optimal $\hat{\beta}$ also has constant risk across the environments, meaning the penalty term is 0, and as a result the objective does not depend on the choice of λ. As in 5.1, invoking Lemma F.1 implies that the overall risk is lower than that of the optimal invariant predictor. □

As mentioned in Section 5, this additional requirement of constant variance is due to the assumptions underlying the design of the objective—REx expects invariance of the conditional distribution $p(y \mid \Phi(x))$, in contrast with the strictly weaker invariance of $\mathbb{E}[y \mid \Phi(x)]$ assumed by IRM. This might seem to imply that REx is a more robust objective, but this does not convey the entire picture. The conditions for the above corollary are just one possible failure case for REx; by extending Theorem 5.3 to this objective, we see that REx is just as prone to bad solutions:

Corollary E.2 (Extension to Theorem 5.3)**.** *Suppose we observe $E \leq d_e$ environments, such that all environmental means are linearly independent. Then there exists a $\Phi, \hat{\beta}$ which uses* only environmental features *and, under any choice of $\lambda \in [0, \infty]$, achieves a lower objective value than the optimal invariant predictor under 0-1 loss on every environment e such that $\tilde{\mu} > \sigma_c^{-1}\|\mu_c\|_2 + \frac{|\beta_0|}{2\sigma_c^{-1}\|\mu_c\|_2}$.*

Proof. We follow the proof of Theorem 5.3, except when solving for p as in Lemma C.1 we instead find the unit-norm vector such that

$$p^T\mu_e = \sigma_e\tilde{\mu} \quad \forall e \in \mathcal{E}. \tag{22}$$

Observe that by setting $\Phi(x) = [p^T z_e]$ and $\hat{\beta} = [1]$, the 0-1 risk in a given environment is

$$\eta F(-\tilde{\mu}\sigma_e/\sigma_e) + (1-\eta)F(-\tilde{\mu}\sigma_e/\sigma_e) = F(-\tilde{\mu}),$$

which is independent of the environment. Further, by carrying through the same proof as in Theorem 5.3, we get that this non-invariant predictor has lower 0-1 risk so long as

$$\alpha + \frac{|\beta_0|}{2\alpha} \leq \tilde{\mu},$$

where $\alpha = \sigma_c^{-1}\|\mu_c\|_2$ □

Though $\tilde{\mu}$ here is not exactly the same value because of the slightly different solution (22), it depends upon the geometry of the training environments in the same way—it is the same as taking the square root of each of the variances. We can therefore expect this condition to hold in approximately the same situations, which we empirically verify by replicating Figure C.2 with the modified equation below.

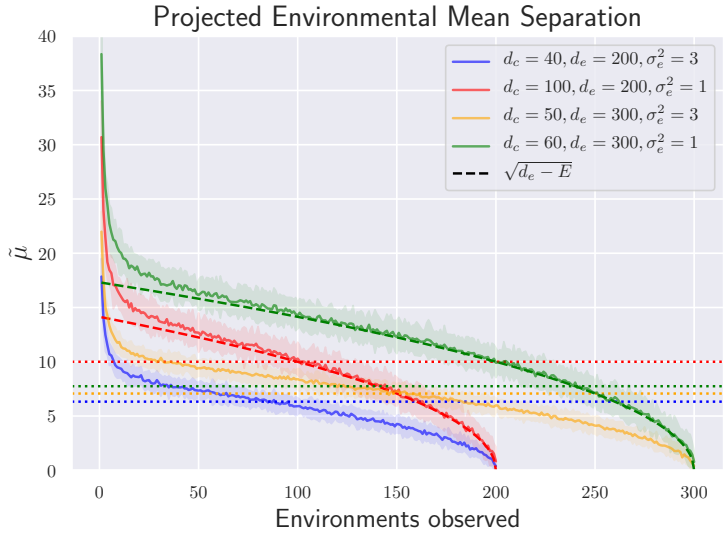

Figure E.1: Simulations to evaluate $\tilde{\mu}$ for varying ratios of $\frac{d_e}{d_c}$. When $\sigma_e^2 = 1$, the value closely tracks $\sqrt{d_e - E}$. Due to the similarity of Equation 22 to Equation 11, it makes sense that the results are very similar to those presented in Figure C.2.

