Malign Overfitting: Interpolation and Invariance are Fundamentally at Odds

Anonymous Author(s)

Abstract

Learned classifiers should often possess certain invariance properties meant to encourage fairness, robustness, or out-of-distribution generalization. Multiple recent works empirically demonstrate that common invariance-inducing regularizers are ineffective in the over-parameterized regime, in which classifiers perfectly fit (i.e. interpolate) the training data. In this work we provide a theoretical justification for these observations. We prove that - even in the simplest of settings - any interpolating classifier (with nonzero margin) will not satisfy these invariance properties. We then propose and analyze an algorithm that - in the same setting - successfully learns a non-interpolating classifier that is provably invariant. Validation of our theoretical observations is performed on simulated data and the Waterbirds dataset.

1 Introduction

Modern machine learning applications often call for models which are not only accurate, but are also robust to distribution shifts and satisfy fairness constraints. For example, we may wish to avoid using hospital specific traces in X-ray images [12, 46], as they rely on spurious correlations that will fail when deployed in a new hospital, or we might seek models with similar error rates across protected demographic groups in the context of loan applications [7]. A developing paradigm for fulfilling such requirements is learning models that satisfy some notion of *invariance* [27, 28] across environments or sub-populations. Many techniques for learning invariant models have been proposed including penalties that encourage notions of invariance [e.g. 3, 40, 43, 30], data re-weighting [34, 44, 17], causal graph analysis [38], and more [1].

While this is a promising approach, many current invariance-inducing methods often fail to improve over naive approaches. This is especially noticeable when these methods are used with overparameterized deep models capable of *interpolating* [13, 14, 25, 41, 10]. Two parallel lines of research address this problem. The first attempts to come up with alternative learning rules that are capable of interpolating while still endowing meaningful invariance properties to the solutions [18, 44]. These works are motivated in part by the phenomenon of "benign overfitting" [6, 5], whereby interpolating overparameterized models achieve excellent generalization performance on an identically-distributed test set [8, 37]. The second line of research forgoes interpolation, and instead applies invariance inducing techniques with small models on top of representations learned by some other means [32, 41, 19, 25, 21], as well as by subsampling techniques [17, 9]. As both lines of research report encouraging empirical results, it is not clear which one is the preferred way forward. In this work we give theoretical arguments to address this question, showing that interpolating models are fundamentally less invariant than non-interpolating ones. In other words, beyond identically-distributed test sets, overfitting is no longer benign. This will be demonstrated on a simple overparaeterized model, similar to those used in [36, 31, 35], as we now turn to describe.

2 Overview of Setting and Results

Our analysis focuses on learning linear models over data collected from a mixture of two Gaussians.

Definition 1. *An* environment *is a distribution parameterized by* $(\boldsymbol{\mu}_c, \boldsymbol{\mu}_s, d, \sigma, \theta)$ *where* $\theta \in [-1, 1]$ *and* $\boldsymbol{\mu}_c, \boldsymbol{\mu}_s \in \mathbb{R}^d$ *satisfy* $\boldsymbol{\mu}_c \perp \boldsymbol{\mu}_s$ *and with samples generated according to:* $\mathbb{P}_\theta(y) = \text{Unif}\{-1, 1\}$, *and* $\mathbb{P}_\theta(\mathbf{x}|y) = \mathcal{N}(y\boldsymbol{\mu}_c + y\theta\boldsymbol{\mu}_s, \sigma^2 I)$.

Submitted to the Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

We focus on problems with two "training environments" [3, 27] $\mathbb{P}_{\theta_e}$ for $e \in \{1, 2\}$, that share all their parameters other than θ.

Definition 2 (Linear Two Environment Problem and Robust Error)**.** *In a Linear Two Environment Problem we have datasets S_1, S_2 of sizes N_1, N_2 drawn from $\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}$ respectively, where $\boldsymbol{\mu}_c$ and $\boldsymbol{\mu}_s$ satisfy $\|\boldsymbol{\mu}_c\| = r_c$ and $\|\boldsymbol{\mu}_s\| = r_s$ and $N := N_1 + N_2$. $S_1 \cup S_2$ is the pooled dataset $S = \{\mathbf{x}_i, y_i\}_{i=1}^N$ and a learning algorithm is a (possibly randomized) mapping from the tuple (S_1, S_2) to $\mathbf{w} \in \mathbb{R}^d$, whose robust error is: $\max_{\theta \in [-1,1]} \epsilon_\theta(\mathbf{w})$, where $\epsilon_\theta(\mathbf{w}) := \mathbb{E}_{\mathbf{x},y\sim\mathbb{P}_\theta}[\mathrm{sign}(\langle \mathbf{w}, \mathbf{x}\rangle) \neq y]$.*

We study settings where θ_1, θ_2 are fixed and d is large compared to N, i.e. the overparameterized regime. The power of this simple model is that many common invariance criteria boil down to the same mathematical constraint:[1] learning a classifier that is orthogonal to $\boldsymbol{\mu}_s$, which induces a spurious correlation between the environment and the label. In terms of predictive accuracy, the goal of learning a linear model that aligns with $\boldsymbol{\mu}_c$ and is orthogonal to $\boldsymbol{\mu}_s$ coincides with providing guarantees on the robust error, i.e. the error when data is generated with values of $\theta \neq \theta_1, \theta_2$

Statement of Main Result. The question we study is whether algorithms that perfectly fit, i.e. interpolate, their training data can learn models with low robust error. To give a meaningful answer, we use the notion of normalized margin. Ideally we would like to give a result on all classifiers that attain training error zero in terms of the 0-1 loss. However, the inherent discontinuity of this loss would make any such statement sensitive to instabilities and pathologies.[2] Hence the margin serves as a surrogate for this notion.

Definition 3 (Normalized margin)**.** *Let $\gamma > 0$, we say a classifier $\mathbf{w} \in \mathbb{R}^d$ separates the set $S = \{\mathbf{x}_i, y_i\}_{i=1}^N$ with* normalized margin *γ if it satisfies for each point in S: $y_i\langle \mathbf{w}, \mathbf{x}_i\rangle / \|\mathbf{w}\| > \gamma\sqrt{\sigma^2 d}$.*

The $\sqrt{\sigma^2 d}$ scaling of γ is roughly proportional to $\|\mathbf{x}\|$ under our data model in Definition 1, and keeps the value of γ comparable across growing values of d. Our main result is as follows.

Theorem 1. *For any sample sizes $N_1, N_2 > 65$, margin lower bound $\gamma < \frac{1}{4\sqrt{N_1+N_2}}$ and target robust error $\epsilon > 0$, there exist parameters $r_c, r_s > 0, d > N_1 + N_2, \sigma, \theta_1, \theta_2$ such that the following holds for the Linear Two Environment Problem (Definition 2) with these parameters.*

1. ***Invariance is attainable**. Algorithm 1 maps (S_1, S_2) to a linear classifier $\mathbf{w}$ such that with probability at least $99/100$ (over the draw S), the* robust error *of $\mathbf{w}$ is less than ϵ.*
2. ***Interpolation is attainable**. With probability at least $99/100$, the signed-sample-mean estimator $\mathbf{w}_{\mathrm{mean}} = N^{-1}\sum_{i\in[N]} y_i \mathbf{x}_i$ separates S with normalized margin greater than $\frac{1}{4}(N_1 + N_2)^{-1/2}$.*
3. ***Interpolation is at odds with invariance.** Given $\boldsymbol{\mu}_c$ uniformly distributed on the sphere of radius r_c and $\boldsymbol{\mu}_s$ uniformly distributed on a sphere of radius r_s in the subspace orthogonal to $\boldsymbol{\mu}_c$, let $\mathbf{w}$ be any classifier learned from (S_1, S_2) as per Definition 2. If $\mathbf{w}$ separates S with normalized margin γ, then with probability at least $99/100$ (over the draw of $\boldsymbol{\mu}_c, \boldsymbol{\mu}_s$, and the sample), the* robust error *of $\mathbf{w}$ is at least $1/2$.*

Essentially, Theorem 1 shows that if a learning algorithm for overparameterized linear classifiers always separates its training data, then there exist natural settings for which the algorithm completely fails to learn a robust classifier. It holds *arbitrarily small* margins γ, where the maximum achievable margin is at least of the order of $1/\sqrt{N}$. Therefore, we believe that Theorem 1 essentially precludes any learning that always fits the data from being consistently invariant. It also shows that failure can be avoided, as there is an algorithm (that *necessarily* does not always separate its training data) which successfully learns an invariant classifier. Appendix A further elaborates on the regimes where failure occurs and how the theorem relates to known results. We establish Theorem 1 with three propositions in Section 4, Appendix E and in Section 3, which we put together by choosing the free parameters in Appendix G so that all the claims hold simultaneously.

3 Interpolating Models Cannot Be Invariant

In this section we prove the third claim in Theorem 1. We set $\sigma^2 d = 1$ and $\theta_1 = 1, \theta_2 = 0$, meaning the spurious correlation is prevalent in the first environment and absent from the second. Our claim

[1] These include Equalized Odds [15], distribution matching [23], multi-domain calibration [16, 43], Risk Extrapolation [20]. See discussion in Appendix H.

[2] For instance, if we do not limit the capacity of our models, we can turn any classifier into an interpolating one by adding "special cases" for the training points, yet intuitively this is not the type of interpolation that we would like to study.

is that, for essentially any nonzero value of γ, there are instances of the Linear Two Environment Problem where with high probability, linear classifiers attaining normalized margin at least γ incur a large robust error. The proof of the following proposition can be found in Appendix D.3.

Proposition 1. *There are universal constants* $c_n \in (0, 1)$ *and* $C_d, C_r \in (1, \infty)$*, such that, for any target normalized* γ *and failure probability* $\delta \in (0, 1)$*, if*

$$\max\{r_s^2, r_c^2\} \le \frac{c_n}{N} \;\; , \;\; \frac{r_s^2}{r_c^2} \ge C_r \left(1 + \frac{\sqrt{N_2}}{N_1 \gamma}\right) \; \textit{and} \;\; d \ge C_d \frac{N}{\gamma^2 N_1^2 r_c^2} \log \frac{1}{\delta}, \tag{1}$$

then with probability at least $1 - \delta$ *over the drawing of* $\boldsymbol{\mu}_c, \boldsymbol{\mu}_s$ *and* (S_1, S_2) *as described in Theorem 1, any* $\hat{\mathbf{w}} \in \mathbb{R}^d$ *that is a measurable function of* (S_1, S_2) *and separates the data with normalized margin larger than* γ *has robust error at least* 0.5.

Proof sketch. The main part of the proof draws a lower bound on the ratio $\langle \mathbf{w}, \boldsymbol{\mu}_s \rangle / \langle \mathbf{w}, \boldsymbol{\mu}_c \rangle$ (with high probability) that is approximately $\left(\|\boldsymbol{\mu}_s\|^2 N_1 \gamma\right) / \left(\|\boldsymbol{\mu}_c\|^2 \sqrt{N_2}\right)$. Therefore, for a classifier that attains margin γ satisfying Equation (1), this ratio is likely to be larger than 1. The ratio directly relates to the robust error: for linear classifiers and Gaussian data, the error $\epsilon_\theta(\mathbf{w})$ is

$$\epsilon_\theta(\mathbf{w}) = Q\left(\frac{\langle \mathbf{w}, \boldsymbol{\mu}_c \rangle + \theta \langle \mathbf{w}, \boldsymbol{\mu}_s \rangle}{\sigma \|\mathbf{w}\|}\right) = Q\left(\frac{\langle \mathbf{w}, \boldsymbol{\mu}_c \rangle}{\sigma \|\mathbf{w}\|}\left(1 + \theta \frac{\langle \mathbf{w}, \boldsymbol{\mu}_s \rangle}{\langle \mathbf{w}, \boldsymbol{\mu}_c \rangle}\right)\right), \tag{2}$$

where $Q(t) := \mathbb{P}(\mathcal{N}(0; 1) > t)$ is the Gaussian tail function. Whenever $\langle \mathbf{w}, \boldsymbol{\mu}_s \rangle / \langle \mathbf{w}, \boldsymbol{\mu}_c \rangle > 1$, it is easy to see that $\epsilon_\theta(\mathbf{w}) = 1/2$ for some $\theta \in [-1, 1]$ and therefore the robust error is at least $\frac{1}{2}$.

To obtain the aforementioned lower bound, we first claim that if we fix a training set $\{\mathbf{x}_i, y_i\}_{i=1}^N$, then the component of $\mathbf{w}$ that is orthogonal to the training set has a negligible contribution to the performance of the classifier (see Corollary 1 in the appendix). This is due to the random generation of $\boldsymbol{\mu}_c, \boldsymbol{\mu}_s$ in our data generating process. Consequently we may write $\mathbf{w} \approx \sum_i \mathbf{x}_i \beta_i$ for some vector $\beta \in \mathbb{R}^N$, and inner products with $\mathbf{w}$ (e.g. $\langle \mathbf{w}, \boldsymbol{\mu}_s \rangle, \langle \mathbf{w}, \mathbf{x}_i \rangle$) can be expressed as linear functions of β. This lets us draw bounds on $\langle \mathbf{w}, \boldsymbol{\mu}_s \rangle$ and $\langle \mathbf{w}, \boldsymbol{\mu}_c \rangle$ under margin constraints via convex duals of the suitable constrained quadratic programs (see Lemma 4 in appendix). These components are put together in Appendix D.3 of the appendix to obtain the bound of interest. □

Implication for invariance-inducing algorithms. Our simulations in Section 5 will show that several popular invariance inducing algorithms interpolate their data in the overparameterized regime. Hence our result predicts that they, as well as any other interpolating algorithm, should fail at learning overparameterized invariant classifiers. It is then natural to ask what type of methods *can* provably learn such models, which leads to our next section and the first part of Theorem 1.

4 A Provably Invariant Overparameterized Estimator

Our approach is a two-staged learning procedure that is conceptually similar to some recently proposed methods [32, 41, 19, 25, 21, 48]. In Section 5 we validate our algorithm on simulations and on the Waterbirds dataset [34], but we leave a thorough empirical evaluation of the techniques described here to future work.

Algorithm 1 (see Appendix F for pseudocode) first evenly[3] splits the data from each environment into the sets $S_e^{\text{trn}}, S_e^{\text{fine}}$, for $e \in \{1, 2\}$. The "Training" stage uses S_e^{trn} to fit an overparameterized, interpolating classifier $\mathbf{w}_e$ *separately* for each environment $e \in \{1, 2\}$. We then use the second portion of the data $S^{\text{fine}} = \{S_1^{\text{fine}}, S_2^{\text{fine}}\}$ to learn an invariant linear classifier over a new representation, which concatenates the outputs of classifiers from the first stage. This classifier is learned by maximizing a score (i.e., minimizing an empirical loss), subject to an empirical version of an invariance constraint. Our analysis uses Equalized Opportunity [15] for convenience (see appendix Appendix F.1 for definition), though any other invariance inducing method can be applied at this stage. Crucially, the invariance penalty is only used in the second stage, in which we are no longer in the overparamterized regime since we are only fitting a two-dimensional classifier. In this way, we overcome the negative result from Section 3.

The guarantees we derive for Algorithm 1 are given in the proposition below, and its full proof is at section F.2 of the appendix.

[3]The even split is used here for simplicity of exposition, and our full proof does not assume it. In practice, allocating more data to the first-stage split would likely perform better.

Proposition 2. *Consider the Linear Two Environment Problem (Definition 2), and further suppose that* $|\theta_1 - \theta_2| > 0.1$.[4] *Let* $\epsilon > 0, \delta \in (0,1)$ *denote the target robust error of the model and failure probability of the algorithm, respectively. Let* $N_{\min} = \min\{N_1, N_2\} \geq C_{\text{opp}} \log(1/\delta)$ *for some* $C_{\text{opp}} \in (1, \infty)$[5] *and assume that for some constants* $C_c, C_s \in (1, \infty)$ *, the following holds:*

$$r_s^2 \geq C_s \sqrt{\log \frac{1}{\delta}} \frac{\sigma^2 \sqrt{d}}{N_{\min}}, \text{ and } r_c^2 \geq C_c \sigma^2 \sqrt{\log \frac{1}{\delta}} \max \left\{ Q^{-1}(\epsilon) \sqrt{\frac{d}{N_{\min}}}, \frac{\sqrt{d}}{N_{\min}}, \frac{r_s^2}{N_{\min} r_c^2} \right\}. \quad (3)$$

Then, with probability at least $1 - \delta$ *over the choice of the training data, the robust error of the model returned by Algorithm 1 does not exceed* ϵ.

