On the Effects of Fairness to Adversarial Vulnerability

Cuong Tran[1] , **Keyu Zhu**[2] , **Pascal Van Hentenryck**[2] and **Ferdinando Fioretto**[1]
[1]University of Virginia
[2]Georgia Institute of Technology
kxb7sd@virginia.edu, keyu.zhu@gatech.edu, phv@isye.gatech.edu, fioretto@virginia.edu

Abstract

Fairness and robustness are two important notions of learning models. Fairness ensures that models do not disproportionately harm (or benefit) some groups over others, while robustness measures the models' resilience against small input perturbations. While equally important properties, this paper illustrates a dichotomy between fairness and robustness, and analyzes when striving for fairness decreases the model robustness to adversarial samples. The reported analysis sheds light on the factors causing such contrasting behavior, suggesting that distance to the decision boundary across groups as a key factor. Experiments on non-linear models and different architectures validate the theoretical findings. In addition to the theoretical analysis, the paper also proposes a simple, yet effective, solution to construct models achieving good tradeoffs between fairness and robustness.

1 Introduction

Data-driven learning systems have become instrumental for decision-making in a variety of consequential contests, including assisting in legal decisions, [Jayatilake and Ganegoda, 2021], lending, [Stevens *et al.*, 2020], hiring, [Schumann *et al.*, 2020], and providing personalized recommendations. [Burke, 2003]. Consequentially, fairness has emerged as a critical requirement for adoption and usage of these systems. Various notions of fairness drawing from legal and philosophical doctrine have been proposed to ensure that the models' errors do not affect specific groups [Mehrabi *et al.*, 2021].

In general, fair models attempt at constraining their hypothesis space so that errors in reported outcomes are uniformly distributed across different protected groups [Mehrabi *et al.*, 2021]. When these fairness constraints are enforced in learning systems, a commonly observed behavior is an overall degradation of the model accuracy. Thus, a growing body of research has been focusing on striking the right balance between fairness and accuracy [Rodolfa *et al.*, 2021]. *This paper shows that fairness may have another important consequence on the deployed models: a reduction of the model robustness*. Given the susceptibility of deep learning models to adversarial attacks in security-sensitive applications, this is an important yet underexplored issue.

Contribution. This paper shows that enforcing fairness may negatively affect the robustness of a model. Specifically, the paper ***(1)*** analyzes when and why fairness and robustness may be misaligned in their objectives, ***(2)*** provides an understanding on the relationship between fair, robust, and "natural" (e.g., non-fair non-robust) models, and ***(3)*** identifies *the distance to the decision boundary* as a key aspect linking fairness and robustness. Moreover, ***(4)*** the paper shows how the distance to the decision boundary can explain the increase of adversarial vulnerability of fair models, providing validation over a variety of tasks and architectures, and verifying the presence of the fairness/robustness dichotomy for multiple techniques aimed at achieving fairness and measuring robustness. Finally, ***(5)*** building from the reported theoretical observations, the paper also proposes a simple, yet effective, strategy to find a good tradeoff between accuracy, fairness, and robustness.

The results show that, without careful considerations, inducing a desired equity property may create significant security challenges. *These results should not be read as an endorsement to avoid constructing fairer or safer models; rather as a call for additional research to achieve appropriate tradeoffs.*

Relation with previous work. We discuss related work in Appendix A[1] and highlight here the distinguishing features of this study in the context of robustness and fairness research. The intersection of fairness and robustness has received limited attention thus far, with only a handful of studies examining this area. For instance, [Xu *et al.*, 2021a] recently showed that adversarially robust models can exhibit significant accuracy disparity among different classes, as opposed to their standard counterparts. To address this issue, they proposed a Fair-Robust-Learning framework for adversarial defense. Meanwhile, [Khani and Liang, 2020] analyzed the impact of noise in features on disparities in error rates when learning regression models. While previous studies highlighted how adversarial training can disproportionately harm certain protected groups [Xu *et al.*, 2021a; Nanda *et al.*, 2021], *we demonstrate that enforcing fairness comes at the expense of reduced robustness*. As a result, the proposed analysis requires a distinct approach from those proposed in earlier studies.

[1] Please refer to [Tran *et al.*, 2022c] for the appendices.

2 Problem Settings

Consider a multi-class classification problem, whose input is a dataset D consisting of n data points (X_i, A_i, Y_i), each of which drawn i.i.d. from an unknown distribution Π and where $X_i \in \mathcal{X}$ is a feature vector, $A_i \in \mathcal{A}$ is a protected attribute, and $Y_i \in \mathcal{Y} = [C]$ is a label, with C being the number of possible class labels. For example, consider the case of a classifier to predict the age range of an individual. The features X_i may describe the pixels associated with the individual headshot and their demographics, the protected attribute A_i may describe the individual gender or ethnicity, and Y_i represents the age range. The goal is to learn a classifier $f_\theta : \mathcal{X} \to \mathcal{Y}$, where θ is a vector of real-valued parameters. The model quality is assessed in terms of a non-negative loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, and the training aims at minimizing the empirical risk function:

$$\overset{\star}{\theta} = \operatorname*{argmin}_{\theta} \mathcal{L}_\theta(D) \left(= \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(X_i), Y_i) \right). \quad (1)$$

For a group $a \in \mathcal{A}$, notation D_a is used to denote the subset of D containing exclusively samples i with $A_i = a$. Importantly, the paper assumes that the attribute A is not part of the model input during inference. The paper focuses on learning classifiers that satisfy group fairness (to be defined shortly) and on analyzing the robustness impact of fairness.

