Provable Tradeoffs in Adversarially Robust Classification

Edgar Dobriban, *Member, IEEE,* Hamed Hassani, *Member, IEEE,* David Hong, *Member, IEEE,* and Alexander Robey

***Abstract*—It is well known that machine learning methods can be vulnerable to adversarially-chosen perturbations of their inputs. Despite significant progress in the area, foundational open problems remain. In this paper, we address several key questions. We derive exact and approximate *Bayes-optimal robust classifiers* for the important setting of two- and three-class Gaussian classification problems with arbitrary imbalance, for ℓ_2 and ℓ_∞ adversaries. In contrast to classical Bayes-optimal classifiers, determining the optimal decisions here cannot be made pointwise and new theoretical approaches are needed. We develop and leverage new tools, including recent breakthroughs from probability theory on robust isoperimetry, which, to our knowledge, have not yet been used in the area. Our results reveal fundamental tradeoffs between standard and robust accuracy that grow when data is imbalanced. We also show further results, including an analysis of classification calibration for convex losses in certain models, and finite sample rates for the robust risk.**

***Index Terms*—adversarial robustness, provable tradeoffs, class imbalance, Gaussian mixtures.**

I. Introduction

MACHINE learning methods, such as deep neural nets, have shown remarkable performance in numerous application domains ranging from computer vision to natural language processing (see, e.g., [1]). However, despite this documented success, it is now well-known that many of these methods are also highly vulnerable to adversarial attacks. Indeed, it has been repeatedly shown that adversarially-chosen, imperceptible changes to the input data at test time can have undesirable effects on the predictions of models that otherwise perform well. For example, imperceptible pixel-wise changes to images are known to severely degrade the performance of state-of-the-art image classifiers [2], [3].

Adversarial training methods (e.g., [4]–[8]) tackle this problem by seeking models that are *robust* to adversarial attacks. A common approach is to replace the standard risk used to assess classifier performance with a *robust risk* that incorporates the possibility of small perturbations to the input. To illustrate this approach, consider the classification problem of assigning labels $y \in \mathcal{C}$ to input vectors (e.g., images) $x \in \mathbb{R}^p$. Traditional, non-adversarial training techniques seek classifiers $\hat{y} : \mathbb{R}^p \to \mathcal{C}$ that minimize the standard risk (misclassification probability)[1]

$$R_{\mathrm{std}}(\hat{y}) \coloneqq \Pr_{x,y}\left\{\hat{y}(x) \neq y\right\} \tag{1}$$
$$= \sum_{c\in\mathcal{C}} \Pr(y=c) \Pr_{x|y=c}\left\{\hat{y}(x) \neq c\right\} = \mathbb{E}_x \Pr_{y|x}\left\{\hat{y}(x) \neq y\right\}.$$

To obtain a classifier robust to ε-perturbations with respect to a given norm $\|\cdot\|$, one can minimize the corresponding robust risk

$$R_{\mathrm{rob}}(\hat{y}, \varepsilon, \|\cdot\|) \coloneqq \Pr_{x,y}\left\{\exists_{\delta:\|\delta\|\leq\varepsilon}\ \ \hat{y}(x+\delta) \neq y\right\} \tag{2}$$
$$= \sum_{c\in\mathcal{C}} \Pr(y=c) \Pr_{x|y=c}\left\{\exists_{\delta:\|\delta\|\leq\varepsilon}\ \ \hat{y}(x+\delta) \neq c\right\}.$$

The robust risk penalizes errors on (x, y) pairs from the data distribution, as well as on data *after ε-sized perturbations* $\delta \in \mathbb{R}^p$. Furthermore, the robust risk defined in (2) generalizes the standard risk (1) since $R_{\mathrm{std}}(\hat{y}) = R_{\mathrm{rob}}(\hat{y}, 0, \|\cdot\|)$.

While minimizing the robust risk has been shown to indeed improve robustness in practice, this approach is not without its drawbacks. Numerous works have argued that there may be a fundamental tradeoff between robustness and standard test risk (e.g., [9], [10]) and that generalization after adversarial training requires significantly more data (e.g., [11]). Moreover, whereas the problem of training a deep neural network typically is overparameterized, finding worst-case perturbations of data as in (2) is severely underparameterized and therefore this problem does not benefit from the benign optimization landscape of standard training [12]–[14]. To this end, a growing body of work has sought to analyze the theoretical properties of these tradeoffs to gain a deeper understanding of the fundamental limits of adversarial robustness (e.g., [8], [9], [15]–[18]).

Despite the progress made toward uncovering the tradeoffs inherent to adversarial training, many fundamental questions remain unresolved. What do adversarially robust classifiers that minimize the robust risk (2) look like in simple settings? How do they depend on properties of the data distribution such

E. Dobriban and D. Hong are with the Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA, 19104 USA (e-mail: dobriban@wharton.upenn.edu; dahong67@wharton.upenn.edu).

H. Hassani and A. Robey are with the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, 19104 USA (e-mail: hassani@seas.upenn.edu; arobey1@seas.upenn.edu).

Equal contribution from all authors. E. Dobriban was supported in part by NSF BIGDATA grant IIS 1837992, and the NSF-Simons Award on the Mathematical and Scientific Foundations of Deep Learning (2031895). D. Hong was supported in part by NSF BIGDATA grant IIS 1837992, the Dean's Fund for Postdoctoral Research of the Wharton School, and NSF Mathematical Sciences Postdoctoral Research Fellowship DMS 2103353. A. Robey and H. Hassani were supported in part by NSF HDR TRIPODS award 1934876, NSF award CPS-1837253, NSF award CIF-1910056, NSF CAREER award CIF-1943064, and the Air Force Office of Scientific Research Young Investigator Program (AFOSR-YIP) under award #FA9550-20-1-0111.

[1]To simplify the discussion, we focus here on the 0-1 loss for which the risk corresponds to misclassification. We consider surrogate losses in Sections VII and VIII.

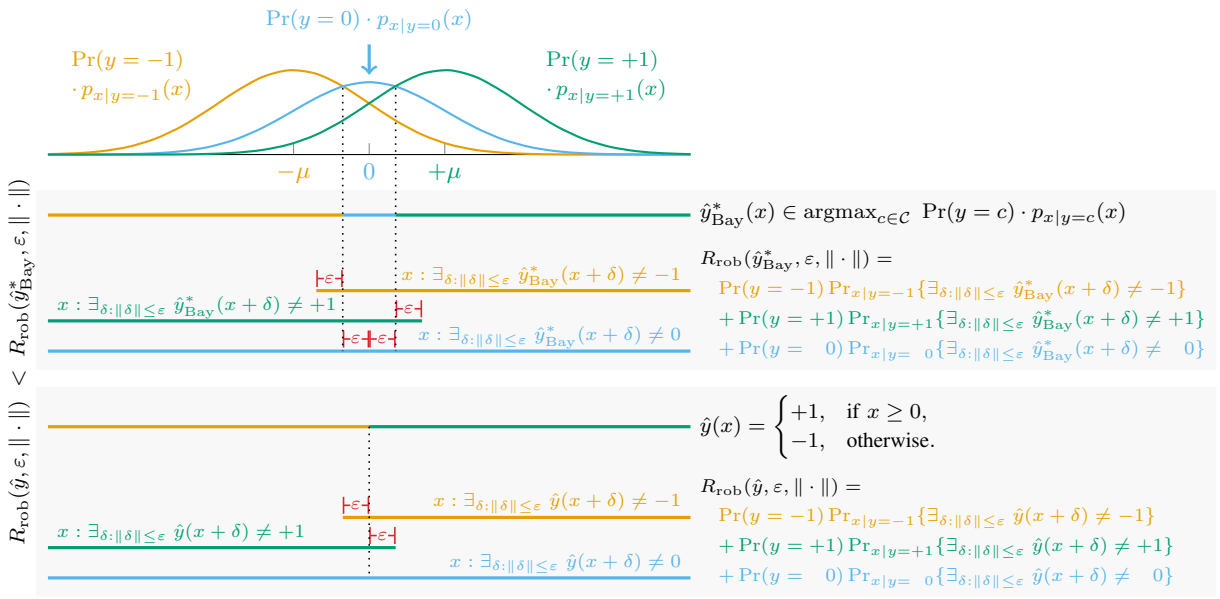


Fig. 1. Illustration of differences between the standard and robust risk. The Bayes classifier $\hat{y}^*_{\text{Bay}}$ minimizes the standard risk by maximizing $\Pr(y = c) \cdot p_{x|y=c}(x)$ for each x pointwise, so it assigns a nontrivial interval around $x = 0$ to the zero class. However, it has worse robust risk than an alternative $\hat{y}$ that drops the zero class. Minimizing the robust risk does not reduce to making optimal pointwise decisions.

as class separation and class imbalance, as well as the choice of perturbation radius ε and norm $\|\cdot\|$? How are they affected when surrogate losses are used or when the classifier is trained from small numbers of samples?

Resolving these issues is complicated by the fact that the robust risk (2) is significantly more challenging to minimize than the standard risk (1). Indeed, in the standard, non-robust setting, much of our understanding stems from knowing the optimal classifier, which minimizes the standard risk. As is well known, e.g., [19, pg. 216], minimizing the standard risk reduces to making an optimal *pointwise* choice for each $x \in \mathbb{R}^p$. In general the minimizer is given by the Bayes optimal classifier

$$\begin{aligned} \hat{y}^*_{\text{Bay}}(x) &\in \operatorname*{argmax}_{c\in\mathcal{C}} \Pr_{y|x}(y = c) \\ &= \operatorname*{argmax}_{c\in\mathcal{C}} \Pr(y = c) \cdot p_{x|y=c}(x). \end{aligned} \tag{3}$$

Unfortunately, an analogous technique has not yet been found for minimizing the robust risk and for deriving expressions for optimal robust classifiers. This is largely because—unlike minimizing the standard risk—minimizing the robust risk does not reduce to making pointwise decisions depending on the data distribution at each given point individually.

To elucidate the differences between minimizing the standard and robust risks, consider the simple yet fundamental setting of the classification of points drawn from Gaussian distributions. In particular, suppose that each of three classes is distributed according to a Gaussian distribution, $\mathcal{N}(-\mu, 1)$, $\mathcal{N}(0, 1)$ and $\mathcal{N}(\mu, 1)$ respectively, with respective proportions $\Pr(c = -1) = \Pr(c = +1) = 0.35$ and $\Pr(c = 0) = 0.30$, as shown in Figure 1. Deciding how to optimally classify a point like $x = 0$ is trivial when minimizing the standard risk. One simply compares $\Pr(y = c) \cdot p_{x|y=c}(0)$ across $c \in \{-1, 0, 1\}$ and selects the class corresponding to the largest term to obtain the Bayes classification of $\hat{y}^*_{\text{Bay}}(0) = 0$. To minimize the robust risk, however, one must also consider the behavior of the classifier on the entire ε-neighborhood of x. In this case, it turns out that dropping the zero class altogether engenders a more robust classifier, meaning that a classifier that assigns $\hat{y}(0) = +1$ has smaller robust risk than the Bayes optimal classifier defined in (3). In this way, minimizing the robust risk does not reduce to a problem depending on the pointwise densities like for the standard risk.

This fundamental difference between the standard and robust risks means that new techniques are needed for deriving optimal robust classifiers. Even for the two-class Gaussian classification setting, while it is well-known that linear classifiers minimize the standard risk, it is not immediately clear that an analogous result holds when minimizing the robust risk.

To address challenges of this type, in this paper we provide new insights and understanding by deriving optimal robust classifiers in the fundamental setting of imbalanced Gaussian distributions. This precise characterization allows us to rigorously investigate the questions enumerated above, and moreover reveals *fundamental tradeoffs* that arise between standard and robust classification. Namely, we show that in this foundational setting, a tradeoff between standard and robust classification arises not merely because we have not yet managed to find a classifier (from some appropriately chosen family) that minimizes both risks. Rather, no such classifier exists in general. The tradeoff holds no matter what training methods are used, how much computation is available, how many training data points are available, or what hypothesis class is chosen. An additional feature of our analysis is that the simplicity of a Gaussian mixture makes it easier to interpret and reason about the results, helping build intuition about adversarially robust learning.

Contributions. Our contributions are as follows:

1) We find the *optimal robust classifiers* for the foundational setting of two- and three-class imbalanced Gaussian classification with respect to ℓ_2 and ℓ_∞ norm-bounded adversaries in the *imbalanced data setting*. These were previously unknown and involve new fundamental challenges. To tackle this problem, we develop a new proof method, carefully combining the characterization of exact Gaussian isoperimetry—i.e., the equality case of Gaussian concentration of measure—with the Neyman-Pearson lemma and Fisher's linear discriminant.[2] This introduces a *new and nontrivial theoretical approach*.
2) We use these results to identify the role of *class imbalance* for tradeoffs between accuracy and robustness for two and three-class ℓ_2 norm-bounded adversaries. In particular, we uncover *fundamental distinctions between balanced and imbalanced classes*. Balanced classes experience no tradeoff with respect to the Bayes risk: the optimal non-robust classifier also turns out to minimize the robust risk. However, an unavoidable tradeoff appears in the imbalanced case: the boundary measure of the larger class expands and gets larger weight, so the optimal boundary necessarily moves. Thus, no classifier simultaneously minimizes both standard and robust risk; the optimal *non-robust* classifier is not the optimal *robust* classifier.
3) We show how the optimal robust classifiers relate to previously proposed randomized classifiers. Additionally, we characterize optimal robust classifiers for data with more general covariances and with low-dimensional structure.
4) We further show that in certain cases *all robust classifiers are approximately linear*, characterizing all approximately optimal robust classifiers. This requires a novel approach, which leverages recent breakthroughs from *robust isoperimetry* [23], [24] and represents one of our main technical innovations.
5) We uncover *surprising phase-transitions arising in three-class robust classification*. Deriving optimal classifiers for this setting requires delicate analysis, but reveals how the optimal classifier can jump discontinuously for small changes in the problem parameters. This does not occur in the two-class setting or for non-robust classification; it arises when combining robustness with a third class.
6) We provide a more comprehensive understanding by analyzing the *non-convex problem of optimizing the robust risk* for linear classifiers. Specifically, we characterize broad settings where classification calibration holds for convex surrogate losses, so the optimizers of surrogate losses coincide with the optimizers of the original objective.
7) We connect our findings to empirical robust risk minimization by providing a *finite-sample analysis* with respect to 0-1 and surrogate loss functionals, which also highlights the key role of geometry in convergence rates. This analysis does not rely on Gaussianity.

The remainder of this paper is organized as follows. In Section II, we review related work concerning algorithms and analysis techniques for adversarial robustness. Next, after some preliminaries in Section III, Section IV derives ℓ_2 robust optimal classifiers for two-class Gaussian classification. In Section V, we tackle three-class Gaussian classification. Following this, in Section VI, we derive ℓ_∞ robust optimal linear classifiers for two-class Gaussian classification models. In Section VII, we study the optimization landscape of the robust risk. Finally, in Section VIII, we connect the results derived in the previous sections to empirical risk minimization by providing a finite sample analysis under broader distributional assumptions.

II. Related work

Adversarial robustness is a very active and rapidly expanding area of research. For this reason, we can only review a collection of some of the most closely related works.

A. Robustness of linear models

Recently, several papers have studied the robustness of linear models. [25] shows that certain robust support vector machines (SVM) are equivalent to regularized SVM. They also give bounds on the standard generalization error based on the regularized empirical hinge risk. [26] shows equivalences between adversarially robust regression and lasso. [27] studies the adversarial robustness of linear models, arguing that random hyperplanes are very close to any data point and that robustness requires strong regularization. Furthermore, the two related works of [17] and [18] study Gaussian and Bernoulli models for data and analytically establish a variety of phenomena regarding robust accuracy and the generalization gap for linear models. They conclude that more data may actually increase the generalization gap. In this paper, we consider the distinct yet related problem of characterizing optimal classifiers in the population setting rather than determining the finite-sample generalization gap. Finally, the recent results in [28] analyze the sample complexity of training adversarially robust linear classifiers on separated data. We study a similar problem, but in our case the data is not separated.

B. Randomized smoothing

The connection between Gaussianity and robustness is one of the key ideas behind randomized smoothing [29]–[32]. While these works provide an interesting and useful alternative to adversarial training, they have little theoretical overlap with this paper. In this paper, we study the different problem of deriving classifiers that are optimal with respect to the robust risk.

C. Generalized likelihood ratio testing

[33] proposes another interesting and useful alternative to adversarial training. The approach is based on the generalized likelihood ratio test (GLRT) and can be applied to general

[2]While Gaussian distributions are in some sense specific, they are also fundamental and broadly applicable. Indeed, many datasets are well-approximated by Gaussian distributions, see textbooks such as [20], [21], including even data distributions generated by GANs [22].

multi-class Gaussian settings. It remains distinct from this paper since it does not seek to optimize the robust risk, instead utilizing a GLRT.

D. Concentration based analyses

Several works use various forms of concentration of measure to explain the existence of adversarial examples in high dimensions [34], [35]. Relatedly, [36] proposes methods for empirically measuring concentration and establishing fundamental limits on intrinsic robustness. These analyses and empirical results generally rely on concentration on the sphere S^{p-1}, whereas we rely on Gaussian concentration in this paper. More related is the work of [37], which studies the adversarial robustness of Bayes-optimal classifiers in two-class Gaussian classification problems with unequal covariance matrices Σ_1 and Σ_2. For instance, when the covariance matrices are strongly asymmetric, so that the smallest eigenvalue of one class tends to zero, they show that almost all points from that class are close to the optimal decision boundary. In contrast, in the symmetric isotropic case, $\Sigma_1 = \Sigma_2 = \sigma^2 I_p$ and $\sigma \to 0$, they show that with high probability all points in both classes are at distance $p/2$ from the boundary. While this is consistent with a portion of our findings, we focus on different problems, namely finding the optimal robust classifiers.

E. Tradeoffs in adversarial robustness

Many works have argued that there are inherent tradeoffs between standard and robust accuracy [15], [38], [39]. Among these works, several consider Gaussian models of data. Notably, [11] studies the two-class Gaussian classification problem $x \sim \mathcal{N}(y\mu, \sigma^2 I_p)$, focusing on the balanced case $\pi = 1/2$, and on signal vectors μ of norm approximately $\sqrt{p}$. In this setting, unlike in ours, it is possible to construct accurate classifiers even from one training data point (x_1, y_1). They show that such classifiers can have high standard accuracy, but low robust accuracy. In contrast, we attack the more challenging problem of deriving *closed-form* expressions for optimal classifiers without strong assumptions on the signal strength.

In [40], the authors study a two-class Gaussian classification problem with balanced classes. They develop sharp minimax bounds on the classification excess risk with a corresponding estimator. In contrast, we derive optimal classifiers and study tradeoffs for the more general, imbalanced-class setting. Similarly, [41] studies robustness defined as the average of the norm of the smallest perturbation that switches the sign of a classifier f. They consider labels that are a deterministic function of the datapoints, which differs from our setting.

Another related work is that of [9], in which the authors consider two-class Gaussian classification where $x = y \cdot (b, \eta 1_{p-1}) + \mathcal{N}(0_p, \mathrm{diag}(0, 1_{p-1}))$, and b is a random sign variable with $P(b = 1) = q \geq 1/2$, while η is a constant. Thus, the first variable contains the correct class y with probability q, while the remaining "non-robust" features contain a weak correlation with y. Our models are related, but distinct from this model. Their work is closest to our results on the optimal robust classifiers for ℓ_∞ perturbations, which are given by soft-thresholding the mean. This will not use the non-robust features, which is consistent with [9]. However, their results are different, as they emphasize the robustness-accuracy tradeoff [9, Theorem 2.1], while we characterize the optimal robust classifiers.

Aside from the tradeoff between accuracy and adversarial robustness, a growing body of work has focused on other naturally-arising tradeoffs in various problem settings. Among such works, [42] studies adversarial robustness in the presence of label noise and explores its relationship to benign overfitting [43]. [44] studies tradeoffs in the distributionally robust setting, where robustness is defined with respect to a family of related data distributions. Finally, in [45], the authors analyze the tradeoff between invariance and sensitivity to adversarial examples. While each of these studies considers tradeoffs in adversarially robust machine learning, the settings and results are different from our setting.

F. Distribution-agnostic results

One notable recent direction is to study the properties of adversarial learning problems in a distribution-agnostic setting. Among such works, [46] introduces the "adversarial VC-dimension" to study the statistical properties of PAC learning in the presence of an adversary. The authors extend the fundamental theorem of statistical learning theory to this setting, and provide sample complexity bounds for this distribution-agnostic setting. These results gave rise to a line of work focusing specifically on the distribution-agnostic setting (see, e.g., [47]–[49]). Building on this, [50] proposes methods for efficiently PAC learning adversarially robust halfspaces with noise. By and large, due to the distribution-agnostic assumption, the PAC-style results from these papers are more conservative and distinct from the results we obtain for the Gaussian setting. However, in very special cases, such as learning in the presence of *random* (e.g., non-adversarial) classification noise, the representation of the risk in terms of the dual norm in [50] agrees with our characterization.

G. Calibration of the adversarial loss

Also of note is a recent line of work which considers the calibration of the robust 0-1 loss. Concretely, a loss is calibrated with respect to a given function class $\mathcal{F}$ if minimizing the excess risk with respect to a surrogate loss over $\mathcal{F}$ implies minimization of the target risk. Following [51], both [52] and [53] consider the calibration of the robust 0-1 loss, showing both positive and negative calibration results for a variety of surrogate losses and function classes. In contrast to these works, the majority of our main results (see Sections IV to VI) focus directly on minimizing the robust 0-1 loss, rather than minimizing a surrogate loss. Furthermore, we note that the calibration results presented in Section VII of this paper are complementary to the results in [52], [53] which show that convex surrogates are in general not calibrated for the robust 0-1 loss. Specifically, we show that surrogate losses can be calibrated in this setting under stronger conditions than convexity. Also related is the work of [54], in which the authors derive lower bounds on the cross-entropy loss

under adversarially-chosen perturbations. This differs from our setting, as we do not consider the cross-entropy loss in this work.

H. Robustness in non-parametric settings

While we study a parametric setting in this paper, there are several notable works that study similar problems in the non-parametric setting. [55] analyzes the robustness of nearest neighbor methods to adversarial examples. In this work, the authors introduce and study quantities called "astuteness" and "r-optimality", which are intimately related to the robust risk. These ideas led to further related studies of attacks and defenses in the non-parametric setting (see, e.g., [56], [57]). In [57], the robust optimal classifier for a non-parametric setting is determined to be the solution of a particular optimization problem. These results are incomparable with ours, since a mixture of Gaussians is not r-separated, and truncating the Gaussians to satisfy r-separation would lead to a large test error. Furthermore, we *explicitly* derive closed-form expressions for optimal robust classifiers in our setting. More related to our paper is the work of [58], in which the authors study robustness through local-Lipschitzness. However, whereas we seek to find statistically optimal robust classifiers in the classical two- and three-class Gaussian setting, [58] considers a completely different model of data.

I. Optimal transport based analyses

Most related to this paper is the recent work of [59], in which the authors develop a framework connecting adversarial risk to optimal transport. As a special case, for balanced two-class Gaussian classification problems with $\pi = 1/2$ and $x_i|y_i \sim \mathcal{N}(\mu y_i, \Sigma)$, and for general perturbations in a closed, convex, absorbing and origin-symmetric set B, they show that linear classifiers are optimal, and characterize these optimal classifiers. Similarly, [60] also characterizes optimal classifiers in various settings; in particular, they focus on the balanced case $\pi = 1/2$, e.g., for two classes with spherical covariances $\mathcal{N}(\mu_i, \sigma^2 I_p)$ or in 1-D with different means and covariances $\mathcal{N}(\mu_i, \sigma_i^2)$. Complementary to these two works, we focus on identifying tradeoffs for *imbalanced* classes. Furthermore, our analyses rely on entirely different proof techniques, and it is unclear whether our results can be obtained from optimal transport theory, or whether results in the imbalanced setting can be derived using proof techniques which use optimal transport. Relatedly, [61] provide lower bounds on the adversarial risk for certain multi-class classification problems whose data distributions satisfy the W_2 Talagrand transportation-cost inequality. In contrast, we find the optimal classifiers for the two- and three-class Gaussian classification problems.

III. Notations and Preliminaries

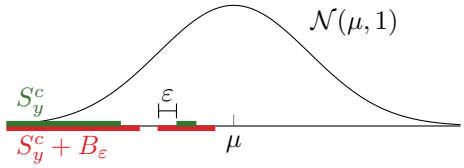
Before we begin, it will be helpful to define some notation. We denote the ball of radius ε (with respect to the norm $\|\cdot\|$) centered at the origin by B_ε, and the indicator function of a set A by $I(A)$. Further, if A and B are sets, then we use the notation $A + B = \{a + b : a \in A, b \in B\}$ for the Minkowski sum; when $A = \{a\}$ contains a single element, we abbreviate it to $a + B$. In these terms, the robust risk with 0-1 loss (2) has another convenient form that we use heavily in the proofs:

$$\begin{aligned} R_{\mathrm{rob}}(\hat{y}, \varepsilon, \|\cdot\|) &= \mathbb{E}_y \Pr_{x|y}\left\{\exists_{\delta:\|\delta\|\le\varepsilon}\ \ \hat{y}(x+\delta) \neq y\right\} \\ &= \mathbb{E}_y \Pr_{x|y}\left\{S_y^c(\hat{y}) + B_\varepsilon\right\}, \end{aligned} \tag{4}$$

where $S_y^c(\hat{y}) := \{x : \hat{y}(x) \neq y\}$ is the *misclassification set* of classifier $\hat{y}$ for class y, and

$$\begin{aligned} S_y^c(\hat{y}) + B_\varepsilon &= \{x + \delta : \hat{y}(x) \neq y \text{ and } \|\delta\| \le \varepsilon\} \\ &= \{x : \exists_{\|\delta\|\le\varepsilon}\ \ \hat{y}(x+\delta) \neq y\}, \end{aligned}$$

is the corresponding *robust misclassification set*, illustrated for a single class by the following diagram.



Note that $S_y(\hat{y}) := \{x : \hat{y}(x) = y\}$ for $y \in \mathcal{C}$ are, correspondingly, the classification sets or decision regions of the classifier $\hat{y}$.

IV. Optimal ℓ_2 robust classifiers for two classes

This section considers the fundamental binary classification setting where data is distributed as a Gaussian for each of the classes $y \in \mathcal{C} = \{+1, -1\}$:

$$\begin{aligned} x|y &\sim \mathcal{N}(y\mu, \sigma^2 I_p), \\ y &= \begin{cases} +1 & \text{with probability } \pi, \\ -1 & \text{with probability } 1-\pi, \end{cases} \end{aligned} \tag{5}$$

where $\mu \in \mathbb{R}^p$ specifies the class means ($+\mu$ and $-\mu$, $\mu \neq 0$), $\sigma^2 \in \mathbb{R}_{>0}$ is the within-class variance, and $\pi \in [0, 1]$ is the proportion of the $y = 1$ class. The means are centered at the origin without loss of generality. By scaling, we will also take $\sigma^2 = 1$ to simplify exposition.

In this setting, the Bayes optimal (non-robust) classifier is linear; in particular, the expression for this classifier is given by the following pointwise calculation:

$$\hat{y}^*_{\mathrm{Bay}}(x) = \mathrm{argmax}_{c\in\mathcal{C}} \Pr_{y|x}(y = c) = \mathrm{sign}(x^\top \mu - q/2),$$

where $q := \ln\{(1-\pi)/\pi\}$ is the log-odds-ratio and we define $\ln(0) := -\infty$ and $\mathrm{sign}(0) = 1$. The classifier is unaffected by any positive rescaling of the argument of sign. Denoting the normal cumulative distribution function $\Phi(x) := (2\pi)^{-1/2} \int_{-\infty}^{x} \exp(-t^2/2)dt$ and $\bar{\Phi} := 1 - \Phi$, the corresponding Bayes risk

$$\begin{aligned} &R_{\mathrm{Bay}}(\mu, \pi) := R_{\mathrm{std}}(\hat{y}^*_{\mathrm{Bay}}) \\ &= \pi \cdot \Phi\left(\frac{q}{2\|\mu\|_2} - \|\mu\|_2\right) + (1-\pi) \cdot \bar{\Phi}\left(\frac{q}{2\|\mu\|_2} + \|\mu\|_2\right), \end{aligned} \tag{6}$$

is the smallest attainable *standard* risk and characterizes the problem difficulty.