5 Empirical Validation

The empirical observations that motivated this work can be found across the literature. We thus focus our simulations on validating the theoretical results in our simplified model and on the popular Waterbirds dataset. Due to space limitations, we defer details on the setup of these experiments to section B and focus this section on evaluation and the results, which are summarized in Figures 3 and 4.

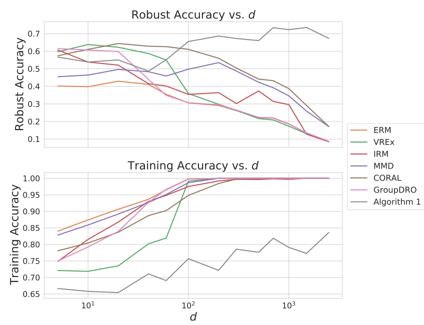

Figure 1: Results for Linear Two Environment Problem simulations. Robust accuracy (top) and training accuracy (bottom) for the different methods.

Linear Two Environment Problem We generate data according to the settings for which we derive our theoretical results, with growing values of d. Robust accuracy and train set accuracy are compared between the learned classifiers, where we use several training meethods implemented in the Domainbed package [13]. First, we observe that all methods except for Algorithm 1 attain perfect accuracy for large enough d, i.e. they interpolate. We further note that while invariance inducing methods give a desirable effect in low dimensions (the non-interpolating regime) – significantly improving the robust error over ERM – they become aligned with ERM in terms of robust accuracy as they go deeper into the interpolation regime (indeed, IRM essentially coincides with ERM for larger d). This is an expected outcome considering our findings in section 3.

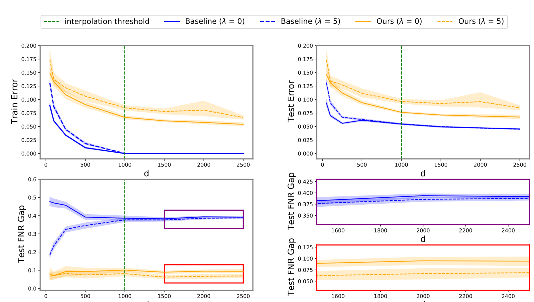

Figure 2: Results for the Waterbirds dataset [34]. **Top row**: Train error (left) and test error (right). **Bottom row**: Comparing the FNR gap on the test set (left), with zoomed-in versions on the right.

Waterbirds. We use the image background type (water or land) as the sensitive feature, denoted by A, and consider the fairness desiderata of Equal Opportunity [15], i.e., similar false negative rate (FNR) for both groups. Towards this, we use the MinDiff penalty [29] with two methods, both learn a linear model over random features extracted from a ResNet-18 representation of the raw image. The baseline trains a regularized logistic regressor with the MinDiff penalty term. Algorithm 1 first learns two logistic regression models, one over data where $A = 0$ and the other where $A = 1$, and then applies regularized risk minimization with MinDiff on a two-dimensional representation obtained as the output of the two logistic regressors. Figure 4 summarizes the results where we run each method with ($\lambda = 5$) and without ($\lambda = 0$) regularization. For the baseline approach, the fairness penalty successfully reduces the FNR gap when the classifier is not interpolating. However, as our negative result predicts and as previously reported in [41], the fairness penalty becomes ineffective in the interpolating regime ($d \geq 1000$). On the other hand, for our two-phased algorithm, the addition of the fairness penalty does reduce the FNR gap with an average relative improvement of 20%); crucially, this improvement is independent of d.

[4] Intuitively, if $|\theta_1 - \theta_2|$ should have a quantifiable effect on our ability to generalize robustly (e.g. when it is 0 robust learning is impossible). the full result in the Appendix takes this item into account

[5] This assumption makes sure we have some positive labels in each environment.

References

[1] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pages 145–155. PMLR, 2020.

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

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[4] Keith Ball et al. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31(1-58):26, 1997.

[5] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[6] Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

[7] Ajay Byanjankar, Markku Heikkilä, and Jozsef Mezei. Predicting credit risk in peer-to-peer lending: A neural network approach. In *2015 IEEE symposium series on computational intelligence*, pages 719–725. IEEE, 2015.

[8] Yuan Cao, Quanquan Gu, and Misha Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=ChWy1anEuow`.

[9] Niladri S. Chatterji, Saminul Haque, and Tatsunori Hashimoto. Undersampling is a minimax optimal robustness intervention in nonparametric classification, 2022. URL `https://arxiv.org/abs/2205.13094`.

[10] Valeriia Cherepanova, Vedant Nanda, Micah Goldblum, John P Dickerson, and Tom Goldstein. Technical challenges for training fair neural networks. *arXiv preprint arXiv:2102.06764*, 2021.

[11] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

[12] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

[13] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[14] Lin Lawrence Guo, Stephen R. Pfohl, Jason Fries, Alistair E. W. Johnson, Jose Posada, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Scientific Reports*, 12(1):2726, 2022.

[15] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[16] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

[17] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, 2022.

[18] Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34, 2021.

[19] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations, 2022. URL `https://arxiv.org/abs/2204.02937`.

[20] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

[21] Ananya Kumar, Tengyu Ma, Percy Liang, and Aditi Raghunathan. Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift. In James Cussens and Kun Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1041–1051. PMLR, 01–05 Aug 2022. URL `https://proceedings.mlr.press/v180/kumar22a.html`.

[22] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, 2019.

[23] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018.

[24] Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the universality of invariant networks. In *International conference on machine learning*, pages 4363–4371. PMLR, 2019.

[25] Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. Overparameterisation and worst-case generalisation: friend or foe? In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=jphnJNOwe36`.

[26] Advait Parulekar, Karthikeyan Shanmugam, and Sanjay Shakkottai. Pac generalization via invariant representations, 2022. URL `https://arxiv.org/abs/2205.15196`.

[27] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

[28] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[29] Flavien Prost, Hai Qian, Qiuwen Chen, Ed H Chi, Jilin Chen, and Alex Beutel. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779*, 2019.

[30] Aahlad Manas Puli, Lily H Zhang, Eric Karl Oermann, and Rajesh Ranganath. Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. In *International Conference on Learning Representations*, 2021.

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

[32] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization, 2022. URL `https://arxiv.org/abs/2202.06856`.

[33] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

[34] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[35] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8346–8356. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/sagawa20a.html`.

[36] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

[37] Ohad Shamir. The implicit bias of benign overfitting. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 448–478. PMLR, 02–05 Jul 2022. URL `https://proceedings.mlr.press/v178/shamir22a.html`.

[38] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019.

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

[40] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=BdKxQp0iBi8`.

[41] Akshaj Kumar Veldanda, Ivan Brugere, Jiahao Chen, Sanghamitra Dutta, Alan Mishler, and Siddharth Garg. Fairness via in-processing in the over-parameterized regime: A cautionary tale. *arXiv preprint arXiv:2206.14853*, 2022.

[42] Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006.

[43] Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. *arXiv preprint arXiv:2102.10395*, 2021.

[44] Ke Alexander Wang, Niladri Shekhar Chatterji, Saminul Haque, and Tatsunori Hashimoto. Is importance weighting incompatible with interpolating classifiers? In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. URL `https://openreview.net/forum?id=pEhpLxVsd03`.

[45] Robert Williamson and Aditya Menon. Fairness risk measures. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6786–6797. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/williamson19a.html`.

[46] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

[47] Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. Understanding why generalized reweighting does not improve over erm, 2022.

[48] Jianyu Zhang, David Lopez-Paz, and Leon Bottou. Rich feature construction for the optimization-generalization dilemma. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26397–26411. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/zhang22u.html`.

A Discussion and Additional Related Work

In terms of formal results, most of the guarantees about invariant learning algorithms rely on the assumption that infinite training data is available [3, 43, 40, 30, 31]. Some exceptions are the works of Ahuja et al. [2] and Parulekar et al. [26] that characterize the sample complexity of methods that learn invariant classifiers, yet they do not analyze the overparameterized cases we are concerned with. Negative results about learning overparameterized robust classifiers have been shown for methods based on importance weighting [47], and negative results on learning with group-robust classifiers have been shown for max-margin classifiers [35]. Our result is thus more general and applies to any learning algorithm that separates the data with arbitrarily small margins, instead of focusing on max-margin classifiers or specific algorithms.

A notable aspect of our result is that it holds for essentially all values of N_2 and N_1. This stands in contrast to prior work such as Sagawa et al. [35], which typically relies on one of the environments being under-represented, i.e., $N_2 \ll N_1$. We are able to sidestep such requirements by making the invariant signal component (r_c) much weaker than the spurious component (r_s), while still allowing for low test error by taking the problem dimension to be sufficiently high. However, when one environment is sufficiently rare (namely $N_2 \leq N_1^2\gamma^2$), we can show that interpolation precludes invariance even when r_s and r_c are of the same order.

Finally, we note that our results hold for classifiers with *arbitrarily small* margin γ, for settings where the maximum achievable margin is always at least of the order of $1/\sqrt{N_1 + N_2}$. Therefore, we believe that Theorem 1 essentially precludes any learning that always fits the data from being consistently invariant. While we focus on the linear case, we believe it is instructive, as any reasonable method is expected to succeed in that case. Nonetheless, we believe our results can be extended to non-linear margins, and we leave this to future work.

One take-away from our result is that while low training loss is not something to avoid, overfitting to the point of interpolation creates a significant difficulty. This means one cannot assume a typical deep learning model with an added invariance penalty will indeed achieve any form of invariance; this fact also motivates using held-out data for imposing invariance, as in our Algorithm 1 as well as several other two-stage approaches mentioned above.

While our focus in this work was on theory underlying a wide array of algorithms, there are many closely related topics that we did not touch upon. For instance, an empirical comparison of two-stage methods along with other methods that avoid interpolation, e.g. by subsampling data [17, 9]. We also note that our focus in this paper was not on types of invariance that are satisfiable by using clever data augmentation techniques (e.g. invariance to image translation), or the design of special architectures (e.g. [11, 22, 24]). These methods cleverly incorporate a-priori known invariances, and their empirical success when applied to large models may suggest that there are lessons to be learned for the type of invariant learning considered in our paper. These connections seem like an exciting avenue for future research.

B Further Details on Empirical Evaluation

Here we provide an extended version of the empirical evaluation section, with more details on the experimental setup and further discussion of the results.

B.1 Simluations

Setup. Our simulation generates data as described in Theorem 1 with two environments where $\theta_1 = 1, \theta_2 = 0$. We further fix $r_c = 1$ and $r_c = 2$, while $N_1 = 800$ and $N_2 = 100$. We then take growing values of d, while adjusting σ so that $(r_c/\sigma)^2 \propto \sqrt{d/N}$.[6] For each value of d we train linear models with IRMv1 [3], VREx [20], MMD [23], CORAL [39], GroupDRO [34], implemented in the Domainbed package [13]. We also train a classifier with the logistic loss to minimize empirical error (ERM), and apply Algorithm 1 where the “fine-tuning" stage trains a linear model over the two-dimensional representation using the VREx penalty to induce invariance. We repeat this for 15 random seeds to set μ_c, μ_s and to draw the training set.

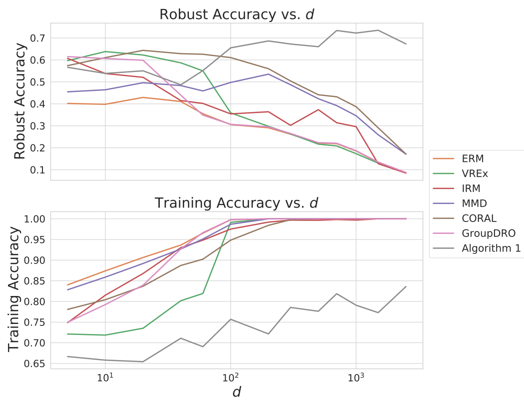

Figure 3: Numerical validation of our theoretical claims. Invariance inducing methods improve robust accuracy compared to ERM in low values of d, but their ability to do so is diminished as d grows (top plot) and they enter the interpolation regime, as seen on the bottom plot for $d > 10^2$. Algorithm 1 learns robust predictors as d grows and does not interpolate.

Evaluation and results. We compare the robust accuracy and the train set accuracy of the learned classifiers as d grows. First, we observe that all methods except for Algorithm 1 attain perfect accuracy for large enough d, i.e. they interpolate. We further note that while invariance inducing methods give a desirable effect in low dimensions (the non-interpolating regime) – significantly improving the robust error over ERM – they become aligned with ERM in terms of robust accuracy as they go deeper into the interpolation regime (indeed, IRM essentially coincides with ERM for larger d). This is an expected outcome considering our findings in section 3, as we set here N_1 to be considerably larger than N_2.

B.2 Waterbirds Dataset

We evaluate Algorithm 1 on the Waterbirds dataset [34], which has been previously used to evaluate the fairness and robustness of deep learning models.

Setup. Waterbirds is a synthetically created dataset containing images of water- and land-birds overlaid on water and land background. Most of the waterbirds (landbirds) appear in water (land) backgrounds, with a smaller minority of waterbirds (landbirds) appearing on land (water) backgrounds. The dataset is split into training, validation and test sets with 4795, 1199 and 5794 images in each set, respectively. We follow previous work [35, 41] in defining a binary task in which waterbirds is the positive class and landbirds are the negative class, and using the following random features setup: for every image, a fixed pre-trained ResNet-18 model is used to extract a d_{rep}-dimensional feature vector $\mathbf{x}'$ ($d_{\mathrm{rep}} = 512$). This feature vector is then converted into an d-dimensional feature vector $\mathbf{x} = \mathrm{ReLU}(U\mathbf{x}')$, where $U \in \mathbb{R}^{d \times d_{\mathrm{rep}}}$ is a random matrix with Gaussian entries. Finally, a logistic regression classifier is trained on $\mathbf{x}$. The extent of over-parameterization in this setup is controlled by varying d, the dimensionality of $\mathbf{x}$. In our experiments we vary d from 50 to 2500, with interpolation empirically observed at $d = 1000$ (which we refer to as the interpolation threshold).

Fairness. We use the image background type (water or land) as the sensitive feature, denoted A, and consider the fairness desiderata of Equal Opportunity [15], i.e., the false negative rate (FNR) should be similar for both groups. Towards this, we use the MinDiff penalty term [29]. It uses the maximum

[6] This is to keep our parameters within the regime where benign overfitting occurs.