E.2 Extensions for the Non-Linear Case

The failure of these objectives in the non-linear regime is even more straightforward, as we can keep unchanged the constructed predictor from Theorem 6.1. Observe that parts 2-4 of the theorem do not involve the objective itself, and therefore do not require modification.

To see that part 1 still holds, note that since the constructed predictor matches the optimal invariant predictor on $1 - p$ of the observations, its risk across environments can only vary on the remaining p fraction: as the 0-1 risk on this fraction is bounded between 0 and p, it is immediate that the variance of the environmental risks is upper bounded by $\frac{p^2}{4}$, which is in $\mathcal{O}(p^2)$ as before. Applying this argument to the other objectives yields similar results.

F Technical Lemmas

Lemma F.1. *Consider solving the standard logistic regression problem*

$$z \sim p(z) \in \mathbb{R}^k,$$
$$y = \begin{cases} +1 & \text{w.p. } \sigma(\beta^T z), \\ -1 & \text{w.p. } \sigma(-\beta^T z). \end{cases}$$

Assume that none of the latent dimensions are degenerate—$\forall S \subseteq [k]$, $\mathbb{P}(\beta_S^T z_S \neq 0) > 0$, and no feature can be written as a linear combination of the other features. Then for any distribution $p(z)$, any classifier $f(z) = \sigma(\beta_S^T z_S)$ that uses a strict subset of the features $S \subsetneq [k]$ has strictly higher risk with logistic loss than the Bayes classifier $f^*(z) = \sigma(\beta^T z)$. This additionally holds for 0-1 loss if $\beta_{-S}^T z_{-S}$ has greater magnitude and opposite sign of $\beta_S^T z_S$ with non-zero probability.

Proof. The Bayes classifier suffers the minimal expected loss for each observation z. Therefore, another classifier has positive excess risk if and only if it disagrees with the Bayes classifier on a set of non-zero measure. Consider the set of values z_{-S} such that $\beta_{-S}^T z_{-S} \neq 0$. Then on this set we have

$$f^*(\beta^T z) = \sigma(\beta_S^T z_S + \beta_{-S}^T z_{-S}) \neq \sigma(\beta_S^T z_S) = f(z).$$

Since these values occur with positive probability, f has strictly higher logistic risk than f^*. By the same argument, there exists a set of positive measure under which

$$f^*(\beta^T z) = \text{sign}(\beta_S^T z_S + \beta_{-S}^T z_{-S}) \neq \text{sign}(\beta_S^T z_S) = f(z),$$

and so f also has strictly higher 0-1 risk. $\square$

Lemma F.2. *For any feature vector which is a linear function of the invariant and environmental features $\tilde{z} = Az_c + Bz_e$, any optimal corresponding coefficient for an environment e is of the form*

$$2(AA^T\sigma_c^2 + BB^T\sigma_e^2)^+(A\mu_c + B\mu_e),$$

where G^+ is a generalized inverse of G.

Proof. We begin by evaluating a closed form for $p^e(y \mid \tilde{z})$. We have:

$$\begin{aligned}
&p^e(y \mid Az_c + Bz_e = \tilde{z}) \\
=&\frac{p(Az_c + Bz_e = \tilde{z} \mid y)p(y)}{p^e(Az_c + Bz_e = \tilde{z})} \\
=&\frac{p^e(Az_c + Bz_e = \tilde{z} \mid y)}{p^e(Az_c + Bz_e = \tilde{z} \mid y) + p^e(Az_c + Bz_e = \tilde{z} \mid -y)} \\
=&\frac{1}{1 + \frac{p^e(Az_c + Bz_e = \tilde{z} \mid -y)}{p^e(Az_c + Bz_e = \tilde{z} \mid y)}}.
\end{aligned}$$