3 Preliminaries

Fairness and fair learning. This paper considers a classifier f satisfying accuracy parity [Zhao and Gordon, 2019], a group fairness notion commonly adopted in machine learning requiring model misclassification rates to be conditionally independent of the protected attribute. That is, $\forall (X, A, Y) \sim \Pi$ and $\forall a \in \mathcal{A}$,

$$\left| \Pr\left(f_\theta(X) \neq Y \mid A = a\right) - \Pr\left(f_\theta(X) \neq Y\right) \right| \leq \alpha, \quad (2)$$

where α denotes the allowed *fairness violation*. In practice, the above is expressed as a difference of empirical expectations of the group and population misclassification rates. That is, $\forall a \in \mathcal{A}$:

$$\left| {}^{1}\!/\!_{|D_a|} \sum_{(X,A,Y) \in D_a} \mathbf{1}\{f_\theta(X) \neq Y\} - {}^{1}\!/\!_{n} \sum_{(X,A,Y) \in D} \mathbf{1}\{f_\theta(X) \neq Y\} \right| \leq \alpha.$$

Several approaches have been proposed in the literature to encourage the satisfaction of accuracy parity. They can be summarized in methods that use *penalty terms* into the empirical risk loss function to capture the fairness violations, and those which *minimize the maximum group loss*. The core of the paper focuses on the first set of methods; the analysis for the second set is presented in Appendix C.

Penalty-based methods. In this category, the model loss function (Equation (1)) is augmented with penalty fairness constraint terms [Agarwal *et al.*, 2018; Tran *et al.*, 2021b] as:

$$\theta_{\mathrm{f}}(\lambda) = \operatorname*{argmin}_{\theta} \mathcal{L}_\theta(D) + \lambda \left(\sum_{a \in \mathcal{A}} \left| \mathcal{L}_\theta(D_a) - \mathcal{L}_\theta(D) \right| \right) \quad (3)$$

where $\mathcal{L}_\theta(D_a) = {}^{1}\!/\!_{|D_a|} \sum_{(X,A,Y) \in D_a} \ell(f_\theta(X), Y)$ is the empirical risk loss of protected group $a \in \mathcal{A}$. In addition, $\lambda > 0$ is the fairness penalty parameter that enforces a tradeoff between fairness and accuracy.

Robustness and robust learning. Following robust learning conventions, the robustness of a model f is measured in terms of the *robust error*:

$$\mathcal{L}^{\mathrm{rob}}_\theta(\epsilon) = \Pr\left(\exists \tau,\ \|\tau\|_p \leq \epsilon, f_\theta(X + \tau) \neq Y\right), \quad (4)$$

which measures the sensitivity of the model errors to small input perturbations $\|\tau\|_p \leq \epsilon$ in ℓ_p norms, with p often considered in $\{0, 1, 2, \infty\}$. Robust errors can be decomposed into two components [Zhang *et al.*, 2019]:

$$\mathcal{L}^{\mathrm{rob}}_\theta(\epsilon) = \mathcal{L}^{\mathrm{nat}}_\theta + \mathcal{L}^{\mathrm{bdy}}_\theta(\epsilon), \quad (5)$$

where the first denotes the *natural error* and the second the *boundary error*. The natural error measures the standard model performance when exposed to *unperturbed* samples (X, A, Y):

$$\mathcal{L}^{\mathrm{nat}}_\theta = \Pr\left(f_\theta(X) \neq Y\right), \quad (6)$$

whose empirical version is defined in Equation (1) with a 0/1 loss function. The boundary error measures the probability that the model predictions change on *perturbed* samples $(X + \|\tau\|_p, A, Y)$:

$$\begin{aligned} \mathcal{L}^{\mathrm{bdy}}_\theta(\epsilon) &= \Pr\big(\exists \|\tau\|_p \leq \epsilon,\ f_\theta(X + \tau) \neq f_\theta(X), \\ &\qquad f_\theta(X) = Y\big). \end{aligned} \quad (7)$$

The concept of boundary error inherently involves a **decision boundary** as well as a *distance* between an input sample and the decision boundary. In the context of linear classifiers, this boundary is typified by a hyperplane. The distance from a sample X to the decision boundary in a classifier f_θ is expressed as:

$$\Delta(X, f_\theta) = \max_{\epsilon} \text{ s.t. } f_\theta(X + \tau) = f_\theta(X),\ \forall \|\tau\| \leq \epsilon.$$

Note that samples close to the decision boundary will be less tolerant to noise than those lying far from it. The analysis in this paper regarding the impact of fairness on robustness is based on this concept. In particular, the results show that imposing fairness constraints may reduce the distance to the decision boundary of the samples $(X, A, Y) \sim \Pi$.

4 Real-World Implications

Prior diving into the analysis, we provide an example showing how robustness errors can be exacerbated when a image classifier is trained to satisfy fairness. When perturbations (either due to noise or by malicious adversaries) are introduced in the model inputs, they may cause harmful effects as they lead the classifier to misclassify targeted inputs.

Figure 1 shows UTKFace dataset examples processed by classifiers trained either on the standard empirical risk loss from Equation (1) (top) or the fair empirical risk loss from Equation (3) (bottom). Although both inputs receive identical

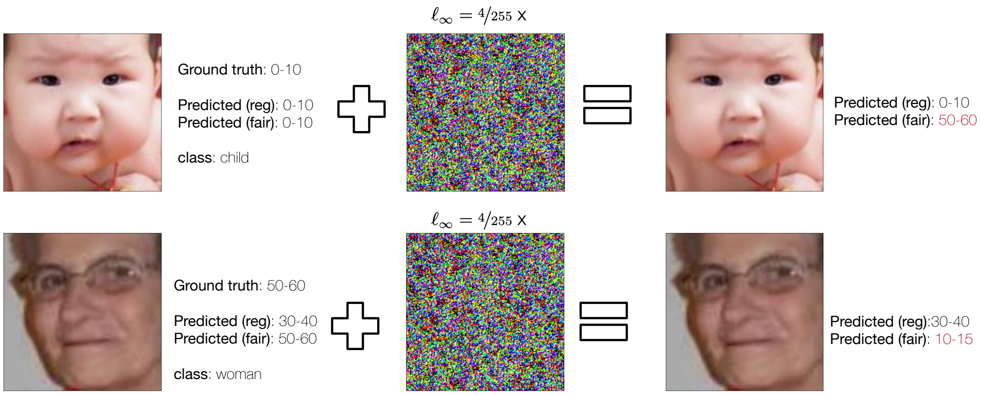

Figure 1: An example of robustness loss in the UTKFace dataset. A regular (reg) and a fair models are trained to predict age group from faces and exposed to adversarial examples generated under an RFGSM attack. The predictions of the regular model do not change under adversarial examples (regardless of their original correctness), while those of the fair models do.