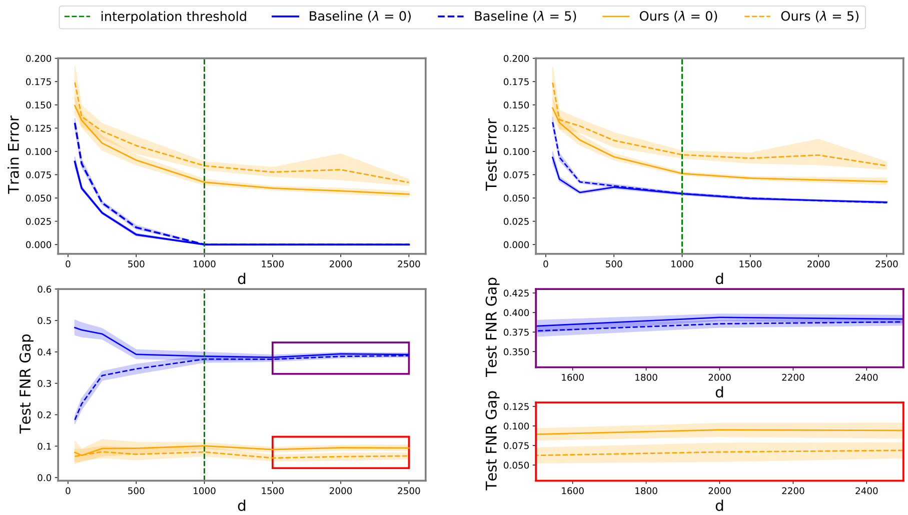

Figure 4: Results for the Waterbirds dataset [34]. **Top row**: Train error (left) and test error (right). The train error is used to identify the interpolation threshold for the baseline method (approximately $d = 1000$). **Bottom row**: Comparing the FNR gap on the test set (left), with zoomed-in versions on the right. For the baseline approach, the fairness penalty successfully reduces the FNR gap when the classifier is not interpolating, but is ineffective in the interpolating regime ($d \geq 1000$). On the other hand, for our two-phased algorithm, the addition of the fairness penalty reduces the FNR gap in a way that is independent of d (average relative improvement 20%).

mean discrepancy (MMD) distance between the model's output for the two sensitive groups when $Y = 1$ as a differentiable proxy to the FNR gap:

$$\mathcal{L}_M(\mathbf{w}) = \text{MMD}\left(\langle \mathbf{w}, X\rangle | A = 0, Y = 1;\ \langle \mathbf{w}, X\rangle | A = 1, Y = 1\right).$$

Evaluation. We compare the following methods: **(1) Baseline**: Learning a linear classifier $\mathbf{w}$ by minimizing $\mathcal{L}_p + \lambda \cdot \mathcal{L}_M$, where $\mathcal{L}_p$ is the standard binary cross entropy loss and $\mathcal{L}_M$ is the MinDiff penalty; **(2) Algorithm 1**: In the first stage, we learn group-specific linear classifiers $\mathbf{w}_0, \mathbf{w}_1$ by minimizing $\mathcal{L}_p$ on the examples from $A = 0$ and $A = 1$, respectively. In the second stage we learn $\boldsymbol{v} \in \mathbb{R}^2$ by minimizing $\mathcal{L}_p + \lambda \cdot \mathcal{L}_M$ on examples the entire dataset, where the new representation of the data is $\tilde{X} = [\langle w_1, X\rangle, \langle w_2, X\rangle] \in \mathbb{R}^2$.[7]

For all the experiments we use the Adam optimizer, a batch size of 128 and a learning rate schedule with initial rate of 0.01 and a decay factor of 10 for every 10,000 gradient steps. Every experiment is repeated 25 times and results are reported over all runs. For the baseline model we train for a total of 30,000 gradient steps whereas for our two-phased algorithm we use 15,000 gradient steps for each model in Phase A and an additional 250 steps for Phase B.

Results. Our main objective is to understand the effect of the fairness penalty. Towards this, for each method we compare both the test error and the test FNR gap when using either $\lambda = 0$ (no regularization) or $\lambda = 5$. The results are summarized in Figure 4. We can see that for the baseline approach, the fairness penalty successfully reduces the FNR gap when the classifier is not interpolating. However, as our negative result predicts and as previously reported in [41], the fairness penalty becomes ineffective in the interpolating regime ($d \geq 1000$). On the other hand, for our two-phased algorithm, the addition of the fairness penalty reduces does reduce the FNR gap with an average relative improvement of 20%); crucially, this improvement is independent of d.

[7]This is basically Algorithm 1 with the following minor modifications: (1) The $\mathbf{w}_e$'s are computed via ERM, rather than simply taken to be the mean estimators; (2) Since the FNR gap penalty is already computed w.r.t a small number of samples, we avoid splitting the data and use the entire training set for both phases; (3) we convert the constrained optimization problem into an unconstrained problem with a penalty term.

C Setting and Helper Lemmas

Notation. Let $\mathbb{U}(\mathrm{O}(d))$ be the uniform distribution over $d \times d$ orthogonal matrices, $\mathrm{Rad}(\alpha)$ the Rademacher distribution with parameter α, and $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ the Gaussian and multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ (the dimension will be clear from context) and $W(\Sigma, d)$ the Wishart distribution with scale matrix Σ and d degrees of freedom. The set $S = [N]$ will denote indices of training examples, $S_1, S_2 \subseteq S$ are the indices of examples in environments $1, 2$ respectively. Our generative process is then:

$$\begin{aligned}
&\mathbf{U} \sim \mathbb{U}(\mathrm{O}(d)) \\
&\boldsymbol{\mu}_c = U_1 \cdot r_c, \boldsymbol{\mu}_s = U_2 \cdot r_s \\
&y_i = \mathrm{Rad}(\frac{1}{2}), n_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d) \quad \forall i \in [N] \\
&\mathbf{x}_i = y_i \boldsymbol{\mu}_c + y_i \theta_e \boldsymbol{\mu}_s + n_i \quad \forall e, i \in S_e.
\end{aligned}$$

The vectors $E_1, E_2 \in \{0,1\}^N$ are binary vectors where $[E_e]_i = 1$ for $i \in S_e$ and $e \in \{1, 2\}$, while $\mathbf{1}$ is the vector of length N whose entries equal 1. We also denote $\mathbf{z}_i = \mathbf{x}_i y_i$ for $i \in S$ and $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]^\top \in \mathbb{R}^{N \times d}$ the matrix that stacks all these vectors. The i-th column of a matrix $\mathbf{M}$ is denoted by M_i, $s_{\min}(\mathbf{M}), s_{\max}(\mathbf{M})$ are its smallest and largest singular values accordingly. The unit matrix of size n is denoted by $\mathbf{I}_n$ and for convenience we denote the direction of any vector $\mathbf{v}$ as $\hat{\mathbf{v}} := \frac{\mathbf{v}}{\|\mathbf{v}\|}$. Finally, for some vector of coefficients $\beta \in \mathbb{R}^N$, we will use the form $\hat{\mathbf{w}} = \sum_{i \in S} \beta_i y_i \mathbf{x}_i + \mathbf{w}_\perp$ where $\mathbf{w}_\perp$ is in the orthogonal complement of $\mathrm{span}(\{\mathbf{x}_i\}_{i \in S})$, to write any linear model (here normalized to unit norm).

For convenience we will write our proofs for the case where $\theta_1 = 1, \theta_2 = 0$ and $\sigma^2 = d^{-1}$, extensions to different settings of these parameters are straightforward but result in a more cumbersome notation.

C.1 Operator Norms of Wishart Matrices

We begin with stating the required events for our results and their occurrence with high-probability:

Lemma 1. *Consider the matrix $\mathbf{G} = \mathbf{Z} - \mathbf{1}\boldsymbol{\mu}_c^\top - E_1 \boldsymbol{\mu}_s^\top$. For any $t > 0$, with probability at least $1 - 6\exp(-t^2/2)$ the following hold simultaneously:*

$$1 - \sqrt{\frac{N}{d}} - \frac{t}{\sqrt{d}} \leq s_{\min}(\mathbf{G}^\top) \leq s_{\max}(\mathbf{G}^\top) \leq 1 + \sqrt{\frac{N}{d}} + \frac{t}{\sqrt{d}} \tag{4}$$

$$\|\mathbf{G}\boldsymbol{\mu}_c\| \leq t\sqrt{\frac{N}{d}}\|\boldsymbol{\mu}_c\| \tag{5}$$

$$\|\mathbf{G}\boldsymbol{\mu}_s\| \leq t\sqrt{\frac{N}{d}}\|\boldsymbol{\mu}_s\| \tag{6}$$

Proof. $\mathbf{G}$ is a random Gaussian matrix with $G_{i,j} \sim \mathcal{N}(0, d^{-1}\mathbf{I}_N)$. By concentration results for random Gaussian matrices [42, Cor. 5.35] we obtain that with probability at least $1 - 2\exp(-t^2/2)$ Equation (4) holds.

Next we note that $\mathbf{G}\boldsymbol{\mu}_c \sim \mathcal{N}(0, d^{-1}\|\boldsymbol{\mu}_c\|^2 \mathbf{I}_N)$ and similarly for $\mathbf{G}\boldsymbol{\mu}_s$. The norm of a Gaussian random vector can be bounded for any $t_2 > 0$:

$$\mathbb{P}\left[\|\mathbf{G}\boldsymbol{\mu}_c\| \geq t_2\right] \leq 2\exp\left(-\frac{dt_2^2}{2N\|\boldsymbol{\mu}_c\|^2}\right)$$

Setting $t_2 = t\sqrt{\frac{N}{d}}\|\boldsymbol{\mu}_c\|$ we get that with probability at least $1 - 2\exp(-t^2/2)$ Equation (5) holds. Repeating the analogous derivation for Equation (6) and taking a union bound over the 3 events, we arrive at the desired result. □

Lemma 2. *Conditioned on the events in Lemma 1 with parameter $t \geq 0$, if*

$$\frac{\sqrt{N} + t}{\sqrt{d}} + \sqrt{N}(\|\boldsymbol{\mu}_c\| + \|\boldsymbol{\mu}_s\|) \leq \frac{1}{2}, \tag{7}$$

then

$$\|\mathbf{Z}\mathbf{Z}^\top - \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]\|_{\mathrm{op}} \le 3\frac{\sqrt{N}+t}{\sqrt{d}} \ \ \textit{and} \ \ \frac{1}{2}I_N \preceq \mathbf{Z}\mathbf{Z}^\top \preceq 2I_N.$$

We note that we already assume $d \gg N$ and $\|\boldsymbol{\mu}_c\| \ll N^{-1/2}$, hence the additional assumption introduced in the conditions of this lemma is regarding the size of $\|\boldsymbol{\mu}_s\|\sqrt{N_1}$.

Proof. Since $\mathbf{G}\mathbf{G}^\top \sim W(d^{-1}\mathbf{I}_N, d)$ we have that $\mathbb{E}[\mathbf{G}\mathbf{G}^\top] = \mathbf{I}_N$. Then from Equation (4) we can also obtain $(1 - \sqrt{\frac{N}{d}} - \frac{t}{\sqrt{d}})^2\,\mathbf{I}_n \preceq \mathbf{G}\mathbf{G}^\top \preceq (1 + \sqrt{\frac{N}{d}} + \frac{t}{\sqrt{d}})^2\,\mathbf{I}_n$, which leads to:

$$\left\|\mathbf{G}\mathbf{G}^\top - \mathbb{E}[\mathbf{G}\mathbf{G}^\top]\right\|_{\mathrm{op}} \le \left(1 + \sqrt{\frac{N}{d}} + \frac{t}{\sqrt{d}}\right)^2 - 1.$$

Combining this with Equation (5) and Equation (6)

$$\begin{aligned}
\|\mathbf{Z}\mathbf{Z}^\top - \mathbb{E}\left[\mathbf{Z}\mathbf{Z}^\top\right]\|_{\mathrm{op}} &\le \|\mathbf{G}\mathbf{G}^\top - \mathbb{E}\left[\mathbf{G}\mathbf{G}^\top\right]\|_{\mathrm{op}} + \|\mathbf{G}\boldsymbol{\mu}_c\mathbf{1}^\top\|_{\mathrm{op}} + \|\mathbf{G}\boldsymbol{\mu}_s E_1^\top\|_{\mathrm{op}} \\
&\le \sqrt{\frac{N}{d}}\left(2\frac{\sqrt{N}+t}{\sqrt{N}} + \frac{(\sqrt{N}+t)^2}{\sqrt{Nd}} + t\sqrt{N}(\|\boldsymbol{\mu}_c\| + \|\boldsymbol{\mu}_s\|)\right) \\
&\le \frac{\sqrt{N}+t}{\sqrt{d}}\left(2 + \frac{\sqrt{N}+t}{\sqrt{d}} + \frac{t}{\sqrt{N}+t}\sqrt{N}(\|\boldsymbol{\mu}_c\| + \|\boldsymbol{\mu}_s\|)\right) \\
&\le \frac{\sqrt{N}+t}{\sqrt{d}} \cdot 2.5,
\end{aligned}$$

where the last transition follows from substituting Equation (7). To obtain the spectral bound on $\mathbf{Z}\mathbf{Z}^\top$ we have that $\mathbf{Z} = \mathbf{G} + \mathbf{1}\boldsymbol{\mu}_c^\top + E_1\boldsymbol{\mu}_s^\top$. From Weyl's inequality for singular values:

$$|s_{\min}(\mathbf{G}^\top + \boldsymbol{\mu}_c\mathbf{1}^\top + \boldsymbol{\mu}_s E_1^\top) - s_{\min}(\mathbf{G}^\top)| \le s_{\max}(\boldsymbol{\mu}_c\mathbf{1}^\top + \boldsymbol{\mu}_s E_1^\top) \le \|\boldsymbol{\mu}_c\|\sqrt{N} + \|\boldsymbol{\mu}_s\|\sqrt{N_1}.$$

Taken together with Equation (4) and the assumption in Equation (7) we get:

$$\begin{aligned}
s_{\min}(\mathbf{Z}^\top) &\ge s_{\min}(\mathbf{G}^\top) - \|\boldsymbol{\mu}_c\|\sqrt{N} - \|\boldsymbol{\mu}_s\|\sqrt{N_1} \\
&\ge 1 - \frac{1}{\sqrt{d}}\left(\sqrt{N}+t\right) - \|\boldsymbol{\mu}_c\|\sqrt{N} - \|\boldsymbol{\mu}_s\|\sqrt{N_1} \\
&\ge \frac{1}{2}.
\end{aligned}$$

To prove that $\mathbf{Z}\mathbf{Z}^\top \preceq 2$ we simply need to follow the same steps while taking notice that Weyl's inequality also holds for $s_{\max}(\mathbf{G}^\top)$. This will give us $s_{\max}(\mathbf{Z}^\top) \le 3/2 \le 2$ from which the upper bound follows. □

C.2 Sufficiency of Linear Classifiers Spanned by Data Points

Note that $\mathbf{w}$ is fixed given $\{\mathbf{x}_i\}_{i\in S}$ since we assume it is the output of a deterministic learning algorithm. Now we wish to bound $\langle \hat{\mathbf{w}}_\perp, \boldsymbol{\mu}_c\rangle = r_c\langle \hat{\mathbf{w}}_\perp, U_1\rangle$. To this end let us take an orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ and let these vectors form the columns of the orthogonal matrix $V \in \mathbb{R}^{d\times N}$. Let P_V be the orthogonal projection matrix on the columns of V. We first claim that conditioned on the data, the component of the mean vectors that is not spanned by the data is distributed uniformly.