Now we need a closed form expression for $p(Az_c + Bz_e = \tilde{z} \mid y)$. Noting that $z_c \perp\!\!\!\perp z_e \mid y$, this is a convolution of the two independent Gaussian densities, which is the density of their sum. In other words,

$$Az_c + Bz_e \mid y \sim \mathcal{N}(y \cdot (\overbrace{A\mu_c + B\mu_e}^{\bar{\mu}}), \overbrace{AA^T\sigma_c^2 + BB^T\sigma_e^2}^{\bar{\Sigma}}).$$

Thus,

$$p^e(Az_c + Bz_e = \tilde{z} \mid y) = \frac{1}{\left(2\pi|\bar{\Sigma}|\right)^{k/2}} \exp\left\{-\frac{1}{2}(\tilde{z} - y \cdot \bar{\mu})^T\bar{\Sigma}^+(\tilde{z} - y \cdot \bar{\mu})\right\}.$$

Canceling common terms, we get

$$\begin{aligned}
p^e(y = 1 \mid Az_c + Bz_e = \tilde{z}) &= \frac{1}{1 + \frac{p^e(Az_c + Bz_e = \tilde{z} \mid -y)}{p^e(Az_c + Bz_e = \tilde{z} \mid y)}} \\
&= \frac{1}{1 + \exp\left\{-y \cdot 2\tilde{z}^T\bar{\Sigma}^+\bar{\mu}\right\}} \\
&= \sigma\left(y \cdot 2\tilde{z}^T\bar{\Sigma}^+\bar{\mu}\right).
\end{aligned}$$

Therefore, given a feature vector $\tilde{z}$, the optimal coefficient vector is $2\bar{\Sigma}^+\bar{\mu}$. $\square$

Lemma F.3. *For any environment e with parameters μ_e, σ_e^2 and any $\epsilon > 1$, define*

$$B := B_{\sqrt{\epsilon\sigma_e^2 d_e}}(\mu_e),$$

where $B_r(\alpha)$ is the ball of ℓ_2-norm radius r centered at α. Then for an observation drawn from p^e, we have

$$\mathop{\mathbb{P}}_{z_e \sim p^e}(z_e \in B^c) \leq \exp\left\{-\frac{d_e \min((\epsilon - 1), (\epsilon - 1)^2)}{8}\right\}.$$

Proof. Without loss of generality, suppose $y = 1$. We have

$$\begin{aligned}
\mathbb{P}(z_e \in B) &\geq \mathop{\mathbb{P}}_{z_e \sim \mathcal{N}(\mu_e, \sigma_e^2 I)}\left(\|z_e - \mu_e\|_2 \leq \sqrt{\epsilon\sigma_e^2 d_e}\right) \\
&= \mathop{\mathbb{P}}_{z_e \sim \mathcal{N}(0, \sigma_e^2 I)}\left(\|z_e\|_2 \leq \sqrt{\epsilon\sigma_e^2 d_e}\right) \\
&= \mathop{\mathbb{P}}_{z_e \sim \mathcal{N}(0, I)}\left(\|z_e\|_2^2 \leq \epsilon d_e\right).
\end{aligned}$$

Each term in the squared norm of z_e is a random variable with distribution χ^2_1, which means their sum has mean d_e and is sub-exponential with parameters $(2\sqrt{d_e}, 4)$. By standard sub-exponential concentration bounds we have

$$\mathop{\mathbb{P}}_{z_e \sim \mathcal{N}(0,I)} \left(\|z_e\|_2^2 \geq \epsilon d_e \right) \leq \exp\left\{ -\frac{d_e \min((\epsilon-1), (\epsilon-1)^2)}{8} \right\},$$

which immediately implies the claim. □

Lemma F.4. *Let $z \sim \mathcal{N}(\mu, \sigma^2 I_d)$ be a multivariate isotropic Gaussian in d dimensions, and for some $r > 0$ define B as the ball of ℓ_2 radius r centered at μ. Then we have*

$$\left\| \mathbb{E}[|z| \mid z \in B^c] \right\|_2^2 \leq 2d \left[\sigma \frac{\phi(r/\sqrt{d})}{F(-r/\sqrt{d})} \right]^2 + 2\|\mu\|_2^2,$$

where ϕ, F are the standard Gaussian PDF and CDF.