ℓ_∞ noise perturbations, the fair network is much more brittle, inducing errors in the classifier's outputs.

It is important to note that while this paper uses datasets such as UTKFace (see details in Appendix D) to demonstrate the effects of fairness to robustness, the very task of predicting gender, race, or other characteristics from a person face is flawed and raises deep ethical concerns [Raji *et al.*, 2020].

5 Why Fairness Weakens Robustness?

This section presents the main results of the paper. It will show that *fairness affects model robustness because the learned decision boundary is* pulled in opposite directions *by fair and robust models*. To render the analysis tractable, the theoretical discussion focuses on linear classifiers, and more specifically on learning a mixture of Gaussians with a linear classifiers. In addition, Section 6 will show that a similar phenomenon occurs in large non-linear models.

5.1 Optimal Models for Mixtures of Gaussians

Consider a binary classification setting (i.e., $\mathcal{Y} = \{-1, 1\}$) with data drawn from a mixture of Gaussian distributions, so that $\Pr(X \mid Y = -1) \propto \mathcal{N}(\underline{\mu}, 1)$ and $\Pr(X \mid Y = 1) \propto \mathcal{N}(\bar{\mu}, K^2)$, with $\underline{\mu} < \bar{\mu}$ and different variances ($K > 1$). These non-restrictive assumptions help simplifying and clarifying exposition, but the appendix generalizes the above to higher-dimensional cases. An illustration of this setting is reported in Figure 2 (top) where the data distributions are highlighted with black dashed curves.

The following analysis poses no restrictions on the relative subgroup sizes $|D_1|$ and $|D_{-1}|$ and focuses on the *balanced* data setting, in which data samples from different protected groups are equally likely. The paper studies a family of parametric classifiers $\{f_\theta\}_\theta$ with $\theta \in [\underline{\mu}, \bar{\mu}] \subseteq \mathbb{R}$, where $f_\theta(X) = \mathbf{1}\{X > \theta\}$ denotes the classification output of the classifier. The optimal models with respect to the natural, fair, and robust losses can be specified as follows:

- **Optimal natural model** ($f_{\overset{\star}{\theta}}$). It is the Bayes classifier which minimizes the natural classification error as defined in Equation (1). In Figure 2 (top), this classifier is represented by vertical blue lines.
- **Optimal fair model** ($f_{\theta_{\mathrm{f}}}$). Intuitively, this classifier is $\theta_{\mathrm{f}}(\infty)$ as defined in Equation (3). Formally speaking, this classifier minimizes a lexicographic function whose first component is $\sum_{a \in \mathcal{A}} |\mathcal{L}_\theta(D_a) - \mathcal{L}_\theta(D))|$ and second component is $\mathcal{L}_\theta(D)$. In Figure 2 (top), this classifier is represented by vertical red lines.
- **Optimal robust model** ($f_{\theta_{\mathrm{r}}^{(\epsilon)}}$). This classifier minimizes the robust classification error in Equation (5), for a given ϵ. In Figure 2 (top), it is depicted by vertical green lines.

5.2 Relationships Between the Optimal Models

The next result characterizes the positional relationship among the three optimal models mentioned above, which can be observed in Figure 2.

Theorem 1. *For any* $\epsilon \in [0, {}^{(\underline{\mu}-\bar{\mu})}/_{2}]$ *and* $K \in (1, B_K]$, *where* $B_K = \min\left\{\exp\left({}^{(\underline{\mu}-\bar{\mu}-2\epsilon)^2}/_{2}\right), {}^{(\underline{\mu}-\bar{\mu})}/_{\epsilon} - 1\right\}$,

$$\underline{\mu} + \epsilon \leq \theta_{\mathrm{f}} \leq \overset{\star}{\theta} \leq \theta_{\mathrm{r}}^{(\epsilon)} \leq \bar{\mu} - \epsilon\,. \tag{8}$$

Besides, $\theta_{\mathrm{r}}^{(\epsilon)}$ *is an increasing function of* ϵ *over* $[0, {}^{(\underline{\mu}-\bar{\mu})}/_{2}]$.

The result follows from the observation that the optimal natural model $f_{\overset{\star}{\theta}}$ can be expressed as

$$\overset{\star}{\theta} = \underline{\mu} - \frac{\underline{\mu} - \bar{\mu}}{K^2 - 1} + \frac{K}{K^2 - 1}\sqrt{2(K^2 - 1)\ln(K) + (\underline{\mu} - \bar{\mu})^2}\,;$$

the fair classifier $f_{\theta_{\mathrm{f}}}$ as:

$$\theta_{\mathrm{f}} = \underline{\mu} + \frac{\underline{\mu} - \bar{\mu}}{K + 1}$$

and the robust classifier $f_{\theta_{\mathrm{r}}^{(\epsilon)}}$ as

$$\theta_{\mathrm{r}}^{(\epsilon)} = \underline{\mu} - \frac{\underline{\mu} - \bar{\mu} - (K^2 + 1)\epsilon}{K^2 - 1} + \frac{K}{K^2 - 1}\sqrt{2(K^2 - 1)\ln(K) + (\underline{\mu} - \bar{\mu} - 2\epsilon)^2}\,.$$

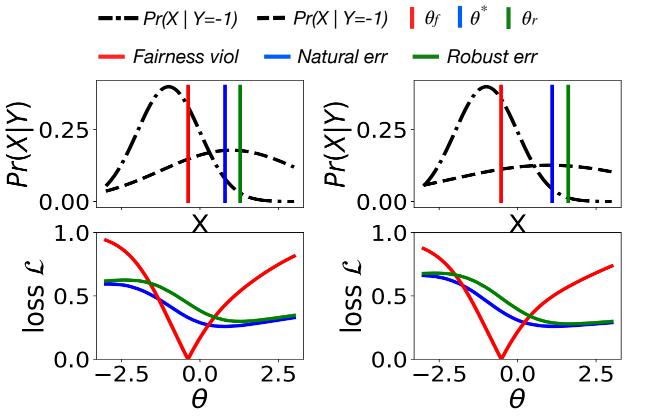

Figure 2: An optimal natural θ^*, fair θ_f, and robust θ_r classifiers for $K = 5$ (left) and $K = 10$ (right) with $\underline{\mu} = -1$ and $\bar{\mu} = 1$.

