No Metric Is an Island: How Algorithmic Fairness Interacts with Other AI Properties

Irene Buselli

Department of Computer Science, Bioengineering, Robotics and System Engineering (DIBRIS)

University of Genova

Supervisor

Luca Oneto

Ph.D. Program in Security, Risk and Vulnerability

Cybersecurity and Reliable Artificial Intelligence

December 11, 2025

Ringraziamenti

Nonostante in copertina io compaia come la sola autrice, queste pagine non esisterebbero allo stesso modo senza il supporto e la sopportazione di un certo numero di persone.

Grazie quindi, innanzitutto, a tutte quelle che negli ultimi anni hanno abitato da vicino la mia vita: alla mia famiglia e ai miei amici, che hanno vegliato sui miei entusiasmi estemporanei e le mie stanchezze insolubili con più affetto che giudizio, anche quando sarebbe stato comprensibile il contrario.

Grazie ai "ragazzi del lab", con cui ho condiviso molti pezzi di strada, dalla guerra dei server alle grigliate, passando per diversi "spegni e riaccendi", qualche centinaio di capsule del caffè e un rischio di arresto in Florida.

Grazie a Dani, lui sa perché.

Grazie a Ludo, perché non poteva capitarmi compagno di fatiche migliori.

Grazie a Carlo, non solo perché ha materialmente reso possibile questo dottorato, ma perché non c'è stato un giorno, in questi anni di lavoro insieme, in cui abbia anteposto una scadenza o un orario alla serenità dei rapporti umani o alle esigenze strampalate della mia vita, con una fiducia e una comprensione che non mi erano dovute, e che pure sono state fondamentali.

Infine, grazie Luca. Otto anni fa sono entrata nel tuo ufficio con del codice matlab orrendo, una soggezione infinita e nessun indizio sul mio futuro — ora che siamo alla terza tesi insieme, mi accorgo di quanto piccoli possono essere i momenti che cambiano il corso delle cose, e di quanto poco allora intuissi della fortuna di averti incrociato lungo la strada. Grazie di tutto. Per aver visto quello che potevo fare prima che lo vedessi io, per tutte le cose che mi hai insegnato e per tutte quelle che ancora spero di imparare da te. E grazie, sopra ogni cosa, per lo sguardo con cui in questi anni hai osservato il mio cammino, con l'apparente certezza che potevo superare da sola ogni ostacolo, eppure, contemporaneamente, con una mano sempre pronta a tendersi in aiuto al momento giusto.

Grazie a tutti voi. Qualunque futuro ci aspetti, ci vediamo là.

Abstract

The rapid integration of Artificial Intelligence (AI) across diverse societal domains has intensified concerns about the trustworthiness of automated systems. Beyond accuracy and efficiency, AI must now also satisfy broader ethical and technical desiderata, such as safety, reliability, equity, or transparency. Yet, these properties are not independent: their intersections often involve trade-offs or interplays that challenge both theoretical analysis and practical deployment.

Stemming from this context, this thesis specifically explores the interaction of algorithmic **fairness** with two critical dimensions of trustworthy AI: **robustness** and **regression**.

The first novel contribution proposes the identification, characterization, and mitigation of unfair regression, a phenomenon whereby model updates, despite improving overall performance, disproportionately harm specific demographic subgroups. The second contribution is the formulation of the Robust Fair Empirical Risk Minimization (RFERM), a theoretical framework designed to account for robustness bias — i.e., the heightened vulnerability of disadvantaged groups to adversarial perturbations.

Taken together, these contributions advance the understanding of fairness as an inherently interdependent property of AI, highlighting the need for joint optimization strategies that move beyond siloed ap-

proaches. In doing so, it provides both conceptual insights and practical methodologies for developing AI systems that are not only effective but also equitable and trustworthy.

Contents

Chapter 1

Introduction

1.1 Motivations and Context

The pervasive integration of Artificial Intelligence (AI) throughout nearly every societal domain — including healthcare [1; 2], financial lending [3; 4], hiring [5; 6], advertising [7; 8], social services [9; 10] — has inaugurated an era of profound transformation, reshaping human interaction and decision-making at an unprecedented scale. As autonomous systems increasingly occupy critical roles, their opacity, vulnerabilities, and propensity for unintended consequences have catalyzed a global debate on the ethical foundations of their design and deployment. Indeed, with this growing influence of machine learning (ML), it has become evident that traditional technical metrics (e.g., accuracy or computational efficiency) are insufficient to capture the full spectrum of real-world behaviors and societal implications associated with ML applications [11]. This recognition has intensified scholarly and public concerns, underscored by the technology's persistent struggles to adhere to established ethical norms and by the record number of AI-related incidents reported in recent years[1]. Models have started to learn

[1] `https://incidentdatabase.ai`

human biases and misbehaviors [12; 13], to violate the privacy of single individuals [14; 15], to show limited robustness to (malicious) data perturbations [16; 17], and to be less and less understandable, undermining the principle of the fundamental right to an explanation [18; 19].

In response to this pressing reality, the notion of **AI trustworthiness** has emerged as a societal imperative — one that extends beyond functionality to embrace essential dimensions such as safety, reliability, equity, and transparency.

One of the core pillars of AI trustworthiness — the one to which the largest part of my doctoral work has been devoted — is the concept of **algorithmic fairness**, i.e., the research area that studies how to prevent deep-seated societal biases and inequities from being inherited, encoded, or even amplified by the AI models themselves.

1.1.1 The Fairness Issue

The discourse around algorithmic fairness is fundamentally rooted in the recognition that AI systems are not neutral, but rather reflexive learners that acquire their operational logic from vast, human-generated datasets. As a result, biases and inequities present in these historical data sources may lead to models that treat individuals unfavorably on the basis of sensitive attributes such as race, gender, disability status, or sexual and political orientation.

A canonical example comes from a tool used by courts in the United States to make pre-trial detention and release decisions. The software, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), estimates the probability that a person will commit another offense [20]. A 2016 investigation published by ProPublica [13] found that black defendants were disproportionately misclassified as being at higher risk of recidivism compared to white defendants, while white defendants were more likely to be incorrectly flagged as low risk,

even if the individual's race was not included in the features fed to COMPAS. Other famous examples include ML hiring systems discriminating against female candidates [21] or commercial face recognition systems with systematically lower accuracy for female and black subjects [22]. These and other evidences have led to a growing interest in the literature in defining, evaluating, and mitigating unfairness in ML [23; 24; 25; 26]. This line of inquiry has not only produced significant theoretical and empirical insights, but also proposed and validated a number of effective approaches and practical solutions aimed at reducing discriminatory outcomes in algorithmic decision-making (see Section 2.1).

1.1.2 Interactions Matter

However, neither fairness nor the other desirable properties of ML models can be considered as "islands", that is, as separate domains without intersections. Indeed, in recent years it has become increasingly evident that the various dimensions of AI research — encompassing both technical and ethical perspectives — cannot be examined in isolation. Rather, they are characterized by a dense network of connections, interactions, frictions, and trade-offs that collectively shape the broader discourse on AI trustworthiness [11].