Lemma 3. *Let $\boldsymbol{\mu}_c^\perp := (I - P_V)\boldsymbol{\mu}_c$ and $\boldsymbol{\mu}_s^\perp := (I - P_V)\boldsymbol{\mu}_c$. Conditional on the training set $\{\mathbf{x}_i, y_i\}_{i\in S}$, the vectors $\frac{\boldsymbol{\mu}_s^\perp}{\|\boldsymbol{\mu}_s^\perp\|}$ and $\frac{\boldsymbol{\mu}_c^\perp}{\|\boldsymbol{\mu}_c^\perp\|}$ are uniformly distributed on unit spheres a subspace of dimension $d - N$.*

Proof. Recalling the notation $\mathbf{z}_i = y_i\mathbf{x}_i$, note that $\{\mathbf{z}_i\}_{i\in S}$ are sufficient statistics for $\boldsymbol{\mu}_s, \boldsymbol{\mu}_c$ given the training data, i.e., $\mathbb{P}(\boldsymbol{\mu}_s, \boldsymbol{\mu}_c \mid \{\mathbf{z}_i\}_{i\in S}) = \mathbb{P}(\boldsymbol{\mu}_s, \boldsymbol{\mu}_c \mid \{\mathbf{x}_i, y_i\}_{i\in S})$. Furthermore, since the joint distribution of $\boldsymbol{\mu}_s, \boldsymbol{\mu}_c, \{\mathbf{z}_i\}_{i\in S}$ is rotationally invariant, we have

$$\mathbb{P}(\boldsymbol{\mu}_s, \boldsymbol{\mu}_c \mid \{\mathbf{z}_i\}_{i\in S}) = \mathbb{P}(\mathbf{R}\boldsymbol{\mu}_s, \mathbf{R}\boldsymbol{\mu}_c \mid \{\mathbf{R}\mathbf{z}_i\}_{i\in S})$$

for any orthogonal matrix $\mathbf{R} \in \mathbb{R}^{d\times d}$. Focusing on matrices $\mathbf{R}$ that presereve that data, i.e., satisfying $\mathbf{R}\mathbf{z}_i = \mathbf{z}_i$ for all $i \in [N]$, we have

$$\mathbb{P}(\boldsymbol{\mu}_s, \boldsymbol{\mu}_c \mid \{\mathbf{z}_i\}_{i\in S}) = \mathbb{P}(\mathbf{R}\boldsymbol{\mu}_s, \mathbf{R}\boldsymbol{\mu}_c \mid \{\mathbf{z}_i\}_{i\in S}).$$

We may also write this equality as

$$\begin{aligned}&\mathbb{P}(P_V\boldsymbol{\mu}_s, P_V\boldsymbol{\mu}_c, (I-P_V)\boldsymbol{\mu}_s, (I-P_V)\boldsymbol{\mu}_c \mid \{\mathbf{z}_i\}_{i\in S})\\ &\quad= \mathbb{P}(P_V\mathbf{R}\boldsymbol{\mu}_s, P_V\mathbf{R}\boldsymbol{\mu}_c, (I-P_V)\mathbf{R}\boldsymbol{\mu}_s, (I-P_V)\mathbf{R}\boldsymbol{\mu}_c \mid \{\mathbf{z}_i\}_{i\in S}).\end{aligned}$$

The fact that R preserves $\{\mathbf{z}_i\}_{i\in S}$ implies that $P_V\mathbf{R} = P_V = \mathbf{R}P_V$ and therefore

$$\mathbb{P}(P_V\boldsymbol{\mu}_s, P_V\boldsymbol{\mu}_c, \boldsymbol{\mu}_s^\perp, \boldsymbol{\mu}_c^\perp \mid \{\mathbf{z}_i\}_{i\in S}) = \mathbb{P}(P_V\boldsymbol{\mu}_s, P_V\boldsymbol{\mu}_c, \mathbf{R}\boldsymbol{\mu}_s^\perp, \mathbf{R}\boldsymbol{\mu}_c^\perp \mid \{\mathbf{z}_i\}_{i\in S}).$$

Marginalizing $P_V\boldsymbol{\mu}_s, P_V\boldsymbol{\mu}_c$, we obtain that, conditional on the training data, the distribution of $\boldsymbol{\mu}_s^\perp, \boldsymbol{\mu}_c^\perp$, is invariant to rotations that preserve the training data. Therefore, the unit vectors in the directions of $\boldsymbol{\mu}_s^\perp$ and $\boldsymbol{\mu}_c^\perp$ must each be uniformly distributed on the sphere orthogonal to the training data, which has dimension $d - N$. □

Now we simply need to derive a bound on $\langle \mathbf{w}_\perp, \boldsymbol{\mu}_s\rangle$:

Corollary 1. *For any $t > 0$ as in Lemma 1, with with probability at least $1 - 10\exp(-t^2/2)$, all the events in Lemma 1 hold and additionally*

$$|\langle \mathbf{w}_\perp, \boldsymbol{\mu}_s\rangle| < \frac{\|\boldsymbol{\mu}_s\|}{\sqrt{d-N}}t \;\text{ and }\; |\langle \mathbf{w}_\perp, \boldsymbol{\mu}_c\rangle| < \frac{\|\boldsymbol{\mu}_c\|}{\sqrt{d-N}}t. \tag{8}$$

Proof. Note that

$$|\langle \mathbf{w}_\perp, \boldsymbol{\mu}_s\rangle| = \left|\langle \mathbf{w}_\perp, \boldsymbol{\mu}_s^\perp\rangle\right| = \|\boldsymbol{\mu}_s^\perp\|\|\mathbf{w}_\perp\| \left|\left\langle \frac{\mathbf{w}_\perp}{\|\mathbf{w}_\perp\|}, \frac{\boldsymbol{\mu}_s^\perp}{\|\boldsymbol{\mu}_s^\perp\|}\right\rangle\right| \le \|\boldsymbol{\mu}_s\| \left|\left\langle \frac{\mathbf{w}_\perp}{\|\mathbf{w}_\perp\|}, \frac{\boldsymbol{\mu}_s^\perp}{\|\boldsymbol{\mu}_s^\perp\|}\right\rangle\right|.$$

Conditional on the training data and the algorithm's randomness, $\frac{\mathbf{w}_\perp}{\|\mathbf{w}_\perp\|}$ is a fixed unit vector in the subspace orthogonal to the training data (of dimension $d - N$), while $\frac{\boldsymbol{\mu}_s^\perp}{\|\boldsymbol{\mu}_s^\perp\|}$ is a spherically uniform unit vector in that subspace. Therefore, standard concentration bounds [4, Lemma 2.2] imply that, for any $t_2 > 0$

$$\mathbb{P}\left(\left|\left\langle \frac{\mathbf{w}_\perp}{\|\mathbf{w}_\perp\|}, \frac{\boldsymbol{\mu}_s^\perp}{\|\boldsymbol{\mu}_s^\perp\|}\right\rangle\right| \ge t_2\right) \le 2\exp(-(d-N)t_2^2/2).$$

The claimed result follows by taking $t_2 = t/\sqrt{d-N}$, applying the same argument for $\boldsymbol{\mu}_c$, taking a union bound. □

D Proofs of Main Result

In this section, we provide the proof of Proposition 1, our main theoretical finding highlighting a fundamental limitation to the robustness of any interpolating classifier. Following the notation of Appendix C, we write a general unit-vector classifier as $\hat{\mathbf{w}} = \sum_{i\in S}\beta_i\mathbf{z}_i + \mathbf{w}_\perp$, where $\mathbf{z}_i = y_i\mathbf{x}_i$. As explained in the proof sketch at Section 3, in order to show a lower bound on robust accuracy, we show a lower bound on the spurious-to-core ratio $\frac{\langle \mathbf{w}, \boldsymbol{\mu}_s\rangle}{\langle \mathbf{w}, \boldsymbol{\mu}_c\rangle}$ or equivalently upper bound $\frac{\langle \mathbf{w}, \boldsymbol{\mu}_s\rangle}{\langle \mathbf{w}, \boldsymbol{\mu}_c\rangle}$, which we can write as

$$\frac{\langle \mathbf{w}, \boldsymbol{\mu}_c\rangle}{\langle \mathbf{w}, \boldsymbol{\mu}_s\rangle} = \frac{\langle \hat{\mathbf{w}}, \boldsymbol{\mu}_c\rangle}{\langle \hat{\mathbf{w}}, \boldsymbol{\mu}_s\rangle} = \frac{\|\boldsymbol{\mu}_c\|^2}{\|\boldsymbol{\mu}_s\|^2} \cdot \frac{\mathbf{1}^\top\beta + \frac{1}{\|\boldsymbol{\mu}_c\|^2}\left[\sum_{i\in S}\beta_i\langle n_i, \boldsymbol{\mu}_c\rangle + \langle \mathbf{w}_\perp, \boldsymbol{\mu}_c\rangle\right]}{E_1^\top\beta + \frac{1}{\|\boldsymbol{\mu}_s\|^2}\left[\sum_{i\in S}\beta_i\langle n_i, \boldsymbol{\mu}_s\rangle + \langle \mathbf{w}_\perp, \boldsymbol{\mu}_s\rangle\right]}. \tag{9}$$

We develop the lower bound - and prove Proposition 1 - in three steps, each corresponsding to a subsection below. First, we give a lower bound on $E_1^\top\beta$ using Lagrange duality (Lemma 4). Second, in Lemma 5, we bound the residual terms of the form $\frac{1}{\|\boldsymbol{\mu}\|^2}\left|\sum_{i\in S}\beta_i\langle n_i, \boldsymbol{\mu}\rangle + \langle \mathbf{w}_\perp, \boldsymbol{\mu}\rangle\right|$ (for $\boldsymbol{\mu} \in \{\boldsymbol{\mu}_c, \boldsymbol{\mu}_s\}$) using concentration of measure arguments from Appendix C. Finally, we combine these two results with the conditions of Proposition 1 to conclude its proof.

D.1 Lower bounding $E_1^\top \beta$

The crux of our proof is showing that the term $E_1^\top \beta$, i.e., the sum of the contributions of elements from the first environment to $\mathbf{w}$, must grow roughly as $N_1\gamma$ for any interpolating classifier. This will in turn imply a large spurious component in the classifier via manipulation of Equation (9).

Lemma 4. *Conditional on the events in Corollary 1 (with parameter $t > 0$), if Equation (7) holds and $\mathbf{w}$ has normalized margin at least γ, we have that*

$$E_1^\top \beta \geq \frac{1}{2}\left(N_1\gamma - \sqrt{2N_2} N_1 \|\boldsymbol{\mu}_c\|^2 - \sqrt{18N_1} \cdot \frac{\sqrt{N}+t}{\sqrt{d}} \right). \tag{10}$$

Proof of Lemma 4. Our strategy for bounding $E_1^\top \beta$ begins with writing down the smallest value it can reach for any unit-norm classifier $\hat{\mathbf{w}}$ with normalized margin at least γ. Recalling that $\hat{\mathbf{w}} = \mathbf{Z}^\top \beta + \mathbf{w}_\perp$ (for $\mathbf{w}_\perp$ such that $\mathbf{Z}\mathbf{w}_\perp = 0$), the smallest possible value of $E_1^\top \beta$ is the solution to the following optimization problem:

$$\begin{aligned} \min_{\beta \in \mathbb{R}^N, \mathbf{w}_\perp \in \ker(\mathbf{Z})} \; & E_1^\top \beta \\ \text{subject to } & \langle \mathbf{Z}^\top \beta + \mathbf{w}_\perp, y_i \mathbf{x}_i \rangle \geq \gamma \;\; \forall i \in [N] \\ & \|\mathbf{Z}^\top \beta + \mathbf{w}_\perp\| = 1. \end{aligned} \tag{11}$$

Since $\mathbf{z}_i = y_i \mathbf{x}_i$ and $\mathbf{Z} w_\perp = 0$, the first constraint is equivalent to the vector inequality $ZZ^\top \beta \geq \gamma \mathbf{1}$, and the second constraint is equivalent to $\beta^\top \mathbf{Z}\mathbf{Z}^\top \beta = 1 - \|\mathbf{w}_\perp\|^2$. Relaxing the second constraint, the smallest value of $E_1^\top \beta$ is bounded from below by the solution to:

$$\begin{aligned} \min_{\beta \in \mathbb{R}^N} \; & \beta^\top E_1 \\ \text{subject to } & \mathbf{Z}\mathbf{Z}^\top \beta \geq \gamma \mathbf{1} \\ & \beta^\top \mathbf{Z}\mathbf{Z}^\top \beta \leq 1. \end{aligned}$$

Take Lagrange multipliers $\lambda \in \mathbb{R}_+^N$ and $\nu \geq 0$, from strong duality the above equals:

$$\max_{\lambda \in \mathbb{R}_+^N, \nu \geq 0} \min_{\beta \in \mathbb{R}^N} \beta^\top E_1 + \lambda^\top (\mathbf{1}\gamma - \mathbf{Z}\mathbf{Z}^\top \beta) + \frac{1}{2}\nu(\beta^\top \mathbf{Z}\mathbf{Z}^\top \beta - 1)$$

Optimizing the quadratic form over β, the above becomes:

$$\max_{\lambda \in \mathbb{R}_+^N, \nu \geq 0} \lambda^\top \mathbf{1}\gamma - \frac{1}{2}\nu - \frac{1}{2}\left(E_1 - \mathbf{Z}\mathbf{Z}^\top \lambda\right)^\top \left(\nu \mathbf{Z}\mathbf{Z}^\top\right)^{-1} \left(E_1 - \mathbf{Z}\mathbf{Z}^\top \lambda\right)$$

Maximizing over ν this becomes:

$$\max_{\lambda \in \mathbb{R}_+^N} \lambda^\top \mathbf{1}\gamma - \sqrt{\left(E_1 - \mathbf{Z}\mathbf{Z}^\top \lambda\right)^\top (\mathbf{Z}\mathbf{Z}^\top)^{-1} (E_1 - \mathbf{Z}\mathbf{Z}^\top \lambda)} := \max_{\lambda \in \mathbb{R}_+^N} \mathcal{L}(\lambda)$$

Thus, $E_1^\top \beta$ is lower bounded by $\mathcal{L}(\lambda)$, for any $\lambda \in \mathbb{R}_+^N$. Taking $\lambda = \alpha E_1$ for $\alpha = \left(1 + \left(\|\boldsymbol{\mu}_c\|^2 + \|\boldsymbol{\mu}_s\|^2\right) N_1\right)^{-1}$, we obtain:

$$\begin{aligned} \mathcal{L}(\lambda) &= N_1\gamma\alpha - \sqrt{E_1^\top \left(\mathbf{I}_N - \alpha \mathbf{Z}\mathbf{Z}^\top\right) (\mathbf{Z}\mathbf{Z}^\top)^{-1} \left(\mathbf{I}_N - \alpha \mathbf{Z}\mathbf{Z}^\top\right) E_1} \\ &\geq N_1\gamma\alpha - \sqrt{2}\| \left(\mathbf{I}_N - \alpha \mathbf{Z}\mathbf{Z}^\top\right) E_1\| \\ &= N_1\gamma\alpha - \sqrt{2}\| \left(\mathbf{I}_N - \alpha \left(\mathbb{E}\left[\mathbf{Z}\mathbf{Z}^\top\right] + \mathbf{Z}\mathbf{Z}^\top - \mathbb{E}\left[\mathbf{Z}\mathbf{Z}^\top\right]\right)\right) E_1\| \\ &\geq N_1\gamma\alpha - \sqrt{2}\| \left(\mathbf{I}_N - \alpha \mathbb{E}\left[\mathbf{Z}\mathbf{Z}^\top\right]\right) E_1\| - \sqrt{2}\|\alpha \left(\mathbf{Z}\mathbf{Z}^\top - \mathbb{E}\left[\mathbf{Z}\mathbf{Z}^\top\right]\right) E_1\| \end{aligned}$$

Here, the first inequality is from our assumption that Equation (7) holds and hence $\mathbf{Z}\mathbf{Z}^\top \succeq \frac{1}{2}\mathbf{I}_N$ and the second is a triangle inequality. Recall the bound $\|\mathbf{Z}\mathbf{Z}^\top - \mathbb{E}\left[\mathbf{Z}\mathbf{Z}^\top\right]\|_{\text{op}} \leq 3\frac{\sqrt{N}+t}{\sqrt{d}}$ from Lemma 2 and apply it to obtain:

$$\mathcal{L}(\lambda) \geq N_1\gamma\alpha - \sqrt{2}\|\left(\mathbf{I}_N - \alpha\mathbb{E}\left[\mathbf{Z}\mathbf{Z}^\top\right]\right)E_1\| - \alpha - \sqrt{18N_1}\cdot\frac{\sqrt{N}+t}{\sqrt{d}}.$$

Let us calculate the second term in the bound above:

$$\begin{aligned}
\|\left(\mathbf{I}_N - \alpha\mathbb{E}\left[\mathbf{Z}\mathbf{Z}^\top\right]\right)E_1\| &= \|\left(1-\alpha-\alpha N_1\|\boldsymbol{\mu}_s\|^2\right)E_1 - \alpha N_1\|\boldsymbol{\mu}_c\|^2\mathbf{1}\| \\
&= \|\left(1-\alpha-\alpha N_1\|\boldsymbol{\mu}_s\|^2\right)E_1 - \alpha N_1\|\boldsymbol{\mu}_c\|^2\left(E_1+E_2\right)\| \\
&= \sqrt{\left(1-\alpha\left(1+N_1(\|\boldsymbol{\mu}_s\|^2+\|\boldsymbol{\mu}_c\|^2)\right)\right)^2 N_1 + \alpha^2N_1^2\|\boldsymbol{\mu}_c\|^4N_2} \\
&= \alpha N_1\|\boldsymbol{\mu}_c\|^2\sqrt{N_2},
\end{aligned}$$

where the final equality used $\alpha\left(1+N_1(\|\boldsymbol{\mu}_s\|^2+\|\boldsymbol{\mu}_c\|^2\right) = 1$. Overall, we get:

$$\beta^\top E_1 \geq \mathcal{L}(\lambda) \geq \alpha\left(N_1\gamma - \sqrt{2N_2}N_1\|\boldsymbol{\mu}_c\|^2 - \sqrt{18N_1}\cdot\frac{\sqrt{N}+t}{\sqrt{d}}\right).$$

The proof is complete by noting that $\alpha \geq 1/2$ due to Equation (7), ☐

D.2 Controlling residual terms

We now provide a bound on the terms in Equation (9) associated with quantities that vanish a the problem dimension grows.