We note that [Xu *et al.*, 2021a] derives expressions for the optimal natural and robust models, which are used to investigate the natural error gap between the two classes. Importantly, however, Theorem 1 above provides a unique comparison among the three classifiers analyzed, *highlighting the difficulty in achieving both robustness and fairness simultaneously, as fairness and robustness pull the optimal classifier in opposing directions.* While [Xu *et al.*, 2021a] focuses on the unfairness resulting from robust training, the remainder of this section examines the cost of adversarial robustness in fair training. Specifically, we measure this reduced robustness cost analyzing robust and boundary errors.

From the relations highlighted in Theorem 1, it follows,

1. *fair classifiers achieve the largest robust errors* while *robust classifiers result in the least error*;
2. *fair classifiers achieve the largest boundary errors* while *robust classifiers result in the smallest boundary error*, as expressed by the following Corollaries.

Corollary 1. *For any* $\epsilon \in [0, (\underline{\mu}-\bar{\mu})/2]$ *and* $K \in (1, B_K]$,

$$\mathcal{L}^{\text{rob}}_{\theta_{\text{f}}}(\epsilon) \geq \mathcal{L}^{\text{rob}}_{\overset{\star}{\theta}}(\epsilon) \geq \mathcal{L}^{\text{rob}}_{\theta_{\text{r}}^{(\epsilon)}}(\epsilon)\,.$$

Corollary 2. *For any* $\epsilon \in [0, (\underline{\mu}-\bar{\mu})/4]$ *and* $K \in \left(1, \bar{B}_K\right]$,

$$\mathcal{L}^{\text{bdy}}_{\theta_{\text{f}}}(\epsilon) \geq \mathcal{L}^{\text{bdy}}_{\overset{\star}{\theta}}(\epsilon) \geq \mathcal{L}^{\text{bdy}}_{\theta_{\text{r}}^{(\epsilon)}}(\epsilon)\ ,$$

where $\bar{B}_K = \min\left\{\exp\left(\frac{(\underline{\mu}-\bar{\mu}-2\epsilon)^2}{2}\right), \phi^{-1}\!\left(\frac{\underline{\mu}-\bar{\mu}}{\epsilon} - 2\right)\right\}$ *and* ϕ^{-1} *is the inverse function associated with* $\phi : [1, +\infty) \mapsto [2, +\infty)$ *such that* $\phi(x) = x + 1/x$.

These results further highlight the impossibility of achieving fairness and robustness simultaneously in this classification task. Fairness and robustness are pulling the classifier in opposite directions.

5.3 The Role of the Decision Boundary

Building on the previous results, this section provides the key theoretical intuitions to explain why fairness increases adversarial vulnerability. It identifies the average distance to the decision boundary as the central aspect linking fairness and robustness, which is formalized in Theorem 2.

Theorem 2. *For any* $\epsilon \in [0, (\underline{\mu}-\bar{\mu})/2]$ *and* $K \in (1, B_K]$,

$$\mathbb{E}\left[\Delta\left(X, f_{\theta_{\text{r}}^{(\epsilon)}}\right)\right] \geq \mathbb{E}\left[\Delta\left(X, f_{\overset{\star}{\theta}}\right)\right] \geq \mathbb{E}\left[\Delta\left(X, f_{\theta_{\text{f}}}\right)\right].$$

In addition, the fair model minimizes the average distance to its decision boundary over all valid classifiers, i.e.,

$$\theta_{\text{f}} = \underset{\theta \in [\underline{\mu}, \bar{\mu}]}{\text{argmin}}\ \mathbb{E}\left[\Delta\left(X, f_\theta\right)\right]\ .$$

Theorem 2 indicates that, among the three considered optimal models, the fair model has the smallest average distance to the decision boundary while the robust model has the largest distance. The result above is exemplified in Figure 2. The bottom plots show the losses associated with the optimal natural, fair, and robust models for two choices of K (left and right) while the top plots show the optimal decision boundaries associated with each of the three models – notice they correspond to the minima of their relative losses.

Observe that class $Y = 1$ has a higher classification error than class $Y = -1$ under the natural (and thus unfair) classifier $f_{\overset{\star}{\theta}}$. This is intuitive since the conditional distribution $\Pr(X \mid Y{=}1)$ has much higher variance than $\Pr(X \mid Y{=}{-1})$. Hence, to balance the classification errors, the fair classifier pushes the decision boundary towards the mean of class $Y{=}{-1}$. This increases the error of class $Y{=}{-1}$ while decreasing the error of class $Y{=}1$. In contrast, the robust classifier pushes the decision boundary far away from the dense input region, i.e., the mean of the data associated with class $Y = -1$.

There are a few points worth emphasizing. First, *robustness and fairness pull the decision boundary into two opposite directions*. Second, the fair model $f_{\theta_{\text{f}}}$ results in predictions with higher robust errors, when compared to the optimal natural model $f_{\overset{\star}{\theta}}$, and it also increases adversarial vulnerability as the variance K increases. The variance K regulates the difference in the standard deviation of the underlying distributions associated with the protected groups and thus controls the overall distance to the decision boundary. *In summary, fairness can reduce the average samples distance of the training samples to the decision boundary which, in turn, makes the model less tolerant to adversarial noise.*

This section concludes with another important result. The previous relationships continue to hold even when the optimality conditions of the fair classifier are relaxed, i.e., when λ is taking values different from ∞. Moreover, the fairness constraints always reduce the distance to the decision boundary among protected groups and this reduction is proportional to the strength of the fairness constraints (or the tightness of the required fairness bound α).

