Understanding the Impact of Adversarial Robustness on Accuracy Disparity

Yuzheng Hu [1] Fan Wu [1] Hongyang Zhang [2] Han Zhao [1]

Abstract

While it has long been empirically observed that adversarial robustness may be at odds with standard accuracy and may have further disparate impacts on different classes, it remains an open question to what extent such observations hold and how the class imbalance plays a role within. In this paper, we attempt to understand this question of accuracy disparity by taking a closer look at linear classifiers under a Gaussian mixture model. We decompose the impact of adversarial robustness into two parts: an inherent effect that will degrade the standard accuracy on all classes due to the robustness constraint, and the other caused by the class imbalance ratio, which will increase the accuracy disparity compared to standard training. Furthermore, we also show that such effects extend beyond the Gaussian mixture model, by generalizing our data model to the general family of stable distributions. More specifically, we demonstrate that while the constraint of adversarial robustness consistently degrades the standard accuracy in the balanced class setting, the class imbalance ratio plays a fundamentally different role in accuracy disparity compared to the Gaussian case, due to the heavy tail of the stable distribution. We additionally perform experiments on both synthetic and real-world datasets to corroborate our theoretical findings. Our empirical results also suggest that the implications may extend to nonlinear models over real-world datasets. Our code is publicly available on GitHub[1].

1. Introduction

The existence and prevalence of adversarial examples (Dalvi et al., 2004; Szegedy et al., 2013; Goodfellow et al., 2015) in the state-of-the-art deep learning models have made adversarial robustness an active field of research, where human-imperceptible perturbations to the original data can arbitrarily disrupt the model prediction. However, it has been empirically observed that the improvement in robustness usually comes with costs in accuracy. In particular, there might exist a trade-off between robust accuracy and standard accuracy (Tsipras et al., 2019; Kolter & Madry, 2018), and it may also lead to the so-called *accuracy disparity* (Chi et al., 2021), a notion of unfairness in the literature of algorithmic fairness. In particular, it is shown that when compared to standard training, the constraint of adversarial robustness might further exacerbate the *discrepancy* of standard accuracy among difference classes (Croce et al., 2021; Benz et al., 2021a;b).

Despite fruitful and intriguing empirical observations, rigorous understanding of how adversarial robustness affects standard accuracy or accuracy disparity has not been extensively explored from a theoretical perspective. In fact, it is not clear whether such a trade-off is inherent, even in the linear setting under a Gaussian mixture model. If it is, then what are the fundamental factors that contribute to this potential drop of accuracy and the increase of accuracy disparity? To the best of our knowledge, there are only a few works (Tsipras et al., 2019; Xu et al., 2021; Ma et al., 2022) that partially attempt to approach these problems. However, the existing analyses are restricted to examples with specific choices of parameters, which oversimplifies the problem and makes it unclear whether the conclusions continue to hold in more general settings. We provide further discussions on the related work in Section 6.

Our Contributions. Towards answering the above questions, we provide a theoretical study of the impact of adversarial robustness on accuracy disparity in the presence of class imbalance (Johnson & Khoshgoftaar, 2019). We consider the classification problem under a common Gaussian mixture model with linear classifiers. We then further generalize our analysis to a broader family of stable distributions. For each data distribution, we decompose the impact of adversarial robustness into two parts, an intrinsic

[1]Department of Computer Science, University of Illinois at Urbana Champaign, Urbana, IL, USA [2]David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada. Correspondence to: Yuzheng Hu <yh46@illinois.edu>.

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

[1] `https://github.com/Accuracy-Disparity/AT-on-AD`

part that leads to the drop of standard accuracy and the class imbalance part which will affect the accuracy disparity. Our main contributions are summarized as follows:

- Under a common Gaussian mixture model and linear classifiers, we reveal two effects of enforcing adversarial robustness in binary classification. The first part is inherent to the constraint of adversarial robustness itself, which will degrade the standard accuracy due to a change of direction of the optimal linear classifier. The second part is caused by the class imbalance ratio between the two classes under consideration, which will increase the accuracy disparity compared to standard training due to a reduction of norm of the optimal linear classifier.
- Inspired by our analysis, we further point out the equivalence between robust training in classification and regularized linear regression. Our observation helps to explain the norm-shrinkage effect that happens in robust learning, which could be of independent theoretical interest.
- Going beyond the Gaussian mixture model, we show that when the data follow a mixture of heavy-tailed stable distributions, the intrinsic effect of robustness persists and the standard accuracy consistently decreases even when the data is balanced among the two classes. On the other hand, the class imbalance ratio plays a fundamentally different role compared to that in the Gaussian case, suggesting that the tail property of the distribution also serves as a crucial factor in the accuracy disparity problem.
- We conduct experiments on both synthetic and real-world datasets. The empirical results not only corroborate our theoretical findings, but also suggest that the implications may extend to nonlinear models over real-world datasets.

2. Preliminaries

We first give an overview of the problem studied in this paper. We then proceed to introduce the necessary background and notation used throughout the paper.

Problem setup. For the ease of presentation, we focus on a binary classification task in this paper. To start with, we assume that the data are generated through a mixture of distribution $\mathcal{P}^+$ and $\mathcal{P}^-$: conditioning on $y = \pm 1$, we have $X \sim \mathcal{P}^\pm$, respectively. As we shall see shortly, one crucial ingredient in understanding the impact of adversarial robustness on accuracy disparity is the imbalance factor $R > 1$ between the marginal probabilities of different classes: $R := \Pr(y = -1)/\Pr(y = +1)$, meaning that there is a larger portion of negative-class examples in the population. Following prior works (Tsipras et al., 2019; Xu et al., 2021), for the model, we consider a linear classifier and couple it with a sign function sgn to obtain the output $f(x; w, b) := \mathrm{sgn}(w^\top x + b)$. Finally, since we are mainly interested in understanding the inherent impact caused by adversarial robustness, throughout the paper we shall focus on the infinite data regime to remove the noise introduced by finite samples, meaning that we will study the population losses instead of their empirical counterparts.

Objective functions. The standard 0-1 population loss (*standard loss*) is defined as follows:

$$\begin{aligned}&\ell_{\text{std}}(w,b) := \Pr(f(x;w,b) \neq y)\\ =&\frac{R}{R+1}\underbrace{\mathbb{P}(w^\top x + b \geq 0 \mid y=-1)}_{\text{Part I: } \ell_{\text{std}}^-} + \frac{1}{R+1}\underbrace{\mathbb{P}(w^\top x + b \leq 0 \mid y=+1)}_{\text{Part II: } \ell_{\text{std}}^+}.\end{aligned}$$

Similarly, the 0-1 adversarial loss (Kurakin et al., 2016; Madry et al., 2018) under ℓ_p-perturbation ($p \geq 1$) and radius ε (*robust ℓ_p loss*), is defined as:

$$\begin{aligned}&\ell_{\text{rob},p,\varepsilon}(w,b) := \Pr(\exists \|\delta\|_p \leq \varepsilon,\ s.t.\ f(x+\delta;w,b) \neq y)\\ =&\frac{R}{R+1}\mathbb{P}(w^\top x + b \geq -\varepsilon\|w\|_q \mid y=-1)\\ +&\frac{1}{R+1}\mathbb{P}(w^\top x + b \leq \varepsilon\|w\|_q \mid y=+1),\end{aligned}$$

where $1/p + 1/q = 1$ and the second equation follows from the Hölder's inequality. When the context is clear, we will omit the use of ε in $\ell_{\text{rob},p,\varepsilon}$. It is straightforward to see that minimizing $\ell_{\text{rob},p}$ will lead to adversarial robustness.

Accuracy disparity. Note that Part II and I in the definition of the standard loss are exactly the population loss for the minority and majority classes, which we denote as ℓ_{std}^+ and ℓ_{std}^-, respectively. The standard accuracy for both classes then writes as $acc^\pm := 1 - \ell_{\text{std}}^\pm$. The key quantity that we will focus on in this paper is the *accuracy disparity* (Chi et al., 2021) between the two classes, defined as

$$AD(w,b) := acc^-(w,b) - acc^+(w,b) = \ell_{\text{std}}^+(w,b) - \ell_{\text{std}}^-(w,b).$$

The notion of accuracy disparity in our context focuses on the performance gap of a model on different sub-groups of the overall population, where each group is indexed by the corresponding class label (Santurkar et al., 2021; Xu et al., 2021). Accuracy disparity has recently gained more attention in the literature of algorithmic fairness (Buolamwini & Gebru, 2018; Chi et al., 2021; Nanda et al., 2021), and we are interested in understanding the role of robustness in accuracy disparity. To proceed, we first characterize the optimal solutions of the standard loss and the robust ℓ_p loss

$$w_{\text{std}}, b_{\text{std}} = \operatorname*{arg\,min}_{w,b} \ell_{\text{std}}(w,b),\ w_{\text{rob},p}, b_{\text{rob},p} = \operatorname*{arg\,min}_{w,b} \ell_{\text{rob},p}(w,b).$$

We then analyze the changes of ℓ_{std}^+ and ℓ_{std}^- as well as the accuracy disparity when we *switch* the measurements from the optimal standard classifier to the optimal robust classifier. Specifically, we are most interested in how these changes depend on the class imbalance ratio R. In other words, we will try to understand the price (*i.e.*, the decrease of the standard loss) paid by a robust classifier in the presence of class imbalance. We emphasize that accuracy disparity is defined over the standard loss for both the standard classifier and robust classifier. There is another concept, known as *robustness disparity* (Nanda et al., 2021), which is defined over the robust loss; it is orthogonal to accuracy disparity and will not be covered in this paper.

Notation. We use Φ to denote the cumulative distribution function of the standard normal distribution. For a vector u, $|u|$ means taking the absolute value per coordinate, u_i denotes the i-th coordinate of u, and $\|u\|_p$ represents the ℓ_p-norm of u. For two vectors v and v', we use the notation $v \parallel v'$ to indicate they are parallel and $v \nparallel v'$ when they are not. We denote 1_d as the all-one vector with dimension d. For $1 \leq p \leq \infty$, we use q to denote its dual index, which is defined through $1/p + 1/q = 1$. Finally, for a multivariate function f, the subdifferential set at x is defined as $\partial f(x) := \{g : \forall y, f(y) \geq f(x) + g^\top (y - x)\}$, and each element within is called a subgradient.

3. Gaussian Mixture: Robustness Implies Accuracy Disparity

In this section, we will closely examine the accuracy disparity of robust classifiers when the data are drawn according to a Gaussian mixture (Reynolds, 2009). Specifically, we are interested in how enforcing the model to be robust affects the accuracy disparity compared to standard training. While it has been shown in a previous work (Xu et al., 2021) that adversarial robustness does introduce severe accuracy disparity when different classes exhibit different "difficulty levels" of learning (*i.e.*, different magnitude of variance) in a toy example (as indicated by specific choices of mean, variance, as well as $p = \infty$), in this section we consider a more general setting where *class imbalance* is present, and we shall provide a comprehensive analysis by considering the following objectives and data distributions: 1) adversarial robustness with general ℓ_p-constraint ($1 \leq p \leq \infty$); 2) a Gaussian mixture distribution with arbitrary mean and covariance matrix, specifically, $\mathcal{P}^+ = \mathcal{N}(\theta^+, \Sigma)$ and $\mathcal{P}^- = \mathcal{N}(\theta^-, \Sigma)$.

In a nutshell, the overall effects of adversarial robustness are separated into two parts: an inherent one that will **decrease the standard accuracy** on all classes, and the other caused additionally by the class imbalance ratio that will **increase the accuracy disparity** compared to standard training.

Before introducing the main results, we highlight that linear models are sufficiently powerful for the classification of Gaussian mixtures in both the standard and adversarial sense. In fact, it is well known that the Bayes-optimal classifier of the standard loss is linear due to the Fisher's linear discriminant (Johnson et al., 2014); and it is further shown in (Dobriban et al., 2020) that the Bayes-optimal robust classifier is also linear.

3.1. Main Results

In what follows we will present a general result (Theorem 3.1) which characterizes the class-wise standard loss for the optimal standard and robust classifiers. We will then discuss several implications. Specifically, Theorem 3.3 (a direct corollary of Proposition 3.2) demonstrates the effect of "decreased standard accuracy", while Theorem 3.6 (derived form Proposition 3.5) demonstrates the effect of "increased accuracy disparity". The proofs are deferred to Appendix A.1.

Theorem 3.1. *Given the means θ^+, θ^-, covariance matrix Σ and ℓ_p-constraint, let $u, v \in \mathbb{R}^d$ satisfy*

$$\Sigma u = \theta^+ - \theta^-, \qquad \Sigma v = \theta^+ - \theta^- - 2\varepsilon \partial \|v\|_q, \quad (1)$$

and q satisfy $1/p + 1/q = 1$. We further set $r^2 := u^\top \Sigma u$, and $s^2 = v^\top \Sigma v$. Then the class-wise standard loss $\ell^\pm_{\text{std}}$ of the optimal standard classifier $(w_{std} := u/r, b_{std})$ satisfy

$$\ell^+_{\text{std}}(w_{\text{std}}, b_{\text{std}}) = \Phi\left(\frac{-\langle u, \theta^+ - \theta^- \rangle + 2\log R}{2r}\right),$$
$$\ell^-_{\text{std}}(w_{\text{std}}, b_{\text{std}}) = \Phi\left(\frac{-\langle u, \theta^+ - \theta^- \rangle - 2\log R}{2r}\right),$$

and the class-wise standard loss $\ell^\pm_{\text{std}}$ of the optimal robust ℓ_p classifier $(w_{rob,p} := v/s, b_{rob,p})$ satisfy

$$\ell^+_{\text{std}}(w_{\text{rob},p}, b_{\text{rob},p}) = \Phi\left(\frac{-\langle v, \theta^+ - \theta^- \rangle + 2\log R}{2s}\right),$$
$$\ell^-_{\text{std}}(w_{\text{rob},p}, b_{\text{rob},p}) = \Phi\left(\frac{-\langle v, \theta^+ - \theta^- \rangle - 2\log R}{2s}\right).$$

It is straightforward to see that the overall effects consist of two parts: one intrinsic due to adversarial robustness ($\frac{\langle u, \theta^+ - \theta^- \rangle}{2r}$ v.s. $\frac{\langle v, \theta^+ - \theta^- \rangle}{2s}$), and the other caused by the class imbalance ratio ($\frac{\log R}{r}$ v.s. $\frac{\log R}{s}$). To compare the optimal robust classifier and the optimal standard classifier, we will show in the following analysis that 1) the intrinsic part corresponds to a change in **direction**, and will degrade the standard performance on both classes — exactly the price paid by a classifier to be robust; 2) the class imbalance part corresponds to a change in **norm**, and will increase the error on the minority class while decreasing the error on the majority class — hence exacerbating the accuracy disparity.

The intrinsic part—decreasing the standard accuracy. The intrinsic part corresponds to a change in *direction* —

in fact, we will show that the optimal solution moves to a direction that incurs larger error due to the constraint of robustness. This helps to explain the long-observed empirical phenomenon that adversarial training (Kurakin et al., 2016; Madry et al., 2018), which is an effective algorithm for empirical robustness, often leads to degraded standard accuracy (Tsipras et al., 2019; Kolter & Madry, 2018).

Proposition 3.2. *For the intrinsic part, we have*

$$\frac{\langle u, \theta^+ - \theta^- \rangle}{2r} \geq \frac{\langle v, \theta^+ - \theta^- \rangle}{2s}.$$

Furthermore, so long as $\Sigma^{\frac{1}{2}} w_{\text{std}} \nparallel \Sigma^{\frac{1}{2}} w_{\text{rob},p}$, the inequality holds strictly.

When the covariance matrix is invertible, we are comparing the directions of the two optimal classifiers in the standard sense. For general covariance matrix, it is therefore natural for us to interpret $\Sigma^{\frac{1}{2}} w$ as a general "direction" of w. Proposition 3.2 now directly leads to the following result.

Theorem 3.3. *When there is no class imbalance, i.e., $R = 1$, enforcing adversarial robustness with ℓ_p-constraint will degrade the standard accuracy on both classes, so long as the "direction" of the optimal robust classifier is not parallel to its counterpart, i.e., $\Sigma^{\frac{1}{2}} w_{\text{std}} \nparallel \Sigma^{\frac{1}{2}} w_{\text{rob},p}$.*

Remark 3.4. Intuitively, we expect that the direction of the optimal classifier to be different for the standard and robust loss. We will provide more discussions about this observation in Section 4.2, where we demonstrate that this is indeed the situation for diagonal matrices; but we also identify a few cases where we can actually get a win-win from both worlds.

The class imbalance part—increasing the accuracy disparity. The class imbalance part corresponds to a change in *norm*, which appears as the square root of quadratic form. We will show: due to a *shrinkage of norm*, the standard loss of the minority class increases while the opposite happens for the majority class.

Proposition 3.5. *For the class imbalance part, we have $u^\top \Sigma u > v^\top \Sigma v$, which implies $r > s$.*

In the special case $\Sigma = \mathbb{I}_d$, the square root of the quadratic form becomes the standard Euclidean norm. It is therefore natural for us to interpret the quantities r and s as realizations of a general "norm" (in fact, it is a seminorm defined by Σ) . With Proposition 3.5 at hand, we are now ready to state the following result, which says that class imbalance will increase the accuracy disparity due to the constraint of robustness, and such growth is monotonic with respect to R in a reasonable range. This demonstrates an **inherent trade-off** of adversarial robustness and accuracy parity.