A. Optimal classifiers with respect to the robust risk

With this background in mind, we now turn to our problem of finding Bayes-optimal *robust* classifiers. Unlike the non-robust setting, one can no longer simply make optimal pointwise decisions for each $x \in \mathbb{R}^p$ depending only on the data distribution at x, because the robust risk is also affected by neighboring datapoints and decisions. Thus, it is not initially obvious how to find provably optimal robust classifiers. Moreover, it is not initially clear how such classifiers might differ from the Bayes-optimal (non-robust) classifier $\hat{y}^*_{\text{Bay}}$, especially in the presence of class imbalance.

The following theorem provides a precise characterization of optimal robust classifiers. Its proof involves a novel approach that combines the result that halfspaces are extremal sets with respect to Gaussian isoperimetry [62]–[64], the Neyman-Pearson lemma, and Fisher's linear discriminant.

Theorem IV.1 (Optimal ℓ_2-robust two-class classifiers)**.** *Suppose the data (x, y) follow the two-class Gaussian model* (5) *and $\varepsilon < \|\mu\|_2$. An optimal ℓ_2 robust classifier is*

$$\hat{y}^*(x) \coloneqq \operatorname{sign}\{x^\top \mu(1 - \varepsilon/\|\mu\|_2)_+ - q/2\}, \tag{7}$$

where $q = \ln\{(1-\pi)/\pi\}$ and $(x)_+ = \max(x, 0)$. Moreover, the corresponding optimal robust risk is

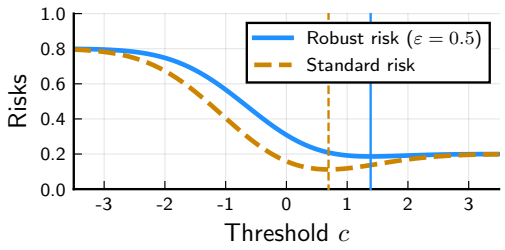
$$R^*_{\text{rob}}(\mu, \pi; \varepsilon) \coloneqq R_{\text{Bay}}\{\mu(1 - \varepsilon/\|\mu\|_2)_+, \pi\}, \tag{8}$$

where R_{Bay} is the Bayes risk defined in (6)*.*

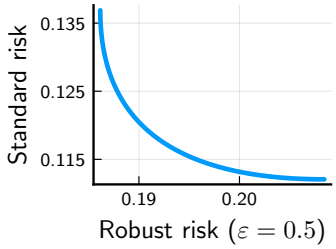
Theorem IV.1 reveals several important insights into the properties of optimal robust classification.

1) Optimality of linear classifiers: Theorem IV.1 shows the nontrivial result that the Bayes optimal robust classifier *is also linear*. Indeed, it shows that the optimal robust classifier corresponds to the classical (non-robust) Bayes optimal classifier with a *reduced effective mean* $\mu \mapsto \mu\,(1 - \varepsilon/\|\mu\|_2)_+$ or, equivalently, an *amplified effective class imbalance*: $q \mapsto q/(1 - \varepsilon/\|\mu\|_2)_+$. Note that if $\varepsilon \geq \|\mu\|_2$, then nontrivial classification is impossible. The effective signal strength reduction is consistent with prior arguments that "adversarially robust generalization requires more data" [11]. However, to the best of our knowledge, such an exact characterization for imbalanced data was previously unknown (see the related work section).

2) Optimal Tradeoffs and Pareto-Frontiers: When the two classes are balanced, i.e., $\pi = 1/2$ (and thus $q = 0$), Theorem IV.1 shows that the Bayes optimal classifier $\hat{y}^*_{\text{Bay}}$ and the optimal robust classifier $\hat{y}^*$ *coincide*. In general, however, there is a *tradeoff*: neither classifier optimizes both standard and robust risks. These insights are important since real datasets are often imbalanced. Indeed, our analysis (see Lemma IV.2) implies that given any classifier, there exists a linear classifier of the form $\operatorname{sign}(x^\top \mu - c)$ with no worse standard risk and no worse robust risk. Using this, we can precisely characterize the *Pareto-frontier (optimal tradeoff)* between the standard risk and the robust risk. Consider a two-dimensional plane in which the x-axis represents the robust risk and the y-axis represents the standard risk (see Figure 2b). Any classifier can be represented as a pair in this plane with its robust risk as the first entry, and its standard risk as the second entry. We further consider the region of all the possible achievable pairs over all classifiers. The *Pareto-frontier (optimal boundary)* of this region shows the fundamental tradeoff between standard and robust risks. For the setting of Theorem IV.1 we can precisely characterize the fundamental tradeoff (Pareto-frontier) between robust and standard risks. Importantly, these tradeoffs hold for any predictor (be it a deep network or a simple linear classifier) including those learned in any way from any size of training data, see Figure 2b.



(a) Risks as functions of threshold c; vertical lines at optimal thresholds.



(b) Pareto-frontier: Standard and robust risk plotted against each other as a function of the threshold c.

Fig. 2. Tradeoffs between optimal classification with respect to standard and robust risks.

This provides new insights. To the best of our knowledge, this is the first work to illustrate tradeoffs due to *class imbalance*, which is prevalent in practice. Moreover, this result proves that a *linear* classifier is *optimal* with respect to robust risk; that is, even if one had access to rich model classes such as deep neural networks, massive computational power, or arbitrarily large datasets, this fundamental tradeoff would remain in place. Such a strong tradeoff (for the Bayes risk) has only been observed for balanced-class settings in a completely separate line of work [59], as discussed in Section II. Relative to this work, our proof techniques are completely different, and we are also able to map the entire Pareto-frontier for imbalanced classes.

Figure 2 illustrates this tradeoff for an example with a mean having norm $\|\mu\|_2 = 1$, a positive class proportion $\pi = 0.2$, and a perturbation radius $\varepsilon = 0.5$. The first figure plots the two risks for the linear classifier $\hat{y}(x) = \operatorname{sign}(x^\top \mu - c)$ as a function of the threshold c. This highlights the difference between the two risks and their corresponding optimal thresholds. The next figure plots the two risks against each other for a sweep of the threshold c.

B. Proof of Theorem IV.1

We now prove Theorem IV.1. Since we cannot simply optimally classify data points based on the data distributions at

each individual point, new techniques are needed for deriving optimal classifiers with respect to the robust risk. Here we introduce a novel approach. First, we prove that there exist optimal linear classifiers by combining the fact that halfspaces are extremal with respect to Gaussian isoperimetry, i.e., the equality case in the Gaussian concentration of measure [62]–[64], with the Neyman-Pearson lemma. Then, we derive the optimal linear classifiers via Fisher's linear discriminant.

Before we begin, note that the robust risk for the two-class setting can be written as

$$\begin{aligned} R(\hat{y},\varepsilon) &\coloneqq R_{\mathrm{rob}}(\hat{y},\varepsilon,\|\cdot\|) \\ &= \pi \cdot \Pr_{x|y=+1}(S_{-1}+B_\varepsilon) + (1-\pi)\cdot \Pr_{x|y=-1}(S_{+1}+B_\varepsilon), \end{aligned}$$

where we drop $\hat{y}$ from S_{-1} and S_{+1} for convenience. This holds for any binary classification problem, and in particular for the two-class Gaussian problem that we consider in this section.

1) Optimality of linear classifiers: The first step is to prove the following claim: for any classifier $\hat{y}$, there exists a linear classifier with robust risk no worse than that of $\hat{y}$. Precisely put, we prove the following lemma.

Lemma IV.2 (Existence of optimal linear classifiers)**.** *Under the assumptions of Theorem IV.1, for any classifier $\hat{y}$, there exists a linear classifier $\hat{y}^*(x) = \mathrm{sign}(x^\top w - c)$, for some w and c, whose standard risk and robust risk with respect to ε-bounded ℓ_2 adversaries is less than or equal to that of the original classifier:*

$$R(\hat{y}^*,\varepsilon) \le R(\hat{y},\varepsilon), \qquad R(\hat{y}^*,0) \le R(\hat{y},0).$$

Moreover, we can take $w = \mu$.

To prove Lemma IV.2, we carefully combine the fact that halfspaces are extremal with respect to Gaussian isoperimetry with the Neyman-Pearson lemma to construct, given any classifier $\hat{y}$, a linear classifier that has robust risk no worse than $\hat{y}$.

Proof of Lemma IV.2: Recall that $x|y \sim \mathcal{N}(y\mu, 1)$, and abbreviate $\Pr_{x|y=\pm1} = \Pr_\pm$. Let $\hat{y}$ be an arbitrary classifier and denote its decision regions as $S_{\pm1}$. We will show there exists a linear classifier with decision regions $\tilde{S}_{\pm1}$, such that:

$$\begin{aligned} \Pr_+(\tilde{S}_{-1}+B_\varepsilon) &\le \Pr_+(S_{-1}+B_\varepsilon), \\ \Pr_-(\tilde{S}_1+B_\varepsilon) &\le \Pr_-(S_1+B_\varepsilon), \\ \Pr_+(\tilde{S}_{-1}) &= \Pr_+(S_{-1}), \\ \Pr_-(\tilde{S}_1) &= \Pr_-(S_1). \end{aligned} \tag{9}$$

Now, the well-known Gaussian concentration of measure (GCM) [62]–[64] states that the sets with minimal "concentration function" are the half-spaces. Specifically, the measurable sets solving the problem

$$\min_S \Pr_\mu(S+B_\varepsilon) \quad \text{s.t.} \quad \Pr_\mu(S) = \alpha, \tag{10}$$

are half-spaces (up to measure zero sets).

For simplicity, let us discuss the one-dimensional problem $p = 1$. Note that the general p case can be reduced to the one-dimensional case. We can take $w = \mu$ and solve the optimal robust classification problem projected into the 1-dimensional line $\{c\mu, c \in \mathbb{R}\}$. Then, projecting back, we can compute probabilities and distances for the multi-dimensional problem. It is not hard to see that the back-projection of the one-dimensional problem also solves the multi-dimensional problem.

In the one-dimensional case, suppose without loss of generality that $\mu > 0$. The GCM states that there is a half-line $\tilde{S}_{-1} = (-\infty, c]$—which will serve as a new set $\{x : \tilde{y}^*(x) = -1\}$—such that

$$\Pr_+(\tilde{S}_{-1}+B_\varepsilon) \le \Pr_+(S_{-1}+B_\varepsilon), \tag{11}$$

$$\Pr_+(\tilde{S}_{-1}) = \Pr_+(S_{-1}). \tag{12}$$

Moreover, we have $c = \mu + \Phi^{-1}(\Pr_+(S_{-1}))$.

Similarly, using the GCM symmetrically, we find that there is a half-line $\tilde{S}_1 = (d, -\infty)$—which will serve as a new set $\{x : \tilde{y}^*(x) = 1\}$—such that

$$\begin{aligned} \Pr_-(\tilde{S}_1+B_\varepsilon) &\le \Pr_-(S_1+B_\varepsilon), \\ \Pr_-(\tilde{S}_1) &= \Pr_-(S_1). \end{aligned}$$

Now, the question is if the two sets $\tilde{S}$ can form the classification regions of a classifier. This would be true if they cover the real line. Here, we claim that they overlap. Thus, they can be shrunken to partition the real line, and the values of the objective in (9) decrease.

To show that they overlap, it is enough to prove that their total probability under one of the two measures, say $\Pr_-$, is at least unity. Thus we need:

$$\begin{aligned} &\Pr_-(\tilde{S}_1) + \Pr_-(\tilde{S}_{-1}) \ge 1 \\ \iff &\Pr_-(S_1) + \Pr_-(\tilde{S}_{-1}) \ge 1 \\ \iff &\Pr_-(\tilde{S}_{-1}) \ge 1 - \Pr_-(S_1) \\ \iff &\Pr_-(\tilde{S}_{-1}) \ge \Pr_-(S_{-1}). \end{aligned}$$

Now note that $\Pr_+(\tilde{S}_{-1}) = \Pr_+(S_{-1})$, so the probability of the two sets coincides under $\Pr_+$. Moreover, $\Pr_+ = \mathcal{N}(\mu, 1)$, $\Pr_- = \mathcal{N}(-\mu, 1)$, $\mu > 0$, and $\tilde{S}_{-1} = (-\infty, c]$. Then, the Neyman-Pearson lemma states that $\tilde{S}_{-1}$ maximizes the function $S \mapsto \Pr_-(S)$ over measurable sets S (i.e., the power of a hypothesis test of $\Pr_+$ against $\Pr_-$), subject to a fixed value of $\Pr_+(S)$. Therefore, the inequality above is true. This shows that the two sets overlap, and thus finishes the proof. ∎

2) Optimal linear classifiers: The proof of Theorem IV.1 now concludes by optimizing among linear classifiers. As in the proof of Lemma IV.2, it is enough to solve the one-dimensional problem. Thus, we want to find the value of the threshold c that minimizes

$$\begin{aligned} &R(\hat{y}_c,\varepsilon) \\ &\quad = \Pr(y=1) \Pr_{x|y=1}(x \le c+\varepsilon) \\ &\qquad + \Pr(y=-1) \Pr_{x|y=-1}(x \ge c-\varepsilon) \end{aligned}$$

$$= \Pr(y=1) \Pr_{\mu}(x \le c+\varepsilon) + \Pr(y=-1) \Pr_{-\mu}(x \ge c-\varepsilon)$$
$$= \Pr(y=1) \Pr_{\mu-\varepsilon}(x \le c) + \Pr(y=-1) \Pr_{-\mu+\varepsilon}(x \ge c).$$

This is exactly the problem of non-robust classification between two Gaussians with means $\mu' = \mu - \varepsilon$ and $-\mu'$. By assumption, $\mu' \ge 0$.

As is well known, the optimal classifier is Fisher's linear discriminant (see, e.g., [19, pg. 216]):

$$\hat{y}^*_\varepsilon(x) = \operatorname{sign}\left[x \cdot (\mu - \varepsilon) - q/2\right],$$

where $q = \ln[(1-\pi)/\pi]$. This concludes the proof of Theorem IV.1. ■

C. Extensions of Theorem IV.1

This section briefly describes a few extensions of Theorem IV.1, with details given in Sections A-A to A-D.

1) Connections to randomized classifiers: The reduced effect size can also be interpreted as adding noise to the data, which relates to previously proposed algorithms [65], [66].

2) Extension to weighted combinations: Given the tradeoff between standard risk and robust risk, one might naturally consider minimizing a weighted combination of the two instead. The techniques used to prove Theorem IV.1 can be extended to this setting, leading to new optimally robust classifiers.

3) Data with a general covariance: We can extend Theorem IV.1 to some settings where the within-class data covariance I_p is replaced with a more general covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$.

4) Data on a low-dimensional subspace: For data that lie in a low-dimensional subspace given by some coordinates being equal to zero, we show the nontrivial result that low-dimensional classifiers are optimal. This is a geometric fact that holds for any norm and does not require Gaussianity. In the Gaussian ℓ_2-robust case, it implies that low-dimensional linear classifiers are optimal.

D. Characterizing approximately optimal robust classifiers via robust isoperimetry

So far, we have explored the problem of characterizing optimal robust classifiers. We conclude this section by briefly characterizing all *approximately* optimal robust classifiers. In particular, we consider robust classifiers that are approximately optimal, i.e., their robust risk is small but potentially suboptimal, and show that they are necessarily close to half-spaces. This is a highly nontrivial question. However, it turns out that we can make some progress by leveraging recent breakthroughs from *robust isoperimetry* (e.g., [23], [24]). In general, these results show that if a set is approximately isoperimetric (in the sense that its boundary measure is close to minimal for its volume), then it has to be close to a half-space. In what follows, we leverage these powerful tools, which, to the best of our knowledge, have not yet been used in machine learning.

Given the difficulty of the problem, we restrict the setting to one-dimensional data. Let $\gamma(S)$ be the Gaussian measure of a measurable set $S \subset \mathbb{R}$. Let also $\gamma^*(S)$ be the *Gaussian deficit* of S, the measure of the error of approximation with a half-line: $\gamma^*(S) \coloneqq \inf_H \gamma(S \Delta H)$, where $S \Delta H = (S \setminus H) \cup (H \setminus S)$ is the symmetric set difference and the infimum is taken over half-lines. Clearly $\gamma^*(S) \ge 0$, with equality when S is a half-line almost surely. The following result concerns a broad family of classifiers whose decision regions are unions of intervals. We show that if $\hat{y}$ has small robust risk, then its decision regions must be close to a half-line; that is, *all robust classifiers are close to linear*.

Theorem IV.3 (Approximately optimal robust classifiers)**.** *For the two-class Gaussian ℓ_2-robust problem, consider classifiers whose classification regions S_+ and S_- are unions of intervals with endpoints in $[-M, M]$ where ε is less than the half-width of all intervals.*

Define $\tau = \tau(\varepsilon, M, \mu) = \varepsilon \exp\left[-((M+\mu)\varepsilon + \varepsilon^2/2)\right]$. Then, for some universal constant $c > 0$,

$$R_{\mathrm{rob}}(\hat{y}, \varepsilon) \ge R_{\mathrm{Bay}} + \tau c \cdot \left[\pi \cdot \gamma^*(S_- - \mu)^2 + (1-\pi) \cdot \gamma^*(S_+ + \mu)^2\right].$$

If the robust risk $R_{\mathrm{rob}}(\hat{y}, \varepsilon)$ is close to the Bayes risk R_{Bay}, then the Gaussian deficits $\gamma^*(S_\pm \pm \mu)$ are small, so the decision regions $S_\pm$ are near half-lines.

A priori, one might suppose that sophisticated classifiers, e.g., deep neural nets, with complicated classification regions, can benefit robustness. Theorem IV.3 shows that these classifiers must also essentially be linear to be even approximately optimally robust. Thus, in this particular case, the complex expressivity of deep neural nets does not bring any clear benefits.

Proof of Theorem IV.3: Denote by ϕ the standard normal density in one dimension, and recall that Φ is the standard normal cumulative distribution function. Let γ^+ be the boundary measure of measurable sets, defined precisely in [23], [24]. While this definition in general poses some technical challenges, we will only use it for unions of intervals $J = \cup_{k \in K}[a_k, b_k]$, where K is a countable index set and $a_k \le b_k < a_{k+1}$ are the endpoints sorted in increasing order. For such sets $\gamma^+(J) = \sum_k [\phi(a_k) + \phi(b_k)]$ is simply the sum of the values of the Gaussian density at the endpoints, which can be finite or infinite.

The *Gaussian isoperimetric profile* is commonly defined as $I = \phi \circ \Phi^{-1}$, and in this language the Gaussian isoperimetric inequality states that $I(\gamma(A)) \le \gamma^+(A)$, with equality if A is a half-line.

Suppose J is a union of intervals in $\mathbb{R}$ with all interval endpoints contained in $[-M, M]$. Then for ε small enough that the ε-expansion of J does not merge any intervals, i.e., $2\varepsilon < (a_{k+1} - b_k)$ for all k, we have

$$\gamma(J + B_\varepsilon) \ge \gamma(J) + \varepsilon \exp\left[-(M\varepsilon + \varepsilon^2/2)\right] \cdot \gamma^+(J). \quad (13)$$

This follows by first considering one interval $J = [a, b]$, and then summing over all intervals, noting that the non-intersection condition on ε guarantees that all terms are

additive. To check the condition for an interval $J = [a, b]$, we write:

$$\begin{aligned}
&\gamma([a,b] + B_\varepsilon) \\
&\quad \geq \gamma([a,b]) + \varepsilon \exp\left[-(M\varepsilon + \varepsilon^2/2)\right] \cdot \gamma^+([a,b]) \\
\iff \\
&\gamma([a-\varepsilon, b+\varepsilon]) \\
&\quad \geq \gamma([a,b]) + \varepsilon \exp\left[-(M\varepsilon + \varepsilon^2/2)\right] \cdot \gamma^+([a,b]) \\
\iff \\
&\gamma([a-\varepsilon, a]) + \gamma([b, b+\varepsilon]) \\
&\quad \geq \varepsilon \exp\left[-(M\varepsilon + \varepsilon^2/2)\right][\phi(a) + \phi(b)].
\end{aligned}$$

Thus, it is enough to verify that

$$\gamma([a-\varepsilon, a]) \geq \varepsilon \exp\left[-(M\varepsilon + \varepsilon^2/2)\right]\phi(a).$$

This follows from

$$\begin{aligned}
\gamma([a-\varepsilon, a]) &= \int_{a-\varepsilon}^{a} \phi(x)dx \\
&= \phi(a) \int_{a-\varepsilon}^{a} \phi(x)/\phi(a)dx \\
&= \phi(a) \int_{a-\varepsilon}^{a} \exp\left[(a^2 - x^2)/2\right]dx \\
&\geq \phi(a) \cdot \varepsilon \cdot \min_{x \in [a-\varepsilon, a]} \exp\left[(a^2 - x^2)/2\right] \\
&= \phi(a) \cdot \varepsilon \cdot \min_{u \in [-\varepsilon, 0]} \exp\left[(a^2 - (a-u)^2)/2\right].
\end{aligned}$$

Now $a^2 - (a-u)^2 = 2au - u^2$. Given that this is a concave function of u, the minimum occurs at one of the two endpoints of the interval $[-\varepsilon, 0]$. Hence, we have

$$\begin{aligned}
\gamma([a-\varepsilon, a]) &\geq \phi(a) \cdot \varepsilon \cdot \min\{\exp\left(a\varepsilon - \varepsilon^2/2\right), 1\} \\
&\geq \phi(a) \cdot \varepsilon \cdot \exp\left[-(M\varepsilon + \varepsilon^2/2)\right].
\end{aligned}$$

This proves the required bound (13) when J is an interval. Thus, by additivity, it also holds for unions of intervals when ε is small enough that the ε-expansion of any two intervals does not merge them. This proves (13) for general sets J.

Suppose now that we have a classifier whose classification regions are unions of intervals. Suppose that the conditions for (13) hold for both S_1 and S_{-1}. Specifically, suppose that all interval endpoints are contained in $[-M, M]$ and $\varepsilon < (a_{k+1} - b_k)$ for all k. Recall that the robust risk can be written as

$$R(\hat{y}, \varepsilon) = \pi \cdot \gamma(S_{-1} - \mu) + (1-\pi) \cdot \gamma(S_1 + \mu).$$

Applying (13) to both classes, and denoting $M' = M + |\mu|$, we find

$$\begin{aligned}
&\gamma(S_1 + \mu + B_\varepsilon) \\
&\quad \geq \gamma(S_1 + \mu) + \varepsilon \exp\left[-(M'\varepsilon + \varepsilon^2/2)\right] \cdot \gamma^+(S_1 + \mu) \\
&\gamma(S_{-1} - \mu + B_\varepsilon) \\
&\quad \geq \gamma(S_{-1} - \mu) + \varepsilon \exp\left[-(M'\varepsilon + \varepsilon^2/2)\right] \cdot \gamma^+(S_{-1} - \mu).
\end{aligned}$$

Let $\tau = \tau(\varepsilon, M, \mu) = \varepsilon \exp\left[-(M'\varepsilon + \varepsilon^2/2)\right]$. Then, by taking a weighted average of the previous two inequalities, we find the bound on the robust risk

$$\begin{aligned}
R(\hat{y}, \varepsilon) \geq\ & R(\hat{y}, 0) \\
&+ \tau \cdot [\pi \cdot \gamma^+(S_{-1} - \mu) + (1-\pi) \cdot \gamma^+(S_1 + \mu)].
\end{aligned}$$

We denote the difference between the boundary measure and isoperimetric profile of a set S as $\delta(S) = \gamma^+(S) - I(\gamma(S))$. The Gaussian isoperimetric inequality states $\delta(S) \geq 0$. Defining $\delta_{\pm 1} = \delta(S_{\pm 1} \pm \mu)$ for the two classification regions, we conclude

$$\begin{aligned}
R(\hat{y}, \varepsilon) \geq\ & R(\hat{y}, 0) \\
&+ \tau[\pi I(\gamma(S_{-1} - \mu)) + (1-\pi) I(\gamma(S_1 + \mu))] \\
&+ \tau \cdot [\pi\delta_{-1} + (1-\pi)\delta_1].
\end{aligned} \tag{14}$$

From [23], it follows for the Gaussian deficit $\gamma^*(S)$ that

$$\gamma^*(S) \leq C\sqrt{\delta(S)},$$

for a universal constant C. Using this inequality for $S_{\pm 1} \pm \mu$, we find that for a universal constant c

$$\begin{aligned}
&\pi\delta_{-1} + (1-\pi)\delta_1 \\
&\quad \geq c \cdot [\pi\gamma^*(S_{-1} - \mu)^2 + (1-\pi)\gamma^*(S_1 + \mu)^2].
\end{aligned}$$

Plugging in to (14), and discarding the second term on the right hand side, we find that

$$\begin{aligned}
R(\hat{y}, \varepsilon) \geq\ & R(\hat{y}, 0) \\
&+ \tau \cdot c \cdot [\pi\gamma^*(S_{-1} - \mu)^2 + (1-\pi)\gamma^*(S_1 + \mu)^2].
\end{aligned}$$

Since $R(\hat{y}, 0) \geq R_{\text{Bay}}$, this gives the desired conclusion. ∎

V. Optimal ℓ_2 robust classifiers for three classes

Having studied two-class robust classification, we now turn to the more general setting of three Gaussian classes $\mathcal{C} = \{-1, 0, 1\}$:

$$\begin{aligned}
x|y &\sim \mathcal{N}(\lambda_y \mu, \sigma^2 I_p), \\
y &= \begin{cases} +1 & \text{with probability } \pi_+, \\ 0 & \text{with probability } \pi_0, \\ -1 & \text{with probability } \pi_-, \end{cases}
\end{aligned} \tag{15}$$

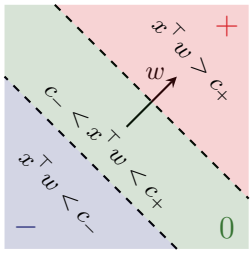
where $\mu \in \mathbb{R}^p \setminus \{0\}$ specifies the line along which the Gaussians lie, λ_y specifies the distance along μ of class y, $\sigma^2 \in \mathbb{R}_{>0}$ is the within-class variance, and $\pi_+, \pi_0, \pi_- \in [0, 1]$ sum to unity and specify the class proportions. Setting $\pi_0 = 0$ produces the two-class model (5). As before, we will take $\sigma^2 = 1$ without loss of generality to simplify the exposition. Further, without loss of generality, we also normalize μ so that $\|\mu\|_2 = 1$, center λ_y so that $\lambda_0 = 0$, and order λ_y so that $\lambda_- := \lambda_{-1} < 0 < \lambda_1 =: \lambda_+$.

The Bayes optimal classifier can again be found via a calculation based on the pointwise densities (see Section B). Here it takes the form of a *linear interval classifier* (illustrated in Figure 3a):

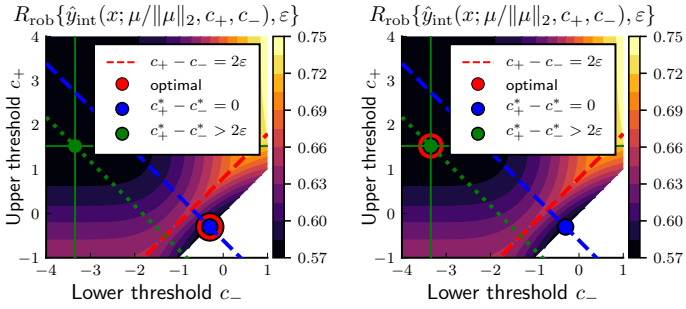
$$\hat{y}_{\text{int}}(x; w, c_+, c_-) := \begin{cases} +1 & \text{if } x^\top w \geq c_+, \\ 0 & \text{if } c_- < x^\top w < c_+, \\ -1 & \text{if } x^\top w \leq c_-, \end{cases} \tag{16}$$

with Bayes optimal thresholds

$$\begin{aligned}
c_+ &:= \max\big\{\lambda_+/2 + \ln(\pi_0/\pi_+)/|\lambda_+|, \\
&\qquad (\lambda_+ + \lambda_-)/2 - \ln(\pi_+/\pi_-)/|\lambda_+ - \lambda_-|\big\}, \\
c_- &:= \min\big\{\lambda_-/2 - \ln(\pi_0/\pi_-)/|\lambda_-|, \\
&\qquad (\lambda_+ + \lambda_-)/2 - \ln(\pi_+/\pi_-)/|\lambda_+ - \lambda_-|\big\},
\end{aligned} \tag{17}$$



(a) Linear interval classifier regions.



