Why does Throwing Away Data Improve Worst-Group Error?

Kamalika Chaudhuri [*1 2] **Kartik Ahuja** [*1] **Martin Arjovsky** [3] **David Lopez-Paz** [1]

Abstract

When facing data with imbalanced classes or groups, practitioners follow an intriguing strategy to achieve best results. They throw away examples until the classes or groups are balanced in size, and then perform empirical risk minimization on the reduced training set. This opposes common wisdom in learning theory, where the expected error is supposed to decrease as the dataset grows in size. In this work, we leverage extreme value theory to address this apparent contradiction. Our results show that the tails of the data distribution play an important role in determining the worst-group-accuracy of max-margin linear classifiers on linearly separable data. When learning on data with heavy tails, throwing away data restores the geometric symmetry of the resulting classifier, and therefore improves its worst-group generalization.

1. Introduction

Imbalances are ubiquitous in real-world data. On the one hand, class imbalance is common in rare-event data such as medical diagnosis, intrusion detection, spam classification, or credit fraud (Johnson & Khoshgoftaar, 2019). On the other hand, imbalances may also exist within the classes of our problem, if each of these comprises hidden groups with different proportions. For instance, in an image classification dataset with balanced classes, most pictures for each class are commonly taken in wealthy countries (Rojas et al., 2022). In all of these situations, simply minimizing average training error may result in classifiers that perform very well on majority groups, while having poor performance on minority groups (Buolamwini & Gebru, 2018). For example, the 1% test error of a classifier could mean that *all* examples from a particular minority group are misclassified. As such, this work is concerned with the 'worst-group-error' of classification rules on imbalanced data.

While many sophisticated methods have been proposed to address imbalanced classification problems (Sagawa et al., 2019), none of these offer a clear advantage to simple subsampling (Idrissi et al., 2022). In subsampling, we throw away data from large groups until they match the smallest group in size. Then, we perform empirical risk minimization on the (often drastically) reduced data. This is a surprising finding, as classical wisdom in learning theory tells us that the expected error of classifiers should decrease as training data grows in size. As it stands, there is no theoretical explanation as to why the popular strategy of subsampling works so well when addressing imbalanced classification problems. In particular, commonplace PAC-learning results (Haussler, 1990) upper-bound the error of the worst of the classifiers correctly classifying all of the training data. While this analysis can be extended to other metrics such as worst-group error, since throwing away examples increases the amount of such compatible classifiers, the error bound can only worsen, falling short in explaining the empirical benefits of data subsampling.

This work is an initial effort to resolve this apparent tension between theory and practice. More specifically, we focus our analysis on linear maximum-margin binary classifiers (Steinwart & Christmann, 2008) on linearly separable data obtained from either logistic regression or hard-margin SVM. Under this setup, it is well known (Bennett & Bredensteiner, 2000) that maximum-margin classifiers are equidistant from the convex hulls delineated by each of the two classes. In turn, the shape of these convex hulls is determined by the extremal properties of the probability distributions of each class. These extremal properties are the subject of study of a branch of mathematics called extreme value theory (EVT, De Haan & Ferreira (2007)). By borrowing results from EVT, we show that the location of the maximum-margin classifier depends on the tail properties of the data distribution. In particular, when facing data distributions with heavy tails, we observe geometric imbalances when groups differ in sample size. These imbalances skew the maximum-margin classifier, leading to suboptimal worst-group-error. This ends up being the reason why subsampling works, as balancing groups in size restores geometric symmetry across groups.

[*]Equal contribution [1]FAIR (Meta AI) [2]UC San Diego [3]Inria. Correspondence to: Kamalika Chaudhuri <kamalika@meta.com>, Kartik Ahuja <kartikahuja@meta.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

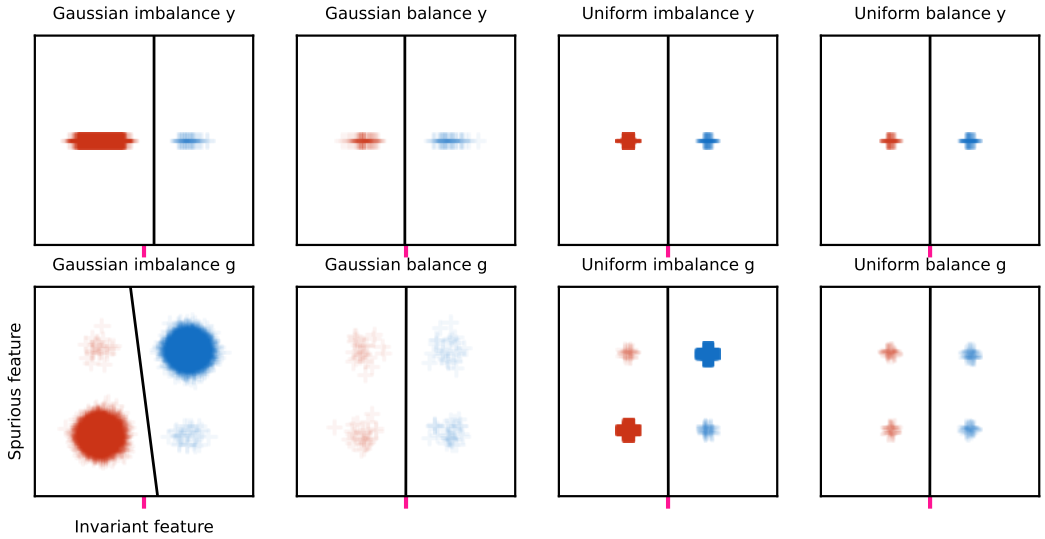

Figure 1: Illustration of the phenomena studied in this paper. The tails of the data distribution can bias the maximum-margin classifier, and throwing away is a tool to restore balance and achieve the optimal solution. The top four plots illustrate a classification problem with two imbalanced classes. In the Gaussian case, throwing away data to balance class sizes leads to the optimal maximum-margin classifier, which is aligned with magenta checkmark. In the Uniform case, throwing away data does not make any difference. The bottom four plots illustrate a binary classification problem, where each class is balanced but contains two groups of different proportions. Once again, in the Gaussian case, throwing away data to balance group sizes leads to the optimal maximum-margin classifier. Subsampling does not provide any benefit in the Uniform case.

We illustrate these phenomena in Figure 1. When dealing with imbalanced classes, we find that the bias term in the maximum-margin classifier is shifted towards the small class. When reducing the data with subsampling to balance the two classes, the maximum-margin classifier converges to the unbiased one with optimal worst-group-error. When dealing with imbalanced groups, the direction of the maximum-margin classifier is biased towards the smaller group, increasing its error in test examples from the tails of the distribution. Figure 1 shows that, while these phenomena happen when dealing with data distributions with Gaussian tails, they are non-existent when dealing with Uniform data distributions, with no tails.

Contributions This work proposes a novel theoretical analysis to understand imbalanced classification, as well as to justify the popularity of data subsampling in this regime (Section 2). To this end, we introduce the use of extreme value theory to construct a new type of generalization analysis that focusses on distributional tails (Section 3). These results allow us to characterize the impact of distributional tails on the worst-group error for both ERM and subsampling strategies. We conduct separate analyses to understand the case of imbalanced classes (Section 4) and groups (Section 5). In particular,

- Subsampling outperforms ERM in worst-group-error when learning from imbalanced classes with tails (such as Gaussians, Theorem 3), while it makes no difference when learning from distributions without tails (such as Uniforms, Theorem 4).
- Similar results follow for balanced classes but imbalanced groups (Theorem 6 for groups with tails, Theorem 7 for groups without tails).
- We extend these results to the high-dimensional case where there exist a multitude of "noise" dimensions polluting the data (Theorems 5, 8, 9) and provide empirical support for our theories using Waterbirds and CelebA, the two most common datasets to benchmark worst-group-error (Section 6).

2. Our learning setup

We consider the binary classification of examples (x, y), where instances $x \in \mathbb{R}^d$ and labels $y \in \{-1, +1\}$. We assume that instances with label y are drawn independently from a class-conditional distribution that we denote by D^y. We look at two settings—imbalanced classes and imbalanced groups. In the setting of imbalanced classes, there is a majority class with $y = -1$ and a minority class with $y = +1$. Training data $\mathcal{S}$ consists of p samples drawn independently and identically distributed (iid) from the majority class-conditional distribution, and m from the minority class-conditional distribution. We assume p is much larger than m. For imbalanced groups, we extend each example to

be a triplet (x, y, a), where $a \in \{-1, 1\}$ is a binary attribute. Label-attribute combinations induce four groups $g = (y, a)$. Then, the data distribution consists of p points drawn iid from the majority groups $g = (1, 1)$ and $(-1, -1)$, and m points drawn from the minority groups $g = (-1, 1)$ and $(1, -1)$. Once again, p is much larger than m. While describing data distributions, we use the notation $D(\rho)$ to denote a distribution with mean ρ.

Commonly, in machine learning we are interested in the generalization error of a classifier θ on some data distribution D, $\mathbf{err}(\theta) = \Pr_D(\theta(x) \neq y)$, where we drop the subscript D when this causes no confusion. For problems with imbalanced classes, we are chiefly interested in measuring *worst-class error*: $\mathbf{wce}(\theta) = \max_{\tilde{y}} \Pr(\theta(x) \neq y \mid y = \tilde{y})$. For imbalanced groups, the metric of interest is the *worst-group error*: $\mathbf{wge}(\theta) = \max_{\tilde{g}} \Pr(\theta(x) \neq y \mid g = \tilde{g})$.

To enable a fine-grained analysis, we focus on a linear classifier, namely $\theta(x) = w^\top x + b$, where w is the weight and b is the bias. We assume that the training data is linearly separable with high probability—even though the entire distribution may not be. This mirrors what happens in deep learning, where zero-training-error is easy to achieve, while zero-test-error may be elusive. The usual way to train a classifier is to follow the Empirical Risk Minimization (ERM) principle to learn the parameters (w, b) separating the training data $\mathcal{S}$. This amounts to finding (w, b) that minimizes

$$L(w, b, \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell\left((w^\top x_i + b) y_i\right), \tag{1}$$

where $\ell : \mathbb{R} \to [0, \infty]$ is a loss function that penalizes classification errors. The most popular loss for classification in deep learning is the logistic loss $\ell(u) = \log(1 + \exp(-u))$.

Support vector machines When our training data is linearly separable, (Soudry et al., 2018) has shown that the logistic loss minimizer converges to the well-known maximum-margin classifier also referred to as the linear hard-margin Support Vector Machine (Vapnik, 1999, SVM). The sequel therefore analyzes the properties of the SVM separator directly. The SVM separator characterized by w^* and b^* is known to be equidistant from the convex hulls of the positive and negative classes (Bennett & Bredensteiner, 2000); it can be obtained by solving:

$$\begin{aligned} w^* &= \arg\max_{\|w\|=1} \Big(\inf_{x \in B} w^\top x - \sup_{x \in A} w^\top x \Big), \\ b^* &= -\frac{1}{2} \Big(\inf_{x \in B} w^\top x + \sup_{x \in A} w^\top x \Big), \end{aligned} \tag{2}$$

where B is set of positively labeled points and A is the set of negatively labeled points. Our analysis connects the optimal w^* and b^* to the tail properties of the data distribution.

ERM versus subsampling We will compare two algorithms. On the one hand, given a training dataset with imbalanced classes or groups, the ERM algorithm directly solves the SVM optimization problem on the entire dataset to get the classifier θ_{erm}. On the other hand, the subsampling algorithm solves the SVM problem on a *subsample* of the training data that balances out classes or groups. Specifically, subsampled training sets consists of the entire minority class or group, and a randomly drawn sample of size m from the majority class or group. Solving the SVM on this reduced dataset gives us the classifier θ_{ss}.

Why are PAC bounds not enough? We will show that subsampling leads in certain cases to a strictly better worst-class or worst-group error compared with plain ERM on the entire training data. In other words, throwing away data strictly helps! This kind of result cannot be directly obtained through standard PAC-style analysis in learning theory. Basic PAC-style analysis in the realizable case provides an upper bound on the worst-case error of any classifier in the *version space*—which is the set of all classifiers that perfectly classify the training points. While this analysis can be adapted to other metrics (such as wce and wge), throwing away data expands the version space, and hence the worst-case error of this expanding set must also increase.

To address this apparent contradiction, the sequel relies on the geometric properties of the maximum-margin classifier. We will use the fact that the maximum-margin separator is equidistant from the convex hulls of the two classes (Bennett & Bredensteiner, 2000), and geometrically analyze properties of these convex hulls. This analysis leverages the fact that properties of the convex hull of a set of random points are related to extreme value statistics (Bennett & Bredensteiner, 2000) of the distribution. To this end, our analysis makes use of extreme value theory, a branch of probability theory concerned with maxima and minima of distributions.

3. Basics of extreme value theory

The branch of mathematics studying extreme deviations is known as Extreme Value Theory (De Haan et al., 2006, EVT). We borrow tools from EVT to analyze the worst-group error in SVMs, which is a first in the research literature. Specifically, we will make use of a central result from EVT: the Fisher-Tippett-Gnedenko theorem (De Haan et al., 2006). Suppose we have n iid examples $X_1, \ldots, X_n$ drawn from an fixed distribution with CDF F; the Fishet-Tippett-Gnedenko theorem characterizes the maximum value M_n of $X_1, \ldots, X_n$ provided the distribution F belongs to one of the following types.

Definition 1. *Let $x_F = \sup_x \{x \mid F(x) < 1\}$ be the largest value not saturating F. Then F is of the family:*

- Frechet, *if* $x_F = \infty$ *and* $\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = x^{-\alpha}$, *for* $\alpha > 0$ *and* $x > 0$.
- Weibull, *if* $x_F < \infty$ *and* $\lim_{h\to 0} \frac{1-F(x_F - xh)}{1-F(x_F-h)} = x^{\alpha}$, *for* $\alpha > 0$ *and* $x > 0$.
- Gumbel, *if* $\lim_{t\to x_F} \frac{1-F(t+xg(t))}{1-F(t)} = e^{-x}$, *for all real* x *where* $g(t) = \frac{\int_t^{x_F}(1-F(u))du}{1-F(t)}$ *for* $t < x_F$.

In this work, we only study Gumbel and Weibull distributions. Gumbel-type distributions have light tails, and include Gaussians and Exponentials. Weibull-type distributions have finite maximum points, and include Uniforms. Frechet-type distributions have heavy tails, including the Pareto and Frechet distributions. Some distributions, such as the Bernoulli, do not belong to any of these types. If the distribution F is of either of these three types, then the Fisher-Tippett-Gnedenko theorem shows that there exist sequence a_n and b_n such that the CDF F' of $(M_n - b_n)/a_n$ converges to a limit distribution G.

$$F'\left(\frac{M_n - b_n}{a_n}\right) \to G,$$

where a_n, b_n and G take values that depend on the specific distribution F. Finally, we define the "tail function" of a distribution with CDF F to measure the spread of its tail.

Definition 2 (Tail Function)**.** *The tail function* U *of a distribution with CDF* F *is* $U(t) = F^{-1}(1 - 1/t)$.

Observe that the tail function $U(t)$ is an increasing function of t; in addition, by definition, we have $F(U(t)) = 1 - 1/t$. We now have all the necessary tools to introduce the Fisher-Tippet-Gnedenko theorem formally.

Theorem 1 (Fisher-Tippett-Gnedenko Theorem)**.** *1. If* F *is of the Frechet type, then* $G(x)$ *is the Frechet distribution with the following CDF:*

$$\begin{aligned} G(x) = &\quad \exp(-x^{-\alpha}), &\quad x \geq 0 \\ = &\quad 0, &\quad x < 0. \end{aligned}$$

Additionally, $a_n = U(n)$ *and* $b_n = 0$.

2. If F *is of the Weibull type, then* $G(x)$ *is the reverse Weibull distribution with the following CDF:*

$$\begin{aligned} G(x) = &\quad 1, &\quad x \geq 0 \\ = &\quad \exp(-(-x)^{\alpha}), &\quad x < 0. \end{aligned}$$

Additionally, $a_n = x_F - U(n)$ *and* $b_n = x_F$.

3. If F *is of the Gumbel type, then* $G(x)$ *is the Gumbel distribution with the following CDF:*

$$G(x) = \exp(-e^{-x}), \quad x \in [-\infty, \infty].$$

Additionally, $a_n = g(U(n))$ *and* $b_n = U(n)$.

4. Analysis of imbalanced classes

We first look at data from imbalanced classes. The basic setup is as follows. To keep the main message of our analysis simple, we assume that the positive (minority) class and the negative (majority) class are both distributed according to a distribution $D(\cdot)$, but with shifted means. Specifically, the positive class is distributed according to $D(\mu)$, and the negative class according to $D(-\mu)$, where $\|\mu\| > 0$. Additionally, we assume that $D(\mu)$ is symmetric about its mean μ – although this symmetry is not strictly needed for our results to hold. Finally, recall that we have p points from the majority class and m from the minority with $p \gg m$. To build intuition, we first look at a simple one-dimensional setting where the feature x_i's are scalars. In this case, weight w is set to one, we use θ to refer to bias b. This case has two interesting properties that makes the analysis intuitive. First, due to symmetry of the class-conditional distributions, the classifier that minimizes worst-class error has a bias of zero. Second, the bias of the maximum-margin classifier θ_{erm} is the mean of the maximum training point with a negative label and the minimum training point with a positive label.

How does ERM behave under these conditions? Geometry suggests that if, due to class imbalance, training data from the positive majority class is *spread out* enough to push the bias of θ_{erm} away from zero, then ERM will have poor worst-class accuracy, and subsampling will help. In contrast, if training data from both classes are equally spread-out, then ERM will retain its symmetry. We formalize this notion of *spreading out* through a Concentration Condition below.

Assumption 1 (Concentration Condition)**.** *Suppose* $x_1, \ldots, x_n$ *are scalars drawn i.i.d from* $D(0)$. *There exist maps* $X_{\max} : \mathbb{Z}_+ \times [0,1] \to \mathbb{R}$, $c : \mathbb{Z}_+ \times [0,1] \to \mathbb{R}$, *and* $C : \mathbb{Z}_+ \times [0,1] \to \mathbb{R}$ *such that for all* $n \geq n_0$, *for every* $\delta \in (0,1)$, *with probability* $\geq 1 - \delta$,

$$\max_{i\in\{1\cdots,n\}} x_i \in \left[X_{\max}(n,\delta) - c(n,\delta), X_{\max}(n,\delta) + C(n,\delta)\right]$$

Also, $\lim_{n\to\infty} C(n,\delta) = 0$, $\lim_{n\to\infty} c(n,\delta) = 0$.

The spread quantities $X_{\max}(n,\delta)$, $c(n,\delta)$ and $C(n,\delta)$ are distribution specific.

For example, for standard Normals, $X_{\max}(n,\delta) = \sqrt{2\log n} - \frac{\log\log n + \log(4\pi)}{\sqrt{2\log n}}$, $c(n,\delta) = \frac{\log\log(6/\delta)}{\sqrt{2\log n}}$ and $C(n,\delta) = \frac{\log(6/\delta)}{\sqrt{2\log n}}$. For standard uniforms, $X_{\max}(n,\delta) = 1$, $C(n,\delta) = 0$ and $c(n,\delta) = \frac{\log(1/\delta)}{n}$.

Using these tools, we now characterize two kinds of classification behavior below. These correspond to the Gumbel and Weibull distribution families, as described in the Fisher-Tippett-Gnedenko theorem.

4.1. Low dimensional Gumbel type

For the Gumbel type, the maximum $M_n = \max(x_1, \ldots, x_n)$ of n iid examples converges to $a_n Z + b_n$, where a_n and b_n are distribution-specific quantities and Z is Gumbel-distributed. For these distributions, $X_{\max}(p, \delta)$ is typically higher than $X_{\max}(m, \delta)$ when $p \gg m$, producing a lack of symmetry in the maximum-margin classifier, and its increasing worst-class error. In this case, throwing away data to balance classes restores the symmetry and helps recover the lost worst-class error. This is formalized below.

Theorem 2. *[General Gumbel Distributions] Let D be a distribution of the Gumbel type with cumulative density function F and tail function $U(\cdot)$ and constants a_n and b_n in the Fisher-Tippett-Gnedenko theorem. Let $\lambda = \frac{\max(a_m, a_p)\log(3/\delta)}{U(p)-U(m)}$, and let $\mu = U(p) + a_p \log(3/\delta)$. Then, for large enough m and n, with probability $\geq 1 - 2\delta$ over the training samples, we have:*

$$\begin{aligned} \theta_{erm} &\geq \frac{1}{2}(U(p) - U(m))(1 - \lambda), \\ |\theta_{ss}| &\leq \frac{1}{2}\lambda(U(p) - U(m)) \end{aligned}$$

In addition, the worst-class errors satisfy:

$$\begin{aligned} \mathbf{wce}(\theta_{erm}) &\geq 1 - F\left(\frac{U(p)(1+3\lambda) + U(m)(1-3\lambda)}{2}\right) \\ \mathbf{wce}(\theta_{ss}) &\leq 1 - F(U(p)(1-\lambda)). \end{aligned}$$

A few remarks are in order. First, ensuring that the training data is linearly separable requires μ to grow with p. Second, observe that usually we expect $\lambda \leq 1$, and it may even be $o(1)$ for certain growth rates of m and p. When this is the case, the theorem implies that $|\theta_{\text{ss}}|$, which is of the order of $\lambda(U(p) - U(m))$ is an order of magnitude closer to the origin than θ_{erm}, which is $\approx -\frac{1}{2}(U(p) - U(m))$. This, in turn, contributes to θ_{ss} smaller worst-class-error. If $\lambda = o(1)$, this worst-class-error would be $\approx 1 - F(U(p)) \approx \frac{1}{p}$. In contrast, the worst-class-error of θ_{erm} would approach $\approx 1 - F(\frac{U(p)+U(m)}{2})$, which is somewhere between $\frac{1}{m}$ and $\frac{1}{p}$, depending on the exact form of F.

Third, observe that both worst-class error are tighter than a standard PAC-style analysis that would give a bound of $\approx 1/m$ on the worst-class error. Next, we present a corollary to tighten the previous result for the important Gaussian case.

Theorem 3. *Let $0 < \epsilon, \delta, \gamma < 1$ be constants and suppose $m = \beta p$. There exists an p_0 such that the following holds. If $p \geq p_0$ and $\beta \geq 1/p^{3/4}$, then with probability greater or equal than $1 - 2\epsilon - 2\delta - 3\gamma$:*

$$\begin{aligned} |\theta_{erm}| &\geq \frac{1}{2\sqrt{2\log(\beta p)}}\Big(\frac{2}{3}\log(1/\beta) - 2\log(1/\gamma)\Big), \\ |\theta_{ss}| &\leq \frac{\log(1/\gamma)}{2\sqrt{2\log(\beta p)}}. \end{aligned}$$

When $p\beta^2 \geq \epsilon$, this implies:

$$\mathbf{wce}(\theta_{ss}) \leq \frac{2\epsilon}{\gamma p}, \quad \mathbf{wce}(\theta_{erm}) \geq \frac{\epsilon\gamma^{1/4}}{2p\beta^{1/12}}.$$

For Gaussians, if $\beta \to 0$ with $\beta p \to \infty$, the relative gap between θ_{ss} and θ_{erm} widens – θ_{ss} lies in an interval of length $\approx \frac{1}{2\sqrt{2\log(\beta p)}}$ around the origin, while θ_{erm} lies $\approx \frac{\log(1/\beta)}{2\sqrt{2\log(\beta p)}}$ away. This leads to a widening of the worst-class error between the two SVM solutions, with θ_{ss} having considerably lower error than θ_{erm}.

4.2. Low dimensional Weibull type

Recall that for these distributions, the extremal point of the distribution is finite, and the maximum M_n converges to $a_n Z + b_n$, where a_n and b_n are distribution-specific quantities and Z is a reverse Weibull random variable with parameter α. For these distributions, $X_{\max}(p, \delta) \approx X_{\max}(m, \delta)$ even when $p \gg m$, and hence the maximum-margin classifier remains symmetric even when the majority class size $p \gg m$. This means that ERM and subsampling perform equally well. The following theorem formalizes the result.