Theorem 3.6. *Define $g(R) := AD(w_{\text{rob},p}, b_{\text{rob},p}) - AD(w_{\text{std}}, b_{\text{std}})$ as the accuracy disparity gap. Then*

- *When $R > 1$, we have $g(R) > 0$, meaning that the accuracy disparity of the optimal robust classifier is larger than that of the optimal standard classifier;*
- *When R satisfies $\ell^+_{\text{std}}(w_{\text{rob},p}, b_{\text{rob},p}) \leq 0.5$, i.e., the four class-wise losses defined in Theorem 3.1 are upper-bounded by 0.5, $g(R)$ is an increasing function w.r.t. R.*

Note that the direction and norm of the normal vector of a linear classifier determines its decision boundary. Hence, Theorem 3.1 and the above discussions completely characterize and provide a fine-grained analysis of the impact of adversarial robustness on the accuracy disparity of linear classifiers over mixture of Gaussian distributions.

3.2. An Illustrating Example

We will now use an example that has been proposed and studied in (Tsipras et al., 2019) to illustrate our main results and demonstrate that Theorem 3.1 can be used to recover and refine existing claims on this example. Specifically, let

$$\theta^+ = (\underbrace{\eta, \cdots, \eta}_{\text{dim}=m}, \underbrace{\gamma, \cdots, \gamma}_{\text{dim}=n})^\top, \quad \theta^- = -\theta^+,$$

where $\gamma < \varepsilon < \eta$ and $m + n = d$. This corresponds to two sets of features in the input space: the *robust features* (coordinates with value η) and the *non-robust features* (coordinates with value γ). Compared to Tsipras et al. (2019), here we slightly modify the setting by making all coordinates Gaussian and allowing the number of robust feature m to be larger than one, though in general we would still expect $m \ll n$. For the other assumptions, we follow Tsipras et al. (2019) to set $\Sigma = \mathbb{I}_d$ and consider the standard perturbation scheme, *i.e.*, $p = \infty$.

The optimal standard and robust classifier. We can assume without loss of generality that $\|w\|_2 = 1$, and the optimal slope and intercept for the standard and robust classifier are given by (details deferred to Appendix A.2):

$$b_{\text{std}} = -\frac{\log R}{2\sqrt{m\eta^2 + n\gamma^2}},$$
$$w_{\text{std},1} = \cdots = w_{\text{std},m} = \frac{\eta}{\sqrt{m\eta^2 + n\gamma^2}},$$
$$w_{\text{std},m+1} = \cdots = w_{\text{std},m+n} = \frac{\gamma}{\sqrt{m\eta^2 + n\gamma^2}},$$
$$b_{\text{rob},\infty} = -\frac{\log R}{2(\eta - \varepsilon)\sqrt{m}},$$
$$w_{\text{rob},\infty,1} = \cdots = w_{\text{rob},\infty,m} = \frac{1}{\sqrt{m}},$$
$$w_{\text{rob},\infty,m+1} = \cdots = w_{\text{rob},\infty,m+n} = 0.$$

Therefore, the standard loss for both classes are

$$\ell^{+}_{\text{std}}(w_{\text{std}}, b_{\text{std}}) = \Phi\left(\frac{\log R}{2\sqrt{m\eta^2 + n\gamma^2}} - \sqrt{m\eta^2 + n\gamma^2}\right),$$
$$\ell^{-}_{\text{std}}(w_{\text{std}}, b_{\text{std}}) = \Phi\left(-\frac{\log R}{2\sqrt{m\eta^2 + n\gamma^2}} - \sqrt{m\eta^2 + n\gamma^2}\right),$$
$$\ell^{+}_{\text{std}}(w_{\text{rob},\infty}, b_{\text{rob},\infty}) = \Phi\left(\frac{\log R}{2\sqrt{m(\eta - \varepsilon)^2}} - \sqrt{m\eta^2}\right),$$
$$\ell^{-}_{\text{std}}(w_{\text{rob},\infty}, b_{\text{rob},\infty}) = \Phi\left(-\frac{\log R}{2\sqrt{m(\eta - \varepsilon)^2}} - \sqrt{m\eta^2}\right).$$

By comparing ℓ^{+}_{std} and ℓ^{-}_{std} of the optimal standard classifier as well as the optimal robust classifier, we shall see that the effects of adversarial robustness indeed match the main results. Specifically,

- When $R = 1$, enforcing adversarial robustness will decrease the standard accuracy on *both* classes due to a shrinkage on the non-robust features ($\sqrt{m\eta^2 + n\gamma^2}$ v.s. $\sqrt{m\eta^2}$), which reflects "a change in direction";
- When $R > 1$, class imbalance will increase the accuracy disparity due to a shrinkage on both the robust features and the non-robust features ($\sqrt{m\eta^2 + n\gamma^2}$ v.s. $\sqrt{m(\eta - \varepsilon)^2}$ in the denominator), which reflects "a reduction of norm".

Furthermore, when the number of robust features is small and the number of non-robust features is relatively large (corresponding to the case $m \ll n$ (Tsipras et al., 2019; Ilyas et al., 2019)), the overall effect tends to be a *significant growth in accuracy disparity*. As a concrete example, if we set $m = 4, n = 48, \eta = 1, \varepsilon = 0.75, \gamma = 0.5, R = e^2$, we have $\ell^{+}_{\text{std}}(w_{\text{std}}, b_{\text{std}}), \ell^{-}_{\text{std}}(w_{\text{std}}, b_{\text{std}}), \ell^{-}_{\text{std}}(w_{\text{rob},\infty}, b_{\text{rob},\infty}) < 0.001$ whereas $\ell^{+}_{\text{std}}(w_{\text{rob},\infty}, b_{\text{rob},\infty}) = 0.5$, which means that $0.5 \approx AD(w_{\text{rob},\infty}, b_{\text{rob},\infty}) \gg AD(w_{\text{std}}, b_{\text{std}}) \approx 0$.

3.3. Connection to (Regularized) Linear Regression

Here we delve deeper into the *fundamental cause* of the norm shrinkage effect as demonstrated in the previous section. Our main findings can be summarized into one sentence: Eq. (1), which is used to solve for the optimal slopes w_{std} and $w_{\text{rob},p}$, enjoys the same form as the optimal conditions of (regularized) linear regression. Specifically, consider the following optimization problems

$$\arg\min_{\beta} \frac{1}{2N}\|Y - X\beta\|_2^2, \text{ and } \arg\min_{\beta} \frac{1}{2N}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_q, \tag{2}$$

where $X \in \mathbb{R}^{N\times d}$ is the design matrix, $Y \in \mathbb{R}^N$ is the label vector, and $\beta \in \mathbb{R}^d$ is the estimator. The first-order conditions of Eq. (2) give us

$$X^\top X\beta = X^\top Y \quad \text{and} \quad X^\top X\beta = X^\top Y - N\lambda\partial\|\beta\|_q.$$

Therefore, Eq. (1) has the same form as Eq. (2) by setting

$$\Sigma = X^\top X, \quad \theta^+ - \theta^- = X^\top Y, \quad \varepsilon = N\lambda/2.$$

As a consequence, solving for the optimal robust ℓ_p classifier is essentially performing linear regression with ℓ_q-regularization, and this explains the norm shrinkage effect in Proposition 3.5.

The connection between robust training and regularized linear regression has been noted in a prior work (Xu et al., 2008) for regression. As a comparison, here we demonstrate that such equivalence also holds true for classification problems under Gaussian mixture distributions, which requires a delicate analysis of the KKT conditions in addition to directly exploiting the duality between the data and parameters as in the regression problem. Furthermore, explicitly formalizing this connection allows us to interpret the tradeoff of adversarial robustness and accuracy parity from the following perspective—A larger ε implies better robustness, but also results in a strong regularization (*i.e.*, norm shrinkage) effect as $\varepsilon \propto \lambda$, hence leading to an increased accuracy disparity.

4. Beyond Gaussian Mixture — Stable Distributions, and Polynomial Tail

In this section, we will go beyond the Gaussian mixture distribution, and examine whether the conclusions we have drawn so far hold true for a broader class of data distributions. In particular, we explore a family of distributions that includes the Gaussian distribution as a special case: the *symmetric α-stable ($S\alpha S$) distribution* (Lévy, 1954; Fama & Roll, 1968; 1971). The motivation for us to study the $S\alpha S$ distribution is twofold. First of all, it is a natural generalization of the Gaussian distribution and preserves an important property of Gaussian: closed under linear transformation since the characteristic function is closed under multiplication. This property then allows us to obtain a precise characterization of the standard/robust loss in terms of the cumulative function. Second, by varying the choice of α, we can better understand whether the findings that we have obtained thus far are specific to Gaussian distribution, or hold true in the presence of heavy-tail.

The results in this section are mixed: while the conclusion of "decreased standard accuracy" generalizes to the $S\alpha S$ distribution, the "increased accuracy disparity" phenomenon disappears, and we shall see that the class imbalance ratio will play a fundamentally different role in affecting accuracy disparity when heavy tail is present. To start with, we assume $\varepsilon \leq \frac{\kappa}{2}\|\theta^+ - \theta^-\|_\infty$ for some $\kappa < 1$ throughout this section, meaning there exists at least one dimension such that the two balls do not intersect.

4.1. A Brief Review of the $S\alpha S$ distribution

The probability distribution function of a univariate $S\alpha S$ distribution with *location*, *scale* and *stability* parameters μ, c, α, is defined through

$$f(x; \alpha, c, \mu) = \frac{1}{2\pi} \int_{\mathbb{R}} \varphi(t; \mu, c, \alpha) e^{-ixt} \mathrm{d}t,$$

with $\varphi(t; \mu, c, \alpha) = \exp\left(it\mu - |ct|^\alpha\right)$ being its characteristic function. The most important quantity here is the tail-index $\alpha \in (0, 2]$ which measures the *concentration* of the corresponding stable distribution, and we recover the Gaussian and Cauchy distribution by setting $\alpha = 2$ and $\alpha = 1$, respectively. We will mainly focus on multivariate $S\alpha S$ distribution with independent components throughout this section, meaning that each coordinate is independent from the others and follows $f(x; \alpha, c_i, \mu_i)$ with $c_i > 0$, and we denote it as $S\alpha S_{IC}(\mu, C)$ where $C = \operatorname{diag}\{c_i\}$. We refer interested readers to Appendix B.3 for discussions of a different multivariate $S\alpha S$ distribution.

4.2. Adversarial Robustness (Still) Hurts the Standard Accuracy for Balanced Dataset

We start with the balanced case, *i.e.*, $R = 1$, and examine whether enforcing adversarial robustness *provably* hurts the standard accuracy. We assume that the data are generated through a mixture of multivariate $S\alpha S$ distributions with independent components and scale parameters $c_i = 1$: $\mathcal{P}^+ = S\alpha S_{IC}(\theta^+, \mathbb{I}_d)$ and $\mathcal{P}^- = S\alpha S_{IC}(\theta^-, \mathbb{I}_d)$. For general choices of c_i, we can scale the coordinates of θ^+ and θ^- inverse-proportionally to obtain the same conclusion.

We first identify two corner cases in Theorem 4.1, where the optimal robust classifier achieves the same standard accuracy as the optimal standard classifier. The detailed analyses are deferred to Appendix B.1.

Theorem 4.1. *Under one of the following conditions: 1) $q = \alpha$, meaning that the dual index equals the tail index; 2) $\bar{\theta} := \theta^+ - \theta^-$ is isotropic, i.e., $|\bar{\theta}| \parallel 1_d$ and $\alpha \geq 1$, the optimal robust classifier enjoys the same standard accuracy as the optimal standard classifier when there is no class imbalance.*

Remark 4.2. It is pointed out in *(Dobriban et al., 2020)* that enforcing adversarial robustness with ℓ_2-constraint will not sacrifice the standard accuracy for balanced Gaussian mixture; this corresponds to $q = \alpha = 2$ in the first corner case. Similarly, we can also get a win-win for Cauchy mixture with ℓ_∞-perturbation, *i.e.*, $q = \alpha = 1$.

Except for the two corner cases above, we show that for general α and q, enforcing adversarial robustness will degrade the standard accuracy as stated in the following theorem.

Theorem 4.3. *Suppose $\alpha > 1$, $1 < q < \infty$, $q \neq \alpha$ and $|\bar{\theta}| \nparallel 1_d$. Then we have $w_{rob,p}^\top(\theta^+ - \theta^-) < w_{std}^\top(\theta^+ - \theta^-)$. In other words, adversarial robustness hurts the accuracy on both classes when there is no class imbalance.*

Remark 4.4. We do not discuss other choices of α and q (e.g. $q = 1, \infty$ or $\alpha \leq 1$) in details as they involve a number of pathological corner cases, mainly due to the non-uniqueness of subgradients. However, we can still expect the conclusion to hold true in general, as the two optimization problems differ by exactly one term, which significantly reduces the possibility of overlapped optimal solution.

4.3. Polynomial Tail Perplexes the Effect of Class Imbalance on Accuracy Disparity

Finally, we will study the effect of class imbalance on accuracy disparity using the multivariate Cauchy distribution (*i.e.*, $\alpha = 1$) with independent components and scale parameters $c_i = 1$. Specifically, the data are generated through $\mathcal{P}^+ = S1S_{IC}(\theta^+, \mathbb{I}_d)$ and $\mathcal{P}^- = S1S_{IC}(\theta^-, \mathbb{I}_d)$, and we focus on the conventional perturbation scheme $p = \infty$. The proofs are deferred to Appendix B.2.

Our first result shows that: when heavy-tail is present, the imbalance ratio will result in a *fundamentally different* behavior in terms of the accuracy disparity compared to the Gaussian case, in the sense that both the optimal standard and robust classifier achieve the *same* accuracy disparity for large R.

Theorem 4.5. *Suppose the class imbalance ratio $R \geq 2 + 4\|\theta^+ - \theta^-\|_\infty^2$, then both the optimal standard classifier as well as the optimal robust ℓ_∞ classifier will assign a negative label to all data. As a result, both classifiers will incur zero loss on the majority class and zero accuracy on the minority class. In terms of accuracy disparity, there is no difference between the optimal standard and robust classifier.*

Remark 4.6. We highlight that the reason for observing such a phenomenon is due to the fact that the Cauchy distribution has a polynomial tail. In contrast, this phenomenon does not exist in distributions with exponentially-decayed tail such as the Gaussian distribution.

Our second result shows that: when the distance between the two means $\|\theta^+ - \theta^-\|_\infty$ is relatively large, adversarial robustness will decrease the accuracy on the majority class while increasing the accuracy on the minority class compared to standard training, hence *reducing* the accuracy disparity.

Theorem 4.7. *Assume the optimal intercepts b_{std} and $b_{rob,\infty}$ are finite and $\|\theta^+ - \theta^-\|_\infty^2 > (R+1)^2/R(1-\kappa)^2$, then adversarial robustness will increase the error on the majority class and decrease the error on the minority class, which further reduces the accuracy disparity.*

Remark 4.8. According to Theorem 4.1 (setting $\alpha = q = 1$ in the first corner case), both the optimal standard and robust classifier achieve the same loss on both classes when $R = 1$. Therefore, the changes of accuracy disparity as shown in Theorem 4.5 and Theorem 4.7 are mainly due to the class imbalance part. Contrary to the Gaussian mixture distribution where this factor **consistently** enlarges the accuracy disparity, here we obtain different observations where it stays the same or even decreases, suggesting that the accuracy disparity not only concerns the class imbalance ratio, but is also heavily influenced by the tail property of the corresponding distribution.

5. Experiments

We corroborate and strengthen our theoretical results regarding *accuracy disparity* and *standard accuracy* via experiments on one synthetic dataset and two real-world datasets. In what follows, we first introduce the experiment setup and then lay out the research questions we shall investigate.

Adversarial Training. Our theoretical findings are algorithm agnostic and only concern the definition of adversarial robustness. In the experiments, we choose a popular algorithm, adversarial training (Kurakin et al., 2016; Madry et al., 2018), to perform the robust training.

Metrics. We use $acc^{R,+}_{\text{std}}, acc^{R,-}_{\text{std}}, acc^{R,\cdot}_{\text{std}}$ to denote the *standard accuracy* of the standard classifier trained on the dataset with imbalance ratio R, measured on the minority class, and the majority class, and both classes, respectively. Likewise, we use $acc^{R,+}_{\text{rob},p,\varepsilon}, acc^{R,-}_{\text{rob},p,\varepsilon}, acc^{R,\cdot}_{\text{rob},p,\varepsilon}$ to denote the standard accuracy of the robust classifier trained with ℓ_p perturbations of scale ε, calculated on the three types of populations. We then use $AD^R_{\text{std}}, AD^R_{\text{rob},p,\varepsilon}$ to denote the *accuracy disparity* of the standard classifiers and robust classifiers, formally defined as $AD^R_{\text{std}} = acc^{R,-}_{\text{std}} - acc^{R,+}_{\text{std}}$ and $AD^R_{\text{rob},p,\varepsilon} = acc^{R,-}_{\text{rob},p,\varepsilon} - acc^{R,+}_{\text{rob},p,\varepsilon}$.

Research Questions. We lay out the research questions (RQs) based on Section 3 and 4.