Lemma 5. *Conditioned on all the events in Corollary 1 with parameter $t > 0$ (which happen with probability at least $1 - 10\exp(-t^2/2)$) and the additional condition of Lemma 2, we have for $\boldsymbol{\mu} \in \{\boldsymbol{\mu}_c, \boldsymbol{\mu}_s\}$:*

$$\frac{1}{\|\boldsymbol{\mu}\|^2}\left|\sum_{i\in S}\beta_i\langle n_i, \boldsymbol{\mu}\rangle + \langle\mathbf{w}_\perp, \boldsymbol{\mu}\rangle\right| \leq \frac{3t}{\|\boldsymbol{\mu}\|}\sqrt{\frac{N}{d-N}} \tag{12}$$

Proof. We prove the claim for $\boldsymbol{\mu}_s$; the proof for $\boldsymbol{\mu}_c$ is analogous. Recall the random matrix $\mathbf{G} = \mathbf{Z} - \mathbf{1}\boldsymbol{\mu}_c^\top - E_1\boldsymbol{\mu}_s^\top \in \mathbb{R}^{N\times d}$ from Lemma 1. From Equation (6) we get that $\|\mathbf{G}\boldsymbol{\mu}_s\| \leq t\sqrt{\frac{N}{d}}\|\boldsymbol{\mu}_s\|$ and then:

$$\sum_{i\in S}\beta_i\langle n_i, \boldsymbol{\mu}_s\rangle = \beta^\top\mathbf{G}\boldsymbol{\mu}_s \leq \|\beta\|\|\mathbf{G}\boldsymbol{\mu}_s\| \leq t\|\beta\|\sqrt{\frac{N}{d}}\|\boldsymbol{\mu}_s\|.$$

To eliminate $\|\beta\|$ from this bound, we use $\mathbf{Z}\mathbf{Z}^\top \preceq \frac{1}{2}I_N$ due to Lemma 2 to write

$$\frac{1}{\sqrt{2}}\|\beta\| \leq \sqrt{\beta^\top\mathbf{Z}\mathbf{Z}^\top\beta} \leq \sqrt{\beta^\top\mathbf{Z}^\top\mathbf{Z}\beta + \|\mathbf{w}_\perp\|^2} = \|\hat{\mathbf{w}}\| = 1.$$

Finally, we use Equation (8) from Corollary 1 to bound $|\langle\mathbf{w}_\perp, \boldsymbol{\mu}\rangle|$. ☐

D.3 Proof of Proposition 1

Proof of Proposition 1. Let $t\sqrt{10\log\frac{10}{\delta}} \geq \sqrt{2\log\frac{10}{\delta}}$, so that the events described in the previous lemmas and corollaries all hold with probability at least $1-\delta$. Note that for $c_r \leq 1/64$ we have

$$\sqrt{N}(\|\boldsymbol{\mu}_c\| + \|\boldsymbol{\mu}_s\|) \leq \frac{1}{4} \tag{13}$$

and (since $\gamma \leq \frac{1}{4\sqrt{N}}$)

$$d \geq \frac{C_d}{10}\frac{1}{\gamma^2}\frac{Nt^2}{N_1^2\|\boldsymbol{\mu}_c\|^2} \geq \frac{C_d}{10c_r}\frac{Nt^2}{N_1\gamma^2} \geq \frac{16C_d}{10c_r}\frac{N^2t^2}{N_1}N \geq \frac{6}{4}C_dNt^2.$$

Consequently, for $C_d \geq 1$

$$\frac{\sqrt{N}+t}{\sqrt{d}} \leq 2\sqrt{\frac{1}{64C_d}} \leq \frac{1}{4}. \tag{14}$$

Combining Equations (13) and (14), we see that the condition in Equation (7) holds.

Therefore, we may apply Lemma 4; we now argue that the assumptions of Proposition 1 the lower bound on $E_1^\top \beta$ simplifies to a constant multiple of $N_1\gamma$. First, taking $c_n \leq 1/8$ and $C_r \geq 1$, we have

$$\sqrt{2N_2}N_1\|\boldsymbol{\mu}_c\|^2 \leq \frac{\sqrt{2N_2}N_1\|\boldsymbol{\mu}_s\|^2}{C_r\left(1+\frac{\sqrt{N_2}}{N_1\gamma}\right)} \leq N_1\gamma\frac{\sqrt{2}N_1\|\boldsymbol{\mu}_s\|^2}{C_r} \leq N_1\gamma\frac{\sqrt{2}c_n}{C_r} \leq \frac{1}{4}N_1\gamma.$$

Second, using again $c_r \leq 1/64$ and taking $C_d \geq 180$,

$$\sqrt{18N_1}\frac{\sqrt{N}+t}{\sqrt{d}} \leq N_1\gamma\frac{\sqrt{18}}{\sqrt{C_d/10}}\frac{\sqrt{N}+t}{t\sqrt{N}}\sqrt{N_1}\|\boldsymbol{\mu}_c\| \leq \frac{1}{4}N_1\gamma.$$

Substituting into Equation (10), we conclude that under our assumptions $E_1^\top\beta \geq \frac{1}{4}N_1\gamma$.

Next, we combine the lower bound on $E_1^\top\beta$ with Lemma 5 to handle the denominator and numerator in the RHS of Equation (9). Beginning with the numerator, we have

$$\mathbf{1}^\top\beta + \frac{1}{\|\boldsymbol{\mu}_c\|^2}\left[\sum_{i\in S}\beta_i\langle n_i, \boldsymbol{\mu}_c\rangle + \langle \mathbf{w}_\perp, \boldsymbol{\mu}_c\rangle\right] \leq E_1^\top\beta + \|E_2\|\|\beta\| + \frac{3t}{\|\boldsymbol{\mu}_c\|}\sqrt{\frac{N}{d-N}}.$$

As argued in the proof pf Lemma 5, we have $\|\beta\| \leq \sqrt{2}$ and therefore $\|E_2\|\|\beta\| \leq \sqrt{2N_2}$. Substituting again our assumptions d (which imply $d > 2N$), using and taking $C_d \geq 64 \cdot 180$, we have

$$\frac{3t}{\|\boldsymbol{\mu}_c\|}\sqrt{\frac{N}{d-N}} \leq \frac{\sqrt{18}t}{\|\boldsymbol{\mu}_c\|}\sqrt{d} \leq N_1\gamma\sqrt{\frac{180}{C_d}} \leq \frac{1}{8}N_1\gamma.$$

For the denominator, noting $\|\boldsymbol{\mu}_c\| \leq \|\boldsymbol{\mu}_s\|$ by our assumption, we may similarly write

$$E_1^\top\beta + \frac{1}{\|\boldsymbol{\mu}_s\|^2}\left[\sum_{i\in S}\beta_i\langle n_i, \boldsymbol{\mu}_s\rangle + \langle \mathbf{w}_\perp, \boldsymbol{\mu}_s\rangle\right] \geq E_1^\top\beta - \frac{1}{8}N_1\gamma.$$

Consequently (since $E_1^\top\beta \geq \frac{1}{4}N_1\gamma$), we have that the denominator is nonnegative. (If the numerator is not positive, $\mathbf{w}$ will have error greater than $1/2$ for $\theta = 0$). Substituting back to Equation (9) and using the lower bound $E_1^\top\beta \geq \frac{1}{4}N_1\gamma$, we get

$$\frac{\langle \mathbf{w}, \boldsymbol{\mu}_c\rangle}{\langle \mathbf{w}, \boldsymbol{\mu}_s\rangle}\frac{\|\boldsymbol{\mu}_s\|^2}{\|\boldsymbol{\mu}_c\|^2} \leq \frac{E_1^\top\beta + \sqrt{2N_2} + \frac{1}{8}N_1\gamma}{E_1^\top\beta - \frac{1}{8}N_1\gamma} \leq \frac{\frac{1}{4}N_1\gamma + \sqrt{2N_2} + \frac{1}{8}N_1\gamma}{\frac{1}{4}N_1\gamma - \frac{1}{8}N_1\gamma} \leq 3 + \frac{\sqrt{128N_2}}{N_1\gamma}.$$

Therefore, for $C_r \geq 16$ we have $\frac{\langle \mathbf{w}, \boldsymbol{\mu}_s\rangle}{\langle \mathbf{w}, \boldsymbol{\mu}_c\rangle} \geq 1$ as required. Since the error of classifier $\mathbf{w}$ in environment with parameter θ is

$$Q\left(\frac{\langle \mathbf{w}, \boldsymbol{\mu}_c\rangle}{\sigma\|\mathbf{w}\|}\left(1 + \theta\frac{\langle \mathbf{w}, \boldsymbol{\mu}_s\rangle}{\langle \mathbf{w}, \boldsymbol{\mu}_c\rangle}\right)\right),$$

(where $Q(t) := \mathbb{P}(\mathcal{N}(0;1) > t)$ is the Gaussian tail function), the fact that $\frac{\langle \mathbf{w}, \boldsymbol{\mu}_s\rangle}{\langle \mathbf{w}, \boldsymbol{\mu}_c\rangle} \geq 1$ implies that there exists $\theta \in [-1, 1]$ for which the error is $Q(0) = 0.5$, implying the stated bound on the robust error. □

E Lower Bound On the Achievable Margin

We now argue that, in our model, a simple signed-sample-mean estimator interpolates the data with normalized margin scaling as $1/\sqrt{N}$. This fact establishes the first part of Theorem 1.

Proposition 3. *There exist universal constants $c_n', C_d' > 0$ such that, in the DGP with parameters $N_1, N_2, d > 0$, $\boldsymbol{\mu}_c, \boldsymbol{\mu}_s \in \mathbb{R}^d$, $\theta_1 = 1$, $\theta_2 = 0$ and $\sigma^2 = 1/d$, for any $\delta \in (0, 1/2)$ if*

$$\max\{\|\boldsymbol{\mu}_c\|, \|\boldsymbol{\mu}_s\|\} \le \frac{c_n'}{N} \text{ and } d \ge C_d' N^2 \log\left(\frac{1}{\delta}\right)$$

then with probability at least $1 - \delta$, the signed-sample-mean estimator $\mathbf{w}_{\text{mean}} = \frac{1}{N}\sum_{i=1}^{N} y_i x_i$ obtains normalized margin of at least $\frac{1}{\sqrt{8N}}$.

Proof. Using the notation defined in the beginning of Appendix C, we note that $\mathbf{w}_{\text{mean}} = \frac{1}{N}\mathbf{Z}^\top \mathbf{1}$ and (for $\sigma^2 d = 1$) its normalized margin is

$$\min_{i \in [N]} \frac{y_i \langle \mathbf{x}_i, \mathbf{w}_{\text{mean}} \rangle}{\|\mathbf{w}_{\text{mean}}\|} = \min_{i \in [N]} \frac{[\mathbf{Z}\mathbf{w}_{\text{mean}}]_i}{\|\mathbf{w}_{\text{mean}}\|} = \min_{i \in [N]} \frac{[\mathbf{Z}\mathbf{Z}^\top \mathbf{1}]_i}{\|\mathbf{Z}^\top \mathbf{1}\|}.$$

Substituting the assumed bounds on d and $\|\boldsymbol{\mu}_c\|, \|\boldsymbol{\mu}_s\|$ into Lemma 2 (with $t = \sqrt{8 \log \frac{1}{\delta}} \ge \sqrt{2 \log \frac{6}{\delta}}$), it is easy to verify that for sufficiently small c_n' and sufficiently large C_d', the condition in Equation (7) holds, and therefore

$$\|\mathbf{Z}\mathbf{Z}^\top - \mathbb{E}\mathbf{Z}\mathbf{Z}^\top\|_{\text{op}} \le 3\frac{\sqrt{N} + t}{\sqrt{d}} \le \frac{1}{\sqrt{4N}},$$

with the final inequality following by choosing C_d' sufficiently large. Lemma 2 then also implies that $\mathbf{Z}\mathbf{Z}^\top \preceq 2I_N$.

Noting that $\mathbb{E}\mathbf{Z}\mathbf{Z}^\top = I_N + \|\mu_c\|^2 \mathbf{1}\mathbf{1}^\top + \|\mu_s\|^2 E_1 E_1^\top$, we have that, for all $i \in [N]$,

$$[\mathbf{Z}\mathbf{Z}^\top \mathbf{1}]_i \ge [\mathbb{E}\mathbf{Z}\mathbf{Z}^\top \mathbf{1}]_i - \|\mathbf{Z}\mathbf{Z}^\top - \mathbb{E}\mathbf{Z}\mathbf{Z}^\top\|_{\text{op}} \|\mathbf{1}\| \ge 1 - \frac{1}{\sqrt{4N}}\|\mathbf{1}\| = \frac{1}{2}.$$

Moreover, $\mathbf{Z}\mathbf{Z}^\top \preceq 2I_N$ implies that

$$\|\mathbf{Z}^\top \mathbf{1}\| = \sqrt{\mathbf{1}^\top \mathbf{Z}\mathbf{Z}^\top \mathbf{1}} \le 2\|\mathbf{1}\| = 2\sqrt{N}.$$

Combining the above two displays yields the claimed margin bound. □

F Two-Stage Algorithm and its Analysis

In this section we give the pseudocode for the algorithm that provably learns an invariant model in our setting (see Algorithm 1) and analyze its performance. For generality, we denote the empirical invariance constraint by membership in some family $\mathcal{F}(S^{\text{fine}})$, though our analysis will concentrate on Equalized Opportunity as described in the next section.

Algorithm 1 Two Phase Learning of Overparameterized Invariant Classifiers

Input: Dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ and a partition S_1, S_2 into environments. Invariance constraint function family $\mathcal{F}(\cdot)$

Output: A classifier $f_{\mathbf{v}}(\mathbf{x})$

Draw subsets of data $S_{\text{trn}} = \cup_{e \in \{1,2\}} S_e^{\text{trn}}$, where $S_e^{\text{trn}} \subset S_e$ for $e \in \{1, 2\}$ and $|S_e^{\text{trn}}| = N_e/2$

Stage 1: Calculate $\mathbf{w}_e = N_e^{-1} \sum_{i \in S_e^{\text{trn}}} \mathbf{x}_i y_i$ for each $e \in \{1, 2\}$

Define $S^{\text{fine}} = S \setminus S^{\text{trn}}$

Stage 2: Return the solution $f_{\mathbf{v}}(\mathbf{x}; S_{\text{trn}}) = \langle v_1 \cdot \mathbf{w}_1 + v_2 \cdot \mathbf{w}_2, \mathbf{x} \rangle$ that solves

$$\text{maximize} \sum_{i \in S^{\text{fine}}} f_{\mathbf{v}}(\mathbf{x}_i) y_i \quad \text{subject to} \quad \|\mathbf{v}\|_\infty = 1 \quad \text{and} \quad f_{\mathbf{v}} \in \mathcal{F}(S^{\text{fine}}) \tag{15}$$

F.1 Analysis of Algorithm 1

The proof that Algorithm 1 indeed achieves a non-trivial robust error will require some definitions and more mild assumptions which we now turn to describe.