Theorem 4. *[General Weibull Distributions] Suppose D is a distribution of the Weibull-type with parameter α and extremal point x_F. Let $\mu = x_F$, and let $m, p \to \infty$. Then, for any $0 < \delta \leq 1/4$, with probability $\geq \frac{1}{4 \cdot 2^{2\alpha}} - \delta$,*

$$\begin{aligned} |\theta_{ss}| &\geq \frac{1}{2}(x_F - U(m))(\ln 2)^{1/\alpha} \\ |\theta_{erm}| &\leq \frac{1}{2}(x_F - U(m))(\ln 2)^{1/\alpha}. \end{aligned}$$

Once again, we pause for some remarks. First, we observe that the extremal value of the Weibull distribution is finite, unlike Gumbel, ensuring that training data is linearly separable only requires $\mu = x_F$; the distributions themselves therefore do not change with p. Second, observe that our theorem states that with constant probability, the $|\theta_{\text{erm}}|$ is lower than $|\theta_{\text{ss}}|$ and so is the worst-class error. This implies that in the Weibull case, we cannot hope to get a high probability theorem such as Theorem 2. A final remark is the dependence of the lower bound on $|\theta_{\text{ss}}|$ on the parameter α of the Weibull distribution; unfortunately, this dependence is inevitable, since the concentration properties of the difference between two Weibull random variables depend on α. The following corollary makes the result concrete for uniform distributions.

Corollary 1. *Suppose $D(0)$ is the uniform distribution on $[-1/2, 1/2]$. Let $\mu = 1/2$, and let $m, p \to \infty$. Then, for any $0 < \delta \leq 1/4$, we have that for m and p large enough, with probability $\geq \frac{1}{16} - \delta$,*

$$|\theta_{ss}| \geq \frac{\ln 2}{2m}, |\theta_{erm}| \leq \frac{\ln 2}{2m}.$$

This implies that with probability $\geq \frac{1}{16} - \delta$,

$$\mathbf{wce}(\theta_{ss}) \geq \frac{\ln 2}{m} \geq \mathbf{wce}(\theta_{erm}).$$

4.3. High dimensional case

We next look at a higher dimensional case where the feature vector $x \in \mathbb{R}^d$. Our basic setup is as follows. As in the previous section, we assume that the class conditional distribution for each class is spherically symmetric around its mean μ; points x from class y follow $D(y\mu)$. The classifier that minimizes worst-class error for this setting is $\theta^*(x) = \mu^\top x$. Additionally, symmetry of the classes ensures that any linear classifier that passes through the origin will have equal error on each class. Thus showing θ_{erm} has high worst-class error involves showing that it has a non-zero bias term. This, in turn, will happen when the tails of the class-conditional distributions "spread out" with increasing sample size. This is formalized by the following high-dimensional concentration condition similar to the low dimensional cases. Notice that the difference here is that the concentration applies to all directions, and that the terms depend on the dimension d in addition to n and δ.

Assumption 2 (Concentration Condition)**.** *Suppose $x_1, \ldots, x_n$ are d dimensional random variables drawn i.i.d from $D(0)$. There exist maps $X_{\max} : \mathbb{Z}_+ \times [0,1] \times \mathbb{Z}_+ \to \mathbb{R}$, $c : \mathbb{Z}_+ \times [0,1] \times \mathbb{Z}_+ \to \mathbb{R}$, and $C : \mathbb{Z}_+ \times [0,1] \times \mathbb{Z}_+ \to \mathbb{R}$ such that for all $n \geq n_0$, for every $\delta \in (0,1)$, and for all directions $v \in \mathbb{R}^d$, with probability $\geq 1 - \delta$,*

$$\max_{i \in \{1 \cdots, n\}} \{v^\top x_i\} \in \big[X_{\max}(n, \delta, d) - c(n, d, \delta), \\ X_{\max}(n, \delta, d) + C(n, \delta, d)\big].$$

Also, $\lim_{n \to \infty \infty} C(n, \delta) = 0$ and $\lim_{n \to \infty} c(n, \delta) = 0$.

Unlike low dimensions, our high dimensional analysis requires one more technical condition to bound the s-th order statistic from the distribution.

Assumption 3. *Let $\zeta : \mathbb{Z}_+ \times \mathbb{R} \times \mathbb{Z}_+ \times \mathbb{Z}_+ \to \mathbb{R}$ be a map such that for all $n \geq n_0$ with probability at least $1 - \delta$ and*

$$\hat{\mu}^\top x^{(s)} \geq \zeta(n, \delta, d, s)$$

where $\hat{\mu}$ is the unit vector along μ, $\hat{\mu}^\top x^{(s)}$ is s^{th} largest value among n iid values of $\hat{\mu}^\top x$, where $x \sim D(0)$.

Define $q = d \log(\log p \max_{x_i \in A} \|x_i\|) + \log(1/\delta)$. We are now ready to state the main result.

Theorem 5. *$D(0)$ satisfies the concentration conditions in Assumption 2 and 3. Suppose $\|\mu\| > X_{\max}(p, \delta, d)$ and $\zeta(p, \delta, d, q) > 4X_{\max}(m, \delta, d)$. If p and m are sufficiently large, then with probability at least $1 - 8\delta$, the worst class error rate achieved by ERM is worse than the worst class error achieved by subsampling the classes.*

We pause for a few remarks. We require $\|\mu\| > X_{\max}(p, \delta, d)$ and $\zeta(p, \delta, d, q) > 4X_{\max}(m, \delta, d)$ to ensure that the direction of the classifier learned by ERM is sufficiently aligned with the direction ($\hat{\mu}$) of the classifier that achieves optimal worst class-error. As a result, we only need to compare the bias term between the subsampling and ERM. Since $\zeta(p, \delta, d, q) > 4X_{\max}(m, \delta, d)$ it ensures that p is sufficiently larger than m, which causes the bias under ERM to be large. Also, $\|\mu\| > X_{\max}(p, \delta, d)$ ensures that the two classes are separable. We now illustrate the above theorem for Gaussians and uniform distribution.

Corollary 2. *Let $D(0)$ be a symmetric Gaussian in $\mathbb{R}^2$. The bias for the ERM classifier θ_{erm} lies in an arbitrarily small interval centered at $\sqrt{2\log(p/\delta)} - \sqrt{2\log(m/\delta)}$. In contrast, the bias for the classifier under subsampling θ_{ss} lies in an arbitrarily small interval centerd at zero. If $m = \log p$ and $\|\mu\| > \sqrt{2\log(p/\delta)}$, then with probability $1 - 8\delta$, ERM has a worse worst class error than subsampling. Let $D(0)$ be a symmetric uniform in $\mathbb{R}^2$. The bias term for both the ERM classifier and the subsampling classifier lies in an arbitrarily small interval centered at zero.*

The above corollary shows how the bias for Gaussian is much larger than in the uniform distribution. As a result, subsampling is guaranteed to help the Gaussians but does not help uniform distributions as shown in Figure 1.

5. Analysis of imbalanced groups

We now look at data from imbalanced groups. Recall our basic setup, where label y and the attribute a induce four groups $g = (y, a)$ that a data point belongs to. To keep our analysis simple, we assume that each of the four groups have the same distribution $D(\cdot)$, but with shifted means, and that $D(\cdot)$ is spherically symmetric about its mean. Specifically, this means that for a group $g = (y, a)$, the group conditional distribution is $D(y\mu + a\psi)$, where μ and ψ are vectors in $\mathbb{R}^d$ with a non-zero norm. Also, recall that we have p points from each of the majority groups $(1, 1)$ and $(-1, -1)$ and m from the minority groups $(1, -1)$ and $(-1, 1)$ with $p \gg m$.

We start with a simple two-dimensional case where each feature vector $x_i \in \mathbb{R}^2$. We set the parameter vector $\mu = \|\mu\|(1, 0)^\top$ and $\psi = \|\psi\|(0, 1)^\top$. As a result, the first coordinate of x is aligned with the label y: $\mathbb{E}[x_1|y] = y\|\mu\|$. Following Nagarajan et al. (2020a), we

call this the *invariant feature*. The corresponding classifier $\theta^*_{\text{inv}} = (w^*_{\text{inv}} = (1,0)^\top, b^*_{\text{inv}} = 0)$ is the ideal linear classifier that achieves the best worst-group accuracy (Theorem 10). We call this the *invariant classifier*. The second coordinate of x is aligned with the attribute a and not the label y: $\mathbb{E}[x_2|a] = a\|\psi\|$. The imbalance in sampling rates of the different groups can cause the SVM classifier to have a component along this coordinate, even though it is trained to target the label y. We call this coordinate the *spurious feature*. If the minority group were entirely absent, the ideal classifier would involve the spurious feature, and would have the weight vector $w^*_{\text{spu}} = \frac{\|\mu\|}{\sqrt{\|\mu\|^2+\|\psi\|^2}}(1,0)^\top + \frac{\|\psi\|}{\sqrt{\|\mu\|^2+\|\psi\|^2}}(0,1)^\top$ and bias 0. We call this the *spurious classifier*. Next we show that the max-margin classifier converges to either the invariant or spurious classifier depending on the tail properties of the group conditional distributions.

5.1. Low dimensional Gumbel type

Finally, we are ready to state our main results. First, we look at the case analogous to the Gumbel types in Section 4, where the distribution tails "spread out" with more data.

Theorem 6. *Suppose $D(0)$ satisfies the concentration condition in Assumption 2 and $X_{\max}(p,\delta,2) - X_{\max}(m,\delta,2) \geq 2\|\psi\| + c(p,\delta,2) + C(m,\delta,2)$. If $p \to \infty$, then with probability at least $1-4\delta$, the ERM solution converges to the spurious solution w^*_{spu}. In addition, with probability at least $1-12\delta$, $\mathbf{wge}(\theta_{ss}) < \mathbf{wge}(\theta_{erm})$.*

First, observe that bounding $X_{\max}(p,\delta,2) - X_{\max}(m,\delta,2)$ from below essentially means that the maximum of p samples is considerably larger than the maximum of m samples when $p \gg m$; this is the *spreading out* condition in Figure 1. The first part of the theorem thus says that ERM has poor worst-group error in this case, while the second part shows that throwing away data through subsampling helps. Second, observe that unlike Section 4, where the bias term in the max-margin solution changes with the tail properties, for imbalanced groups, it is the direction of the max-margin classifier that changes (as illustrated in Figure 1).

We now make this result concrete when the group-conditional distributions are symmetric Gaussians centered at zero in $\mathbb{R}^2$. Let $\|\mu\| = \sqrt{3\log\frac{p}{\delta}}$ and $\|\psi\| = \sqrt{\kappa/4\log\frac{p}{\delta}}$. The ratio of the weight associated with the spurious feature to the invariant feature in w^*_{spu} is $\sqrt{\frac{\kappa}{12}}$. Let $m = p^\tau$, where $\tau < 1$. In this case, since $\|\mu\| > X_{\max}(p,\delta,2)$, the two classes are linearly separable. If $\kappa < 2\big(1+\tau-2\sqrt{\tau}\big)$, then the condition in the above theorem is satisfied and thus we can conclude that for this family of Gaussians the max-margin solution converges to the spurious solution. Let us contrast this with uniform $D(0)$, which has no tails. Since $X_{\max}(p,\delta,2) = X_{\max}(m,\delta,2) = 1$ the condition in the above theorem is not satisfied for uniform distribution.

5.2. Low dimensional Weibull type

We next look at our analogue of the Weibull case in Section 4, where the tails of the group-conditional distributions grow slowly with more data. In this case, we show that the max-margin solution converges to the invariant classifier that achieves the optimal worst-group error.

Theorem 7. *Suppose $D(0)$ satisfies the concentration condition in Assumption 2 and as p and m approach ∞*

$$\frac{X_{\max}(p,\delta,2) - X_{\max}(m,\delta,2)}{2\|\psi\|} \to 0.$$

If $m,p \to \infty$, then with probability at least $1-4\delta$, the ERM solution converges to the invariant solution w^*_{inv}.

Some remarks are in order. Observe that this theorem requires that the difference between the tails $X_{\max}(p,\delta,2) - X_{\max}(m,\delta,2)$ shrinks as p and m go to infinity. A concrete example where Theorem 7 applies is symmetric uniform $D(0)$, where $X_{\max}(p,\delta,2) = X_{\max}(m,\delta,2) = 1$.

5.3. Higher dimensional case

Moving on to higher dimensions, we again consider a setup where the feature vectors $x \in \mathbb{R}^d$. We assume that the group-conditional distributions D have the same form but with shifted means, and are spherically symmetric around their means. We select the first unit vector $e_1 = (1,0,\ldots)$ as the invariant feature, and the second one $e_2 = (0,1,0,\ldots)$ as the spurious feature. Thus the group-conditional distribution for points in group $g = (y,a)$ is $D(y\|\mu\|e_1 + a\|\psi\|e_2)$. As earlier in the Section 5, we can similarly define the invariant classifier and the spurious classifiers. The main results here, which we state below, mirror the theorems for the low dimensional cases.

Theorem 8. *[Gumbel type] Suppose $D(0)$ satisfies the concentration condition in Assumption 2 and $X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d) \geq 2\|\psi\| + c(p,\delta,d) + C(m,\delta,d)$. If $p \to \infty$, then with probability at least $1-4\delta$, the ERM solution converges to the spurious solution w^*_{spu}. In addition, with probability at least $1-12\delta$, $\mathbf{wge}(\theta_{ss}) < \mathbf{wge}(\theta_{erm})$.*

Theorem 9. *[Weibull type] Suppose $D(0)$ satisfies the concentration condition in Assumption 2 and as p and $m \to \infty$*

$$\frac{X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d)}{2\|\psi\|} \to 0,$$

$$\frac{X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d)}{2\|\mu\|} \to 0,$$

and $\frac{C(m,\delta,d)+c(m,\delta,d)+\frac{1}{2}\sqrt{2C(m,\delta,d)+c(m,\delta,d)\|\psi\|}}{\|\mu\|} \to 0$. *If $m,p \to \infty$, then with probability at least $1-4\delta$, the ERM solution converges to the invariant solution w^*_{inv}.*

Table 1: Performance of ERM on imbalanced and balanced datasets on high and low-dimensional versions of Waterbirds and CelebA. Worst group accuracy of ERM trained on balanced data substantially improves.

Dataset	Method	Avg. Accuracy	WG accuracy
Waterbirds	ERM	0.89 ± 0.00	0.63 ± 0.01
Waterbirds	SS	0.93 ± 0.00	0.88 ± 0.01
Waterbirds	ERM-PCA	0.88 ± 0.01	0.66 ± 0.03
Waterbirds	SS-PCA	0.94 ± 0.01	0.89 ± 0.01
CelebA	ERM	0.95 ± 0.00	0.36 ± 0.05
CelebA	SS	0.91 ± 0.00	0.83 ± 0.01
CelebA	ERM-PCA	0.95 ± 0.00	0.40 ± 0.01
CelebA	SS-PCA	0.90 ± 0.00	0.83 ± 0.01

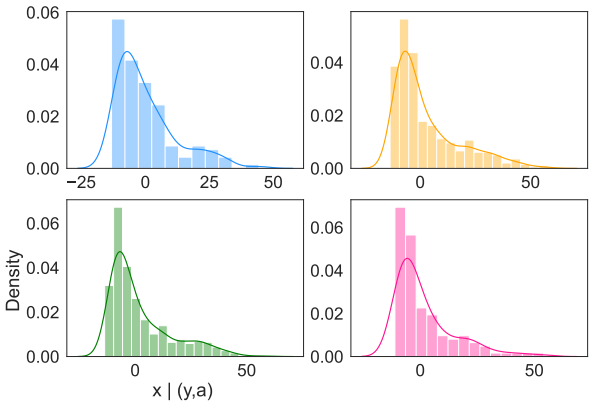

Figure 2: Waterbirds: Distribution of the top features.

6. Empirical Implications

We next investigate the empirical implications of the proposed theory. Specifically, we ask:

• Our theory is developed for linear SVMs. Does it bear relevance to neural networks trained on imbalanced data?

• Our theory is most applicable in low to moderate dimensions. Does throwing away data improve the worst group error in real data when applied to the top few features?

• What do the tails of the top feature distributions look like?

Datasets & Baselines. These questions are considered in the context of Waterbirds (Sagawa et al., 2019) and CelebA (Liu et al., 2015), the two most commonly used datasets for studying group imbalance. The Waterbirds data consists of two target classes – Waterbirds and Landbirds, and two background types – Water and Land. Most waterbirds appear on water, and most landbirds appear on land. In CelebA data, the target is to predict hair type – Blond or Non-Blond, where the frequency of blond women is much higher than blond men. Our ERM baseline consists of a ImageNet-pretrained ResNet-50 model that is finetuned on Waterbirds (4795 data points) and CelebA datasets (162770 data points) respectively. To understand the impact of dimensionality reduction, we compare this with the ERM-PCA baseline, that takes the PCA of the last layer (2048 dimension) of ResNet-50 model and trains a linear classifier on the first four principal components that explain $\approx 99\%$ of the variance in the data. Kirichenko et al. (2022) showed that if we freeze the fine-tuned representations and retrain just the last linear layer on balanced data obtained by subsampling that suffices to improve the worst group error. We call this method SS. We compare it with SS-PCA that trains a linear layer on balanced four-dimensional data (top four PCA components).

Results. Table 1 shows the results. We see that as expected for the high dimensional data, ERM performs much worse in terms of worst-group accuracy (WG accuracy) than SS. This confirms the findings of prior work (Kirichenko et al., 2022; Idrissi et al., 2022). We also see that the same pattern holds for ERM-PCA and SS-PCA. This confirms that throwing away data improve the worst group error in real data when applied to the top few features. In Figure 2, we visualize the tails of the top PCA feature for Waterbirds. Specifically, we plot a histogram of data from each group projected along the feature with the highest PCA value. The results on other features and CelebA are plotted in the Appendix. The results show that the groups are indeed long-tailed, in the sense that they do not look like the uniform distribution. This suggests that the theoretical phenomenon that we describe in this paper might contribute to the success of subsampling.

7. Discussion

The vast majority of learning theory literature including works carried out in the context of multi group analysis have focussed on measuring concentrations of bounded functions (Rothblum & Yona, 2021; Tosh & Hsu, 2022; Haghtalab et al., 2022; Sagawa et al., 2019; 2020), and tail properties of distributions feature rarely. A handful of papers have looked at designing algorithms with performance guarantees for tasks carried out on heavy-tailed distributions. For example, (Hsu & Sabato, 2016) proposes a regression algorithm based on the median-of-means estimator that concentrates well under heavy-tailed distributions. In contrast, we analyze algorithms under distributions with different kinds of tail properties. Other works on theory of imbalanced classification have focused on metrics other than accuracy (Menon et al., 2013; Natarajan et al., 2017), such as precision and recall (Diochnos & Trafalis, 2021). An example is (Narasimhan et al., 2015), which proposes

a new Bayes Optimal classifier for different functions of the confusion matrix when there are imbalanced classes, together with consistency guarantees. However, the bounds in these works are coarser than ours, as they do not reflect changing behavior depending on the tail distributions of the classes.

Our work is also related to Nagarajan et al. (2020b), which analyzes max-margin classifiers to explain the failure of of ERM under group imbalance. Our work complements their findings. While the authors derive a lower bound on the weight associated with the spurious feature, they do not specify conditions on the distribution under which this bound is positive, which is crucial towards explaining when the model relies on spurious features. We fill this gap as we provide a characterization of the max-margin classifiers in terms of the tails of the distribution. To the best of our knowledge, this is a first characterization of the SVM classifiers as a function of the tails of the distribution. Therefore, we believe the proof techniques developed here can be of independent interest. Looking forward, we believe that while data balancing is powerful it still requires access to the knowledge of spurious attributes. Therefore, it is important to formalize what is achievable in the absence of such knowledge.

References

Bennett, K. P. and Bredensteiner, E. J. Duality and geometry in svm classifiers. In *ICML*, volume 2000, pp. 57–64. Citeseer, 2000.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

De Haan, L. and Ferreira, A. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.

De Haan, L., Ferreira, A., and Ferreira, A. *Extreme value theory: an introduction*, volume 21. Springer, 2006.

Diochnos, D. I. and Trafalis, T. B. Learning reliable rules under class imbalance. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 28–36. SIAM, 2021.

Feller, W. *An Introduction to Probability Theory and Its Applications*. Wiley, 1968.

Haghtalab, N., Jordan, M. I., and Zhao, E. On-demand sampling: Learning optimally from multiple distributions. *arXiv preprint arXiv:2210.12529*, 2022.

Haussler, D. *Probably approximately correct learning*. University of California, 1990.

Hsu, D. and Sabato, S. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.

Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. *CLeaR*, 2021.

Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. *CLeaR*, 2022. URL https://arxiv.org/abs/2110.14503.

Johnson, J. M. and Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1): 1–54, 2019.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pp. 1302–1338, 2000.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Menon, A., Narasimhan, H., Agarwal, S., and Chawla, S. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, pp. 603–611. PMLR, 2013.

Nagarajan, V., Andreassen, A., and Neyshabur, B. Understanding the failure modes of out-of-distribution generalization. *arXiv*, 2020a.

Nagarajan, V., Andreassen, A., and Neyshabur, B. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020b.

Narasimhan, H., Ramaswamy, H., Saha, A., and Agarwal, S. Consistent multiclass algorithms for complex performance measures. In *International Conference on Machine Learning*, pp. 2398–2407. PMLR, 2015.

Natarajan, N., Dhillon, I. S., Ravikumar, P., and Tewari, A. Cost-sensitive learning with noisy labels. *J. Mach. Learn. Res.*, 18(1):5666–5698, 2017.

Rojas, W. A. G., Diamos, S., Kini, K. R., Kanter, D., Reddi, V. J., and Coleman, C. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=qnfYsave0U4.

Rothblum, G. N. and Yona, G. Multi-group agnostic pac learnability. In *International Conference on Machine Learning*, pp. 9107–9115. PMLR, 2021.

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

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *ICML*, 2020.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Tosh, C. J. and Hsu, D. Simple and near-optimal algorithms for hidden stratification and multi-group learning. In *International Conference on Machine Learning*, pp. 21633–21657. PMLR, 2022.

Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 1999.

Why Throwing Away Data Improves Worst-Group Error?

Appendices

We organize this section as follows.

- In Appendix A, we derive the results for imbalanced classification.
  - In Appendix A.1, we derive results for imbalanced classification in the one-dimensional setting.
  - In Appendix A.2, we derive results for imbalanced classification in higher dimensions.
- In Appendix B, we derive the results for classification with imbalanced groups.
  - In Appendix B.1, we derive the results for imbalanced groups in two-dimensional setting.
  - In Appendix B.2, we derive the results for imbalanced groups in higher-dimensional setting.
- In Appendix C, we present the supplementary materials for the empirical findings.