> RQ1. Will adversarial training exacerbate accuracy disparity compared with standard training, *i.e.*, $AD^R_{\text{rob},p,\varepsilon} > AD^R_{\text{std}}$, and when?
>
> RQ2. Will a more severe class imbalance (*i.e.*, a larger imbalance ratio R) lead to a more significant accuracy disparity gap (*i.e.*, a larger $(AD^R_{\text{rob},p,\varepsilon} - AD^R_{\text{std}})$), and when?
>
> RQ3. Will adversarial training worsen the standard accuracy, $acc^{R,\cdot}_{\text{rob},p,\varepsilon} < acc^{R,\cdot}_{\text{std}}$, and when?

The basis of RQ1 and RQ2 from the theoretical side are Theorems 3.6 and 4.5; the basis of RQ3 are Theorems 3.3 and 4.3. We next investigate these questions from the empirical side to gain insights into how well and how far the theoretical results can be supported and extended.

Experimental Setup. We evaluate the above three questions using three datasets: a synthetic dataset of a mixture of Gaussians, as well as two real-world datasets MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky et al., 2009). Results on additional datasets, including two synthetic datasets featuring stable distributions (Cauchy and Holtsmark), as well as two more real-world datasets, Fashion-MNIST (Xiao et al., 2017) and ImageNet (Deng et al., 2009), are provided in Appendices C.2 and C.3. For each dataset, we investigate both the balanced case ($R = 1$) and the imbalanced cases $R \in \{2, 5, 10\}$. For the synthetic dataset and MNIST, we use a linear classifier; for CIFAR, we use a neural network with two linear layers. When performing adversarial training, we use the fast gradient method (FGM) (Goodfellow et al., 2014) and projected gradient descent (PGD) (Madry et al., 2018) to craft the adversarial examples. We experiment with both $p = 2, \infty$, each with multiple perturbation scales ε. For each set of experiments, we report results averaged over 5 runs with different random seeds to account for variability. More details are deferred to Appendix C.1.

5.1. Analysis of the Increased Accuracy Disparity (RQ1 and RQ2)

By comparing the accuracy disparity of standard classifier and a variety of adversarial classifiers, we offer the following answers to RQ1 and RQ2.

> A1: Yes, when $R > 1$.
>
> A2: The increase of the accuracy disparity gap with the imbalance ratio consistently happens for the synthetic dataset and MNIST, but not for CIFAR.

We now present the concrete experimental results along with detailed discussions. We plot the accuracy disparity gap in the 1st row of Figure 1; the raw numbers can be referred to in Appendix C.5 We draw the following conclusions.

Regarding RQ1, the accuracy disparity gap is invariably larger than 0 in the class imbalance setting (*i.e.*, $R > 1$) on all three datasets. This provides an affirmative answer for RQ1 and matches our theoretical result in Theorem 3.6. Actually, in the balanced case, the gap is also close to 0 in most of the cases, apart from an intriguingly low number for CIFAR. We figure that this is associated with the relative difficulty of learning the two classes—in the synthetic dataset and MNIST, the difficulty levels of learning the two classes are alike; for CIFAR, the class 'cat' is much harder to learn than the class 'dog', as already demonstrated in previous work (Croce et al., 2021).

Regarding RQ2, the accuracy disparity gap grows with the class imbalance ratio on both the synthetic dataset and

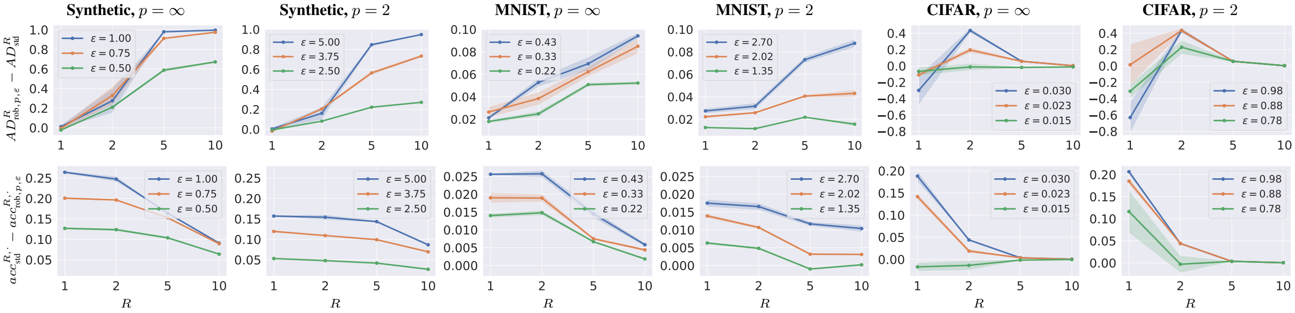

Figure 1. The gap of accuracy disparity $AD^R_{\text{rob},p,\varepsilon} - AD^R_{\text{std}}$ (1st row, RQ2) and standard accuracy $acc^{R,\cdot}_{\text{std}} - acc^{R,\cdot}_{\text{rob},p,\varepsilon}$ (2nd row, RQ3) between robust and standard classifiers, w.r.t. the imbalance ratio R. Different columns correspond to different datasets or ℓ_p norms of adversarial training. For each ℓ_p norm, we experiment with multiple perturbation scales ε. The shaded area in each subfigure represents the standard error of 5 runs.

MNIST. However, the same phenomenon does not occur for CIFAR, where the gap drops to almost 0 for larger R (*i.e.*, $R = 5, 10$). We note that the result on the synthetic dataset and MNIST aligns well with Theorem 3.6; this serves as a verification for the theoretical result on exact Gaussian mixtures and an implication on the potential of extending the result to real-world datasets that can be roughly modeled as Gaussian mixtures. In comparison, what we observe on CIFAR resembles the theoretical analysis on the heavy-tail distribution (Theorem 4.5), where the standard classifier can only achieve accuracy close to 0 on the minority class, accounting for the accuracy disparity gap of nearly 0. We conjecture that the *distributional difference* between MNIST and CIFAR contributes to the distinction in the experimental results. To strengthen our hypothesis, we first perform two additional experiments in Appendix C.4, ruling out alternative explanation regarding the insufficiency of training on CIFAR (caused either by the small dataset size or the limited model capacity, in a relative sense compared to MNIST). We additionally provide an empirical comparison of the statistical properties between MNIST and CIFAR in Appendix C.4, demonstrating that the empirical distribution of CIFAR is indeed more scattered than MNIST.

5.2. Analysis of the Decreased Standard Accuracy (RQ3)

In this part, we shift our focus from the accuracy *disparity* (defined as the gap of the per class standard accuracy) to the *overall* standard accuracy (measured on both classes). In a nutshell:

> A3: Yes, adversarial training almost always hurts the standard accuracy in the scenarios we experiment with.

We plot the gap of the standard accuracy between the standard classifier and the robust classifier in the 2nd row of Figure 1 (raw numbers in Appendix C.5). First, we see that in the class imbalance setting (*i.e.*, $R = 1$), the gap is almost always larger than 0, consistent with the theoretical results in Theorems 3.3 and 4.3. Furthermore, in the class imbalance setting which is beyond the scope of our theoretical results, we observe an interesting decrease of the accuracy gap with the increase of the imbalance ratio R. We offer the following explanation. The impact of class imbalance gradually takes the dominance (in the sense of encouraging the prediction to favor the majority); in this case, whether performing standard or adversarial training will not have much influence on the outcome.

6. Related Work

We will mainly review related work focusing on the relationship between fairness, adversarial robustness, and accuracy.

Robustness and Fairness. What initially motivates this work is the observation of a trade-off between adversarial robustness and fairness (Liu et al., 2021). Here, fairness refers to the *class-wise* performance of the robust classifier (a broader definition would be the performance across subgroups defined by sensitive attributes (Hardt et al., 2016; Zafar et al., 2017)), and is measured by either the robust accuracy or the standard accuracy. The former corresponds to a phenomenon known as “robustness disparity” (Nanda et al., 2021), and is verified on a wide range of datasets, model architectures, as well as attacks and defenses (Nanda et al., 2021; Tian et al., 2021). The latter concerns the “accuracy disparity” (Chi et al., 2021) of the robust classifier, and it is empirically shown that not only does such accuracy disparity exists (Croce et al., 2021; Benz et al., 2021a), but is further exacerbated compared to the standard classifier (Benz et al., 2021b). As a complement to these empirical observations, we aim to provide an in-depth theoretic study towards understanding the impact of adversarial robustness on accuracy disparity.

Detailed comparisons with two closest works. Xu et al. (2021); Ma et al. (2022) identify and analyze the significant disparity of standard accuracy and robust accuracy among different classes or subgroups of data for adversarially trained models. Our work differs from theirs in several key aspects. *First*, their theoretical analysis is restricted to specific choices of parameters, resulting in an oversimplified problem and less convincing conclusions. In contrast, our approach accommodates arbitrary means, covariance matrices, and perturbation types. Additionally, differences exist in the settings and targets of study. Xu et al. (2021) focus on a balanced class setting but with varying "difficulty levels" as measured by the magnitude of variance, whereas we address class imbalance. Ma et al. (2022) measure fairness through the variance of robust risk, while we measure fairness by class-wise accuracy disparity (corresponding to the standard risk). Importantly, we have identified *critical flaws* in the proof of Theorem 5.7 in Ma et al. (2022). On page 18 of their publication[2], a non-trivial gap exists between their Equations (51) and (52) even given their unnatural assumptions on ϵ_{train} and ϵ_{test}, which the authors made no attempt to address. Furthermore, the last inequality in Equation (52) is not correct, as the term is plainly smaller than 1 for small ϵ_{train} and ϵ_{test}. These critical issues undermine the validity of their findings.

Robustness and (Standard) Accuracy. Ever since the seminal work (Tsipras et al., 2019), there has been a line of research studying the fundamental trade-off between adversarial robustness and standard accuracy. This includes some empirical evaluations (Raghunathan et al., 2019; Su et al., 2018), theoretical results regarding the statistical/information limit or sample complexity for robust classification (Bhagoji et al., 2019; Chen et al., 2020; Dan et al., 2020), as well as algorithms that explicitly exploit such trade-off based on theoretical insights (Zhang et al., 2019; Raghunathan et al., 2020; Zhang et al., 2020; Yang et al., 2020). Despite the fruitful results that have been achieved in this field thus far, a rigorous understanding towards how enforcing adversarial robustness decreases the standard accuracy is still lacking. Our work does not target this problem directly, but as a by-product, we show that for balanced dataset and stable distributions, robustness *in general* comes at a cost of degraded performance for standard accuracy due to a change of "direction" (see Subsection 3.1), thus offering a new perspective towards interpreting this intriguing phenomenon.

7. Conclusion, Limitation and Future Directions

In this work, we provide an in-depth and fine-grained study towards understanding the impact of adversarial robustness on accuracy disparity when class imbalance is present. To this end, we offer a complete characterization regarding the classification of a Gaussian mixture with linear models, and decompose the overall effect of enforcing adversarial robustness into two disjoint parts: an inherent one that will degrade the standard accuracy due to a change of "direction", and the other caused additionally by the class imbalance ratio that will increase the accuracy disparity due to a change of "norm". We proceed to analyze the general stable distribution. While the intrinsic effect of robustness can generalize and consistently decrease the standard accuracy even for the balanced class setting, we uncover that the imbalance ratio plays a fundamentally different role in the accuracy disparity due to the heavy tail of the stable distribution. Finally, we support and strengthen our theoretical results with experiments on both synthetic and real-world datasets.

Limitation. An obvious limitation of the paper is that the analyses are restricted to binary classification. Generalization to multi-class classification requires modifying the decision rule; for instance, using argmax of the logits. However, this will lead to a Voronoi diagram partition of the space for the $k > 2$ classes, which is challenging to precisely and analytically characterize (*e.g.*, it is no longer easy to compute the probability mass on each convex body within the diagram and in general this partition does not have analytical characterization).

Future directions. As real-world datasets contain both class imbalance and discrepancy of class-wise distributions, a more complete theory should consider the usage of different covariance matrices and analyze its interaction with the class imbalance ratio. Some additional future directions include 1) introduce the protected attribute A under each label Y, which will make the results more appealing to the fairness community; 2) allow for different test and training distributions, and check whether robustness provably helps in the presence of distribution/subpopulation shifts. On the empirical side, our theoretical insights could lead to the design of future robust training algorithms that aim to achieve a certain notion of accuracy parity among classes. Furthermore, the distributions of real-world datasets in the feature space might be closer to a GMM, and hence one could enforce the robustness constraint on the feature distributions. Last but not least, MNIST and CIFAR exhibit significantly different conclusions on `RQ2`; the nice correspondences between them and the findings in Gaussian and stable distributions could motivate a deeper understanding towards the distributional characteristics of real-world datasets.

Acknowledgements

The work of HZ was supported in part by the Defense Advanced Research Projects Agency (DARPA) under Cooperative Agreement Number: HR00112320012, a Facebook Research Award, and Amazon AWS Cloud Credit. Hongyang Zhang was supported in part by NSERC Discovery Grant RGPIN-2022-03215, DGECR-2022-00357.

[2] `https://openreview.net/attachment?id=LqGA2JMLwBw&name=supplementary_material`

References

Benz, P., Zhang, C., Ham, S., Karjauv, Adil Cho, G., and Kweon, I. S. The triangular trade-off between accuracy, robustness, and fairness. *Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges (AML-CV) at CVPR*, 2021a.

Benz, P., Zhang, C., Karjauv, A., and Kweon, I. S. Robustness may be at odds with fairness: An empirical study on class-wise accuracy. In *NeurIPS 2020 Workshop on Preregistration in Machine Learning*, pp. 325–342. PMLR, 2021b.

Bertsekas, D. P. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

Bhagoji, A. N., Cullina, D., and Mittal, P. Lower bounds on adversarial robustness from optimal transport. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Chen, L., Min, Y., Zhang, M., and Karbasi, A. More data can expand the generalization gap between adversarially robust and standard models. In *International Conference on Machine Learning*, pp. 1670–1680. PMLR, 2020.

Chi, J., Tian, Y., Gordon, G. J., and Zhao, H. Understanding and mitigating accuracy disparity in regression. In *International Conference on Machine Learning*, pp. 1866–1876. PMLR, 2021.

Chrabaszcz, P., Loshchilov, I., and Hutter, F. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL `https://openreview.net/forum?id=SSKZPJCt7B`.

Dalvi, N., Domingos, P., Sanghai, S., and Verma, D. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 99–108, 2004.

Dan, C., Wei, Y., and Ravikumar, P. Sharp statistical guaratees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pp. 2345–2355. PMLR, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dobriban, E., Hassani, H., Hong, D., and Robey, A. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020.

Fama, E. F. and Roll, R. Some properties of symmetric stable distributions. *Journal of the American Statistical Association*, 63(323):817–836, 1968.

Fama, E. F. and Roll, R. Parameter estimates for symmetric stable distributions. *Journal of the American Statistical Association*, 66(334):331–338, 1971.

Friedman, J. H. Exploratory projection pursuit. *Journal of the American statistical association*, 82(397):249–266, 1987.

Gawronski, W. On the bell-shape of stable densities. *The Annals of Probability*, pp. 230–242, 1984.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL `http://arxiv.org/abs/1412.6572`.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Holtsmark, J. Über die verbreiterung von spektrallinien. *Annalen der Physik*, 363(7):577–630, 1919.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

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

Johnson, R. A., Wichern, D. W., et al. *Applied multivariate statistical analysis*, volume 6. Pearson London, UK:, 2014.

Kiefer, J. and Wolfowitz, J. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pp. 462–466, 1952.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Kolter, Z. and Madry, A. Adversarial robustness - theory and practice, 2018. URL https://adversarial-ml-tutorial.org/.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lévy, P. *Théorie de l'addition des variables aléatoires*. Gauthier-Villars, 1954.

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A. K., and Tang, J. Trustworthy ai: A computational perspective. *arXiv preprint arXiv:2107.06641*, 2021.

Ma, X., Wang, Z., and Liu, W. On the tradeoff between robustness and fairness. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=LqGA2JMLwBw.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Nanda, V., Dooley, S., Singla, S., Feizi, S., and Dickerson, J. P. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 466–477, 2021.

Prechelt, L. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pp. 55–69. Springer, 1998.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J. C., and Liang, P. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J., and Liang, P. Understanding and mitigating the tradeoff between robustness and accuracy. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 7909–7919. PMLR, 2020.

Reynolds, D. A. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

Samorodnitsky, G., Taqqu, M. S., and Linde, R. Stable non-gaussian random processes: stochastic models with infinite variance. *Bulletin of the London Mathematical Society*, 28(134):554–555, 1996.

Santurkar, S., Tsipras, D., and Madry, A. Breeds: Benchmarks for subpopulation shift. In *International Conference on Learning Representations*, 2021.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y. (eds.), *ICLR*, 2015. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2015.html#SimonyanZ14a.

Su, D., Zhang, H., Chen, H., Yi, J., Chen, P.-Y., and Gao, Y. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 631–648, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Tian, Q., Kuang, K., Jiang, K., Wu, F., and Wang, Y. Analysis and applications of class-wise robustness in adversarial training. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1561–1570, 2021.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyxAb30cY7.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Xu, H., Caramanis, C., and Mannor, S. Robust regression and lasso. *Advances in neural information processing systems*, 21, 2008.