Researchers have started to draw attention to the interdependence and interplay among the various properties of AI systems [27]: for instance, the need for data privacy might interfere with the desire to explain the system output in detail [28], and the pursuit of algorithmic fairness may be detrimental to the accuracy and robustness experienced by some groups [29]. As a result, the trivial combination of systems that individually target different aspects of trustworthiness does not ensure a more trustworthy and effective overall result. Instead, achieving trustworthiness requires joint optimization and careful management of trade-offs across multiple dimensions [11; 30].

Stemming from this research context, the aim of my doctoral work is to investigate the interactions between fairness and other key ML metrics, as well as to develop approaches for mitigating the effects of these interplays.

1.2 Objectives and Contributions

Starting from the recognition that fairness cannot be comprehensively understood in isolation from other properties of AI systems, this thesis sets out to explore how fairness interacts with two particular ML dimensions: **robustness** and **regression**.

The overarching aim is to advance both the theoretical understanding and the methodological treatment of fairness in relation to these two metrics of AI trustworthiness. More specifically, the research objectives are (i) to investigate the forms that these interactions take, (ii) to propose specific metrics to evaluate them, providing both conceptual clarity and quantitative measures that can be applied across domains, and (iii) to design and propose mitigation strategies and optimization frameworks aimed at alleviating the negative effects of these phenomena, while taking into account the trade-off with accuracy.

In particular, the main original contributions of this thesis (see Chapters 3 and 4) are:

1. The identification and characterization of the phenomenon of *unfair regression*, documented here for the first time. To address this issue, we propose a set of ethical metrics specifically designed to measure unfairness in regression settings, together with two mitigation strategies: one acting on the learning algorithm itself and another on the hyperparameter tuning phase. Experiments on real-world datasets confirm both the existence of the unfair regression phenomenon and the effectiveness of the proposed mitigation

approaches.

2. The theoretical characterization of a novel optimization framework, the Robust Fair Empirical Risk Minimization (RFERM), for training models that are explicitly aware of robustness bias. This contribution is twofold: (i) we establish the theoretical consistency of the RFERM paradigm, and (ii) we demonstrate its applicability by implementing it across three prominent families of ML models — linear, kernelized, and deep learning approaches.

Together, these contributions aim to advance the understanding of fairness as an inherently interdependent property of machine learning systems and provide actionable pathways toward the design and deployment of more trustworthy AI. Beyond the specific findings, this work underscores the importance of addressing fairness not as an isolated desideratum, but as part of a broader ecosystem of interacting values that collectively determine the ethical and societal impact of AI technologies.

1.3 Overview of the Thesis

The following chapters provide a comprehensive analysis of current practices and emerging directions in the field of trustworthy ML, along with the original contributions developed throughout my PhD research. Specifically, Chapter 2 reviews the state of the art of the research pertaining fairness, robustness, regression and their mutual interactions; Chapter 3 and Chapter 4 present the original contributions of the thesis; finally, Chapter 5 concludes the work, summarizing its main findings and highlighting the key points of interest.

Chapter 2

Background and Related Works

Summary

Since this work addresses the intersections between fairness, robustness, and regression, the following paragraphs provide a concise overview of these concepts, supported by a comprehensive review of the relevant literature. Furthermore, particular attention is devoted to the ways in which these notions interact with one another, with an emphasis on the findings that form the foundation of this thesis as well as the main gaps that persist within the existing literature.

2.1 Fairness

Research in algorithmic fairness began gaining formal recognition in the mid-to-late 2010s, although its intellectual roots trace back to earlier work in statistics, ethics, and social sciences on discrimination and bias in decision-making [31]. The rise of large-scale machine learning applications — especially in high-stakes areas such as hiring [5; 6], lending [3; 4], policing [32], and healthcare [1; 2] — exposed instances where algorithms amplified existing societal inequalities [33]. Seminal

investigations, such as the aforementioned ProPublica analysis of COMPAS recidivism prediction tools [34] or the case of the Amazon ML hiring system discriminating against female candidates [21], catalyzed widespread concern, prompting the development of formal definitions of fairness and bias mitigation methods [24; 35]. During this period, dedicated research fields were established [36], marking the shift from scattered discussions to a focused, interdisciplinary research community addressing fairness as a core dimension of responsible AI.

2.1.1 Types of Bias

There exists a rich literature on the different types of ML biases. In general, they can arise from various stages in the data–algorithm–user interaction feedback loop. In the following, I will summarize the main ones using the categorization introduced by Mehrabi et al. in [20], distinguishing these biases into three main pathways: *data to algorithm*, *algorithm to user*, and *user to data*.

2.1.1.1 Data to Algorithm

When training data contain biases, algorithms can learn and propagate them. Key types of data biases include *measurement biases*, which stem from the choice and measurement of features (e.g., using prior arrests as a proxy for criminality) [37], *omitted variable biases*, where missing relevant variables lead to biased prediction [38], and *representation or sampling biases* [37; 39], when non-representative data samples omit subgroups (e.g., in geographically skewed datasets).

2.1.1.2 Algorithm to User

As algorithms modulate user behavior, any bias in algorithms might introduce biases in user behavior. In particular, we can distinguish algorithm–related biases

in:

- *Algorithmic biases*, when the bias is added purely by the algorithm: the algorithmic design choices, such as use of certain optimization functions, regularizations, choices in applying models on the data as a whole or considering subgroups, and the general use of statistically biased estimators in algorithms [40] can all contribute to biased algorithmic decisions;
- *Evaluation biases*, which happen during model evaluation, especially through the use of inappropriate or disproportionate benchmarks for evaluation of applications (e.g., benchmarks used in the evaluation of facial recognition systems that were biased toward skin color and gender [41]).

2.1.1.3 User to Data

Eventually, many data sources used for training ML models are user-generated. Thus, any inherent or societal bias in users might be reflected in the data, producing for instance *historical biases* [37], *population biases* [42], *social biases* [43], or *behavioral biases* [44].
As a final remark, as has been well established in the literature, all these categories — *data to algorithm*, *algorithm to user*, and *user to data* — are interconnected through a feedback loop, whereby biases in one stage can influence and amplify biases in the others [20].

2.1.2 Fairness: Definitions and Metrics

To address the problem of bias in ML, one of the first issues that deserves attention is the need to find a definition of fairness. Since algorithms operate mathematically, the fairness research community has worked to formalize society's inherently ambiguous concepts of fairness and discrimination into rigorous

mathematical definitions, with the aim of embedding these principles into machine learning systems [45; 46]. In general, fairness is the absence of prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision making [47]. To translate this concept into a quantitative definition, researchers have mainly worked on the idea of minimizing the difference in the predictions across sensitive groups or individuals.

The notion of *disparate impact* [48], for instance, requires a high ratio between the positive prediction rates of the groups, thus ensuring that the proportion of positive predictions is similar. Similarly, *demographic parity* [49] imposes that the difference in the positive prediction rates is small. One disadvantage of these two measures is that a fully accurate classifier may be considered unfair, when the base rates (i.e., the proportion of actual positive outcomes) of the various groups are significantly different. Moreover, to satisfy demographic parity, two similar individuals may be treated differently since they belong to two different groups — which in some cases may even be prohibited by law [50].

In [35], to overcome the disadvantages of measures such as disparate impact and demographic parity, Hardt et al. proposed the definitions of *equalized odds* and *equal opportunity* in supervised learning. These measures compute the difference between the false-positive rates (FPRs) and the difference between the true-positive rates (TPRs) of the two groups.