Theorem 3. *Consider the fair classifier* $f_{\theta_{\text{f}}(\lambda)}$ *that optimizes Eq.* (3)*. It follows that, for any* $\lambda \in \left(\frac{K-1}{K+1}, +\infty\right)$,

$$\theta_{\text{f}}(\lambda) = \theta_{\text{f}}\ ,$$

which means that the fair classifier $f_{\theta_{\text{f}}(\lambda)}$ *coincides with the optimal fair classifier* $f_{\theta_{\text{f}}}$*, when the fairness penalty* λ *is large. while for any* $\lambda \in [0, {}^{K-1}/_{K+1}]$,

$$\theta_{\text{f}}(\lambda) = \underline{\mu} - \frac{\underline{\mu} - \bar{\mu}}{K^2 - 1} +$$

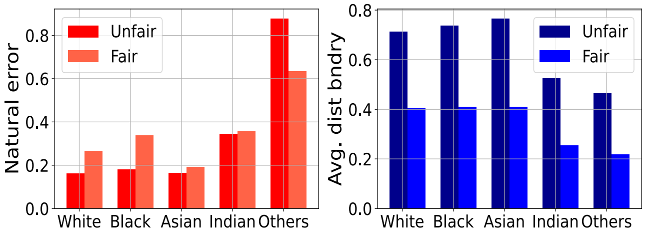

Figure 3: Comparing group's natural accuracy (left) and its average distance to the decision boundary (right) in fair and unfair models (UTK-Face dataset).

$$\frac{K}{K^2-1}\sqrt{2(K^2-1)\ln\left(\frac{1-\lambda}{1+\lambda}\cdot K\right)+(\underline{\mu}-\bar{\mu})^2}.$$

Moreover, the parameter $\theta_{\text{f}}(\lambda)$ associated with the fair classifier and the average distance to its decision boundary $\mathbb{E}\left[\Delta\left(X, f_{\theta_{\text{f}}(\lambda)}\right)\right]$ are both decreasing as λ increases.

Informally speaking, Theorem 3 states that applying fairness constraint with large enough penalty λ will push the decision boundary towards the negative class (group with smallest variance). As a result, the average distance to the decision boundary of all samples will be reduced.

While this analysis applies to the setting considered in this section, the results are empirically validated on large nonlinear models. For example, Figure 3 compares the performance of a fair CNN model (bottom plots) with $\lambda = 1.0$ against a natural (non-fair) CNN classifier (top plots). The left plots report the task accuracy by each subgroup (denoting races) and average distance to decision boundary (right) of each subgroup. Note how the fair classifier reduces the disparities in task accuracy experienced by the various subgroups. This effect, however, also reduces the *overall* average distance to the decision boundary. As a consequence, fair models will be more vulnerable to adversarial perturbations.

The next sections focus on assessing these theoretical intuitions onto general non-linear classifiers in a variety of settings and on devising a possible mitigation strategy to balance a good tradeoff between fairness and robustness.

6 Beyond the Linear Case

This section empirically validates the theoretical insights discussed earlier, extending them to more complex architectures, datasets, and loss functions. Our experiments focus on exploring the interplay between fairness, robustness, and error rates in relation to decision boundary proximity. For nonlinear models f_θ, calculating this proximity becomes a computational challenge. Hence, we employ a widely-used proxy metric that quantifies the difference between the two highest order statistics of the softmax output [Wang and Loog, 2022].

Datasets. The experiments of this section focus on three vision datasets: *UTK-Face* [Zhang *et al.*, 2017], *FMNIST* [Xiao *et al.*, 2017] and *CIFAR-10* [Krizhevsky *et al.*, 2009]. The adopted protected groups and labels in the UTK-Face datasets are *ethnicity* (White/Black/Indian/Asian/Others) or *age* (nine age bins), resulting in two distinct tasks. For FMNIST and CIFAR, the experiments use their standard labels and assume that labels are also protected groups, mirroring the setting of previous work [Tran *et al.*, 2022a; Ma *et al.*, 2022]. A complete description of the dataset and settings is found in Appendix E.

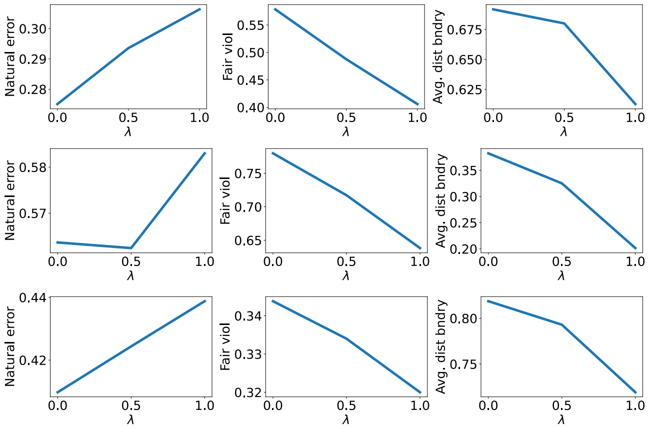

Figure 4: Natural errors, fairness deviations, and average decision-boundary distances for the UTK-Face (ethnicity and age bins) and CIFAR datasets, with variations in the fairness coefficient λ on a CNN model.

Settings. The experiments consider several deep neural network architectures, including CNN, ResNet 50 [He *et al.*, 2016] and VGG-13 [Simonyan and Zisserman, 2014]. The former uses 3 convolutional layers followed by 3 fully connected layers. Models trained on the UTK-Face data use a learning rate of $1e^{-3}$ and 70 epochs. Those trained on FMNIST and CIFAR, use a learning rate of $1e^{-1}$ and 200 epochs, as suggested in previous work [Xu *et al.*, 2021a]. The experiments analyze penalty-based fairness method, RFGSM attacks [Tramèr *et al.*, 2017], and the VGG-13 network, unless specified otherwise. Additional experiments using group-loss focused method (see Appendix C), additional network architectures, and adversarial attacks are reported in Appendix E.

Fairness impacts on the decision boundary. As shown by Theorem 3, fairness reduces the average distance of the testing samples to the decision boundary. This section illustrates how this result carries over to larger non-linear models. Figure 4 reports results obtained by executing the penalty-based fair models on the UTK-Face datasets for ethnicity (top) and age (middle) classification and on CIFAR (bottom). A clear trend emerges: As more fairness is enforced (larger λ values), the natural errors (left plots) increase, while the fairness violations (center plots) decrease. Importantly, and in agreement with the theoretical results, the experiments report a sharp reduction to the average distance to the decision boundary (right plots). This behavior renders fair models more vulnerable to adversarial attacks, as will be highlighted shortly. Similar results are also observed for the group-loss based models and other architectures (see Appendix E).