Definitions. Denote the first-stage training set indices by S, where $|S| = N$ and second stage "fine-tuning" set by $|D| = M$. Let us denote:

$$\bar{\mathbf{n}}_e = \frac{1}{N_e}\sum_{i\in S_e} n_i,\ \bar{\mathbf{m}}_e = \frac{1}{M_e}\sum_{i\in D_e} n_i,\ \bar{\mathbf{m}}_{e,1} = \frac{1}{M_{e,1}}\sum_{i\in D_{e,1}} n_i.$$

Models will be defined by:

$$\begin{aligned}\mathbf{w}_e &:= \frac{1}{N_e}\sum_{i\in S_e} y_i\mathbf{x}_i = \mu_c + \theta_e\mu_s + \bar{\mathbf{n}}_e, \quad e\in\{1,2\},\\ f_{\mathbf{v}}(x;S) &= \langle v_1\cdot\mathbf{w}_1 + v_2\cdot\mathbf{w}_2, \mathbf{x}\rangle.\end{aligned}$$

The Equalized Opportunity (EOpp) constraint is:

$$\begin{aligned}\hat{T}_1(f_{\mathbf{v}};D,S) &= \hat{T}_2(f_{\mathbf{v}};D)\\ \hat{T}_e(f_{\mathbf{v}};D,S) &= \frac{1}{M_{e,1}}\sum_{i\in D_{e,1}} f_{\mathbf{v}}(\mathbf{x}_i)\end{aligned}$$

Additional Assumptions We assume w.l.o.g $\theta_2 > \theta_1$, define $\Delta := \theta_2 - \theta_1 > 0$ and $r_\mu = \frac{\|\mu_s\|}{\|\mu_c\|} > 1$. We consider r_μ, Δ as fixed numbers. That is, they do not depend on N, d and other parameters of the problem. Also define $r := \frac{\Delta\theta_{\max}}{\Delta+4\theta_{\max}}$, where $\theta_{\max} := \mathrm{argmax}\{|\theta_1|, |\theta_2|\} \le 1$. The following additional assumptions will be required for our concentration bounds.

Assumption 1. *Let $t > 0$ be a fixed user specified value, which we define later and will control the success probability of the algorithm. We will assume that for each $e\in\{1,2\}$ and some universal constants $c_c, c_s > 0$:*

$$\begin{aligned}\|\mu_s\|^2 &\ge t\sigma^2 c_s \max\left\{\frac{1}{r^2N_e}, \frac{1}{(r\Delta)^2M_{e,1}}, \frac{\sqrt{d}}{M_{e,1}r\Delta}\right\}\\ \|\mu_c\|^2 &\ge t\sigma^2 c_c \max\left\{\frac{1}{\Delta^2N_e}, \frac{r_\mu^2}{(\Delta^2M_{e,1})}, \frac{r_\mu^2}{\Delta^2M_e}, \frac{\sqrt{d}}{M_{e,1}\Delta^2}, \frac{\sqrt{d}}{M_e\Delta}\right\}\end{aligned}$$

Analyzing the EOpp constraint. Writing the terms defined above in more detailed form gives:

$$\begin{aligned}\epsilon_e(\mathbf{v}) =& \langle\bar{\mathbf{m}}_{e,1}, v_1(\mu_c + \theta_1\mu_s + \bar{\mathbf{n}}_1) + v_2(\mu_c + \theta_2\mu_s + \bar{\mathbf{n}}_2)\rangle\\ \delta_e(\mathbf{v}) =& \langle\bar{\mathbf{m}}_e, v_1(\mu_c + \theta_1\mu_s + \bar{\mathbf{n}}_1) + v_2(\mu_c + \theta_2\mu_s + \bar{\mathbf{n}}_2)\rangle\\ \hat{T}_e(f_{\mathbf{v}};D,S) =& (v_1+v_2)\|\mu_c\|^2 + (v_1\theta_1 + v_2\theta_2)\theta_e\|\mu_s\|^2 +\\ & \langle\mu_c + \theta_e\mu_s, v_1\bar{\mathbf{n}}_1 + v_2\bar{\mathbf{n}}_2\rangle + \epsilon_e(\mathbf{v})\end{aligned}$$

So the EOpp constraint is:

$$\begin{aligned}& v_1\left[\theta_1\|\mu_s\|^2 + \langle\bar{\mathbf{n}}_1,\mu_s\rangle\right]\theta_1 + v_2\left[\theta_2\|\mu_s\|^2 + \langle\bar{\mathbf{n}}_2,\mu_s\rangle\right]\theta_1 + \epsilon_1(\mathbf{v}) =\\ & \quad v_1\left[\theta_1\|\mu_s\|^2 + \langle\bar{\mathbf{n}}_1,\mu_s\rangle\right]\theta_2 + v_2\left[\theta_2\|\mu_s\|^2 + \langle\bar{\mathbf{n}}_2,\mu_s\rangle\right]\theta_2 + \epsilon_2(\mathbf{v}) \end{aligned} \tag{16}$$

Lemma 6. *Consider all the solutions $\mathbf{v} = (v_1, v_2)$ that satisfy EOpp and have $\|\mathbf{v}\|_\infty = 1$. With probability 1 there are exactly two such solutions $\mathbf{v}_{\mathrm{pos}}, \mathbf{v}_{\mathrm{neg}}$, where $\mathbf{v}_{\mathrm{pos}} = -\mathbf{v}_{\mathrm{neg}}$.*

We will consider $\mathbf{v}_{\mathrm{pos}}$ as the solution that satisfies $v_{\mathrm{pos},1} + v_{\mathrm{pos},2} > 0$.

Proof. Is it easy to see that the EOpp constraint is a linear equation in v_1, v_2 and with probability 1 the coefficients in this linear equations are nonzero. Therefore the solutions to this equation form a line in $\mathbb{R}^2$ that passes through the origin. Consequently, this line intersects the l_∞ unit ball at two points, that we denote $\mathbf{v}_{\mathrm{pos}}, \mathbf{v}_{\mathrm{neg}}$, which are negations of one another. ☐

The proposed algorithm. Now we can restate our algorithm in terms of v_{pos} and v_{neg} and analyze its retrieved solution.

- Calculate $\mathbf{w}_1$ and $\mathbf{w}_2$ according to their definitions.
- Consider the solutions $\{\mathbf{v}_{\text{pos}}, \mathbf{v}_{\text{neg}}\}$ that satisfy EOpp and also $\|\mathbf{v}\|_\infty = 1$.
- Return the solution: $\mathbf{v} \in \{\mathbf{v}_{\text{pos}}, \mathbf{v}_{\text{neg}}\}$ which has the higher score, where the score is:

$$\mathbf{v}^* \in \arg \max_{\mathbf{v} \in \{\mathbf{v}_{\text{pos}}, \mathbf{v}_{\text{neg}}\}} \sum_{i \in D} \langle v_1 \mathbf{w}_1 + v_2 \mathbf{w}_2, y_i \mathbf{x}_i \rangle$$

We first analyze the two possible solution v_{pos} and v_{neg} and show that their coordinates cannot be negations of each other. Intuitively, in an ideal scenario with infinite data, the EOpp constraint will enforce $v_1\theta_1 = -v_2\theta_2$. Then $v_1 = -v_2$ is only possible if $\theta_1 = \theta_2$, which we assume is not the case (if it is, we cannot identify the spurious correlation from data). The assumption of a fixed $\Delta > 0$, will let us show that indeed with high probability $v_1 = -v_2$ does not occur.

Lemma 7. *Let $t > 0$ and consider the solutions $v_{\text{neg}}, v_{\text{pos}}$ that the algorithm may return. With probability at least $1 - 34\exp(-t^2/2)$, the solutions satisfy $|v_1 + v_2| \geq \frac{\Delta}{2}$.*

Proof. Assume that for $e \in \{1, 2\}$ the following events occur:

$$|\langle \bar{\mathbf{n}}_e, \mu_s \rangle| \leq r\|\mu_s\|^2 \tag{17}$$

$$|\langle \bar{\mathbf{m}}_{1,1} - \bar{\mathbf{m}}_{2,1}, \mu_c + \theta_e \mu_s + \bar{\mathbf{n}}_e \rangle| \leq r\Delta\|\mu_s\|^2 \tag{18}$$

Corollary 3 will show that they occur with the desired probability in our statement. Let us incorporate these events into the EOpp constraint. We group the items multiplied by v_1 and those multiplied by v_2:

$$\begin{aligned} -v_1^* \left[\theta_1\|\mu_s\|^2\Delta + \langle \bar{\mathbf{n}}_1, \mu_s \rangle\Delta + \langle \bar{\mathbf{m}}_{1,1} - \bar{\mathbf{m}}_{2,1}, \mu_c + \theta_1\mu_s + \bar{\mathbf{n}}_1 \rangle\right] = \\ v_2^* \left[\theta_2\|\mu_s\|^2\Delta + \langle \bar{\mathbf{n}}_2, \mu_s \rangle\Delta + \langle \bar{\mathbf{m}}_{2,1} - \bar{\mathbf{m}}_{1,1}, \mu_c + \theta_2\mu_s + \bar{\mathbf{n}}_2 \rangle\right] \end{aligned}$$

Let us denote for convenience (where we drop the dependence on parameters in the notation):

$$\begin{aligned} a &= \|\mu_s\|^{-2}\Delta \left(\langle \bar{\mathbf{n}}_1, \mu_s \rangle + \Delta^{-1}\langle \bar{\mathbf{m}}_{1,1} - \bar{\mathbf{m}}_{2,1}, \mu_c + \theta_1\mu_s + \bar{\mathbf{n}}_1 \rangle\right) \\ b &= \|\mu_s\|^{-2}\Delta \left(\langle \bar{\mathbf{n}}_2, \mu_s \rangle + \Delta^{-1}\langle \bar{\mathbf{m}}_{2,1} - \bar{\mathbf{m}}_{1,1}, \mu_c + \theta_2\mu_s + \bar{\mathbf{n}}_2 \rangle\right) \end{aligned}$$

Now the EOpp constraint can be written as $-v_1^*\|\mu_s\|^2\Delta\left(\theta_1 + a\right) = v_2^*\|\mu_s\|^2\Delta\left(\theta_2 + b\right)$. Plugging in Equation (17) and Equation (18), we see that $\max\{|a|, |b|\} \leq r$.

Assume that $|\theta_1 + b| \geq |\theta_2 + a|$, and note that since $\|\mathbf{v}^*\|_\infty = 1$ we have that $|v_1^*| = 1$ (the proof for the other case is analogous). [8] We note that by definition $\Delta \leq 2\theta_{\max}$, hence if $v_2^* = 0$ we have $|v_1^* + v_2^*| = 1 \geq \frac{\Delta}{2\theta_{\max}}$ and our claim holds. Otherwise, we can write:

$$\begin{aligned} |v_1^* + v_2^*| &= \left|1 - \frac{\theta_2 + b}{\theta_1 + a}\right| = \left|\frac{\Delta + a - b}{\theta_1 + a}\right| \geq \frac{\Delta - 2r}{\theta_{\max} + r} = \frac{\Delta - 2\frac{\Delta\theta_{\max}}{\Delta + 4\theta_{\max}}}{\theta_{\max} + \frac{\Delta\theta_{\max}}{\Delta + 4\theta_{\max}}} \\ &= \frac{\Delta\left(\Delta + 4\theta_{\max} - 2\theta_{\max}\right)}{\theta_{\max}\left(\Delta + 4\theta_{\max} + \Delta\right)} = \frac{\Delta}{2\theta_{\max}} \geq \frac{\Delta}{2} \end{aligned}$$

□

The result above will be useful for proving the rest of our claims towards the performance guarantees of the algorithm. We first show that the retrieved solution is the one that is positively aligned with μ_c.

Lemma 8. *With probability at least $1 - 34\exp(-t^2/2)$, between the two solutions considered at the second stage of our algorithm, the one with $v_1 + v_2 \geq 0$ achieves a higher score.*

[8] In the case where $|\theta_2 + a| \geq |\theta_1 + b|$ then $|v_2^*| = 1$ would hold.

Proof. Let's write down the score on environment $e \in \{1, 2\}$ in detail:

$$\sum_{i \in D_e} \mathbf{w}^\top \mathbf{x}_i y_i = (v_1 + v_2)\|\mu_c\|^2 + \langle \mu_c, v_1 \bar{\mathbf{n}}_1 + v_2 \bar{\mathbf{n}}_2 \rangle + \\ (v_1\theta_1 + v_2\theta_2)\theta_e \|\mu_s\|^2 + \langle \mu_s, \theta_e (v_1 \bar{\mathbf{n}}_1 + v_2 \bar{\mathbf{n}}_2) \rangle + \\ \langle \bar{\mathbf{m}}_e, (v_1 + v_2)\mu_c + (\theta_1 v_1 + \theta_2 v_2)\mu_s + v_1 \bar{\mathbf{n}}_1 + v_2 \bar{\mathbf{n}}_2 \rangle \tag{19}$$

We will bound all the items other than $(v_1 + v_2)\|\mu_s\|^2$ with concentration inequalities, and for the second line also use the EOpp constraint. Regrouping items in Equation (16) we have:

$$\left| (v_1\theta_1 + v_2\theta_2) \|\mu_s\|^2 + \langle \mu_s, v_1 \bar{\mathbf{n}}_1 + v_2 \bar{\mathbf{n}}_2 \rangle \right| \cdot \Delta = |\epsilon_2(\mathbf{v}) - \epsilon_1(\mathbf{v})|$$

In Corollary 3 we will prove that with probability at least $1 - 34\exp(-t^2/2)$, it holds that $|\epsilon_2(\mathbf{v}) - \epsilon_1(\mathbf{v})| \leq \frac{\Delta}{6}|v_1 + v_2| \cdot \|\mu_c\|^2$. Combined with $|\theta_e| < 1$, we get that the magnitude of the terms in the second line of Equation (19) is bounded by $\frac{1}{6}|v_1 + v_2| \cdot \|\mu_c\|^2$. We will also show in Corollary 3 that the other two terms in Equation (19) besides $(v_1 + v_2)\|\mu_c\|^2$, are bounded by $\frac{1}{6}|v_1 + v_2| \cdot \|\mu_c\|^2$. Hence we have for some b such that $|b| \leq \frac{1}{2}|(v_1 + v_2)| \cdot \|\mu_c\|^2$ that:

$$\sum_{i \in D_e} \mathbf{w}^\top \mathbf{x}_i y_i = (v_1 + v_2)\|\mu_c\|^2 + b$$

We note that the score in the algorithm is a weighted average of the scores over the training environments, yet the derivation above holds regardless of e. That is, θ_e did not play a role in the derivation other than the assumption that its magnitude is smaller than 1. Hence it is clear that the solution $\mathbf{v}^* = \mathbf{v}_{\text{pos}}$ will be chosen over $\mathbf{v}_{\text{neg}}$. □

Once we have characterized our returned solution, it is left to show its guaranteed performance over all environments $\theta \in [-1, 1]$. We can draw a similar argument to Lemma 8 to reason about the expected score obtained in each environment.

Lemma 9. *Let $t > 0$ and consider the retrieved solution $\mathbf{v}^*$. With probability at least $1 - 34\exp(-t^2/2)$, the expected score of $\mathbf{v}^*$ over any environment corresponding to $\theta \in [-1, 1]$ is larger than $\frac{\Delta}{3}\|\mu_c\|^2$.*

Proof. The expected score can be written same as in Equation (19), except we can drop the last item since it has expected value 0. We let $\theta \in [-1, 1]$ and write:

$$\mathbb{E}_{\mathbf{x}, y \sim P_\theta}\left[\mathbf{w}^\top \mathbf{x} y\right] = (v_1^* + v_2^*)\|\mu_c\|^2 + \langle \mu_c, v_1^* \bar{\mathbf{n}}_1 + v_2^* \bar{\mathbf{n}}_2 \rangle + \\ (v_1^*\theta_1 + v_2^*\theta_2)\theta \|\mu_s\|^2 + \langle \mu_s, \theta (v_1^* \bar{\mathbf{n}}_1 + v_2^* \bar{\mathbf{n}}_2) \rangle \geq \frac{2}{3}(v_1^* + v_2^*)\|\mu_c\|^2.$$

The inequality follows from the arguments already stated in Lemma 8, where the second and third items in the above expression have magnitude at most $\frac{1}{6}(v_1^* + v_2^*)\|\mu_c\|^2$. Now it is left to conclude that $(v_1^* + v_2^*) \geq \frac{\Delta}{2}$, which is a direct consequence of Lemma 7 and Lemma 8. □

F.2 Proof of Proposition 2

Now we are in place to prove the guarantee given in the main paper on the robust error of the model returned by the algorithm. We will restate it here with compatible notation to the earlier parts of this section which slightly differ from those in the main paper (e.g. by incorporating Δ). We also note that to obtain the statement in the main paper we should eliminate the dependence of Assumption 1 on $M_{e,1}$. We do this by assuming that our algorithm draws M_e as half of the original dataset for environment e. Then we have that $\mathbb{P}(M_{e,1} \leq N_{\min}/8)$ is bounded by the cumulative probability of a Binomial variable with $k = N_{\min}/8$ successes and at least $N_{\min}$ trials. This may be bounded with a Hoeffding bound by $1 - 2\exp(\frac{1}{2}N_{\min})$ and with a union bound over the two environments. To absorb this into our failure probability we require $N_{\min} > c_{eo}\log(1/\delta)$, leading to this added constraint in the main paper.