A. Proofs for imbalanced classes

A.1. One-dimensional case

Theorem 2. *[General Gumbel Distributions] Let D be a distribution of the Gumbel type with cumulative density function F and tail function $U(\cdot)$ and constants a_n and b_n in the Fisher-Tippett-Gnedenko theorem. Let $\lambda = \frac{\max(a_m, a_p)\log(3/\delta)}{U(p)-U(m)}$, and let $\mu = U(p) + a_p \log(3/\delta)$. Then, for large enough m and n, with probability $\geq 1 - 2\delta$ over the training samples, we have:*

$$\begin{aligned} \theta_{erm} &\geq \frac{1}{2}(U(p) - U(m))(1-\lambda), \\ |\theta_{ss}| &\leq \frac{1}{2}\lambda(U(p) - U(m)) \end{aligned}$$

In addition, the worst-class errors satisfy:

$$\begin{aligned} \mathbf{wce}(\theta_{erm}) &\geq 1 - F\left(\frac{U(p)(1+3\lambda) + U(m)(1-3\lambda)}{2}\right) \\ \mathbf{wce}(\theta_{ss}) &\leq 1 - F(U(p)(1-\lambda)). \end{aligned}$$

Proof. Let M_m denote the maximum of m random variables drawn from $D(0)$. The bias term is the mean of maximum negative point and minimum positive training point. In the case of subsampled data, where both classes have the same number of datapoints, the solution to the bias term is

$$\theta_{\text{ss}} = \frac{1}{2}(M_m - M'_m),$$

From the Gumbel concentration lemma (Lemma 8), with probability $\geq 1 - \delta$, $b_m + a_m \log\log(3/\delta) \leq M_m \leq b_m + a_m \log(3/\delta)$. Therefore,

$$|\theta_{\text{ss}}| \leq \frac{a_m}{2}\log(3/\delta) \leq \lambda(U(p) - U(m)),$$

where the second step follows from the definition of λ.

Similarly, we can show that:

$$\theta_{\text{erm}} = \frac{1}{2}(M_p - M_m),$$

where M_p and M_m are the maxima of p and m independent draws from $D(0)$. From the Fisher-Tippett-Gnedenko theorem, $M_n = b_n + a_n Z$ where Z is an unit Gumbel random variable, and $b_n = U(n)$. Therefore,

$$\theta_{\text{erm}} = \frac{1}{2}(U(p) - U(m) + a_p Z - a_m Z'),$$

where Z and Z' are independent Gumbel variables. This is at least:

$$\theta_{\text{erm}} \geq \frac{1}{2}(U(p) - U(m) - \max\{a_m, a_p\}\log(3/\delta)) \geq \frac{1}{2}(U(p) - U(m))(1-\lambda),$$

where the last step follows from the definition of λ. The first part of the lemma follows.

For the second part of the lemma, let $\Phi(x) = 1 - F(x)$.

Observe that

$$\begin{aligned}
\mathbf{wce}(\theta_{\text{ss}}) &\leq \Phi(\mu - |\theta_{\text{ss}}|) \\
&\leq \Phi(U(p) + a_p \log(3/\delta) - \frac{a_m}{2}\log(3/\delta)) \\
&\leq \Phi(U(p)(1-\lambda))
\end{aligned}$$

where the first step follows by definition of **wce** (worse class error is determined by error conditioned on data drawn from negative class). The second step follows by plugging in the values of μ_n and an upper bound on $|\theta_{\text{ss}}|$ and using the fact that $\Phi(x)$ is a decreasing function of x. The third step follows because $\frac{a_m}{2}\log(3/\delta) - a_p \log(3/\delta) \leq \lambda U(p)$, and because $\Phi(x)$ is a decreasing function of x.

In contrast,

$$\begin{aligned}
\mathbf{wce}(\theta_{\text{erm}}) &\geq \max(\Phi(\mu - \theta_{\text{erm}}), \Phi(\mu + \theta_{\text{erm}})) \\
&\geq \Phi(U(p) + a_n \log(3/\delta) - \frac{1}{2}(U(p) - U(m))(1-\lambda)) \\
&\geq \Phi\left(\frac{U(p)(1+3\lambda) + U(m)(1-3\lambda)}{2}\right),
\end{aligned}$$

where the first step follows from the fact that Φ is a decreasing function, the second step by plugging in the value of μ, and the third step from the observation that $\mu_n \leq U(p)(1+\lambda)$. The second part of the theorem follows. □

Lemma 1. *Let Z_3 and Z_4 be two independent standard Gumbel variables. Then with probability $\geq 1 - \frac{2}{1+e^\tau}$, $|Z_3 - Z_4| \leq \tau$.*

Proof. Since Z_3 and Z_4 are independent standard Gumbel variables, $Z_3 - Z_4$ is distributed according to a logistic distribution with location parameter 0 and scale parameter 1. This means that $Z_3 - Z_4$ is symmetric about 0, and also that for any τ,

$$\Pr(Z_3 - Z_4 \in [-\tau, \tau]) = 1 - \frac{2}{1+e^\tau} \tag{3}$$

The lemma follows. □

Lemma 2. *(Feller, 1968) Let $X \sim N(0,1)$. Then, for any t,*

$$\frac{1}{\sqrt{2\pi}} e^{-t^2/2}\left(\frac{1}{t} - \frac{1}{t^3}\right) \leq \Pr(X \geq t) \leq \frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t}$$

Theorem 3. *Let $0 < \epsilon, \delta, \gamma < 1$ be constants and suppose $m = \beta p$. There exists an p_0 such that the following holds. If $p \geq p_0$ and $\beta \geq 1/p^{3/4}$, then with probability greater or equal than $1 - 2\epsilon - 2\delta - 3\gamma$:*

$$\begin{aligned}
|\theta_{erm}| &\geq \frac{1}{2\sqrt{2\log(\beta p)}}\left(\frac{2}{3}\log(1/\beta) - 2\log(1/\gamma)\right), \\
|\theta_{ss}| &\leq \frac{\log(1/\gamma)}{2\sqrt{2\log(\beta p)}}.
\end{aligned}$$

When $p\beta^2 \geq \epsilon$, this implies:

$$\mathbf{wce}(\theta_{ss}) \leq \frac{2\epsilon}{\gamma p}, \quad \mathbf{wce}(\theta_{erm}) \geq \frac{\epsilon \gamma^{1/4}}{2p\beta^{1/12}}.$$

Proof. Fix ϵ, γ and δ. We first show that with the given value of μ_n, the probability that the training samples are realizable is at least $1-2\epsilon$. To see this, observe that from Lemma 2, the probability a sample z from the positive class has value < 0 is at most:

$$\Pr(X \geq \sqrt{2\log(n/\epsilon)}|X \sim N(0,1)) \leq \frac{e^{-\mu_n^2/2}}{\sqrt{2\pi}\mu_n} \leq \epsilon/n$$

An union bound over all n samples establishes that the probability that any of the n positive samples lie below zero is at most ϵ. A similar argument can also be applies to the negative class to show that with probability $\geq 1-2\epsilon$, the training data is linearly separable.

To establish the first part of the theorem, we will separately look at θ_{ss} and θ_{erm}. For p large enough, we can write

$$\Pr\left(|\theta_{\text{ss}}| \geq \frac{1}{2}\cdot\frac{\log(1/\gamma)}{\sqrt{2\log(\beta p)}}\right) \leq \Pr\left(|Z_3 - Z_4| \geq \log(1/\gamma)\right) + \delta$$

From Lemma 1, this probability is at most $\delta + \frac{2}{1+(1/\gamma)} \leq \delta + 2\gamma$.

Similarly, for p large enough, we have that:

$$\theta_{\text{erm}} \to \frac{1}{2}(b_{\beta p} - b_p) + \frac{1}{2}a_{\beta p}(Z_1 - Z_2) + \frac{1}{2}(a_p - a_{\beta p})Z_2$$

where Z_1 and Z_2 are standard Gumbel random variables. This means that for p large enough, and for any threshold τ, we have that:

$$\Pr(\theta_{\text{erm}} \leq \tau) \quad \leq \quad \Pr(\frac{1}{2}(b_{\beta p} - b_p) + \frac{1}{2}a_{\beta p}(Z_1 - Z_2) + \frac{1}{2}(a_p - a_{\beta p})Z_2 \leq \tau) + \delta \tag{4}$$

Now, observe that for Gaussians:

$$\begin{aligned}
b_{\beta p} - b_p &= \sqrt{2\log\beta p} - \frac{\log\log(\beta p) + \log 4\pi}{\sqrt{2\log\beta p}} - \sqrt{2\log p} + \frac{\log\log p + \log 4\pi}{\sqrt{2\log p}} \\
&\leq \sqrt{2\log\beta p} - \sqrt{2\log p} \\
&= \sqrt{2\log\beta p}(1 - \sqrt{\frac{\log p}{\log\beta p}}) \\
&= \sqrt{2\log\beta p}(1 - (1 + \frac{\log(1/\beta)}{\log\beta p})^{1/2}) \\
&\leq \sqrt{2\log\beta p}(-\frac{\log(1/\beta)}{3\log\beta p}) \\
&\leq -\frac{2\log(1/\beta)}{3\sqrt{2\log\beta p}}
\end{aligned} \tag{5}$$

Here the first step follows as $\frac{\log\log p + \log(4\pi)}{\sqrt{\log p}}$ is a decreasing function of p, and the second and third steps follow from algebra. The fourth step follows from the fact that $(1+x)^{1/2} \geq 1 + x/3$ when $x \leq 3$; as $p^3 \geq \beta^4$, $\frac{\log(1/\beta)}{\log\beta p} \leq 3$. The final step then follows from algebra. Additionally, from Lemma 1,

$$\Pr(a_{\beta p}(Z_1 - Z_2) \geq \frac{\log(1/\gamma)}{\sqrt{2\log\beta p}}) \leq \Pr(Z_1 - Z_2 \geq \log(1/\gamma)) \leq \frac{1}{1+1/\gamma} \leq \gamma$$

Finally, observe that

$$\begin{aligned}
a_{\beta p} - a_n &= \frac{1}{\sqrt{2\log\beta p}} - \frac{1}{\sqrt{2\log p}} \\
&= \frac{1}{\sqrt{2\log\beta p}}(1 - \sqrt{\frac{\log\beta p}{\log p}}) \\
&= \frac{1}{\sqrt{2\log\beta p}}(1 - (1 - \frac{\log(1/\beta)}{\log p})^{1/2}) \\
&\leq \frac{1}{\sqrt{2\log\beta p}}\cdot\frac{\log(1/\beta)}{\log p}
\end{aligned} \tag{6}$$

where the last step follows because for $0 < x < 1$, $\sqrt{1-x} \geq 1 - x$. Therefore,

$$\begin{aligned}
\Pr((a_{\beta p} - a_p)Z_2 \geq \frac{\log(1/\gamma)}{\sqrt{2\log\beta p}}) &\leq \Pr(Z_2 \geq \log(1/\gamma) \cdot \frac{\log p}{\log(1/\beta)}) \\
&\leq \Pr(Z_2 \geq \log(1/\gamma)) \\
&\leq 1 - e^{-\gamma} \leq \gamma
\end{aligned} \tag{7}$$

where the first step follows from Equation 6, the second step from the fact that $\log(1/\beta) \leq \log n$, and the final step from the standard Gumbel cdf and the fact that $1 - e^{-x} \leq x$. The first part of the theorem follows from combining Equations 5, 6 and 7.

To prove the third part of the theorem, we use Lemma 2. From the second part of the theorem and Lemma 2, observe that

$$\mathbf{err}(\theta_{\text{ss}}) \leq \Phi\Big(\mu - \frac{\log(1/\gamma)}{2\sqrt{2\log\beta p}}\Big)$$

From Lemma 2, the right hand side is, in turn, at most:

$$\begin{aligned}
&\leq \frac{1}{\sqrt{2\pi}(\mu - \frac{\log(1/\gamma)}{2\sqrt{2\log\beta p}})} \cdot \exp\Big(-\frac{1}{2}\Big(\mu - \frac{\log(1/\gamma)}{2\sqrt{2\log\beta p}}\Big)^2\Big) \\
&\leq \frac{2}{\sqrt{2\pi}\mu_n} \exp\Big(-\frac{1}{2}\mu^2 + \frac{\mu\log(1/\gamma)}{2\sqrt{2\log\beta n}}\Big) \\
&\leq \frac{2\epsilon}{n} \cdot \exp\Big(\frac{\mu\log(1/\gamma)}{2\sqrt{2\log\beta p}}\Big)
\end{aligned}$$

Here the first step follows because for p large enough, $\mu - \frac{\log(1/\gamma)}{2\sqrt{2\log\beta p}} \geq \frac{1}{2}\mu$; this is because μ is an increasing function of p. The second step follows because by design $\frac{1}{\sqrt{2\pi}\mu}e^{-\mu^2/2} = \epsilon/p$ and the third step follows from simple algebra.

We observe that $\mu \leq \sqrt{2\log(p/\epsilon)}$ – this is because $e^{-\sqrt{2\log(p/\epsilon)}^2/2}/\sqrt{4\pi\log(p/\epsilon)} < \epsilon/p$. This implies that:

$$\frac{\mu_n}{\sqrt{2\log\beta p}} \leq \Big(1 + \frac{\log(1/\epsilon\beta)}{\log\beta p}\Big)^{1/2} \leq 2,$$

provided $\frac{1}{\epsilon} \leq \beta^2 p$. Therefore,

$$\mathbf{err}(\theta_{\text{ss}}) \leq \frac{2\epsilon}{\gamma p}$$

In contrast, from Lemma 2,

$$\begin{aligned}
\mathbf{err}(\theta_{\text{erm}}) &\geq \Phi\Big(\mu - \frac{\frac{2}{3}\log(1/\beta) - 2\log(1/\gamma)}{2\sqrt{2\log\beta p}}\Big) \\
&= \Phi\Big(\mu - \frac{\log(\gamma^2/\beta^{2/3})}{2\sqrt{2\log\beta p}}\Big)
\end{aligned}$$

From Lemma 2, this is at least:

$$\frac{1}{\sqrt{2\pi}}\Big(\frac{1}{(\mu - \frac{\log(\gamma^2/\beta^{2/3})}{2\sqrt{2\log\beta p}})} - \frac{1}{(\mu - \frac{\log(\gamma^2/\beta^{2/3})}{2\sqrt{2\log\beta p}})^3}\Big) \cdot \exp(-\frac{1}{2}\Big((\mu - \frac{\log(\gamma^2/\beta^{2/3})}{2\sqrt{2\log\beta p}})^2)\Big)$$

For p large enough, $\frac{1}{(\mu - \frac{\log(\gamma^2/\beta^{2/3})}{2\sqrt{2\log\beta p}})^3} \leq \frac{1}{2}\frac{1}{(\mu - \frac{\log(\gamma^2/\beta^{2/3})}{2\sqrt{2\log\beta p}})}$; also $\frac{1}{(\mu_n - \frac{\log(\gamma^2/\beta^{2/3})}{2\sqrt{2\log\beta p}})} \geq \frac{1}{\mu}$. This implies that the right hand side is at least:

$$\frac{1}{\sqrt{2\pi}} \cdot \frac{1}{2\mu} \cdot \exp(-\frac{1}{2}\Big((\mu - \frac{\log(\gamma^2/\beta^{2/3})}{2\sqrt{2\log\beta p}})^2)\Big)$$

Observe that:

$$\begin{aligned}
& \frac{1}{2\sqrt{2\pi}\mu} \cdot \exp(-\frac{1}{2}\Big((\mu - \frac{\log(\gamma^2/\beta^{2/3})}{2\sqrt{2\log\beta p}})^2)\Big) \\
\geq\ & \frac{1}{2\sqrt{2\pi}\mu} \cdot \exp(-\frac{1}{2}\mu^2) \cdot \exp(\frac{\mu\log(\gamma^2/\beta^{2/3})}{2\sqrt{2\log\beta p}} - (\frac{\log(\gamma^2/\beta^{2/3})}{2\sqrt{2\log\beta p}})^2) \\
\geq\ & \frac{1}{2} \cdot \frac{\epsilon}{p} \cdot \exp(\frac{1}{2}\frac{\mu\log(\gamma^2/\beta^{2/3})}{2\sqrt{2\log\beta p}})
\end{aligned}$$

where the first step follows since $\frac{1}{\sqrt{2\pi}\mu} \cdot e^{-\mu^2/2} = \epsilon/p$ and the second step follows since for large enough p, $\mu/2 \geq \frac{\log(\gamma^2/\beta^{2/3})}{2\sqrt{2\log\beta p}}$.

Also: observe that for p large enough $\frac{\mu}{\sqrt{2\log\beta p}} \geq \frac{1}{2}$ – since $\mu \geq \frac{1}{2}\sqrt{2\log(p/\epsilon)}$. Putting these all together, the entire error is at least:

$$\frac{\epsilon}{2p}\exp(\frac{1}{8}\log(\gamma^2/\beta^{2/3})) \geq \frac{\epsilon\gamma^{1/4}}{2p\beta^{1/12}}$$

The theorem now follows. □

Theorem 4. *[General Weibull Distributions] Suppose D is a distribution of the Weibull-type with parameter α and extremal point x_F. Let $\mu = x_F$, and let $m, p \to \infty$. Then, for any $0 < \delta \leq 1/4$, with probability $\geq \frac{1}{4\cdot 2^{2^\alpha}} - \delta$,*

$$\begin{aligned}
|\theta_{ss}| &\geq \frac{1}{2}(x_F - U(m))(\ln 2)^{1/\alpha} \\
|\theta_{erm}| &\leq \frac{1}{2}(x_F - U(m))(\ln 2)^{1/\alpha}.
\end{aligned}$$

Proof. Let M_m denote the maximum of m random variables drawn from $D(0)$. We observe that

$$\theta_{\text{ss}} = \frac{1}{2}(M_m - M'_m),$$

which for the Weibull case converges to $\frac{a_m}{2}(W - W') = \frac{a_m}{2}(Z' - Z)$; here, W and W' are reverse Weibull random variables with parameters α and 1, which makes $Z \sim \textbf{Weibull}(\alpha, 1)$ and $Z' \sim \textbf{Weibull}(\alpha, 1)$ are independent Weibull variables, and $a_m = x_F - U(m)$. Combining this with Lemma 3, and accounting for the distributional convergence, we get that for m large enough,

$$\Pr(|\theta_{\text{ss}}| \geq \frac{(x_F - U(m))}{2}(\ln 2)^{1/\alpha}) \geq \frac{1}{2^{2^\alpha}} - \delta$$

Similarly,

$$\theta_{\text{erm}} = \frac{1}{2}(M_m - M_n),$$

which in the Weibull case converges to $\frac{a_m}{2}(Z' - Z)$; here $Z \sim \textbf{Weibull}(\alpha, 1)$ and $Z' \sim \textbf{Weibull}(\alpha, \frac{a_n}{a_m})$. Observe that as $U(\cdot)$ is an increasing function, $\frac{a_n}{a_m} = \frac{x_F - U(n)}{x_F - U(m)} \leq 1$. We can therefore apply Lemma 4 to conclude that for large enough m and n,

$$\Pr(|\theta_{\text{erm}}| \leq \frac{x_F - U(m)}{2}(\ln 2)^{1/\alpha}) \geq \frac{1}{4} - \delta$$

The theorem follows. □

Lemma 3. *Let $Z \sim \textbf{Weibull}(\alpha, 1)$ and $Z' \sim \textbf{Weibull}(\alpha, 1)$ be independent Weibull variables. Then,*

$$\Pr(|Z - Z'| \geq (\ln 2)^{1/\alpha}) \geq \frac{1}{2^{2^\alpha}}$$

Proof. For any constants a and b,

$$\Pr(|Z - Z'| \geq a) \geq 2\Pr(Z \geq a + b, Z' \leq b) = 2\Pr(Z \geq a + b)\Pr(Z' \leq b) = 2e^{-(a+b)^\alpha}(1 - e^{-b^\alpha}),$$

where the first step follows because Z and Z' come from the same distribution, the second step follows from the independence of Z and Z', and the third step from plugging in the CDF for the Weibull distribution. Next, we plug in $a = (\ln 2)^{1/k}$ and $b = (\ln 2)^{1/k}$ – and bound the final quantity as:

$$\geq 2 \cdot \frac{1}{2} \cdot \exp(-(2(\ln 2)^{1/\alpha})^\alpha) \geq \frac{1}{2^{2^\alpha}},$$

which is a constant for constant α. □

Lemma 4. *Let $\lambda < 1$ and let $Z \sim \textbf{Weibull}(\alpha, 1)$ and $Z' \sim \textbf{Weibull}(\alpha, \lambda)$ be independent Weibull variables. Then,*

$$\Pr(|Z - Z'| \leq \textit{median}(Z)) \geq \frac{1}{4}$$

Proof. Since both Z and Z' are positive random variables, $\Pr(|Z - Z'| \leq a) \geq \Pr(Z \leq a, Z' \leq a) = \Pr(Z \leq a)\Pr(Z' \leq a)$. Now, $\lambda < 1$, we can establish a coupling between Z and Z' to show that for any a, $\Pr(Z' \leq a) \geq \Pr(Z = a)$. The lemma follows by plugging this in, and setting $a = \text{median}(Z)$. □

Finally, the proof to Corollary 1 follows by substituting the expression for a_n and b_n for the Uniform distribution in Theorem 4.

A.2. Imbalanced Classes: Higher Dimensional Case

Denote $\hat{v}$ as a unit vector in the direction of the vector v. In the lemmas below, we prove some facts that are necessary to prove the main Theorem 5. $\hat{\mu}$ is a unit vector in the direction of μ (conditional mean of positive class) and $\hat{y}$ is a unit vector orthogonal to $\hat{\mu}$.

Lemma 5. *If $D(0)$ is spherically symmetric, then the median $\hat{y}^\top x$ conditioned on $\hat{\mu}^\top x$ is zero, i.e., Median$[\hat{y}^\top x | \hat{\mu}^\top x] = 0$, where $\hat{\mu} \perp \hat{y}$.*

Proof. To show the above lemma, we will first prove that the joint probability $\Pr(\hat{\mu}^\top x, \hat{y}^\top x) = \Pr(\hat{\mu}^\top x, \hat{z}^\top x)$, where $\hat{y}$ and $\hat{z}$ are any two unit vectors perpendicular to $\hat{\mu}$. Consider two orthonormal basis $\mathcal{B}_1$ and $\mathcal{B}_2$ to express vectors $x \in \mathbb{R}^d$. We write the coordinates of x in $\mathcal{B}_1$ as $\{\upsilon_1^1, \cdots, \upsilon_d^1\}$ and in $\mathcal{B}_2$ as $\{\upsilon_1^2, \cdots, \upsilon_d^2\}$. Since x is drawn from a spherically symmetric distribution $\Pr(\upsilon_1^1, \cdots, \upsilon_d^1) = \Pr(\upsilon_1^2, \cdots, \upsilon_d^2)$. As a result, $\Pr(\upsilon_1^1, \upsilon_2^1) = \Pr(\upsilon_1^2, \upsilon_2^2)$. Suppose the first two vectors in $\mathcal{B}_1$ are $\hat{\mu}$ and $\hat{y}$ and suppose the first two vectors in $\mathcal{B}_2$ are $\hat{\mu}$ and $\hat{z}$. Thus, $\upsilon_1^1 = \hat{\mu}^\top x$, $\upsilon_2^1 = \hat{y}^\top x$, and $\upsilon_1^2 = \hat{\mu}^\top x$, $\upsilon_2^2 = \hat{z}^\top x$. As a result, $\Pr(\hat{\mu}^\top x, \hat{y}^\top x) = \Pr(\hat{\mu}^\top x, \hat{z}^\top x)$. This implies

$$\Pr(\hat{y}^\top x \leq a | \hat{\mu}^\top x) = \Pr(\hat{z}^\top x \leq a | \hat{\mu}^\top x)$$

Substitute $\hat{z} = -\hat{y}$ to get

$$\Pr(\hat{y}^\top x \leq a | \hat{\mu}^\top x) = \Pr(-\hat{y}^\top x \leq a | \hat{\mu}^\top x)$$
$$\Pr(\hat{y}^\top x \leq a | \hat{\mu}^\top x) = \Pr(\hat{y}^\top x \geq -a | \hat{\mu}^\top x)$$

If $a = 0$, then $\Pr(\hat{y}^\top x \leq 0 | \hat{\mu}^\top x) = \Pr(\hat{y}^\top x \geq 0 | \hat{\mu}^\top x)$. This implies that the conditional median is zero. □

Lemma 6. *If $D(0)$ is spherically symmetric, then the median $\hat{y}^\top x^{(i)}$ conditioned on $\hat{\mu}^\top x^{(i)}$ is zero, i.e., Median$[\hat{y}^\top x^{(i)} | \hat{\mu}^\top x^{(i)}] = 0$, where $\hat{\mu} \perp \hat{y}$ and $x^{(i)}$ corresponds to the x in $\{x_1, \cdots, x_n\}$ with i^{th} largest projection on $\hat{\mu}$.*

Proof. For cleaner exposition, without loss of generality we will say x_j takes the jth highest projection on $\hat{\mu}$ for all $j \in \{1, \cdots, n\}$.