Xu, H., Liu, X., Li, Y., Jain, A., and Tang, J. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning*, pp. 11492–11501. PMLR, 2021.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R. R., and Chaudhuri, K. A closer look at accuracy vs. robustness. *Advances in neural information processing systems*, 33:8588–8601, 2020.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970. PMLR, 2017.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Zhang, J., Xu, X., Han, B., Niu, G., Cui, L., Sugiyama, M., and Kankanhalli, M. Attacks which do not kill training make adversarial learning stronger. In *International conference on machine learning*, pp. 11278–11287. PMLR, 2020.

A. Omitted Proofs from Section 3

A.1. Proof of the Main Results

Proof of Theorem 3.1. For the standard loss, we have

$$\ell_{\text{std}}(w,b)=\frac{R}{R+1}\Phi\left(\frac{b+w^\top\theta^-}{\sqrt{w^\top\Sigma w}}\right)+\frac{1}{R+1}\Phi\left(\frac{-b-w^\top\theta^+}{\sqrt{w^\top\Sigma w}}\right).$$

Since ℓ_{std} is scale-invariant, we can assume w.l.o.g. that $w^\top\Sigma w=1$. We then obtain an equivalent constrained optimization problem

$$\begin{aligned}\min_{w,b}\quad & R\Phi(b+w^\top\theta^-)+\Phi(-b-w^\top\theta^+)\\ \text{s.t.}\quad & w^\top\Sigma w=1.\end{aligned}$$

The Lagrangian can be written as

$$\mathcal{L}_{\text{std}}(w,b,\nu)=R\Phi(b+w^\top\theta^-)+\Phi(-b-w^\top\theta^+)+\nu(w^\top\Sigma w-1),$$

and the KKT conditions give us

$$\frac{R}{\sqrt{2\pi}}\exp\left(-\frac{(b+w^\top\theta^-)^2}{2}\right)\theta^- -\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(b+w^\top\theta^+)^2}{2}\right)\theta^+ +2\nu\Sigma w=0 \tag{3}$$

and

$$\frac{R}{\sqrt{2\pi}}\exp\left(-\frac{(b+w^\top\theta^-)^2}{2}\right)-\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(b+w^\top\theta^+)^2}{2}\right)=0. \tag{4}$$

Plugging Eq. (4) into Eq. (3), we can conclude that

$$\Sigma w_{\text{std}}=(\theta^+-\theta^-)\cdot C_{\text{std}}$$

for some positive constant C_{std}, and w_{std} additionally satisfies

$$w_{\text{std}}^\top\Sigma w_{\text{std}}=1$$

Therefore, following the statement of Theorem 3.1, suppose $\Sigma u=\theta^+-\theta^-$, then we can pick

$$w_{\text{std}}=\frac{u}{r},$$

where $u^\top\Sigma u=r^2$. After determining w_{std}, b_{std} can be solved directly from Eq. (4), which gives

$$b_{\text{std}}=-\frac{2\log R+(w_{\text{std}}^\top\theta^+)^2-(w_{\text{std}}^\top\theta^-)^2}{2(w_{\text{std}}^\top\theta^+-w_{\text{std}}^\top\theta^-)}.$$

Note

$$w_{\text{std}}^\top\theta^+-w_{\text{std}}^\top\theta^-=\left\langle\frac{u}{r},\Sigma u\right\rangle=r.$$

Therefore,

$$\begin{aligned}\ell_{\text{std}}^+(w_{\text{std}},b_{\text{std}})&=\Phi\left(-b_{\text{std}}-w_{\text{std}}^\top\theta^+\right)\\&=\Phi\left(\frac{-r^2+2\log R}{2r}\right)\\&=\Phi\left(\frac{-\langle u,\theta^+-\theta^-\rangle+2\log R}{2r}\right),\end{aligned}$$

whereas

$$\begin{aligned}\ell^-_{\text{std}}(w_{\text{std}}, b_{\text{std}}) &= \Phi\left(b_{\text{std}} + w_{\text{std}}^\top \theta^-\right) \\ &= \Phi\left(\frac{-r^2 - 2\log R}{2r}\right) \\ &= \Phi\left(\frac{-\langle u, \theta^+ - \theta^-\rangle - 2\log R}{2r}\right).\end{aligned}$$

Similarly, for the robust ℓ_p loss

$$\ell_{\text{rob}}(w,b) = \frac{R}{R+1}\Phi\left(\frac{b + w^\top\theta^- + \varepsilon\|w\|_q}{\sqrt{w^\top\Sigma w}}\right) + \frac{1}{R+1}\Phi\left(\frac{-b - w^\top\theta^+ + \varepsilon\|w\|_q}{\sqrt{w^\top\Sigma w}}\right),$$

we can assume w.l.o.g. that $w^\top\Sigma w = 1$ and write down the Lagrangian

$$\mathcal{L}_{\text{rob}}(w,b,\lambda) = R\Phi(b + w^\top\theta^- + \varepsilon\|w\|_q) + \Phi(-b - w^\top\theta^+ + \varepsilon\|w\|_q) + \mu(w^\top\Sigma w - 1),$$

and the KKT conditions give us

$$\Sigma w_{\text{rob},p} = (\theta^+ - \theta^- - 2\varepsilon\partial\|w_{\text{rob},p}\|_q)\cdot C_{\text{rob},p}.$$

for some positive constant $C_{\text{rob},p}$. A crucial observation here is that the subdifferential set $\partial\|w_{\text{rob},p}\|_q$ is invariant when scaled by a positive constant (guaranteed by Danskin's theorem), so if we follow the statement of Theorem 3.1 and suppose $\Sigma v = \theta^+ - \theta^- - 2\varepsilon\partial\|v\|_q$, then to satisfy the constraint $w_{\text{rob},p}^\top\Sigma w_{\text{rob},p} = 1$, we can pick

$$w_{\text{rob},p} = \frac{v}{s},$$

where $v^\top\Sigma v = s^2$. Similarly, $b_{\text{rob},p}$ can be derived from the KKT conditions as

$$b_{\text{rob},p} = -\frac{2\log R + (w_{\text{rob},p}^\top\theta^+)^2 - (w_{\text{rob},p}^\top\theta^-)^2 - 2\epsilon\|w_{\text{rob},p}\|_q(w_{\text{rob},p}^\top\theta^+ + w_{\text{rob},p}^\top\theta^-)}{2(w_{\text{rob},p}^\top\theta^+ - w_{\text{rob},p}^\top\theta^- - 2\epsilon\|w_{\text{rob},p}\|_q)}.$$

Note

$$\begin{aligned}\langle w_{\text{rob},p}, \theta^+ - \theta^-\rangle &= \left\langle \frac{v}{s}, \Sigma v + 2\epsilon\partial\|v\|_q\right\rangle \\ &= s + 2\epsilon\frac{\langle v, \partial\|v\|_q\rangle}{s} \\ &= s + 2\epsilon\frac{\|v\|_q}{s} \\ &= s + 2\epsilon\|w_{\text{rob},p}\|_q,\end{aligned}$$

where we use Danskin's theorem again in the second-to-last inequality. Therefore,

$$\ell^+_{\text{std}}(w_{\text{rob},p}, b_{\text{rob},p}) = \Phi\left(-b_{\text{rob},p} - w_{\text{rob},p}^\top\theta^+\right) = \Phi\left(\frac{-\langle v, \theta^+ - \theta^-\rangle + 2\log R}{2s}\right),$$

whereas

$$\ell^-_{\text{std}}(w_{\text{rob},p}, b_{\text{rob},p}) = \Phi\left(b_{\text{rob},p} + w_{\text{rob},p}^\top\theta^-\right) = \Phi\left(\frac{-\langle v, \theta^+ - \theta^-\rangle - 2\log R}{2s}\right).$$

This finishes the proof as desired. ■

Proof of Proposition 3.2. Plugging in the equation $\Sigma u = \theta^+ - \theta^-$ as well as the definitions of r and s, it suffices to show

$$\sqrt{u^\top\Sigma u}\sqrt{v^\top\Sigma v} \geq v^\top\Sigma u.$$

Since Σ is positive semi-definite, $\Sigma^{\frac{1}{2}}$ is well-defined. Now denote $u' = \Sigma^{\frac{1}{2}} u$ and $v' = \Sigma^{\frac{1}{2}} v$, then the above inequality is equivalent to

$$\|u'\|_2 \|v'\|_2 \geq v'^{\top} u',$$

which holds due to the Cauchy-Schwarz inequality. Furthermore, the inequality holds strictly as long as u' and v' are not parallel. Combining the fact that u is parallel to w_{std}, and v is parallel to $w_{\text{rob},p}$ finishes the proof as desired.

■

Proof of Theorem 3.3. It follows directly from Proposition 3.2 since Φ is monotonic. ■

Proof of Proposition 3.5. We have

$$\begin{aligned} u^{\top}\Sigma u - v^{\top}\Sigma v &= (u-v)^{\top}\Sigma(u-v) + 2(u-v)^{\top}\Sigma v \\ &\geq 2v^{\top}\Sigma(u-v) \\ &= 2\langle v, 2\varepsilon\partial\|v\|_q\rangle \\ &= 4\varepsilon\|v\|_q > 0, \end{aligned}$$

where we take the difference between

$$\Sigma u = \theta^{+} - \theta^{-} \quad \text{and} \quad \Sigma v = \theta^{+} - \theta^{-} - 2\varepsilon\partial\|v\|_q$$

in the second-to-last equality, and use Danskin's theorem (Bertsekas, 1997) in the last equality. ■

Proof of Theorem 3.6. Denote

$$p = \frac{-\langle u, \theta^{+} - \theta^{-}\rangle}{2r}, \quad q = \frac{-\langle v, \theta^{+} - \theta^{-}\rangle}{2s}.$$

Note $\langle u, \theta^{+} - \theta^{-}\rangle = u^{\top}\Sigma u \geq 0$, $\langle v, \theta^{+} - \theta^{-}\rangle = v^{\top}\Sigma v + 2\varepsilon\|v\|_q > 0$, so Proposition 3.2 implies $p \leq q < 0$. Further denote

$$m = \frac{\log R}{r} > 0, \quad k = \frac{r}{s} > 1,$$

then $\frac{\log R}{s} = km$. We first show the "increased accuracy disparity" part in the theorem, which is equivalent to

$$\Phi(q+km) - \Phi(q-km) > \Phi(p+m) - \Phi(p-m).$$

In fact, we have

$$\Phi(q+km) - \Phi(q-km) > \Phi(q+m) - \Phi(q-m),$$

so it suffices to show

$$\Phi(q+m) - \Phi(q-m) \geq \Phi(p+m) - \Phi(p-m).$$

Define $F(x) = \Phi(x+m) - \Phi(x-m)$, and we will show $F(x)$ is increasing on $(-\infty, 0]$. In fact, we have

$$\begin{aligned} F'(x) &= \frac{1}{\sqrt{2\pi}}\left(\exp\left(-\frac{(x+m)^2}{2}\right) - \exp\left(-\frac{(x-m)^2}{2}\right)\right) \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-m)^2}{2}\right)\left(e^{-2xm} - 1\right) \geq 0 \end{aligned}$$

since $m \geq 0$ and $x < 0$.

To show that the gap between the two accuracy disparities is monotonic, it suffices to prove

$$g(R) = G(m) := \Phi(q + km) - \Phi(q - km) - \Phi(p + m) + \Phi(p - m)$$

is increasing w.r.t. m when $q + km < 0$. In fact, we have

$$\begin{aligned}\sqrt{2\pi}G'(m) = k &\left(\exp\left(-\frac{(q+km)^2}{2}\right) + \exp\left(-\frac{(q-km)^2}{2}\right)\right) \\ &- \left(\exp\left(-\frac{(p+m)^2}{2}\right) + \exp\left(-\frac{(p-m)^2}{2}\right)\right).\end{aligned}$$

Define

$$L(k) := k\left(\exp\left(-\frac{(q+km)^2}{2}\right) + \exp\left(-\frac{(q-km)^2}{2}\right)\right),$$

then it suffices to show that $L'(k) > 0$ when $q + km < 0$. In fact, we have

$$\begin{aligned}L'(k) &= \exp\left(-\frac{(q+km)^2}{2}\right) + \exp\left(-\frac{(q-km)^2}{2}\right) \\ &\quad - km(q+km)\exp\left(-\frac{(q+km)^2}{2}\right) - km(km-q)\exp\left(-\frac{(q-km)^2}{2}\right) \\ &= \exp\left(-\frac{(q-km)^2}{2}\right)\left[e^{-2qkm}(1 - km(q+km)) + (1 - km(km-q))\right].\end{aligned}$$

Let $km = a$ and $-q = b$, it suffices to show

$$e^{2ab} > \frac{ab + (a^2-1)}{ab - (a^2-1)}$$

when $b > a > 0$. In fact, we have

$$\frac{ab + (a^2-1)}{ab - (a^2-1)} \le \frac{ab + (ab-1)}{ab - (ab-1)} = 2ab - 1 < 2ab + 1 \le e^{2ab}.$$

■

A.2. Derivation of the Optimal Classifiers for the Toy Example

Given $X \sim \mathcal{N}(\theta, \Sigma)$ and some $w \in \mathbb{R}^d$, we have $w^\top X \sim \mathcal{N}(w^\top\theta, w^\top\Sigma w)$ due to the fact that Gaussian distribution is closed under linear transformation.

The optimal standard classifier. Using the property above, we have

$$\begin{aligned}\ell_{\text{std}}(w, b) = &\frac{R}{R+1}\Phi\left(\frac{b - \eta\sum_{i=1}^m w_i - \gamma\sum_{j=m+1}^{m+n} w_j}{\sqrt{\sum_{k=1}^{m+n} w_k^2}}\right) \\ &+ \frac{1}{R+1}\Phi\left(\frac{-b - \eta\sum_{i=1}^m w_i - \gamma\sum_{j=m+1}^{m+n} w_j}{\sqrt{\sum_{k=1}^{m+n} w_k^2}}\right),\end{aligned}$$

We can assume w.l.o.g. that $\sum_{k=1}^{m+n} w_k^2 = 1$ as ℓ_{std} is scale-invariant. Hence, by Cauchy-Schwarz,

$$\eta\sum_{i=1}^m w_i + \gamma\sum_{j=m+1}^{m+n} w_j \le \sqrt{\sum_{k=1}^{m+n} w_k^2}\sqrt{m\eta^2 + n\gamma^2} = \sqrt{m\eta^2 + n\gamma^2}.$$

Further calculating the derivative w.r.t. b, we have

$$b_{\text{std}} = -\frac{\log R}{2\sqrt{m\eta^2 + n\gamma^2}}.$$

The optimal robust classifier. Similarly, for the robust loss, we have

$$\begin{aligned}\ell_{\text{rob},\infty}(w,b) = \frac{R}{R+1}\Phi\left(\frac{b-(\eta\sum_{i=1}^m w_i - \varepsilon\sum_{i=1}^m |w_i|) - (\gamma\sum_{j=m+1}^{m+n} w_j - \varepsilon\sum_{j=m+1}^{m+n}|w_j|)}{\sqrt{\sum_{k=1}^{m+n} w_k^2}}\right) \\ + \frac{1}{R+1}\Phi\left(\frac{-b-(\eta\sum_{i=1}^m w_i - \varepsilon\sum_{i=1}^m |w_i|) - (\gamma\sum_{j=m+1}^{m+n} w_j - \varepsilon\sum_{j=m+1}^{m+n}|w_j|)}{\sqrt{\sum_{k=1}^{m+n} w_k^2}}\right).\end{aligned}$$

We assume w.l.o.g. that $\sum_{k=1}^{m+n} w_k^2 = 1$. Using $\gamma < \varepsilon < \eta$ and Cauchy-Schwarz, we have

$$\left(\eta\sum_{i=1}^m w_i - \varepsilon\sum_{i=1}^m |w_i|\right) + \left(\gamma\sum_{j=m+1}^{m+n} w_j - \varepsilon\sum_{j=m+1}^{m+n}|w_j|\right) \le (\eta-\varepsilon)\sum_{i=1}^m |w_i| \le (\eta-\varepsilon)\sqrt{m},$$

and the inequalities are achieved when $w_1 = \cdots = w_m = \frac{1}{\sqrt{m}}$, and $w_{m+1} = \cdots = w_{m+n} = 0$. Further calculating the derivative w.r.t. b, we have

$$b_{\text{rob},\infty} = -\frac{\log R}{2(\eta-\varepsilon)\sqrt{m}}.$$

B. Omitted Proofs from Section 4

Throughout this section, we will use Φ_α to denote the cumulative distribution function of the standard $S\alpha S$ distribution $f(x;\alpha,1,0)$.

B.1. Detailed Analysis of Subsection 4.2 — Multivariate $S\alpha S$ Distribution with Independent Components

Using the "closed under linear transformation" property, we have

$$\ell_{\text{std}}(w,b) = \frac{1}{2}\Phi_\alpha\left(\frac{b+w^\top\theta^-}{\|w\|_\alpha}\right) + \frac{1}{2}\Phi_\alpha\left(\frac{-b-w^\top\theta^+}{\|w\|_\alpha}\right)$$

for the standard loss and

$$\ell_{\text{rob},p}(w,b) = \frac{1}{2}\Phi_\alpha\left(\frac{b+w^\top\theta^- + \varepsilon\|w\|_q}{\|w\|_\alpha}\right) + \frac{1}{2}\Phi_\alpha\left(\frac{-b-w^\top\theta^+ + \varepsilon\|w\|_q}{\|w\|_\alpha}\right)$$

for the robust ℓ_p loss. Following the same procedure as in Section 3, we assume w.l.o.g. that $\|w_{\text{std}}\|_\alpha = \|w_{\text{rob},\infty}\|_\alpha = 1$. We use $v_1 \propto v_2$ to describe two vectors v_1 and v_2 differing by some *positive* constant coordinate-wisely.