Proposition 4. *Under Assumption 1, let $\epsilon > 0$ be the target maximum error of the model and $t > 0$. If $\|\mu_c\|^2 \geq tQ^{-1}(\epsilon)\frac{15}{\Delta}\sigma^2\sqrt{\frac{d}{N_{\min}}}$, then with probability at least $1 - 34\exp(-t^2/2)$ the robust accuracy error of the model is at most ϵ.*

Proof. The error of the model in the environment defined by $\theta \in [-1, 1]$ is given by the Gaussian tail function:

$$Q\left(\frac{\langle \mathbf{w}, \mu_c + \theta\mu_s \rangle}{\sigma\|\mathbf{w}\|}\right)$$

The nominator of this expression is simply the expected score from Lemma 9, which we already proved is at least $\frac{\Delta}{3}\|\mu_c\|^2$. Then we need to bound $\|\mathbf{w}\|$ from above to get a bound on the robust accuracy. According to Corollary 3, if we denote $N_{\min} = \min\{N_1, N_2\}$, this upper bound can be taken as $5t\sqrt{\sigma^2 d/N_{\min}}$. We plug this in to get:

$$\frac{\langle \mathbf{w}, \mu_c + \theta\mu_s \rangle}{\sigma\|\mathbf{w}\|} \geq \frac{\Delta}{15t}\|\mu_c\|^2 \frac{1}{\sigma^2}\sqrt{\frac{N_{\min}}{d}}$$

Since Q is a monotonically decreasing function, if $\|\mu_c\|^2 \geq tQ^{-1}(\epsilon)\frac{15}{\Delta}\sigma^2\sqrt{\frac{d}{N_{\min}}}$ our model achieves the desired performance. □

F.3 Required Concentration Bounds

To conclude the proof we now show all the concentration results used in the above derivation. Note that $\mathbf{v}^*$ is determined by all the other random factors in the problem, hence we should be careful when using them in our bounds. We will only use the fact that $\|\mathbf{v}^*\|_\infty = 1$ and hence $\|\mathbf{v}^*\|_1 \leq 2$.

To bound the inner product of noise vectors, we use [33, Theorem 1.1]:

Theorem 2. *(Hanson-Wright inequality). Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent components X_i which satisfy $\mathbb{E}X_i = 0$ and $\|X_i\|_{\psi_2} \leq K$. Let A be an $n \times n$ matrix. Then, for every $t \geq 0$,*

$$\mathbb{P}\left\{\left|X^\top AX - \mathbb{E}[X^\top AX]\right| > t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{K^4\|A\|_{\text{HS}}^2}, \frac{t}{K^2\|A\|}\right)\right]$$

We can apply this theorem to get the following result.

Corollary 2. *for some universal constant $c > 0$ (when we assume w.l.o.g that $M_{e'} \leq N_e$):*

$$\mathbb{P}\left\{|\langle \bar{\mathbf{n}}_e, \bar{\mathbf{m}}_{e'} \rangle| > t\right\} \leq 2\exp\left[-c\min\left(\frac{M_{e'}^2 t^2}{\sigma^4 d}, \frac{M_{e'}t}{\sigma^2\sqrt{d}}\right)\right] \tag{20}$$

Proof. We take X as the concatenation of $\bar{\mathbf{n}}_e$ and $\bar{\mathbf{m}}_{e'}$, then A is set such that $X^\top AX = \langle \bar{\mathbf{n}}_e, \bar{\mathbf{m}}_{e'} \rangle$ (e.g. $A_{i,i+d} = 1$ for $1 \leq i \leq d$ and 0 elsewhere). Then $\|A\|_{HS}^2 = d$ and $\|A\| = \sqrt{d}$. Since entries in $\bar{\mathbf{n}}_e, \bar{\mathbf{m}}_{e'}$ are distributed as $\mathcal{N}(0, \frac{\sigma^2}{N_e}), \mathcal{N}(0, \frac{\sigma^2}{M_e})$ respectively, we have $K \leq C\frac{\sigma}{\sqrt{\min\{N_e, M_{e'}\}}}$ (assume w.l.o.g that $M_{e'} < N_e$) for some universal constant C which we can incorporate into the constant c in the theorem. This gives:

$$\mathbb{P}\left\{|\langle \bar{\mathbf{n}}_e, \bar{\mathbf{m}}_{e'} \rangle| > t\right\} \leq 2\exp\left[-c\min\left(\frac{M_{e'}^2 t^2}{\sigma^4 d}, \frac{M_{e'}t}{\sigma^2\sqrt{d}}\right)\right]$$

□

The next statement collects all of the concentration results we require for the other parts of the proof.

Lemma 10. *Define $r := \frac{\Delta\theta_{\max}}{\Delta + 4\theta_{\max}}$ where $\theta_{\max} := \arg\max_{e\in\{1,2\}}\{|\theta_e|\}$, denote by v^* the solution retrieved by the algorithm, and let $t > 0$. When Assumption 1 holds, then with probability at least*

$1 - 34\exp(-t^2/2)$ *we have that all the following events occur simultaneously (for all* $e, e' \in \{1, 2\}$*):*

$$|\langle \bar{\mathbf{n}}_e, \mu_s \rangle| \leq r\|\mu_s\|^2 \tag{21}$$

$$|\langle \bar{\mathbf{n}}_e, \mu_c \rangle| \leq \frac{\Delta}{24}\|\mu_c\|^2 \tag{22}$$

$$|\langle \bar{\mathbf{m}}_{e,1}, \mu_c + \theta_{e'}\mu_s \rangle| \leq \min\left\{\frac{1}{4}r\Delta\|\mu_s\|^2, \frac{\Delta}{36}\|\mu_c\|^2\right\} \tag{23}$$

$$|\langle \bar{\mathbf{m}}_{e,1}, \mu_s \rangle| \leq \frac{\Delta}{64}\|\mu_c\|^2 \tag{24}$$

$$|\langle \bar{\mathbf{n}}_e, \bar{\mathbf{m}}_{e',1} \rangle| \leq \min\left\{\frac{1}{4}r\Delta\|\mu_s\|^2, \frac{\Delta^2}{288}\|\mu_c\|^2\right\} \tag{25}$$

$$|\langle \bar{\mathbf{m}}_e, (\mu_c + \theta_{e'}\mu_s) \rangle| \leq \frac{1}{48}\Delta \cdot \|\mu_c\|^2 \tag{26}$$

$$|\langle \bar{\mathbf{n}}_e, \bar{\mathbf{m}}_{e'} \rangle| \leq \frac{1}{48}\Delta \cdot \|\mu_c\|^2 \tag{27}$$

$$\|\bar{\mathbf{n}}_e\| \leq t\sqrt{\frac{2\sigma^2 d}{N_e}} \tag{28}$$

Proof. We first treat Equation (21) with a tail bound for Gaussian variables:

$$\langle \bar{\mathbf{n}}_e, \mu_s \rangle \sim \mathcal{N}(0, \frac{\sigma^2\|\mu_s\|^2}{N_e}) \Rightarrow \mathbb{P}\left(|\langle \bar{\mathbf{n}}_e, \mu_s \rangle| > t_2\right) \leq 2\exp\left(-\frac{t_2^2 N_e}{2\sigma^2\|\mu_s\|^2}\right)$$

Hence as long as $\|\mu_s\|^2 \geq t\frac{2\sigma^2}{r^2 N_e}$, Equation (21) holds with probability at least $1 - 4\exp\{-t^2\}$ (since we take a union bound on the two environments). Following the same inequality and taking a union bound, Equation (22) also hold with probability at least $1 - 8\exp\{-t^2\}$ if $\|\mu_c\|^2 \geq t\frac{1152\sigma^2}{\Delta^2 N_e}$.

We use the same bound for Equation (23), Equation (24) and Equation (26) while using $|\theta_e| \leq 1$. Hence for $t_2 = \frac{1}{4}r\Delta\|\mu_s\|^2$ and $t_2 = \frac{\Delta}{36}\|\mu_c\|^2$:

$$\begin{aligned}
\mathbb{P}\left(|\langle \bar{\mathbf{m}}_{e,1}, \mu_c + \theta_{e'}\mu_s \rangle| > t_2\right) &\leq 2\exp\left(-\frac{t_2^2 M_{e,1}}{2\sigma^2\|\mu_c + \theta_{e'}\mu_s\|^2}\right) = 2\exp\left(-\frac{(r\Delta)^2\|\mu_s\|^4 M_{e,1}}{32\sigma^2\|\mu_c + \theta_{e'}\mu_s\|^2}\right) \\
&\leq 2\exp\left(-\frac{(r\Delta)^2\|\mu_s\|^2 M_{e,1}}{128\sigma^2}\right) \\
\mathbb{P}\left(|\langle \bar{\mathbf{m}}_{e,1}, \mu_c + \theta_{e'}\mu_s \rangle| > t_2\right) &\leq 2\exp\left(-\frac{\Delta^2\|\mu_c\|^4 M_{e,1}}{2592\sigma^2\|\mu_c + \theta_{e'}\mu_s\|^2}\right) = 2\exp\left(-\frac{\Delta^2\|\mu_c\|^2 M_{e,1}}{10368\sigma^2 r_\mu^2}\right)
\end{aligned}$$

Similarly with $t_2 = \frac{1}{48}\Delta \cdot \|\mu_c\|^2$:

$$\mathbb{P}\left(|\langle \bar{\mathbf{m}}_e, (\mu_c + \theta_{e'}\mu_s) \rangle| > t_2\right) \leq 2\exp\left(-\frac{\Delta^2\|\mu_c\|^4 M_e}{(48\sigma\|\mu_c + \theta_{e'}\mu_s\|)^2}\right)$$

Taking the required union bounds we get that with probability at least $1 - 24\exp\left(-t^2/2\right)$ Equation (23), Equation (24) and Equation (26) hold, as long as $\|\mu_s\|^2 \geq t \cdot 128\sigma^2((r\Delta)^2 M_{e,1})^{-1}$ and $\|\mu_c\|^2 \geq t \cdot \max\left\{10368\sigma^2 r_\mu^2\left(\Delta^2 M_{e,1}\right)^{-1}, (96\sigma r_\mu)^2(\Delta^2 M_e)^{-1}\right\}$.

For Equation (25) and Equation (27) we use Corollary 2: [9]

$$\mathbb{P}\left\{|\langle \bar{\mathbf{n}}_e, \bar{\mathbf{m}}_{e',1} \rangle| \geq t_2\right\} \leq 2\exp\left[-c\frac{M_{e',1}^2 t_2^2}{\sigma^4 d}\right]$$

[9] For simplicity, assume we have $\sqrt{M_{1,1}^{-2} + M_{2,1}^{-2}} \leq N_1^{-1}$ and that we set t large enough such that $\left(M_{1,1}^{-1} + M_{2,1}^{-1}\right)^{-2} t^2/(\sigma^4 d) \geq \left(M_{1,1}^{-1} + M_{2,1}^{-1}\right)^{-1} t/(\sigma^2\sqrt{d})$

Setting $t_2 = \frac{r\Delta}{4}\|\mu_s\|^2$ or $t_2 = \frac{\Delta^2}{288}\|\mu_c\|^2$ we will get that:

$$\mathbb{P}\left(|\langle \bar{\mathbf{n}}_e, \bar{\mathbf{m}}_{e',1}\rangle| \geq \min\left\{\frac{r\Delta}{4}\|\mu_s\|^2, \frac{\Delta^2}{288}\|\mu_c\|^2\right\}\right) \leq 2\exp\left(-c\frac{M_{e',1}^2}{\sigma^4 d}\min\left\{\frac{(r\Delta)^2}{16}\|\mu_s\|^4, \frac{\Delta^4}{288^2}\|\mu_c\|^4\right\}\right)$$

Hence we require $\|\mu_c\|^2 \geq t \cdot c \cdot (M_{e',1}\Delta^2)^{-1} \cdot (288\sigma^2\sqrt{d})$ and $\|\mu_s\|^2 \geq t \cdot c \cdot (M_{e',1}r\Delta)^{-1} \cdot (4\sigma^2\sqrt{d})$ for Equation (25) to hold. For Equation (27) we can get in a similar manner that it holds in case that $\|\mu_c\|^2 \geq t \cdot c \cdot (M_{e'}\Delta)^{-1}(48\sigma^2\sqrt{d})$. The probability for all the events listed so far to occur is at last $1 - 32\exp\left(-t^2/2\right)$. Finally, for Equation (28) we simply use the bound on a norm of Gaussian vector:

$$\mathbb{P}\left(\|\bar{\mathbf{n}}_e\| \geq t_2\right) \leq 2\exp\left(-\frac{t_2^2 N_e}{2\sigma^2 d}\right)$$

Plugging in $t\sqrt{\frac{2\sigma^2 d}{N_e}}$ we arrive at the desired result with a final union bound that give the overall probability of at least $1 - 34\exp\left(-t^2/2\right)$. □

We now use the bounds above to write down the specific bounds on expressions that we used during proof.