We write the joint probability over $\hat{y}^\top x_i$, and all $\hat{\mu}^\top x_j$ as

$$\Pr\left(\hat{y}^\top x_i, \hat{\mu}^\top x_i = u, \{\hat{\mu}^\top x_j \leq u, \forall j \leq i - 1\}, \{\hat{\mu}^\top x_k \geq u, \forall k \geq i + 1\}\right)$$

$$= \Pr(\hat{y}^\top x_i, \hat{\mu}^\top x_i = u) F_{\hat{\mu}}(u)^{i-1}(1 - F_{\hat{\mu}}(u))^{n-i} \tag{8}$$

where $F_{\hat{\mu}}$ is the CDF of x projected on $\hat{\mu}$.

The conditional probability simplifies as follows

$$\Pr(\hat{y}^\top x^{(i)} | \hat{\mu}^\top x^{(i)}) = \frac{\Pr(\hat{y}^\top x_i, \hat{\mu}^\top x_i = u) F_{\hat{\mu}}(u)^{i-1}(1 - F_{\hat{\mu}}(u))^{n-i}}{\Pr(\hat{\mu}^\top x_i = u) F_{\hat{\mu}}(u)^{i-1}(1 - F_{\hat{\mu}}(u))^{n-i}} = \Pr(\hat{y}^\top x_i | \hat{\mu}^\top x_i) \tag{9}$$

The rest of the lemma now follows from the previous Lemma 5. □

Lemma 7. *Suppose n iid samples $\{x_1, \cdots, x_n\}$ are sampled from a spherically symmetric $D(0)$. Let $s = d\log(2N/\epsilon + 1) + \log(1/\delta)$ where $N = \max_{i \in \{1,\cdots,n\}} \|x_i\|$. Then, with probability $\geq 1 - \delta$, for all directions $\hat{y}$ in $\mathbb{R}^d$ there exists an $i \in \{1, \ldots, s\}$ such that*

$$\hat{y}^\top x^{(i)} \geq -\epsilon$$

where $\{x^{(1)}, \cdots, x^{(n)}\}$ are in the decreasing order of their projection on $\hat{\mu}$, where $\hat{\mu} \perp \hat{y}$.

Proof. We select an ϵ/N cover $C = \{y_1, \ldots, y_M\}$ over the surface of the sphere; this is possible for $M = (\frac{2N}{\epsilon} + 1)^d$.[1] We will prove the lemma in two steps. First, we will show that with probability $\geq 1 - \delta$, for all $y_j \in C$, there exists a x_i in $x^{(1)}, \ldots, x^{(s)}$ for which $y_j^\top x_i \geq 0$. To show this, first let us fix a particular y_j, and consider $x^{(1)}, \ldots, x^{(s)}$. For any of these $x^{(i)}$'s, $\Pr(y_j^\top x^{(i)} \geq 0 | \hat{\mu}^\top x^{(i)}) = 1/2$ (from Lemma 6) and as a result $\Pr(y_j^\top x^{(i)} \geq 0) = 1/2$. This means that the probability that $y_j^\top x^{(i)} < 0$ for all $x^{(i)}$ in $x^{(1)}, \ldots, x^{(s)}$ is at most $1/2^s$. For $s = d\log(2N/\epsilon + 1) + \log(1/\delta)$, this probability is at most $\delta/(\frac{2N}{\epsilon} + 1)^d$. Now, let $\hat{y}$ be any vector on the surface of the sphere. Suppose y_j is its closest vector in C, and x_i is the corresponding x such that $y_j^\top x_i \geq 0$. Since C is an ϵ/N-cover of the sphere, this means that:

$$\hat{y}^\top x_i \geq y_j^\top x_i - \frac{\epsilon}{N}\|x_i\| \geq -\epsilon$$

The lemma follows. □

Denote set of points in positive class as B and the set of points in negative class as A. We simplify the optimal solution to SVM as follows.

$$\begin{aligned} w^* &= \arg\max_{\|w\|=1} \inf_{x \in B} w^\top x - \sup_{x \in A} w^\top x \\ &= \arg\min_{\|w\|=1} - \inf_{x \in B} w^\top x + \sup_{x \in A} w^\top x \\ &= \arg\min_{\|w\|=1} \sup_{x \in B} -w^\top x + \sup_{x \in A} w^\top x \\ &= \arg\min_{\|w\|=1} \sup_{x \in -B} w^\top x + \sup_{x \in A} w^\top x \end{aligned} \tag{10}$$

Additionally,

$$-b^* = \frac{1}{2}(\sup_{x \in A}(\alpha\hat{\mu} + \beta\hat{y})^\top)x - \sup_{x \in -B}(\alpha\hat{\mu} + \beta\hat{y})^\top)x)$$

Define the set $A_\mu = \{x + \mu, \forall x \in A\}$, and the set $-B_\mu = \{x + \mu, \forall x \in -B\}$.

With this, and some algebraic simplification the SVM optimization problem becomes:

$$\arg\min_{\alpha \in [-1,1], \hat{y}} \sup_{x \in A_\mu} (\alpha\hat{\mu} + \beta\hat{y})^\top (x - \mu) + \sup_{x \in -B_\mu} (\alpha\hat{\mu} + \beta\hat{y})^\top (x - \mu)$$

$$\arg\min_{\alpha \in [-1,1], \hat{y}} -2\alpha\|\mu\| + \sup_{x \in A_\mu} (\alpha\hat{\mu} + \beta\hat{y})^\top)x + \sup_{x \in -B_\mu} (\alpha\hat{\mu} + \beta\hat{y})^\top)x$$

[1] https://people.eecs.berkeley.edu/~bartlett/courses/281b-sp08/19.pdf

$$w^* = \arg\min_{\alpha,\hat{y}} -2\alpha\|\mu\| + \sup_{x\in A_\mu} (\alpha\hat{\mu} + \beta\hat{y})^\top)x + \sup_{x\in -B_\mu} (\alpha\hat{\mu} + \beta\hat{y})^\top)x \tag{11}$$

Additionally,

$$-b^* = \frac{1}{2}(\sup_{x\in A_\mu} (\alpha\hat{\mu} + \beta\hat{y})^\top)x - \sup_{x\in -B_\mu} (\alpha\hat{\mu} + \beta\hat{y})^\top)x)$$

Theorem 5. *$D(0)$ satisfies the concentration conditions in Assumption 2 and 3. Suppose $\|\mu\| > X_{\max}(p,\delta,d)$ and $\zeta(p,\delta,d,q) > 4X_{\max}(m,\delta,d)$. If p and m are sufficiently large, then with probability at least $1-8\delta$, the worst class error rate achieved by ERM is worse than the worst class error achieved by subsampling the classes.*

Proof. We start with expression for optimal SVM solution derived above

$$w^* = \arg\min_{\alpha,\hat{y}} -2\alpha\|\mu\| + \sup_{x\in A_\mu} (\alpha\hat{\mu} + \beta\hat{y})^\top)x + \sup_{x\in -B_\mu} (\alpha\hat{\mu} + \beta\hat{y})^\top)x \tag{12}$$

Additionally,

$$-b^* = \frac{1}{2}(\sup_{x\in A_\mu} (\alpha\hat{\mu} + \beta\hat{y})^\top)x - \sup_{x\in -B_\mu} (\alpha\hat{\mu} + \beta\hat{y})^\top)x)$$

Let us try to bound $-b^*$. From the concentration condition in Assumption 2, we know the first term above lies in

$$\frac{1}{2}[X_{\max}(p,\delta,d) - c(p,\delta,d), X_{\max}(p,\delta,d) + C(p,\delta,d)]$$

The second term lies in

$$\frac{1}{2}[X_{\max}(m,\delta,d) - c(m,\delta,d), X_{\max}(m,\delta,d) + C(m,\delta,d)]$$

As a result, the lower bound on $-b^*$ is

$$\frac{1}{2}\Big(X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d) - c(p,\delta,d) - C(m,\delta,d)\Big) \tag{13}$$

The upper bound on $-b^*$ is

$$\frac{1}{2}\Big(X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d) + C(p,\delta,d) + c(m,\delta,d)\Big) \tag{14}$$

Note $-b^*$ lies in the above interval with probability at least $1-2\delta$.

The expression for the error of a classifier $w^\top x + b$ is given as follows. We assume $\|w\| = 1$.

$$\mathsf{Err}_+ = \mathbb{P}(w^\top X + b \le 0 | X \sim D(\mu)) \tag{15}$$

where $D(\mu)$ is the distribution of samples for the positive class centered at μ. We assume D is spherically symmetric about zero so we simplify the above expression as follows.

$$\mathsf{Err}_+ = \mathbb{P}(w^\top(\mu + \tilde{X}) + b \le 0 | \tilde{X} \sim D(0)) \tag{16}$$

Since $\tilde{X}$ is sampled from $D(0)$ which is spherically symmetric, its projection on $w^T\tilde{X}$ will have a distribution that does not depend on the direction w. Let us denote $w^\top\tilde{X} = W$. The above expression becomes. Let us denote the CDF of W as F_W.

$$\mathsf{Err}_+ = \mathbb{P}(W \le -w^\top\mu - b) = F_W(-w^\top\mu - b) \tag{17}$$

We now plug in expression for the max-margin solution. In the analysis above, we showed that $-b^* \in [a_{\min}, a_{\max}]$. Therefore, the error for the positive class

$$\begin{aligned}
&F_W(-w^\top \mu + a_{\min}) \leq \mathsf{Err}_+ \leq F_W(-w^\top \mu + a_{\max}) \\
&F_W(-\alpha\|\mu\| + a_{\min}) \leq \mathsf{Err}_+ \leq F_W(-\alpha\|\mu\| + a_{\max}) \\
&F_W(-\|\mu\| + \frac{1}{2}\big(X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d) - c(p,\delta,d) - C(m,\delta,d)\big)) \leq \mathsf{Err}_+ \leq \\
&F_W(-\alpha\|\mu\| + \frac{1}{2}\big(X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d) + C(p,\delta,d) + c(m,\delta,d)\big))
\end{aligned} \tag{18}$$

In the upper bound, we invoked a condition that for the optimal $w = \alpha\hat{\mu} + \beta\hat{y}$, where $\beta = \sqrt{1-\alpha^2}$. In the lower bound, we set α to one. Consider the following two cases.

- **Balanced class case:** $m = p$

$$F_W\bigg(-\|\mu\| - \frac{1}{2}c(p,\delta,d) - \frac{1}{2}C(m,\delta,d)\bigg) \leq \mathsf{Err}_+ \leq F_W\bigg(-\alpha\|\mu\| + \frac{1}{2}C(p,\delta,d) + \frac{1}{2}c(m,\delta,d)\bigg) \tag{19}$$

- **Imbalanced class case:** $p >> m$. In this case, the error at least grows as

$$F_W\big(-\|\mu\| + \frac{1}{2}\big(X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d)\big)\big)$$

We need to show that the upper bound of the balanced case is better than the lower bound of the imbalanced case, which boils down to the following

$$\begin{aligned}
&F_W\bigg(-\alpha\|\mu\| + \frac{1}{2}C(p,\delta,d) + \frac{1}{2}c(m,\delta,d)\bigg) \leq \\
&F_W\bigg(-\|\mu\| + \frac{1}{2}\big(X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d) - c(p,\delta,d) - C(m,\delta,d)\big)\bigg)
\end{aligned} \tag{20}$$

If $\alpha \geq 1 - \frac{X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d) - \bar{c}(p,m,\delta,d)}{2\|\mu\|}$, where $\bar{c}(p,m,\delta,d) = C(p,\delta,d) + c(p,\delta,d) + C(m,\delta,d) + c(m,\delta,d)$ then the above inequality holds true.

We now show that $\alpha \geq 1-\eta$, where $\eta = \frac{X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d) - \bar{c}(p,m,\delta,d)}{2\|\mu\|}$. Since $\|\mu\| > X_{\max}(p,\delta,d)$, $\eta < \frac{1}{2}$. To confirm that $1 - \frac{X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d) - \bar{c}(p,m,\delta,d)}{2\|\mu\|} \leq 1$, we need to check that $X_{\max}(p,\delta,d) \geq X_{\max}(m,\delta,d) + \bar{c}(p,m,\delta,d) \geq 0$. Observe that $\zeta(p,\delta,d,q) > 4X_{\max}(m,\delta,d)$, which implies $X_{\max}(p,\delta,d,q) > 4X_{\max}(m,\delta,d)$. $X_{\max}(p,\delta,d) - \big(X_{\max}(m,\delta,d) + \bar{c}(p,m,\delta,d)\big)$, which is lower bounded $3X_{\max}(m,\delta,d) - \bar{c}(p,m,\delta,d)$. Note that the second term $\bar{c}(p,m,\delta,d)$ diminishes to zero for sufficiently large m and p while the first term is positive, which shows that $\alpha \leq 1$.

Recall the SVM objective is

$$-2\alpha\|\mu\| + \sup_{x \in A_\mu} (\alpha\hat{\mu} + \beta\hat{y})^\top x + \sup_{x \in -B_\mu} (\alpha\hat{\mu} + \beta\hat{y})^\top x$$

where $x \in D(0)$. For a fixed α, let $\hat{y}(\alpha)$ denote the minimizer of the above.

We compare the SVM objective when $\alpha = 1$ to a lower bound on the optimal value achievable if $\alpha < 1 - \eta$. When $\alpha = 1$ the objective becomes

$$-2\|\mu\| + \sup_{x \in A_\mu} (\hat{\mu}^\top x) + \sup_{x \in -B_\mu} (\hat{\mu}^\top x) \tag{21}$$

Recall $q = d\log(N/\epsilon + 1) + \log(1/\delta)$, where $N = \max_{i \in A} \|x_i\|$, where $\epsilon = \frac{1}{\log p}$. Similarly, define $\tilde{q} = d\log(\tilde{N}/\epsilon + 1) + \log(1/\delta)$, where $\tilde{N} = \max_{i \in B} \|x_i\|$. Consider the qth and $\tilde{q}^{th}$ highest value for $\hat{\mu}^\top x$ on set A and set B respectively. For a fixed α, we obtain a lower bound for the SVM objective in terms of qth and $\tilde{q}^{th}$ highest values as follows.

$$-2\alpha\|\mu\| + \alpha\hat{\mu}^\top x_+^{(i)} + \alpha\hat{\mu}^\top x_-^{(j)} + \beta\hat{y}(\alpha)^\top x_+^{(i)} + \beta\hat{y}(\alpha)^\top x_-^{(j)}$$

where $x_-^{(j)}$ has one of the top q projections on $\hat{\mu}$ among the positive samples, where $x_+^{(i)}$ has one of the top $\tilde{q}$ projections on $\hat{\mu}$ among the negative samples. We use Lemma 7 to arrive at a lower bound on the SVM objective. To use Lemma 7, we need $\hat{y}^\top x_+^{(i)}$ to have a median of zero conditional on $\hat{\mu}^\top x_+^{(i)}$. We also need a similar condition for $\hat{y}^\top x_-^{(j)}$ conditional on $\hat{\mu}^\top x_-^{(j)}$. These conditions follow from Lemma 6.

With probability $1-2\delta$ the lower bound on the objective is

$$\alpha(-2\|\mu\| + \hat{\mu}^\top x_+^{(i)} + \hat{\mu}^\top x_-^{(j)}) - 2\epsilon$$

We minimize this lower bound for $\alpha \in [-1, 1-\eta)$ and obtain the following

$$(1-\eta)(-2\|\mu\| + \hat{\mu}^\top x_+^{(i)} + \hat{\mu}^\top x_-^{(j)}) - 2\epsilon$$

where we use the following fact $\|\mu\| > X_{\max}(p,\delta,d) \geq X_{\max}(m,\delta,d) + \bar{c}(p,m,\delta,d)$. We will now show that the lower bound above has a very low probability to improve upon the objective value for $\alpha = 1$. As a result, optimal α will have to be more than $1-\eta$. Let us consider the event

$$\begin{aligned}(1-\eta)(-2\|\mu\| + \hat{\mu}^\top x_+^{(i)} + \hat{\mu}^\top x_-^{(j)}) - 2\epsilon \leq \\ -2\|\mu\| + \sup_{x\in A_\mu}(\alpha\hat{\mu}^\top x) + \sup_{x\in -B_\mu}(\alpha\hat{\mu}^\top x)\end{aligned} \tag{22}$$

After rearrangement we get

$$\eta\|\mu\| + \frac{1}{2}(1-\eta)\Big(\hat{\mu}^\top x_+^{(i)} + \hat{\mu}^\top x_-^{(j)})\Big) \leq \frac{1}{2}\Big(\sup_{x\in A_\mu}(\hat{\mu}^\top x) + \sup_{x\in -B_\mu}(\hat{\mu}^\top x) + 2\epsilon\Big)$$

We substitute $\epsilon = 1/\log p$ and use the expression for $\eta\|\mu\|$ to get

$$\begin{aligned}&\big(X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d) - \bar{c}(p,m,\delta,d)\big) + (1-\eta)\Big(\hat{\mu}^\top x_+^{(i)} + \hat{\mu}^\top x_-^{(j)})\Big) \leq \\ &\Big(\sup_{x\in A_\mu}(\hat{\mu}^\top x) + \sup_{x\in -B_\mu}(\hat{\mu}^\top x) + \frac{2}{\log p}\Big)\end{aligned}$$

After further rearrangement we get

$$\begin{aligned}&(1-\eta)\Big(\hat{\mu}^\top x_+^{(i)} + \hat{\mu}^\top x_-^{(j)})\Big) \leq \\ &\Big(\sup_{x\in A_\mu}(\hat{\mu}^\top x) + \sup_{x\in -B_\mu}(\hat{\mu}^\top x) - X_{\max}(p,\delta,d) + X_{\max}(m,\delta,d) + \frac{2}{\log p}\Big) + \bar{c}(p,m,\delta,d)\end{aligned}$$

From the above we get

$$\begin{aligned}&(1-\eta)\Big(\hat{\mu}^\top x_+^{(i)} + \hat{\mu}^\top x_-^{(j)})\Big) \leq \\ &\Big(\sup_{x\in A_\mu}(\hat{\mu}^\top x) + \sup_{x\in -B_\mu}(\hat{\mu}^\top x) - X_{\max}(p,\delta,d) + X_{\max}(m,\delta,d) + \frac{2}{\log p}\Big) + \bar{c}(p,m,\delta,d)\end{aligned}$$

Since $\eta < \frac{1}{2}$ we can further simplify the LHS with a weaker lower bound

$$\frac{1}{2}\left(\hat{\mu}^\top x_+^{(i)} + \hat{\mu}^\top x_-^{(j)})\right) \leq \left(\sup_{x\in A_\mu}(\hat{\mu}^\top x) + \sup_{x\in -B_\mu}(\hat{\mu}^\top x) - X_{\max}(p,\delta,d) + X_{\max}(m,\delta,d) + \frac{2}{\log p}\right) + \bar{c}(p,m,\delta,d)$$

We write an upper bound for RHS using the concentration condition.

$$\frac{1}{2}\left(\hat{\mu}^\top x_+^{(i)} + \hat{\mu}^\top x_-^{(j)})\right) \leq 2X_{\max}(m,\delta,d) + \bar{c}(p,m,\delta,d) + C(m,\delta,d) + C(p,\delta,d) + \frac{2}{\log p}$$

Using the lower bound from the concentration condition in Assumption 3, we get the following lower bound

$$\zeta(m,\delta,d,\tilde{q}) + \zeta(p,\delta,d,q) \leq 4X_{\max}(m,\delta,d) + 2\bar{c}(p,m,\delta,d) + 2C(m,\delta,d) + 2C(p,\delta,d) + \frac{4}{\log p}$$

Since $\zeta(p,\delta,d,q) > 4X_{\max}(m,\delta,d)$ for a sufficiently large m and p we gather that $\zeta(p,\delta,d,q) + \zeta(m,\delta,d,\tilde{q}) \geq 4X_{\max}(m,\delta,d) + 2\bar{c}(p,m,\delta,d) + 2C(m,\delta,d) + 2C(p,\delta,d) + \frac{4}{\log p}$. As a result, the above inequality does not hold. Therefore, the event in equation 22 occurs with a probability at most 2δ. As a result, we obtain that $\alpha \geq 1-\eta$ with a probability at least $1-4\delta$. We showed above that if $\alpha \geq 1-\eta$, then with probability at least $1-4\delta$, worst class error improves under data balancing. The intersection of these two events occurs with a probability at least $1-8\delta$.

□

Expression for ζ In this section, our goal is to derive a lower bound on the qth maximum projection of $\hat{\mu}$ across different data samples. We denote $V = \hat{\mu}^\top X$.

We first make some observations that we use subsequently. Consider the event $V^{(q)} > t$, where $V^{(q)}$ is q^{th} highest value of V among p samples. Suppose the CDF of V is F_V. Find a value r such that $F_V(t) = 1-r$. This denotes r fraction of V is greater than t. Define $U_i = I(V_i > t)$, where I is the indicator function and U_i is one when $V_i > r$ and zero otherwise. Consider the event

$$\sum_{i=i}^{p} U_i > q$$

If the above event is true, then that implies there are at least q values that are above t and thus $V^{(q)} > t$. Also, if $V^{(q)} > t$, then there exist at least q U_i's that are one. Thus the above two events are equivalent. The expectation $\mathbb{E}[\sum_{i=1}^{p} U_i] = pr$. Let $q = \frac{pr}{2}$. We use Chernoff bound to arrive at the following bound

$$\Pr(\sum_{i=i}^{p} U_i < q) < e^{-\frac{pr}{8}}$$

We set $\frac{pr}{8} = d\log(\log p \max_{x_i\in A}\|x_i\|) + \log(1/\delta)$.

Therefore, $r = 8\frac{d\log p\log(\max_{x_i\in A}\|x_i\|/\epsilon)+\log(1/\delta)}{p}$. For the case of symmetric d dimensional Gaussians centered at zero we get, $\frac{pr}{8} = d\log(\log p\sqrt{d\log p})) + \log(1/\delta)$. We simplify the bound $e^{-\frac{pr}{8}}$ as follows.

$$e^{-\frac{pr}{8}} \leq e^{-d\log(\log p\sqrt{d\log p}))} \leq \frac{1}{(\log p\sqrt{d\log p})^d}$$

For sufficiently large p, the probability falls below any δ.

We now derive a bound on t.

Recall $F(t) = 1-r$, which simplifies for a Gaussian to $Q(t) = r$, where Q is the Q function. Since $Q(t) \leq e^{-t^2}$. We get $e^{-t^2} \geq r$, which implies

$$t \leq \sqrt{2\log\frac{1}{r}} = \sqrt{2\log\frac{p}{8(d\log(\log p\sqrt{d\log p}) + \log(1/\delta))}}$$

Hence, we can use $\zeta(p,\delta,d) = \sqrt{2\log\frac{p}{8(d\log(\log p\sqrt{d\log p})+\log(1/\delta))}}$ for symmetric Gaussian distributions.