(b) $\pi_0 = 42.00\%$ (c) $\pi_0 = 42.01\%$

Fig. 3. Optimal linear interval ℓ_2-robust classifiers for three classes; (b) and (c) show the thresholds (19) and (20), circling the optimal, where $\mu = 1$, $\lambda_\pm = \pm 1$, $\gamma = 1.2$, $\varepsilon = 0.4$ and $\pi_\pm = (1-\pi_0)\{\gamma^{\pm 1}/(\gamma+\gamma^{-1})\}$.

and weights $w = \mu$. As in the two class setting, the positive and negative classes are half-spaces, but now the zero class in between is a slab. As we will see, this feature produces new phenomena and new challenges in robust classification. Unlike the standard risk, for which pointwise calculation of the Bayes classifier trivially generalizes to multi-class settings, going from two-class robust classification to even three classes requires some new methods.

A. Optimal linear interval classifiers with respect to the robust risk

We now present the main result of this section: optimal linear interval classifiers with respect to the robust risk. As in the two-class setting, this minimization does not simply reduce to pointwise comparisons of posterior probabilities. We focus here on the regime $\varepsilon < \min\{|\lambda_+|, |\lambda_-|\}/2$ where nontrivial classification to all three classes occurs; when $\varepsilon \geq \min\{|\lambda_+|, |\lambda_-|\}/2$ the problem essentially reduces to two-class problems (see Section C).

Theorem V.1 (Optimal linear interval ℓ_2-robust three-class classifiers)**.** *Suppose the data (x, y) follow the three-class Gaussian model* (15) *and $\varepsilon < \min\{|\lambda_+|, |\lambda_-|\}/2$. Among linear interval classifiers, an optimal ℓ_2 robust classifier is:*

$$\hat{y}^*_{\rm int}(x) := \hat{y}_{\rm int}(x; \mu, c^*_+, c^*_-), \tag{18}$$

where the thresholds $c^*_+ \geq c^*_-$ are:

Case 1 (rare zero class). If $\pi_0 \leq \alpha^* \sqrt{\pi_-\pi_+}$ for α^* specified below in (21)*, then the optimal classifier ignores the zero class. The thresholds are equal, i.e., $c^*_+ = c^*_-$, with value*

$$c^*_+ = c^*_- = \frac{\lambda_+ + \lambda_-}{2} + \frac{\ln(\pi_-/\pi_+)}{|\lambda_+ - \lambda_-| - 2\varepsilon}, \tag{19}$$

and the corresponding robust risk is

$$R_{\rm rob}(\hat{y}^*_{\rm int}, \varepsilon) = \pi_0 + (\pi_+ + \pi_-) R^*_{\rm rob}\left(\frac{|\lambda_+ - \lambda_-|}{2}, \frac{\pi_+}{\pi_+ + \pi_-}; \varepsilon\right),$$

where $R^*_{\rm rob}$ is the two-class optimal robust risk in (8)*.*

Case 2 (frequent zero class). Otherwise, the thresholds are at least 2ε apart, i.e., $c^*_+ - c^*_- > 2\varepsilon$, with

$$c^*_\pm = \frac{\lambda_\pm}{2} \pm \frac{\ln(\pi_0/\pi_\pm)}{|\lambda_\pm| - 2\varepsilon}, \tag{20}$$

and the corresponding robust risk is

$$R_{\rm rob}(\hat{y}^*_{\rm int}, \varepsilon) = (\pi_+ + \pi_0) R^*_{\rm rob}\left(\frac{|\lambda_+|}{2}, \frac{\pi_+}{\pi_+ + \pi_0}; \varepsilon\right) + (\pi_- + \pi_0) R^*_{\rm rob}\left(\frac{|\lambda_-|}{2}, \frac{\pi_-}{\pi_- + \pi_0}; \varepsilon\right).$$

The cutoff α^* is the unique solution with $\alpha \geq \bar{\alpha}$ to the equation:

$$\begin{aligned}(\gamma + \gamma^{-1}) R^*_{\rm rob}\left(\frac{|\lambda_+ - \lambda_-|}{2}, \frac{\gamma}{\gamma + \gamma^{-1}}; \varepsilon\right) \\ = (\gamma + \alpha) R^*_{\rm rob}\left(\frac{|\lambda_+|}{2}, \frac{\gamma}{\gamma + \alpha}; \varepsilon\right) \\ + (\gamma^{-1} + \alpha) R^*_{\rm rob}\left(\frac{|\lambda_-|}{2}, \frac{\gamma^{-1}}{\gamma^{-1} + \alpha}; \varepsilon\right) - \alpha,\end{aligned} \tag{21}$$

where $\gamma := \sqrt{\pi_+/\pi_-}$, and

$$\bar{\alpha} := \exp\left[-\frac{(|\lambda_+| - 2\varepsilon)(|\lambda_-| - 2\varepsilon)}{2} - \frac{\lambda_+ + \lambda_-}{|\lambda_+ - \lambda_-| - 4\varepsilon} \ln\gamma\right].$$

Theorem V.1 reveals several properties of optimally robust linear interval classification in the three-class setting.

1) New three-class phenomenon (discontinuity of the optimal thresholds): The optimal thresholds in Theorem V.1 fall into two cases. In the first case, the thresholds (19) coincide with the one from two-class robust classification between the positive and negative classes, *ignoring the zero class*. In the second case, the thresholds (20) coincide with two-class robust classification: i) between the zero and positive classes, and ii) between the zero and negative classes.

However, the robust setting with $\varepsilon > 0$ introduces a new and perhaps surprising property: optimal thresholds can now be *discontinuous* in the problem parameters, as they jump between the cases. For example, Figures 3b and 3c show nearly identical setups, with only a *slight* change in class imbalance, but the optimal thresholds jump *discontinuously* from (19) to (20). To understand why this occurs, the optimal thresholds are never in $0 < c_+ - c_- < 2\varepsilon$ since any such choice can be improved by moving c_+ and c_- closer together. This discontinuity does *not* arise for the standard risk ($\varepsilon = 0$), since (19) and (20) coincide at the cutoff in that case. It also does *not* arise in the two-class setup discussed previously.

The three-class Gaussian setting (15) considered here is an exemplar for general multi-class settings, and we expect similar discontinuous transitions to appear more generally. Essentially, they arise from a general property of the ε-robust risk that our analysis draws into focus. Namely, classification

regions must either be: i) empty or ii) have radius at least ε (otherwise it is better to remove them). Jumping between these cases produces the observed discontinuity.

2) Tradeoffs and the impact of class imbalance: As in the two-class setting, we see a tradeoff in general between the standard risk and the robust risk. No classifier optimizes both *even when restricted to linear interval classifiers*, which are optimal for the standard risk. A new three-class manifestation of this phenomenon arises in the jump from (19) to (20), i.e., from ignoring to including the zero class. The class ratios in (19) and (20) are also effectively inflated by $1/(1-2\varepsilon/|\lambda_+ - \lambda_-|)$ and $1/(1-2\varepsilon/|\lambda_\pm|)$, respectively, amplifying the effects of class imbalance.

B. Proof of Theorem V.1

We now prove Theorem V.1. The three-class setting introduces a new challenge in the proof: the optimization landscape will turn out to have two qualitatively different regions. Essentially, the robust risk produces a dichotomy between classifiers where the central zero-class slab in Figure 3a is either "thin" or "thick". Optimizing over each region separately yields two candidate optimizers. We will need to carefully analyze their properties to characterize when each is globally optimal.

Precisely put, our goal here is to find $w \in \mathbb{R}^p \setminus \{0\}$ and $c_+ \geq c_-$ that minimize the robust risk

$$\begin{aligned} R_{\text{rob}}\{\hat{y}_{\text{int}}(\cdot; w, c_+, c_-), \varepsilon, \|\cdot\|_2\} &= \pi_+ \mathcal{M}_+(w, c_+, c_-, \varepsilon) \\ &\quad + \pi_0 \mathcal{M}_0(w, c_+, c_-, \varepsilon) \\ &\quad + \pi_- \mathcal{M}_-(w, c_+, c_-, \varepsilon), \end{aligned}$$

where

$$\begin{aligned} &\mathcal{M}_+(w, c_+, c_-, \varepsilon) \\ &\quad := \Pr_{x|y=+1}\big\{\exists_{\delta:\|\delta\|_2\leq\varepsilon}\ \hat{y}_{\text{int}}(x+\delta; w, c_+, c_-) \neq +1\big\}, \\ &\mathcal{M}_-(w, c_+, c_-, \varepsilon) \\ &\quad := \Pr_{x|y=-1}\big\{\exists_{\delta:\|\delta\|_2\leq\varepsilon}\ \hat{y}_{\text{int}}(x+\delta; w, c_+, c_-) \neq -1\big\}, \\ &\mathcal{M}_0(w, c_+, c_-, \varepsilon) \\ &\quad := \Pr_{x|y=0}\big\{\exists_{\delta:\|\delta\|_2\leq\varepsilon}\ \hat{y}_{\text{int}}(x+\delta; w, c_+, c_-) \neq 0\big\}, \end{aligned}$$

are the associated robust misclassification probabilities as functions of w, c_+, c_- and ε.

1) Optimizing over $\boldsymbol{w}$: First we optimize over $w \in \mathbb{R}^p \setminus \{0\}$. Since any positive scaling of w can be absorbed into c_+ and c_-, it suffices to consider $\|w\|_2 = 1$. Then we have

$$\begin{aligned} \mathcal{M}_0(w, c_+, c_-, \varepsilon) &= \Pr_{x|y=0}(x^\top w \leq c_- + \varepsilon \text{ or } x^\top w \geq c_+ - \varepsilon) \\ &= \Pr_{\tilde{x}\sim\mathcal{N}(0,1)}(\tilde{x} \leq c_- + \varepsilon \text{ or } \tilde{x} \geq c_+ - \varepsilon), \end{aligned}$$

which is a constant with respect to w. Meanwhile,

$$\begin{aligned} &\mathcal{M}_+(w, c_+, c_-, \varepsilon) \\ &\quad = \Pr_{x|y=+1}(x^\top w < c_+ + \varepsilon) = \Pr_{\tilde{x}\sim\mathcal{N}(\lambda_+\mu^\top w, 1)}(\tilde{x} < c_+ + \varepsilon), \\ &\mathcal{M}_-(w, c_+, c_-, \varepsilon) \\ &\quad = \Pr_{x|y=-1}(x^\top w > c_- - \varepsilon) = \Pr_{\tilde{x}\sim\mathcal{N}(\lambda_-\mu^\top w, 1)}(\tilde{x} > c_- - \varepsilon), \end{aligned}$$

which are both minimized by taking $w = \mu$. Hence $w^* := \mu$ optimizes the robust risk.

2) Finding candidate optimizers with respect to c_+ and c_-: We now proceed to derive optimal thresholds $c_+ \geq c_-$ given optimal weights $w^* = \mu$. Substituting w^* into the robust risk yields the following function of c_- and c_+:

$$\begin{aligned} R_{\text{rob}}(c_-, c_+) &:= R_{\text{rob}}\{\hat{y}_{\text{int}}(\cdot; w^*, c_+, c_-), \varepsilon, \|\cdot\|_2\} \\ &= \pi_+ \mathcal{M}_+(w^*, c_+, c_-, \varepsilon) \\ &\quad + \pi_0 \mathcal{M}_0(w^*, c_+, c_-, \varepsilon) \\ &\quad + \pi_- \mathcal{M}_-(w^*, c_+, c_-, \varepsilon) \\ &= \pi_+ \Pr_{\tilde{x}\sim\mathcal{N}(\lambda_+\mu^\top w^*, 1)}(\tilde{x} < c_+ + \varepsilon) \\ &\quad + \pi_0 \Pr_{\tilde{x}\sim\mathcal{N}(0,1)}(\tilde{x} \leq c_- + \varepsilon \text{ or } \tilde{x} \geq c_+ - \varepsilon) \\ &\quad + \pi_- \Pr_{\tilde{x}\sim\mathcal{N}(\lambda_-\mu^\top w^*, 1)}(\tilde{x} > c_- - \varepsilon) \\ &= \pi_- \Pr_{-}(\tilde{x} > c_- - \varepsilon) + \pi_+ \Pr_{+}(\tilde{x} < c_+ + \varepsilon) \\ &\quad + \pi_0 \Pr_{0}(\tilde{x} \leq c_- + \varepsilon \text{ or } \tilde{x} \geq c_+ - \varepsilon), \end{aligned}$$

where we drop the arguments ε and $\|\cdot\|_2$ from R_{rob} for simplicity, and use the following shorthands for the class-conditional probabilities

$$\begin{aligned} \Pr_{-} &:= \Pr_{\tilde{x}\sim\mathcal{N}(\lambda_-\mu^\top w^*, 1)} = \Pr_{\tilde{x}\sim\mathcal{N}(\lambda_-, 1)}, \\ \Pr_{+} &:= \Pr_{\tilde{x}\sim\mathcal{N}(\lambda_+\mu^\top w^*, 1)} = \Pr_{\tilde{x}\sim\mathcal{N}(\lambda_+, 1)}, \\ \Pr_{0} &:= \Pr_{\tilde{x}\sim\mathcal{N}(0,1)}. \end{aligned}$$

Next we will explain the unique feature of the robust risk: the constraint set $\Omega := \{(c_-, c_+ : c_- \leq c_+)\}$ contains two qualitatively different regions. Minimizing over each region results in two candidate optimizers.

First, consider $\Omega_0 := \{(c_-, c_+) : c_- \leq c_+ \leq c_- + 2\varepsilon\}$. In this region, $\Pr_0(\tilde{x} \leq c_- + \varepsilon \text{ or } \tilde{x} \geq c_+ - \varepsilon) = 1$ so the robust risk simplifies to

$$R_{\text{rob}}(c_-, c_+) = \pi_- \Pr_{-}(\tilde{x} > c_- - \varepsilon) + \pi_+ \Pr_{+}(\tilde{x} < c_+ + \varepsilon) + \pi_0.$$

Since this function is decreasing in c_- and increasing in c_+, it is minimized by $c_- = c_+$. This yields a two-class problem between the negative and positive classes with minimizer

$$c_- = c_+ = \tilde{c} := \frac{\lambda_+ + \lambda_-}{2} + \frac{\ln(\pi_-/\pi_+)}{|\lambda_+ - \lambda_-| - 2\varepsilon}. \tag{22}$$

This is the minimizer over Ω_0.

Second, consider $\Omega_1 := \{(c_-, c_+) : c_+ \geq c_- + 2\varepsilon\}$. In this region, $\Pr_0(\tilde{x} \leq c_- + \varepsilon \text{ and } \tilde{x} \geq c_+ - \varepsilon) = 0$ so

$$\begin{aligned} &\Pr_{0}(\tilde{x} \leq c_- + \varepsilon \text{ or } \tilde{x} \geq c_+ - \varepsilon) \\ &\quad = \Pr_{0}(\tilde{x} \leq c_- + \varepsilon) + \Pr_{0}(\tilde{x} \geq c_+ - \varepsilon), \end{aligned}$$

and the robust risk simplifies to $R_{\text{rob}}(c_-, c_+) = R^-_{\text{rob}}(c_-) + R^+_{\text{rob}}(c_+)$, where

$$\begin{aligned} R^-_{\text{rob}}(c_-) &:= \pi_- \Pr_{-}(\tilde{x} > c_- - \varepsilon) + \pi_0 \Pr_{0}(\tilde{x} \leq c_- + \varepsilon), \\ R^+_{\text{rob}}(c_+) &:= \pi_0 \Pr_{0}(\tilde{x} \geq c_+ - \varepsilon) + \pi_+ \Pr_{+}(\tilde{x} < c_+ + \varepsilon). \end{aligned}$$

Now, $R^-_{\text{rob}}(c_-)$ is proportional to the ε-robust risk for a two-class problem between the negative and zero classes. Hence,

it is a decreasing function for $c_- < \tilde{c}_-$ and an increasing function for $c_- > \tilde{c}_-$, where the critical point $\tilde{c}_-$ is

$$\tilde{c}_- \coloneqq \frac{\lambda_-}{2} - \frac{\ln(\pi_0/\pi_-)}{|\lambda_-| - 2\varepsilon}. \tag{23}$$

Likewise $R^+_{\mathrm{rob}}(c_+)$ is a decreasing function for $c_+ < \tilde{c}_+$ and an increasing function for $c_+ > \tilde{c}_+$, where the critical point is

$$\tilde{c}_+ \coloneqq \frac{\lambda_+}{2} + \frac{\ln(\pi_0/\pi_+)}{|\lambda_+| - 2\varepsilon}. \tag{24}$$

If $(\tilde{c}_-, \tilde{c}_+) \in \Omega_1$, then it follows that $(\tilde{c}_-, \tilde{c}_+)$ is globally optimal within Ω_1. On the other hand, if $(\tilde{c}_-, \tilde{c}_+) \notin \Omega_1$ then the optimal value in Ω_1 must occur on the boundary $c_+ = c_- + 2\varepsilon$. By the monotonicity analysis above, any point off that boundary can necessarily be improved either by increasing c_- or by decreasing c_+ since either $c_- \le \tilde{c}_-$ or $c_+ \ge \tilde{c}_+$ for any $(c_-, c_+) \in \Omega_1$ when $(\tilde{c}_-, \tilde{c}_+) \notin \Omega_1$.

3) Characterizing when each candidate optimizer is globally optimal: Now we compare the minimizers from the regions Ω_0 and Ω_1 to find globally optimal thresholds. For this purpose, it turns out that $\alpha = \pi_0/\sqrt{\pi_-\pi_+}$ and $\gamma = \sqrt{\pi_+/\pi_-}$ provide a more convenient parameterization than π_-, π_0 and π_+. Recall that $\pi_0 + \pi_- + \pi_+ = 1$ necessarily constrains those parameters. Rewriting (22) to (24) in terms of α and γ yields

$$\tilde{c} = \frac{\lambda_+ + \lambda_-}{2} - \frac{\ln\gamma}{|\lambda_+ - \lambda_-|/2 - \varepsilon},$$
$$\tilde{c}_\pm = \frac{\lambda_\pm}{2} \pm \frac{\ln\alpha}{|\lambda_\pm| - 2\varepsilon} - \frac{\ln\gamma}{|\lambda_\pm| - 2\varepsilon}.$$

Furthermore,

$$\begin{aligned}
&\tilde{c}_+ > \tilde{c}_- + 2\varepsilon \\
&\iff 0 < \tilde{c}_+ - \tilde{c}_- - 2\varepsilon \\
&\qquad = \frac{|\lambda_+ - \lambda_-| - 4\varepsilon}{2} + \frac{|\lambda_+ - \lambda_-| - 4\varepsilon}{(|\lambda_+| - 2\varepsilon)(|\lambda_-| - 2\varepsilon)} \ln\alpha \\
&\qquad\quad + \frac{\lambda_+ + \lambda_-}{(|\lambda_+| - 2\varepsilon)(|\lambda_-| - 2\varepsilon)} \ln\gamma \\
&\iff \alpha > \bar{\alpha},
\end{aligned}$$

yielding an equivalent condition for $(\tilde{c}_-, \tilde{c}_+) \in \mathbf{int}\,\Omega_1$, where $\mathbf{int}\,A$ denotes the interior of the set A with respect to the standard topology induced by the Euclidean metric. Thus, when $\alpha \le \bar{\alpha}$, the optimal value in Ω_1 occurs on the boundary $c_+ = c_- + 2\varepsilon$, but this boundary is also contained in Ω_0 so it is no worse than (22). Namely, (22) is optimal when $\alpha \le \bar{\alpha}$.

Now suppose $\alpha > \bar{\alpha}$. In this case, $(\tilde{c}_-, \tilde{c}_+) \in \mathbf{int}\,\Omega_1$ is optimal in Ω_1, so we compare $R_{\mathrm{rob}}(\tilde{c}, \tilde{c})$ with $R_{\mathrm{rob}}(\tilde{c}_-, \tilde{c}_+) = R^-_{\mathrm{rob}}(\tilde{c}_-) + R^+_{\mathrm{rob}}(\tilde{c}_+)$. For this comparison, we study the sign of

$$\Delta(\alpha) \coloneqq \widetilde{R}^-_{\mathrm{rob}}(\tilde{c}_-, \alpha) + \widetilde{R}^+_{\mathrm{rob}}(\tilde{c}_+, \alpha) - \widetilde{R}_{\mathrm{rob}}(\tilde{c}, \alpha), \tag{25}$$

where we define

$$\widetilde{R}^-_{\mathrm{rob}}(\tilde{c}_-, \alpha) \coloneqq \gamma^{-1} \Pr_-(\tilde{x} > \tilde{c}_- - \varepsilon) + \alpha \Pr_0(\tilde{x} \le \tilde{c}_- + \varepsilon),$$
$$\widetilde{R}^+_{\mathrm{rob}}(\tilde{c}_+, \alpha) \coloneqq \alpha \Pr_0(\tilde{x} \ge \tilde{c}_+ - \varepsilon) + \gamma \Pr_+(\tilde{x} < \tilde{c}_+ + \varepsilon),$$
$$\widetilde{R}_{\mathrm{rob}}(\tilde{c}, \alpha) \coloneqq \gamma^{-1} \Pr_-(\tilde{x} > \tilde{c} - \varepsilon) + \gamma \Pr_+(\tilde{x} < \tilde{c} + \varepsilon) + \alpha.$$

Note first that $\Delta(\bar{\alpha}) \ge 0$ since, as established above, $R_{\mathrm{rob}}(\tilde{c}, \tilde{c}) \le R^-_{\mathrm{rob}}(\tilde{c}_-) + R^+_{\mathrm{rob}}(\tilde{c}_+)$ in this case.

Next, when $\alpha > \bar{\alpha}$

$$\begin{aligned}
\frac{\partial\Delta(\alpha)}{\partial\alpha} &= \frac{\partial\widetilde{R}^-_{\mathrm{rob}}(\tilde{c}_-, \alpha)}{\partial\tilde{c}_-}\frac{\partial\tilde{c}_-}{\partial\alpha} + \frac{\partial\widetilde{R}^-_{\mathrm{rob}}(\tilde{c}_-, \alpha)}{\partial\alpha} \\
&\quad + \frac{\partial\widetilde{R}^+_{\mathrm{rob}}(\tilde{c}_+, \alpha)}{\partial\tilde{c}_+}\frac{\partial\tilde{c}_+}{\partial\alpha} + \frac{\partial\widetilde{R}^+_{\mathrm{rob}}(\tilde{c}_+, \alpha)}{\partial\alpha} \\
&\quad - \frac{\partial\widetilde{R}_{\mathrm{rob}}(\tilde{c}, \alpha)}{\partial\alpha} \\
&= \Pr_0(\tilde{x} \le \tilde{c}_- + \varepsilon) + \Pr_0(\tilde{x} \ge \tilde{c}_+ - \varepsilon) - 1 < 0.
\end{aligned}$$

Both $\tilde{c}_-$ and $\tilde{c}_+$ are differentiable functions of α for $\alpha > 0$, and $\widetilde{R}^-_{\mathrm{rob}}$, $\widetilde{R}^+_{\mathrm{rob}}$ and $\widetilde{R}_{\mathrm{rob}}$ are continuously differentiable functions of $\tilde{c}_-$, $\tilde{c}_+$ and α. Hence, Δ is also a differentiable function of α. The equality in the second line holds because $\tilde{c}_-$ and $\tilde{c}_+$ are critical points of R^-_{rob} and R^+_{rob}, respectively, and the inequality holds because $\tilde{c}_- + \varepsilon < \tilde{c}_+ - \varepsilon$ for $\alpha > \bar{\alpha}$. Namely, $\Delta(\alpha)$ is a decreasing function for $\alpha > \bar{\alpha}$.

Next, $\tilde{c}_-$ is a decreasing function of α, and $\tilde{c}_+$ is an increasing function of α. Hence, $\Pr_0(\tilde{x} \le \tilde{c}_- + \varepsilon)$ and $\Pr_0(\tilde{x} \ge \tilde{c}_+ - \varepsilon)$ are both decreasing functions of α, and so $\partial\Delta(\alpha)/\partial\alpha$ is also a decreasing function of α. As a result, $\Delta < 0$ eventually and Δ has exactly one root with respect to α in the domain $\alpha \ge \bar{\alpha}$. Specifically, there is a unique $\alpha^* \ge \bar{\alpha}$ for which $\Delta(\alpha^*) = 0$. If $\alpha < \alpha^*$ then $\Delta > 0$, and if $\alpha > \alpha^*$ then $\Delta < 0$.