Corollary 3. *Conditioned on all the events in Lemma 10, we have for* $e \in \{1, 2\}$ *that:*

$$\frac{\Delta}{6}|v_1 + v_2| \cdot \|\mu_c\|^2 \geq |\epsilon_2(\mathbf{v}) - \epsilon_1(\mathbf{v})| \tag{29}$$

$$\frac{1}{6}|v_1 + v_2| \cdot \|\mu_c\|^2 \geq |\langle \mu_c, v_1\bar{\mathbf{n}}_1 + v_2\bar{\mathbf{n}}_2\rangle| \tag{30}$$

$$\frac{1}{6}|v_1 + v_2| \cdot \|\mu_c\|^2 \geq |\langle \bar{\mathbf{m}}_e, (v_1 + v_2)\mu_c + (\theta_1 v_1 + \theta_2 v_2)\mu_s + v_1\bar{\mathbf{n}}_1 + v_2\bar{\mathbf{n}}_2\rangle| \tag{31}$$

$$r\Delta\|\mu_s\|^2 \geq |\langle \bar{\mathbf{m}}_{1,1} - \bar{\mathbf{m}}_{2,1}, \mu_c + \theta_e\mu_s + \bar{\mathbf{n}}_e\rangle| \tag{32}$$

$$r\|\mu_s\|^2 \geq |\langle \bar{\mathbf{n}}_e, \mu_s\rangle| \tag{33}$$

$$5t\sqrt{\frac{\sigma^2 d}{\min_e N_e}} \geq \|\mathbf{w}\| \tag{34}$$

Proof. Equation (33) is just Equation (21) restated for convenience. Equation (32) is a combination of Equation (23) and Equation (25):

$$|\langle \bar{\mathbf{m}}_{1,1} - \bar{\mathbf{m}}_{2,1}, \mu_c + \theta_e\mu_s + \bar{\mathbf{n}}_e\rangle| \leq \sum_{e'} |\langle \bar{\mathbf{m}}_{e',1}, \mu_c + \theta_e\mu_s\rangle| + |\langle \bar{\mathbf{m}}_{e',1}, \bar{\mathbf{n}}_e\rangle| \leq r\Delta\|\mu_s\|^2$$

These are the events required for Lemma 7, hence from now on we can now assume that:

$$|v_1 + v_2| \geq \frac{\Delta}{2} = \frac{\Delta}{4} \cdot 2 \geq \frac{\Delta}{4}\|\mathbf{v}\|_1$$

Now we can combine with Equation (22) to prove Equation (30):

$$\langle \mu_c, v_1\bar{\mathbf{n}}_1 + v_2\bar{\mathbf{n}}_2\rangle \leq \sum_e |v_e| \cdot |\langle \mu_c, \bar{\mathbf{n}}_e\rangle| \leq \|\mathbf{v}\|_1 \frac{\Delta}{24}\|\mu_c\|^2 \leq \frac{1}{6}|v_1 + v_2| \cdot \|\mu_c\|^2$$

Next we prove Equation (31) in a similar manner using Equation (26) and Equation (27):

$$|\langle \bar{\mathbf{m}}_e, (v_1 + v_2)\mu_c + (\theta_1 v_1 + \theta_2 v_2)\mu_s + v_1\bar{\mathbf{n}}_1 + v_2\bar{\mathbf{n}}_2\rangle| \leq \sum_{e'} |v_{e'}| \cdot (|\langle \bar{\mathbf{m}}_e, \mu_c + \theta_{e'}\mu_s\rangle| + |\langle \bar{\mathbf{m}}_e, \bar{\mathbf{n}}_{e'}\rangle|) \leq \|\mathbf{v}\|_1 \cdot 2 \cdot \frac{1}{48}\Delta\|\mu_c\|^2 \leq \frac{1}{6}|v_1 + v_2| \cdot \|\mu_c\|^2$$

For Equation (29), let us write the right hand side:

$$\begin{aligned}
|\epsilon_2(\mathbf{v}) - \epsilon_1(\mathbf{v})| &= |\langle \bar{\mathbf{m}}_{2,1} - \bar{\mathbf{m}}_{1,1}, v_1(\mu_c + \theta_1\mu_s + \bar{\mathbf{n}}_1) + v_2(\mu_c + \theta_2\mu_s + \bar{\mathbf{n}}_2)\rangle| \\
&= |(v_1 + v_2) \cdot \langle \bar{\mathbf{m}}_{2,1} - \bar{\mathbf{m}}_{1,1}, \mu_c + \frac{1}{2}(\theta_1 + \theta_2)\mu_s\rangle \\
&+ \langle \bar{\mathbf{m}}_{2,1} - \bar{\mathbf{m}}_{1,1}, v_1\bar{\mathbf{n}}_1 + v_2\bar{\mathbf{n}}_2\rangle + \frac{1}{2}(v_1 - v_2)\langle \bar{\mathbf{m}}_{2,1} - \bar{\mathbf{m}}_{1,1}, \Delta\mu_s\rangle| \\
&\leq |v_1 + v_2| \cdot \sum_e |\langle \bar{\mathbf{m}}_{e,1}, \mu_c + \frac{1}{2}(\theta_1 + \theta_2\mu_s)\rangle| + \|\mathbf{v}\|_1 \sum_{e,e'} |\langle \bar{\mathbf{m}}_{e,1}, \bar{\mathbf{n}}_{e'}\rangle| \\
&+ \frac{1}{2}\Delta\|\mathbf{v}\|_1 \sum_e |\langle \bar{\mathbf{m}}_{e,1}, \mu_s\rangle| \\
&\leq |v_1 + v_2| \cdot \sum_e |\langle \bar{\mathbf{m}}_{e,1}, \mu_c + \frac{1}{2}(\theta_1 + \theta_2\mu_s)\rangle| + \frac{4}{\Delta}|v_1 + v_2| \sum_{e,e'} |\langle \bar{\mathbf{m}}_{e,1}, \bar{\mathbf{n}}_{e'}\rangle| \\
&+ 2|v_1 + v_2| \sum_e |\langle \bar{\mathbf{m}}_{e,1}, \mu_s\rangle| \\
&\leq \frac{1}{6}\Delta|v_1 + v_2|
\end{aligned}$$

The first inequality is simply a triangle inequality, the second plugs in the bound we obtained for $\|\mathbf{v}\|_1$ and the last uses the relevant inequalities from Lemma 10.

For Equation (34), we write the weights of the returned linear classifier as:

$$\mathbf{w} = v_1^*(\mu_c + \theta_1\mu_s + \bar{\mathbf{n}}_1) + v_2^*(\mu_c + \theta_2\mu_s + \bar{\mathbf{n}}_2)$$

Hence we can bound:

$$\begin{aligned}
\|\mathbf{w}\| - (v_1^* + v_2^*)\|\mu_c\| &\leq \|(v_1^*\theta_1 + v_2^*\theta_2)\mu_s + v_1^*\bar{\mathbf{n}}_1 + v_2^*\bar{\mathbf{n}}_2\| \\
&= \sqrt{(v_1^*\theta_1 + v_2^*\theta_2)^2\|\mu_s\|^2 + 2\langle v_1^*\bar{\mathbf{n}}_1 + v_2^*\bar{\mathbf{n}}_2, (v_1^*\theta_1 + v_2^*\theta_2)\mu_s\rangle + \|v_1^*\bar{\mathbf{n}}_1 + v_2^*\bar{\mathbf{n}}_2\|^2} \\
&= \sqrt{(v_1^*\theta_1 + v_2^*\theta_2)\left((v_1^*\theta_1 + v_2^*\theta_2)\|\mu_s\|^2 + 2\langle v_1^*\bar{\mathbf{n}}_1 + v_2^*\bar{\mathbf{n}}_1, \mu_s\rangle\right) + \|v_1^*\bar{\mathbf{n}}_1 + v_2^*\bar{\mathbf{n}}_2\|^2}
\end{aligned}$$

We also proved in Lemma 8, that under the events we assumed and the EOpp constraint:

$$\begin{aligned}
(v_1^*\theta_1 + v_2^*\theta_2)\|\mu_s\|^2 + 2\langle v_1^*\bar{\mathbf{n}}_1 + v_2^*\bar{\mathbf{n}}_2, \mu_s\rangle &\leq 2\left((v_1^*\theta_1 + v_2^*\theta_2)\|\mu_s\|^2 + |\langle v_1^*\bar{\mathbf{n}}_1 + v_2^*\bar{\mathbf{n}}_2, \mu_s\rangle|\right) \\
&\leq \frac{1}{3}(v_1^* + v_2^*)\|\mu_c\|^2
\end{aligned}$$

Incorporating with $v_1^*\theta_1 + v_2^*\theta_2 \leq 2(v_1^* + v_2^*)$, the concavity of the square root and Equation (28), we get:

$$\begin{aligned}
\|\mathbf{w}\| &\leq \left(1 + \sqrt{2/3}\right)(v_1^* + v_2^*)\|\mu_c\| + \|v_1^*\bar{\mathbf{n}}_1 + v_2^*\bar{\mathbf{n}}_2\| \\
&\leq \left(1 + \sqrt{2/3}\right)(v_1^* + v_2^*)\|\mu_c\| + \|\bar{\mathbf{n}}_1\| + \|\bar{\mathbf{n}}_2\| \\
&\leq \left(1 + \sqrt{2/3}\right)(v_1^* + v_2^*)\|\mu_c\| + t \cdot \sqrt{\frac{\sigma^2 d}{\min_e N_e}} \\
&\leq 4\|\mu_c\| + t \cdot \sqrt{\frac{\sigma^2 d}{\min_e N_e}} \\
&\leq 5t \cdot \sqrt{\frac{\sigma^2 d}{\min_e N_e}}
\end{aligned}$$

□

G Proof of Theorem 1

Proof of Theorem 1. Our proof simply consists of choosing the free parameters in Theorem 1 ($r_c, r_s, d, \sigma, \theta_1$ and θ_2) based on Propositions 1, 2 and 3 such that all the claims in the theorem hold simultaneously. Keeping in line with the setting of Propositions 1 and 3, we take $\sigma^2 = 1/d$, $\theta_1 = 1$ and $\theta_2 = 0$. Next, our strategy is to pick r_s and r_c so as to satisfy the requirements of Propositions 1 and 3, and then pick a sufficiently large d so that the requirements of Proposition 2 hold as well. Throughout, we set $\delta = 99/100$ so as to meet the failure probability requirement stated in the theorem; it is straightforward to adjust the proof to guarantee lower error probabilities.

Starting with the value of r_s, we let

$$r_s^2 = \frac{\min\{c_n, c_n'\}}{N}$$

where the parameters c_n, c_m and c_n' are as given by Propositions 1 and 3, respectively. Next, we pick r_c to be

$$r_c^2 = \frac{r_s^2}{C_r\left(1 + \frac{\sqrt{N_2}}{N_1\gamma}\right)} = \frac{\min\{c_n, c_n'\}}{C_r N\left(1 + \frac{\sqrt{N_2}}{N_1\gamma}\right)}$$

with C_r from Proposition 1 (this setting guarantees $r_c \le r_s$ as $C_r \ge 1$). Thus, we have satisfied the requirements in Equation (1) in Proposition 1, as well as the requirement $\max\{r_c, r_s\} \le \frac{c_n'}{N}$ in Proposition 3; it remains to choose d so that the remaining requirements hold.

Proposition 1 requires the dimension to satisfy $d \ge C_d \frac{N}{\gamma^2 N_1^2 r_c^2} \log\frac{1}{\delta}$ and Proposition 3 requires $d \ge C_d' N^2 \log\frac{1}{\delta}$. Substituting our choices of $\sigma^2 = 1/d$, r_s and r_c above, let us rewrite the requirements of Proposition 2 as lower bounds on d. The requirement in Equation (G) reads

$$d \ge C_s^2 \frac{\log\frac{1}{\delta}}{N_{\min}^2 r_s^4},$$

while the requirement in (with minor simplifications) reads

$$d \ge \frac{C_c^2 \log\frac{1}{\delta}}{N_{\min} r_c^4} \max\left\{(Q^{-1}(\epsilon))^2, \frac{1}{N_{\min}}, r_s^2\right\}.$$

Using $r_s \ge r_c$ and $r_s^2 \le \frac{1}{N_{\min}}$, the above two displays simplify to

$$d \ge \frac{\max\{C_c, C_s\}^2 \log\frac{1}{\delta}}{N_{\min} r_c^4} \max\left\{(Q^{-1}(\epsilon))^2, \frac{1}{N_{\min}}\right\}.$$

Therefore, taking

$$d = \max\{C_d, C_d', C_s^2, C_c^2\} \max\left\{N^2, \frac{N}{\gamma^2 N_1^2 r_c^2}, \frac{(Q^{-1}(\epsilon))^2}{N_{\min} r_c^4}, \frac{1}{N_{\min}^2 r_c^4}\right\} \log\frac{1}{\delta}$$

fulfills all the requirements and completes the proof. □

H Definitions of Invariance and Their Manifestation In Our Model

In section 4 we show that the Equalized Odds principle in our setting reduces to the demand that $\langle \mathbf{w}, \boldsymbol{\mu}_s\rangle = 0$. Here we provide short derivations that show this is also the case for some other invariance principles from the literature. We will show this in the population setting, that is in expectation over the training data. We also assume that $\theta_1 \neq \theta_2$.

Calibration over environments [43] Assume $\sigma(\langle \mathbf{w}, \mathbf{x}\rangle)$ is a probabilistic classifier with some invertible function $\sigma : \mathbb{R} \to [0, 1]$ such as a sigmoid, that maps the output of the linear function to a probability that $y = 1$. Calibration can be written as the condition that:

$$\mathbb{P}_\theta(y = 1 \mid \sigma(\langle \mathbf{w}, \mathbf{x}\rangle - b) = \hat{p}) = \hat{p} \quad \forall \hat{p} \in [0, 1].$$

Calibration on training environments in our setting then requires that this holds simultaneously for $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$. We can write the conditional probability of y on the prediction (when the prior over y is uniform) as:

$$\mathbb{P}_{\theta_e}(y = 1 \mid \langle \mathbf{w}, \mathbf{x} \rangle - b = \alpha) = \frac{\exp\left(\frac{(\alpha - \langle \mathbf{w}, \boldsymbol{\mu}_c + \theta_1 \boldsymbol{\mu}_s \rangle + b)^2}{2\sigma^2 \|\mathbf{w}\|^2}\right)}{\exp\left(\frac{(\alpha - \langle \mathbf{w}, \boldsymbol{\mu}_c + \theta_1 \boldsymbol{\mu}_s \rangle + b)^2}{2\sigma^2 \|\mathbf{w}\|^2}\right) + \exp\left(\frac{(\alpha + \langle \mathbf{w}, \boldsymbol{\mu}_c + \theta_1 \boldsymbol{\mu}_s \rangle + b)^2}{2\sigma^2 \|\mathbf{w}\|^2}\right)}$$

Now it is easy to see that if the classifier is calibrated across environments, we must have equality in the log-odds ratio for the above with $e = 1$ and $e = 2$ and all $\alpha \in \mathbb{R}$:

$$\frac{(\alpha - \langle \mathbf{w}, \boldsymbol{\mu}_c + \theta_1 \boldsymbol{\mu}_s \rangle + b)^2}{2\sigma^2 \|\mathbf{w}\|^2} - \frac{(\alpha + \langle \mathbf{w}, \boldsymbol{\mu}_c + \theta_1 \boldsymbol{\mu}_s \rangle + b)^2}{2\sigma^2 \|\mathbf{w}\|^2} =$$
$$\frac{(\alpha - \langle \mathbf{w}, \boldsymbol{\mu}_c + \theta_2 \boldsymbol{\mu}_s \rangle + b)^2}{2\sigma^2 \|\mathbf{w}\|^2} - \frac{(\alpha + \langle \mathbf{w}, \boldsymbol{\mu}_c + \theta_2 \boldsymbol{\mu}_s \rangle + b)^2}{2\sigma^2 \|\mathbf{w}\|^2}.$$

After dropping all the terms that cancel out in the subtractions we arrive at:

$$\langle \mathbf{w}, \boldsymbol{\mu}_c + \theta_1 \boldsymbol{\mu}_s \rangle = \langle \mathbf{w}, \boldsymbol{\mu}_c + \theta_2 \boldsymbol{\mu}_s \rangle.$$

Clearly this holds if and only if $\langle \mathbf{w}, \boldsymbol{\mu}_s \rangle = 0$, hence calibration on both environments entails invariance in the context of the data generating process of Definition 2.

Conditional Feature Matching [23, 40] Treating the environment index as a random variable, the conditional independence relation $\langle \mathbf{w}, \mathbf{x} \rangle \perp\!\!\!\perp e \mid y$ is a popular invariance criterion in the literature. Other works besides the ones mentioned in the title of this paragraph have used this, like the Equalized Odds criterion [15]. This independence is usually enforced w.r.t available training distributions, hence in our case w.r.t $\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}$. Writing this down we can see that:

$$\mathbb{P}_{\theta_e}(\langle \mathbf{w}, \mathbf{x} \rangle \mid y = 1) = \mathcal{N}(\langle \mathbf{w}, \mu_c + \theta_e \mu_s \rangle, \|\mathbf{w}\|^2 \sigma^2 I).$$

Hence requiring conditional independence in the sense of $\mathbb{P}_{\theta_1}(\langle \mathbf{w}, \mathbf{x} \rangle \mid y = 1) = \mathbb{P}_{\theta_2}(\langle \mathbf{w}, \mathbf{x} \rangle \mid y = 1)$ means we need to have equality of the expectations, i.e. $\langle \mathbf{w}, \mu_c + \theta_1 \mu_s \rangle = \langle \mathbf{w}, \mu_c + \theta_2 \mu_s \rangle$ which happens only if $\langle \mathbf{w}, \mu_s \rangle = 0$.

Other notions of invariance. It is easy to see that even without conditioning on y, the independence relation $\langle \mathbf{w}, \mathbf{x} \rangle \perp\!\!\!\perp e$ used in Veitch et al. [40] among many others will also require that $\langle \mathbf{w}, \boldsymbol{\mu}_s \rangle = 0$. For the last invariance principle we discuss here, we note that VREx and CVaR Fairness essentially require equality in distribution of losses [45, 20] under both environments. Examining the expression for the error of $\mathbf{w}$ under our setting (Equation (2)) reveals immediately that these conditions will also impose $\langle \mathbf{w}, \boldsymbol{\mu}_s \rangle = 0$.