Analysis of w. Introducing the Lagrangians and the KKT conditions give us

$$\partial\|w_{\text{std}}\|_\alpha \propto \theta^+ - \theta^-. \tag{5}$$

Similarly, for the robust ℓ_p loss, we have

$$\partial\|w_{\text{rob},p}\|_\alpha \propto \theta^+ - \theta^- - 2\varepsilon\partial\|w_{\text{rob}}\|_q. \tag{6}$$

Analysis of b. Fixing some optimal w with $\|w\|_\alpha = 1$, we can take partial derivatives w.r.t. b, and obtain

$$\frac{\partial\ell_{\text{std}}(w,b)}{\partial b} = \frac{1}{2}\varphi_\alpha(b+w^\top\theta^-) - \frac{1}{2}\varphi_\alpha(-b-w^\top\theta^-),$$

where we use φ_α to denote the probability density function of $f(x;\alpha,1,0)$. Since φ_α is symmetric and monotonically decreasing on $(0,\infty)$ (see Theorem 1 in (Gawronski, 1984)), we have either

$$b_{\text{std}} + w_{\text{std}}^\top\theta^- = -b_{\text{std}} - w_{\text{std}}^\top\theta^+ \quad \text{or} \quad b_{\text{std}} + w_{\text{std}}^\top\theta^- = b_{\text{std}} + w_{\text{std}}^\top\theta^+,$$

and the latter is impossible due to Eq. (5) and Danskin's theorem. As a consequence, we have $b_{\text{std}} = -\frac{w_{\text{std}}^\top(\theta^+ + \theta^-)}{2}$, and similarly $b_{\text{rob},p} = -\frac{w_{\text{rob},p}^\top(\theta^+ + \theta^-)}{2}$.

Putting together. Combining the analysis for w and b, we have

$$\ell^+_{\text{std}}(w_{\text{std}}, b_{\text{std}}) = \ell^-_{\text{std}}(w_{\text{std}}, b_{\text{std}}) = \Phi_\alpha\left(-\frac{w^\top_{\text{std}}(\theta^+ - \theta^-)}{2}\right)$$

and

$$\ell^+_{\text{nat}}(w_{\text{rob},p}, b_{\text{rob},p}) = \ell^-_{\text{nat}}(w_{\text{rob},p}, b_{\text{rob},p}) = \Phi_\alpha\left(-\frac{w^\top_{\text{rob},p}(\theta^+ - \theta^-)}{2}\right).$$

To understand whether enforcing adversarial robustness provably hurt standard accuracy, it suffices to check whether the inequality

$$w^\top_{\text{rob},p}(\theta^+ - \theta^-) < w^\top_{\text{std}}(\theta^+ - \theta^-)$$

holds. Note under the constraint $\|w\|_\alpha = 1$, w_{std} and $w_{\text{rob},p}$ solve the following optimization problems respectively:

$$w_{\text{std}} = \arg\max_w w^\top(\theta^+ - \theta^-), \quad w_{\text{rob},p} = \arg\max_w w^\top(\theta^+ - \theta^-) - 2\varepsilon\|w\|_q.$$

We can then prove Theorem 4.1 and 4.3 based on such formulation.

Proof of Theorem 4.1. We will discuss the two corner cases separately.

Case 1. When $q = \alpha$, $w_{\text{rob},p}$ has the same optimization formulation as w_{std} since $\|w\|_q = \|w\|_\alpha = 1$ is a constant, so they have the same optimal values.

Case 2. Suppose $\bar\theta := \theta^+ - \theta^-$ is isotropic, *i.e.*, $|\bar\theta| \parallel 1_d$ and $\alpha \geq 1$. By Hölder's inequality, we have $w_{\text{std}} \parallel \bar\theta$. For $w_{\text{rob},p}$, denote $C = |\bar\theta_1|$, then

$$\begin{aligned} w^\top(\theta^+ - \theta^-) - 2\varepsilon\|w\|_q &\leq (C - 2\varepsilon C_1)\|w\|_1 \\ &\leq C_2(C - 2\varepsilon C_1)\|w\|_\alpha \\ &= C_2(C - 2\varepsilon C_1) \end{aligned}$$

by Hölder's inequality, where $C_1 = d^{\frac{1}{q}-1}$, $C_2 = d^{\frac{1}{\alpha}-1}$ and

$$C - 2\varepsilon C_1 \geq C - 2\varepsilon > 0$$

by our assumption on the perturbation radius. The two inequalities hold simultaneously when $w_{\text{rob},p} \parallel \bar\theta$. As a consequence, both the optimal standard and robust classifier achieve the same value. ■

Proof of Theorem 4.3. Denote the dual index of α as α', then by Hölder's inequality,

$$w^\top(\theta^+ - \theta^-) \leq \|w\|_\alpha\|\theta^+ - \theta^-\|_{\alpha'} = \|\theta^+ - \theta^-\|_{\alpha'},$$

and there is exactly one w with $\|w\|_\alpha = 1$ that makes the equality hold. As a consequence, to show

$$w^\top_{\text{rob},p}(\theta^+ - \theta^-) < w^\top_{\text{std}}(\theta^+ - \theta^-)$$

it suffices to show $w_{\text{std}} \nparallel w_{\text{rob},p}$, under the conditions listed in the theorem statement. In fact, if $w_{\text{std}} \parallel w_{\text{rob},p}$ (we use w to represent the corresponding direction), then by the KKT conditions, we have

$$\partial\|w\|_q \parallel \partial\|w\|_\alpha,$$

implying that $|w_1| = \cdots = |w_d|$ (since $\alpha > 1$, $1 < q < \infty$ and $q \neq \alpha$). By the KKT condition of the standard classifier, we have

$$\partial\|w\|_\alpha \propto \bar\theta$$

which further implies that $|\bar\theta| \parallel 1_d$, and this is already precluded by the assumption in the theorem statement. As a consequence, we must have $w_{\text{std}} \nparallel w_{\text{rob},p}$, and

$$w^\top_{\text{rob},p}(\theta^+ - \theta^-) < w^\top_{\text{std}}(\theta^+ - \theta^-).$$

■

B.2. Detailed analysis of Subsection 4.3

We will now analyze the intercept b and the slope w respectively.

Analysis of b. Following the same procedure as in Section 3 and the previous subsection, we assume w.l.o.g. that $\|w_{\text{std}}\|_1 = \|w_{\text{rob},\infty}\|_1 = 1$, then the standard loss and the robust ℓ_∞ loss write as

$$\ell_{\text{std}}(w,b) = \frac{R}{R+1}\Phi_1\left(b + w^\top\theta^-\right) + \frac{1}{R+1}\Phi_1\left(-b - w^\top\theta^+\right)$$

and

$$\ell_{\text{rob},\infty}(w,b) = \frac{R}{R+1}\Phi_1\left(b + w^\top\theta^- + \varepsilon\right) + \frac{1}{R+1}\Phi_1\left(-b - w^\top\theta^+ + \varepsilon\right).$$

Fixing w with $\|w\|_1 = 1$, we are interested in finding the optimal b that minimizes ℓ_{std} and $\ell_{\text{rob},\infty}$. Taking partial derivatives w.r.t. b and simplifying the expressions, we have

$$\text{sgn}\left(\frac{\partial \ell_{\text{std}}(w,b)}{\partial b}\right) = \text{sgn}\left(q_1(b)\right), \quad \text{sgn}\left(\frac{\partial \ell_{\text{rob},\infty}(w,b)}{\partial b}\right) = \text{sgn}\left(q_2(b)\right),$$

where $q_1(b)$ and $q_2(b)$ are two quadratic functions defined through

$$q_1(b) := (R-1)b^2 + \left(2Rw^\top\theta^+ - 2Rw^\top\theta^-\right)b + \left(R(w^\top\theta^+)^2 - (w^\top\theta^-)^2 + R - 1\right)$$

and

$$\begin{aligned} q_2(b) := {} & (R-1)b^2 + \left(2Rw^\top\theta^+ - 2Rw^\top\theta^- - (2R+2)\varepsilon\right)b \\ & + \left(R(w^\top\theta^+)^2 - (w^\top\theta^-)^2 + R - 1 + (R-1)\varepsilon^2 - 2Rw^\top\theta^+\varepsilon - 2w^\top\theta^-\varepsilon\right), \end{aligned}$$

whose discriminants are given by

$$\Delta_1 := R\left(w^\top\theta^+ - w^\top\theta^-\right)^2 - (R-1)^2$$

and

$$\Delta_2 = R\left(w^\top\theta^+ - w^\top\theta^- - 2\varepsilon\right)^2 - (R-1)^2$$

respectively. The expressions immediately lead us to the following proposition.

Proposition B.1. *When the class imbalance ratio satisfies*

$$R \geq 2 + 4\|\theta^+ - \theta^-\|_\infty^2,$$

we have $\Delta_1, \Delta_2 < 0$, implying that the quadratic functions $q_1(b)$ and $q_2(b)$ are always positive.

Proof. By Hölder's inequality, we have

$$|w^\top\theta^+ - w^\top\theta^-| \leq \|w\|_1\|\theta^+ - \theta^-\|_\infty = \|\theta^+ - \theta^-\|_\infty.$$

The result then follows from the assumption on ε, as well as the fact that

$$\frac{(R-1)^2}{R} \geq R + 2.$$

■

Without further digging into the analysis of w, Proposition B.1 itself is sufficient to characterize the behavior of the optimal standard classifier as well as the robust ℓ_∞ classifier — since the quadratic functions are always positive and have the same signs as the partial derivatives, the optimal value is attained at $b = -\infty$ for both losses. Theorem 4.5 is a direct consequence of such observation.

Analysis of w. We now switch to analyzing w. Following similar procedures as in Section 3 and the previous subsection, we can introduce the Lagrangians, and the KKT conditions reveal that both w_{std} and $w_{\text{rob},\infty}$ satisfy the relation

$$\partial\|w\|_1 \propto \theta^+ - \theta^-.$$

This further implies

$$\langle w_{\text{std}}, \theta^+ - \theta^- \rangle = \langle w_{\text{rob},\infty}, \theta^+ - \theta^- \rangle = \|\theta^+ - \theta^-\|_\infty. \tag{7}$$

Now

$$\Delta_1 = R\|\theta^+ - \theta^-\|_\infty^2 - (R-1)^2 \quad \Delta_2 = R\left(\|\theta^+ - \theta^-\|_\infty - 2\varepsilon\right)^2 - (R-1)^2,$$

and we assume they are both non-negative. Plugging the optimal b_{std} and $b_{\text{rob},\infty}$ (the larger root of the quadratic functions) back to the standard loss, we have

$$\ell^+_{\text{std}}(w_{\text{std}}, b_{\text{std}}) = \Phi_1\left(\frac{-\|\theta^+ - \theta^-\|_\infty - \sqrt{\Delta_1}}{R-1}\right), \quad \ell^-_{\text{std}}(w_{\text{std}}, b_{\text{std}}) = \Phi_1\left(\frac{-R\|\theta^+ - \theta^-\|_\infty + \sqrt{\Delta_1}}{R-1}\right), \tag{8}$$

whereas

$$\begin{aligned} \ell^+_{\text{std}}(w_{\text{rob},\infty}, b_{\text{rob},\infty}) &= \Phi_1\left(\frac{-\|\theta^+ - \theta^-\|_\infty - (R+1)\varepsilon - \sqrt{\Delta_2}}{R-1}\right) \\ \ell^-_{\text{std}}(w_{\text{rob},\infty}, b_{\text{rob},\infty}) &= \Phi_1\left(\frac{-R\|\theta^+ - \theta^-\|_\infty + (R+1)\varepsilon + \sqrt{\Delta_2}}{R-1}\right). \end{aligned} \tag{9}$$

Define

$$d(\varepsilon) := (R+1)\varepsilon + \sqrt{\Delta_2},$$

then it is straightforward to see that $d(0) = \sqrt{\Delta_1}$. By comparing Eq. (8) and (9), we know that it is essentially the relation between $d(\varepsilon)$ and $d(0)$ that determines the change of accuracy disparity; specifically, to prove Theorem 4.7 it suffices to show $d(\varepsilon) < d(0)$.

Proof of Theorem 4.7. Calculating the derivative, we have

$$d'(s) = R + 1 - \frac{2R}{\sqrt{R - \left(\frac{R-1}{\|\theta^+ - \theta^-\|_\infty - 2s}\right)^2}}.$$

When $s \le \frac{\kappa}{2}\|\theta^+ - \theta^-\|_\infty$, it is straightforward to see that $d'(s) < 0$, hence $d(\varepsilon) < d(0)$ by our assumption on the perturbation radius.

■

B.3. Elliptically-Contoured Multivariate $S\alpha S$ Distribution

Here we will introduce another multivariate $S\alpha S$ distribution whose joint characteristic function has closed form.

Elliptically-Contoured. We say a multivariate distribution X is elliptically-contoured (Samorodnitsky et al., 1996), if it has joint characteristic function

$$\mathbb{E}\exp\left(is^\top X\right) = \exp\{-(s^\top \Sigma s)^{\alpha/2} + is^\top \mu\},$$

where Σ is a positive semi-definite *shape matrix* and μ is the location parameter. We use $S\alpha S_{EC}(\mu, \Sigma)$ to denote such elliptically-contoured multivariate $S\alpha S$ distribution.

Suppose the data are generated through a mixture of elliptically-contoured multivariate $S\alpha S$ distributions: $\mathcal{P}^+ = S\alpha S_{EC}(\theta^+, \Sigma)$ and $\mathcal{P}^- = S\alpha S_{EC}(\theta^-, \Sigma)$. We will now show in the following theorem that, since the shape matrix is an analogy to the covariance matrix as in the Gaussian case, we can obtain a similar conclusion of Corollary 3.3 when the data are drawn from a mixture of elliptically-contoured $S\alpha S$ distribution.

Theorem B.2. *Suppose the data are generated through a mixture of elliptically-contoured multivariate $S\alpha S$ distributions. When there is no class imbalance, i.e., $R = 1$, enforcing adversarial robustness with ℓ_p-constraint will increase the error on both classes, so long as the "direction" of the optimal robust classifier is not parallel to its counterpart, i.e., $\Sigma^{\frac{1}{2}} w_{\text{std}} \nparallel \Sigma^{\frac{1}{2}} w_{\text{rob},p}$. Furthermore, if the shape matrix Σ is positive definite, then the requirement $\Sigma^{\frac{1}{2}} w_{\text{std}} \nparallel \Sigma^{\frac{1}{2}} w_{\text{rob},p}$ is equivalent to $w_{\text{std}} \nparallel w_{\text{rob},p}$, i.e., a shift in direction in the standard sense.*

Proof of Theorem B.2. Note for $X \sim S\alpha S_{EC}(\mu, \Sigma), w \in \mathbb{R}^d$ and arbitrary $t \in \mathbb{R}$, we can set $s = tw$, and the definition of elliptically-contoured multivariate $S\alpha S$ distribution gives us $\mathbb{E} \exp\left(itw^\top X\right) = \exp\{-|t\sqrt{w^\top \Sigma w}|^\alpha + itw^\top \mu\}$, so according to the definition of univariate $S\alpha S$ distribution,

$$w^\top X \sim f(x; \alpha, \sqrt{w^\top \Sigma w}, w^\top \mu).$$

Therefore, the "closed under linear transformation" property yields

$$\ell_{\text{std}}(w, b) = \frac{1}{2} \Phi_\alpha \left(\frac{b + w^\top \theta^-}{\sqrt{w^\top \Sigma w}} \right) + \frac{1}{2} \Phi_\alpha \left(\frac{-b - w^\top \theta^+}{\sqrt{w^\top \Sigma w}} \right),$$

and similarly

$$\ell_{\text{rob},p}(w, b) = \frac{1}{2} \Phi_\alpha \left(\frac{b + w^\top \theta^- + \varepsilon \|w\|_q}{\sqrt{w^\top \Sigma w}} \right) + \frac{1}{2} \Phi_\alpha \left(\frac{-b - w^\top \theta^+ + \varepsilon \|w\|_q}{\sqrt{w^\top \Sigma w}} \right).$$

The only difference compared to the Gaussian case discussed in Section 3 is that the cumulative function is replaced by Φ_α; moreover, a closer examination of the proof of Theorem 3.1 (see Appendix A) shows that we only resort to the symmetry and monotonicity of Φ, and these properties are still preserved in Φ_α. As a consequence, we obtain Theorem B.2 as desired. ■

C. Additional Experimental Details, Results, and Analyses

C.1. Omitted Details of Experimental Setup

In this part, we provide a concrete description of our experimental setups. Additionally, we release our code at `https://github.com/Accuracy-Disparity/AT-on-AD`, where we include the datasets, the code, and the instructions for reproducing the experiments.