In parallel, [51] introduced the notion of *individual fairness*, which requires that similar individuals receive similar predictions. The authors formalize this using a similarity metric and develop a framework to incorporate this concept into algorithm design.

Interestingly, several recent studies have shown that it is not possible to satisfy multiple notions of fairness simultaneously [33; 52].

2.1.3 Proposed Mitigations

There have been numerous attempts to address biases in artificial intelligence in order to achieve fairness; the proposed methods are typically categorized into three types: *pre-processing*, *in-processing*, and **post-processing**.

2.1.3.1 Pre-processing Mitigations

Pre-processing methodologies aim to modify the training data before they are used to train the algorithm. The objective is to ensure that the data provided to the model are balanced and do not systematically favor any particular group. Early mitigation strategies, such as those proposed by Kamiran and Calders [53], rely on relabeling instances or re-weighting samples prior to training in order to achieve fairer classifications. In practice, label adjustments are often applied to instances near the decision boundary, as these are the most susceptible to discrimination [50]. More recent approaches instead focus on modifying feature representations. For example, Feldman et al. [48] propose altering dataset features so that the resulting distributions of privileged and unprivileged groups become more similar. Building on this idea of fair representation learning, Chierichetti et al. [54] adapt it to fair clustering, while Samadi et al. [55] apply it to fair dimensionality reduction (PCA).

2.1.3.2 In-processing Mitigations

In-processing methods are based on the idea of modifying the ML algorithms to account for fairness during the training time. For example, Kamishima et al. [56] suggest adding a regularization term to the objective function that penalizes the mutual information between the sensitive feature and the classifier's predictions. Zafar et al. [57] develop a technique that incorporates fairness constraints directly into the optimization objective of a classifier, targeting the reduction of disparate

mistreatment by balancing error rates across groups. Oneto et al. [58] propose a new optimization framework that extends the standard Empirical Risk Minimization proposition by constraining the optimization with task-related fairness criteria. Kamiran et al. [59] suggest adjusting a decision tree split criterion to maximize information gain between the split attribute and the class label while minimizing information gain with respect to the sensitive attribute. Zemel et al. [60] combine fair representation learning with an in-processing mitigation by applying a multi-objective loss function based on logistic regression, and Louizos et al. [61] apply this notion using a variational autoencoder. In [62], the notion of *privileged learning* is proposed in order to face cases where the sensitive information is available at training time but not at testing time.

2.1.3.3 Post-processing Mitigations

In post-processing approaches, a fair surrogate model is trained by observing the input-output relations of a pre-trained biased model. For example, Hardt et al. [35] propose a technique for flipping some decisions of a classifier to enhance equalized odds or equalized opportunity. In [63] and [64], authors suggest selecting separate thresholds for each group separately, in a manner that maximizes accuracy and minimizes demographic parity. Dwork et al. [65] propose a decoupling technique to learn a different classifier for each group, also combining a transfer learning technique with their procedure to learn from out-of-group samples.

2.1.4 Emerging Trends

As AI research evolves, algorithmic fairness is encountering a new set of challenges and emerging trends. These concerns and opportunities stem primarily from the simultaneous rise in both performance and complexity of AI models, which, on

one hand, has abruptly accelerated the widespread integration of AI systems into new and highly sensitive domains, while, on the other hand, has made interpreting and ensuring fairness across diverse contexts increasingly challenging.

2.1.4.1 New Research Frontiers and Challenges

It is central to recognize the unprecedented specificities brought about by recent vision and language models, such as more sophisticated data generation procedures (e.g., generative models), different training techniques (e.g., transfer learning and multi-step training), unusual inference procedures (e.g., prompting in LLMs), and the ever-increasing presence of large-scale models (e.g., Foundation Models), both in research and industry settings [66].

For instance, recent advances in the field of **image generation** have made it possible to create images of almost anything given a simple description [67]. Generative models are often trained on colossal amounts of unbalanced data, which may easily carry historical and societal biases. Even though there are currently no objective metrics specific for fairness in image generation, it is possible to measure whether a generative model is creating a similar number of samples for different groups by using a classifier on a large set of generated samples, verifying which group they belong to and then comparing these proportions [66]. Choi et al. [68] employed this strategy with a binary gender classifier to assess their face generation model, while Cho et al. [69] undertook a more extensive evaluation addressing both gender and ethnicity.

At the same time, these advances in image generation can also be an opportunity for unfairness mitigation: Generative Adversarial Networks (GANs) can be an option when looking for ways to increase a dataset with synthetic data [70; 71], since they can create high-quality new images when properly trained, balancing the potential misrepresentations in training or validation datasets for other mod-

els.

With regard to text generation, the rise of **LLMs** has introduced new dimensions to fairness concerns. For instance, a recent study [72] found that Google's Gemma model exhibited gender bias by downplaying women's health issues. In this context, most fairness metrics assess bias by analyzing model outputs and their associated probabilities. In addition, several tailor-made metrics have been proposed. For example, Direct Bias (DB) [73] is a gender-specific metric that quantifies bias as the projection of embeddings onto a gender subspace. The Word Embedding Association Test (WEAT) [74] and its extensions [75] measure bias via permutation tests, comparing associations between two sets of target words (e.g., male-dominated professions such as *engineer* versus female-dominated ones such as *nurse*) and two sets of attribute words (e.g., {*man*, *male*, ...} versus {*woman*, *female*, ...}). Similarly, co-occurrence metrics [76] rely on word co-occurrence patterns in generated text to quantify bias. Finally, Discovery of Correlations (DisCo) [77] combines template-based and generation-based methods to identify systematic output differences across groups of sensitive terms.

Analogously to the image generation case, also in the language context some new debiasing opportunities can be found. For instance, Qian et al. [78] propose a sequence-to-sequence model that generates perturbations regarding the protected variables in dataset instances: the new dataset thus comprises both original and demographically perturbed text snippets, creating a fairer and more balanced set. The authors show that pretraining and fine-tuning LLMs on demographically perturbed datasets lead to fairer models than using only the original data, while not decreasing performance.

Finally, with regard to **Deep Learning** more broadly, a further challenge arises from the widespread practice of initializing neural networks with pre-trained weights rather than training them from scratch: this approach accelerates training

and frequently enhances both performance and generalization, but it also makes the application of debiasing methods during the initial training stages impractical or undesirable. Consequently, debiasing strategies that adapt and refine already existing models, rather than starting from scratch, become particularly attractive to improve fairness [66].

2.1.4.2 Evolving Societal Demands and Perspectives

The recent evolution of AI models and performance has easily led to a more pervasive presence of automated "intelligent" systems even in high-stakes domains. At the same time, society's focus on equality and anti-discrimination issues has probably reached unprecedented levels, and discussions surrounding the ethical limitations to be imposed on automated systems are becoming increasingly prominent and often polarized.

Between the most natural emerging demands there is the need for continuous **monitoring and maintenance** of AI systems, to ensure sustained fairness in deployed systems. Indeed, while performance drift of prediction models is well-documented, the potential for algorithmic biases to emerge post-deployment has had limited characterization. A better understanding of how temporal model performance may shift across subpopulations is required to incorporate fairness drift into model maintenance strategies. In [79], authors highlight that algorithmic fairness cannot be assured through one-time assessments during model development: temporal changes in fairness may take multiple forms and interact with model updating strategies in unanticipated ways.