In particular, we must solve the equation $\Delta(\alpha) = 0$ to obtain α^*. We conclude by deriving the alternative form (21) for this equation. First, recall that $\tilde{c}$ is the optimal two-class threshold between the negative and positive classes so

$$\begin{aligned}
&R_{\mathrm{rob}}(\tilde{c}, \tilde{c}) \\
&\quad = \pi_0 + \pi_- \Pr_-(\tilde{x} > \tilde{c} - \varepsilon) + \pi_+ \Pr_+(\tilde{x} < \tilde{c} + \varepsilon) \\
&\quad = \pi_0 + (\pi_+ + \pi_-)\bigg\{\frac{\pi_-}{\pi_+ + \pi_-}\Pr_-(\tilde{x} > \tilde{c} - \varepsilon) \\
&\qquad\qquad + \frac{\pi_+}{\pi_+ + \pi_-}\Pr_+(\tilde{x} < \tilde{c} + \varepsilon)\bigg\} \\
&\quad = \pi_0 + (\pi_+ + \pi_-)R^*_{\mathrm{rob}}\bigg(\frac{|\lambda_+ - \lambda_-|}{2}, \frac{\pi_+}{\pi_+ + \pi_-}; \varepsilon\bigg).
\end{aligned}$$

Next, when $\alpha \ge \bar{\alpha}$, we have $\tilde{c}_+ \ge \tilde{c}_- + 2\varepsilon$ so

$$\begin{aligned}
&R_{\mathrm{rob}}(\tilde{c}_-, \tilde{c}_+) = R^-_{\mathrm{rob}}(\tilde{c}_-) + R^+_{\mathrm{rob}}(\tilde{c}_+) \\
&\quad = \pi_+ \Pr_+(\tilde{x} < \tilde{c}_+ + \varepsilon) + \pi_0 \Pr_0(\tilde{x} \ge \tilde{c}_+ - \varepsilon) \\
&\qquad + \pi_0 \Pr_0(\tilde{x} \le \tilde{c}_- + \varepsilon) + \pi_- \Pr_-(\tilde{x} > \tilde{c}_- - \varepsilon) \\
&\quad = (\pi_+ + \pi_0)\bigg\{\frac{\pi_0}{\pi_+ + \pi_0}\Pr_0(\tilde{x} \ge \tilde{c}_+ - \varepsilon) \\
&\qquad\qquad + \frac{\pi_+}{\pi_+ + \pi_0}\Pr_+(\tilde{x} < \tilde{c}_+ + \varepsilon)\bigg\} \\
&\qquad + (\pi_- + \pi_0)\bigg\{\frac{\pi_-}{\pi_- + \pi_0}\Pr_-(\tilde{x} > \tilde{c}_- - \varepsilon) \\
&\qquad\qquad + \frac{\pi_0}{\pi_- + \pi_0}\Pr_0(\tilde{x} \le \tilde{c}_- + \varepsilon)\bigg\} \\
&\quad = (\pi_+ + \pi_0)R^*_{\mathrm{rob}}\bigg(\frac{|\lambda_+|}{2}, \frac{\pi_+}{\pi_+ + \pi_0}; \varepsilon\bigg)
\end{aligned}$$

$$+ (\pi_- + \pi_0) R^*_{\text{rob}}\left(\frac{|\lambda_-|}{2}, \frac{\pi_-}{\pi_- + \pi_0}; \varepsilon\right),$$

since $\tilde{c}_+$ and $\tilde{c}_-$ are, respectively, optimal two-class thresholds: i) between the zero and the positive class and ii) between the zero and the negative class. Substituting these into (25) and simplifying yields

$$\begin{aligned}\Delta &= \frac{1}{\sqrt{\pi_- \pi_+}}\big\{R^-_{\text{rob}}(\tilde{c}_-) + R^+_{\text{rob}}(\tilde{c}_+) - R_{\text{rob}}(\tilde{c}, \tilde{c})\big\} \\ &= (\gamma + \alpha) R^*_{\text{rob}}\left(\frac{|\lambda_+|}{2}, \frac{\gamma}{\gamma + \alpha}; \varepsilon\right) \\ &\quad + (\gamma^{-1} + \alpha) R^*_{\text{rob}}\left(\frac{|\lambda_-|}{2}, \frac{\gamma^{-1}}{\gamma^{-1} + \alpha}; \varepsilon\right) - \alpha \\ &\quad - (\gamma + \gamma^{-1}) R^*_{\text{rob}}\left(\frac{|\lambda_+ - \lambda_-|}{2}, \frac{\gamma}{\gamma + \gamma^{-1}}; \varepsilon\right),\end{aligned}$$

and re-arranging gives (21). This concludes the proof of Theorem V.1. ■

C. On the optimality of linear interval classifiers

Theorem V.1 gave an optimal linear interval robust classifier for the three-class setting (15), i.e., we derived a classifier that minimizes the robust risk subject to the constraint that it is a linear interval classifier (16). Naturally, one may wonder if this classifier is only optimal among linear interval classifiers or if it is in fact optimal across all classifiers. In other words, are linear interval classifiers optimal? It is important to note here that answering this question was not needed to demonstrate a tradeoff between the standard risk and the robust risk, as discussed above. Nevertheless, this is an important and interesting question in its own right. Indeed, given the symmetry of the three-class setting we consider, we expect that linear interval classifiers are optimal.[3] Namely, we have the following conjecture:

Conjecture V.2 (Linear interval classifiers are optimal across all classifiers)**.** *Under the assumptions of Theorem V.1, the optimal linear interval classifier from Theorem V.1 is also optimal across all classifiers.*

While this conjecture is very natural, a proof for it is still unknown and appears to be surprisingly nontrivial. The three-class setting introduces new challenges beyond the two-class setting. Indeed we are unaware of any existing optimality results like this for robust classification beyond two-class settings. The following subsections make progress towards rigorously establishing the conjecture, providing a pair of first results on the optimality of linear interval classifiers in multi-class settings. The two results are obtained using two different theoretical approaches and are somewhat complementary as a result: the first applies for any model parameters but does not consider all classifiers, while the second considers all classifiers but only applies for certain model parameters. For each result, we explain not only the approach used to prove it but also why that approach falls short of fully establishing the conjecture. Taken together, we hope these results and insights will provide a foundation for future work on proving the conjecture.

1) Optimality across "ignore/separate" classifiers: Recall that the optimal linear interval classifier derived in Theorem V.1 takes one of two forms. It either: 1) *ignores* the zero class (in the "rare zero class" case), or 2) *separates* the positive and negative classes by at least 2ε (in the "frequent zero class" case). In other words, the optimal linear interval classifier belongs to the following general family of classifiers that we will refer to as "ignore/separate" classifiers.

Definition V.3 (Ignore/separate classifiers)**.** *A classifier $\hat{y} : \mathbb{R}^p \to \{-1, 0, 1\}$ is an ignore/separate classifier if either*

$$\forall_x \;\; \hat{y}(x) \neq 0$$

or

$$\inf\big\{\|x_+ - x_-\|_2 : \hat{y}(x_+) = 1 \text{ and } \hat{y}(x_-) = -1\big\} > 2\varepsilon,$$

i.e., $\hat{y}$ either ignores the zero class or separates its positive and negative decision regions by at least 2ε.

This is a very large family of classifiers and importantly includes numerous nonlinear classifiers. Moreover, the classifiers it omits are all unlikely to be optimal. The omitted classifiers all include the zero class but still have positive and negative decision regions within 2ε. Recall that for linear interval classifiers, doing so is always suboptimal; the robust risk can be improved by absorbing the zero decision region into one of the others, i.e., by ignoring the zero class. While this does not necessarily imply that the same holds for nonlinear classifiers, it does make the family of ignore/separate classifiers a very natural one to consider.

The main result of this subsection is that the classifier from Theorem V.1 is not only optimal among linear interval classifiers but also across the family of ignore/separate classifiers. The result places no additional conditions on the model parameters; it holds for any class proportions (π_-, π_0, π_+), any spacing of the means (λ_+, λ_-), and so on.

Theorem V.4 (Linear interval classifiers are optimal across ignore/separate classifiers)**.** *Under the assumptions of Theorem V.1, the optimal linear interval classifier from Theorem V.1 is also optimal across all ignore/separate classifiers.*

We prove Theorem V.4 using the same overall strategy as the proof of Theorem IV.1. Namely, Theorem V.4 follows directly from the optimal linear interval classifier (Theorem V.1) combined with the following three-class variant of Lemma IV.2 that shows that linear interval classifiers dominate all ignore/separate classifiers.

Lemma V.5 (Linear interval classifiers dominate all ignore/separate classifiers)**.** *Under the assumptions of Theorem V.1, for any ignore/separate classifier $\hat{y} : \mathbb{R}^p \to \{-1, 0, 1\}$, there exists a linear interval classifier $\tilde{y}$ that dominates its robust misclassification on all three classes simultaneously, i.e.,*

$$\forall_{y \in \{-1,0,1\}} \quad \Pr_{x|y}\{\exists_{\delta : \|\delta\|_2 \leq \varepsilon} \;\; \tilde{y}(x + \delta) \neq y\}$$

[3] For Gaussians in arbitrary positions, i.e., with means that do not lie on a line, we do not expect linear interval classifiers to be optimal. See Section D for a more detailed discussion with some conjectures.

$$\leq \Pr_{x|y}\{\exists_{\delta:\|\delta\|_2\leq\varepsilon}\ \ \hat{y}(x+\delta)\neq y\}.$$

Thus, $R_{\rm rob}(\tilde{y},\varepsilon)\leq R_{\rm rob}(\hat{y},\varepsilon)$, i.e., the linear interval classifier $\tilde{y}$ also dominates the ignore/separate classifier $\hat{y}$ with respect to the robust risk.

The proof of this lemma follows a similar strategy as the proof of Lemma IV.2.

Proof of Lemma V.5: Let $\hat{y}:\mathbb{R}^p\to\{-1,0,1\}$ be an ignore/separate classifier and define

$$c_- \coloneqq \Phi^{-1}\Big(1-\Pr_-\{\hat{y}(x)\neq -1\}\Big)+\lambda_-,$$
$$c_+ \coloneqq \Phi^{-1}\Big(\Pr_+\{\hat{y}(x)\neq 1\}\Big)+\lambda_+,$$

where Φ is the cumulative distribution functions of the standard normal distribution (with $\Phi^{-1}(0)=-\infty$ and $\Phi^{-1}(1)=\infty$), and where we denote the conditional probabilities $\Pr_{x|y=-1}$, $\Pr_{x|y=0}$ and $\Pr_{x|y=1}$ by $\Pr_-$, $\Pr_0$ and $\Pr_+$, respectively. As one can readily verify, c_- and c_+ match the misclassification probabilities of $\hat{y}$ on the negative and positive classes:

$$\Pr_-\{x^\top\mu>c_-\}=\Pr_-\{\hat{y}(x)\neq -1\},\qquad(26)$$
$$\Pr_+\{x^\top\mu<c_+\}=\Pr_+\{\hat{y}(x)\neq 1\}.$$

We first show that $c_-\leq c_+$ so that they define a valid linear interval classifier. Let

$$S_-^c \coloneqq \{x: x^\top\mu>c_-\},\qquad \widehat{S}_-^c \coloneqq \{x:\hat{y}(x)\neq -1\},$$

where we view S_-^c as the rejection region of a hypothesis test of the negative class against the alternative of the zero class; likewise for $\widehat{S}_-^c$. Under this interpretation, it follows from (26) that the two tests have matching significance levels $\Pr_-(S_-^c)=\Pr_-(\widehat{S}_-^c)$. However, S_-^c is the rejection region of the likelihood ratio test here. Thus, it follows from the Neyman-Pearson lemma that the test corresponding to $\widehat{S}_-^c$ must have power less than or equal to that of the test corresponding to S_-^c. Namely,

$$\Pr_0\{x^\top\mu>c_-\}=\Pr_0(S_-^c)\geq\Pr_0(\widehat{S}_-^c)=\Pr_0\{\hat{y}(x)\neq -1\}.$$

Repeating the analogous argument for c_+ yields

$$\Pr_0\{x^\top\mu<c_+\}\geq\Pr_0\{\hat{y}(x)\neq 1\}.$$

As a result,

$$\Pr_0\{x^\top\mu>c_-\}+\Pr_0\{x^\top\mu<c_+\}$$
$$\geq\Pr_0\{\hat{y}(x)\neq -1\}+\Pr_0\{\hat{y}(x)\neq 1\}\geq 1,$$

from which we conclude that $c_-\leq c_+$.

Since $c_-\leq c_+$, the following linear interval classifier is well defined:

$$\tilde{y}(x)=\begin{cases}-1 & \text{if } x^\top\mu\leq c_-,\\ 0 & \text{if } c_-<x^\top\mu<c_+,\\ +1 & \text{if } x^\top\mu\geq c_+,\end{cases}$$

and it remains to show that

$$\forall_{y\in\{-1,0,1\}}\quad \Pr_{x|y}\{S_y^c(\tilde{y})+B_\varepsilon\}\leq\Pr_{x|y}\{S_y^c(\hat{y})+B_\varepsilon\}.$$

Applying the equality case of the Gaussian concentration of measure, similar to equations (11) and (12) in the proof of Lemma IV.2 in the two-class case, to the construction of c_- and c_+ immediately yields

$$\Pr_-\{S_-^c(\tilde{y})+B_\varepsilon\}\leq\Pr_-\{S_-^c(\hat{y})+B_\varepsilon\},$$
$$\Pr_+\{S_+^c(\tilde{y})+B_\varepsilon\}\leq\Pr_+\{S_+^c(\hat{y})+B_\varepsilon\},$$

so the result is shown for the negative and positive classes.

For the zero class, suppose first that $\hat{y}:\mathbb{R}^p\to\{-1,0,1\}$ ignores the zero class, i.e., $\forall_x\ \ \hat{y}(x)\neq 0$. Then, the decision region $S_0(\hat{y})$ is empty so we immediately have

$$\Pr_0\{S_0^c(\tilde{y})+B_\varepsilon\}\leq 1=\Pr_0\{S_0^c(\hat{y})+B_\varepsilon\}.$$

So it only remains to consider the case where $\hat{y}$ separates its positive and negative decision regions by at least 2ε. For this case, using $\Pr\{C\cup D\}\leq\Pr\{C\}+\Pr\{D\}$ for any measurable sets C,D, yields

$$\Pr_0\{S_0^c(\tilde{y})+B_\varepsilon\}=\Pr_0[\{S_-(\tilde{y})+B_\varepsilon\}\cup\{S_+(\tilde{y})+B_\varepsilon\}]$$
$$\leq\Pr_0\{S_-(\tilde{y})+B_\varepsilon\}+\Pr_0\{S_+(\tilde{y})+B_\varepsilon\},$$

and another application of the equality case in the Gaussian concentration of measure yields

$$\Pr_0\{S_-(\tilde{y})+B_\varepsilon\}\leq\Pr_0\{S_-(\hat{y})+B_\varepsilon\},$$
$$\Pr_0\{S_+(\tilde{y})+B_\varepsilon\}\leq\Pr_0\{S_+(\hat{y})+B_\varepsilon\}.$$

Putting these together yields

$$\begin{aligned}\Pr_0\{S_0^c(\tilde{y})+B_\varepsilon\}&\leq\Pr_0\{S_-(\tilde{y})+B_\varepsilon\}+\Pr_0\{S_+(\tilde{y})+B_\varepsilon\}\\&\leq\Pr_0\{S_-(\hat{y})+B_\varepsilon\}+\Pr_0\{S_+(\hat{y})+B_\varepsilon\}\\&=\Pr_0\{S_-(\hat{y})+B_\varepsilon\}+\Pr_0\{S_+(\hat{y})+B_\varepsilon\}\\&\quad-\Pr_0\{(S_-(\hat{y})+B_\varepsilon)\cap(S_+(\hat{y})+B_\varepsilon)\}\\&=\Pr_0\{(S_-(\hat{y})+B_\varepsilon)\cup(S_+(\hat{y})+B_\varepsilon)\}\\&=\Pr_0\{S_0^c(\hat{y})+B_\varepsilon\},\end{aligned}$$

where the first equality holds because the 2ε-separation of the positive and negative decision regions of $\hat{y}$ implies that $(S_-(\hat{y})+B_\varepsilon)\cap(S_+(\hat{y})+B_\varepsilon)=\emptyset$, the second equality is the inclusion-exclusion principle, and

$$\begin{aligned}(S_-(\hat{y})+B_\varepsilon)\cup(S_+(\hat{y})+B_\varepsilon)&=(S_-(\hat{y})\cup S_+(\hat{y}))+B_\varepsilon\\&=S_0^c(\hat{y})+B_\varepsilon.\end{aligned}$$

yields the third equality. ∎

Rigorously extending Theorem V.4 beyond ignore/separate classifiers turns out to be quite nontrivial. The proof technique used here encounters a major obstacle, as we will explain in the remainder of this subsection. The main issue is that the proof of the core lemma (Lemma V.5) crucially uses the fact that linear interval classifiers can match the robust misclassification of any ignore/separate classifier with respect to *all three classes simultaneously*. It is tempting to expect this fact to extend beyond ignore/separate classifiers to all classifiers. However, this is surprisingly not the case. Namely, there can exist a (nonlinear) classifier for which *all* linear

interval classifiers have worse robust misclassification on at least one class, as shown by the following counter-example.

Example V.6 (A classifier for which no linear interval classifier has matching robust classification performance on all three classes simultaneously)**.** *Let $p = 1$, $\mu = 1$, $\lambda_\pm = \pm 1$, $\varepsilon = 0.3$, $\pi_\pm = 1/4$ and $\pi_0 = 1/2$, and consider the classifier*

$$\hat{y}(x) = \begin{cases} -1, & \text{if } x < 1, \\ 1, & \text{if } 1 \le x < 2.15, \\ 0, & \text{if } 2.15 \le x < 4, \\ 1, & \text{if } x \ge 4. \end{cases}$$

For this classifier, the robust misclassification probabilities are:

$$\begin{aligned} \mathcal{M}_- &= \Pr_{x|y=-1}\{\exists_{\delta:\|\delta\|_2\le\varepsilon}\ \hat{y}(x+\delta) \neq -1\} \\ &= \Pr_{x|y=-1}(x \ge 1 - 0.3) \\ &\approx 0.0446, \\ \mathcal{M}_0 &= \Pr_{x|y=0}\{\exists_{\delta:\|\delta\|_2\le\varepsilon}\ \hat{y}(x+\delta) \neq 0\} \\ &= \Pr_{x|y=0}(x < 2.15 + 0.3) + \Pr_{x|y=0}(x \ge 4 - 0.3) \\ &\approx 0.9930, \\ \mathcal{M}_+ &= \Pr_{x|y=1}\{\exists_{\delta:\|\delta\|_2\le\varepsilon}\ \hat{y}(x+\delta) \neq 1\} \\ &= \Pr_{x|y=1}(x < 1 + 0.3) \\ &\quad + \Pr_{x|y=1}(2.15 - 0.3 \le x < 4 + 0.3) \\ &\approx 0.8151, \end{aligned}$$

where we use $\approx$ to indicate that the calculated values have been rounded to the digits shown.

Now for a linear interval classifier

$$\tilde{y}(x) = \begin{cases} -1, & \text{if } x \le c_-, \\ 0, & \text{if } c_- < x < c_+, \\ 1, & \text{if } x \ge c_+, \end{cases}$$

to match the robust misclassification probabilities $\mathcal{M}_-$ and $\mathcal{M}_+$, i.e., to have robust misclassification no worse on the negative and positive classes, the thresholds must satisfy

$$\begin{aligned} c_- \ge \tilde{c}_- &\coloneqq \bar{\Phi}^{-1}(\mathcal{M}_-) + \lambda_- + \varepsilon \approx 1.000, \\ c_+ \le \tilde{c}_+ &\coloneqq \Phi^{-1}(\mathcal{M}_+) + \lambda_+ - \varepsilon \approx 1.597, \end{aligned}$$

where Φ is the cumulative distribution function of the standard normal distribution, $\bar{\Phi} \coloneqq 1 - \Phi$, and Φ^{-1} and $\bar{\Phi}^{-1}$ are their inverses. Otherwise, if $c_- < \tilde{c}_-$ then

$$\begin{aligned} &\Pr_{x|y=-1}\{\exists_{\delta:\|\delta\|_2\le\varepsilon}\ \tilde{y}(x+\delta) \neq -1\} \\ &\quad = \Pr_{x|y=-1}(x > c_- - \varepsilon) > \Pr_{x|y=-1}(x > \tilde{c}_- - \varepsilon) = \mathcal{M}_-, \end{aligned}$$

and likewise if $c_+ > \tilde{c}_+$ then

$$\begin{aligned} &\Pr_{x|y=1}\{\exists_{\delta:\|\delta\|_2\le\varepsilon}\ \tilde{y}(x+\delta) \neq 1\} \\ &\quad = \Pr_{x|y=1}(x < c_+ + \varepsilon) > \Pr_{x|y=-1}(x < \tilde{c}_+ + \varepsilon) = \mathcal{M}_+. \end{aligned}$$

However, if $c_- \ge \tilde{c}_-$ and $c_+ \le \tilde{c}_+$ then

$$\begin{aligned} &\Pr_{x|y=0}\{\exists_{\delta:\|\delta\|_2\le\varepsilon}\ \tilde{y}(x+\delta) \neq 0\} \\ &\quad = \Pr_{x|y=0}(x \le c_- + \varepsilon \text{ or } x \ge c_+ - \varepsilon) \\ &\quad \ge \Pr_{x|y=0}(x \le \tilde{c}_- + \varepsilon \text{ or } x \ge \tilde{c}_+ - \varepsilon) = 1 > \mathcal{M}_0, \end{aligned}$$

since $\tilde{c}_- + \varepsilon \ge \tilde{c}_+ - \varepsilon$ here. Hence, there is no choice of c_- and c_+, i.e., there is no linear interval classifier $\tilde{y}$, that matches the robust misclassification probabilities of $\hat{y}$ for all *classes simultaneously.*

As a result, the strategy used to prove Lemma V.5 (and consequently Theorem V.4) cannot be directly used to go beyond ignore/separate classifiers. Note, however, that this does not mean that linear interval classifiers are not in fact optimal; it simply shows that the approach used before cannot be used here. Indeed, the nonlinear classifier $\hat{y}$ considered in Example V.6 has robust risk

$$R_{\mathrm{rob}}(\hat{y}, \varepsilon) = \pi_- \mathcal{M}_- + \pi_0 \mathcal{M}_0 + \pi_+ \mathcal{M}_+ \approx 0.7114,$$

while the corresponding optimal linear interval classifier $\hat{y}^*_{\mathrm{int}}$ from Theorem V.1 has worse robust misclassification probabilities on the positive and negative classes but still a better robust risk of $R_{\mathrm{rob}}(\hat{y}^*_{\mathrm{int}}, \varepsilon) \approx 0.4953$.

2) Optimality in the "sufficiently rare zero class" case: This subsection derives an optimality result that does not restrict the family of classifiers, but instead considers a restricted subset of model parameters. In particular, we show that the classifier from Theorem V.1 is optimal across all classifiers if the zero class is sufficiently rare.

Theorem V.7 (Linear interval classifiers are optimal if the zero class is sufficiently rare)**.** *Under the assumptions of Theorem V.1, suppose further that $\pi_0 \le \hat{\alpha}\sqrt{\pi_-\pi_+}$, where*

$$\hat{\alpha} \coloneqq \exp\left\{ -\frac{(|\lambda_+| - \varepsilon)(|\lambda_-| - \varepsilon)}{2} - \frac{\lambda_+ + \lambda_-}{|\lambda_+ - \lambda_-| - 2\varepsilon} \ln\gamma \right\}.$$

Then the optimal linear interval classifier from Theorem V.1 (using the "rare zero class" case) is also optimal across all classifiers.

This result does not have a restriction on the classifier family like Theorem V.4; we find the optimal classifier across all classifiers. However, the additional condition that $\pi_0 \le \hat{\alpha}\sqrt{\pi_-\pi_+}$ essentially limits this result to a subset of the "rare zero class" case in Theorem V.1; it does not apply to cases where we expect the optimal robust classifier to assign points to all three classes. As we explain at the end of this subsection (after proving Theorem V.7), extending this result to remove the condition turns out to be quite nontrivial.

We obtain Theorem V.7 through a different approach than Theorems IV.1 and V.4. In particular, we prove Theorem V.7 by using the same overall strategy as was used in [40] for two-class settings. In [40], an optimal robust classifier was derived by identifying a careful ε-perturbation of the means for which the corresponding Bayes optimal classifier has standard risk matching the robust risk with respect to the unperturbed means. Optimality then followed by exploiting

the insight that the ε-robust risk of any classifier is lower bounded by its standard risk with respect to any ε-perturbed means, which is in turn lower bounded by the standard risk of the corresponding Bayes optimal classifier. The proof of Theorem V.7 follows the same approach.

Proof of Theorem V.7: Consider the following ε-perturbed spacings for the means:

$$\begin{aligned}\lambda'_- &\coloneqq \lambda_- + \varepsilon = -(|\lambda_-| - \varepsilon) < 0,\\ \lambda'_0 &\coloneqq \lambda_0 = 0,\\ \lambda'_+ &\coloneqq \lambda_+ - \varepsilon = |\lambda_+| - \varepsilon > 0.\end{aligned}$$

We first show that the Bayes optimal classifier $\hat{y}^*_{\mathrm{Bay},\lambda'}(x)$ for these perturbed means has standard risk $R_{\mathrm{std},\lambda'}(\hat{y}^*_{\mathrm{Bay},\lambda'})$ with respect to the perturbed means matching its robust risk $R_{\mathrm{rob}}(\hat{y}^*_{\mathrm{Bay},\lambda'},\varepsilon)$ with respect to the unperturbed means. Note first that $\hat{y}^*_{\mathrm{Bay},\lambda'}(x)$ ignores the zero class when $\pi_0 \le \tilde{\alpha}\sqrt{\pi_-\pi_+}$. Indeed, if $\pi_0 \le \tilde{\alpha}\sqrt{\pi_-\pi_+}$, then

$$\begin{aligned}&\ln\left(\frac{\pi_0}{\sqrt{\pi_-\pi_+}}\right) \le \ln\tilde{\alpha}\\ &= -\frac{(|\lambda_+| - \varepsilon)(|\lambda_-| - \varepsilon)}{2} - \frac{\lambda_+ + \lambda_-}{|\lambda_+ - \lambda_-| - 2\varepsilon}\ln\gamma\\ &= \frac{\lambda'_+\lambda'_-}{2} - \frac{\lambda'_+ + \lambda'_-}{\lambda'_+ - \lambda'_-}\ln\gamma.\end{aligned}$$

Now, adding $\ln\gamma$ then simplifying, and likewise subtracting $\ln\gamma$ then simplifying, yields the following two inequalities

$$\begin{aligned}\ln\left(\frac{\pi_0}{\pi_+}\right) &= \ln\left(\frac{\pi_0}{\sqrt{\pi_-\pi_+}}\right) - \ln\gamma\\ &\le \frac{\lambda'_+\lambda'_-}{2} - \frac{\lambda'_+ + \lambda'_-}{\lambda'_+ - \lambda'_-}\ln\gamma - \ln\gamma\\ &= +\left[\frac{\lambda'_-}{2} - \frac{\ln(\pi_+/\pi_-)}{|\lambda'_+ - \lambda'_-|}\right]|\lambda'_+|,\\ \ln\left(\frac{\pi_0}{\pi_-}\right) &= \ln\left(\frac{\pi_0}{\sqrt{\pi_-\pi_+}}\right) + \ln\gamma\\ &\le \frac{\lambda'_+\lambda'_-}{2} - \frac{\lambda'_+ + \lambda'_-}{\lambda'_+ - \lambda'_-}\ln\gamma + \ln\gamma\\ &= -\left[\frac{\lambda'_+}{2} - \frac{\ln(\pi_+/\pi_-)}{|\lambda'_+ - \lambda'_-|}\right]|\lambda'_-|.\end{aligned}$$

Next, dividing the first inequality by $+|\lambda'_+|$ then adding $\lambda'_+/2$, and dividing the second inequality by $-|\lambda'_-|$ then adding $\lambda'_-/2$, yields

$$\begin{aligned}\frac{\lambda'_+}{2} + \frac{\ln(\pi_0/\pi_+)}{|\lambda'_+|} &\le \frac{\lambda'_+ + \lambda'_-}{2} - \frac{\ln(\pi_+/\pi_-)}{|\lambda'_+ - \lambda'_-|},\\ \frac{\lambda'_-}{2} - \frac{\ln(\pi_0/\pi_-)}{|\lambda'_-|} &\ge \frac{\lambda'_+ + \lambda'_-}{2} - \frac{\ln(\pi_+/\pi_-)}{|\lambda'_+ - \lambda'_-|},\end{aligned}$$

so the Bayes optimal classifier for the perturbed means is the following linear interval classifier with Bayes optimal thresholds from (17):

$$\hat{y}^*_{\mathrm{Bay},\lambda'}(x) = \hat{y}_{\mathrm{int}}(x;\mu,c'_+,c'_-),$$

where

$$\begin{aligned}c'_+ = c'_- &= \frac{\lambda'_+ + \lambda'_-}{2} - \frac{\ln(\pi_+/\pi_-)}{|\lambda'_+ - \lambda'_-|}\\ &= \frac{\lambda_+ + \lambda_-}{2} + \frac{\ln(\pi_-/\pi_+)}{|\lambda_+ - \lambda_-| - 2\varepsilon}.\end{aligned}$$

This classifier is exactly the optimal linear interval robust classifier from Theorem V.1 in the "rare zero class" case, and it ignores the zero class. Since $\hat{y}^*_{\mathrm{Bay},\lambda'}(x)$ ignores the zero class, we finally have

$$\begin{aligned}R_{\mathrm{rob}}(\hat{y}^*_{\mathrm{Bay},\lambda'},\varepsilon) &= \pi_0 + \pi_+ \Pr_{x|y=+1}(x^\top\mu < c'_+ + \varepsilon)\\ &\quad + \pi_- \Pr_{x|y=-1}(x^\top\mu > c'_- - \varepsilon)\\ &= \pi_0 + \pi_+ \Pr_{\tilde{x}\sim\mathcal{N}(\lambda_+,1)}(\tilde{x} < c'_+ + \varepsilon)\\ &\quad + \pi_- \Pr_{\tilde{x}\sim\mathcal{N}(\lambda_-,1)}(\tilde{x} > c'_- - \varepsilon)\\ &= \pi_0 + \pi_+ \Pr_{\tilde{x}\sim\mathcal{N}(\lambda'_+,1)}(\tilde{x} < c'_+)\\ &\quad + \pi_- \Pr_{\tilde{x}\sim\mathcal{N}(\lambda'_-,1)}(\tilde{x} > c'_-)\\ &= R_{\mathrm{std},\lambda'}(\hat{y}^*_{\mathrm{Bay},\lambda'}).\end{aligned} \tag{27}$$

Namely, the standard risk $R_{\mathrm{std},\lambda'}(\hat{y}^*_{\mathrm{Bay},\lambda'})$ of $\hat{y}^*_{\mathrm{Bay},\lambda'}(x)$ with respect to the perturbed means matches the corresponding robust risk $R_{\mathrm{rob}}(\hat{y}^*_{\mathrm{Bay},\lambda'},\varepsilon)$ with respect to the unperturbed means.