Datasets. We altogether experiment with seven groups of datasets, including three groups of synthetic datasets and four groups of real-world datasets. Here, for ease of reference, we refer to the balanced and imbalanced datasets ($R = 1, 2, 5, 10$) constructed from the original balanced dataset as a *group* of dataset, which consists of four datasets corresponding to four choices of R.

We will first describe the dataset properties of the original balanced dataset, and then explain how we construct the imbalanced datasets in the dataset group. In the end, we introduce the training, validation, and test split for performing training and evaluation.

Dataset Properties. The three balanced **synthetic** datasets are constructed to be a mixture of two Gaussian, Cauchy, or other stable distributions with $1 < \alpha < 2$. We set the sample size of both classes as $N = 10,000$ and the sample dimension size as $d = 100$. For the Gaussian case, the two means are sampled from $U[0, 1]^d$ and $U[-1, 0]^d$, and their (same) variances are set as $AA^\top$ where $A \sim \mathcal{N}(0, I_d)$. For the Cauchy case and the $S\alpha S$ stable distribution, we construct the data with *independent components*. In both cases, for each dimension, we sample the location parameter from $U[0, 0.5]$ (or $U[-0.5, 0]$) and set the scale parameter as 1. We set the parameter $\alpha = 1$ to construct the Cauchy dataset and $\alpha = 1.5$ for the other one. We release our synthetic datasets at `https://github.com/Accuracy-Disparity/AT-on-AD`. The four balanced **real-world** datasets are built upon the handwritten digits dataset *MNIST* (LeCun et al., 1998) under the Creative Commons Attribution-Share Alike 3.0 license, the fashion products dataset *Fashion-MNIST* (Xiao et al., 2017) under the MIT license, *CIFAR-10* (Krizhevsky et al., 2009) under the MIT license, and *ImageNet*. MNIST and Fashion-MNIST consist of grey-scale images of dimensionality 28×28; CIFAR-10 consists of colored images of dimensionality $32 \times 32 \times 3$. Originally, the three datasets are used for 10-class classification. To adapt for the binary classification task we consider here, we choose two classes from all ten — digit 1 and digit 7 for MNIST, T-shirt and trouser for Fashion-MNSIT, and cat and dog for CIFAR-10. ImageNet consists of colored images of various dimensionality. We use the downsampled dataset with image dimensionality

$64 \times 64 \times 3$ (Chrabaszcz et al., 2017). The dataset contains $1,000$ classes following a semantic hierarchy[3]. For our binary classification task, we choose two "macro" classes in the hierarchy, "car" and "edible fruit", each consisting of 10 classes. We do not choose two lowest level of classes (among all $1,000$ classes) from ImageNet, since the number of samples per class (around $1,000$) is insufficient for training the deep neural network. By their original construction, these real-world datasets are class balanced.

Construction of the Imbalanced Dataset. We next explain how we obtain the dataset group consisting of datasets of various class imbalance ratios ($R = 2, 5, 10$) from the balanced one ($R = 1$) we initially construct. Concretely, we obtain datasets with increasing class imbalance ratio sequentially. For each dataset, we hold the majority class samples as fixed (*i.e.*, same as the $R = 1$ case), and subsample the minority class samples from the previously constructed set of minority class samples. (For example, we sample from the minority set in the $R = 2$ dataset to obtain the minority set for the $R = 5$ dataset.)

Dataset Partition. For each dataset in each dataset group, we split the dataset into three disjoint partitions: training, validation, and testing. For the synthetic dataset, we set the ratio of the three partitions to be 8:1:1, which gives us a total number of 8000 training samples, 1000 validation samples, and 1000 testing samples for the majority class in each dataset. (The sizes of the three partitions for the minority class are the sizes for the majority class divided by the imbalance ratio R). For the real-world datasets MNIST, Fashion-MNIST and CIFAR, since the datasets are originally split into training and testing, we further split the training set into training and validation with a ratio of 8:1. For MNIST and Fashion-MNIST, the numbers of training, validation, and testing samples for the majority class are 5333, 533, and 1000 respectively. For CIFAR-10, the numbers are 4444, 556, and 1000 respectively. For ImageNet, there are in all 13000 images for each "macro" class in the training set. For the majority class, the numbers of training, validation, and testing samples are 10000, 2000, and 1000 respectively. We release all our datasets at `https://github.com/Accuracy-Disparity/AT-on-AD`.

Models. For most experiments, we mainly adopt simple models on these datasets — linear classifiers on the three synthetic datasets, MNIST, and Fashion-MNIST, as well as networks with two linear layers on CIFAR-10.

The reason we mainly experiment with simple models is that they are sufficiently powerful for the binary classification task, as we can see from the accuracy results in Table 5 and 6 in Appendix C.5. Actually, on CIFAR-10, we also experiment with a more complicated convolutional neural network VGG-11 (Simonyan & Zisserman, 2015); we derive similar conclusions on this complicated network as we did on simple networks. The details can be found in Appendix C.4.

Out of interest for deep neural networks on complicated datasets, we also experiment with ImageNet and apply a deep neural network ResNet18 (He et al., 2016).

Training protocols. We perform *standard training* and *adversarial training* on the binary classification task. We first describe the common training protocols for both training schemes, and then go into details on the specificity of adversarial training.

We adopt stochastic gradient descent (SGD) optimizer (Kiefer & Wolfowitz, 1952) or Adam optimizer (Kingma & Ba, 2015) for network training. We take the cross entropy loss as the training objective. We perform model selection on a held-out validation set (as introduced in the data part). We perform a grid search for the hyper-parameters including learning rate, batch size, and hidden layer size (when applicable) based on the model's performance on the validation set. The search space for learning rate is $\{0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0002, 0.0001\}$, for batch size is $\{32, 64, 128\}$, and for hidden layer size is $\{100, 200, 500, 1000, 2000\}$. For each model, we perform training for a maximum of 500 training epochs; we keep track of the best model throughout the training based on the validation loss and apply early stopping (Prechelt, 1998) when the lowest validation loss does not decrease for the past 50 epochs. For ImageNet specifically, we perform early stopping when the loss does not decrease for the past 10 epochs.

For adversarial training, we follow Goodfellow et al. (Goodfellow et al., 2014) and Madry et al. (Madry et al., 2018) to craft the adversarial examples via the fast gradient method (FGM) or the projected gradient descent (PGD). We adopt the former for linear classifiers and the latter for two layer neural networks and deep neural networks (VGG-11 and ResNet18). For all datasets and all ℓ_p norms ($p \in \{2, \infty\}$), we experiment with three perturbation scales ε. The perturbation scales are selected based on the ℓ_p distances between the empirical means of the two classes. For most of the datasets, we choose the three values as $1/4, 3/8, 1/2$ of the distance. For CIFAR-10 and ℓ_2 only, we select slightly larger perturbation scales, following the practice in Tsipras et al. (Tsipras et al., 2019). For PGD attack specifically, we set the step number to be 50 and the limit on the per step size to be $2.5 \cdot \varepsilon/50$ following Madry et al. (Madry et al., 2018). On ImageNet, we limit the step number to

[3] A diagram of the hierarchy can be found at `https://observablehq.com/@mbostock/imagenet-hierarchy`

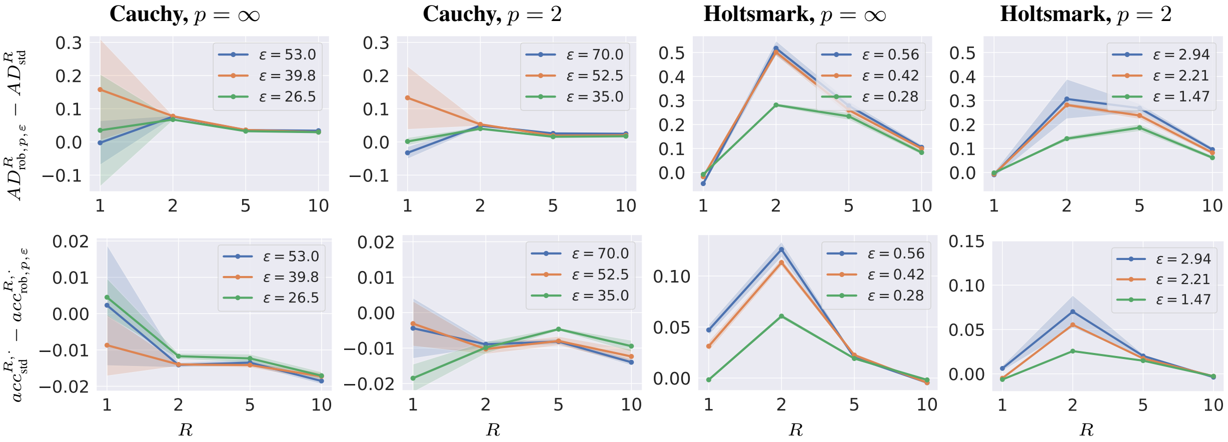

Figure 2. The gap of accuracy disparity $AD^R_{\text{rob},p,\varepsilon} - AD^R_{\text{std}}$ (1st row, RQ2) and the gap of standard accuracy $acc^{R,\cdot}_{\text{std}} - acc^{R,\cdot}_{\text{rob},p,\varepsilon}$ (2nd row, RQ3) w.r.t. the imbalance ratio R. For the robust classifiers, we consider $p \in \{2, \infty\}$ and multiple perturbation scales. We present the results on the two synthetic datasets of $S\alpha S$ distribution where $\alpha = 1$ or 1.5. The shaded area represents the standard error of 5 runs.

10 considering the high computation cost.

C.2. Additional Experiments on Synthetic Datasets of Stable Distributions

As a complement to the Gaussian case evaluated in Section 5 in the main paper, here, we evaluate general symmetric α-stable distributions corresponding to the theoretical results in Section 4. We consider two values of α — the special case $\alpha = 1$ which is commonly known as the Cauchy distribution, as well as the case $\alpha = 1.5$ which is known as the Holtsmark distribution (Holtsmark, 1919) and can be viewed as an "intermediate" between Gaussian ($\alpha = 2$) and Cauchy. The construction details of the dataset are described in Appendix C.1.

We present the results on these two datasets in Figure 2.

In the Cauchy case, the two classes in our constructed dataset are barely separable, so in the balanced case $R = 1$, the classification accuracy for both classes are close to random guess with value around 0.5. (Detailed numbers can be found in Table 3.) In the class imbalance settings ($R = 2, 5, 10$), the predictions of both the standard and the robust classifiers favor the majority class, and so the accuracy disparity is close to 1 in both cases, leading to the accuracy disparity gap of nearly 0. This result provides the answers to RQ1 and RQ2 and aligns well with the theoretical result in Theorem 4.5. Regarding RQ3 which asks whether and when adversarial training will worsen the standard accuracy, we comment that on this Cauchy dataset we investigate, the classification outcomes are similar for the standard and robust classifiers. In the balanced class setting, the hardness of the dataset dominates and leads to random guesses in both training scenarios; in the class imbalance setting, the class imbalance dominates and leads to unanimous preference towards the majority class. Thus, it is difficult to draw conclusions regarding the factor of adversarial training.

For the Holtsmark distribution, unsurprisingly, we find that the phenomenon is in the between of the Guassian and Cauchy in terms of accuracy disparity. Concretely, we do see that adversarial training will exacerbate the accuracy disparity in the class imbalance setting, but the gap does not increase with the imbalance ratio R. This observation aligns partially with Theorem 3.6 and partially with Theorem 4.5, but not both. In terms of the standard accuracy, we see that adversarial training leads to a decrease of standard accuracy, which is consistent with Theorem 4.3.

In all, from these experiments on synthetic datasets of stable distributions, we see that 1) class imbalance will be a dominating factor which surpasses the influence of adversarial training for the Cauchy case; 2) α value between 1 and 2 will lead to intermediate behavior of Gaussian (Theorem 3.6) and Cauchy (Theorem 4.5) regarding the accuracy disparity; 3) adversarial training will invariably decrease the standard accuracy empirically.

C.3. Additional Experiments on Real-World Datasets Fashion-MNIST and ImageNet

We evaluate two additional real-world datasets Fashion-MNIST (Xiao et al., 2017) and ImageNet (Deng et al., 2009). Fashion-MNIST is often used as a drop-in replacement for MNIST (LeCun et al., 1998). ImageNet is a large-scale dataset

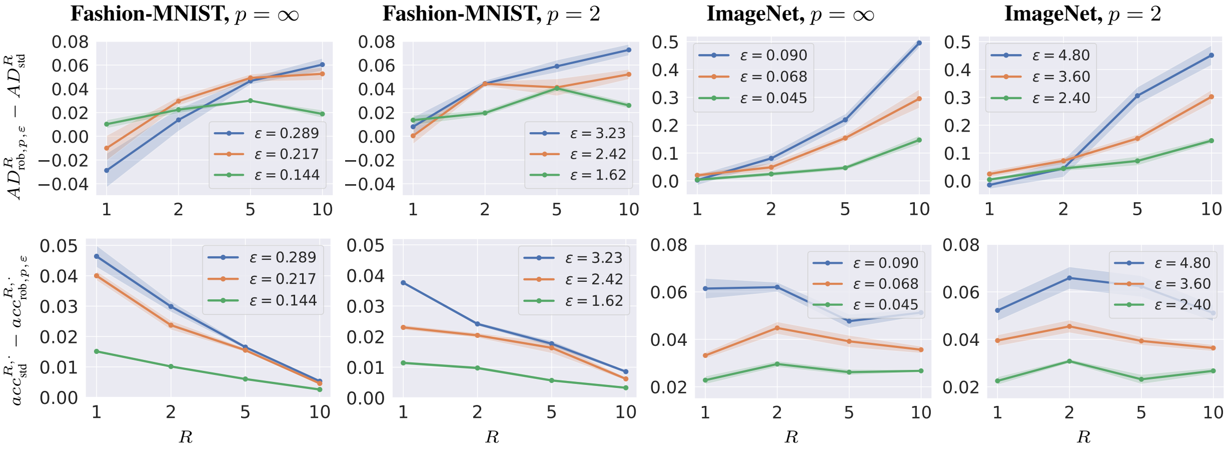

Figure 3. The gap of accuracy disparity $AD^R_{\text{rob},p,\varepsilon} - AD^R_{\text{std}}$ (1st row, RQ2) and the gap of standard accuracy $acc^{R,\cdot}_{\text{std}} - acc^{R,\cdot}_{\text{rob},p,\varepsilon}$ (2nd row, RQ3) w.r.t. the imbalance ratio R. For the robust classifiers, we consider $p \in \{2, \infty\}$ and multiple perturbation scales. We present the results on two real-world datasets **Fashion-MNIST** and **ImageNet**. The shaded area represents the standard error of 5 runs.

consisting of $1,000$ classes of high-dimensional images. The dataset description and the details on the construction of the dataset are provided in Appendix C.1.

We present the results in Figure 3. Comparing the results with that in Figure 1 in Section 5 of the main paper, we obtain highly similar observations and conclusions w.r.t. all our three research questions. Concretely, both Fashion-MNIST and ImageNet display similar behavior with the Gaussian mixture case — adversarial training exacerbates accuracy disparity compared with standard training which is more severe with increased imbalance (1st row), and adversarial training worsens the standard accuracy when $R = 1$ (2nd row).

In order to understand how "close" these two real-world datasets are to the Gaussian mixture, we follow the approach described in Appendix C.4 to compute the outlier ratio (*i.e.*, the ratio to samples that are 1 sigma away from the empirical mean after performing preconditioning). The outlier ratios are 0.30 and 0.26 for the two classes of Fashion-MNIST, and 0.00 and 0.00 for ImageNet. Compared with the outlier ratios 0.52 and 0.52 for CIFAR (see Appendix C.4), Fashion-MNIST and ImageNet are indeed less heavy-tailed.

Thus, we show that Fashion-MNIST and ImageNet are additional evidences for the potential of extending the theoretical results to real-world datasets that can be roughly modeled as Gaussian mixtures.

C.4. Additional Analysis on the Statistical Properties of Real-World Datasets

From Section 5.1 in the main paper, we see that the results on MNIST resemble the theoretical analysis on the Gaussian distribution, while the results on CIFAR resemble the analysis on the Cauchy distribution. In order to understand whether CIFAR is indeed more heavy-tailed than MNIST, we look into the statistical properties of the two datasets and make a comparison.

We compute the ratio of the outlier of the dataset as a proxy of how "heavy" the tail is. Concretely, we first perform a preconditioning on the dataset such that the covariance of the preconditioned dataset becomes an identity matrix. We leverage the PCA whitening approach (Friedman, 1987) to achieve the goal. Then, we compute an empirical mean of the dataset and check for how many instances are 1 sigma away from the empirical mean (*i.e.*, the distance between the instance and the empirical mean is larger than $\sqrt{d}$, where d is the dataset dimensionality).

We follow the above approach to compute the outlier ratio for two classes separately in both datasets. The outlier ratios are as high as 0.52 and 0.52 for the two classes of CIFAR, while only 0.13 and 0.16 for MNIST. This means that CIFAR is indeed much more heavy-tailed than MNIST, supporting the experimental results in Section 5.1.