Another emerging challenge lies in the need to adopt an **intersectional perspective** on discrimination. Disadvantaged groups defined by combinations of sensitive attributes often exhibit dynamics that diverge substantially from those observed when attributes are analyzed individually. For instance, the biases ex-

perienced by dark-skinned women cannot be explained as a linear aggregation of biases against women and biases against dark-skinned individuals [41]. The detection and mitigation of such intersectional biases are particularly difficult due to the complex interdependencies among attributes [80; 81], thereby generating both statistical and ethical dilemmas in the assessment of fairness

As a final note, recent US and worldwide political shifts are sometimes leading to **skepticism** about AI fairness initiatives. Critics argue that efforts to promote fairness and inclusivity may be ideologically driven [82], potentially stalling progress in equitable AI development.

All of these developments highlight the dynamic nature of algorithmic fairness, emphasizing the need for ongoing research, interdisciplinary collaboration, and ethical considerations to navigate the complexities of AI in society.

2.2 Robustness

The study of robust and adversarial machine learning traces back to the early 2000s [16]. Early pioneering works, such as Dalvi et al. [83] and Lowd and Meek [84], investigated adversarial manipulation in spam filtering. They showed that linear classifiers could be deceived by modifying only a few words in spam messages, without compromising readability. Shortly after, Barreno et al. [85] broadened the perspective by questioning the reliability of machine learning in adversarial settings. They introduced a taxonomy of attacks, targeting both training and testing phases, and outlined possible countermeasures. A major turning point came in 2014 with the seminal work of Szegedy et al. [86], who demonstrated that deep neural networks for image recognition are vulnerable to imperceptible input perturbations. Since then, research in adversarial machine learning has expanded rapidly: developing attacks at both training [16; 87; 88; 89]

and testing [83; 90; 91; 92] stages, establishing systematic methodologies for assessing the security of learning algorithms [93; 94; 95], and proposing effective defense mechanisms [17; 96].

2.2.1 Defining Threats

2.2.1.1 Adversarial Robustness

Adversarial robustness refers to the ability of models to maintain their performance under potential adversarial attacks and perturbations [97].

Several studies [16; 85; 94] have proposed framing threats to ML systems in terms of adversarial attacks, by characterizing the attacker's objectives, knowledge, and capabilities.

- *Attacker's goals.* The attacker may aim to (i) evade detection without compromising the system, (ii) compromise the system integrity related to legitimate users, or (iii) obtain private information about the system, its users, or the data. Furthermore, the attack can be targeted or indiscriminate, e.g., the attacker may aim to cause misclassification of a specific set of samples (for instance, belonging to a particular class) or any sample.

- *Attacker's knowledge.* The attacker can have different levels of knowledge about the targeted ML system, including the training data and their features, the learning algorithm, the loss function, and the model parameters. This leads to white-, grey-, and black-box attacks.

- *Attacker's capabilities.* The attacks can also be categorized based on their influence on the data and whether they must adapt to some consistency constraints. Specifically, attacks are defined as **poisoning** if they can influence the model learning process through the manipulation of the training

data, or **evasion** if their influence is limited to just the manipulation of test samples.

2.2.1.2 Natural Robustness

Natural robustness (a.k.a. robustness against natural perturbations) is the capability of a model to preserve its performance under naturally-induced image corruptions or alterations [98]. Natural Perturbations are introduced through different types of commonly witnessed natural noise [99] (e.g., Gaussian noise in low lighting conditions [100]), and represent conditions more likely to occur in the real world compared to adversarial perturbations [98]. For instance, temporal perturbations are natural perturbations that hinder the capability of a model to detect objects in nearby frames in videos [101]. These perturbations result in a condition where the distribution of the test set differs from the one of the training set [102]. This condition is typically referred to in the literature with distribution shift [103], Out-of-Distribution (OOD) data [104], or data outside the training set [105].

2.2.2 Proposed Approaches for Improving Robustness

A large part of the literature explores methods to improve the robustness of AI models across their lifecycle: training data augmentation with malicious samples, in-model designing strategies, and post-training methodologies.

2.2.2.1 Processing the Training Data

The first class of approaches focuses on data manipulation, either by creating perturbations or by augmenting training sets with adversarial or naturally corrupted samples.

- *Generating adversarial attacks.* A number of papers tackle the challenge of developing methods to generate adversarial attacks that prove deep learning models brittle. The proposed methods vary w.r.t. the type of task targeted (e.g., NLP models [106], image classification [107], or object detection models [108]), the type of attack (e.g., sticking patches on the physical object to be recognized [99], or processing its digital input sample [108]), and the type of brittleness targeted (e.g., cases where the identified important features are not the correct ones [109]).

- *Augmenting data for adversarial robustness.* Many researchers focused on transforming [110; 111; 112], generating [113; 114; 115; 116], or employing ready-to-use [117] data and/or adversarial samples to extend or create datasets to train more robust models. When defending against adversarial attacks, GAN-based solutions are proven useful to generate adversarial samples [118], perturbations [116], and boundary samples [115]. While most methods apply complex transformations to improve robustness, simple transformations, like rotation [112] and image background removal [119], are still proven effective.

2.2.2.2 Designing Tn-model Robustness Strategies

This second cluster of methods deals with changes to training procedures and model architectures to embed robustness directly into the model.

- *Adversarial training.* Adversarial Training has proven to be a fundamental tool to build robust models: the basic idea is to simply create and incorporate adversarial examples into the training process, thus integrating an attack into the inner maximization procedure. Researchers have proposed a plethora of strategies [120; 121; 122; 123], also borrowing different ML

paradigms like self-supervised and unsupervised learning [124; 125]. The most common attack used in this context is the Projected Gradient Descent (PGD) [121], a common white-box algorithm (i.e., the attacker knows everything about the model). In recent years, Terzi et al. [126] and Gupta et al. [120] proposed extensions of PGD by using Wasserstein distance in the adversarial search space.

- *Robust architecture design.* Researchers have also investigated ways to make neural models robust from an architectural perspective [127]. For example, Jin et al. [128] introduced additive stochastic noise in the input layer of a CNN and re-parametrized the subsequent layers to take advantage of this additional information. Alternatively, Momeny et al. [129] introduce a CNN variant that is robust to noise by dynamically adapting both the striding of convolutions and the following pooling operations. Work by Xu et al. [130] operates on the classification layer by constraining its weights to be orthogonal.

2.2.2.3 Post-processing Methods

The last cluster looks at post-training modifications that adapt or combine models to enhance robustness — the main idea is that even after training, models can be "patched" with additional mechanisms that mitigate vulnerabilities.

- *Pruning.* Pruning (i.e., removing neurons and/or connections from a model) has become a popular compression approach that aims at reducing the computational cost of training models [131]. Recent literature in Robust AI has explored the use of pruning techniques or methodologies inspired by pruning to enhance model robustness. An example is the work by Chen et al. [106], which designed a methodology for selectively replacing ReLU neurons that are identified as unstable.