Corollary 2. *Let $D(0)$ be a symmetric Gaussian in $\mathbb{R}^2$. The bias for the ERM classifier θ_{erm} lies in an arbitrarily small interval centered at $\sqrt{2\log(p/\delta)} - \sqrt{2\log(m/\delta)}$. In contrast, the bias for the classifier under subsampling θ_{ss} lies in an arbitrarily small interval centerd at zero. If $m = \log p$ and $\|\mu\| > \sqrt{2\log(p/\delta)}$, then with probability $1-8\delta$, ERM has a worse worst class error than subsampling. Let $D(0)$ be a symmetric uniform in $\mathbb{R}^2$. The bias term for both the ERM classifier and the subsampling classifier lies in an arbitrarily small interval centered at zero.*

Proof. To prove this Corollary, we leverage Theorem 5 and its proof. In equation 13 and equation 14 we derive the upper and the lower bounds for the bias. For a d dimensional spherically symmteric Gaussian, the expressions for concentration condition are derived in Lemma 18. For a sufficiently large m, p, the bias is centered at

$$\sqrt{2\log\frac{p}{\delta}} - \sqrt{2\log\frac{m}{\delta}}$$

Observe that $X_{\max}(p,\delta,d) \approx \sqrt{2\log\frac{p}{\delta}}$ and $X_{\max}(m,\delta,d) \approx \sqrt{2\log\frac{m}{\delta}}$. If we subsample, then we are in the case, where $X_{\max}(p,\delta,d) = X_{\max}(m,\delta,d)$ and as a result the bias term is centered at zero. If m grows as $\log p$, then the upper bound on $X_{\max}(m,\delta,d)$ is $\sqrt{2\log\log p}$. As a result, the condition that $\zeta(p,\delta,d,q) > 4X_{\max}(m,\delta,d)$ is satisfied for sufficiently large p. Finally, if $\|\mu\|$ for the mean of the Gaussian is more than $\sqrt{2\log(p/\delta)} - \sqrt{2\log(m/\delta)}$, then it follows from the previous theorem that subsampling improves the worst group error. For a 2 dimensional symmetric uniform, the expressions for the concentration condition are derived in Lemma 13. For a sufficiently large m, p, the bias term is centered at zero with the interval given as

$$\left[-\frac{\varrho}{p^{2/3}} - \frac{1}{p}, \frac{1}{m} + \frac{\varrho}{m^{2/3}}\right]$$

where ϱ is a constant whose expression can be obtained from Lemma 13. □

B. Proofs for imbalanced groups

B.1. Two-dimensional case

Lemma 8. *Let $x_1,\ldots,x_n$ be n i.i.d unit Gaussians and let $X_{\max} = \max(x_1,\ldots,x_n)$. Then for n large enough, with probability $\geq 1-3\delta$, we have that:*

$$X_{\max} \leq b_n + a_n\log(1/\delta), \quad X_{\max} \geq b_n - a_n\log\log(1/\delta)$$

where $a_n = \frac{1}{\sqrt{2\log(n)}}$ and $b_n = \sqrt{2\log(n)} - \frac{\log\log n + \log(4\pi)}{\sqrt{2\log n}}$ are the constants in the Fisher-Tippett-Gnedenko theorem when applied to Gaussians.

Proof. From the Fisher-Tippett-Gnedenko theorem, when n is large enough, we have that $X_{\max} \xrightarrow{d} a_nZ + b_n$, where $\xrightarrow{d}$ stands for convergence in distribution, and Z is a standard Gumbel random variable. If n is sufficiently large, then for any $t \in \mathbb{R}$, $|\Pr(X_{\max} \leq t) - \Pr(a_nZ + b_n \leq t)| \leq \delta/2$.

For a standard Gumbel variable Z, we have that:

$$\Pr(Z \leq \log(1/\delta)) = \exp(-\exp(-\log(1/\delta))) = \exp(-\delta) \geq 1-\delta$$

As a result,

$$\Pr(X_{\max} \leq b_n + a_n\log(1/\delta) \geq 1 - \frac{3\delta}{2}$$

Additionally, we have:

$$\Pr(Z \geq -\log\log(1/\delta)) = 1 - \exp(-\exp(\log\log(1/\delta))) = 1-\delta$$

As a result,

$$\Pr(X_{\max} \geq b_n - a_n\log\log(1/\delta)) \geq 1 - \frac{3\delta}{2}$$

Finally, if we take a union bound on the complement of the above two events and then the lemma follows. □

Lemma 9. *Let $x_1, \ldots, x_n$ be vectors in $\mathbb{R}^2$ drawn i.i.d from $N(0, I_2)$. Then, with probability $\geq 1-\delta$,*

$$\max_i \|x_i\| \leq \sqrt{2\log(n/\delta)}$$

Proof. $\|x_i\|$ follows a Rayleigh distribution and we use this observation to arrive at the above result.

$$\begin{aligned} &\Pr\left(\max_i \|x_i\| \leq \sqrt{2\log(n/\delta)}\right) = 1 - \Pr\left(\max_i \|x_i\| \geq \sqrt{2\log(n/\delta)}\right) \geq \\ &1 - n\Pr\left(\|x_i\| \geq \sqrt{2\log(n/\delta)}\right) = 1 - ne^{-\log(n/\delta)} = 1-\delta \end{aligned} \tag{23}$$

□

Lemma 10. *Let $x_1, \ldots, x_n$ be n i.i.d unit Gaussians with covariance I_2. Then for n large enough, with probability $\geq 1-\delta$, we have that for all directions $v \in \mathbb{R}^2$:*

$$\max_{i\in\{1,\cdots,n\}} \{v^\top x_i\} \leq b_n + a_n + a_n \log\left(\frac{6\sqrt{2\log(2n/\delta)}}{a_n\delta}\right),$$

$$\max_{i\in\{1,\cdots,n\}} \{v^\top x_i\} \geq b_n - a_n - a_n \log\log\left(\frac{6\sqrt{2\log(2n/\delta)}}{a_n\delta}\right)$$

where a_n and b_n are the constants in the Fisher-Tippett-Gnedenko theorem when applied to Gaussians.

Proof. Suppose $f(v) = \max_i v^\top x_i$ where v is a unit vector in $\mathbb{R}^2$. Then,

$$f(v) - f(u) = \max_i v^\top x_i - \max_i u^\top x_i \leq \max_i (v-u)^\top x_i \leq \|v-u\| \cdot \max_i \|x_i\|$$

where the first step follows from definition, the second step from subtracting a smaller quantity, and the last step from the Cauchy-Schwartz inequality.

From Lemma 9, with probability $\geq 1-\delta/2$, $\max_i \|x_i\| \leq \sqrt{2\log(2n/\delta)}$, which gives us:

$$f(v) - f(u) \leq \sqrt{2\log(2n/\delta)} \cdot \|v-u\|$$

Now, we can build an ϵ-cover $C(\epsilon)$ over unit vectors on the circle so that successive vectors v_i and v_{i+1} have the property that $\|v_i - v_{i+1}\| \leq \epsilon$. The size of such an ϵ-cover is $N(\epsilon) = 1/\epsilon$; additionally, for any unit vector v in $\mathbb{R}^2$, there exists some v_i in the cover such that

$$f(v_i) - \sqrt{2\log(2n/\delta)}\epsilon \leq f(v) \leq f(v_i) + \sqrt{2\log(2n/\delta)}\epsilon$$

$f(v_i)$ is the maximum over n i.i.d. standard Gaussians $N(0,1)$. From Lemma 8, we know

$$b_n - a_n \log\log(1/\delta) \leq f(v_i) \leq b_n + a_n \log(1/\delta)$$

Now we can apply Lemma 8 with $\delta = \frac{\delta}{6N(\epsilon)}$ plus an union bound over the cover $C(\epsilon)$ to get that for all v_i in the cover,

$$b_n - a_n \log\log(6/(\epsilon\delta)) \leq f(v_i) \leq b_n + a_n \log(6/(\epsilon\delta))$$

For all directions $v \in \mathbb{R}^2$

$$b_n - a_n \log\log(6/(\epsilon\delta)) - \sqrt{2\log(2n/\delta)}\epsilon \leq f(v) \leq b_n + a_n \log(6/(\epsilon\delta)) + \sqrt{2\log(2n/\delta)}\epsilon$$

Plugging in $\epsilon = \frac{a_n}{\sqrt{2\log(2n/\delta)}}$ in the above expression we get.

$$b_n - a_n - a_n \log\log\left(\frac{6\sqrt{2\log(2n/\delta)}}{a_n\delta}\right) \leq f(v) \leq b_n + a_n + a_n \log\left(\frac{6\sqrt{2\log(2n/\delta)}}{a_n\delta}\right)$$

□

Lemma 11. *Consider the density: $f(t) = \frac{2}{\pi}\sqrt{1-t^2}$ for $t \in [0,1]$ and $f(t) = 0$ otherwise. Let F be the corresponding CDF and let $U(t) = F^{-1}(1-1/t)$. Then, the following facts hold:*

1. $$\lim_{h\to 0} \frac{1-F(1-xh)}{1-F(1-h)} = x^{3/2}$$

2. $1 - U(n) \geq \left(\frac{3\pi}{4\sqrt{2}n}\right)^{2/3}$.

Proof. The first part follows by integration by substitution and Taylor expansion of $\theta - \frac{\sin 2\theta}{2}$ around $\theta = 0$.

To see the first part, observe that:

$$1 - F(1-h) = \int_{1-h}^{1} \frac{2}{\pi} \cdot \sqrt{1-t^2}dt$$

We now calculate this integral by substitution. Let $t = \cos\theta$, then $dt = -\sin\theta d\theta$, and the limits of the integral become $\cos^{-1}(1-h)$ to 0. The integral becomes:

$$\int_0^{\cos^{-1}(1-h)} \frac{2}{\pi} \cdot \sin^2\theta d\theta = \int_0^{\cos^{-1}(1-h)} \frac{1}{\pi} \cdot (1-\cos 2\theta)\, d\theta = \frac{1}{\pi} \cdot \left(\theta - \frac{\sin 2\theta}{2}\right)\Big|_0^{\cos^{-1}(1-h)}$$

A Taylor series expansion of $\sin 2\theta$ shows that $\sin 2\theta = 2\theta - \frac{8\theta^3}{3!} + o(\theta^3)$; therefore

$$\theta - \frac{\sin 2\theta}{2} = \frac{2\theta^3}{3} + o(\theta^3),$$

which brings the result of the integral to $\frac{4(\cos^{-1}(1-h))^3}{3\pi} + o((\cos^{-1}(1-h))^3)$. Observe through a Taylor expansion that

$$\cos^{-1}(1-h) = \sin^{-1}(\sqrt{h(2-h)}) = \sqrt{h(2-h)} - o(h(2-h)),$$

and hence

$$\lim_{h\to 0} \frac{1}{\pi} \cdot \left(\theta - \frac{\sin 2\theta}{2}\right)\Big|_0^{\cos^{-1}(1-h)} = \frac{1}{\pi} \cdot \frac{2(2h)^{3/2}}{3}, \tag{24}$$

from which the first part of the lemma follows. For the second part, we observe that from the definition of $U(n)$, we have that $1 - U(n) = h$, where:

$$\int_{1-h}^{1} \frac{2}{\pi} \cdot \sqrt{1-t^2}dt = \frac{1}{n}$$

From equation 24, observe that for small enough h (which corresponds to large enough n), the left hand side is at most $\frac{4\sqrt{2}}{3\pi}h^{3/2}$. This implies that $h = 1 - U(n) \geq \left(\frac{3\pi}{4\sqrt{2}n}\right)^{2/3}$ and the lemma follows. □

Lemma 12. *Consider the density: $f(t) = \frac{2}{\pi}\sqrt{1-t^2}$ for $t \in [0,1]$ and $f(t) = 0$ otherwise. Let $x_1, \ldots, x_n$ be n drawn i.i.d from f and let $X_{\max} = \max(x_1, \ldots, x_n)$. Then for n large enough, with probability $\geq 1-\delta$, we have that:*

$$X_{\max} \leq 1, \quad X_{\max} \geq 1 - \left(\frac{3\pi\log(2/\delta)}{4\sqrt{2}n}\right)^{2/3}$$

Proof. Observe that for this distribution, $x_F = 1$. From this, and the first part of Lemma 11, it follows that this distribution is of the Weibull type with $\alpha = 3/2$. From the Fisher-Tippett-Gnedenko Theorem, this means that the maximum of n points converges to $a_n Z + b_n$ in distribution, where $a_n = 1 - U(n)$, $b_n = 1$, and Z is a reverse Weibull distributed variable with $\alpha = 3/2$. Setting $X_{\max}(n,\delta) = 1$, we get that $C(n,\delta) = 0$.

To calculate $c(n, \delta)$, we observe that from the second part of Lemma 11, $a_n \geq \left(\frac{3\pi}{4\sqrt{2}n}\right)^{2/3}$. Additionally, if Z is a reverse Weibull variable with parameter $\alpha = 3/2$, then,

$$\Pr(Z \leq -(\log(2/\delta)^{2/3})) = \exp(-(\log(2/\delta))^{2/3})^{3/2} = \exp(-(\log(2/\delta)) = \delta/2$$

Therefore, $\Pr\left(a_n Z + b_n \leq 1 - \left(\frac{3\pi \log(2/\delta)}{4\sqrt{2}n}\right)^{2/3}\right) \leq \delta/2$. We get another $\delta/2$ from the distributional convergence of the maximum of n random variables to the limit for large enough n. □

Lemma 13. *Let $x_1, \ldots, x_n$ be n drawn i.i.d from symmetric uniform distribution centered at zero. Then for n large enough, with probability $\geq 1 - \delta$, we have that for all directions $v \in \mathbb{R}^2$:*

$$\max_{i \in \{1, \cdots, n\}} \{v^\top x_i\} \leq 1$$

$$\max_{i \in \{1, \cdots, n\}} \{v^\top x_i\} \geq 1 - \left(\frac{3\pi \log(2n/\delta)}{4\sqrt{2}n}\right)^{2/3} - \frac{1}{n}$$

Proof. Suppose $f(v) = \max_i v^\top x_i$ where v is a unit vector in $\mathbb{R}^2$. Then,

$$f(v) - f(u) = \max_i v^\top x_i - \max_i u^\top x_i \leq \max_i (v - u)^\top x_i \leq \|v - u\| \cdot \max_i \|x_i\|$$

where the first step follows from definition, the second step from subtracting a smaller quantity, and the last step from the Cauchy-Schwartz inequality.

Note that $\max_i \|x_i\| \leq 1$, which gives us:

$$f(v) - f(u) \leq \cdot \|v - u\|$$

Now, we can build an ϵ-cover $C(\epsilon)$ over unit vectors on the circle so that successive vectors v_i and v_{i+1} have the property that $\|v_i - v_{i+1}\| \leq \epsilon$. The size of such an ϵ-cover is $N(\epsilon) = 1/\epsilon$; additionally, for any unit vector v in $\mathbb{R}^2$, there exists some v_i in the cover such that

$$f(v_i) - \epsilon \leq f(v) \leq f(v_i) + \epsilon$$

Observe that $f(v_i)$ is a maximum over n i.i.d. random variables drawn from a distribution $f(t) = \frac{2}{\pi}\sqrt{1 - t^2}$ for $t \in [0, 1]$ and $f(t) = 0$ otherwise. Now we can apply Lemma 12 with $\delta = \frac{\delta}{N(\epsilon)}$ plus an union bound over the cover $C(\epsilon)$ to get that for all v_i in the cover,

$$1 - \left(\frac{3\pi \log(2N(\epsilon)/\delta)}{4\sqrt{2}n}\right)^{2/3} \leq f(v_i) \leq 1$$

For all directions $v \in \mathbb{R}^2$

$$1 - \left(\frac{3\pi \log(2N(\epsilon)/\delta)}{4\sqrt{2}n}\right)^{2/3} - \epsilon \leq f(v) \leq 1$$

Plugging in $\epsilon = \frac{1}{n}$ in the above expression we get.

$$1 - \left(\frac{3\pi \log(2n/\delta)}{4\sqrt{2}n}\right)^{2/3} - \frac{1}{n} \leq f(v) \leq 1$$

□

Lemma 14 (Approximate Maximization Lemma - I)**.** *Let $F(\alpha) = f(\alpha) + g(\alpha)$ where $g(\alpha) = \alpha u + \sqrt{1 - \alpha^2} v$, $u, v > 0$, and $f(\alpha)$ is an arbitrary function of α that lies in the interval $[-L, U]$. Let α_F be the value of α that maximizes $F(\alpha)$, and let $\alpha_g = \frac{u}{\sqrt{u^2+v^2}}$ be the value of α that maximizes $g(\alpha)$.*

Then, the angle between $(\alpha_F, \sqrt{1 - \alpha_F^2})$ and $(\alpha_g, \sqrt{1 - \alpha_g^2})$ is at most $\cos^{-1}\left(1 - \frac{L+U}{\sqrt{u^2+v^2}}\right)$. Additionally, the maximum value of $F(\alpha)$ is at least $\sqrt{u^2 + v^2} - L$.

Proof. For convenience, we can do a quick change of variables – we let $\alpha = \cos\theta$. Then $g(\theta) = u\cos\theta + v\sin\theta$, and is maximized at $\theta_g = \cos^{-1}\left(\frac{u}{\sqrt{u^2+v^2}}\right)$. This means we can re-write g as follows:

$$\begin{aligned} g(\theta) &= \sqrt{u^2+v^2}\cdot(\cos\theta_g\cos\theta + \sin\theta_g\sin\theta) \\ &= \sqrt{u^2+v^2}\cdot\cos(\theta_g - \theta) \end{aligned}$$

Similarly, we can do a change of variables on F and f as well. Suppose the value of θ that maximizes F is θ_F. Then we have that:

$$f(\theta_g) + \sqrt{u^2+v^2} \leq f(\theta_F) + \sqrt{u^2+v^2}\cos(\theta_g - \theta_F)$$

Since $f(\theta_g) \geq -L$ and $f(\theta_F) \leq U$, this gives us:

$$-L + \sqrt{u^2+v^2} \leq U + \sqrt{u^2+v^2}\cos(\theta_g - \theta_F)$$

The lemma follows from simple algebra. □

Lemma 15 (Approximate Maximixation Lemma - II)**.** *Let $F(\alpha) = f(\alpha) + g(\alpha)$ where $g(\alpha) = \alpha u - \sqrt{1-\alpha^2}v$, $u, v > 0$, and $f(\alpha)$ is an arbitrary function of α that lies in the interval $[-L, U]$. Let α_F be the value of α that maximizes $F(\alpha)$, and let $\alpha_g = 1$ be the value of α that maximizes $g(\alpha)$. Then, $\alpha_F \geq 1 - \frac{U+L}{u+v}$.*

Proof. To show the lemma, we observe that since $f(\alpha) \in [-L, U]$,

$$-L + u \leq U + \alpha_F u - \sqrt{1-\alpha_F^2}v$$

which implies $u(1-\alpha_F) + v\sqrt{1-\alpha_F^2} \leq L + U$. This will hold when $1 - \alpha_F \leq \frac{U+L}{u+v}$. The lemma follows. □

Denote set of points in positive class as B and the set of points in negative class as A. We simplify the optimal solution to SVM as follows.

$$\begin{aligned} w &= \arg\max_{\|w\|=1} \inf_{x\in B} w^\top x - \sup_{x\in A} w^\top x \\ &= \arg\min_{\|w\|=1} -\inf_{x\in B} w^\top x + \sup_{x\in A} w^\top x \\ &= \arg\min_{\|w\|=1} \sup_{x\in B} -w^\top x + \sup_{x\in A} w^\top x \\ &= \arg\min_{\|w\|=1} \sup_{x\in -B} w^\top x + \sup_{x\in A} w^\top x \end{aligned} \tag{25}$$

We write the classifier as $w = \alpha\hat{\mu} + \sigma\beta\hat{\psi}$, where $\hat{\mu}$ is a unit vector in the direction μ, $\hat{\psi}$ is a unit vector in the direction ψ, $\alpha \in [-1, 1]$, $\beta = \sqrt{1-\alpha^2}$ and σ is either $+1$ or -1.

Define the set $A_\mu = \{x + \mu, \forall x \in A\}$, and the set $-B_\mu = \{x + \mu, \forall x \in -B\}$.

With this, and some algebraic simplification the SVM optimization problem becomes:

$$\alpha^* = \arg\min_{\alpha\in[-1,1],\sigma\in\{-1,1\}} \sup_{x\in A_\mu} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top(x-\mu) + \sup_{x\in -B_\mu} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top(x-\mu)$$

Theorem 7. *Suppose $D(0)$ satisfies the concentration condition in Assumption 2 and as p and m approach ∞*

$$\frac{X_{\max}(p,\delta,2) - X_{\max}(m,\delta,2)}{2\|\psi\|} \to 0.$$

If $m, p \to \infty$, then with probability at least $1 - 4\delta$, the ERM solution converges to the invariant solution w^*_{inv}.

Proof. We write $w^* = \alpha^* \hat{\mu} + \sigma\beta^* \hat{\psi}$. Recall that α^* is a solution to:

$$\alpha^* = \underset{\alpha\in[-1,1],\sigma\in\{-1,1\}}{\arg\min} \sup_{x\in A_\mu} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top (x-\mu) + \sup_{x\in -B_\mu} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top (x-\mu)$$

where $\beta = \sqrt{1-\alpha^2}$. We next consider a further split of the positive class into the majority and minority groups – A_μ^M and A_μ^m.

$$\sup_{x\in A_\mu} v^\top x = \max\left(\sup_{x\in A_\mu^M} v^\top x, \sup_{x\in A_\mu^m} v^\top x \right),$$

Define sets $A_{\mu,\psi}^M = -\psi + A_\mu^M$ and $A_{\mu,\psi}^m = \psi + A_\mu^m$. We can write

$$\begin{aligned}
&\sup_{x\in A_\mu} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top (x-\mu) \\
= \quad & \max\left(\sup_{x\in A_\mu^M} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top (x-\mu), \sup_{x\in A_\mu^m} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top (x-\mu) \right) \\
= \quad & \max\left(\sup_{x\in A_{\mu,\psi}^M} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top x - \alpha\|\mu\| + \sigma\beta\|\psi\|, \sup_{x\in A_{\mu,\psi}^m} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top x - \alpha\|\mu\| - \sigma\beta\|\psi\| \right),
\end{aligned}$$

A similar expression will hold for B_μ. Define sets $B_{\mu,\psi}^M = \psi + B_\mu^M$ and $B_{\mu,\psi}^m = \psi + B_\mu^m$. We can write

$$\sup_{x\in -B_\mu} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top (x-\mu)$$

$$\max\left(\sup_{x\in -B_{\mu,\psi}^M} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top x - \alpha\|\mu\| + \sigma\beta\|\psi\|, \sup_{x\in -B_{\mu,\psi}^m} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top x - \alpha\|\mu\| - \sigma\beta\|\psi\| \right)$$

Define

$$f_1(\alpha) = \sup_{x\in A_{\mu,\psi}^M} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top x, \quad f_2(\alpha) = \sup_{x\in A_{\mu,\psi}^m} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top x$$

$$f_3(\alpha) = \sup_{x\in -B_{\mu,\psi}^M} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top x, \quad f_4(\alpha) = \sup_{x\in -B_{\mu,\psi}^m} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top x$$

We split up the SVM objective, and begin with two cases:

Case 1: $\sigma = 1$**.** Here, the SVM objective becomes:

$$\begin{aligned}
F(\alpha) = \min_\alpha \Big\{ & \max(f_1(\alpha) - \alpha\|\mu\| + \beta\|\psi\|, f_2(\alpha) - \alpha\|\mu\| - \beta\|\psi\|) + \\
& + \max(f_3(\alpha) - \alpha\|\mu\| + \beta\|\psi\|, f_4(\alpha) - \alpha\|\mu\| - \beta\|\psi\|) \Big\}
\end{aligned}$$

Recall the concentration condition holds for $D(0)$. Since the size of the majority group and minority group both $\geq n_0$, we obtain

$$f_1(\alpha), f_3(\alpha) \in [X_{\max}(p,\delta,2) - c(p,\delta,2), X_{\max}(p,\delta,2) + C(p,\delta,2)], \tag{26}$$

and also:

$$f_2(\alpha), f_4(\alpha) \in [X_{\max}(m,\delta,2) - c(m,\delta,2), X_{\max}(m,\delta,2) + C(m,\delta,2)] \tag{27}$$

We now look at two possible cases for α to determine what the inside maximum will look like. The first case is for large β – where

$$\beta \geq \frac{C(m,\delta,2) - c(p,\delta,2) - (X_{max}(p,\delta,2) - X_{\max}(m,\delta,2))}{2\|\psi\|}$$

and the objective simplifies to.

$$F(\alpha) = \min_{\alpha} f_1(\alpha) + f_3(\alpha) - 2\alpha\|\mu\| + 2\beta\|\psi\|,$$

The solution to this, from Lemma 15 is $\alpha^* \geq 1 - \frac{C(p,\delta,2)+c(p,\delta,2)}{\|\mu\|+\|\psi\|}$, which converges to the relevant solution as $p \to \infty$. The other case is where β is small – namely, $\beta \leq \frac{C(m,\delta,2)-c(p,\delta,2)}{2\|\psi\|}$ (here we use the fact that $X_{\max}(p,\delta,2) \geq X_{\max}(m,\delta,2)$. Since $\alpha^2 + \beta^2 = 1$, this implies that α is now close to 1 – specifically, $\alpha \geq \sqrt{1 - \left(\frac{C(m,\delta,2)-c(p,\delta,2)}{2\|\psi\|})\right)^2} \to 1$ as $p \to \infty$ by the conditions of the theorem. This means that in both cases, the inner maximum is achieved when α is close to 1.