Boundary errors increase as fairness decreases. This section highlights the key consequence of the sharp reduction to the average distance to the decision boundary: *the increase of*

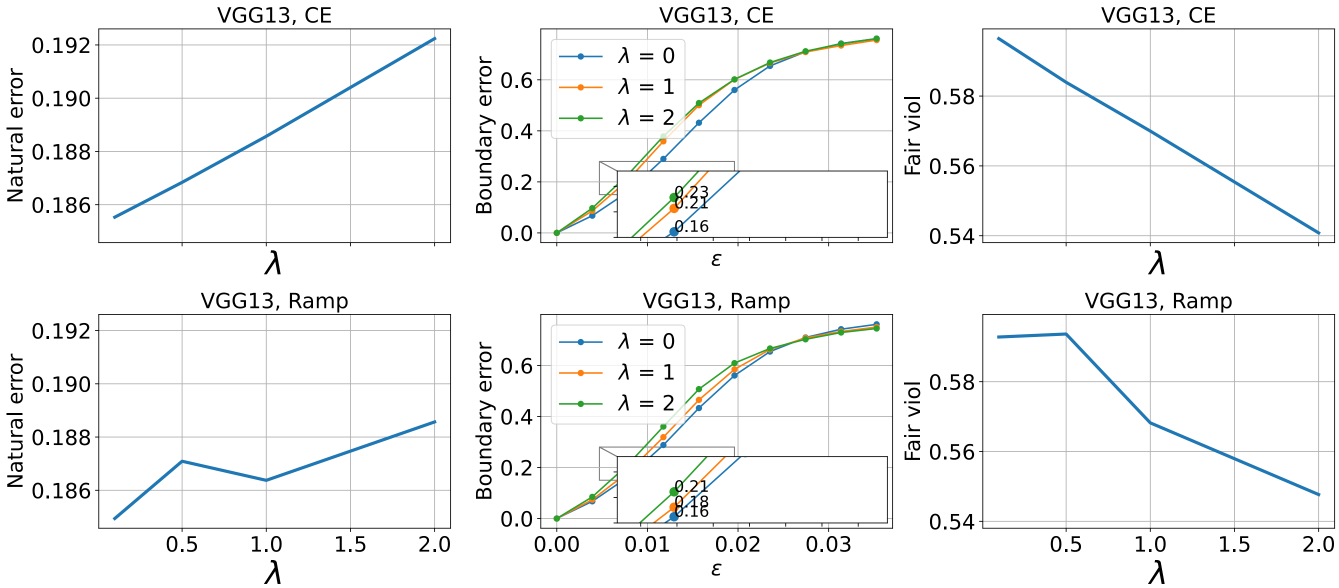

Figure 5: **Top**: Natural errors (left) and fairness violations (right) on the UTKFace *ethnicity* task at varying of the fairness parameters λ. The middle plots compares the robustness of fair ($\lambda > 0$) vs. natural ($\lambda = 0$) classifiers to different RFGSM attack levels. **Bottom**: Mitigating solution using the bounded Ramp loss.

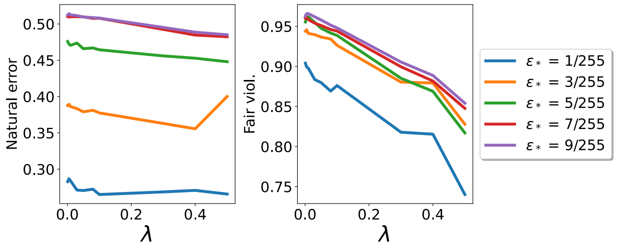

Figure 6: Natural error (left) and fairness violation (right) at varying of the margin perturbation ϵ_* and fairness parameters λ.

the vulnerability to adversarial attacks. Figure 5 (top) reports the natural errors (left), boundary errors (middle), and fairness violations (right) for a VGG-13 model trained on UTKFace dataset on the *ethnicity* task using a standard cross-entropy (CE) loss. Once again, other architectures[2] and datasets are reported in the appendix and the results follow the same trends as those reported here. Error rates and fairness violations are evaluated for *fair* classifiers across different fairness coefficients, denoted as λ. We also report boundary errors for classifiers that meet various fairness levels (λ in $\{0, 1, 2\}$) and robustness strengths (ϵ), as defined in Equation (4).

Notice how, compared to natural models, fair models incur much higher natural and boundary errors. In particular, the fairness models have boundary errors that are up to 9% larger than their natural counterparts. These observations match the theoretical analysis and highlight a significant increase in vulnerability to adversarial examples by the fair models, even for moderate selections of the fairness violation parameters λ.

6.1 Enforcing Both Fairness and Robustness

This section considers an additional experiment that highlights the potential negative impact of fairness on robustness. The experiment involves a classifier attempting to achieve both fairness and robustness. similar to [Xu *et al.*, 2021a], which incorporates two fairness components to align per-class natural/robust accuracy per class with overall natural/robust accuracy (see Equation (9) in [Xu *et al.*, 2021a]), our approach adds both robustness and fairness regularization terms to the standard classification objective function. The resulting model aims at solving the following regularized empirical risk problem:

$$\min_{\theta} \mathcal{L}_\theta(D) + \frac{1}{n}\sum_{i=1}^{n} \max_{\|\tau\|_p \le \epsilon_*} \ell(f_\theta(X_i+\tau), Y_i)$$

$$+ \lambda \sum_{a \in \mathcal{A}} \left| {}^{1}\!/\!{}_{|D_a|} \sum_{(X,A,Y) \in D_a} \ell(f_\theta(X), Y) - \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(X_i), Y_i) \right|$$

using stochastic gradient descent. The second component aims at increasing the robustness of the classifier under a margin perturbation ϵ_*, following the PGD training [Madry *et al.*, 2017] with perturbation norm $p = \infty$. It works by first generating adversarial samples $X_i + \tau$, where $\|\tau\|_\infty \le \epsilon_*$, and then the learning progress aims at minimizing the loss between the model prediction for that adversarial samples and the ground-truth $\ell(f_\theta(X_i + \tau), Y_i)$. The larger the margin perturbation ϵ_*, the more robust the resulting classifier. The third component implements a penalty-based fairness strategy [Fioretto *et al.*, 2020; Tran *et al.*, 2022b; Agarwal *et al.*, 2018],

[2] With the caveat that VGG-13 could not be used for FMNIST since the 28x28 pixel resolution of FMNIST is smaller than that required by some VGG filters.