Ruling out Other Possible Influencing Factors. As we can observe from Table 4, the reason why there is only small accuracy disparity gap for CIFAR in the class imbalance case is that the accuracy of the standard classifier on the minority class is close to 0. The heavy tailed property is one possible explanation (Theorem 4.5); the other straightforward hypothesis is that the standard classifier is not well trained on CIFAR, either because the dataset size is relatively small, or because the model capacity is limited. We then separately study the influence of the two factors:

- **Dataset Size.** Since we cannot enlarge the dataset size for CIFAR, we instead shrink the dataset size for MNIST. We use 53 majority class samples and 5 minority class samples (*i.e.*, $R = 10$) to train the model, which is $1/100$ of the size of the original training data. Compared with the original result, the majority class accuracy remains 1.00, and the minority class accuracy drops from 0.97 to 0.88.
- **Model Capacity.** Instead of the original two linear layer network, we adopt a deep convolutional network VGG-11 (Simonyan & Zisserman, 2015) to train the standard classifier on the $R = 10$ case for CIFAR. As a result, the minority class accuracy increases from 0.00 to 0.23.

From the above results, we see that neither shrinking the dataset size nor increasing the model capacity can significantly impact the accuracy disparity gap. Thus, we can confidently rule out these alternative explanations. We conclude that the main contributor that leads to the distinction is the distributional characteristic (specifically, the tail property) of the dataset.

C.5. Full Experimental Results

In this part, we present the full experimental results on seven dataset groups—accuracy disparity gap in Table 1, standard accuracy gap in Table 2, per class accuracy in Table 3 and 4, and overall accuracy in Table 5 and 6. In each table, we present results for the standard and robust classifiers (with various p and ε) on datasets with different imbalance ratios R.

C.6. Computational Resources and Runtime

We perform experiments on a machine with AMD EPYC 7352 24-Core Processor CPU and 8 NVIDIA RTX A6000 GPUs. The computational costs for training both the standard and the robust classifiers for both the synthetic datasets and the real world datasets are low. For adversarial training on CIFAR (the most expensive case), each run of training would take less than 20 minutes. We parallel the training tasks on all 8 GPU cards.

Table 1. **Accuracy disparity gap** $AD_R^{\text{rob},p,\varepsilon} - AD_R^{\text{std}}$ for various choices of p and ε on seven dataset groups. Results are averaged over 5 runs with different random seeds. (Corresponding to 1st rows of Figures 1, 2 and 3.)

Synthetic Gaussian	$p=\infty$			$p=2$		
	$\varepsilon=1.00$	$\varepsilon=0.75$	$\varepsilon=0.50$	$\varepsilon=5.00$	$\varepsilon=3.50$	$\varepsilon=2.50$
$R=1$	0.00 ± 0.01	-0.02 ± 0.00	-0.01 ± 0.00	0.01 ± 0.02	-0.01 ± 0.02	-0.02 ± 0.00
$R=2$	0.16 ± 0.06	0.20 ± 0.02	0.08 ± 0.00	0.28 ± 0.12	0.33 ± 0.10	0.21 ± 0.02
$R=5$	0.85 ± 0.02	0.56 ± 0.02	0.22 ± 0.01	0.98 ± 0.00	0.91 ± 0.01	0.59 ± 0.02
$R=10$	0.95 ± 0.00	0.73 ± 0.02	0.27 ± 0.01	1.00 ± 0.00	0.97 ± 0.00	0.67 ± 0.01

Synthetic Cauchy	$p=\infty$			$p=2$		
	$\varepsilon=53.00$	$\varepsilon=39.75$	$\varepsilon=26.50$	$\varepsilon=70.0$	$\varepsilon=52.5$	$\varepsilon=35.0$
$R=1$	-0.03 ± 0.02	0.13 ± 0.10	0.00 ± 0.01	-0.00 ± 0.07	0.16 ± 0.15	0.03 ± 0.17
$R=2$	0.05 ± 0.01	0.05 ± 0.00	0.04 ± 0.01	0.08 ± 0.00	0.08 ± 0.00	0.07 ± 0.00
$R=5$	0.03 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.04 ± 0.00	0.04 ± 0.00	0.03 ± 0.00
$R=10$	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00

Synthetic Holtsmark	$p=\infty$			$p=2$		
	$\varepsilon=0.56$	$\varepsilon=0.42$	$\varepsilon=0.28$	$\varepsilon=2.94$	$\varepsilon=2.21$	$\varepsilon=1.47$
$R=1$	-0.01 ± 0.01	-0.01 ± 0.00	-0.00 ± 0.00	-0.05 ± 0.01	-0.02 ± 0.00	-0.01 ± 0.00
$R=2$	0.31 ± 0.08	0.28 ± 0.01	0.14 ± 0.01	0.52 ± 0.03	0.50 ± 0.02	0.28 ± 0.01
$R=5$	0.27 ± 0.02	0.24 ± 0.02	0.19 ± 0.02	0.28 ± 0.02	0.26 ± 0.02	0.23 ± 0.02
$R=10$	0.10 ± 0.01	0.08 ± 0.01	0.06 ± 0.01	0.10 ± 0.01	0.10 ± 0.01	0.08 ± 0.01

MNIST	$p=\infty$			$p=2$		
	$\varepsilon=0.43$	$\varepsilon=0.33$	$\varepsilon=0.22$	$\varepsilon=2.70$	$\varepsilon=2.02$	$\varepsilon=1.35$
$R=1$	0.03 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.03 ± 0.00	0.02 ± 0.00
$R=2$	0.03 ± 0.00	0.03 ± 0.00	0.01 ± 0.00	0.05 ± 0.00	0.04 ± 0.01	0.02 ± 0.00
$R=5$	0.07 ± 0.00	0.04 ± 0.00	0.02 ± 0.00	0.07 ± 0.01	0.06 ± 0.00	0.05 ± 0.00
$R=10$	0.09 ± 0.00	0.04 ± 0.00	0.02 ± 0.00	0.09 ± 0.00	0.08 ± 0.01	0.05 ± 0.00

Fashion-MNIST	$p=\infty$			$p=2$		
	$\varepsilon=0.29$	$\varepsilon=0.22$	$\varepsilon=0.14$	$\varepsilon=3.23$	$\varepsilon=2.42$	$\varepsilon=1.62$
$R=1$	0.01 ± 0.01	0.00 ± 0.01	0.01 ± 0.00	-0.03 ± 0.01	-0.01 ± 0.01	0.01 ± 0.00
$R=2$	0.04 ± 0.00	0.04 ± 0.00	0.02 ± 0.00	0.01 ± 0.01	0.03 ± 0.00	0.02 ± 0.00
$R=5$	0.06 ± 0.01	0.04 ± 0.01	0.04 ± 0.00	0.05 ± 0.00	0.05 ± 0.00	0.03 ± 0.00
$R=10$	0.07 ± 0.00	0.05 ± 0.00	0.03 ± 0.00	0.06 ± 0.01	0.05 ± 0.01	0.02 ± 0.00

CIFAR	$p=\infty$			$p=2$		
	$\varepsilon=0.030$	$\varepsilon=0.023$	$\varepsilon=0.015$	$\varepsilon=0.98$	$\varepsilon=0.88$	$\varepsilon=0.78$
$R=1$	-0.63 ± 0.19	0.01 ± 0.25	-0.31 ± 0.09	-0.30 ± 0.18	-0.11 ± 0.03	-0.07 ± 0.05
$R=2$	0.43 ± 0.03	0.43 ± 0.03	0.23 ± 0.09	0.43 ± 0.03	0.19 ± 0.04	-0.01 ± 0.04
$R=5$	0.06 ± 0.01	0.06 ± 0.01	0.06 ± 0.01	0.06 ± 0.01	0.06 ± 0.01	-0.02 ± 0.01
$R=10$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	-0.01 ± 0.01

ImageNet	$p=\infty$			$p=2$		
	$\varepsilon=0.090$	$\varepsilon=0.068$	$\varepsilon=0.045$	$\varepsilon=4.80$	$\varepsilon=3.60$	$\varepsilon=2.40$
$R=1$	-0.01 ± 0.01	0.02 ± 0.01	0.00 ± 0.01	0.00 ± 0.02	0.02 ± 0.01	0.00 ± 0.01
$R=2$	0.04 ± 0.03	0.07 ± 0.01	0.04 ± 0.01	0.08 ± 0.02	0.05 ± 0.02	0.02 ± 0.01
$R=5$	0.31 ± 0.03	0.15 ± 0.01	0.07 ± 0.02	0.22 ± 0.02	0.15 ± 0.01	0.05 ± 0.01
$R=10$	0.45 ± 0.04	0.30 ± 0.02	0.14 ± 0.01	0.50 ± 0.02	0.30 ± 0.03	0.15 ± 0.02

Table 2. **Standard accuracy gap** $acc_R^{\text{std}} - acc_R^{\text{rob},p,\varepsilon}$ for various choices of p and ε on seven dataset groups. Results are averaged over 5 runs with different random seeds. (Corresponding to 2nd rows of Figures 1, 2 and 3.)

Synthetic Gaussian	$p = \infty$			$p = 2$		
	$\varepsilon = 1.00$	$\varepsilon = 0.75$	$\varepsilon = 0.50$	$\varepsilon = 5.00$	$\varepsilon = 3.50$	$\varepsilon = 2.50$
$R = 1$	0.16 ± 0.00	0.12 ± 0.00	0.05 ± 0.00	0.26 ± 0.00	0.20 ± 0.00	0.13 ± 0.00
$R = 2$	0.15 ± 0.01	0.11 ± 0.00	0.05 ± 0.00	0.25 ± 0.01	0.20 ± 0.00	0.12 ± 0.00
$R = 5$	0.14 ± 0.00	0.10 ± 0.00	0.04 ± 0.00	0.16 ± 0.00	0.15 ± 0.00	0.10 ± 0.00
$R = 10$	0.09 ± 0.00	0.07 ± 0.00	0.03 ± 0.00	0.09 ± 0.00	0.09 ± 0.00	0.06 ± 0.00

Synthetic Cauchy	$p = \infty$			$p = 2$		
	$\varepsilon = 53.00$	$\varepsilon = 39.75$	$\varepsilon = 26.50$	$\varepsilon = 70.0$	$\varepsilon = 52.5$	$\varepsilon = 35.0$
$R = 1$	-0.00 ± 0.01	-0.00 ± 0.01	-0.02 ± 0.00	0.00 ± 0.02	-0.01 ± 0.01	0.00 ± 0.01
$R = 2$	-0.01 ± 0.00	-0.01 ± 0.00	-0.01 ± 0.00	-0.01 ± 0.00	-0.01 ± 0.00	-0.01 ± 0.00
$R = 5$	-0.01 ± 0.00	-0.01 ± 0.00	-0.00 ± 0.00	-0.01 ± 0.00	-0.01 ± 0.00	-0.01 ± 0.00
$R = 10$	-0.01 ± 0.00	-0.01 ± 0.00	-0.01 ± 0.00	-0.02 ± 0.00	-0.02 ± 0.00	-0.02 ± 0.00

Synthetic Holtsmark	$p = \infty$			$p = 2$		
	$\varepsilon = 0.56$	$\varepsilon = 0.42$	$\varepsilon = 0.28$	$\varepsilon = 2.94$	$\varepsilon = 2.21$	$\varepsilon = 1.47$
$R = 1$	0.01 ± 0.00	-0.00 ± 0.00	-0.01 ± 0.00	0.05 ± 0.01	0.03 ± 0.00	-0.00 ± 0.00
$R = 2$	0.07 ± 0.02	0.06 ± 0.00	0.03 ± 0.00	0.13 ± 0.01	0.11 ± 0.00	0.06 ± 0.00
$R = 5$	0.02 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00
$R = 10$	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00

MNIST	$p = \infty$			$p = 2$		
	$\varepsilon = 0.43$	$\varepsilon = 0.33$	$\varepsilon = 0.22$	$\varepsilon = 2.70$	$\varepsilon = 2.02$	$\varepsilon = 1.35$
$R = 1$	0.02 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.03 ± 0.00	0.02 ± 0.00	0.01 ± 0.00
$R = 2$	0.02 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.03 ± 0.00	0.02 ± 0.00	0.01 ± 0.00
$R = 5$	0.01 ± 0.00	0.00 ± 0.00	-0.00 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
$R = 10$	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

Fashion-MNIST	$p = \infty$			$p = 2$		
	$\varepsilon = 0.29$	$\varepsilon = 0.22$	$\varepsilon = 0.14$	$\varepsilon = 3.23$	$\varepsilon = 2.42$	$\varepsilon = 1.62$
$R = 1$	0.04 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	0.05 ± 0.00	0.04 ± 0.00	0.02 ± 0.00
$R = 2$	0.02 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	0.03 ± 0.00	0.02 ± 0.00	0.01 ± 0.00
$R = 5$	0.02 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.01 ± 0.00
$R = 10$	0.01 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

CIFAR	$p = \infty$			$p = 2$		
	$\varepsilon = 0.030$	$\varepsilon = 0.023$	$\varepsilon = 0.015$	$\varepsilon = 0.98$	$\varepsilon = 0.88$	$\varepsilon = 0.78$
$R = 1$	0.21 ± 0.01	0.18 ± 0.01	0.12 ± 0.05	0.19 ± 0.01	0.14 ± 0.01	-0.02 ± 0.01
$R = 2$	0.04 ± 0.00	0.04 ± 0.00	-0.00 ± 0.02	0.04 ± 0.00	0.02 ± 0.00	-0.01 ± 0.01
$R = 5$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	-0.00 ± 0.00
$R = 10$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	-0.00 ± 0.00

ImageNet	$p = \infty$			$p = 2$		
	$\varepsilon = 0.090$	$\varepsilon = 0.068$	$\varepsilon = 0.045$	$\varepsilon = 4.80$	$\varepsilon = 3.60$	$\varepsilon = 2.40$
$R = 1$	0.05 ± 0.00	0.04 ± 0.00	0.02 ± 0.00	0.06 ± 0.00	0.03 ± 0.00	0.02 ± 0.00
$R = 2$	0.07 ± 0.00	0.05 ± 0.00	0.03 ± 0.00	0.06 ± 0.00	0.04 ± 0.00	0.03 ± 0.00
$R = 5$	0.06 ± 0.00	0.04 ± 0.00	0.02 ± 0.00	0.05 ± 0.00	0.04 ± 0.00	0.03 ± 0.00
$R = 10$	0.05 ± 0.00	0.04 ± 0.00	0.03 ± 0.00	0.05 ± 0.00	0.04 ± 0.00	0.03 ± 0.00

Table 3. **Per class accuracy** for the standard and robust classifiers on synthetic datasets. We denote the majority class as "class $-$" and the minority class as "class $+$" following Section 2. "std" refers to the standard classifier and "rob" refers to the robust classifier with the specified p and ε. The presented results are averaged over 5 runs.

Synthetic Gaussian	$R = 1$		$R = 2$		$R = 5$		$R = 10$	
	class $-$	class $+$	class $-$	class $+$	class $-$	class $+$	class $-$	class $+$
std	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
rob ($p = 2, \varepsilon = 5.00$)	0.85 ± 0.01	0.84 ± 0.01	0.89 ± 0.02	0.73 ± 0.05	0.99 ± 0.00	0.15 ± 0.02	1.00 ± 0.00	0.05 ± 0.00
rob ($p = 2, \varepsilon = 3.75$)	0.87 ± 0.00	0.89 ± 0.00	0.95 ± 0.00	0.75 ± 0.01	0.99 ± 0.00	0.43 ± 0.02	1.00 ± 0.00	0.26 ± 0.01
rob ($p = 2, \varepsilon = 2.50$)	0.94 ± 0.00	0.95 ± 0.00	0.98 ± 0.00	0.89 ± 0.00	0.99 ± 0.00	0.77 ± 0.01	0.99 ± 0.00	0.72 ± 0.01
rob ($p = \infty, \varepsilon = 1.00$)	0.74 ± 0.01	0.73 ± 0.01	0.84 ± 0.03	0.56 ± 0.09	1.00 ± 0.00	0.02 ± 0.00	1.00 ± 0.00	0.00 ± 0.00
rob ($p = \infty, \varepsilon = 0.75$)	0.80 ± 0.01	0.80 ± 0.01	0.90 ± 0.03	0.58 ± 0.07	1.00 ± 0.00	0.09 ± 0.01	1.00 ± 0.00	0.02 ± 0.00
rob ($p = \infty, \varepsilon = 0.50$)	0.86 ± 0.00	0.88 ± 0.00	0.95 ± 0.01	0.74 ± 0.01	0.99 ± 0.00	0.40 ± 0.02	0.99 ± 0.00	0.32 ± 0.01

Synthetic Cauchy	$R = 1$		$R = 2$		$R = 5$		$R = 10$	
	class $-$	class $+$	class $-$	class $+$	class $-$	class $+$	class $-$	class $+$
std	0.50 ± 0.00	0.52 ± 0.01	0.95 ± 0.00	0.04 ± 0.00	0.98 ± 0.00	0.02 ± 0.00	0.98 ± 0.00	0.01 ± 0.00
rob ($p = 2, \varepsilon = 70.00$)	0.48 ± 0.01	0.54 ± 0.01	0.98 ± 0.00	0.02 ± 0.00	0.99 ± 0.00	0.00 ± 0.00	0.99 ± 0.00	0.00 ± 0.00
rob ($p = 2, \varepsilon = 52.50$)	0.57 ± 0.05	0.46 ± 0.06	0.98 ± 0.00	0.02 ± 0.00	0.99 ± 0.00	0.01 ± 0.00	0.99 ± 0.00	0.01 ± 0.00
rob ($p = 2, \varepsilon = 35.00$)	0.51 ± 0.00	0.54 ± 0.01	0.98 ± 0.00	0.02 ± 0.00	0.99 ± 0.00	0.01 ± 0.00	0.99 ± 0.00	0.01 ± 0.00
rob ($p = \infty, \varepsilon = 53.00$)	0.49 ± 0.03	0.52 ± 0.04	1.00 ± 0.00	0.01 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00
rob ($p = \infty, \varepsilon = 39.75$)	0.58 ± 0.08	0.45 ± 0.09	1.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00
rob ($p = \infty, \varepsilon = 26.50$)	0.51 ± 0.09	0.50 ± 0.09	0.99 ± 0.00	0.01 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00