The proof now concludes by observing that for any classifier $\hat{y}:\mathbb{R}^p \to \{-1,0,1\}$

$$\begin{aligned}&R_{\mathrm{rob}}(\hat{y},\varepsilon)\\ &= \pi_+ \Pr_{x\sim\mathcal{N}(\lambda_+\mu,I_p)}\left\{\exists_{\delta:\|\delta\|_2\le\varepsilon}\ \hat{y}(x+\delta) \ne +1\right\}\\ &\quad + \pi_0 \Pr_{x\sim\mathcal{N}(0,I_p)}\left\{\exists_{\delta:\|\delta\|_2\le\varepsilon}\ \hat{y}(x+\delta) \ne 0\right\}\\ &\quad + \pi_- \Pr_{x\sim\mathcal{N}(\lambda_-\mu,I_p)}\left\{\exists_{\delta:\|\delta\|_2\le\varepsilon}\ \hat{y}(x+\delta) \ne -1\right\}\\ &\ge \pi_+ \Pr_{x\sim\mathcal{N}(\lambda_+\mu,I_p)}\left\{\hat{y}(x-\varepsilon\mu) \ne +1\right\}\\ &\quad + \pi_0 \Pr_{x\sim\mathcal{N}(0,I_p)}\left\{\hat{y}(x) \ne 0\right\}\\ &\quad + \pi_- \Pr_{x\sim\mathcal{N}(\lambda_-\mu,I_p)}\left\{\hat{y}(x+\varepsilon\mu) \ne -1\right\}\\ &= \pi_+ \Pr_{\tilde{x}\sim\mathcal{N}(\lambda'_+\mu,I_p)}\left\{\hat{y}(\tilde{x}) \ne +1\right\}\\ &\quad + \pi_0 \Pr_{\tilde{x}\sim\mathcal{N}(0,I_p)}\left\{\hat{y}(\tilde{x}) \ne 0\right\}\\ &\quad + \pi_- \Pr_{\tilde{x}\sim\mathcal{N}(\lambda'_-\mu,I_p)}\left\{\hat{y}(x) \ne -1\right\}\\ &= R_{\mathrm{std},\lambda'}(\hat{y}),\end{aligned} \tag{28}$$

so for any classifier $\hat{y}:\mathbb{R}^p \to \{-1,0,1\}$ we finally have that

$$\begin{aligned}R_{\mathrm{rob}}(\hat{y},\varepsilon) &\ge R_{\mathrm{std},\lambda'}(\hat{y})\\ &\ge R_{\mathrm{std},\lambda'}(\hat{y}^*_{\mathrm{Bay},\lambda'}) = R_{\mathrm{rob}}(\hat{y}^*_{\mathrm{Bay},\lambda'},\varepsilon),\end{aligned}$$

where the first inequality is (28), the second inequality is from the definition of the Bayes optimal classifier, and the final equality is from (27). ∎

As mentioned above, rigorously extending Theorem V.7 beyond the "sufficiently rare zero class" case turns out to be

quite nontrivial. The proof technique used here encounters a major obstacle, as we will explain in the remainder of this subsection. The main issue is that a crucial step in the above proof was to find ε-perturbations of the means for which the corresponding Bayes optimal classifier has standard risk matching the robust risk with respect to the unperturbed means. Such perturbations always exist in the two-class setting studied by [40], and it is tempting to hope that the same may hold for the three-class setting we consider. However, this is not the case, as the following counter-example illustrates.

Example V.8 (A case where no set of ε-perturbed means produce matching robust and standard risk.)**.** *Let $p = 1$, $\mu = 1$, $\lambda_\pm = \pm 1$, $\varepsilon = 0.3$, and $\pi_\pm = \pi_0 = 1/3$. Setting $\mu = 1$ is without loss of generality. Then for any ε-perturbed means*

$$\lambda'_- \in [-1.3, -0.7], \quad \lambda'_0 \in [-0.3, 0.3], \quad \lambda'_+ \in [0.7, 1.3],$$

the robust risk $R_{\mathrm{rob}}(\hat{y}^*_{\mathrm{Bay},\lambda'}, \varepsilon)$ is lower bounded by the robust risk of the optimal linear interval classifier $\hat{y}^*_{\mathrm{int}}$ derived in Theorem V.1, i.e.,

$$\forall_{\lambda'_- \in [-1.3,-0.7]} \ \forall_{\lambda'_0 \in [-0.3,0.3]} \ \forall_{\lambda'_+ \in [0.7,1.3]} \quad R_{\mathrm{rob}}(\hat{y}^*_{\mathrm{Bay},\lambda'}, \varepsilon) \geq R_{\mathrm{rob}}(\hat{y}^*_{\mathrm{int}}, \varepsilon) \approx 0.56,$$

since $\hat{y}^*_{\mathrm{Bay},\lambda'}$ is always a linear interval classifier here and hence sub-optimal with respect to $\hat{y}^*_{\mathrm{int}}$.

However, the standard risk $R_{\mathrm{std},\lambda'}(\hat{y}^*_{\mathrm{Bay},\lambda'})$ with respect to the ε-perturbed means λ' is upper bounded as

$$\begin{aligned}
R_{\mathrm{std},\lambda'}(\hat{y}^*_{\mathrm{Bay},\lambda'}) &\leq R_{\mathrm{std},\lambda'}(\hat{y}^*_{\mathrm{Bay},\lambda}) \\
&= \pi_- \Pr_{x \sim \mathcal{N}(\lambda'_-, 1)} \left\{ x > -\frac{1}{2} \right\} \\
&\quad + \pi_0 \Pr_{x \sim \mathcal{N}(\lambda'_0, 1)} \left\{ x < -\frac{1}{2} \text{ or } x > +\frac{1}{2} \right\} \\
&\quad + \pi_+ \Pr_{x \sim \mathcal{N}(\lambda'_+, 1)} \left\{ x < +\frac{1}{2} \right\} \\
&\leq \frac{1}{3} \Pr_{x \sim \mathcal{N}(-0.7, 1)} \left\{ x > -\frac{1}{2} \right\} \\
&\quad + \frac{1}{3} \Pr_{x \sim \mathcal{N}(0.3, 1)} \left\{ x < -\frac{1}{2} \text{ or } x > +\frac{1}{2} \right\} \\
&\quad + \frac{1}{3} \Pr_{x \sim \mathcal{N}(+0.7, 1)} \left\{ x < +\frac{1}{2} \right\} \\
&\approx 0.49.
\end{aligned}$$

Thus, there is no choice of ε-perturbed means λ'_-, λ'_0, and λ'_+ for which $R_{\mathrm{rob}}(\hat{y}^*_{\mathrm{Bay},\lambda'}, \varepsilon) = R_{\mathrm{std},\lambda'}(\hat{y}^*_{\mathrm{Bay},\lambda'})$.

Essentially, the issue is that the robust misclassification probabilities for the positive and negative classes can be matched by perturbing the positive and negative means, respectively, but the same cannot be done for the zero class in general. Finding these perturbations is central to the approach used to prove Theorem V.7. As a result, this approach cannot be directly used to generalize beyond the sufficiently rare zero class case.

VI. Optimal ℓ_∞ robust classifiers

We now shift our attention from ℓ_2 to ℓ_∞ adversaries, i.e., perturbations up to an ℓ_∞ radius, and seek to minimize the robust risk $R_{\mathrm{rob}}(\hat{y}, \varepsilon, \|\cdot\|_\infty)$. Doing so introduces new challenges: the rotational invariant geometry of ℓ_2 allowed a reduction to the simpler one-dimensional case, but this does not apply here. However, the next result captures one setting where the geometry is favorable and the ℓ_2 findings of Section IV extend to ℓ_∞ robustness. We provide its proof in Section E-A.

Corollary VI.1 (Optimal ℓ_∞ robust classifiers for one-sparse means)**.** *Let the data (x, y) follow the two-class Gaussian model* (5)*, and let μ have exactly one non-zero coordinate $\mu_j > 0$ and $\varepsilon < \mu_j$. An optimal ℓ_∞ robust classifier is*

$$\hat{y}^*(x) \coloneqq \mathrm{sign}\{x_j(\mu_j - \varepsilon) - q/2\}, \tag{29}$$

where $q = \ln\{(1-\pi)/\pi\}$.

In essence, the ℓ_2 and ℓ_∞ norms agree when restricted to the nonzero coordinate, enabling us to extend Theorem IV.1. The same applies to the three-class setting of Section V with a similar extension of Theorem V.1 as follows. We provide its proof in Section E-B.

Corollary VI.2 (Optimal linear interval ℓ_∞ robust classifiers for one-sparse means – three classes)**.** *Suppose data (x, y) are from the three-class Gaussian model of Section V, μ has exactly one non-zero coordinate $\mu_j > 0$ and $\varepsilon < \mu_j/2$. An optimal linear interval ℓ_∞ robust classifier is:*

$$\hat{y}^*_{\mathrm{int}}(x) \coloneqq \hat{y}_{\mathrm{int}}(x_j; 1, c^*_+, c^*_-),$$

where the thresholds $c^*_+ \geq c^*_-$ are as follows:
Case 1. If $\pi_0 \leq \alpha^* \sqrt{\pi_- \pi_+}$, then

$$c^*_+ = c^*_- = \ln(\pi_-/\pi_+)/(2\mu_j - 2\varepsilon).$$

Case 2. Otherwise, $c^*_+ - c^*_- > 2\varepsilon$, with

$$\begin{aligned}
c^*_+ &= +\mu_j/2 + \ln(\pi_0/\pi_+)/(\mu_j - 2\varepsilon), \\
c^*_- &= -\mu_j/2 - \ln(\pi_0/\pi_-)/(\mu_j - 2\varepsilon).
\end{aligned}$$

The cutoff α^* is the unique solution to the equation:

$$\begin{aligned}
&(\gamma + \gamma^{-1}) R^*_{\mathrm{rob}}\{\mu_j, \gamma/(\gamma + \gamma^{-1}); \varepsilon\} \\
&\quad = (\gamma + \alpha) R^*_{\mathrm{rob}}\{\mu_j/2, \gamma/(\gamma + \alpha); \varepsilon\} \\
&\qquad + (\gamma^{-1} + \alpha) R^*_{\mathrm{rob}}\{\mu_j/2, \gamma^{-1}/(\gamma^{-1} + \alpha); \varepsilon\} - \alpha,
\end{aligned}$$

in the domain $\alpha \geq \exp\{-(\mu_j - 2\varepsilon)^2/2\}$ with $\gamma \coloneqq \sqrt{\pi_+/\pi_-}$; $\alpha^* = \exp(-\mu_j^2/2)$ when $\varepsilon = 0$.

Removing the restriction that μ be one-sparse is highly nontrivial in general, but it turns out to be possible if we instead consider only *linear* classifiers: $\hat{y}_{\mathrm{lin}}(x; w, c) = \mathrm{sign}(x^\top w - c)$. This statement for the balanced case has been derived in [67, Lemma 1]. However our result generalizes it to the imbalanced case.

Theorem VI.3 (Optimal linear ℓ_∞ robust classifiers)**.** *Suppose data (x, y) are from the two-class Gaussian model* (5)*. An optimal linear ℓ_∞ robust classifier is: $\hat{y}^*(x) \coloneqq \mathrm{sign}\{x^\top \eta_\varepsilon(\mu) -$*

$q/2\}$, *where* $q = \ln\{(1-\pi)/\pi\}$ *and the soft-thresholding operator*

$$\eta_\varepsilon(x) \coloneqq \begin{cases} x-\varepsilon, & \text{if } x \geq \varepsilon, \\ 0, & \text{if } x \in (-\varepsilon, \varepsilon), \\ x+\varepsilon, & \text{if } x \leq -\varepsilon, \end{cases} \tag{30}$$

is applied element-wise to the vector $\mu \in \mathbb{R}^p$.

The proof is based on a connection to the well-known water-filling optimization problem, and is provided in Section E-C. The analogous extension again holds for three classes.

Theorem VI.4 (Optimal linear interval ℓ_∞ robust classifiers – three classes)**.** *Suppose data* (x, y) *are from the three-class Gaussian model of Section V and* $\varepsilon < \|\mu\|_\infty/2$*. An optimal linear interval* ℓ_∞ *robust classifier is either:*

1) $\hat{y}_{\text{int}}\{x; \eta_\varepsilon(\mu), c^*, c^*\}$*, where* $c^* = \ln(\pi_-/\pi_+)/2$*, or*
2) $\hat{y}_{\text{int}}\{x; \eta_{2\varepsilon}(\mu), c^*_+, c^*_-\}$*, where* $c^*_\pm = \pm\eta_{2\varepsilon}(\mu)^\top\mu/2 \pm \ln(\pi_0/\pi_\pm)$*,*

where the second case applies only when $c^*_+ \geq c^*_-$.

We provide its proof in Section E-D.

VII. Landscape of the robust risk

Sections IV to VI theoretically optimized the robust risk, but left open important questions about its *optimization landscape*, which can be non-convex and challenging to optimize. For example, what happens if we use surrogate losses as is commonly done in practice? This section makes progress on this question.

Consider data (x, y) from the two-class Gaussian model (5) with linear classifiers and corresponding ℓ-robust risk as a function of weights $w \in \mathbb{R}^p$ and bias $c \in \mathbb{R}$:

$$\widetilde{R}_{\varepsilon,\|\cdot\|,\ell}(w, c) \coloneqq \mathbb{E}_{x,y} \sup_{\|\delta\| \leq \varepsilon} \ell[\{w^\top(x+\delta) - c\} \cdot y]. \tag{31}$$

The 0-1 loss $\bar{\ell}(z) = I(z \leq 0)$ yields $\widetilde{R}_{\varepsilon,\|\cdot\|,\bar{\ell}}(w, c) = R_{\text{rob}}\{\text{sign}(w^\top x - c), \varepsilon, \|\cdot\|\}$.

It is common to use surrogate losses ℓ in (31) such as the logistic loss $\ell(z) = \log(1+\exp(-z))$, the exponential loss $\ell(z) = \exp(-z)$, or the hinge loss $\ell(z) = (1-z)_+$. The impact of doing so is well-studied in standard settings [68], but has remained an important open problem in the adversarial setting. Minimizing a surrogate loss here does not in general produce optimal weights for the 0-1 loss, but it does so in a few settings which the next result describes.

Theorem VII.1 (Classification calibration)**.** *Let* $w^* \in \mathbb{R}^p$ *be the optimal weights for a linear classifier with no bias term, i.e.,* w^* *minimizes* $\widetilde{R}_{\varepsilon,\|\cdot\|,\bar{\ell}}(w, 0)$ *with the 0-1 loss* $\bar{\ell}$*. Any strictly decreasing surrogate loss* ℓ *is classification calibrated; minimizing the* ℓ*-robust risk* $\widetilde{R}_{\varepsilon,\|\cdot\|,\ell}(w, 0)$ *recovers* w^*.

Furthermore, calibration extends to the case with bias, i.e., jointly minimizing $\widetilde{R}_{\varepsilon,\|\cdot\|,\ell}(w, c)$ *produces* $(w^*, 0)$ *if either: i)* ℓ *is convex, or ii) the classes are balanced, i.e.,* $\pi = 1/2$.

Theorem VII.1 partially extends to surrogate losses ℓ that are decreasing but not strictly so. In this case, w^* still minimizes the ℓ-robust risks but might not do so uniquely.

Proof of Theorem VII.1: Let $\|\cdot\|_*$ be the dual norm of $\|\cdot\|$. This is defined as $\|w\|_* = \sup w^\top z$, subject to $\|z\| \leq 1$. Since ℓ is decreasing, as is well known (see, e.g., [69]), we have

$$\begin{aligned} R(\ell, w, b, \varepsilon, \|\cdot\|) &= \mathbb{E}_{x,y} \sup_{\|\delta\| \leq \varepsilon} \ell([w^\top(x+\delta) + b] \cdot y) \\ &= \mathbb{E}_{x,y} \ell(y \cdot [w^\top x + b] - \varepsilon \cdot \|w\|_*). \end{aligned}$$

This shows that for any candidate w, the worst-case perturbations are equal to the conjugate of w, with respect to the $\|\cdot\|$ norm, namely $\delta^*(x) = -\hat{y}(x) \cdot \varepsilon \cdot w^*$, where w^* solves $\|w\|_* = \sup w^\top z$, subject to $\|z\| \leq 1$.

Now, in our case, due to the distributional assumption on the data, we have $y \cdot x \sim \mathcal{N}(\mu, I_p)$. Moreover, $y \cdot w^\top x \sim \mathcal{N}(w^\top\mu, \|w\|_2^2 I_p)$. It is readily verified that $y \cdot w^\top x$ is probabilistically independent of y. Therefore, we can write, for some $z \sim \mathcal{N}(0,1)$ independent of y

$$\begin{aligned} &R(\ell, w, b, \varepsilon, \|\cdot\|) \\ &\quad = \mathbb{E}_{z,y} \ell(w^\top\mu - \varepsilon \cdot \|w\|_* + by + \sigma \cdot \|w\|_2 \cdot z). \end{aligned}$$

Now we discuss the cases considered in the theorem.

1) If minimizing restricted to $b = 0$, the inner term reduces to $\ell(w^\top\mu - \varepsilon \cdot \|w\|_* + \sigma \cdot \|w\|_2 \cdot z)$.
2) When the loss is strictly convex, then by Jensen's inequality we obtain

$$\begin{aligned} &\mathbb{E}_y \ell(w^\top\mu - \varepsilon \cdot \|w\|_* + by + \sigma \cdot \|w\|_2 \cdot z) \\ &\quad \geq \ell(w^\top\mu - \varepsilon \cdot \|w\|_* + \sigma \cdot \|w\|_2 \cdot z). \end{aligned}$$

In both cases it is enough to minimize the objective

$$R(\ell, w, \varepsilon, \|\cdot\|) = \mathbb{E}_{z,y} \ell(w^\top\mu - \varepsilon \cdot \|w\|_* + \sigma \cdot \|w\|_2 \cdot z).$$

Now fix $\|w\|_2 = 1$. It is readily verified that, when the loss is strictly decreasing and as the normal random variable is symmetric, this is equivalent to maximizing the inner argument. When the loss is decreasing but not necessarily strictly monotonic, maximizing the inner argument is still a sufficient condition that guarantees the risk is minimized; however in this latter case there may be other minimizers of the risk. Therefore, it is enough to maximize the inner argument.

That is, we study maximizing, subject to $\|w\|_2 = c > 0$,

$$w^\top\mu - \varepsilon \cdot \|w\|_*.$$

Given the homogeneity of the norms, we thus conclude that the optimal w minimizing the robust ℓ-risk

$$R(\ell, w, b, \varepsilon, \|\cdot\|) = \mathbb{E}_{x,y} \sup_{\|\delta\| \leq \varepsilon} \ell([w^\top(x+\delta) + b] \cdot y). \tag{32}$$

maximize

$$\frac{w^\top\mu - \varepsilon \cdot \|w\|_*}{\|w\|_2}. \tag{33}$$

Next, we study how to minimize the true robust risk. This is similar to the derivation for the optimal robust classifier. We will assume without loss of generality that $\sigma = 1$. As above, recall our general formula:

$$\begin{aligned} R(\hat{y}, \varepsilon) = \pi \cdot P_{x|y=1}(S_{-1} + B_\varepsilon) \\ + (1-\pi) \cdot P_{x|y=-1}(S_1 + B_\varepsilon). \end{aligned}$$

For a linear classifier $\hat{y}^*(x) = \text{sign}(x^\top w + b)$, we can restrict without loss of generality to w such that $\|w\| = 1$. The classifiers are scale invariant, and so we get the same predictions for all scaled versions of the weights w, by changing b appropriately. Then $S_1 + B_\varepsilon$ is the set of datapoints such that $x^\top w + b \geq -\varepsilon\|w\|_*$. Thus,

$$\begin{aligned} R(w, b; \varepsilon) &= \pi \cdot P_{\mathcal{N}(\mu, I)}(x^\top w + b \leq \varepsilon\|w\|_*) \\ &\quad + (1-\pi) \cdot P_{\mathcal{N}(-\mu, I)}(x^\top w + b \geq -\varepsilon\|w\|_*) \\ &= \pi \cdot \Phi\left(\varepsilon\|w\|_* - b - \mu^\top w\right) \\ &\quad + (1-\pi) \cdot \Phi\left(\varepsilon\|w\|_* + b - \mu^\top w\right). \end{aligned}$$

Now we examine the cases of unrestricted bias (general b), and zero bias (b constrained to zero) in turn. For the zero bias case we find

$$R(w; \varepsilon) = \Phi\left(\varepsilon\|w\|_* - \mu^\top w\right).$$

Another way to put this is that for a weight w with unit norm $\|w\| = 1$, a linear classifier reduces the effect size from $\mu^\top w$ (which we can assume to be positive, without loss of generality, by flipping the sign if needed), to $\mu^\top w - \varepsilon\|w\|_*$. So the optimal w minimizing the true robust risk solves

$$\sup_w \ \mu^\top w - \varepsilon\|w\|_* \quad \text{s.t.} \quad \|w\|_2 = 1.$$

Recalling again that the original problem is scale-invariant, it follows that this is equivalent to maximizing (33). Therefore, the optimal linear classifier for the true and surrogate robust risks coincide.

For the general bias case, we recall that the minimizer of $b \to \pi \cdot \Phi(c-b) + (1-\pi) \cdot \Phi(c+b)$ occurs at $b = \ln[(1-\pi)/\pi]/c$. Plugging back, we find that the "profile risk", minimized over b, equals, with $c(w) := \varepsilon\|w\|_* - \mu^\top w$, and $q := \ln[(1-\pi)/\pi]$,

$$\begin{aligned} R_{prof}(w; \varepsilon) &= \pi \cdot \Phi\left(c(w) - q/c(w)\right) \\ &\quad + (1-\pi) \cdot \Phi\left(c(w) + q/c(w)\right). \end{aligned}$$

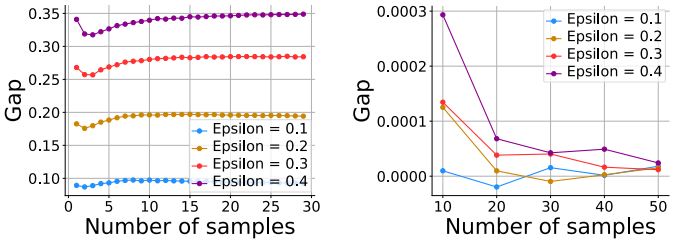
Clearly, this may in general minimizers other than the ones above. This shows that in general, surrogate loss minimization is not consistent. An exception is when $\pi = 1/2$, in which case $q = 0$, and the optimal bias in the robust risk is $b = 0$. This finishes the proof. ∎

VIII. Finite sample analysis

Having studied optimal population robust classifiers, we now consider robust linear classifiers learned from finitely many samples $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \{\pm 1\}$. This section does not assume Gaussianity; much of our subsequent analysis turns out to not rely on it. Here, we learn classifiers by minimizing the *empirical* ℓ-robust risk with a decreasing loss functional ℓ:

$$\begin{aligned} \widehat{R}^{(n)}_{\varepsilon, \|\cdot\|, \ell}(w, c) &:= \frac{1}{n} \sum_{i=1}^n \sup_{\|\delta\| \leq \varepsilon} \ell[\{w^\top(x_i + \delta) - c\} \cdot y_i] \quad (34) \\ &= \frac{1}{n} \sum_{i=1}^n \ell\{(w^\top x_i - c) \cdot y_i - \varepsilon\|w\|_*\}, \end{aligned}$$

where $\|\cdot\|_*$ is the dual norm and the equality holds because ℓ is decreasing; see, e.g., [69]. Using the 0-1 loss $\bar{\ell}(z) = I(z \leq 0)$ yields a non-convex and discontinuous empirical robust risk $\widehat{R}^{(n)}_{\varepsilon, \|\cdot\|, \bar{\ell}}(w, c)$, making optimization challenging. So one often uses *convex* surrogates instead, making $\widehat{R}^{(n)}_{\varepsilon, \|\cdot\|, \ell}(w, c)$ convex (decreasing convex functions of concave functions are convex).



(a) Linear surrogate loss. (b) Hinge surrogate loss.

Fig. 4. Mean gap between robust and standard risks of optimal finite-sample ℓ_∞ robust classifiers obtained via empirical robust risk minimization. Here we set the dimension $p = 5$, mean vector $\mu = 1/2 \cdot \mathbb{1}$, and class proportion $\pi = 1/2$.

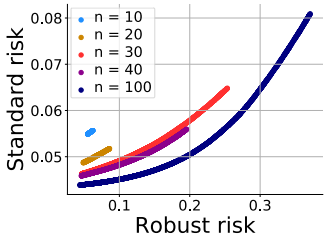


Fig. 5. Tradeoff between (population) standard and robust risk for $\varepsilon \in [0, 1]$ for classifiers obtained via Proposition VIII.1. Here we set $p = 5$, $\mu = 1/2 \cdot \mathbb{1}$, $\pi = 1/2$.

Hence, the empirical ℓ-robust risk (34) can be efficiently minimized for ℓ_∞ adversaries with convex decreasing surrogates such as the linear and hinge losses. Given n samples, this gives optimal weights $\hat{w}_n \in \mathbb{R}^p$ and a classifier $\hat{y}_n(x) = \text{sign}(x^\top \hat{w}_n)$, where throughout we will fix the bias $c = 0$. We will study these classifiers for the linear and hinge losses.

To study the tradeoff between standard and robust classifiers in finite samples, inspired by [17], in Figure 4 we plot the mean gaps between the population robust and standard risks, i.e., $R_{\text{rob}}(\hat{y}_n, \varepsilon, \|\cdot\|_\infty) - R_{\text{std}}(\hat{y}_n)$, as a function of the number of samples n, in the two-class Gaussian model (5). If the gap is large, then the robust risk is much greater than the standard risk for the optimal robust classifiers, yielding an unfavorable tradeoff. For the linear loss (constrained to $\|w\|_2 \leq 1$ to ensure boundedness), the gap between the standard and robust risks is large as n grows, consistent with [17]. However under the hinge loss, regardless of the value of ε, the gap decreases, which had not been investigated in [17]. This shows that the loss functional matters in robust risk minimization, consistent with our landscape results and expanding on the observations in [17].

A. Optimal empirical robust classifiers

The empirical risk-minimization perspective we have described gives an effective procedure for obtaining robust classifiers. Moreover, in some special cases we can also

derive explicit optimal empirical ℓ-robust classifiers. The next proposition does so for ℓ_∞ adversaries with linear loss where we again drop the bias term, i.e., $c = 0$.

Proposition VIII.1. *For the linear loss $\ell(z) = -z$, the empirical ℓ_∞ robust risk $\widehat{R}^{(n)}_{\varepsilon,\|\cdot\|_\infty,\ell}(w, 0)$ constrained to $\|w\|_2 \le 1$ is minimized by $w^* := \eta_\varepsilon(\hat{\mu})/\left\|\eta_\varepsilon(\hat{\mu})\right\|_2$. Here η is the soft-thresholding operator* (30)*, which is applied element-wise to the empirical mean vector $\hat{\mu} := (1/n)\sum_{i=1}^n y_i x_i \in \mathbb{R}^p$.*

Interestingly, these finite-sample weights can be viewed as plug-in estimates of the *population* optimal weights $\eta_\varepsilon(\mu)$ from Theorem VI.3 for the two-class Gaussian model (5), where the *empirical* mean $\hat{\mu}$ is substituted for the population mean μ. In Figure 4, we illustrate the tradeoff between population standard and robust risk for classifiers obtained via Proposition VIII.1.