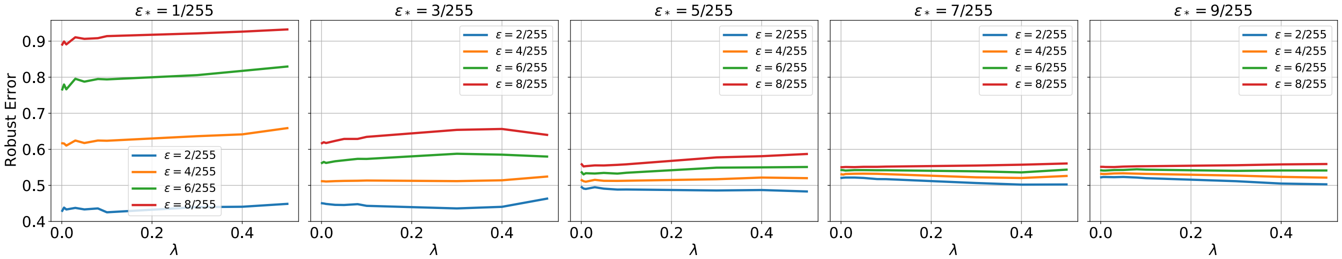

Figure 7: Robust errors for different attack levels ϵ of a robust and fair classifier at varying of the margin perturbation ϵ_* and fairness value λ.

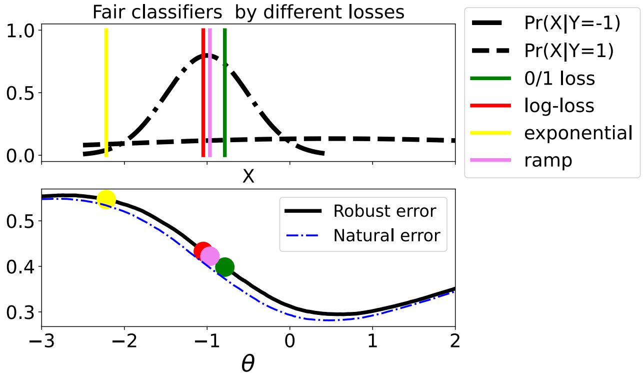

Figure 8: Classifiers using different losses (top) and the associated natural and robust errors (bottom).

which promotes fairness by penalizing the difference among each groups' average loss and the overall's average loss.

The experiments vary the margin perturbation ϵ_* (robustness) and the penalty value λ (fairness). Figure 6 reports the (natural) error (left) and fairness violations (right) for different levels of the margin perturbation ϵ_* on the UTK-Face (ethnicity) dataset. As expected, enforcing larger margin perturbations ϵ_* increases model robustness, *but at the cost of significantly increasing the natural errors*. Increasing the fairness parameter λ decreases the fairness violation.

Figure 7 reports the robust errors under different levels of adversarial attacks, specified by the perturbation value ϵ. Notice how the level of defense ϵ_* correlates with higher robustness (smaller robust errors) for all fairness parameters λ tested. *These results show the challenge to achieving simultaneously robustness, fairness, and accuracy*. They also suggest that the incorporation of both robustness and fairness, as proposed in [Xu *et al.*, 2021a], may not effectively reduce the trade-off between accuracy, robustness, and fairness.

7 A Mitigating Solution

While previous sections established the inherent trade-off between fairness and robustness, we now introduce a theoretically-grounded solution to mitigate this conflict. Note that in standard (unbounded) loss functions, misclassified samples far from the decision boundary incur much higher losses than those near it. Given that the decision boundary is a pivotal factor connecting fairness and robustness, we propose using a bounded loss function [Goh *et al.*, 2016; Collobert *et al.*, 2006]:

$$\ell_{Ramp}(f_\theta(X), Y) = \min(1, \max(0, 1 - Y f_\theta(X))),$$

and referred to as *Ramp loss*, with domain $(0, 1]$. Our strategy incorporates this bounded loss function into a fair classifier, as outlined in Equation (3). The benefits are evident in Figure 5, where it shows decreased natural (left) and boundary errors (middle) for fair classifiers with $\lambda > 0$.

Figure 8, further illustrates the strenghts of the proposed strategy. The results depict the same setting used in the previous section and compare a fair classifier trained using the ramp loss with one trained using a 0/1-loss (which is also bounded but not differentiable), a log-loss , and an exponential loss (both unbounded) (top). The results show that the fair classifier trained using a ramp-loss is the least impacted by misclassified samples, resulting in lower robust errors compared to unbounded losses. It can be observed in the bottom subplot, where its associated loss is the closest, among all differentiable losses, to the local minima. Further analysis reported in Appendix E illustrates the strength of the proposed solution.

8 Conclusions

This paper was motivated by two key challenges brought by the the adoption of modern machine learning systems in consequential domains: *fairness* and *robustness*. The paper observed and analyzed the relationship between these two important machine-learning properties and showed that fairness increases vulnerability to adversarial examples. Through a theoretical analysis on linear models, this work provided a new understanding of why such tension arises and identified the distance to the decision boundary as a key explanation factor linking fairness and robustness. These theoretical findings were validated on non-linear models through extensive experiments on a variety of vision tasks. Finally, building from this new understanding, the paper proposed a simple, yet effective, strategy to find a better balance between accuracy, fairness and robustness.