Synthetic Holtsmark	$R = 1$		$R = 2$		$R = 5$		$R = 10$	
	class $-$	class $+$	class $-$	class $+$	class $-$	class $+$	class $-$	class $+$
std	0.81 ± 0.00	0.82 ± 0.00	0.94 ± 0.00	0.60 ± 0.00	0.97 ± 0.00	0.27 ± 0.02	0.99 ± 0.00	0.10 ± 0.01
rob ($p = 2, \varepsilon = 2.94$)	0.80 ± 0.01	0.82 ± 0.00	0.97 ± 0.01	0.32 ± 0.07	1.00 ± 0.00	0.02 ± 0.00	1.00 ± 0.00	0.01 ± 0.00
rob ($p = 2, \varepsilon = 2.21$)	0.81 ± 0.00	0.83 ± 0.00	0.97 ± 0.00	0.35 ± 0.01	0.99 ± 0.00	0.05 ± 0.01	1.00 ± 0.00	0.03 ± 0.00
rob ($p = 2, \varepsilon = 1.47$)	0.81 ± 0.00	0.83 ± 0.00	0.95 ± 0.00	0.48 ± 0.01	0.99 ± 0.00	0.10 ± 0.01	0.99 ± 0.00	0.04 ± 0.00
rob ($p = \infty, \varepsilon = 0.56$)	0.74 ± 0.01	0.79 ± 0.00	0.98 ± 0.00	0.13 ± 0.03	1.00 ± 0.00	0.02 ± 0.00	1.00 ± 0.00	0.01 ± 0.00
rob ($p = \infty, \varepsilon = 0.42$)	0.77 ± 0.01	0.80 ± 0.00	0.98 ± 0.00	0.14 ± 0.01	0.99 ± 0.00	0.03 ± 0.01	1.00 ± 0.00	0.01 ± 0.00
rob ($p = \infty, \varepsilon = 0.28$)	0.81 ± 0.00	0.82 ± 0.00	0.97 ± 0.00	0.35 ± 0.00	0.99 ± 0.00	0.05 ± 0.01	0.99 ± 0.00	0.02 ± 0.00

Table 4. **Per class accuracy** for the standard and robust classifiers on real-world datasets. We denote the majority class as "class $-$" and the minority class as "class $+$" following Section 2. "std" refers to the standard classifier and "rob" refers to the robust classifier with the specified p and ε. The presented results are averaged over 5 runs.

MNIST	$R = 1$		$R = 2$		$R = 5$		$R = 10$	
	class $-$	class $+$	class $-$	class $+$	class $-$	class $+$	class $-$	class $+$
std	1.00 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	1.00 ± 0.00	0.97 ± 0.00
rob ($p = 2$, $\varepsilon = 2.70$)	0.99 ± 0.00	0.96 ± 0.00	1.00 ± 0.00	0.95 ± 0.00	1.00 ± 0.00	0.91 ± 0.00	1.00 ± 0.00	0.88 ± 0.00
rob ($p = 2$, $\varepsilon = 2.02$)	1.00 ± 0.00	0.97 ± 0.00	1.00 ± 0.00	0.96 ± 0.00	1.00 ± 0.00	0.94 ± 0.00	1.00 ± 0.00	0.93 ± 0.00
rob ($p = 2$, $\varepsilon = 1.35$)	1.00 ± 0.00	0.98 ± 0.00	1.00 ± 0.00	0.97 ± 0.00	1.00 ± 0.00	0.96 ± 0.00	1.00 ± 0.00	0.95 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.43$)	0.98 ± 0.00	0.96 ± 0.00	0.99 ± 0.00	0.93 ± 0.00	0.99 ± 0.00	0.91 ± 0.01	1.00 ± 0.00	0.87 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.33$)	0.99 ± 0.00	0.96 ± 0.00	1.00 ± 0.00	0.94 ± 0.01	1.00 ± 0.00	0.92 ± 0.00	1.00 ± 0.00	0.88 ± 0.01
rob ($p = \infty$, $\varepsilon = 0.22$)	0.99 ± 0.00	0.97 ± 0.00	0.99 ± 0.00	0.96 ± 0.00	1.00 ± 0.00	0.93 ± 0.00	1.00 ± 0.00	0.92 ± 0.00

Fashion-MNIST	$R = 1$		$R = 2$		$R = 5$		$R = 10$	
	class $-$	class $+$	class $-$	class $+$	class $-$	class $+$	class $-$	class $+$
std	0.98 ± 0.00	0.99 ± 0.00	0.98 ± 0.00	0.99 ± 0.00	0.97 ± 0.00	1.00 ± 0.00	0.95 ± 0.00	1.00 ± 0.00
rob ($p = 2$, $\varepsilon = 3.23$)	0.94 ± 0.00	0.96 ± 0.00	0.92 ± 0.00	0.98 ± 0.00	0.90 ± 0.00	0.99 ± 0.00	0.87 ± 0.00	1.00 ± 0.00
rob ($p = 2$, $\varepsilon = 2.42$)	0.96 ± 0.00	0.97 ± 0.00	0.93 ± 0.00	0.99 ± 0.00	0.92 ± 0.00	0.99 ± 0.00	0.89 ± 0.00	1.00 ± 0.00
rob ($p = 2$, $\varepsilon = 1.62$)	0.97 ± 0.00	0.99 ± 0.00	0.95 ± 0.00	0.99 ± 0.00	0.93 ± 0.00	1.00 ± 0.00	0.92 ± 0.00	1.00 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.29$)	0.95 ± 0.00	0.93 ± 0.01	0.94 ± 0.01	0.97 ± 0.00	0.91 ± 0.00	0.99 ± 0.00	0.89 ± 0.01	1.00 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.22$)	0.95 ± 0.01	0.94 ± 0.01	0.93 ± 0.00	0.98 ± 0.00	0.91 ± 0.00	0.99 ± 0.00	0.89 ± 0.00	1.00 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.14$)	0.96 ± 0.00	0.98 ± 0.00	0.95 ± 0.00	0.99 ± 0.00	0.94 ± 0.00	1.00 ± 0.00	0.93 ± 0.00	1.00 ± 0.00

CIFAR	$R = 1$		$R = 2$		$R = 5$		$R = 10$	
	class $-$	class $+$	class $-$	class $+$	class $-$	class $+$	class $-$	class $+$
std	0.71 ± 0.01	0.74 ± 0.01	0.90 ± 0.01	0.33 ± 0.02	0.99 ± 0.00	0.05 ± 0.01	1.00 ± 0.00	0.00 ± 0.00
rob ($p = 2$, $\varepsilon = 0.98$)	0.19 ± 0.10	0.85 ± 0.08	1.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00
rob ($p = 2$, $\varepsilon = 0.88$)	0.53 ± 0.12	0.54 ± 0.14	1.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00
rob ($p = 2$, $\varepsilon = 0.78$)	0.44 ± 0.08	0.77 ± 0.04	0.98 ± 0.01	0.18 ± 0.07	1.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.030$)	0.37 ± 0.10	0.70 ± 0.09	1.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.023$)	0.51 ± 0.02	0.65 ± 0.01	0.94 ± 0.01	0.18 ± 0.02	1.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.015$)	0.69 ± 0.03	0.79 ± 0.02	0.91 ± 0.01	0.36 ± 0.04	0.99 ± 0.00	0.07 ± 0.01	1.00 ± 0.00	0.01 ± 0.01

ImageNet	$R = 1$		$R = 2$		$R = 5$		$R = 10$	
	class $-$	class $+$	class $-$	class $+$	class $-$	class $+$	class $-$	class $+$
std	0.99 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	1.00 ± 0.00	0.97 ± 0.00	1.00 ± 0.00	0.95 ± 0.00
rob ($p = 2$, $\varepsilon = 4.80$)	0.93 ± 0.01	0.94 ± 0.01	0.95 ± 0.01	0.89 ± 0.02	0.99 ± 0.00	0.66 ± 0.03	1.00 ± 0.00	0.49 ± 0.04
rob ($p = 2$, $\varepsilon = 3.60$)	0.96 ± 0.00	0.93 ± 0.01	0.98 ± 0.00	0.89 ± 0.01	0.99 ± 0.00	0.81 ± 0.02	0.99 ± 0.00	0.64 ± 0.03
rob ($p = 2$, $\varepsilon = 2.40$)	0.97 ± 0.00	0.96 ± 0.00	0.98 ± 0.00	0.92 ± 0.01	0.99 ± 0.00	0.89 ± 0.01	0.99 ± 0.00	0.79 ± 0.01
rob ($p = \infty$, $\varepsilon = 0.090$)	0.93 ± 0.01	0.92 ± 0.01	0.96 ± 0.01	0.87 ± 0.01	0.99 ± 0.00	0.74 ± 0.02	1.00 ± 0.00	0.45 ± 0.02
rob ($p = \infty$, $\varepsilon = 0.068$)	0.97 ± 0.00	0.94 ± 0.00	0.97 ± 0.00	0.91 ± 0.01	0.99 ± 0.00	0.81 ± 0.01	0.99 ± 0.00	0.65 ± 0.03
rob ($p = \infty$, $\varepsilon = 0.045$)	0.97 ± 0.00	0.96 ± 0.00	0.98 ± 0.00	0.94 ± 0.00	0.98 ± 0.00	0.91 ± 0.01	0.99 ± 0.00	0.79 ± 0.02

Table 5. **Overall accuracy** for the standard and robust classifiers on synthetic datasets. “std” refers to the standard classifier and “rob” refers to the robust classifier with the specified p and ε. The presented results are averaged over 5 runs.

Synthetic Gaussian	$R = 1$	$R = 2$	$R = 5$	$R = 10$
std	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
rob ($p = 2$, $\varepsilon = 5.00$)	0.87 ± 0.01	0.86 ± 0.01	0.87 ± 0.01	0.86 ± 0.01
rob ($p = 2$, $\varepsilon = 3.75$)	0.90 ± 0.01	0.90 ± 0.01	0.90 ± 0.01	0.90 ± 0.01
rob ($p = 2$, $\varepsilon = 2.50$)	0.96 ± 0.01	0.96 ± 0.00	0.96 ± 0.00	0.95 ± 0.01
rob ($p = \infty$, $\varepsilon = 1.00$)	0.82 ± 0.03	0.81 ± 0.03	0.81 ± 0.03	0.80 ± 0.03
rob ($p = \infty$, $\varepsilon = 0.75$)	0.84 ± 0.02	0.84 ± 0.02	0.84 ± 0.02	0.85 ± 0.02
rob ($p = \infty$, $\varepsilon = 0.50$)	0.90 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.90 ± 0.01

Synthetic Cauchy	$R = 1$	$R = 2$	$R = 5$	$R = 10$
std	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
rob ($p = 2$, $\varepsilon = 5.00$)	0.84 ± 0.00	0.85 ± 0.01	0.86 ± 0.00	0.91 ± 0.00
rob ($p = 2$, $\varepsilon = 3.75$)	0.88 ± 0.00	0.89 ± 0.00	0.90 ± 0.00	0.93 ± 0.00
rob ($p = 2$, $\varepsilon = 2.50$)	0.95 ± 0.00	0.95 ± 0.00	0.96 ± 0.00	0.97 ± 0.00
rob ($p = \infty$, $\varepsilon = 1.00$)	0.74 ± 0.00	0.75 ± 0.01	0.84 ± 0.00	0.91 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.75$)	0.80 ± 0.00	0.80 ± 0.00	0.85 ± 0.00	0.91 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.50$)	0.87 ± 0.00	0.88 ± 0.00	0.90 ± 0.00	0.94 ± 0.00

Synthetic Holtsmark	$R = 1$	$R = 2$	$R = 5$	$R = 10$
std	0.51 ± 0.00	0.65 ± 0.00	0.82 ± 0.00	0.89 ± 0.00
rob ($p = 2$, $\varepsilon = 70.00$)	0.51 ± 0.00	0.66 ± 0.00	0.83 ± 0.00	0.90 ± 0.00
rob ($p = 2$, $\varepsilon = 52.50$)	0.51 ± 0.01	0.66 ± 0.00	0.83 ± 0.00	0.90 ± 0.00
rob ($p = 2$, $\varepsilon = 35.00$)	0.53 ± 0.00	0.66 ± 0.00	0.82 ± 0.00	0.90 ± 0.00
rob ($p = \infty$, $\varepsilon = 53.00$)	0.51 ± 0.01	0.67 ± 0.00	0.83 ± 0.00	0.91 ± 0.00
rob ($p = \infty$, $\varepsilon = 39.75$)	0.52 ± 0.01	0.67 ± 0.00	0.83 ± 0.00	0.91 ± 0.00
rob ($p = \infty$, $\varepsilon = 26.50$)	0.50 ± 0.01	0.66 ± 0.00	0.83 ± 0.00	0.91 ± 0.00

Table 6. **Overall accuracy** for the standard and robust classifiers on real-world datasets. "std" refers to the standard classifier and "rob" refers to the robust classifier with the specified p and ε. The presented results are averaged over 5 runs.

MNIST	$R = 1$	$R = 2$	$R = 5$	$R = 10$
std	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00
rob ($p = 2$, $\varepsilon = 2.70$)	0.98 ± 0.00	0.98 ± 0.00	0.98 ± 0.00	0.98 ± 0.00
rob ($p = 2$, $\varepsilon = 2.02$)	0.98 ± 0.00	0.98 ± 0.00	0.99 ± 0.00	0.99 ± 0.00
rob ($p = 2$, $\varepsilon = 1.35$)	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.43$)	0.97 ± 0.00	0.97 ± 0.00	0.98 ± 0.00	0.99 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.33$)	0.98 ± 0.00	0.98 ± 0.00	0.99 ± 0.00	0.99 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.22$)	0.98 ± 0.00	0.98 ± 0.00	0.99 ± 0.00	0.99 ± 0.00

Fashion-MNIST	$R = 1$	$R = 2$	$R = 5$	$R = 10$
std	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00
rob ($p = 2$, $\varepsilon = 3.23$)	0.95 ± 0.00	0.96 ± 0.00	0.97 ± 0.00	0.98 ± 0.00
rob ($p = 2$, $\varepsilon = 2.42$)	0.96 ± 0.00	0.96 ± 0.00	0.97 ± 0.00	0.98 ± 0.00
rob ($p = 2$, $\varepsilon = 1.62$)	0.98 ± 0.00	0.98 ± 0.00	0.99 ± 0.00	0.99 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.29$)	0.94 ± 0.00	0.96 ± 0.00	0.97 ± 0.00	0.98 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.22$)	0.95 ± 0.00	0.96 ± 0.00	0.98 ± 0.00	0.99 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.14$)	0.97 ± 0.00	0.98 ± 0.00	0.99 ± 0.00	0.99 ± 0.00

CIFAR	$R = 1$	$R = 2$	$R = 5$	$R = 10$
std	0.72 ± 0.00	0.71 ± 0.00	0.84 ± 0.00	0.91 ± 0.00
rob ($p = 2$, $\varepsilon = 0.98$)	0.52 ± 0.01	0.67 ± 0.00	0.83 ± 0.00	0.91 ± 0.00
rob ($p = 2$, $\varepsilon = 0.88$)	0.54 ± 0.01	0.67 ± 0.00	0.83 ± 0.00	0.91 ± 0.00
rob ($p = 2$, $\varepsilon = 0.78$)	0.61 ± 0.05	0.71 ± 0.02	0.83 ± 0.00	0.91 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.030$)	0.53 ± 0.01	0.67 ± 0.00	0.83 ± 0.00	0.91 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.023$)	0.58 ± 0.01	0.69 ± 0.00	0.83 ± 0.00	0.91 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.015$)	0.74 ± 0.01	0.72 ± 0.01	0.84 ± 0.00	0.91 ± 0.00

ImageNet	$R = 1$	$R = 2$	$R = 5$	$R = 10$
std	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	1.00 ± 0.00
rob ($p = 2$, $\varepsilon = 4.80$)	0.94 ± 0.00	0.93 ± 0.00	0.93 ± 0.01	0.95 ± 0.00
rob ($p = 2$, $\varepsilon = 3.60$)	0.95 ± 0.00	0.95 ± 0.00	0.95 ± 0.00	0.96 ± 0.00
rob ($p = 2$, $\varepsilon = 2.40$)	0.97 ± 0.00	0.96 ± 0.00	0.97 ± 0.00	0.97 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.090$)	0.93 ± 0.00	0.93 ± 0.00	0.94 ± 0.00	0.95 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.068$)	0.95 ± 0.00	0.95 ± 0.00	0.95 ± 0.00	0.96 ± 0.00
rob ($p = \infty$, $\varepsilon = 0.045$)	0.97 ± 0.00	0.96 ± 0.00	0.97 ± 0.00	0.97 ± 0.00