Case 2: $\sigma = -1$**.** In this case, the SVM objective becomes:

$$F(\alpha) = \min_{\alpha} \Big\{ \max(f_1(\alpha) - \alpha\|\mu\| - \beta\|\psi\|, f_2(\alpha) - \alpha\|\mu\| + \beta\|\psi\|) + \\ \max(f_3(\alpha) - \alpha\|\mu\| - \beta\|\psi\|, f_4(\alpha) - \alpha\|\mu\| + \beta\|\psi\|) \Big\}$$

We again do a case by case analysis. We say that β is large if

$$\beta \geq \beta_{\mathsf{th}} = \frac{X_{\max}(p,\delta,2) - X_{\max}(m,\delta,2) + C(p,\delta,2) + c(m,\delta,2)}{2\|\psi\|}$$

In this case, the SVM objective becomes:

$$F(\alpha) = f_2(\alpha) + f_4(\alpha) - 2\|\mu\|\alpha + 2\beta\|\psi\|$$

Since $\beta \geq \beta_{\mathsf{th}}$, $-\sqrt{1-\beta_{\mathsf{th}}^2} \leq \alpha \leq \sqrt{1-\beta_{\mathsf{th}}^2}$. From Lemma 23, the solution to this is $\alpha \geq \sqrt{1-\beta_{\mathsf{th}}^2} - \frac{C(m,\delta,2)+c(m,\delta,2)}{\|\mu\|+\|\psi\|} \to 1$ as $m \to \infty$. The other case is for small β, where

$$\beta \leq \frac{X_{\max}(p,\delta,2) - X_{\max}(m,\delta,2) + C(p,\delta,2) + c(m,\delta,2)}{2\|\psi\|}$$

Since $\alpha^2 + \beta^2 = 1$, here α by definition satisfies

$$\alpha \geq \sqrt{1 - \left(\frac{X_{\max}(p,\delta,2) - X_{\max}(m,\delta,2) + C(p,\delta,2) + c(m,\delta,2)}{4\|\psi\|}\right)^2} \to 1$$

As $p, m \to \infty$ from the condition in the theorem $\alpha \to 1$. This means that in all four cases, the optimum is achieved when α is close to 1. The theorem follows. □

Theorem 6. *Suppose $D(0)$ satisfies the concentration condition in Assumption 2 and $X_{\max}(p,\delta,2) - X_{\max}(m,\delta,2) \geq 2\|\psi\| + c(p,\delta,2) + C(m,\delta,2)$. If $p \to \infty$, then with probability at least $1 - 4\delta$, the ERM solution converges to the spurious solution w^*_{spu}. In addition, with probability at least $1 - 12\delta$, $\mathbf{wge}(\theta_{ss}) < \mathbf{wge}(\theta_{erm})$.*

Proof. Recall $w^* = \alpha^*\hat{\mu} + \sigma\beta^*\hat{\psi}$, where α^* is a solution to:

$$\alpha^* = \underset{\alpha\in[-1,1],\sigma\in\{-1,1\}}{\arg\min} \sup_{x\in A_\mu} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top (x-\mu) + \sup_{x\in -B_\mu} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top (x-\mu)$$

where $\beta = \sqrt{1-\alpha^2}$. We follow a similar strategy as the previous proof of Theorem 7 and look at two cases – $\sigma = 1$ and -1.

Case 1: $\sigma = 1$**.** Here, the SVM objective becomes:

$$F(\alpha) = \min_{\alpha} \Big\{ \max(f_1(\alpha) - \alpha\|\mu\| + \beta\|\psi\|, f_2(\alpha) - \alpha\|\mu\| - \beta\|\psi\|) + \\ + \max(f_3(\alpha) - \alpha\|\mu\| + \beta\|\psi\|, f_4(\alpha) - \alpha\|\mu\| - \beta\|\psi\|) \Big\}$$

where $f_1(\alpha) = \sup_{x \in A^M_{\mu,\psi}} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top x$, $f_2(\alpha) = \sup_{x \in A^m_{\mu,\psi}} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top x$, $f_3(\alpha) = \sup_{x \in -B^M_{\mu,\psi}} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top x$, $f_2(\alpha) = \sup_{x \in -B^m_{\mu,\psi}} (\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top x$. From conditions on the majority and the minority class, and the Concentration Condition, with probability $1 - 4\delta$

$$f_1(\alpha), f_3(\alpha) \in [X_{\max}(p, \delta, 2) - c(p, \delta, 2), X_{\max}(p, \delta, 2) + C(p, \delta, 2)], \tag{28}$$

and also:

$$f_2(\alpha), f_4(\alpha) \in [X_{\max}(m, \delta, 2) - c(m, \delta, 2), X_{\max}(m, \delta, 2) + C(m, \delta, 2)] \tag{29}$$

Observe that from the conditions of the theorem, the first terms will dominate for all values of α, and hence the SVM objective will become:

$$F(\alpha) = \min_{\alpha} f_1(\alpha) + f_3(\alpha) - 2\alpha\|\mu\| + 2\beta\|\psi\|,$$

From Lemma 15, the optimal solution $\alpha^* \geq 1 - \frac{C(p,\delta,2)+c(p,\delta,2)}{\|\mu\|+\|\psi\|}$, with a lower bound on the optimal value $2X_{\max}(p, \delta, 2) - 2c(p, \delta, 2) - 2\|\mu\|$.

Case 2: $\sigma = -1$**.** Here, the SVM objective becomes:

$$F(\alpha) = \min_{\alpha} \Big\{ \max(f_1(\alpha) - \alpha\|\mu\| - \beta\|\psi\|, f_2(\alpha) - \alpha\|\mu\| + \beta\|\psi\|) + \\ \max(f_3(\alpha) - \alpha\|\mu\| - \beta\|\psi\|, f_4(\alpha) - \alpha\|\mu\| + \beta\|\psi\|) \Big\}$$

This time, from the conditions of the theorem, the first terms will dominate the maximum for all values of α, and hence the objective will become:

$$F(\alpha) = \min_{\alpha} f_1(\alpha) + f_3(\alpha) - 2\alpha\|\mu\| - 2\beta\|\psi\|$$

From Lemma 14, the optimal solution vector $(\alpha, \sqrt{1-\alpha^2})$ will be close to the spurious solution vector $(\frac{|\mu\|}{\sqrt{\|\mu\|^2+\|\psi\|^2}}, \frac{\|a\|}{\sqrt{\|\mu\|^2+\|\psi\|^2}})$, with the angle being at most $\cos^{-1}(1 - \frac{C(p,\delta,2)+c(p,\delta,2)}{\sqrt{\|\mu\|^2+\|\psi\|^2}})$. The optimal solution value will be at most $2X_{\max}(p, \delta, 2) + 2C(p, \delta, 2) - 2\sqrt{\|\mu\|^2 + \|\psi\|^2}$. From the conditions of the theorem, this value is lower than the lower bound on the optimal solution for $\sigma = 1$, and hence the optimal SVM solution will be achieved at this value. Thus the result follows, from the additional condition that $C(p, \delta, 2) + c(p, \delta, 2) \to 0$ as $p \to \infty$. The comparison of the worst group errors is carried out in Lemma 16.

□

Illustrating Theorem 6 using Gaussians.

- For sufficiently large p, $\big(X_{\max}(p, \delta, 2) - X_{\max}(p^\tau, \delta, 2)\big)^2$ gets arbitarily close to $(\sqrt{2\log\frac{p}{\delta}}) - \sqrt{2\tau\log\frac{p}{\delta}})^2$, which when simplified gives

  $$(\sqrt{2\log\frac{p}{\delta}} - \sqrt{2\tau\log\frac{p}{\delta}})^2 = 2\log\frac{p}{\delta}(1 + \tau - 2\sqrt{\tau})$$

  For sufficiently large p, $(2\|\psi\| + c(p, \delta, 2) + C(p^\tau, \delta, 2))^2$ gets arbitrarily close to $\log(\frac{p}{\delta})^\kappa$. Now if $\kappa < 2(1+\tau-2\sqrt{\tau})$ the condition $X_{\max}(p, \delta, 2) - X_{\max}(p^\tau, \delta, 2)\big) \geq 2\|\psi\| + c(p, \delta, 2) + C(p^\tau, \delta, 2)$ is satisfied.

- The objective value $F(\alpha)$ when $\alpha = 1$ is at most $2X_{\max}(p, \delta, 2) + 2C(p, \delta, 2) - 2\|\mu\|$. This expression simplifies to $2\big((\sqrt{2} - \sqrt{3}) \log p + C(p, \delta, 2)\big)$. For a sufficiently large p, the objective is negative. This implies that the data is perfectly separable in the invariant feature.
- We also need to check $\sqrt{\|\mu\|^2 + \|\psi\|^2} - \|\mu\| > C(p, \delta, 2) + c(p, \delta, 2)$. The expression in the LHS simplifies $\sqrt{\|\mu\|^2 + \|\psi\|^2} - \|\mu\| = \sqrt{\log \frac{p}{\delta}}(\sqrt{3 + \frac{\kappa}{4}} - \sqrt{3})$. The expression in the LHS is an increasing function of p and grows to infinity and the RHS decreases to zero. For sufficiently large p, the condition has to be satisfied. Finally, the ratio spurious feature to invariant feature is $\sqrt{\kappa/12}$.

Lemma 16 (Data balancing helps improve worst group error under heavy tails)**.** *Consider the same set of assumptions as in Theorem 6. With probability at least* $1 - 12\delta$, $\mathbf{wge}(\theta^*_{ss}) < \mathbf{wge}(\theta^*_{erm})$.

Proof. We start with analyzing the worst group error for the standard SVM solution, i.e., without any data balancing. Recall

$$-b^* = \frac{1}{2}(\sup_{x \in A}(\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top)x - \sup_{x \in -B}(\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top)x)$$

Let us try to bound $-b^*$. From the concentration condition and the fact that $p \geq n_0$, with probability at least $1 - \delta$, the first term above $\sup_{x \in A}(\alpha\hat{\mu} + \sigma\beta\hat{\psi})^\top x)$ lies in

$$[X_{\max}(p + m, \delta, 2) - c(p + m, \delta, 2), X_{\max}(p + m, \delta) + C(p + m, \delta, 2)]$$

The second term also lies in

$$[X_{\max}(p + m, \delta, 2) - c(p + m, \delta, 2), X_{\max}(p + m, \delta, 2) + C(p + m, \delta, 2)]$$

As a result, with probability $1 - 2\delta$, $-b^*$ is in

$$\left[\frac{-c(p + m, \delta, 2) - C(p + m, \delta, 2))}{2}, \frac{c(p + m, \delta, 2) + C(p + m, \delta, 2)}{2}\right]$$

We denote $a_{\min} = \frac{-c(p+m,\delta,2)-C(p+m,\delta,2))}{2}$ and $a_{\max} = \frac{c(p+m,\delta,2)+C(p+m,\delta,2))}{2}$

Consider a classifier $w^\top x + b$. We write the error for different groups. $\mathsf{Err}_{y,a}$ is the error for the group $g = (y, a)$.

$$\begin{aligned}
\mathsf{Err}_{1,-1} &= \mathbb{P}(w^\top X + b \leq 0 | X \sim D(\mu - \psi)) \\
\mathsf{Err}_{1,-1} &= \mathbb{P}(w^\top(\mu - \psi + \tilde{X}) - b \leq 0 | \tilde{X} \sim D(0)) \\
&= \mathbb{P}(w^\top X \leq w^\top(\psi - \mu) - b | \tilde{X} \sim D(0)) \\
&= F_W(w^\top(\psi - \mu) - b)
\end{aligned} \tag{30}$$

Denote $w^\top X = W$, F_W is the CDF of W. Also, observe that since $\tilde{X}$ is spherically symmetric, the distribution $w^\top X$ is the same as distribution of another $w^{',\top}\tilde{X}$, where $\|w\| = \|w^{'}\| = 1$. We now plug in the value of $-b^*$ for the max-margin classifier to arrive at the bounds for the error for each of the groups. We write

$$\begin{aligned}
&F_W(w^\top(\psi - \mu) + a_{\min}) \leq \mathsf{Err}_{1,-1} \leq F_W(w^\top(\psi - \mu) + a_{\max}) \\
&F_W(-\alpha\|\mu\| + \sigma\beta\|\psi\| + a_{\min}) \leq \mathsf{Err}_{1,-1} \leq F_W(-\alpha\|\mu\| + \sigma\beta\|\psi\| + a_{\max})
\end{aligned} \tag{31}$$

Similarly, we write

$$F_W(-\alpha\|\mu\| - \sigma\beta\|\psi\| + a_{\min}) \leq \mathsf{Err}_{1,1} \leq F_W(-\alpha\|\mu\| - \sigma\beta\|\psi\| + a_{\max}) \tag{32}$$

Observe that as p grows, $a_{\max}$ and $a_{\min}$ converge to zero (from Assumption 1). Also as p grows, from Theorem 6, we know that the optimal α approaches $\frac{\|\mu\|}{\sqrt{\|\mu\|^2 + \|\psi\|^2}}$ and $\sigma = -1$.

As a result, we can say that with probability at least $1-6\delta$, $\mathsf{Err}_{1,-1}$ and $\mathsf{Err}_{1,-1}$ approach the following quantities.

$$\mathsf{Err}_{1,-1} \to F_W\left((-\|\mu\|^2 - \|\psi\|^2)\frac{1}{\sqrt{\|\mu\|^2+\|\psi\|^2}}\right)$$

$$\mathsf{Err}_{1,1} \to F_W\left((-\|\mu\|^2 + \|\psi\|^2)\frac{1}{\sqrt{\|\mu\|^2+\|\psi\|^2}}\right)$$

We now turn our attention to the optimal SVM solution achieved after balancing the data. In this case, we throw the data out so all groups have same size m. In this case, the optimal b^* lies in the interval $a_{\min} = \frac{-c(2m,\delta,2)-C(2m,\delta,2))}{2}$ and $a_{\max} = \frac{c(2m,\delta,2)+C(2m,\delta,2))}{2}$

Observe that as m grows, $a_{\max}$ and $a_{\min}$ converge to zero (from Assumption 1). Also as m grows, from Theorem 7, we know that the optimal α approaches 1. We denote the error for a group y, a under balancing as $\mathsf{Err}^{\mathsf{bal}}_{y,a}$.

As a result, we can say that with probability at least $1-6\delta$, $\mathsf{Err}^{\mathsf{bal}}_{1,-1}$ and $\mathsf{Err}^{\mathsf{bal}}_{1,-1}$ approach the following quantities.

$$\mathsf{Err}^{\mathsf{bal}}_{1,-1} \to F_W(-\|\mu\|)$$

$$\mathsf{Err}^{\mathsf{bal}}_{1,1} \to F_W(-\|\mu\|)$$

We compare the error achieved by the two approaches. With probability $1-12\delta$ (We need to account for the joint probability that for imbalanced case the optimal solution is the spurious one and under the balanced case the optimal solution is the invariant one. From union bound it follows that at least one of them does not occur with probability at most 12δ).

$$\mathsf{Err}^{\mathsf{bal}}_{1,1} \to F_W(-\|\mu\|)$$

and

$$\mathsf{Err}_{1,1} \to F_W(-\|\mu\|)$$

We want to show

$$\mathsf{Err}^{\mathsf{bal}}_{1,1} < \mathsf{Err}_{1,1} \tag{33}$$

To show the above, is equivalent to showing

$$(-\|\mu\|^2 + \|\psi\|^2)\frac{1}{\sqrt{\|\mu\|^2+\|\psi\|^2}} > -\|\mu\|$$

Suppose $\|\psi\| \geq \|\mu\|$, then the LHS is non negative and RHS is non positive. Thus the claim is true in that case.

Suppose $\|\psi\| < \|\mu\|$, then both the LHS and RHS are negative. As a result, we want to show that

$$\frac{(-\|\mu\|^2 + \|\psi\|^2)^2}{\|\mu\|^2+\|\psi\|^2} < \|\mu\|^2 \tag{34}$$

Further simplification yields

$$\frac{(-\|\mu\|^2 + \|\psi\|^2)^2}{\|\mu\|^2+\|\psi\|^2} < \|\mu\|^2 \iff \|\psi\|^2(\|\psi\|^2 - 3\|\mu\|^2) < 0 \tag{35}$$

Since $\|\psi\| < \|\mu\|$, the above condition is satisfied. □

B.2. Higher-dimensional case

Lemma 17. *Let $x_1, \ldots, x_n$ be vectors in $\mathbb{R}^d$ drawn i.i.d from $N(0, I_d)$. Then, with probability $\geq 1-\delta$,*

$$\max_i \|x_i\| \leq \sqrt{d + 2\sqrt{d\log(n/\delta)} + 2\log(n/\delta)}$$

Proof. Observe that

$$\begin{aligned} &\|x_i\| \leq \sqrt{d + 2\sqrt{d\log(n/\delta)} + 2\log(n/\delta)}, \forall i \in \{1, \cdots, n\} \\ &\|x_i\|^2 \leq d + 2\sqrt{d\log(n/\delta)} + 2\log(n/\delta), \forall i \in \{1, \cdots, n\} \end{aligned} \tag{36}$$

We bound the probability of the above

$$\begin{aligned} &\mathbb{P}(\max_i \|x_i\|^2 \leq d + 2\sqrt{d\log(n/\delta)} + 2\log(n/\delta)) = \\ &1 - \mathbb{P}(\max_i \|x_i\|^2 \geq d + 2\sqrt{d\log(n/\delta)} + 2\log(n/\delta)) \geq 1 - n\mathbb{P}(\|x_i\|^2 \geq \\ &d + 2\sqrt{d\log(n/\delta)} + 2\log(n/\delta)) = 1 - ne^{-\log(n/\delta)} = 1 - \delta \end{aligned} \tag{37}$$

For the last step in the above, we leverage the fact that $\|x_i\|^2$ follows the Chi-square $\mathcal{X}^2(d)$ distribution and use the tail bound in Lemma 1 from (Laurent & Massart, 2000). □

Define $Q(n, \delta, d) = \sqrt{d + 2\sqrt{d\log(n/\delta)} + 2\log(n/\delta)}$

Lemma 18. *Let $x_1, \ldots, x_n$ be n i.i.d unit Gaussians with covariance I_d. Then for n large enough, with probability $\geq 1-\delta$, we have that for all directions $v \in \mathbb{R}^d$:*

$$\max_{i \in \{1, \cdots, n\}} \{v^\top x_i\} \leq b_n + a_n + a_n d \log\left(\frac{\frac{12Q(n,\delta/2,d)}{a_n} + 6}{\delta}\right),$$

$$\max_{i \in \{1, \cdots, n\}} \{v^\top x_i\} \geq b_n - a_n - a_n \log\left(d \log\left(\frac{\frac{12Q(n,\delta/2,d)}{a_n} + 6}{\delta}\right)\right)$$

where a_n and b_n are the constants in the Fisher-Tippett-Gnedenko theorem when applied to Gaussians.

Proof. Suppose $f(v) = \max_i v^\top x_i$ where v is a unit vector in $\mathbb{R}^d$. Then,

$$f(v) - f(u) = \max_i v^\top x_i - \max_i u^\top x_i \leq \max_i (v-u)^\top x_i \leq \|v-u\| \cdot \max_i \|x_i\|$$

where the first step follows from definition, the second step from subtracting a smaller quantity, and the last step from the Cauchy-Schwartz inequality.

From Lemma 17, with probability $\geq 1 - \delta/2$, $\max_i \|x_i\| \leq Q(n, \delta/2, d)$, which gives us:

$$f(v) - f(u) \leq Q(n, \delta/2, d) \cdot \|v-u\|$$

Now, we can build an ϵ-cover $C(\epsilon)$ over unit vectors on the circle so that successive vectors v_i and v_{i+1} have the property that $\|v_i - v_{i+1}\| \leq \epsilon$. The size of such an ϵ-cover is $N(\epsilon) \leq (\frac{2}{\epsilon} + 1)^d$ [2]; additionally, for any unit vector v in $\mathbb{R}^d$, there exists some v_i in the cover such that

$$f(v_i) - Q(n, \delta/2, d)\epsilon \leq f(v) \leq f(v_i) + Q(n, \delta/2, d)\epsilon$$

$f(v_i)$ is the maximum over n i.i.d. standard Gaussians $N(0, 1)$. From Lemma 8, we know

[2] https://www.stat.berkeley.edu/~bartlett/courses/2013spring-stat210b/notes/12notes.pdf

$$b_n - a_n \log\log(1/\delta) \leq f(v_i) \leq b_n + a_n \log(1/\delta)$$

Now we can apply Lemma 8 with $\delta = \frac{\delta}{6N(\epsilon)}$ plus an union bound over the cover $C(\epsilon)$ to get that for all v_i in the cover,

$$b_n - a_n \log\log(6N(\epsilon)/(\delta)) \leq f(v_i) \leq b_n + a_n \log(6N(\epsilon)/(\delta))$$

$$b_n - a_n \log\left(d\log\left(\frac{(\frac{12}{\epsilon}+6)}{\delta}\right)\right) \leq f(v_i) \leq b_n + a_n d\log\left(\frac{(\frac{12}{\epsilon}+6)}{\delta}\right)$$

For all directions $v \in \mathbb{R}^d$

$$b_n - a_n \log\left(d\log\left(\frac{(\frac{12}{\epsilon}+6)}{\delta}\right)\right) - Q(n,\delta/2,d)\epsilon \leq f(v) \leq b_n + a_n d\log\left(\frac{(\frac{12}{\epsilon}+6)}{\delta}\right) + Q(n,\delta/2,d)\epsilon$$

Plugging in $\epsilon = \frac{a_n}{Q(n,\delta/2,d)}$ in the above expression we get.

$$b_n - a_n - a_n \log\left(d\log\left(\frac{\frac{12Q(n,\delta/2,d)}{a_n}+6}{\delta}\right)\right) \leq f(v) \leq b_n + a_n + a_n d\log\left(\frac{\frac{12Q(n,\delta/2,d)}{a_n}+6}{\delta}\right)$$

□

Lemma 19. *Consider the density: $f(t) = \frac{3}{2}(1-t^2)$ for $t \in [0,1]$ and $f(t) = 0$ otherwise. Let F be the corresponding CDF and let $U(t) = F^{-1}(1-1/t)$. Then, the following facts hold:*

1. $$\lim_{h\to 0} \frac{1-F(1-xh)}{1-F(1-h)} = x$$

2. $1 - U(n) \geq \frac{2}{3n}$.