- *Fusing models.* Another approach for achieving post-model-training robustness consists of plugging additional models into a trained model. These additional models can be used to identify and deal with problematic data instances (e.g., OOD, mistaken [132], noisy [133], or adversarially modified [134] occurrences). In the first case, in the context of NLP, Pruthi et al. [132] attach a task-agnostic word recognition model to a classification model as a means to defend the main classifier against spelling mistakes. In the context of Computer Vision, Ye et al. [134] use an additional classifier to determine real vs. adversarially manipulated data instances. A prominent line of work in this field consists in using GANs as auxiliary models— e.g., Sun et al. [115] use a Boundary Conditional GAN to generate boundary samples.

2.2.3 Emerging Trends

With the proliferation of generative AI, the threat landscape has expanded significantly beyond traditional models to target new unique complexities. In LLMs, primary attack vectors include jailbreaking [135], which aims to circumvent safety alignment to generate harmful content, and data poisoning [136], which manipulates training data to induce malicious behavior. Recent research on data poisoning has shown that increasing an LLM's parameter size does not inherently enhance its resilience [137] and that hallucinations [138] emerge even when more sophisticated prompting strategies are used, e.g., in-context learning [139] or chain-of-thought prompting [140]. Only retrieval-augmented generation [141] appears to mitigate hallucinations.

The robustness field is also moving toward a deeper understanding of the underlying nature of AI vulnerabilities. A new line of research proposes a symbiotic relationship between adversarial robustness and interpretability. These works

rely on the idea that when the model features extracted via an explainability method are aligned with human reasoning (i.e., the features are meaningful for a human to make a prediction for a data sample), then the model should be more robust. This research agenda seeks to transition the field from a reactive, cybersecurity-like posture to a more foundational science by investigating the causal relationship between a model's internal workings and its vulnerability. In terms of evaluation, Li et al. [142] have investigated how explainability can be used in order to evaluate the robustness of a model. In terms of improvement, Kortylewski et al. [143] proposed Compositional Neural Networks, a unification of CNNs with part-based models (inherently interpretable models), and show that these new networks increase model robustness.

2.3 Regression

While the notions of fairness and robustness are by now quite well-known ones, the concept of regression in ML is still relatively new in the literature. In classical software development, the term *regression* refers to a decline in software performance or functionality following an update. Also ML-based systems often require updates for various reasons [144], such as the availability of new data or models and the need to optimize different technical or ethical metrics. As a consequence, the practice of delivering frequent ML model updates is opening up to the *regression* challenge also for ML predictions [145; 146]. Indeed, updated versions of ML models are usually designed to improve the average performance (e.g., the performance on a series of predictions) not taking into account the sample-wise performance (e.g., performance on specific predictions), possibly introducing a *regressive* behavior (also called *backward-compatibility* error [147]). For example, in classification, an update may be able to decrease the average

number of misclassifications while introducing misclassifications on samples that were previously correctly classified. These newly introduced misclassifications — also called negative flips (NFs) — are thoroughly studied by Yan et al. in [145], and the need for reducing them is now starting to pose challenges in different applications. A paradigmatic example is malware detection via ML-based classifiers: if the updated model misses malwares that were previously detected, this is clearly an important security issue, even if the overall detection performance improves [148].

2.3.1 Proposed Mitigations

In [145], the authors introduce focal distillation, a method that trains the new model to partly match the predictions of the old model by minimizing the Kullback-Leibler divergence between their outputs, and Xie et al. [149] show the applicability of this technique on text-based models. Building on these works, Zhao et al. [150] propose Ensemble Logit Difference Inhibition (ELODI), a technique to distill the ensemble of new models into a single model.

In [151], the authors analyze the problem of regression during data updates and propose Backward Compatible Weight Interpolation (BCWI), a method that interpolates between the weights of the old and new model reducing negative flips without sacrificing the improved accuracy of the new model.

In [148], the problem of malware detection models in a Continual Learning setting is addressed by proposing a learning strategy that combines a replay-based continual learning method with a regression-aware penalty to preserve the correct decisions of earlier models. In [152], the authors show that negative flips and overall error rates are decoupled, and then develop a plug-in algorithm that leverages a tailor-made validation set to significantly reduce the negative flip rate.

In [153], the Prediction Update Problem is formalized and a probabilistic

approach for backward-compatible prediction updates is developed. Matsuno et al. [154] propose a backward compatibility metric focusing on the compatibility of local predictive performance. Furthermore, they propose a model-retraining method based on the proposed metric. Finally, [147] studies challenges with backward compatibility across different ML architectures and datasets, focusing on common settings including data shifts with structured noise and ML employed in inferential pipelines.

2.4 Mutual Interactions

In the past decade, it has become increasingly evident that there is a pressing need to unify the currently fragmented approaches to trustworthy AI. One of the most comprehensive works in this area, by Oneto et al. [11], investigate for the first time what happens when multiple technical and ethical metrics — often in tension with one another [30] — need to be simultaneously optimized. Their findings are promising, suggesting that it is theoretically feasible to design consistent algorithms capable of satisfying multiple trustworthiness criteria.

As discussed previously, the central aim of this work is to analyze and mitigate the ways in which fairness interacts with robustness and regression. While the interplay between fairness and robustness has recently begun to receive growing attention, giving rise to some interesting emerging research (see Section 2.4.1), the interaction between fairness and regression (see Section 2.4.2) remains largely unexplored.

2.4.1 Fairness and Robustness

The relationship between fairness and robustness, and the ways in which one contributes to the other, has been receiving increasing attention [155]. Rezaei

et al. [156] aim to make classifications that have robust fairness without relying on previously labeled data, as these may carry some inherent biases, formulating the fair prediction problem as a game between two adversaries. Similarly, Yurochkin et al. [157] propose an adversarial approach to fairness, using a distributionally robust approach to enforce individual fairness during training. Some recent works [158; 159] show that it may be easier for an attacker to target a particular subgroup, resulting in a form of *robustness bias*, and proposing methods to measure this form of bias.

Also the robustness tradeoff with fairness has been explored in recent literature. In [160], the authors show that fairness can be in conflict with robustness, and Benz et al. [161] show that adaptively adjusting class-wise loss weights negatively affects overall robustness and accuracy performance. Xu et al. [29] hypothesize that adversarial training algorithms tend to introduce severe disparity in accuracy and robustness between different groups of data. Stemming from this work, Ma et al. [162] show that adversarial training easily leads to more serious robust fairness issue than natural training and propose a fairly adversarial training. On the other hand, Pruksachatkun et al. [163] studied if an increase in robustness can improve fairness in text-classification tasks. They found that certified robustness methods improve fairness, and using both robustness and bias mitigation methods in training may result in an improvement for both.

2.4.2 Fairness and Regression

In [164], Bertran et al. note that the current notion of regression may have shortcomings in measuring how negative flips arise in different populations, and discuss the possibility to update a model over a newly augmented dataset in a way that guarantees a desired backward compatibility measure across all groups. While not tackling regression specifically, Davis et al. [79] focus on the concept

of fairness drift — either due to differences in dataset shift across subpopulations or to the model updating process — and raise concerns that models previously assessed as fair may then begin to perpetuate inequities.

Contributions Overview

The works presented in this section describe two novel contributions exploring, respectively, the interplay between Fairness and Regression and the one between Fairness and Robustness. In particular:

(1) In the first work, *Mitigating Unfair Regression in Machine Learning Model Updates* [165], we show, for the first time, the existence of the phenomenon of *unfair regression* and propose different ethical metrics to measure it. Additionally, we offer two mitigation strategies — one focused on modifying the learning algorithm and one focused on modifying the tuning phase — to address the issue. Our results on real-world datasets confirm the existence of the unfair regression phenomenon and demonstrate the effectiveness of the proposed mitigation strategies.