Proof. Observe that

$$\begin{aligned}
\mathop{\mathbb{E}}_{z \sim \mathcal{N}(\mu, \sigma^2 I_d)} [|z| \mid z \in B^c] &= \mathop{\mathbb{E}}_{z \sim \mathcal{N}(\mu, \sigma^2 I_d)} [|z| \mid \|z - \mu\|_2 > r] \\
&= \mathop{\mathbb{E}}_{z \sim \mathcal{N}(0, \sigma^2 I_d)} [|z + \mu| \mid \|z\|_2 > r] \\
&\leq \mathop{\mathbb{E}}_{z \sim \mathcal{N}(0, \sigma^2 I_d)} [|z| \mid \|z\|_2 > r] + |\mu|.
\end{aligned}$$

Now, consider the expectation for an individual dimension, and note that $|z_i| > \frac{r}{\sqrt{d}} \, \forall i \implies \|z\|_2 > r$. So because the dimensions are independent, conditioning on this event can only increase the expectation:

$$\begin{aligned}
\mathop{\mathbb{E}}_{z \sim \mathcal{N}(0, \sigma^2 I_d)} [|z_i| \mid \|z\|_2 > r] &\leq \mathop{\mathbb{E}}_{z_i \sim \mathcal{N}(0, \sigma^2)} \left[|z_i| \mid |z_i| > \frac{r}{\sqrt{d}} \right] \\
&= \mathop{\mathbb{E}}_{z_i \sim \mathcal{N}(0, \sigma^2)} \left[z_i \mid z_i > \frac{r}{\sqrt{d}} \right],
\end{aligned}$$

where the equality is because the distribution is symmetric about 0. This last term is known as the *conditional tail expectation* of a Gaussian and is available in closed form:

$$\mathop{\mathbb{E}}_{z_i \sim \mathcal{N}(0, \sigma^2)} \left[z_i \mid z_i > \frac{r}{\sqrt{d}} \right] = \sigma \frac{\phi(F^{-1}(\alpha))}{1 - \alpha},$$

where $\alpha = F(r/\sqrt{d})$. Combining the above results, squaring with $(a+b)^2 \leq 2(a^2 + b^2)$, and summing over dimensions, we get

$$\begin{aligned}
\left\| \mathbb{E}[|z| \mid z \in B^c] \right\|_2^2 &\leq 2 \sum_{i=1}^{d} \mathop{\mathbb{E}}_{z_i \sim \mathcal{N}(0, \sigma^2)} \left[z_i \mid z_i > \frac{r}{\sqrt{d}} \right]^2 + 2\|\mu\|_2^2 \\
&= 2d \left[\sigma \frac{\phi(r/\sqrt{d})}{F(-r/\sqrt{d})} \right]^2 + 2\|\mu\|_2^2,
\end{aligned}$$

as desired. □

Lemma F.5. *For $\sigma, \epsilon > 0$, define $r = \sqrt{\epsilon}\sigma$. Then*

$$\left[\frac{\phi(r)}{F(-r)} \right]^2 \leq \frac{1}{2} \exp\left\{ 2\epsilon(\sigma^2 - 1/2) \right\} \left[\epsilon\sigma^2 + 1 \right].$$

Proof. We have

$$\phi(r) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\epsilon}{2}\right\}$$

and

$$F(-r) \geq \frac{2\exp\{-\epsilon\sigma^2\}}{\sqrt{\pi}(\sqrt{\epsilon}\sigma + \sqrt{\epsilon\sigma^2 + 2})}$$

(see [21]). Dividing them gives

$$\frac{\phi(r)}{F(-r)} \leq \frac{1}{2\sqrt{2}} \exp\{\epsilon(\sigma^2 - 1/2)\} \left[\sqrt{\epsilon}\sigma + \sqrt{\epsilon\sigma^2 + 2}\right].$$

Squaring and using $(a+b)^2 \leq 2(a^2 + b^2)$ yields the result. □