Proof. To see the first part, observe that:

$$1 - F(1-h) = \int_{1-h}^{1} \frac{3}{2}(1-t^2) = \frac{3}{2}(h - \frac{h^3 - 3h^2 + 3h}{3}$$

$$\lim_{h\to 0} \frac{1-F(1-xh)}{1-F(1-h)} = \frac{xh - \frac{(xh)^3 - 3(xh)^2 + 3xh}{3}}{h - \frac{h^3-3h^2+3h}{3}} = x$$

For the second part, we observe that from the definition of $U(n)$, we have that $1 - U(n) = h$, where:

$$\int_{1-h}^{1} \frac{3}{2}(1-t^2) = \frac{1}{n}$$

Observe that for small enough h (which corresponds to large enough n), the left hand side is at most $\frac{3}{2}h$. This implies that $h = 1 - U(n) \geq \frac{2}{3n}$ and the lemma follows. □

Lemma 20. *Consider the density: $f(t) = \frac{3}{2}(1-t^2)$ for $t \in [0,1]$ and $f(t) = 0$ otherwise. Let $x_1, \ldots, x_n$ be n drawn i.i.d from f and let $X_{\max} = \max(x_1, \ldots, x_n)$. Then for n large enough, with probability $\geq 1-\delta$, we have that:*

$$X_{\max} \leq 1, \quad X_{\max} \geq 1 - \left(\frac{2\log(2/\delta)}{3n}\right)$$

Proof. Observe that for this distribution, $x_F = 1$. From this, and the first part of Lemma 11, it follows that this distribution is of the Weibull type with $\alpha = 1$. From the Fisher-Tippett-Gnedenko Theorem, this means that the maximum of n points converges to $a_n Z + b_n$ in distribution, where $a_n = 1 - U(n)$, $b_n = 1$, and Z is a reverse Weibull distributed variable with $\alpha = 1$. Setting $X_{\max}(n, \delta) = 1$, we get that $C(n, \delta) = 0$.

To calculate $c(n, \delta)$, we observe that from the second part of Lemma 19, $a_n \geq \frac{2}{3n}$. Additionally, if Z is a reverse Weibull variable with parameter $\alpha = 1$, then,

$$\Pr(Z \leq -(\log(2/\delta))) = \exp(-(\log(2/\delta))) = \delta/2$$

Therefore, $\Pr\left(a_n Z + b_n \leq 1 - \left(\frac{2\log(2/\delta)}{3n}\right)\right) \leq \delta/2$. We get another $\delta/2$ from the distributional convergence of the maximum of n random variables to the limit for large enough n. □

Lemma 21. *Let $x_1, \ldots, x_n$ be n drawn i.i.d from symmetric uniform distribution centered at zero in $\mathbb{R}^3$. Then for n large enough, with probability $\geq 1 - \delta$, we have that for all directions $v \in \mathbb{R}^2$:*

$$\max_{i \in \{1, \cdots, n\}} \{v^\top x_i\} \leq 1$$

$$\max_{i \in \{1, \cdots, n\}} \{v^\top x_i\} \geq 1 - \frac{6\log(4n+2)/\delta)}{3n} - \frac{1}{n}$$

Proof. Suppose $f(v) = \max_i v^\top x_i$ where v is a unit vector in $\mathbb{R}^2$. Then,

$$f(v) - f(u) = \max_i v^\top x_i - \max_i u^\top x_i \leq \max_i (v-u)^\top x_i \leq \|v - u\| \cdot \max_i \|x_i\|$$

where the first step follows from definition, the second step from subtracting a smaller quantity, and the last step from the Cauchy-Schwartz inequality.

Note that $\max_i \|x_i\| \leq 1$, which gives us:

$$f(v) - f(u) \leq \cdot \|v - u\|$$

Now, we can build an ϵ-cover $C(\epsilon)$ over unit vectors on the circle so that successive vectors v_i and v_{i+1} have the property that $\|v_i - v_{i+1}\| \leq \epsilon$. The size of such an ϵ-cover is $N(\epsilon) = 1/\epsilon$; additionally, for any unit vector v in $\mathbb{R}^2$, there exists some v_i in the cover such that

$$f(v_i) - \epsilon \leq f(v) \leq f(v_i) + \epsilon$$

Observe that $f(v_i)$ is a maximum over n i.i.d. random variables drawn from a distribution $f(t) = \frac{3}{2}(1 - t^2)$ for $t \in [0, 1]$ and $f(t) = 0$ otherwise. Now we can apply Lemma 20 with $\delta = \frac{\delta}{N(\epsilon)}$ plus an union bound over the cover $C(\epsilon)$ to get that for all v_i in the cover,

$$1 - \frac{2\log(2N(\epsilon)/\delta)}{3n} \leq f(v_i) \leq 1$$

For all directions $v \in \mathbb{R}^3$

$$1 - \frac{2\log(2N(\epsilon)/\delta)}{3n} - \epsilon \leq f(v) \leq 1$$

Plugging in $\epsilon = \frac{1}{n}$ in the above expression we get.

$$1 - \frac{6\log(4n+2)/\delta)}{3n} - \frac{1}{n} \leq f(v) \leq 1$$

□

Lemma 22 (Approximate Maximization Lemma - I)**.** *Let $F(\alpha) = f(\alpha) + g(\alpha)$ where $g(\alpha) = \alpha u + \sqrt{\eta^2 - \alpha^2} v$, $u, v > 0$, and $f(\alpha)$ is an arbitrary function of α that lies in the interval $[-L, U]$. Let α_F be the value of α that maximizes $F(\alpha)$, and let $\alpha_g = \eta \frac{u}{\sqrt{u^2+v^2}}$ be the value of α that maximizes $g(\alpha)$. Then, the angle between $(\alpha_F, \sqrt{\eta^2 - \alpha_F^2})$ and $(\alpha_g, \sqrt{1 - \alpha_g^2})$ is at most $\cos^{-1}\left(1 - \frac{L+U}{\sqrt{u^2+v^2}}\right)$. Additionally, the maximum value of $F(\alpha)$ is at least $\eta\sqrt{u^2+v^2} - L$.*

Proof. For convenience, we can do a quick change of variables – we let $\alpha = \eta \cos\theta$. Then $g(\theta) = \eta(u\cos\theta + v\sin\theta)$, and is maximized at $\theta_g = \cos^{-1}\left(\frac{u}{\sqrt{u^2+v^2}}\right)$. This means we can re-write g as follows:

$$\begin{aligned} g(\theta) &= \eta\sqrt{u^2+v^2}\cdot(\cos\theta_g\cos\theta + \sin\theta_g\sin\theta) \\ &= \eta\sqrt{u^2+v^2}\cdot\cos(\theta_g - \theta) \end{aligned}$$

Similarly, we can do a change of variables on F and f as well. Suppose the value of θ that maximizes F is θ_F. Then we have that:

$$f(\theta_g) + \eta\sqrt{u^2+v^2} \leq f(\theta_F) + \eta\sqrt{u^2+v^2}\cos(\theta_g - \theta_F)$$

Since $f(\theta_g) \geq -L$ and $f(\theta_F) \leq U$, this gives us:

$$-L + \eta\sqrt{u^2+v^2} \leq U + \eta\sqrt{u^2+v^2}\cos(\theta_g - \theta_F)$$

The lemma follows from simple algebra. □

Lemma 23 (Approximate Maximixation Lemma - II)**.** *Let $F(\alpha) = f(\alpha) + g(\alpha)$ where $g(\alpha) = \alpha u - \sqrt{\eta^2 - \alpha^2}v$, $u, v > 0$, and $f(\alpha)$ is an arbitrary function of α that lies in the interval $[-L, U]$. Let α_F be the value of $\alpha \in [-\eta, \eta]$ that maximizes $F(\alpha)$, and let $\alpha_g = \eta$ be the value of α that maximizes $g(\alpha)$. Then, $\alpha_F \geq \eta - \frac{U+L}{u+v}$.*

Proof. To show the lemma, we observe that since $f(\alpha) \in [-L, U]$,

$$-L + \eta u \leq U + \alpha_F u - \sqrt{\eta^2 - \alpha_F^2}v$$

which implies $u(\eta - \alpha_F) + v\sqrt{\eta^2 - \alpha_F^2} \leq L + U$. This will hold when $\eta - \alpha_F \leq \frac{U+L}{u+v}$. The lemma follows. □

Recall that the SVM solution is stated as

$$w^* = \arg\max_{\|w\|=1} \inf_{x\in B} w^\top x - \sup_{x\in A} w^\top x$$

We rewrite $w = \alpha\hat{\mu} + \beta\hat{\psi} + \gamma^\top\hat{\Gamma}$, where $\hat{\mu}$, $\hat{\psi}$ denote unit vectors along μ, ψ respectively. $\hat{\Gamma} \in \mathbb{R}^{d\times d-2}$ is a matrix of $d-2$ vectors that span the subspace orthogonal to the subspace spanned μ and ψ. We introduce an additional parameter $\eta \in [0,1]$ and define a set as follows $\mathcal{S}_\eta = \{(\alpha,\beta,\gamma),\ \alpha^2+\beta^2 = \eta^2, \|\gamma\|^2 = 1-\eta^2\}$. Note that $\mathcal{S} = \cup_{n\in[0,1]}\mathcal{S}_\eta$ is the set of all vectors of norm 1. We divide the standard SVM optimization into an optimization over the set $\mathcal{S}_\eta$ and then choosing the best η.

$$\begin{aligned} \alpha^*(\eta) = &\arg\min_{\alpha\in[-\eta,\eta],\sigma\in\{-1,1\},\|\gamma\|=\sqrt{1-\eta^2}} \sup_{x\in A_\mu}\left\{(\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top(x-\mu)\right\} \\ &+ \sup_{x\in -B_\mu}\left\{(\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top(x-\mu)\right\}, \end{aligned} \tag{38}$$

where $\beta = \sqrt{1-\alpha^2}$. We compare the SVM objective for all η and pick the set of $\alpha^*(\eta)$ that lead to the optimal value.

Theorem 8. *[Gumbel type] Suppose $D(0)$ satisfies the concentration condition in Assumption 2 and $X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d) \geq 2\|\psi\| + c(p,\delta,d) + C(m,\delta,d)$. If $p \to \infty$, then with probability at least $1-4\delta$, the ERM solution converges to the spurious solution w^*_{spu}. In addition, with probability at least $1-12\delta$, $\mathbf{wge}(\theta_{ss}) < \mathbf{wge}(\theta_{erm})$.*

Proof. We fix an η and write the optimal solution for the η as

$$\begin{aligned} \alpha^*(\eta) = &\arg\min_{\alpha\in[-\eta,\eta],\sigma\in\{-1,1\},\|\gamma\|=\sqrt{1-\eta^2}} \sup_{x\in A_\mu}\left\{(\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top(x-\mu)\right\} \\ &+ \sup_{x\in -B_\mu}\left\{(\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top(x-\mu)\right\}, \end{aligned} \tag{39}$$

where $\beta = \sqrt{1-\alpha^2}$. We next consider a further split of the positive class into the majority and minority groups – A_μ^M and A_μ^m. This means that:

$$\sup_{x \in A_\mu} \boldsymbol{v}^\top x = \max\left(\sup_{x \in A_\mu^M} \boldsymbol{v}^\top x,\ \sup_{x \in A_\mu^m} \boldsymbol{v}^\top x\right),$$

and hence, we can write:

$$\begin{aligned}
&\sup_{x \in A_\mu} ((\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top (x-\mu) \\
&= \max\left(\sup_{x \in A_\mu^M} (\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top (x-\mu),\ \sup_{x \in A_\mu^m} ((\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top (x-\mu)\right) \\
&= \max\left(\sup_{x \in A_{\mu,\psi}^M} (\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top x - \alpha\|\mu\| + \sigma\beta\|\psi\|,\right. \\
&\left.\sup_{x \in A_{\mu,\psi}^m} (\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top x - \alpha\|\mu\| - \sigma\beta\|\psi\|\right),
\end{aligned}$$

and a similar expression will hold for B_μ. We look at two cases – $\sigma = 1$ and -1.

Case 1: $\sigma = 1$**.** Here, the SVM objective becomes:

$$\begin{aligned}
\min_{\alpha,\gamma} \Big\{ &\max(f_1(\alpha,\gamma) - \alpha\|\mu\| + \beta\|\psi\|, f_2(\alpha,\gamma) - \alpha\|\mu\| - \beta\|\psi\|) + \\
&\max(f_3(\alpha,\gamma) - \alpha\|\mu\| + \beta\|\psi\|, f_4(\alpha,\gamma) - \alpha\|\mu\| - \beta\|\psi\|)\Big\}
\end{aligned}$$

where $f_1(\alpha,\gamma) = \sup_{x \in A_{\mu,\psi}^M} (\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top x$, $f_2(\alpha,\gamma) = \sup_{x \in A_{\mu,\psi}^m} (\alpha\hat{\mu} + \sigma\beta\hat{\psi} + +\gamma^\top\hat{\Gamma})^\top x$, $f_3(\alpha,\gamma) = \sup_{x \in -B_{\mu,\psi}^M} (\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top x$, $f_4(\alpha,\gamma) = \sup_{x \in -B_{\mu,\psi}^m} (\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top x$. From conditions on the majority and the minority class, and the Concentration Condition with probability $\geq 1 - 4\delta$,

$$f_1(\alpha,\gamma), f_3(\alpha,\gamma) \in [X_{\max}(p,\delta,d) - c(p,\delta,d), X_{\max}(p,\delta,d) + C(p,\delta,d)], \tag{40}$$

and also:

$$f_2(\alpha,\gamma), f_4(\alpha,\gamma) \in [X_{\max}(m,\delta,d) - c(m,\delta,d), X_{\max}(m,\delta,d) + C(m,\delta,d)] \tag{41}$$

Observe that from the conditions of the theorem, the first terms will dominate for all values of α, and hence the SVM objective will become:

$$\begin{aligned}
&\min_{\alpha \in [-\eta,\eta], \|\gamma\| = \sqrt{1-\eta^2}} f_1(\alpha,\gamma) + f_3(\alpha,\gamma) - 2\alpha\|\mu\| + 2\beta\|\psi\| = \\
&\min_{\alpha \in [-\eta,\eta]} h(\alpha,\eta) - 2\alpha\|\mu\| + 2\beta\|\psi\|
\end{aligned} \tag{42}$$

where $h(\alpha,\eta) = \min_{\|\gamma\| = \sqrt{1-\eta^2}} f_1(\alpha,\gamma) + f_3(\alpha,\gamma)$. Observe that $h(\alpha,\eta) \in [2X_{\max}(p,\delta,d) - 2c(p,\delta,d), 2X_{\max}(p,\delta,d) + 2C(p,\delta,d)]$. From Lemma 23, the optimal solution $\alpha^* \geq \eta - \frac{C(p,\delta,d)+c(p,\delta,d)}{\|\mu\|+\|\psi\|}$, with an optimal value is lower bounded by $2X_{\max}(p,\delta,d) - 2c(p,\delta,d) - 2\eta\|\mu\|$.

Case 2: $\sigma = -1$**.** Here, the SVM objective becomes:

$$\begin{aligned}
\min_{\alpha \in [-\eta,\eta], \|\gamma\| = \sqrt{1-\eta^2}} \Big\{ &\max(f_1(\alpha,\gamma) - \alpha\|\mu\| - \beta\|\psi\|, f_2(\alpha,\gamma) - \alpha\mu + \beta\psi) + \\
&\max(f_3(\alpha,\gamma) - \alpha\|\mu\| - \beta\|\psi\|, f_4(\alpha,\gamma) - \alpha\|\mu\| + \beta\|\psi\|)\Big\}
\end{aligned}$$

This time, from the conditions of the theorem, the first terms will dominate the maximum for all values of α, and hence the objective will become:

$$\begin{aligned}&\min_{\alpha\in[-\eta,\eta],\|\gamma\|=\sqrt{1-\eta^2}} f_1(\alpha,\gamma)+f_3(\alpha,\gamma)-2\alpha\|\mu\|-2\beta\|\psi\|=\\&\min_{\alpha\in[-\eta,\eta]} h(\alpha,\eta)-2\alpha\|\mu\|-2\beta\|\psi\|\end{aligned} \tag{43}$$

Observe that $h(\alpha,\eta)\in[2X_{\max}(p,\delta,d)-2c(p,\delta,d),2X_{\max}(p,\delta,d)+2C(p,\delta,d)]$. From Lemma 22, the optimal solution $(\alpha,\sqrt{1-\alpha^2})$ will be close to the spurious solution vector $(\eta\frac{\|\mu\|}{\sqrt{\|\mu\|^2+\|\psi\|^2}},\eta\frac{\|\psi\|}{\sqrt{\|\mu\|^2+\|\psi\|^2}})$, with the angle being at most $\cos^{-1}(1-\frac{C(p,\delta,d)+c(p,\delta,d)}{\sqrt{\|\mu\|^2+\|\psi\|^2}})$. The optimal solution value will be at most $2X_{\max}(p,\delta,d)+2C(p,\delta,d)-2\eta\sqrt{\|\mu\|^2+\|\psi\|^2}$. From the conditions of the theorem, this value is lower than the lower bound on the optimal solution for $\sigma=1$, and hence the optimal SVM solution for a fixed η will be achieved at this value. We now compare the lower bound for the optimal η with the upper bound of the optimal value at $\eta=1$.

$$\begin{aligned}&2X_{\max}(p,\delta,d)-2C(p,\delta,d)-2\eta\sqrt{\|\mu\|^2+\|\psi\|^2}\le 2X_{\max}(p,\delta,d)+2C(p,\delta,d)-2\sqrt{\|\mu\|^2+\|\psi\|^2}\\&\eta\ge 1-\frac{C(p,\delta,d)}{\sqrt{\|\mu\|^2+\|\psi\|^2}}\end{aligned} \tag{44}$$

The result follows from the additional condition that $C(p,\delta,d)+c(p,\delta,d)\to 0$ as $p\to\infty$. The comparison of worst group errors is carried out in Lemma 24 □

Lemma 24 (Data balancing helps improve worst group error under heavy tails)**.** *Consider the same set of assumptions as Theorem 8. With probability at least* $1-12\delta$, $\mathbf{wge}(\theta^*_{ss})<\mathbf{wge}(\theta^*_{erm})$.

Proof. We start with analyzing the worst group error for the standard SVM solution, i.e., without any data balancing. Recall

$$-b^*=\frac{1}{2}(\sup_{x\in A}(\alpha\hat{\mu}+\sigma\beta\hat{\psi}+\gamma^\top\hat{\Gamma})^\top x-\sup_{x\in -B}(\alpha\hat{\mu}+\sigma\beta\hat{\psi}+\gamma^\top\hat{\Gamma})x)$$

Let us try to bound $-b^*$. From the concentration condition and the fact that $p\ge n_0$, with probability at least $1-\delta$, the first term above $\sup_{x\in A}(\alpha\hat{\mu}+\sigma\beta\hat{\psi}+\gamma^\top\hat{\Gamma})^\top x)$ lies in

$$[X_{\max}(p+m,\delta,d)-c(p+m,\delta,d),X_{\max}(p+m,\delta,d)+C(p+m,\delta,d)]$$

The second term also lies in

$$[X_{\max}(p+m,\delta,d)-c(p+m,\delta,d),X_{\max}(p+m,\delta,d)+C(p+m,\delta,d)]$$

As a result, with probability $1-2\delta$, $-b^*$ is in

$$\left[\frac{-c(p+m,\delta,d)-C(p+m,\delta,d))}{2},\frac{c(p+m,\delta,d)+C(p+m,\delta,d)}{2}\right]$$

We denote $a_{\min}=\frac{-c(p+m,\delta,d)-C(p+m,\delta,d))}{2}$ and $a_{\max}=\frac{c(p+m,\delta,d)+C(p+m,\delta,d))}{2}$

Consider a classifier $w^\top x+b$. We write the error for different groups. $\mathsf{Err}_{y,a}$ is the error for the group $g=(y,a)$.

$$\begin{aligned}\mathsf{Err}_{1,-1}&=\mathbb{P}(w^\top X+b\le 0|X\sim D(\mu-\psi))\\\mathsf{Err}_{1,-1}&=\mathbb{P}(w^\top(\mu-\psi+\tilde{X})+b\le 0|\tilde{X}\sim D(0))\\&=\mathbb{P}(w^\top X\le w^\top(\psi-\mu)-b|\tilde{X}\sim D(0))\\&=F_W(w^\top(\psi-\mu)+b)\end{aligned} \tag{45}$$

Denote $w^\top X = W$, F_W is the CDF of W. Also, observe that since X is spherically symmetric, the distribution $w^\top X$ is the same as distribution of another $w^{',\top} X$, where $\|w\| = \|w^{'}\| = 1$. We now plug in the value of b^* for the max-margin classifier to arrive at the bounds for the error for each of the groups. We write

$$\begin{aligned} &F_W(w^\top(\psi-\mu)+a_{\min}) \leq \mathsf{Err}_{1,-1} \leq F_W(w^\top(\psi-\mu)+a_{\max}) \\ &F_W(-\alpha\|\mu\|+\sigma\beta\|\psi\|+a_{\min}) \leq \mathsf{Err}_{1,-1} \leq F_W(-\alpha\|\mu\|+\sigma\beta\|\psi\|+a_{\max}) \end{aligned} \tag{46}$$

Similarly, we write

$$F_W(-\alpha\|\mu\|-\sigma\beta\|\psi\|+a_{\min}) \leq \mathsf{Err}_{1,1} \leq F_W(-\alpha\|\mu\|-\sigma\beta\|\psi\|+a_{\max}) \tag{47}$$

Observe that as p grows, $a_{\max}$ and $a_{\min}$ converge to zero (from Assumption 1). Also as p grows, from Theorem 8, we know that the optimal α approaches $\frac{\|\mu\|}{\sqrt{\|\mu\|^2+\|\psi\|^2}}$ and $\sigma = -1$ and $\gamma = 0$.

As a result, we can say that with probability at least $1-6\delta$, $\mathsf{Err}_{1,-1}$ and $\mathsf{Err}_{1,-1}$ approach the following quantities.

$$\mathsf{Err}_{1,-1} \to F_W\left((-\|\mu\|^2-\|\psi\|^2)\frac{1}{\sqrt{\|\mu\|^2+\|\psi\|^2}}\right)$$

$$\mathsf{Err}_{1,1} \to F_W\left((-\|\mu\|^2+\|\psi\|^2)\frac{1}{\sqrt{\|\mu\|^2+\|\psi\|^2}}\right)$$

We now turn our attention to the optimal SVM solution achieved after balancing the data. In this case, we throw the data out so all groups have same size m. In this case, the optimal b^* lies in the interval $a_{\min} = \frac{-c(2m,\delta,d)-C(2m,\delta,d))}{2}$ and $a_{\max} = \frac{c(2m,\delta,d)+C(2m,\delta,d))}{2}$

Observe that as m grows, $a_{\max}$ and $a_{\min}$ converge to zero (from Assumption 1). Also as m grows, from Theorem 9, we know that the optimal α approaches 1. We denote the error for a group y, a under balancing as $\mathsf{Err}^{\mathsf{bal}}_{y,a}$.

As a result, we can say that with probability at least $1-6\delta$, $\mathsf{Err}^{\mathsf{bal}}_{1,-1}$ and $\mathsf{Err}^{\mathsf{bal}}_{1,-1}$ approach the following quantities.

$$\mathsf{Err}^{\mathsf{bal}}_{1,-1} \to F_W(-\|\mu\|)$$

$$\mathsf{Err}^{\mathsf{bal}}_{1,1} \to F_W(-\|\mu\|)$$

We compare the error achieved by the two approaches. With probability $1-12\delta$ (We need to account for the joint probability that for imbalanced case the optimal solution is the spurious one and under the balanced case the optimal solution is the invariant one. From union bound it follows that at least one of them does not occur with probability at most 12δ).

$$\mathsf{Err}^{\mathsf{bal}}_{1,1} \to F_W(-\|\mu\|)$$

and

$$\mathsf{Err}_{1,1} \to F_W(-\|\mu\|)$$

We want to show

$$\mathsf{Err}^{\mathsf{bal}}_{1,1} < \mathsf{Err}_{1,1} \tag{48}$$

To show the above, is equivalent to showing

$$(-\|\mu\|^2+\|\psi\|^2)\frac{1}{\sqrt{\|\mu\|^2+\|\psi\|^2}} > -\|\mu\|$$

Suppose $\|\psi\| \geq \|\mu\|$, then the LHS is non negative and RHS is negative. Thus the claim is true in that case.