B. Convergence of robust risk minimization

In this section, we quantify the concentration of the empirical robust risk $\widehat{R}^{(n)}_{\varepsilon,\|\cdot\|,\bar{\ell}}(w, c)$ around its population analogue $\widetilde{R}_{\varepsilon,\|\cdot\|,\bar{\ell}}(w, c)$, where $\bar{\ell}$ is again the 0-1 loss. Notably, in this result $x_i|y_i$ need not be Gaussian.

Theorem VIII.2 (Convergence of empirical robust risk for linear classifiers)**.** *For any $\delta > 0$,*

$$\Pr\left\{\forall_{(w,c)\in\mathbb{R}^p\times\mathbb{R}} \quad \left|\widehat{R}^{(n)}_{\varepsilon,\|\cdot\|,\bar{\ell}}(w, c) - \widetilde{R}_{\varepsilon,\|\cdot\|,\bar{\ell}}(w, c)\right| \le \delta\right\} \ge 1 - \exp(C(p - \delta^2 n)),$$

where C is a constant independent of n, d, and the probability is with respect to the n independent identically distributed samples $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \{\pm 1\}$.

Put another way, the empirical robust risk concentrates *uniformly* across all linear classifiers at a rate $O(\sqrt{p/n})$. The proof uses that the ε-expansions of half-spaces are still half-spaces, enabling arguments by VC-dimension. Characterizing more general classifiers is highly nontrivial since the ε-expansion of a finite VC dimension hypothesis class can have infinite VC dimension [70]. However, for one-dimensional data it turns out that we can generalize to classifiers that assign *finite unions of intervals* to each class.

Theorem VIII.3 (Convergence rate of empirical robust risk in 1D)**.** *In the setting of Theorem VIII.2 for any $\delta > 0$, we have uniformly over all classifiers $\hat{y}$ whose classification regions are unions of at most $2k$ intervals that $|R_n(\hat{y}, \varepsilon) - R(\hat{y}, \varepsilon)| \le \delta$ with probability at least $1 - 4\exp(-2n\delta^2/k^2)$, for the empirical robust risk R_n and the population robust risk R.*

Thus uniform concentration for k intervals occurs at rate $O(k/\sqrt{n})$. The proof is based on the Dvoretzky-Kiefer-Wolfowitz inequality.

Proof of Theorem VIII.3: As we have previously argued, the robust risk can be expressed as follows

$$R(\hat{y}, \varepsilon) = P(y = 1)P_{x|y=1}(S_{-1} + B_\varepsilon) + P(y = -1)P_{x|y=-1}(S_1 + B_\varepsilon).$$

Now let (x_i, y_i) for $i = 1, \ldots, n$ be sampled iid from a joint distribution $P_{x,y}$ for $i = 1, \ldots, n$. Let the fraction of 1-s be $\pi_n \in [0, 1]$. Let $P_{n\pm}$ be the empirical distributions of x_i given $y_i = 1$ and -1, respectively. We can write the finite sample robust risk as

$$R_n(\hat{y}, \varepsilon) = \pi_n \cdot P_{n+}(S_{-1} + B_\varepsilon) + (1 - \pi_n) \cdot P_{n-}(S_1 + B_\varepsilon).$$

Now $P_{n\pm}$ are empirical distributions that will converge to the limiting distributions under certain conditions.

Consider classifiers $\hat{y}$ whose decision boundaries are at most k points. For instance, if $k = 1$, then these are linear classifiers. Then $S_{\pm 1}$ each consist of a union of at most $j = \lceil k/2 \rceil$ disjoint intervals (finite or semi-infinite). Let I_j denote the collection of all such subsets of the real line, unions of at most j disjoint finite or semi-infinite intervals. Thus $S_{\pm 1} \in I_j$. Critically, the ε-expansions also have this property: by expanding the intervals, we still get intervals, merging them as needed. Thus, $S_{\pm 1} + B_\varepsilon \in I_j$.

Now, the classical Dvoretzky-Kiefer-Wolfowitz inequality [71]–[73] states the following. Let F_n be the CDF of n iid samples with CDF F. For every $\delta > 0$,

$$\Pr(\sup_x |F_n(x) - F(x)| > \delta) \le 2\exp(-2n\delta^2).$$

Let $\delta_n(x) = P_{n+}(-\infty, x] - P_+(-\infty, x]$. Consider the event $\sup_c |\delta_n(c)| \le \delta$, which happens with probability at least $1 - 2\exp(-2n\delta^2)$. On this event, we have

$$\begin{aligned}
&|P_{n+}(S_{-1} + B_\varepsilon) - P_+(S_{-1} + B_\varepsilon)| \\
&\le \sup_{A\in I_j} |P_{n+}(A) - P_+(A)| \\
&= \sup_{c_1<c_2<\ldots<c_j} |\delta_n(c_1) - \delta_n(c_2) + \cdots + (-1)^{j-1}\delta_n(c_j)| \\
&\le j \cdot \sup_c |\delta_n(c)| \le j\delta.
\end{aligned}$$

A similar argument applies to S_1. Then, on the intersection of the two events, which happens with probability $1 - 4\exp(-2n\delta^2)$, we find that

$$\begin{aligned}
&|R_n(\hat{y}, \varepsilon) - R(\hat{y}, \varepsilon)| \\
&\qquad \le \max_i |P_{n+}(S_i + B_\varepsilon) - P_+(S_i + B_\varepsilon)| \le j\delta.
\end{aligned}$$

as was to be shown. ∎

IX. Conclusion

In this paper, we studied the tradeoffs inherent to robust classification in the fundamental setting of two- and three-class Gaussian classification models. In particular, we leveraged that half-spaces are extremal sets with respect to Gaussian isoperimetry to derive ℓ_2 and ℓ_∞ optimal robust classifiers in the imbalanced data setting. This analysis revealed a fundamental tradeoff between accuracy and robustness, which depends on the level of class imbalance in the data. Indeed, we showed that in this setting, no classifier minimizes both the standard and robust risks simultaneously. Furthermore, we analyzed the optimization landscape of the robust risk, demonstrating that the optimizers of various convex surrogate losses coincide with the nonconvex robust 0-1 loss. Finally, we connected our results to empirical robust risk minimization by providing a finite-sample analysis with respect to the 0-1 and surrogate loss functionals.

APPENDIX A
EXTENSIONS OF THEOREM IV.1

A. Connections to randomized classifiers

Adding random noise has been used as a heuristic to obtain robust classifiers (see, e.g., [65], [66]). While it has been shown to be attackable via gradient based methods [74], we can still study it as a heuristic. It turns out that it has connections to optimal robust classifiers in our models.

In this section, we suppose that the noise level in the data is σ^2, so $x_i|y_i \sim \mathcal{N}(y\mu, \sigma^2 I_p)$. Suppose we add noise $Z \sim \mathcal{N}(0, \tau^2 I_p)$, for some $\tau^2 > 0$, and then train a standard classifier. Note that the Bayes-optimal classifier depends only on the SNR $s(\mu, \sigma^2) = \|\mu\|_2/\sigma$. Thus we get that the Bayes-optimal classifier with noise is the optimal ε-robust classifier if (assuming $\varepsilon < \|\mu\|_2$)

$$s(\|\mu\|_2, \sigma^2 + \tau^2) = s(\|\mu\|_2 - \varepsilon, \sigma^2),$$

or equivalently if

$$\tau = \sigma \sqrt{\frac{\|\mu\|_2^2}{(\|\mu\|_2 - \varepsilon)^2} - 1}.$$

Put it another way, our results show that robust classifiers reduce the signal strength. Equivalently, randomized classifiers increase the noise level. However, note that for this the added noise level has to be tuned very carefully.

B. Extension to weighted combinations

Given a distribution Q over ε, we can try to minimize $R(\hat{y}, Q) = \mathbb{E}_{\varepsilon \sim Q} R(\hat{y}, \varepsilon)$. This leads to classifiers that can achieve various tradeoffs between robustness to different sizes of perturbations. For instance, we can minimize $R(\hat{y}, 0) + \lambda \cdot R(\hat{y}, \varepsilon)$ for some $\lambda > 0$.

It is readily verified that Lemma IV.2 still holds for $R(\hat{y}, Q)$, as long as Q is supported on $[0, \|\mu\|)$. This is because the linear classifier $\hat{y}$ found in the proof of that result does not depend on ε, and reduces the ε robust risk for all $\varepsilon < \|\mu\|$. Hence, linear classifiers are admissible for $R(\hat{y}, Q)$.

However, in general there is no analytical expression for the optimal linear classifier. Following Theorem IV.1, it is readily verified that the threshold c in the optimal linear classifier is the unique solution of the equation

$$\mathbb{E}_{\mu'} \exp(-\mu'^2/2)[\pi \exp(c\mu') + (1-\pi)\exp(-c\mu')] = 0,$$

where $\mu' = \mu - \varepsilon$ and $\varepsilon \sim Q$.

As before, it is enough to solve the 1-D problem. Thus, we want to find the value of the threshold c that minimizes

$$\begin{aligned} R(\hat{y}_c, Q) &= P(y=1)\mathbb{E}_{\varepsilon\sim Q} P_{\mu-\varepsilon}(x \le c) \\ &\quad + P(y=-1)\mathbb{E}_{\varepsilon\sim Q} P_{-\mu+\varepsilon}(x \ge c) \\ &= \pi \cdot \mathbb{E}_{\varepsilon\sim Q} P_{\mu-\varepsilon}(x \le c) \\ &\quad + (1-\pi)\cdot \mathbb{E}_{\varepsilon\sim Q} P_{-\mu+\varepsilon}(x \ge c) \\ &= \pi \cdot \mathbb{E}_{\mu'} P_{\mu'}(x \le c) + (1-\pi)\cdot \mathbb{E}_{\mu'} P_{-\mu'}(x \ge c). \end{aligned}$$

Differentiating with respect to c, we find that

$$\begin{aligned} R'(c) &:= dR(\hat{y}_c, Q)/dc \\ &= \pi \cdot \mathbb{E}_{\mu'}\phi(c-\mu') - (1-\pi)\cdot \mathbb{E}_{\mu'}\phi(c+\mu') \\ &= (2\pi)^{-1/2}\Big[\pi \cdot \mathbb{E}_{\mu'} \exp[-(c-\mu')^2/2] \\ &\qquad - (1-\pi)\cdot \mathbb{E}_{\mu'} \exp[-(c+\mu')^2/2]\Big]. \end{aligned}$$

Up to the factor $(2\pi)^{-1/2}$, and also factoring out the term $\exp[-c^2/2]$, which cannot be zero, we find that $R'(c) = 0$ iff

$$\begin{aligned} a(c) &= \mathbb{E}_{\mu'} \exp[-\mu'^2/2][\pi \cdot \exp(c\mu') - (1-\pi)\cdot \exp(-c\mu')] \\ &= 0. \end{aligned}$$

This is exactly the claimed equation for c. Now, it is not hard to see that $a(c)$ is strictly increasing with limits $\pm\infty$ at $\pm\infty$. Hence, the solution c exists and is unique.

C. Data with a general covariance

A natural question is whether optimality extends to data with general covariance. To this end, suppose that the data is distributed according to the two-class Gaussian model with an invertible covariance matrix Σ so that $x_i \sim \mathcal{N}(y_i\mu, \Sigma)$. In this setting, we can study the setting when the difference between the population means aligns with the smallest eigenvectors of Σ.

Theorem A.1 (Optimal robust classifiers, general covariance)**.** *Consider finding ℓ_2 robust classifiers in the two-class Gaussian classification problem with data (x_i, y_i), $i = 1, \ldots, n$, where $y_i = \pm 1$, $x_i \sim \mathcal{N}(y_i\mu, \Sigma)$, where Σ is an invertible covariance matrix. Let V be the span of eigenvectors of Σ corresponding to its smallest eigenvalue, and note that this is a nonempty linear space. Suppose that $\mu \in V$. The optimal ℓ_2 robust classifiers are linear classifiers*

$$\hat{y}^*(x) = \operatorname{sign}\left(x^\top \mu \left[1 - \frac{\varepsilon}{\|\mu\|}\right] - \lambda^{1/2}\frac{q}{2}\right),$$

where λ is the smallest eigenvalue of Σ, and the other symbols are as in Theorem IV.1.

This theorem generalizes the result of Theorem IV.1. When the covariance matrix $\Sigma = \sigma^2 I_p$ is diagonal, the optimal robust ℓ_2 classifier from Theorem A.1 is identical to that from Theorem IV.1.

Proof of Theorem A.1: The proof proceeds along the lines of Theorem IV.1, checking that it extends to this setting. We will only sketch the key steps.

We define the ε-expansion of a set A in a norm $\|\cdot\|$ to be the Minkowski sum $A + B_\varepsilon = \{a + b : a \in A, b \in B_\varepsilon\}$, where $B_\varepsilon = \{x : \|x\| \le \varepsilon\}$ is the ε-ball in the given norm.

The key insight is that ε-expansions in ℓ_2 norm will turn into ε-expansions in the Mahalanobis metric $d_\Sigma(a, b) = [(a-b)^\top \Sigma (a-b)]^{1/2}$. To put it another way, by changing coordinates from $x \to \Sigma^{-1/2}x$, the ℓ_2 ball transforms to a Mahalanobis ball, i.e., an ellipsoid. We explain this below.

The first critical step was the existence of optimal linear classifiers (Lemma IV.2). For any fixed set S, the key is to be able to solve the problem (10). This requires us to find the

optimal isoperimetric set, i.e., the one with minimal probability under ε-expansion, with respect to the probability measure $\mathcal{N}(\mu, \Sigma)$.

Now, letting $z - \mu = \Sigma^{-1/2}(x - \mu)$, we can write

$$P_{x\sim\mathcal{N}(\mu,\Sigma)}(x \in S) = P_{z\sim\mathcal{N}(\mu,I)}(z \in \mu + \Sigma^{-1/2}(S - \mu)),$$

and

$$\begin{aligned} &P_{x\sim\mathcal{N}(\mu,\Sigma)}(x \in S + B_\varepsilon) \\ &\quad = P_{z\sim\mathcal{N}(\mu,I)}(z \in \mu + \Sigma^{-1/2}(S + B_\varepsilon - \mu)). \end{aligned}$$

Let $S' = \mu + \Sigma^{-1/2}(S - \mu)$. We can write

$$\mu + \Sigma^{-1/2}(S + B_\varepsilon - \mu) = S' + \Sigma^{-1/2}B_\varepsilon.$$

Now, $\Sigma^{-1/2}B_\varepsilon$ is precisely the ε-ball in the Mahalanobis metric, $\Sigma^{-1/2}B_\varepsilon = \{\Sigma^{-1/2}x : \|x\|_2 \le \varepsilon\} = \{z : \|z\|_\Sigma \le \varepsilon\}$. Let us call this set $B_{\Sigma,\varepsilon}$.

So problem (10) can equivalently be written as

$$\min_{S'} P_{z\sim\mathcal{N}(\mu,I)}(S' + B_{\Sigma,\varepsilon}) \ s.t. \ P_{z\sim\mathcal{N}(\mu,I)}(S') = \alpha.$$

Consider any set S' of the form $v^\top z \le c$, with $\|v\|_2 = 1$ (i.e., a hyperplane). Then (as can be readily seen by drawing a picture)

$$S' + B_{\Sigma,\varepsilon} = \{z + z' : v^\top z \le c, \|\Sigma^{-1/2}z'\|_2 \le \varepsilon\},$$

equals the set

$$S'' = \{v^\top z \le c + \varepsilon \cdot \|\Sigma^{1/2}v\|_2\}.$$

For fixed c, ε, Σ, the size of this set (with respect to any probability measure absolutely continuous with respect to Lebesgue measure) can be minimized in v by taking v to lie in the span of the eigenvectors of Σ with smallest eigenvalue.

Thus the sets S' minimizing the expansion are hyperplanes orthogonal to the eigenvectors with *smallest* eigenvalues of Σ. Then $S = \Sigma^{1/2} \cdot \{z : v^\top z \le c\} = \{x : v^\top \Sigma^{-1/2} x \le c'\}$. Now, since v is an eigenvector of Σ, i.e., $\Sigma v = \lambda v$, we have that $v' = \Sigma^{-1/2}v = \lambda^{-1/2}v$ is still a scaled version of v. Hence, even in the original coordinate system, the sets S' have the same interpretation.

Next, the second critical step in the proof was to reduce the problem to a one-dimensional classification along the direction of μ. This is the case if the eigenvectors align with μ. In that case, after the one-dimensional projection, we have the same problem that we already solved in Theorem IV.1. One can easily verify that the remaining steps go through. This finishes the proof. ∎

D. Data on a low-dimensional subspace

We can extend the above analysis to low-dimensional data. Suppose that the data x_i, y_i live in a lower dimensional linear space. For simplicity, suppose that $x_i = (x_i^1, 0_d)$, so only the first p' coordinates are nonzero, and the remaining $d := p - p'$ dimensions are zero. This is a model of a low-dimensional manifold. For rotationally invariant problems like ℓ_2 norm robustness, we can consider instead any low-dimensional affine space, and the same conclusions apply. However, for non-rotationally invariant problems like ℓ_∞ norm robustness (studied in detail later), the conclusions only apply to this specific space.

Intuitively, decision boundaries that are not perpendicular to the manifold $M = (x, 0_d)$ can have a larger "expansion" projected down into the manifold. Hence, their adversarial risk can be larger. This will imply that we can restrict to decision boundaries perpendicular to the manifold, and thus reduce the problem to the previous case. To this end, we provide the following lemma. We emphasize that this is a purely geometric fact, and holds for any classification problem (not just Gaussian), and any norm (not just ℓ_2.)

Lemma A.2 (Low-dimensional classifiers are admissible)**.** *Consider any classification problem and robust classifiers for the low-dimensional data model above. For any classifier $\hat{y}$, the low-dimensional classifier $\hat{y}^*(x_i^1, x_i^2) = \hat{y}(x_i^1, 0_d)$ has robust risk is less than or equal to that of the original classifier with respect to any norm $\|\cdot\|$:*

$$R(\hat{y}, \varepsilon) \ge R(\hat{y}^*, \varepsilon).$$

The above claim shows that for low-dimensional data as above, even if we have the data represented as full-length vectors, we can restrict to classifiers that depend only on the first coordinates. This reduces the problem to the one considered before, and all the results derived above are applicable. In particular, for a low-dimensional two-class Gaussian mixture, low-dimensional linear classifiers are optimal, under the previous conditions.

Proof of Lemma A.2: Suppose S_1 is the decision region $x : \hat{y}(x) = 1$ where the original classifier outputs the first class. The modified classifier $\hat{y}^*$ makes the same decision as $\hat{y}$ restricted to the first p' coordinates.

Then the decision region S_1^* where the modified classifier outputs the first class is the set of vectors $x = (x^1, x^2)$ such that $(x^1, 0) \in S_1 \cap M$. We can write S_1^* as the direct product $S_1^* = S_1^{*,p'} \times \mathbb{R}^d$.

Then, the ε-expansion of S_1^* within M is $S_1^{*,p'} + B_\varepsilon^{p'}$, where $B_\varepsilon^{p'}$ is a p'-dimensional ball, and we can compute the sum in p'-dimensional space. Then, it is readily verified that, by denoting R_d the restriction to the first p' coordinates of a subset of M (i.e., ignoring the last d zero coordinates),

$$S_1^{*,p'} + B_\varepsilon^{p'} \subset R_d[(S_1 + B_\varepsilon^p) \cap M],$$

or equivalently, viewing this as embedded in the p-dimensional space,

$$[S_1 \cap M] + (B_\varepsilon^{p'}, 0_d) \subset (S_1 + B_\varepsilon^p) \cap M.$$

Indeed, if $z \in [S_1 \cap M] + (B_\varepsilon^{p'}, 0)$, then $z = x + \delta$, where $x \in S_1 \cap M$ and $\delta \in (B_\varepsilon^{p'}, 0)$. Then it is clear that $z \in (S_1 + B_\varepsilon^p) \cap M$. Here we only use that $B_\varepsilon^{p'}$ is the restriction of the p-dimensional ε-ball B_ε^p onto the first p' coordinates. This shows that the ε-expansion of S_1 is contained within the ε-expansion of S_1^*. The same reasoning applies to S_{-1}.

This shows that the classifier $\hat{y}^*$ has robust risk at most as large as that of the original classifier $\hat{y}$. This finishes the proof. ∎

APPENDIX B
POINTWISE CALCULATION OF A BAYES OPTIMAL CLASSIFIER

Here we derive a Bayes optimal classifier, i.e., a classifier that minimizes the standard non-robust risk R_{std} for the three-class setting (15) of Section V. Recall that Bayes optimal classification is achieved by maximizing the posterior probability pointwise:

$$\begin{aligned}\hat{y}^*_{\mathrm{Bay}}(x) &\in \operatorname*{argmax}_{c\in\mathcal{C}} \Pr_{y|x}(y=c)\\ &= \operatorname*{argmax}_{c\in\mathcal{C}} \frac{p_{x|y=c}(x)\Pr(y=c)}{p(x)}\\ &= \operatorname*{argmax}_{c\in\mathcal{C}} \Pr(y=c)p_{x|y=c}(x).\end{aligned}$$

Hence it remains to identify the associated classification regions

$$S_+ \coloneqq \left\{ x\in\mathbb{R}^p : \begin{array}{l}\Pr(y=+1)p_{x|y=+1}(x)\\ \quad\geq \max\left\{\begin{array}{l}\Pr(y=0)p_{x|y=0}(x),\\ \Pr(y=-1)p_{x|y=-1}(x)\end{array}\right\}\end{array}\right\},$$

$$S_- \coloneqq \left\{ x\in\mathbb{R}^p : \begin{array}{l}\Pr(y=-1)p_{x|y=-1}(x)\\ \quad\geq \max\left\{\begin{array}{l}\Pr(y=0)p_{x|y=0}(x),\\ \Pr(y=+1)p_{x|y=+1}(x)\end{array}\right\}\end{array}\right\},$$

where the complementary region $(S_+\cup S_-)^c$ will be classified as the remaining zero class.

Starting with S_+, note that

$$\begin{aligned}&\Pr(y=+1)p_{x|y=+1}(x) \geq \Pr(y=0)p_{x|y=0}(x)\\ &\iff \pi_+\exp\left[-\frac{\|x-\lambda_+\mu\|_2^2}{2}\right] \geq \pi_0\exp\left[-\frac{\|x\|_2^2}{2}\right]\\ &\iff \|x-\lambda_+\mu\|_2^2 - \|x\|_2^2 \leq 2\ln(\pi_+/\pi_0)\\ &\iff -2\lambda_+x^\top\mu + \lambda_+^2 \leq 2\ln(\pi_+/\pi_0)\\ &\iff x^\top\mu \geq \lambda_+/2 - \ln(\pi_+/\pi_0)/|\lambda_+|,\end{aligned}$$

and similarly

$$\begin{aligned}&\Pr(y=+1)p_{x|y=+1}(x) \geq \Pr(y=-1)p_{x|y=-1}(x)\\ &\iff \pi_+\exp\left[-\frac{\|x-\lambda_+\mu\|_2^2}{2}\right] \geq \pi_-\exp\left[-\frac{\|x-\lambda_-\mu\|_2^2}{2}\right]\\ &\iff \|x-\lambda_+\mu\|_2^2 - \|x-\lambda_-\mu\|_2^2 \leq 2\ln(\pi_+/\pi_-)\\ &\iff -2(\lambda_+-\lambda_-)x^\top\mu + \lambda_+^2 - \lambda_-^2 \leq 2\ln(\pi_+/\pi_-)\\ &\iff x^\top\mu \geq (\lambda_++\lambda_-)/2 - \ln(\pi_+/\pi_-)/|\lambda_+-\lambda_-|.\end{aligned}$$

Therefore, $S_+ = \{x\in\mathbb{R}^p : x^\top\mu \geq c_+\}$ where

$$c_+ \coloneqq \max\left\{\frac{\lambda_+}{2} + \frac{\ln(\pi_0/\pi_+)}{|\lambda_+|}, \frac{\lambda_++\lambda_-}{2} - \frac{\ln(\pi_+/\pi_-)}{|\lambda_+-\lambda_-|}\right\}.$$

In the same way, we obtain $S_- = \{x\in\mathbb{R}^p : x^\top\mu \leq c_-\}$ where

$$c_- \coloneqq \min\left\{\frac{\lambda_-}{2} - \frac{\ln(\pi_0/\pi_-)}{|\lambda_-|}, \frac{\lambda_++\lambda_-}{2} - \frac{\ln(\pi_+/\pi_-)}{|\lambda_+-\lambda_-|}\right\}.$$

Putting these together yields

$$\hat{y}^*_{\mathrm{Bay}}(x) = \begin{cases}+1 & \text{if } x\in S_+,\\ -1 & \text{if } x\in S_-,\\ 0 & \text{otherwise,}\end{cases} = \begin{cases}+1 & \text{if } x^\top\mu \geq c_+,\\ 0 & \text{if } c_- < x^\top\mu < c_+,\\ -1 & \text{if } x^\top\mu \leq c_-,\end{cases}$$

which is a linear interval classifier (16), as illustrated in Figure 3a. At the boundaries between regions, the posterior probabilities of the two corresponding classes are equal. Indeed, assigning the boundary to either class yields a Bayes optimal classifier, since the boundaries have zero probability under the marginal distribution of x.

APPENDIX C
THREE-CLASS LINEAR INTERVAL CLASSIFICATION FOR $\varepsilon \geq \min\{|\lambda_+|, |\lambda_-|\}/2$

Here we show that three-class linear interval classification reduces down to several two-class problems when $\varepsilon \geq \min\{|\lambda_+|, |\lambda_-|\}/2$. Specifically, we show the following proposition.

Proposition C.1. *Suppose $\varepsilon \geq \min\{|\lambda_+|, |\lambda_-|\}/2$. Then for any $w\neq 0$ and $c_+ \geq c_-$,*

$$\begin{aligned}&R_{\mathrm{rob}}\{\hat{y}_{\mathrm{int}}(\cdot; w, c_+, c_-), \varepsilon\}\\ &\quad\geq \min\Big\{R_{\mathrm{rob}}(\hat{y}^*_{-1,+1}, \varepsilon), R_{\mathrm{rob}}(\hat{y}^*_{0,+1}, \varepsilon), R_{\mathrm{rob}}(\hat{y}^*_{0,-1}, \varepsilon)\Big\},\end{aligned}$$

where

$$\hat{y}^*_{-1,+1}(x) \coloneqq \begin{cases}+1, & \textit{if } [x-\mu(\lambda_++\lambda_-)/2]^\top\\ & \qquad [\mu(\lambda_+-\lambda_-)/2]\\ & \qquad (1-2\varepsilon/|\lambda_+-\lambda_-|)_+\\ & \qquad -\ln(\pi_-/\pi_+)/2\\ & \qquad \geq 0,\\ -1, & \textit{otherwise},\end{cases} \tag{35}$$

$$\hat{y}^*_{0,+1}(x) \coloneqq \begin{cases}+1, & \textit{if } [x-\mu\lambda_+/2]^\top[\mu\lambda_+/2]\\ & \qquad (1-2\varepsilon/|\lambda_+|)_+\\ & \qquad -\ln(\pi_0/\pi_+)/2\\ & \qquad \geq 0,\\ 0, & \textit{otherwise},\end{cases} \tag{36}$$

$$\hat{y}^*_{0,-1}(x) \coloneqq \begin{cases}-1, & \textit{if } [x-\mu\lambda_-/2]^\top[\mu\lambda_-/2]\\ & \qquad (1-2\varepsilon/|\lambda_-|)_+\\ & \qquad -\ln(\pi_0/\pi_-)/2\\ & \qquad \geq 0,\\ 0, & \textit{otherwise},\end{cases} \tag{37}$$

are the optimal robust two-class classifiers obtained by applying (7) *from Theorem IV.1 to each pair of classes.*

Proof of Proposition C.1: Let $\varepsilon \geq \min\{|\lambda_+|, |\lambda_-|\}/2$ be arbitrary. Considering the negative and positive classes alone (i.e., ignoring the zero class) and applying (7) yields the optimal robust two-class classifier (35). Similarly, considering the zero and positive classes alone and applying (7) to $x-\mu\lambda_+/2$ (for which the two classes have means $\pm\mu\lambda_+/2$) yields the optimal robust two-class classifier (36). Likewise with $x-\mu\lambda_-/2$ for the optimal robust two-class classifier (37).