This research has been conducted with the valuable contribution of Anna Pallarès López, Eduard Martín Jiménez, Davide Anguita, Fabio Roli, and Luca Oneto.

(2) In *Towards Robust Fair Empirical Risk Optimization: a Theoretical Perspective*, we address the challenge of mitigating the robustness bias of ML models — i.e., to reduce the effectiveness of malicious adversarial attacks on discriminated subgroups within the population. Our approach involves establishing a specific threat model, which is the case of white-box attacks

under an L_2 ball constraint, and employing a recently introduced fairness notion, $\epsilon-$Fairness Revised [166] to address robustness bias under the specified threat model. We then introduce the Robust Fair Empirical Risk Optimization (RFERM) framework and examine its theoretical consistency.

The research stems from the fundamental contribution of Danilo Franco and Luca Oneto.

Chapter 3

Mitigating Unfair Regression in Machine Learning Model Updates

Summary

Machine learning systems often require updates for various reasons, such as the availability of new data or models and the need to optimize different technical or ethical metrics. Typically, these metrics reflect an average performance rather than sample-wise behavior. Indeed, improvements in metrics like accuracy can introduce negative flips, where the updated model makes errors that the previous model did not make. In certain applications, these negative flips can be perceived by developers or users as a regression in performance, contributing to the hidden technical debt of machine learning systems. Moreover, if the distribution of negative flips is biased with respect to some sensitive attribute (e.g., gender or race), it may be perceived as discrimination, termed unfair regression. In this paper we show, for the first time, the existence of the phenomenon of unfair regression

and propose different ethical metrics to measure it. Additionally, we offer two mitigation strategies — one focused on modifying the learning algorithm and one focused on modifying the tuning phase — to address this issue. Our results on real-world datasets confirm the existence of the unfair regression phenomenon and demonstrate the effectiveness of the proposed mitigation strategies.

3.1 Introduction

As previously discussed, while ML solutions seem to become day by day more powerful and easy to deploy, it is also becoming more and more evident that this does not come without shortcomings. An effective metaphor to describe this dichotomy is the concept of *hidden debt*, introduced by Ward Cunningham in 1992 [167] to reason about the long-term costs of ML systems. This concept is multifaceted and includes by now both a technical debt [144] (i.e., the one related to the need for models to be updated and maintained over time) and an ethical debt [168] (i.e., the one deriving from not taking proactively into account the ethical implications of a system in the design phase).

Our work tackles the intersection between technical and ethical debt, focusing in particular on the study and mitigation of the mutual interaction between *regression* (i.e., the tendency of updated models to fail predictions that were correctly performed by older versions of the model [145]) and *unfairness* (i.e., the tendency of algorithms to perpetuate or amplify historical biases against sensitive groups [50]).

Indeed, updated versions of ML models are usually designed to improve the average performance (e.g., the performance on a series of predictions) not taking into account the sample-wise performance (e.g., performance on specific predictions), possibly introducing a *regressive* behavior. Regression is, per se, already

a problem when dealing with specific applications (e.g., a person may see its loan incorrectly blocked because of a change in the model), but it becomes even more problematic when it disproportionately affects historically discriminated subgroups (e.g., dark-skinned people). While both regression and unfairness have been already studied individually, we propose, for the first time in the literature, to focus on their interaction.

Hence, in this chapter we (i) formally define the unfair regression phenomenon proposing different metrics to describe and measure it, (ii) show that this phenomenon actually happens in real-world applications, and (iii) proposes two mitigation strategies, one focused on modifying the learning algorithm while the other one focused on modifying the tuning phase (specifying both strategies to the case of Support Vector Machines [169]), to address the unfair regression phenomenon showing its actual effectiveness in real-world applications.

The rest of the work is organized as follows. Section 3.2 introduces some preliminary concepts and defines the unfair regression phenomenon proposing different metrics to describe and measure it. Section 3.3 proposes two mitigation strategies, one focused on modifying the learning algorithm and one focused on modifying the tuning phase, to address the unfair regression phenomenon. Section 3.4 will show the actual existence of the unfair regression phenomenon and the effectiveness of our proposed mitigation strategies on a series of real-world datasets coming from the financial, medical, and legal fields. Finally, Section 3.5 concludes the chapter, highlighting strengths and weaknesses of the work, and identifying possible future developments.

3.2 Unfair Regression

Let us consider the dataset $\mathcal{D}_n = \{(x_1, s_1, y_1), \ldots, (x_n, s_n, y_n)\}$ to be a sequence of n samples identically and independently drawn from a joint probability distribution μ over $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space[1] [170; 171], $\mathcal{S} = \{a, b\}$ is the space of the sensitive attributes[2] [20], and $\mathcal{Y} = \{-1, +1\}$ is the output space[3] [170]. Ideally, an ML algorithm $\mathscr{A}_{\mathcal{H}}$, characterized by a set of hyperparameters $\mathcal{H}$, selects a model $f : \mathcal{X} \to \mathcal{Y}$ in a space of possible ones $\mathcal{F}$ based on $\mathcal{D}_n$ to well approximate $\mathbb{P}\{y|x\}$ [170]. $\mathcal{F}$ is implicitly or explicitly induced by the choice of $\mathscr{A}$ and $\mathcal{H}$ [170; 171]. Examples of ways to modify $\mathcal{F}$ are the regularizers (e.g., p-norm of the weights) [171], the model functional form (e.g., convolutions or transformers) [171] or the optimization process (e.g., stochastic gradient descent with early stopping and dropout) [171; 172]. The quality of the approximation f is measured with respect to a loss function $\ell(f(x), y)$ which allows to define the risk of the model[4] $\mathtt{L}(f) = \mathbb{E}_{x,y}\ell(f(x), y)$ [173]. Since μ is unknown we can only measure the empirical risk $\hat{\mathtt{L}}(f, \mathcal{D}_n) = \mathbb{E}^{\mathcal{D}_n}_{x,y}\ell(f(x), y)$, where $\mathbb{E}^{\mathcal{D}_n}$ denotes the empirical expectation computed over $\mathcal{D}_n$. Given the set of all measurable functions $\mathcal{G}$ the oracle algorithm [174] can be formulated as

$$g^* = \arg\min_{g \in \mathcal{G}} \mathtt{L}(g), \tag{3.1}$$

[1] Which can be unstructured (i.e., tabular data) or structured (e.g., images or graphs).

[2] For the sake of notation simplicity we restrict to a binary-valued sensitive attribute (e.g., male/female or dark-/light-skinned individuals) but the approach can be easily generalized to any space.

[3] For the sake of notation simplicity we restrict to the binary classification but the results can be easily generalized to the whole supervised learning setting.