Suppose $\|\psi\| < \|\mu\|$, then both the LHS and RHS are negative. As a result, we want to show that

$$\frac{(-\|\mu\|^2 + \|\psi\|^2)^2}{\|\mu\|^2 + \|\psi\|^2} < \|\mu\|^2 \tag{49}$$

Further simplification yields

$$\frac{(-\|\mu\|^2 + \|\psi\|^2)^2}{\|\mu\|^2 + \|\psi\|^2} < \|\mu\|^2 \iff \|\psi\|^2(\|\psi\|^2 - 3\|\mu\|^2) < 0 \tag{50}$$

Since $\|\psi\| < \|\mu\|$, the above condition is satisfied. □

Theorem 9. *[Weibull type] Suppose $D(0)$ satisfies the concentration condition in Assumption 2 and as p and $m \to \infty$*

$$\frac{X_{\max}(p, \delta, d) - X_{\max}(m, \delta, d)}{2\|\psi\|} \to 0,$$

$$\frac{X_{\max}(p, \delta, d) - X_{\max}(m, \delta, d)}{2\|\mu\|} \to 0,$$

and $\frac{C(m,\delta,d)+c(m,\delta,d)+\frac{1}{2}\sqrt{2C(m,\delta,d)+c(m,\delta,d)\|\psi\|}}{\|\mu\|} \to 0$. If $m, p \to \infty$, then with probability at least $1 - 4\delta$, the ERM solution converges to the invariant solution w^*_{inv}.

Proof. We fix an η and write the optimal solution for the η as

$$\begin{aligned}
&\alpha^*(\eta) = \\
&\quad \underset{\alpha\in[-\eta,\eta],\sigma\in\{-1,1\},\|\gamma\|=\sqrt{1-\eta^2}}{\arg\min} \sup_{x\in A_\mu} \Big\{(\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top(x-\mu) + \\
&\quad \sup_{x\in -B_\mu} (\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top(x-\mu)\Big\}
\end{aligned} \tag{51}$$

where $\beta = \sqrt{1-\alpha^2}$. We next consider a further split of the positive class into the majority and minority groups – A^M_μ and A^m_μ. This means that:

$$\sup_{x\in A_\mu} \boldsymbol{v}^\top x = \max\left(\sup_{x\in A^M_\mu} \boldsymbol{v}^\top x, \sup_{x\in A^m_\mu} \boldsymbol{v}^\top x\right),$$

and hence, we can write:

$$\begin{aligned}
&\sup_{x\in A_\mu} ((\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top(x-\mu) \\
&= \max\left(\sup_{x\in A^M_\mu} ((\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top(x-\mu), \sup_{x\in A^m_\mu} ((\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top(x-\mu)\right) \\
&\max\left(\sup_{x\in A^M_{\mu,\psi}} ((\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top x - \alpha\|\mu\| + \right. \\
&\left. \sigma\beta\|\psi\|, \sup_{x\in A^m_{\mu,\psi}} ((\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top x - \alpha\|\mu\| - \sigma\beta\|\psi\|\right)
\end{aligned}$$

and a similar expression will hold for B_μ. We look at two cases – $\sigma = 1$ and -1.

Case 1: $\sigma = 1$**.** Here, the SVM objective becomes:

$$\min_{\alpha,\gamma} \Big\{ \max(f_1(\alpha,\gamma) - \alpha\|\mu\| + \beta\|\psi\|, f_2(\alpha,\gamma) - \alpha\|\mu\| - \beta\|\psi\|) + \\ \max(f_3(\alpha,\gamma) - \alpha\|\mu\| + \beta\|\psi\|, f_4(\alpha,\gamma) - \alpha\|\mu\| - \beta\|\psi\|) \Big\}$$

where $f_1(\alpha,\gamma) = \sup_{x\in A^M_{\mu,\psi}}(\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top x$, $f_2(\alpha,\gamma) = \sup_{x\in A^m_{\mu,\psi}}(\alpha\hat{\mu} + \sigma\beta\hat{\psi} + +\gamma^\top\hat{\Gamma})^\top x$, $f_3(\alpha,\gamma) = \sup_{x\in -B^M_{\mu,\psi}}(\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top x$, $f_4(\alpha,\gamma) = \sup_{x\in -B^m_{\mu,\psi}}(\alpha\hat{\mu} + \sigma\beta\hat{\psi} + \gamma^\top\hat{\Gamma})^\top x$. From conditions on the majority and the minority class, and the Concentration Condition with probability $\geq 1 - 4\delta$,

$$f_1(\alpha,\gamma), f_3(\alpha,\gamma) \in [X_{\max}(p,\delta,d) - c(p,\delta,d), X_{\max}(p,\delta,d) + C(p,\delta,d)], \tag{52}$$

and also:

$$f_2(\alpha,\gamma), f_4(\alpha,\gamma) \in [X_{\max}(m,\delta,d) - c(m,\delta,d), X_{\max}(m,\delta,d) + C(m,\delta,d)] \tag{53}$$

We now look at two possible cases for α to determine what the inside maximum will look like. The first case is for large β – where

$$\beta \geq \frac{C(m,\delta,d) - c(n,\delta,d) - (X_{max}(p,\delta,d) - X_{\max}(m,\delta,d))}{2\|\psi\|}$$

and the objective simplifies to.

$$F(\alpha) = \min_{\alpha} f_1(\alpha,\gamma) + f_3(\alpha,\gamma) - 2\alpha\|\mu\| + 2\beta\|\psi\|,$$

$$\begin{aligned} &\min_{\alpha\in[-\eta,\eta],\|\gamma\|=\sqrt{1-\eta^2}} f_1(\alpha,\gamma) + f_3(\alpha,\gamma) - 2\alpha\|\mu\| + 2\beta\|\psi\| = \\ &\min_{\alpha\in[-\eta,\eta]} h(\alpha,\eta) - 2\alpha\|\mu\| + 2\beta\|\psi\| \end{aligned} \tag{54}$$

where $h(\alpha,\eta) = \min_{\|\gamma\|=\sqrt{1-\eta^2}} f_1(\alpha,\gamma) + f_3(\alpha,\gamma)$. Observe that $h(\alpha,\eta) \in [2X_{\max}(p,\delta,d) - 2c(p,\delta,d), 2X_{\max}(p,\delta,d) + 2C(p,\delta,d)]$. From Lemma 23, the optimal solution $\alpha^* \geq \eta - \frac{C(p,\delta,d)+c(p,\delta,d)}{\|\mu\|+\|\psi\|}$, with an optimal value is lower bounded by $2X_{\max}(p,\delta,d) - 2c(p,\delta,d) - 2\eta\|\mu\|$.

The other case is where β is small – namely, $\beta \leq \frac{C(m,\delta,d)-c(p,\delta,d)}{2\|\psi\|}$ (here we use the fact that $X_{\max}(p,\delta,d) \geq X_{\max}(m,\delta,d)$. Since $\alpha^2 + \beta^2 = \eta^2$, this implies that α is now close to 1 – specifically, $\alpha \geq \sqrt{\eta^2 - \left(\frac{C(m,\delta,d)-c(p,\delta,d)}{2\|\psi\|})\right)^2} \to 1$ as $p \to \infty$ by the conditions of the theorem. This means that in both cases, the inner maximum is achieved when α is close to η.

Case 2: $\sigma = -1$**.** In this case, the SVM objective becomes:

$$F(\alpha) = \min_{\alpha,\gamma} \Big\{ \max(f_1(\alpha,\gamma) - \alpha\|\mu\| - \beta\|\psi\|, f_2(\alpha,\gamma) - \alpha\|\mu\| + \beta\|\psi\|) + \\ \max(f_3(\alpha,\gamma) - \alpha\|\mu\| - \beta\|\psi\|, f_4(\alpha,\gamma) - \alpha\|\mu\| + \beta\|\psi\|) \Big\}$$

We again do a case by case analysis. We say that β is large if

$$\beta \geq \beta_{\mathsf{th}} = \frac{X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d) + C(p,\delta,d) + c(m,\delta,d)}{2\|\psi\|}$$

In this case, the SVM objective becomes:

$$F(\alpha) = f_2(\alpha, \gamma) + f_4(\alpha, \gamma) - 2\|\mu\|\alpha + 2\beta\|\psi\|$$

As a result of the above, $-\sqrt{1-\beta_{\mathsf{th}}^2} \leq \alpha \leq \sqrt{1-\beta_{\mathsf{th}}^2}$. We divide the analysis into two cases. Case 1. $\eta < \sqrt{1-\beta_{\mathsf{th}}^2}$ From Lemma 23, the solution to this is $\alpha \geq \eta - \frac{C(m,\delta,d)+c(m,\delta,d)}{\|\mu\|+\|\psi\|} \to \eta$ as $m \to \infty$. Case 2. $\eta \geq \sqrt{1-\beta_{\mathsf{th}}^2}$ From Lemma 23, the solution to this is $\alpha \geq \sqrt{1-\beta_{\mathsf{th}}^2} - \frac{C(m,\delta,d)+c(m,\delta,d)}{\|\mu\|+\|\psi\|} \to 1$ as $m \to \infty$.

Now we are left with analyzing the setting when β is small, where

$$\beta \leq \frac{X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d) + C(p,\delta,d) + c(m,\delta,d)}{2\|\psi\|}$$

Since $\alpha^2 + \beta^2 = \eta^2$, here α by definition satisfies

$$\alpha \geq \sqrt{\eta^2 - \left(\frac{X_{\max}(p,\delta,d) - X_{\max}(m,\delta,d) + C(p,\delta,d) + c(m,\delta,d)}{4\|\psi\|}\right)^2} \to \eta$$

This means that in all four cases, the inner maximum is achieved when α is close to η. We now compare the lower bound for optimal value achieved by η with the upper bound on the objective for $\alpha = \sqrt{1-\beta_{\mathsf{th}}^2}$ to show that η approaches 1 as $m \to \infty$.

For a fixed η, α takes value arbitrarily close to η. A lower bound on the SVM objective when $\alpha \in [-\eta, \eta]$ and $\eta \leq \sqrt{1-\beta_{\mathsf{th}}^2}$ is

$$\begin{aligned}
&2X_{\max}(m,\delta,d) - 2c(m,\delta,d) - 2\eta\|\mu\| - \beta\|\psi\| = \\
&2X_{\max}(m,\delta,d) - 2c(m,\delta,d) - 2\eta\|\mu\| - \sqrt{2\eta\frac{C(m,\delta,d)+c(m,\delta,d)}{\|\mu\|+\|\psi\|}}\|\psi\| \leq \\
&2X_{\max}(m,\delta,d) - 2c(m,\delta,d) - 2\eta\|\mu\| - \sqrt{2C(m,\delta,d)+c(m,\delta,d)\|\psi\|}
\end{aligned}$$

When $\alpha = \sqrt{1-\beta_{\mathsf{th}}^2}$ the SVM objective can be at most

$$2X_{\max}(m,\delta,d) + 2C(m,\delta,d) - 2\sqrt{1-\beta_{\mathsf{th}}^2}\|\mu\| + 2\beta_{\mathsf{th}}\|\psi\|$$

Comparing the above to the lower bound on the SVM objective we get

$$\begin{aligned}
&2X_{\max}(m,\delta,d) - 2c(m,\delta,d) - 2\eta\|\mu\| - \sqrt{2C(m,\delta,d)+c(m,\delta,d)\|\psi\|} < \\
&2X_{\max}(m,\delta,d) + 2C(m,\delta,d) - 2\sqrt{1-\beta_{\mathsf{th}}^2}\|\mu\| + 2\beta_{\mathsf{th}}\|\psi\| \\
&\eta \geq \sqrt{1-\beta_{\mathsf{th}}^2} - \frac{C(m,\delta,d)+c(m,\delta,d)+\beta_{\mathsf{th}}\|\psi\| + \frac{1}{2}\sqrt{2C(m,\delta,d)+c(m,\delta,d)\|\psi\|}}{\|\mu\|}
\end{aligned} \tag{55}$$

Owing to the conditions in the theorem, as $m \to \infty$, $\eta \to 1$. □

Theorem 10. *The ideal invariant classifier $\hat{\mu}$ achieves the minimum worst group error.*

Proof. We write down the error expressions for the four groups as follows. We consider a general classifier $w^\top x + b$, where $\|w\| = 1$.

$$\begin{aligned}
\mathsf{Err}_{1,-1} &= \mathbb{P}(w^\top X + b \leq 0 | X \sim D(\mu - \psi)) \\
\mathsf{Err}_{1,-1} &= F_W(w^\top(\psi - \mu) - b)
\end{aligned} \tag{56}$$

$$\begin{aligned}
\mathsf{Err}_{1,1} &= \mathbb{P}(w^\top X + b \leq 0 | X \sim D(\mu + \psi)) \\
\mathsf{Err}_{1,1} &= F_W(w^\top(-\psi - \mu) - b)
\end{aligned} \tag{57}$$

$$\begin{aligned}\mathsf{Err}_{-1,1} &= \mathbb{P}(w^\top X + b \geq 0 | X \sim D(-\mu + \psi)) \\ &= \mathbb{P}(w^\top(\tilde{X} - \mu + \psi) + b \geq 0 | \tilde{X} \sim D(0)) \\ &= \mathbb{P}(w^\top \tilde{X} \geq w^\top(\mu - \psi) - b | \tilde{X} \sim D(0)) \\ &= F_W(w^\top(\psi - \mu) + b)\end{aligned} \tag{58}$$

In the above simplification, we exploit the fact that $\tilde{X}$ is symmetric and as a result the distribution of $\tilde{X}$ is same as $-\tilde{X}$.

$$\begin{aligned}\mathsf{Err}_{-1,-1} &= \mathbb{P}(w^\top X + b \geq 0 | X \sim D(-\mu - \psi)) \\ &= F_W(w^\top(-\psi - \mu) + b)\end{aligned} \tag{59}$$

- Case 1. $w^\top \psi \geq 0$, $b \geq 0$. In this case, observe that $\mathsf{Err}_{-1,1}$ achieves the worst group error. Observe that F_W is monotonic in b so $b = 0$ is optimal. Now we want to minimize $F_W(w^\top(\psi - \mu)$ subject to $\|w\| = 1$ and $w^\top \psi \geq 0$. The first term takes smallest value when $w^\top(\psi) = 0$ and second term takes smallest value when $-w^\top \mu = -\|\mu\|$. If $w = \hat{\mu}$, then both constraints are simultaneously satisfied as $\mu \perp \psi$. The error achieved as a result is $F_W(-\|\mu\|)$
- Case 2. $w^\top \psi \leq 0$, $b \geq 0$. In this case, observe that $\mathsf{Err}_{-1,-1}$ achieves the worst group error. Observe that F_W is monotonic in b so $b = 0$ is optimal. Now we want to minimize $F_W(w^\top(-\psi - \mu)$ subject to $\|w\| = 1$ and $w^\top \psi \leq 0$. The first term takes smallest value when $w^\top(\psi) = 0$ and second term takes smallest value when $-w^\top \mu = -\|\mu\|$. If $w = \hat{\mu}$, then both constraints are simultaneously satisfied as $\mu \perp \psi$. The error achieved as a result is $F_W(-\|\mu\|)$
- Case 3. $w^\top \psi \geq 0$, $b \leq 0$. In this case, observe that $\mathsf{Err}_{1,-1}$ achieves the worst group error. Observe that F_W is monotonic in b so $b = 0$ is optimal. Now we want to minimize $F_W(w^\top(\psi - \mu)$ subject to $\|w\| = 1$ and $w^\top \psi \geq 0$. The first term takes smallest value when $w^\top(\psi) = 0$ and second term takes smallest value when $-w^\top \mu = -\|\mu\|$. If $w = \hat{\mu}$, then both constraints are simultaneously satisfied as $\mu \perp \psi$. The error achieved as a result is $F_W(-\|\mu\|)$
- Case 4. $w^\top \psi \leq 0$, $b \leq 0$. In this case, observe that $\mathsf{Err}_{1,1}$ achieves the worst group error. Observe that F_W is monotonic in b so $b = 0$ is optimal. Now we want to minimize $F_W(w^\top(-\psi - \mu)$ subject to $\|w\| = 1$ and $w^\top \psi \leq 0$. The first term takes smallest value when $w^\top(\psi) = 0$ and second term takes smallest value when $-w^\top \mu = -\|\mu\|$. If $w = \hat{\mu}$, then both constraints are simultaneously satisfied as $\mu \perp \psi$. The error achieved as a result is $F_W(-\|\mu\|)$

Therefore, $F_W(-\|\mu\|)$ is the lowest value for the error and is achieved by $w = \hat{\mu}$. In fact, if the cdf of F_W is strictly increasing, then $w = \hat{\mu}$ is the unique optimal solution. □

C. Supplementary Materials for Empirical Findings

C.1. Training details for the experiments

The training procedure consists of two steps. We use the training strategy very similar to that in (Kirichenko et al., 2022). We process CelebA and Waterbirds dataset using the procedure used in (Idrissi et al., 2021). We train in Pytorch using the same environment from (Idrissi et al., 2021) provided at `https://github.com/facebookresearch/BalancingGroups`.

We first explain training of ERM and SS.

- **Feature Learning** We take a pretrained ResNet-50 and fine tune a fresh linear layer on the target data (Waterbirds or CelebA). We use Adam optimizer with a learning rate of 10^{-4} and a weight decay of 10^{-3} and train for 10 epochs with a batch size of 128.
- **Linear Layer Learning** In this step, we train a fresh linear layer. The only difference between ERM and SS is that SS is trained on a balanced dataset obtained by subsampling. We use Adam optimizer with a learning rate of 10^{-2} and train for 100 epochs with a batch size of 128.

In ERM-PCA and SS-PCA the first step is exactly the same. Before the second step of linear layer learning, we carry out PCA on the representations input to the last linear layer and retain the first four components as they explain 99 percent of variance in the data. After this we carry out the second step with same parameters as above.

C.2. Supplementary figures

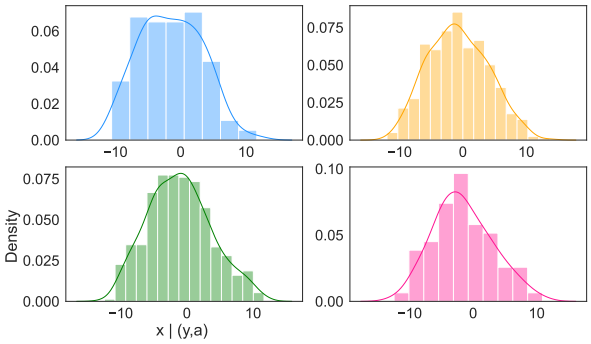

Figure 3: Waterbirds: Distribution of the second highest PCA feature.

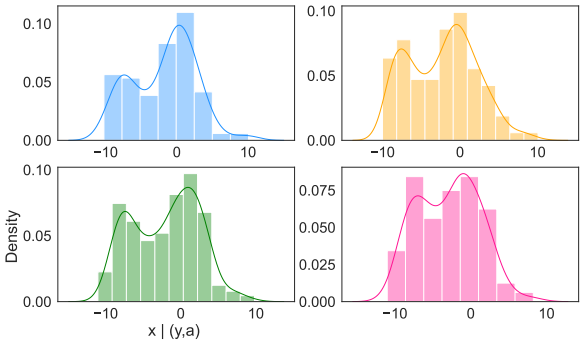

Figure 4: Waterbirds: Distribution of the third highest PCA feature.

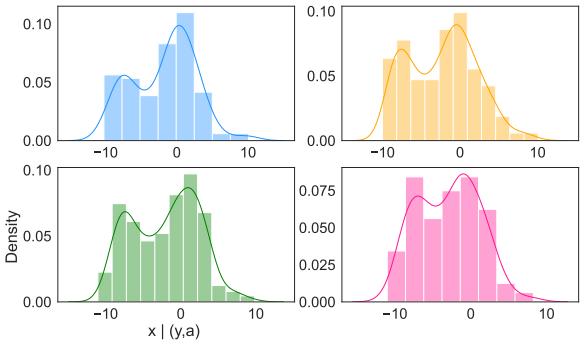

Figure 5: Waterbirds: Distribution of the fourth highest PCA feature

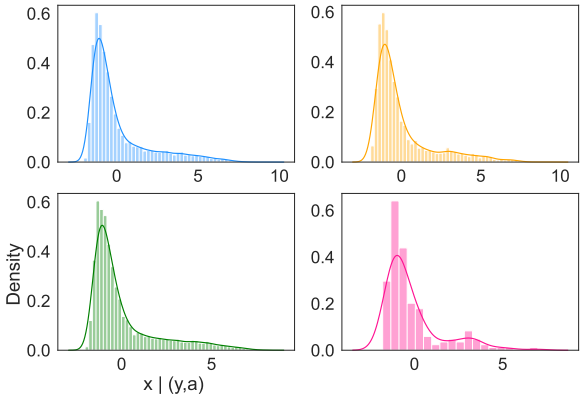

Figure 6: CelebA: Distribution of the highest PCA feature.

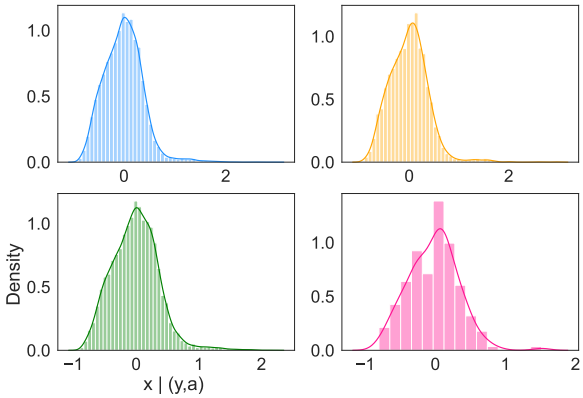

Figure 7: CelebA: Distribution of the second highest PCA feature.

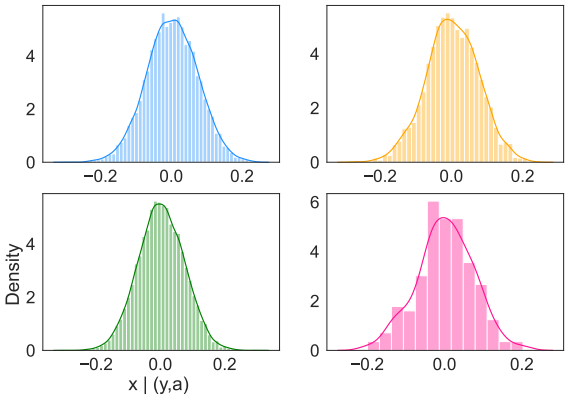

Figure 8: CelebA: Distribution of the third highest PCA feature.

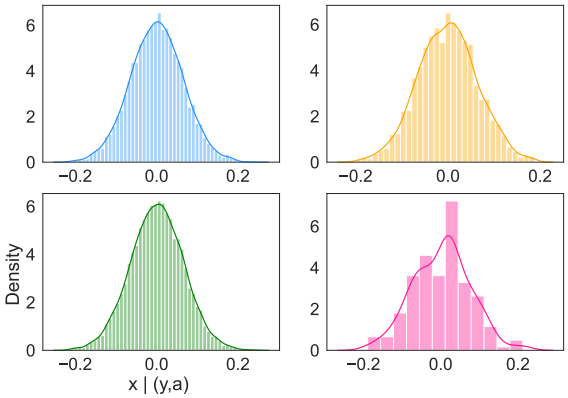

Figure 9: CelebA: Distribution of the fourth highest PCA feature.