Overall, our results show that, without a careful consideration, inducing a desired equity property on a learning task may create significant security challenges. We stress that this should not be read as an endorsement to satisfy a single property, but as a call for additional research at the intersection of fairness and robustness in order to design appropriate tradeoffs and hope that our results could stimulate such a needed discussion.

Acknowledgements

This research is partially supported by NSF grants 2227082, 2232054 and NSF CAREER Award 2143706. Fioretto is also supported by an Amazon Research Award and a Google Research Scholar Award. Its views and conclusions are those of the authors only.

References

[Agarwal *et al.*, 2018] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[Bagdasaryan *et al.*, 2019] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

[Buolamwini and Gebru, 2018] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[Burke, 2003] Robin Burke. Hybrid systems for personalized recommendations. In *IJCAI Workshop on Intelligent Techniques for Web Personalization*, pages 133–152. Springer, 2003.

[Caton and Haas, 2020] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

[Collobert *et al.*, 2006] Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou. Trading convexity for scalability. In *Proceedings of the 23rd international conference on Machine learning*, pages 201–208, 2006.

[Fioretto *et al.*, 2020] Ferdinando Fioretto, Pascal Van Hentenryck, Terrence W. K. Mak, Cuong Tran, Federico Baldo, and Michele Lombardi. Lagrangian duality for constrained deep learning. In *European Conference on Machine Learning*, volume 12461 of *Lecture Notes in Computer Science*, pages 118–135. Springer, 2020.

[Friedler *et al.*, 2019] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.

[Galloway *et al.*, 2019] Angus Galloway, Anna Golubeva, Thomas Tanay, Medhat Moussa, and Graham W Taylor. Batch normalization is a cause of adversarial vulnerability. *arXiv preprint arXiv:1905.02161*, 2019.

[Goh *et al.*, 2016] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. *Advances in Neural Information Processing Systems*, 29, 2016.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[Hastie *et al.*, 2009] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hooker *et al.*, 2019] Sara Hooker, Yann Dauphin, Aaron Courville, and Andrea Frome. Selective brain damage: Measuring the disparate impact of model pruning. 2019.

[Hooker *et al.*, 2020] Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.

[Jayatilake and Ganegoda, 2021] Senerath Mudalige Don Alexis Chinthaka Jayatilake and Gamage Upeksha Ganegoda. Involvement of machine learning tools in healthcare decision making. *Journal of Healthcare Engineering*, 2021, 2021.

[Khani and Liang, 2020] Fereshte Khani and Percy Liang. Feature noise induces loss discrepancy across groups. In *International Conference on Machine Learning*, pages 5209–5219. PMLR, 2020.

[Kim, 2020] Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020.

[Krishnan *et al.*, 2020] Anoop Krishnan, Ali Almadan, and Ajita Rattani. Understanding fairness of gender classification algorithms across gender-race groups. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1028–1035. IEEE, 2020.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[Li *et al.*, 2019] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.

[Ma *et al.*, 2022] Xinsong Ma, Zekai Wang, and Weiwei Liu. On the tradeoff between robustness and fairness. *Advances in Neural Information Processing Systems*, 35, 2022.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[Nanda *et al.*, 2021] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 466–477, 2021.

[Raji *et al.*, 2020] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151, 2020.

[Rodolfa *et al.*, 2021] Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, 3(10):896–904, 2021.

[Schumann *et al.*, 2020] Candice Schumann, Jeffrey Foster, Nicholas Mattei, and John Dickerson. We need fairness and explainability in algorithmic hiring. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2020.

[Shamir *et al.*, 2019] Adi Shamir, Itay Safran, Eyal Ronen, and Orr Dunkelman. A simple explanation for the existence of adversarial examples with small hamming distance. *arXiv preprint arXiv:1901.10861*, 2019.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Stevens *et al.*, 2020] Alexander Stevens, Peter Deruyck, Ziboud Van Veldhoven, and Jan Vanthienen. Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1241–1248. IEEE, 2020.

[Subramanian *et al.*, 2021] Shivashankar Subramanian, Afshin Rahimi, Timothy Baldwin, Trevor Cohn, and Lea Frermann. Fairness-aware class imbalanced learning. *arXiv preprint arXiv:2109.10444*, 2021.

[Sukthanker *et al.*, 2022] Rhea Sukthanker, Samuel Dooley, John P Dickerson, Colin White, Frank Hutter, and Micah Goldblum. On the importance of architectures and hyperparameters for fairness in face recognition. *arXiv preprint arXiv:2210.09943*, 2022.

[Szegedy *et al.*, 2013] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[Tramèr *et al.*, 2017] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[Tran *et al.*, 2021a] Cuong Tran, My Dinh, and Ferdinando Fioretto. Differentially private empirical risk minimization under the fairness lens. In *Advances in Neural Information Processing Systems*, 2021.

[Tran *et al.*, 2021b] Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. Differentially private and fair deep learning: A lagrangian dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9932–9939, 2021.

[Tran *et al.*, 2022a] Cuong Tran, Ferdinando Fioretto, Jung-Eun Kim, and Rakshit Naidu. Pruning has a disparate impact on model accuracy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[Tran *et al.*, 2022b] Cuong Tran, Ferdinando Fioretto, Jung-Eun Kim, and Rakshit Naidu. Pruning has a disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

[Tran *et al.*, 2022c] Cuong Tran, Keyu Zhu, Ferdinando Fioretto, and Pascal Van Hentenryck. Fairness increases adversarial vulnerability. *CoRR*, abs/2211.11835, 2022.

[Wang and Loog, 2022] Ziqi Wang and Marco Loog. Enhancing classifier conservativeness and robustness by polynomiality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13327–13336, 2022.

[Wang *et al.*, 2020] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020.

[Xiao *et al.*, 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[Xu *et al.*, 2021a] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning*. PMLR, 2021.

[Xu *et al.*, 2021b] Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, and Zhen Cui. Consistent instance false positive improves fairness in face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 578–586, 2021.

[Yao *et al.*, 2018] Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. *arXiv preprint arXiv:1802.08241*, 2018.

[Zhang *et al.*, 2017] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[Zhang *et al.*, 2019] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[Zhao and Gordon, 2019] Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. *Advances in neural information processing systems*, 32:15675–15685, 2019.