[4] Note that the presentation can be easily generalized to multiple technical and ethical metrics [11] but this is out of the scope of this paper.

but, since μ is unknown and $\mathcal{G}$ too large, $\mathscr{A}_{\mathcal{H}}$ is generally formulated as an Empirical Risk Minimization (ERM) problem [170]

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{\mathrm{L}}(f, \mathcal{D}_n). \tag{3.2}$$

The latter problem is affected by both approximation (wrong choice of $\mathcal{F} \subseteq \mathcal{G}$) and estimation (limited amount of data in $\mathcal{D}_n$) errors preventing to find exactly g^* [174]. By tuning $\mathcal{F}$, namely $\mathscr{A}$ and $\mathcal{H}$, we can strongly influence the performance (risk) of $\hat{f}$, getting closer to the one of g^* [174; 175]. In order to perform this tuning phase, the most common approach is to use resampling methods [175]: $\mathcal{D}_n$ is split, one or r times with or without replacement, into learning $\mathcal{L}_l^i$ and validation $\mathcal{V}_v^i$ sets with $i \in \{1, \cdots, r\}$ such that $\mathcal{L}_l^i \cup \mathcal{V}_v^i = \mathcal{D}_n$ and $\mathcal{L}_l^i \cap \mathcal{V}_v^i = \oslash$. Then, in order to select the best algorithm and hyperparameters configurations $\mathscr{A}_{\mathcal{H}^*}^*$, in a set of possible ones $\mathcal{A}_{\mathcal{H}} = \{\mathscr{A}_{\mathcal{H}_1}^1, \mathscr{A}_{\mathcal{H}_2}^1, \cdots, \mathscr{A}_{\mathcal{H}_1}^2, \mathscr{A}_{\mathcal{H}_2}^2, \cdots\}$ the following procedure has to be applied

$$\begin{aligned} f^* &= \mathscr{A}_{\mathcal{H}^*}^*(\mathcal{D}_n), \\ \mathscr{A}_{\mathcal{H}^*}^* &= \arg\min_{\mathscr{A}_{\mathcal{H}} \in \mathcal{A}_{\mathcal{H}}} \sum_{i=1}^{r} \hat{\mathrm{L}}(\mathscr{A}_{\mathcal{H}}(\mathcal{L}_l^i), \mathcal{V}_v^i), \end{aligned} \tag{3.3}$$

where $\mathscr{A}_{\mathcal{H}}(\mathcal{L}_l^i)$ is a model $\hat{f}$ built with the algorithm $\mathscr{A}$ with its set of hyperparameters $\mathcal{H}$ and with the data $\mathcal{L}_l^i$ and $\mathscr{A}_{\mathcal{H}^*}^*(\mathcal{D}_n)$ is a model f^* built with the algorithm $\mathscr{A}^*$ with its set of hyperparameters $\mathcal{H}^*$ and with the data $\mathcal{D}_n$. Since the data in $\mathcal{L}_l^i$ are independent of the data in $\mathcal{V}_v^i$, $\mathscr{A}_{\mathcal{H}^*}^*$ should be an algorithm and the associated set of hyperparameters which allows achieving a small risk on previously unseen data, i.e., f^* should be close to g^*.

When new data becomes available or new algorithms are proposed in the literature, it is then mandatory to ensure optimal quality of the deployed model by updating it [145; 146]. For this reason, let us define f_{old}^* the "old" model (i.e.,

the model trained with the original data and algorithms) and f^*_{new} the "new" model (i.e., the model trained with the updated data and set of algorithms). Even if μ is unknown we expect that the performance of f^*_{new} improves the one of f^*_{old}

$$\mathtt{L}(f^*_{\text{new}}) \leq \mathtt{L}(f^*_{\text{old}}), \tag{3.4}$$

since, e.g., we tested f^*_{old} and f^*_{new} on a fresh set of previously unseen data (test set). Even if assuming Eq. (3.4) to be true we may experience the following phenomenon: for certain $(x, y) \in (\mathcal{X} \times \mathcal{Y})$

$$\ell(f^*_{\text{new}}(x), y) > \ell(f^*_{\text{old}}(x), y), \tag{3.5}$$

namely there is a regression in performance for certain samples even if the average performance increased (Eq. (3.4)). As we deal with binary classification, the problem arises when x was correctly classified by f^*_{old} but not by f^*_{new}, i.e., there is a negative flip [145]:

$$f^*_{\text{new}}(x) \neq y \wedge f^*_{\text{old}}(x) = y. \tag{3.6}$$

The greater the number of negative flips the worse the developers or the users perceive the behavior of f^*_{new}.

In this work, we argue that the regression, apart from being a problem for developers or the users per se, may lead to other problems like discrimination for subgroups of the population. For example, one can observe a different rate of negative flips in the different subgroups of the population. More formally, we

would like to observe a small value for the following quantity

$$\mathtt{UR}^{\mathrm{DP}}_{f^*_{\mathrm{old}}}(f^*_{\mathrm{new}}) = \left|\mathbb{E}^{\mathrm{old},a}_{x,y}[f^*_{\mathrm{new}}(x) \neq y] - \mathbb{E}^{\mathrm{old},b}_{x,y}[f^*_{\mathrm{new}}(x) \neq y]\right| \tag{3.7}$$

named Unfair Regression (UR), where the Iverson braked notation has been used and $\mathbb{E}^{\mathrm{old},\circ}_{x,y} = \mathbb{E}_{\{(x,s,y):f^*_{\mathrm{old}}(x)=y, s=\circ\}}$ with $\circ \in \mathcal{S}$. Of course, Eq. (3.7) is a measure of discrimination as, in case of model update, people in a subgroup will observe a larger fraction of negative flips with respect to the other subgroup. The discrimination measure of Eq. (3.7) is the counterpart, for negative flips, of the Demographic Parity (DP) in algorithmic fairness [49]. Another notion, coming in this case from the Equal Opportunity (EO) in algorithmic fairness [35], can be formalized as

$$\mathtt{UR}^{\mathrm{EO}}_{f^*_{\mathrm{old}}}(f^*_{\mathrm{new}}) = \left|\mathbb{E}^{\mathrm{old},a,1}_{x,y}[f^*_{\mathrm{new}}(x) \neq y] - \mathbb{E}^{\mathrm{old},b,1}_{x,y}[f^*_{\mathrm{new}}(x) \neq y]\right| \tag{3.8}$$

where $\mathbb{E}^{\mathrm{old},\circ,1}_{x,y} = \mathbb{E}_{\{(x,s,y):f^*_{\mathrm{old}}(x)=y, s=\circ, y=1\}}$ for $\circ \in \mathcal{S}$ — i.e., we care about the negative flips only when they appear in the positively labeled samples (e.g., we do not grant a credit to someone who deserves it and that f^*_{old} actually granted).

3.3 Mitigation Strategies

In this section we will propose two ways to make updates, i.e., going from f^*_{old} to f^*_{new}, in such a way to continue to improve the accuracy of f^*_{new} with respect to f^*_{old}, in the sense of Eq. (3.4), but also minimizing the unfair regression, in the sense of Eq. (3.7) or Eq. (3.8).