Now, let $w\neq 0$ and $c_+ \geq c_-$ be arbitrary. Recall first that $\hat{y}_{\mathrm{int}}(x; w, c_+, c_-) = \hat{y}_{\mathrm{int}}(x; \tilde{w}, \tilde{c}_+, \tilde{c}_-)$, where

$$\tilde{w} = w/\|w\|_2, \qquad \tilde{c}_+ = c_+/\|w\|_2, \qquad \tilde{c}_- = c_-/\|w\|_2,$$

and so

$$R_{\rm rob}\{\hat{y}_{\rm int}(\cdot; w, c_+, c_-), \varepsilon\} = R_{\rm rob}\{\hat{y}_{\rm int}(\cdot; \tilde{w}, \tilde{c}_+, \tilde{c}_-), \varepsilon\}.$$

Next, recall from Section V-B1 that for any $\tilde{c}_+ \geq \tilde{c}_-$, the robust risk is optimized subject to $\|\tilde{w}\|_2 = 1$ by $\tilde{w} = \mu$ so we have

$$\begin{aligned}
&R_{\rm rob}\{\hat{y}_{\rm int}(\cdot; \tilde{w}, \tilde{c}_+, \tilde{c}_-), \varepsilon\} \\
&\quad \geq R_{\rm rob}\{\hat{y}_{\rm int}(\cdot; \mu, \tilde{c}_+, \tilde{c}_-), \varepsilon\} \\
&\quad = \pi_- \Pr_-(\tilde{x} > \tilde{c}_- - \varepsilon) + \pi_+ \Pr_+(\tilde{x} < \tilde{c}_+ + \varepsilon) \\
&\qquad + \pi_0 \Pr_0(\tilde{x} \leq \tilde{c}_- + \varepsilon \text{ or } \tilde{x} \geq \tilde{c}_+ - \varepsilon),
\end{aligned}$$

where the shorthand notations

$$\Pr_- \coloneqq \Pr_{\tilde{x} \sim \mathcal{N}(\lambda_-, 1)}, \quad \Pr_+ \coloneqq \Pr_{\tilde{x} \sim \mathcal{N}(\lambda_+, 1)}, \quad \Pr_0 \coloneqq \Pr_{\tilde{x} \sim \mathcal{N}(0, 1)},$$

are taken from Section V-B2. It now remains to bound the robust risk $R_{\rm rob}\{\hat{y}_{\rm int}(\cdot; \mu, \tilde{c}_+, \tilde{c}_-), \varepsilon\}$.

To begin, consider the case where $\tilde{c}_- \leq \tilde{c}_+ \leq \tilde{c}_- + 2\varepsilon$. In this case, note that $\Pr_0(\tilde{x} \leq \tilde{c}_- + \varepsilon \text{ or } \tilde{x} \geq \tilde{c}_+ - \varepsilon) = 1$, so the robust risk can be rewritten and bounded as

$$\begin{aligned}
&R_{\rm rob}\{\hat{y}_{\rm int}(\cdot; \mu, \tilde{c}_+, \tilde{c}_-), \varepsilon\} \\
&\quad = \pi_- \Pr_-(\tilde{x} > \tilde{c}_- - \varepsilon) + \pi_+ \Pr_+(\tilde{x} < \tilde{c}_+ + \varepsilon) + \pi_0 \\
&\quad \geq \pi_- \Pr_-(\tilde{x} > \tilde{c}_+ - \varepsilon) + \pi_+ \Pr_+(\tilde{x} < \tilde{c}_+ + \varepsilon) + \pi_0.
\end{aligned}$$

The final expression is the robust risk of a linear two-class classifier that assigns points to only the positive and negative classes with threshold $\tilde{c}_+$, so is no better than that of $\hat{y}^*_{-1,+1}$. As a result, we have that

$$\begin{aligned}
&R_{\rm rob}\{\hat{y}_{\rm int}(\cdot; \tilde{w}, \tilde{c}_+, \tilde{c}_-), \varepsilon\} \\
&\quad \geq \pi_- \Pr_-(\tilde{x} > \tilde{c}_+ - \varepsilon) + \pi_+ \Pr_+(\tilde{x} < \tilde{c}_+ + \varepsilon) + \pi_0 \\
&\quad \geq R_{\rm rob}(\hat{y}^*_{-1,+1}, \varepsilon),
\end{aligned}$$

which completes the proof in this case.

Next, consider the case where $\tilde{c}_+ \geq \tilde{c}_- + 2\varepsilon$, in which case we can rewrite the robust risk as

$$\begin{aligned}
&R_{\rm rob}\{\hat{y}_{\rm int}(\cdot; \mu, \tilde{c}_+, \tilde{c}_-), \varepsilon\} \\
&\quad = 1 - \Big\{\pi_- \Pr_-(\tilde{x} \leq \tilde{c}_- - \varepsilon) + \pi_+ \Pr_+(\tilde{x} \geq \tilde{c}_+ + \varepsilon) \\
&\qquad\qquad + \pi_0 \Pr_0(\tilde{c}_- + \varepsilon < \tilde{x} < \tilde{c}_+ - \varepsilon)\Big\}.
\end{aligned}$$

Now note that $\varepsilon \geq |\lambda_+|/2$ or $\varepsilon \geq |\lambda_-|/2$ since $\varepsilon \geq \min\{|\lambda_+|, |\lambda_-|\}/2$. Suppose first that $\varepsilon \geq |\lambda_+|/2$, and note that if $\pi_+ \leq \pi_0$ then

$$\begin{aligned}
&\pi_+ \Pr_+(\tilde{x} \geq \tilde{c}_+ + \varepsilon) + \pi_0 \Pr_0(\tilde{c}_- + \varepsilon < \tilde{x} < \tilde{c}_+ - \varepsilon) \\
&\quad \leq \pi_0 \Big\{ \Pr_+(\tilde{x} \geq \tilde{c}_+ + \varepsilon) + \Pr_0(\tilde{c}_- + \varepsilon < \tilde{x} < \tilde{c}_+ - \varepsilon) \Big\} \\
&\quad = \pi_0 \Big\{ \Pr_0(\tilde{x} \geq \tilde{c}_+ - (\lambda_+ - \varepsilon)) \\
&\qquad\qquad + \Pr_0(\tilde{c}_- + \varepsilon < \tilde{x} < \tilde{c}_+ - \varepsilon) \Big\} \\
&\quad \leq \pi_0 \Pr_0(\tilde{x} > \tilde{c}_- + \varepsilon),
\end{aligned}$$

where the final inequality used the fact that $\tilde{c}_+ - (\lambda_+ - \varepsilon) \geq \tilde{c}_+ - \varepsilon$ when $\varepsilon \geq |\lambda_+|/2$. Consequently,

$$\begin{aligned}
&R_{\rm rob}\{\hat{y}_{\rm int}(\cdot; \mu, \tilde{c}_+, \tilde{c}_-), \varepsilon\} \\
&\quad \geq \pi_+ + \pi_- \Pr_-(\tilde{x} > \tilde{c}_- - \varepsilon) + \pi_0 \Pr_0(\tilde{x} \leq \tilde{c}_- + \varepsilon) \\
&\quad \geq R_{\rm rob}(\hat{y}^*_{0,-1}, \varepsilon),
\end{aligned}$$

where the final inequality holds because the middle term is the robust risk of a linear two-class classifier with threshold $\tilde{c}_-$ that assigns points to only the negative and zero classes, so is no better than that of $\hat{y}^*_{0,-1}$. Likewise, if $\pi_+ > \pi_0$ then $\pi_+ \Pr_+(\tilde{x} \geq \tilde{c}_+ + \varepsilon) + \pi_0 \Pr_0(\tilde{c}_- + \varepsilon < \tilde{x} < \tilde{c}_+ - \varepsilon) \leq \pi_+ \Pr_+(\tilde{x} > \tilde{c}_- + \varepsilon)$ and

$$\begin{aligned}
&R_{\rm rob}\{\hat{y}_{\rm int}(\cdot; \mu, \tilde{c}_+, \tilde{c}_-), \varepsilon\} \\
&\quad \geq \pi_0 + \pi_- \Pr_-(\tilde{x} > \tilde{c}_- - \varepsilon) + \pi_+ \Pr_+(\tilde{x} \leq \tilde{c}_- + \varepsilon) \\
&\quad \geq R_{\rm rob}(\hat{y}^*_{-1,+1}, \varepsilon).
\end{aligned}$$

As a result, whether $\pi_+ \leq \pi_0$ or $\pi_+ > \pi_0$,

$$\begin{aligned}
&R_{\rm rob}\{\hat{y}_{\rm int}(\cdot; \tilde{w}, \tilde{c}_+, \tilde{c}_-), \varepsilon\} \\
&\quad \geq \min\Big\{R_{\rm rob}(\hat{y}^*_{0,-1}, \varepsilon), R_{\rm rob}(\hat{y}^*_{-1,+1}, \varepsilon)\Big\},
\end{aligned}$$

completing the proof when $\varepsilon \geq |\lambda_+|/2$.

Repeating the analogous argument for $\varepsilon \geq |\lambda_-|/2$ yields that in that case

$$\begin{aligned}
&R_{\rm rob}\{\hat{y}_{\rm int}(\cdot; \tilde{w}, \tilde{c}_+, \tilde{c}_-), \varepsilon\} \\
&\quad \geq \min\Big\{R_{\rm rob}(\hat{y}^*_{0,+1}, \varepsilon), R_{\rm rob}(\hat{y}^*_{-1,+1}, \varepsilon)\Big\},
\end{aligned}$$

completing the proof. ∎

Appendix D
Beyond Gaussians that lie on a line

For the one-dimensional setting $p = 1$, the three Gaussian are always as in (15), i.e., their means lie along a line. However, when $p > 1$, the means can be in more general locations and one naturally wonders what the optimal robust classifier would be in such settings.

When the means do not lie on a line, the Bayes optimal classifier is no longer linear in general and so the optimal robust classifier is likely no longer linear as well. Moreover, as shown above, the optimal robust classifier does not coincide with the Bayes optimal classifier in general, making it non-trivial to even conjecture what the optimal robust classifier is in such cases. However, we do expect them to coincide when the location of the means is symmetric, as described by the following conjecture.

Conjecture D.1 (Optimal robust classifier for means arranged in a triangle)**.** *Consider three balanced Gaussian classes $\mathcal{C} = \{0, 1, 2\}$ with means at the corners of an equilateral triangle:*

$$x|y \sim \mathcal{N}\Big(\mu \cdot [\cos(y \cdot 2\pi/3)e_1 + \sin(y \cdot 2\pi/3)e_2], I_p\Big),$$

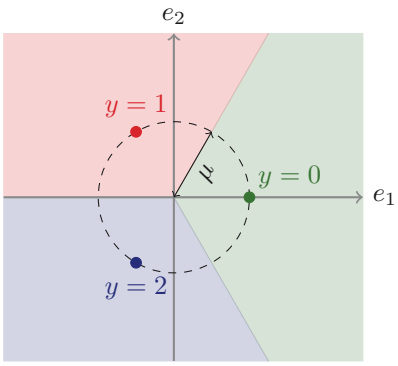
$$y = \begin{cases} 0 & \textit{with probability } 1/3, \\ 1 & \textit{with probability } 1/3, \\ 2 & \textit{with probability } 1/3, \end{cases}$$

where $\mu \in \mathbb{R}_{>0}$ defines the size of the triangle, and $e_1, e_2 \in \{0,1\}^p$ are the canonical basis vectors.

We conjecture that the optimal robust classifier is the same as the Bayes optimal classifier, i.e.,

$$\hat{y}^*(x) \coloneqq \operatorname*{argmax}_y \|x - \mu \cdot [\cos(y \cdot 2\pi/3)e_1 + \sin(y \cdot 2\pi/3)e_2]\|_2,$$

as shown in the following diagram:



As before, we have taken the within-class variance to be unity without loss of generality. Furthermore, we have centered the triangle at the origin without loss of generality.

This conjecture is natural, but turns out to be nontrivial to rigorously establish. In particular, the Gaussian concentration of measure approach used to prove optimality in Theorems IV.1 and V.4 does not directly apply here since the conjectured optimal classifier is not composed of half-spaces. It would appear that the method used in [40] may be used here since the Bayes classifier for perturbed means $\mu' = \mu - \varepsilon/\sin(\pi/3)$ has standard risk with respect to the perturbed means matching the robust risk with respect to the unperturbed means. However, the $\varepsilon/\sin(\pi/3)$ size of the perturbation is too large and cannot be used.

Going beyond three-classes to settings with even more classes, we conjecture that the same holds in corresponding symmetric settings, as described by the following conjecture.

Conjecture D.2 (Optimal robust classifier for n means arranged in a regular polygon)**.** *Consider n balanced Gaussian classes $\mathcal{C} = \{0, 1, \ldots, n-1\}$ with means at the corners of a regular n-sided polygon:*

$$x|y \sim \mathcal{N}\Big(\mu \cdot [\cos(y \cdot 2\pi/n)e_1 + \sin(y \cdot 2\pi/n)e_2], I_p\Big),$$
$$y = \begin{cases} 0 & \textit{with probability } 1/n, \\ \vdots & \vdots \\ n-1 & \textit{with probability } 1/n, \end{cases}$$

where $\mu \in \mathbb{R}_{>0}$ defines the size of the polygon, and $e_1, e_2 \in \{0,1\}^p$ are the canonical basis vectors.

We conjecture that the optimal robust classifier is the same as the Bayes optimal classifier, i.e.,

$$\hat{y}^*(x) \coloneqq \operatorname*{argmax}_y \|x - \mu \cdot [\cos(y \cdot 2\pi/n)e_1 + \sin(y \cdot 2\pi/n)e_2]\|_2.$$

As with Conjecture D.1, this conjecture is natural but highly nontrivial to rigorously establish. Once again, the Gaussian concentration of measure approach used to prove optimality in Theorems IV.1 and V.4 does not directly apply since the conjectured optimal classifier is not composed of half-spaces. Likewise, the method used in [40] does not seem to apply since the Bayes classifier for perturbed means $\mu' = \mu - \varepsilon/\sin(\pi/n)$ has standard risk with respect to the perturbed means matching the robust risk with respect to the unperturbed means. However, the $\varepsilon/\sin(\pi/n)$ size of the perturbation is again too large and cannot be used.

Developing new theoretical approaches that can prove Conjectures D.1 and D.2 is an exciting direction for future work.

Appendix E
Proofs for optimal ℓ_∞ robust classifiers

A. Proof of Corollary VI.1

This follows because ℓ_∞ norm is upper bounded by the ℓ_2 norm. Thus for any fixed ε, the ℓ_∞ robust risk is upper bounded by the ℓ_2 robust risk:

$$R(\hat{y}, \varepsilon, \|\cdot\|_\infty) \geq R(\hat{y}, \varepsilon, \|\cdot\|_2),$$
$$R^*(\varepsilon, \|\cdot\|_\infty) \geq R^*(\varepsilon, \|\cdot\|_2).$$

From Theorem IV.1, we know the optimal ℓ_2 robust classifiers, i.e., the ones minimizing the upper bound, are based on $x_j \cdot [\mu_j - \varepsilon]$. Now, it follows that for the decision sets S_i of these classifiers (axis aligned half-planes), $S_i + B_{2,\varepsilon} = S_i + B_{\infty,\varepsilon}$, where $B_{q,\varepsilon}$ denotes the ε-ball in the ℓ_q norm. Thus, the ℓ_∞ robust risk is *equal* to the ℓ_2 risk for this specific classifier. This implies that it also minimizes the ℓ_∞ risk, and that the two risks are the same:

$$R^*(\varepsilon, \|\cdot\|_\infty) = R^*(\varepsilon, \|\cdot\|_2).$$

This finishes the proof. ■

B. Proof of Corollary VI.2

As before, $R_{\text{rob}}(\hat{y}, \varepsilon, \|\cdot\|_\infty) \geq R_{\text{rob}}(\hat{y}, \varepsilon, \|\cdot\|_2)$ for any classifier $\hat{y}$ and radius ε. Now, by Theorem V.1 the weights $w^* \coloneqq \mu/\|\mu\|_2$ optimize $R_{\text{rob}}\{\hat{y}_{\text{int}}(x; w^*, c_+^*, c_-^*), \varepsilon, \|\cdot\|_2\}$ where the formulae for the two cases of c_+^* and c_-^*, as well as the cutoff α^*, are simplified by noting that $\|\mu\|_2 = \mu_j$. Moreover,

$$\begin{aligned}\hat{y}^*_{\text{int}}(x) &\coloneqq \hat{y}_{\text{int}}(x; w^*, c_+^*, c_-^*) \\ &= \hat{y}_{\text{int}}(x^\top w^*; 1, c_+^*, c_-^*) = \hat{y}_{\text{int}}(x_j; 1, c_+^*, c_-^*),\end{aligned}$$

since w^* has one non-zero coordinate $w_j^* = 1$ (w^* is 1-sparse). Finally,

$$R_{\text{rob}}(\hat{y}^*_{\text{int}}, \varepsilon, \|\cdot\|_\infty) = R_{\text{rob}}(\hat{y}^*_{\text{int}}, \varepsilon, \|\cdot\|_2),$$

since $S_y^c(\hat{y}^*_{\text{int}}) + B_{2,\varepsilon} = S_y^c(\hat{y}^*_{\text{int}}) + B_{\infty,\varepsilon}$ for $y \in \{-1, 0, 1\}$; the misclassification sets are coordinate-aligned. Thus, it follows that $\hat{y}^*_{\text{int}}$ also optimizes $R_{\text{rob}}(\hat{y}^*_{\text{int}}, \varepsilon, \|\cdot\|_\infty)$. ■

C. Proof of Theorem VI.3

Recall our general formula:

$$R(\hat{y},\varepsilon) = \pi \cdot P_{x|y=1}(S_{-1}+B_\varepsilon) + (1-\pi)\cdot P_{x|y=-1}(S_1+B_\varepsilon).$$

Take a linear classifier $\hat{y}^*(x)=\operatorname{sign}(x^\top w - c)$ for some some w, c, and $P_{x|y}=\mathcal{N}(y\mu, I_p)$. Then S_1 is the set of datapoints such that $x^\top w - c \geq 0$. So, S_1+B_ε is the set of datapoints such that $x^\top w - c \geq -\varepsilon\|w\|_1$. Thus, restricting without loss of generality to w such that $\|w\|_2=1$,

$$\begin{aligned} R(w,c;\varepsilon) &= \pi\cdot P_{\mathcal{N}(\mu,I)}(x^\top w - c \leq \varepsilon\|w\|_1) \\ &\quad + (1-\pi)\cdot P_{\mathcal{N}(-\mu,I)}(x^\top w - c \geq -\varepsilon\|w\|_1) \\ &= \pi\cdot\Phi\left(c+\varepsilon\|w\|_1 - \mu^\top w\right) \\ &\quad + (1-\pi)\cdot\Phi\left(-c+\varepsilon\|w\|_1 - \mu^\top w\right). \end{aligned}$$

The minimizer is

$$c^* = \frac{q}{2\cdot(\mu^\top w - \varepsilon\|w\|_1)},$$

where recall that $q=\log[(1-\pi)/\pi]$. This applies when $\mu^\top w - \varepsilon\|w\|_1 > 0$. If that does not happen, then the weight w is not aligned properly with the problem, in the sense that it reduces the "effective" effect size to a negative value. Thus, we do not need to consider those cases.

Another way to put this is that for a weight w with unit norm $\|w\|_2=1$, a linear classifier reduces the effect size from $\mu^\top w$ (which we can assume to be positive, without loss of generality, by flipping the sign if needed), to $\mu^\top w - \varepsilon\|w\|_1$. So we can solve the problem:

$$\sup_w \mu^\top w - \varepsilon\|w\|_1 \quad \text{s.t.} \quad \|w\|_2 = 1.$$

First, we can WLOG restrict to weights w which have the same sign as μ, because for any w, flipping a sign of a coordinate such that it has the same sign as μ_i increases (or does not decrease, in the extreme case where μ_i or w_i are zero), the objective. Moreover, we can also solve first the problem where all coordinates of μ are non-negative. (Then we can flip the signs of w according to the sign of μ to recover the solution).

These simplifications lead to the problem with $\mu_i \geq 0$

$$\sup_w \sum_i [\mu_i - \varepsilon]\cdot w_i \quad \text{s.t.} \quad \|w\|_2 = 1,\ w_i \geq 0.$$

If, for some i, $\mu_i - \varepsilon \leq 0$, then we need to set $w_i = 0$. For the remaining coordinates, we can upper bound the objective value by the Cauchy-Schwarz inequality: $v^\top w \leq \|v\|_2 \cdot \|w\|_2 = \|v\|_2$; with $v = \mu - \varepsilon\cdot 1$ restricted to the positive coordinates. Moreover, to satisfy the unit norm constraint, we need to set $w^* = v/\|v\|_2$.

More generally, with negative coordinates, the solution will depend on the *soft thresholding* operator $v=\eta(\mu,\lambda)$ well known in signal processing and statistics.

Specifically, we will have $v=\eta(\mu,\varepsilon)$, and $w = v/\|v\|_2$. Then we also get

$$c^* = \frac{q}{2\cdot(\mu^\top w - \varepsilon\|w\|_1)} = \frac{q}{2\cdot\|\eta(\mu,\varepsilon)\|}.$$

This shows that the optimal classifier is $\operatorname{sign}\{\eta(\mu,\varepsilon)^\top x - q/2\}$, as desired. ■

D. Proof of Theorem VI.4

If $\|w\|_2=1$, then the linear interval classifier is

$$\hat{y}_{\text{int}}(x; w, c_+, c_-) = \hat{y}_{\text{int}}(x^\top w; 1, c_+, c_-),$$

and the problem effectively reduces to a one-dimensional problem with new variable $\tilde{x}_w := x^\top w \in \mathbb{R}$, which is the mixture of Gaussians $\tilde{x}_w | y \sim \mathcal{N}(yw^\top\mu, 1)$, where $\varepsilon\|w\|_1$ is the corresponding one-dimensional perturbation.

Hence, the robust risk to minimize with respect to weights $\|w\|_2=1$ and thresholds $c_+ \geq c_-$ is

$$\begin{aligned} &\widetilde{R}(w, c_+, c_-) \\ &:= R_{\text{rob}}\left\{\hat{y}_{\text{int}}(\tilde{x}_w; 1, c_+, c_-), \varepsilon\|w\|_1\right\} \\ &= \pi_- \Pr_{\tilde{x}_w|y=-1}(\tilde{x}_w > c_- - \varepsilon\|w\|_1) \\ &\quad + \pi_+ \Pr_{\tilde{x}_w|y=1}(\tilde{x}_w < c_+ + \varepsilon\|w\|_1) \\ &\quad + \pi_0 \Pr_{\tilde{x}_w|y=0}(\tilde{x}_w \leq c_- + \varepsilon\|w\|_1 \text{ or } \tilde{x}_w \geq c_+ - \varepsilon\|w\|_1) \\ &= \pi_- \Pr_{\tilde{x}_w|y=-1}(\tilde{x}_w > c_- - \varepsilon\|w\|_1) \\ &\quad + \pi_+ \Pr_{\tilde{x}_w|y=1}(\tilde{x}_w < c_+ + \varepsilon\|w\|_1) \\ &\quad + \pi_0 \min\Big\{1, \Pr_{\tilde{x}_w|y=0}(\tilde{x}_w \leq c_- + \varepsilon\|w\|_1) \\ &\qquad\qquad + \Pr_{\tilde{x}_w|y=0}(\tilde{x}_w \geq c_+ - \varepsilon\|w\|_1)\Big\} \\ &= \pi_-\bar{\Phi}(c_- - \varepsilon\|w\|_1 + w^\top\mu) + \pi_+\Phi(c_+ + \varepsilon\|w\|_1 - w^\top\mu) \\ &\quad + \pi_0 \min\left\{1, \Phi(c_- + \varepsilon\|w\|_1) + \bar{\Phi}(c_+ - \varepsilon\|w\|_1)\right\} \\ &= \min\{\widetilde{R}_1(w, c_+, c_-), \widetilde{R}_2(w, c_+, c_-)\}, \end{aligned}$$

where Φ is the normal CDF, its complement is $\bar{\Phi} := 1-\Phi$, and

$$\begin{aligned} \widetilde{R}_1(w, c_+, c_-) &:= \pi_-\bar{\Phi}(c_- - \varepsilon\|w\|_1 + w^\top\mu) \\ &\quad + \pi_+\Phi(c_+ + \varepsilon\|w\|_1 - w^\top\mu) + \pi_0, \\ \widetilde{R}_2(w, c_+, c_-) &:= \pi_-\bar{\Phi}(c_- - \varepsilon\|w\|_1 + w^\top\mu) \\ &\quad + \pi_+\Phi(c_+ + \varepsilon\|w\|_1 - w^\top\mu) \\ &\quad + \pi_0\left\{\Phi(c_- + \varepsilon\|w\|_1) + \bar{\Phi}(c_+ - \varepsilon\|w\|_1)\right\}. \end{aligned}$$

Now, $\widetilde{R}_1$ amounts to the two-class setting in Theorem VI.3 and is likewise minimized by

$$\tilde{w}_1^* = \frac{\eta_\varepsilon(\mu)}{\|\eta_\varepsilon(\mu)\|_2}, \qquad c_+ = c_- = \tilde{c}^* = \frac{\ln(\pi_-/\pi_+)}{2\|\eta_\varepsilon(\mu)\|_2},$$

since $\widetilde{R}_1$ is a decreasing function (for $c_+ \geq c_-$ fixed) in $w^\top\mu - \varepsilon\|w\|_1$, which is itself maximized by $\eta_\varepsilon(\mu)$. Assuming $\varepsilon < \|\mu\|_\infty/2$ prevents the degenerate case where $\eta_\varepsilon(\mu)=0$, and with $w=\tilde{w}_1^*$ fixed, minimization with respect to $c_+ \geq c_-$ is as in the proof of Theorem V.1; note that $\eta_\varepsilon(\mu)^\top\mu - \varepsilon\|\eta_\varepsilon(\mu)\|_1 = \|\eta_\varepsilon(\mu)\|_2^2$. Thus,

$$\inf_{\substack{\|w\|_2=1 \\ c_+\geq c_-}} \widetilde{R}_1(w, c_+, c_-) = \widetilde{R}_1(\tilde{w}_1^*, \tilde{c}^*, \tilde{c}^*).$$

Next, note that

$$\widetilde{R}_2(w, c_+, c_-) \geq \widetilde{R}_1(w, c_+, c_-),$$

when $c_- + \varepsilon\|w\|_1 \geq c_+ - \varepsilon\|w\|_1$ so we need only minimize $\widetilde{R}_2(w, c_+, c_-)$ over $c_- + \varepsilon\|w\|_1 \leq c_+ - \varepsilon\|w\|_1$, which is equivalently expressed via change of variables as

$$\begin{aligned}&\inf_{\substack{\|w\|_2=1\\ c_+ \geq c_- + 2\varepsilon\|w\|_1}} \widetilde{R}_2(w, c_+, c_-)\\ &\quad= \inf_{\substack{\|w\|_2=1\\ \tau_+ \geq \tau_-}} \widetilde{R}_2(w, \tau_+ + \varepsilon\|w\|_1, \tau_- - \varepsilon\|w\|_1).\end{aligned}$$

For any $\tau_+ \geq \tau_-$,

$$\begin{aligned}&\widetilde{R}_2(w, \tau_+ + \varepsilon\|w\|_1, \tau_- - \varepsilon\|w\|_1)\\ &\quad= \pi_- \bar{\Phi}(\tau_- - 2\varepsilon\|w\|_1 + w^\top\mu)\\ &\qquad+ \pi_+ \Phi(\tau_+ + 2\varepsilon\|w\|_1 - w^\top\mu) + \pi_0\big\{\Phi(\tau_-) + \bar{\Phi}(\tau_+)\big\},\end{aligned}$$

is a decreasing function of $w^\top\mu - 2\varepsilon\|w\|_1$, which is maximized by $\tilde{w}_2^* := \eta_{2\varepsilon}(\mu)/\|\eta_{2\varepsilon}(\mu)\|_2$; again the case $\eta_{2\varepsilon}(\mu) = 0$ is prevented by $\varepsilon < \|\mu\|_\infty/2$. Fixing $w = \tilde{w}_2^*$, minimization with respect to $c_+ \geq c_- + 2\varepsilon\|w\|_1$ is as in the proof of Theorem V.1. Namely,

$$\begin{aligned}\tilde{c}_+^* &:= +\frac{(\tilde{w}_2^*)^\top\mu}{2} + \frac{\ln(\pi_0/\pi_+)}{\|\eta_{2\varepsilon}(\mu)\|_2},\\ \tilde{c}_-^* &:= -\frac{(\tilde{w}_2^*)^\top\mu}{2} - \frac{\ln(\pi_0/\pi_-)}{\|\eta_{2\varepsilon}(\mu)\|_2},\end{aligned}$$

are optimal if $\tilde{c}_+^* \geq \tilde{c}_-^* + 2\varepsilon\|\tilde{w}_2^*\|_1$, and setting $c_+ = c_- + 2\varepsilon\|\tilde{w}_2^*\|_1$ is optimal otherwise. Thus

$$\begin{aligned}&\inf_{\substack{\|w\|_2=1\\ c_+ \geq c_- + 2\varepsilon\|w\|_1}} \widetilde{R}_2(w, c_+, c_-)\\ &\quad= \begin{cases}\widetilde{R}_2(\tilde{w}_2^*, \tilde{c}_+^*, \tilde{c}_-^*), & \text{if } \tilde{c}_+^* \geq \tilde{c}_-^* + 2\varepsilon\|\tilde{w}_2^*\|_1,\\ \inf_{c\in\mathbb{R}} \widetilde{R}_2(\tilde{w}_2^*, c + 2\varepsilon\|\tilde{w}_2^*\|_1, c), & \text{otherwise.}\end{cases}\end{aligned}$$

Putting it all together, we conclude that

$$\begin{aligned}&\inf_{\substack{\|w\|_2=1\\ c_+ \geq c_-}} \widetilde{R}(w, c_+, c_-)\\ &\quad= \min\left\{\begin{array}{l}\inf\limits_{\substack{\|w\|_2=1\\ c_+ \geq c_-}} \widetilde{R}_1(w, c_+, c_-),\\ \inf\limits_{\substack{\|w\|_2=1\\ c_+ \geq c_- + 2\varepsilon\|w\|_1}} \widetilde{R}_2(w, c_+, c_-)\end{array}\right\}\\ &\quad= \begin{cases}\min\left\{\begin{array}{l}\widetilde{R}_1(\tilde{w}_1^*, \tilde{c}^*, \tilde{c}^*),\\ \widetilde{R}_2(\tilde{w}_2^*, \tilde{c}_+^*, \tilde{c}_-^*)\end{array}\right\}, & \text{if } \tilde{c}_+^* \geq \tilde{c}_-^* + 2\varepsilon\|\tilde{w}_2^*\|_1,\\ \min\left\{\begin{array}{l}\widetilde{R}_1(\tilde{w}_1^*, \tilde{c}^*, \tilde{c}^*),\\ \inf\limits_{c\in\mathbb{R}} \widetilde{R}_2(\tilde{w}_2^*, c + 2\varepsilon\|\tilde{w}_2^*\|_1, c)\end{array}\right\}, & \text{otherwise,}\end{cases}\\ &\quad= \begin{cases}\min\left\{\begin{array}{l}\widetilde{R}_1(\tilde{w}_1^*, \tilde{c}^*, \tilde{c}^*),\\ \widetilde{R}_2(\tilde{w}_2^*, \tilde{c}_+^*, \tilde{c}_-^*)\end{array}\right\}, & \text{if } \tilde{c}_+^* \geq \tilde{c}_-^* + 2\varepsilon\|\tilde{w}_2^*\|_1,\\ \widetilde{R}_1(\tilde{w}_1^*, \tilde{c}^*, \tilde{c}^*), & \text{otherwise,}\end{cases}\end{aligned}$$

where the final equality follows from the observation that

$$\begin{aligned}\widetilde{R}_2(\tilde{w}_2^*, c + 2\varepsilon\|\tilde{w}_2^*\|_1, c) &= \widetilde{R}_1(\tilde{w}_2^*, c + 2\varepsilon\|\tilde{w}_2^*\|_1, c)\\ &\geq \widetilde{R}_1(\tilde{w}_1^*, \tilde{c}^*, \tilde{c}^*).\end{aligned}$$

Hence, we have optimal linear interval classifiers given by two cases: i) $\tilde{w}_1^*$ and $\tilde{c}^*$, or ii) $\tilde{w}_2^*$ and $\tilde{c}_\pm^*$. Note that (ii) remains a valid/feasible choice so long as $\tilde{c}_+^* \geq \tilde{c}_-^*$ even if $\tilde{c}_+^* < \tilde{c}_-^* + 2\varepsilon\|\tilde{w}_2^*\|_1$; it may just be sub-optimal in that case. Finally, noting that weights and thresholds can be scaled, i.e.,

$$\begin{aligned}&\hat{y}_{\text{int}}(x; \tilde{w}_1^*, \tilde{c}^*, \tilde{c}^*)\\ &\quad= \hat{y}_{\text{int}}(x; \tilde{w}_1^*\|\eta_\varepsilon(\mu)\|_2, \tilde{c}^*\|\eta_\varepsilon(\mu)\|_2, \tilde{c}^*\|\eta_\varepsilon(\mu)\|_2),\\ &\hat{y}_{\text{int}}(x; \tilde{w}_2^*, \tilde{c}_+^*, \tilde{c}_-^*)\\ &\quad= \hat{y}_{\text{int}}(x; \tilde{w}_2^*\|\eta_{2\varepsilon}(\mu)\|_2, \tilde{c}_+^*\|\eta_{2\varepsilon}(\mu)\|_2, \tilde{c}_-^*\|\eta_{2\varepsilon}(\mu)\|_2),\end{aligned}$$

and simplifying completes the proof. ■

APPENDIX F
PROOFS FOR FINITE SAMPLE ANALYSIS

A. Proof of Proposition VIII.1

The first portion of this proof follows that of [17, Lemma 10], in that we first consider the following problem:

$$\begin{aligned}w_n^* &\in \operatorname*{argmin}_{\|w\|_2\leq 1} \sum_{i=1}^n \max_{\|\delta_i\|_\infty \leq \varepsilon} -y_i\left(\langle x_i + \delta_i, w\rangle\right)\\ &= \operatorname*{argmax}_{\|w\|_2\leq 1} \sum_{i=1}^n \min_{\|\delta_i\|_\infty \leq \varepsilon} y_i\left(\langle x_i + \delta_i, w\rangle\right).\end{aligned}$$

In the inner minimization problem, it holds that

$$\min_{\|\delta_i\|_\infty \leq \varepsilon} y_i\langle x_i + \delta_i, w\rangle = y_i\langle x_i, w\rangle - \varepsilon\left\|w\right\|_1,$$

by the definition of the dual norm. Therefore the original problem takes the form

$$\begin{aligned}w_n^* &\in \operatorname*{argmax}_{\|w\|_2\leq 1} \sum_{i=1}^n y_i\langle x_i, w\rangle - \varepsilon\left\|w\right\|_1\\ &= \operatorname*{argmax}_{\|w\|_2\leq 1} n\left(\langle u, w\rangle - \varepsilon\left\|w\right\|_1\right).\end{aligned}$$

where we have defined $u := \frac{1}{n}\sum_{i=1}^n y_i x_i$. Now if we let $w(j)$ and $u(j)$ denote the j^{th} components of the vectors w and u respectively, we have

$$w_n^* \in \operatorname*{argmax}_{\|w\|_2\leq 1} \sum_{j=1}^d u(j)w(j) - \varepsilon|w(j)|,$$

Notice that if $u(j) \neq 0$, then $\operatorname{sign}(u(j)) = \operatorname{sign}(w_n^*(j))$ as flipping the signs will only make the j^{th} term smaller. On the other hand, if $u(j) = 0$, then the maximum is achieved when $w_n^*(j) = 0$. Thus $\operatorname{sign}(u) = \operatorname{sign}(w_n^*)$. Now in a similar way to what was done in the proof of Theorem VI.3, let us assume WLOG that $u \succeq 0$, which implies that $w_n^* \succeq 0$ as well. Then we wish to solve

$$\operatorname*{argmax}_w \langle u - \varepsilon\mathbb{1}, w\rangle \quad \text{s.t.} \quad \|w\|_2 \leq 1, w \succeq 0.$$

It follows that $w_n^\star = \eta(u, \varepsilon)/\left\|\eta(u, \varepsilon)\right\|$ where η is the soft-thresholding operator. ■

B. Proof of Theorem VIII.2

The formula of the robust risk for a classifier $\hat{y}$ is

$$R(\hat{y},\varepsilon) = P(y=1)P_{x|y=1}(S_{-1}+B_\varepsilon) + P(y=-1)P_{x|y=-1}(S_1+B_\varepsilon).$$

This expression holds for any classification problem, and the set S_1 (resp. S_{-1}) denotes the set of all $x \in \mathbb{R}^p$ which are classified to $+1$ (resp. -1) by the classifier $\hat{y}$. When $\hat{y}$ is a linear classifier, both sets S_1 and S_{-1} are half-spaces, e.g., $S_1 = \{x \in \mathbb{R}^p : w^T x - c \geq 0\}$. Furthermore, it is easy to see that the sets $S_{+1}+B_\varepsilon$ and $S_{-1}+B_\varepsilon$ are also half-spaces. E.g., we have $S_1 + B_\varepsilon = \{x \in \mathbb{R}^p : w^T x - c + \varepsilon\|w\|_* \geq 0\}$ where $\|\cdot\|_*$ is the dual norm. In other words, we can interpret $S_1 + B_e p$ as the set of all the points that are classified as $+1$ by a slightly shifted *linear* classifier $(w, c-\varepsilon\|w\|_*)$. Hence, the term $P_{x|y=-1}(S_1+B_\varepsilon)$ is the probability that the the new linear classifier $(w, c-\varepsilon\|w\|_*)$ labels a point x as $+1$ while x is generated conditioned on $y=-1$.

Let now (x_i, y_i) for $i = 1, \ldots, n$ be sampled iid from a joint distribution $P_{x,y}$ for $i = 1, \ldots, n$. Let the fraction of 1-s be $\pi_n \in [0,1]$. Let $P_{n\pm}$ be the empirical distributions of x_i given $y_i = 1$ and -1, respectively. We can write the finite sample robust risk as

$$R_n(\hat{y},\varepsilon) = \pi_n \cdot P_{n+}(S_{-1}+B_\varepsilon) + (1-\pi_n)\cdot P_{n-}(S_1+B_\varepsilon). \tag{38}$$

As explained above, for any linear classifier (w,c) the sets S_1+B_ε and $S_{-1}+B_\varepsilon$ are equivalent to half-spaces created by slightly shifted linear classifiers. Hence, considering the hypothesis class of all linear classifiers, the complexity of the sets S_1 (resp. S_{-1}) is the same as the complexity of the sets $S_{+1}+B_\varepsilon$ (resp. $S_{-1}+B_\varepsilon$). Now, by using standard arguments from uniform-convergence theory and PAC learning, and noting that the class of halfspaces has VC-dimension $p+1$, we conclude that for any $\delta > 0$,

$$\Pr\Big\{\forall_{(w,c)\in\mathbb{R}^p\times\mathbb{R}} \Big|P_{n+}(S_1+B_\varepsilon) - P_{x|y=1}(S_{-1}+B_\varepsilon)\Big| \leq \delta\Big\} \geq 1-\exp(C(p-n\delta^2)),$$

where C is a constant independent of n, p. A similar result can be obtained for uniform concentration on the sets $S_{-1}+B_\varepsilon$. We also note (using, e.g., Hoeffding's inequality) that $\Pr(|\pi_n - P(y=1)| < \delta) \geq 1-2\exp(-n\delta^2)$. The result of the theorem now follows by incorporating the bounds obtained above into (38) and choosing C sufficiently large but independent of n, p. ■

ACKNOWLEDGMENT

The authors thank the editor and anonymous reviewers for their helpful comments and suggestions that led to significant improvements in the strength and clarity of the paper.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learning Representations*, 2014.

[3] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Proc. Joint European Conf. Mach. Learning and Knowledge Discovery in Databases*, 2013, pp. 387–402.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learning Representations*, 2015.

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learning Representations*, 2018.

[6] E. Wong and J. Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *Proc. 35th Int. Conf. Mach. Learning*, vol. 80, 2018, pp. 5286–5295.

[7] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. 2016 IEEE Conf. Comput. Vision Pattern Recognition*, 2016, pp. 2574–2582.

[8] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. 36th Int. Conf. Mach. Learning*, vol. 97, 2019, pp. 7472–7482.

[9] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *Proc. Int. Conf. Learning Representations*, 2019.

[10] D. Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao, "Is robustness the cost of accuracy? – a comprehensive study on the robustness of 18 deep image classification models," in *Proc. European Conf. Comput. Vision*, 2018, pp. 644–661.

[11] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry, "Adversarially robust generalization requires more data," in *Adv. Neural Inform. Process. Syst.*, vol. 31, 2018.

[12] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 742–769, 2019.

[13] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *Proc. Int. Conf. Learning Representations*, 2017.

[14] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien, "A closer look at memorization in deep networks," in *Proc. 34th Int. Conf. Mach. Learning*, vol. 70, 2017, pp. 233–242.

[15] A. Raghunathan, S. M. Xie, F. Yang, J. C. Duchi, and P. Liang, "Adversarial training can hurt generalization," in *Proc. ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*, 2019.

[16] A. Javanmard, M. Soltanolkotabi, and H. Hassani, "Precise tradeoffs in adversarial training for linear regression," in *Proc. 33rd Conf. on Learning Theory*, vol. 125, 2020, pp. 2034–2078.

[17] L. Chen, Y. Min, M. Zhang, and A. Karbasi, "More data can expand the generalization gap between adversarially robust and standard models," in *Proc. 37th Int. Conf. Mach. Learning*, vol. 119, 2020, pp. 1670–1680.

[18] Y. Min, L. Chen, and A. Karbasi, "The curious case of adversarially robust models: More data can help, double descend, or hurt generalization," in *Proc. 37th Conf. Uncertainty in Artificial Intell.*, vol. 161, 2021, pp. 129–139.

[19] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. Wiley New York, 2003.

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer New York, 2009.

[21] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer New York, 2006.

[22] M. E. A. Seddik, C. Louart, M. Tamaazousti, and R. Couillet, "Random matrix theory proves that deep learning representations of GAN-data behave as Gaussian mixtures," in *Proc. 37th Int. Conf. Mach. Learning*, vol. 119, 2020, pp. 8573–8582.

[23] A. Cianchi, N. Fusco, F. Maggi, and A. Pratelli, "On the isoperimetric deficit in Gauss space," *Amer. J. Math.*, vol. 133, no. 1, pp. 131–186, 2011.

[24] E. Mossel and J. Neeman, "Robust dimension free isoperimetry in Gaussian space," *Ann. Probability*, vol. 43, no. 3, pp. 971–991, 2015.

[25] H. Xu, C. Caramanis, and S. Mannor, "Robustness and regularization of support vector machines," *J. Mach. Learning Research*, vol. 10, no. 51, pp. 1485–1510, 2009.

[26] ——, "Robust regression and lasso," in *Adv. Neural Inform. Process. Syst.*, vol. 21, 2008.

[27] I. Megyeri, I. Hegedűs, and M. Jelasity, "Adversarial robustness of linear models: Regularization and dimensionality," in *Proc. European Symp. Artificial Neural Networks, Comput. Intell. and Mach. Learning*, 2019, pp. 61–66.

[28] R. Bhattacharjee, S. Jha, and K. Chaudhuri, "Sample complexity of robust linear classification on separated data," in *Proc. 38th Int. Conf. Mach. Learning*, vol. 139, 2021, pp. 884–893.

[29] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proc. 36th Int. Conf. Mach. Learning*, vol. 97, 2019, pp. 1310–1320.

[30] H. Salman, G. Yang, J. Li, P. Zhang, H. Zhang, I. Razenshteyn, and S. Bubeck, "Provably robust deep learning via adversarially trained smoothed classifiers," in *Adv. Neural Inform. Process. Syst.*, vol. 32, 2019.

[31] A. Blum, T. Dick, N. Manoj, and H. Zhang, "Random smoothing might be unable to certify ℓ_∞ robustness for high-dimensional images," *J. Mach. Learning Research*, vol. 21, no. 211, pp. 1–21, 2020.

[32] J. Mohapatra, C.-Y. Ko, L. Weng, P.-Y. Chen, S. Liu, and L. Daniel, "Hidden cost of randomized smoothing," in *Proc. 24th Int. Conf. Artificial Intell. and Stat.*, vol. 130, 2021, pp. 4033–4041.

[33] B. Puranik, U. Madhow, and R. Pedarsani, "Adversarially robust classification based on GLRT," in *Proc. 2021 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 3785–3789.

[34] J. Gilmer, L. Metz, F. Faghri, S. S. Schoenholz, M. Raghu, M. Wattenberg, and I. Goodfellow, "Adversarial spheres," in *Proc. Int. Conf. Learning Representations*, 2018.

[35] A. Shafahi, W. R. Huang, C. Studer, S. Feizi, and T. Goldstein, "Are adversarial examples inevitable?" in *Proc. Int. Conf. Learning Representations*, 2019.

[36] S. Mahloujifar, X. Zhang, M. Mahmoody, and D. Evans, "Empirically measuring concentration: Fundamental limits on intrinsic robustness," in *Adv. Neural Inform. Process. Syst.*, vol. 32, 2019.

[37] E. Richardson and Y. Weiss, "A bayes-optimal view on adversarial examples," *J. Mach. Learning Research*, vol. 22, no. 221, pp. 1–28, 2021.

[38] A. Raghunathan, S. M. Xie, F. Yang, J. Duchi, and P. Liang, "Understanding and mitigating the tradeoff between robustness and accuracy," in *Proc. 37th Int. Conf. Mach. Learning*, vol. 119, 2020, pp. 7909–7919.

[39] J. Hayes, "Provable trade-offs between private & robust machine learning," 2020, unpublished. [Online]. Available: https://arxiv.org/abs/2006.04622v1

[40] C. Dan, Y. Wei, and P. Ravikumar, "Sharp statistical guarantees for adversarially robust Gaussian classification," in *Proc. 37th Int. Conf. Mach. Learning*, vol. 119, 2020, pp. 2345–2355.

[41] A. Fawzi, O. Fawzi, and P. Frossard, "Analysis of classifiers' robustness to adversarial perturbations," *Mach. Learning*, vol. 107, no. 3, pp. 481–508, 2018.

[42] A. Sanyal, P. K. Dokania, V. Kanade, and P. H. S. Torr, "How benign is benign overfitting?" in *Proc. Int. Conf. Learning Representations*, 2021.

[43] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *Proc. Nat. Academy Sci.*, vol. 117, no. 48, pp. 30 063–30 070, 2020.

[44] M. Mehrabi, A. Javanmard, R. A. Rossi, A. Rao, and T. Mai, "Fundamental tradeoffs in distributionally adversarial training," in *Proc. 38th Int. Conf. Mach. Learning*, vol. 139, 2021, pp. 7544–7554.

[45] F. Tramèr, J. Behrmann, N. Carlini, N. Papernot, and J.-H. Jacobsen, "Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations," in *Proc. 37th Int. Conf. Mach. Learning*, vol. 119, 2020, pp. 9561–9571.

[46] D. Cullina, A. N. Bhagoji, and P. Mittal, "PAC-learning in the presence of evasion adversaries," in *Adv. Neural Inform. Process. Syst.*, vol. 31, 2018.

[47] D. Yin, R. Kannan, and P. Bartlett, "Rademacher complexity for adversarially robust generalization," in *Proc. 36th Int. Conf. Mach. Learning*, vol. 97, 2019, pp. 7085–7094.

[48] I. Attias, A. Kontorovich, and Y. Mansour, "Improved generalization bounds for robust learning," in *Proc. 30th Int. Conf. Algorithmic Learning Theory*, vol. 98, 2019, pp. 162–183.

[49] P. Awasthi, N. Frank, and M. Mohri, "Adversarial learning guarantees for linear hypotheses and neural networks," in *Proc. 37th Int. Conf. Mach. Learning*, vol. 119, 2020, pp. 431–441.

[50] O. Montasser, S. Goel, I. Diakonikolas, and N. Srebro, "Efficiently learning adversarially robust halfspaces with noise," in *Proc. 37th Int. Conf. Mach. Learning*, vol. 119, 2020, pp. 7010–7021.

[51] I. Steinwart, "How to compare different loss functions and their risks," *Constructive Approximation*, vol. 26, no. 2, pp. 225–287, 2007.

[52] P. Awasthi, A. Mao, M. Mohri, and Y. Zhong, "A finer calibration analysis for adversarial robustness," 2021, unpublished. [Online]. Available: https://arxiv.org/abs/2105.01550v2

[53] H. Bao, C. Scott, and M. Sugiyama, "Calibrated surrogate losses for adversarially robust classification," in *Proc. 33rd Conf. on Learning Theory*, vol. 125, 2020, pp. 408–451.

[54] A. N. Bhagoji, D. Cullina, V. Sehwag, and P. Mittal, "Lower bounds on cross-entropy loss in the presence of test-time adversaries," in *Proc. 38th Int. Conf. Mach. Learning*, vol. 139, 2021, pp. 863–873.

[55] Y. Wang, S. Jha, and K. Chaudhuri, "Analyzing the robustness of nearest neighbors to adversarial examples," in *Proc. 35th Int. Conf. Mach. Learning*, vol. 80, 2018, pp. 5133–5142.

[56] R. Bhattacharjee and K. Chaudhuri, "When are non-parametric methods robust?" in *Proc. 37th Int. Conf. Mach. Learning*, vol. 119, 2020, pp. 832–841.

[57] Y.-Y. Yang, C. Rashtchian, Y. Wang, and K. Chaudhuri, "Robustness for non-parametric classification: A generic attack and defense," in *Proc. 23rd Int. Conf. Artificial Intell. and Stat.*, vol. 108, 2020, pp. 941–951.

[58] Y.-Y. Yang, C. Rashtchian, H. Zhang, R. Salakhutdinov, and K. Chaudhuri, "A closer look at accuracy vs. robustness," in *Adv. Neural Inform. Process. Syst.*, vol. 33, 2020, pp. 8588–8601.

[59] A. N. Bhagoji, D. Cullina, and P. Mittal, "Lower bounds on adversarial robustness from optimal transport," in *Adv. Neural Inform. Process. Syst.*, vol. 32, 2019.

[60] M. S. Pydi and V. Jog, "Adversarial risk via optimal transport and optimal couplings," in *Proc. 37th Int. Conf. Mach. Learning*, vol. 119, 2020, pp. 7814–7823.

[61] E. Dohmatob, "Limitations of adversarial robustness: Strong no free lunch theorem," in *Proc. 36th Int. Conf. Mach. Learning*, vol. 97, 2019, pp. 1646–1654.

[62] C. Borell, "The Brunn-Minkowski inequality in Gauss space," *Inventiones Mathematicae*, vol. 30, no. 2, pp. 207–216, 1975.

[63] V. N. Sudakov and B. S. Tsirel'son, "Extremal properties of half-spaces for spherically invariant measures," *J. Soviet Math.*, vol. 9, no. 1, pp. 9–18, 1978.

[64] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[65] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," in *Proc. Int. Conf. Learning Representations*, 2018.

[66] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proc. 35th Int. Conf. Mach. Learning*, vol. 80, 2018, pp. 284–293.

[67] M. Goibert and E. Dohmatob, "Adversarial robustness via label-smoothing," 2019, unpublished. [Online]. Available: https://arxiv.org/abs/1906.11567v2

[68] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *J. Amer. Stat. Assoc.*, vol. 101, no. 473, pp. 138–156, 2006.

[69] J. Khim and P.-L. Loh, "Adversarial risk bounds via function transformation," 2018, unpublished. [Online]. Available: https://arxiv.org/abs/1810.09519v2

[70] O. Montasser, S. Hanneke, and N. Srebro, "VC classes are adversarially robustly learnable, but only improperly," in *Proc. 32nd Conf. on Learning Theory*, vol. 99, 2019, pp. 2512–2530.

[71] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *Ann. Math. Stat.*, vol. 27, no. 3, pp. 642–669, 1956.

[72] M. R. Kosorok, *Introduction to Empirical Processes and Semiparametric Inference*. Springer New York, 2008.

[73] G. R. Shorack and J. A. Wellner, *Empirical Processes with Applications to Statistics*. Society for Industrial and Applied Mathematics, 2009.

[74] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proc. 35th Int. Conf. Mach. Learning*, vol. 80, 2018, pp. 274–283.

Edgar Dobriban is an assistant professor of Statistics and Data Science, with a secondary appointment in Computer and Information Science, at the University of Pennsylvania. Prior to that, he received a PhD degree in Statistics from Stanford University in 2017, and MS degree in Electrical Engineering from Stanford University in 2015. Prior to that, he received a Bachelor of Arts degree in Mathematics from Princeton University in 2012 (with highest honors). His research interests include the statistical analysis of large-scale data, including dimension reduction and high-dimensional statistics; and the theoretical foundations of machine learning, including uncertainty quantification, the use of symmetry, and fairness, among other topics. He is a recipient of the National Science Foundation (NSF) CAREER Award (2021), a 2022 Junior Research award from the International Chinese Statistical Association (ICSA), and the Theodore W. Anderson award for the best thesis in theoretical statistics from the Department of Statistics at Stanford University in 2017.

Hamed Hassani is currently an assistant professor of Electrical and Systems Engineering department at the University of Pennsylvania. Prior to that, he was a research fellow at Simons Institute for the Theory of Computing (UC Berkeley) affiliated with the program of Foundations of Machine Learning, and a post-doctoral researcher in the Institute of Machine Learning at ETH Zurich. He received a Ph.D. degree in Computer and Communication Sciences from EPFL, Lausanne. He is the recipient of the 2014 IEEE Information Theory Society Thomas M. Cover Dissertation Award, 2015 IEEE International Symposium on Information Theory Student Paper Award, 2017 Simons-Berkeley Fellowship, 2018 NSF-CRII Research Initiative Award, 2020 Air Force Office of Scientific Research (AFOSR) Young Investigator Award, 2020 National Science Foundation (NSF) CAREER Award, and 2020 Intel Rising Star award. He has recently been selected as the distinguished lecturer of the IEEE Information Theory Society in 2022-2023.

David Hong is a National Science Foundation Post-Doctoral Fellow in the Department of Statistics and Data Science at the University of Pennsylvania. He received the BS degree in ECE and Mathematics from Duke University in 2013, the MS degree in EECS from the University of Michigan in 2015, and the PhD degree in EECS from the University of Michigan in 2019, where he was a National Science Foundation Graduate Fellow. His research interests are in the theoretical foundations for low-dimensional modeling of modern high-dimensional and heterogeneous data, and its many applications.

Alexander Robey is a Ph.D. student in the Department of Electrical and Systems Engineering at the University of Pennsylvania, where he is advised by Professors Hamed Hassani and George J. Pappas. At Penn, he is affiliated with the GRASP robotics laboratory and the Warren Center for Network and Data Sciences. During his Ph.D., he has also spent time as part of the AI research team at Google Cloud. Prior to beginning his Ph.D., he received a B.S. in engineering and a B.A. in mathematics from Swarthmore College in 2018. His research interests include convex optimization and robust machine learning.