The first step in this direction is to note that, analogously to the risk of a model, $\mathtt{UR}^{\mathrm{DP}}_{f^*_{\mathrm{old}}}(f^*_{\mathrm{new}})$ and $\mathtt{UR}^{\mathrm{EO}}_{f^*_{\mathrm{old}}}(f^*_{\mathrm{new}})$ are unknown since μ is unknown. As a

consequence, we can only compute their empirical counterparts, namely

$$\hat{\mathtt{UR}}^{\mathrm{DP}}_{f^*_{\mathrm{old}}}(f^*_{\mathrm{new}}, \mathcal{D}_n) = \left| \mathbb{E}^{\mathcal{D}_n^{\mathrm{old},a}}_{x,y}[f^*_{\mathrm{new}}(x) \neq y] - \mathbb{E}^{\mathcal{D}_n^{\mathrm{old},b}}_{x,y}[f^*_{\mathrm{new}}(x) \neq y] \right| \tag{3.9}$$

with $\mathcal{D}_n^{\mathrm{old},\circ} = \{(x,y,s) : (x,y,s) \in \mathcal{D}_n, f^*_{\mathrm{old}}(x) = y, s = \circ\}$ for $\circ \in \mathcal{S}$ and

$$\hat{\mathtt{UR}}^{\mathrm{EO}}_{f^*_{\mathrm{old}}}(f^*_{\mathrm{new}}, \mathcal{D}_n) = \left| \mathbb{E}^{\mathcal{D}_n^{\mathrm{old},a,1}}_{x,y}[f^*_{\mathrm{new}}(x) \neq y] - \mathbb{E}^{\mathcal{D}_n^{\mathrm{old},b,1}}_{x,y}[f^*_{\mathrm{new}}(x) \neq y] \right| \tag{3.10}$$

with $\mathcal{D}_n^{\mathrm{old},\circ,1} = \{(x,y,s) : (x,y,s) \in \mathcal{D}_n, f^*_{\mathrm{old}}(x) = y, s = \circ, y = 1\}$ for $\circ \in \mathcal{S}$.

To mitigate unfair regression, we have two primary strategies: (i) modifying the learning algorithm to account for the unfair regression phenomenon, similarly to the positive congruent training approach described by [145] for handling negative flips; or (ii) adjusting the tuning (model selection) phase, since we aim to optimize two different metrics simultaneously (accuracy and unfair regression).

In pursuing these strategies, it is crucial to recognize that accuracy and unfair regression can often be in conflict. Indeed, new algorithms and datasets might not be able to both improve accuracy and minimize unfair regression at the same time. This conflict can arise from various factors, such as the original dataset not being representative, shifts in the distribution of the sensitive feature in the new dataset, shifts in the distribution over the input space or labeling process, or the new algorithm's inability to approximate the function identified by the old algorithm.

Since any strategy to mitigate unfair regression will require careful tuning of hyperparameters or choosing between algorithms, we will start with (ii), i.e., adjusting the tuning phase, and then (i), i.e., propose a mitigation strategy modifying the learning algorithm.

3.3.1 Mitigation (i) - Adjusting the Tuning Phase

Mitigating unfair regression, in all cases, requires careful tuning of hyperparameters and, in our case, we cannot rely solely on one metric (such as accuracy, as we did in Eq. (3.3)), because this would likely neglect the unfair regression aspect. This issue is similar to challenges encountered when optimizing for both accuracy and fairness, as discussed by [26], or when balancing multiple ethical and technical metrics, as highlighted by [11].

With this problem in mind, as in general we cannot simultaneously optimize accuracy and unfair regression, the simplest approach is to set a level of improvement in accuracy that we are willing to lose to mitigate the unfair regression. More formally, the procedure of Eq. (3.3) can be modified as follows

$$
\begin{aligned}
&f^*_{\text{new}} = \mathscr{A}^*_{\mathcal{H}^*}(\mathcal{D}_n), \\
&\mathscr{A}^*_{\mathcal{H}^*} = \arg\min_{\mathscr{A}^\delta_{\mathcal{H}} \in \mathcal{A}_{\mathcal{H}}} \textstyle\sum_{i=1}^{r} \hat{\mathtt{UR}}^{\diamond}_{f^*_{\text{old}}}(\mathscr{A}_{\mathcal{H}}(\mathcal{L}^i_l), \mathcal{V}^i_v), \\
&\mathcal{A}^\delta_{\mathcal{H}} = \{\mathscr{A}_{\mathcal{H}} : \mathscr{A}_{\mathcal{H}} \in \mathcal{A}_{\mathcal{H}}, \textstyle\sum_{i=1}^{r} \hat{\mathtt{L}}(\mathscr{A}_{\mathcal{H}}(\mathcal{L}^i_l), \mathcal{V}^i_v) \leq (1+\delta)\hat{\mathtt{L}}^*\}, \\
&\hat{\mathtt{L}}^* = \min_{\mathscr{A}_{\mathcal{H}} \in \mathcal{A}_{\mathcal{H}}} \textstyle\sum_{i=1}^{r} \hat{\mathtt{L}}(\mathscr{A}_{\mathcal{H}}(\mathcal{L}^i_l), \mathcal{V}^i_v),
\end{aligned}
\tag{3.11}
$$

where δ is a fraction of additional error that we are willing to bear in order to reduce the unfair regression, and $\diamond \in \{\text{DP}, \text{EO}\}$ depending on the notion of unfair regression we want to use.

Once the procedure is concluded we need to estimate, e.g., using a test set, $\mathtt{L}(f^*_{\text{old}})$, $\mathtt{L}(f^*_{\text{new}})$, and $\mathtt{UR}^{\diamond}_{f^*_{\text{old}}}(f^*_{\text{new}})$ and check that

$$
\mathtt{L}(f^*_{\text{old}}) \geq \mathtt{L}(f^*_{\text{new}}) \tag{3.12}
$$

otherwise the update does not make sense, and that $\mathtt{UR}^{\diamond}_{f^*_{\text{old}}}(f^*_{\text{new}})$ is acceptable to the specific application.

3.3.2 Mitigation (ii) - Modifying the Learning Algorithm

The limitation of acting on just the tuning phase, as described in Section 3.3.1, is that the training phase, i.e., the ERM of Eq. (3.2), remains unaware of our desire to mitigate the unfair regression. Instead, our goal would be to find the following new oracle

$$g^*_{\text{new}} = \arg\min_{g \in \mathcal{G}} \mathtt{L}(g), \text{ s.t. } \mathtt{UR}^{\diamond}_{f^*_{\text{old}}}(g) = 0, \tag{3.13}$$

with $\diamond \in \{\text{DP}, \text{EO}\}$ depending on the notion of unfair regression exploited. In other words, we search for the best model in all the measurable functions with no unfair regression with respect to f^*_{old}. As for ERM, since μ is unknown we can just compute the empirical counterparts of $\mathtt{L}$ and $\mathtt{UR}^{\diamond}_{f^*_{\text{old}}}$ in Eq. (3.13). Then, as minimizing with respect to $\mathcal{G}$ would surely lead to overfitting (as $\mathcal{G}$ is too large), we have to replace $\mathcal{G}$ with a carefully tuned $\mathcal{F} \subset \mathcal{G}$. Another problem is the constraint on the unfair regression set to zero, which again may lead to overfitting in the sense of low unfair regression on the data but not in the population, and so we have to relax it. The result is our proposal for Unfair-Regression-Free ERM (URFERM) that can be defined as follows

$$\hat{f}_{\text{new}} = \arg\min_{f \in \mathcal{F}} \hat{\mathtt{L}}(f, \mathcal{D}_n), \text{ s.t. } \hat{\mathtt{UR}}^{\diamond}_{f^*_{\text{old}}}(f, \mathcal{D}_n) \leq \epsilon, \tag{3.14}$$

where $\diamond \in \{\text{DP}, \text{EO}\}$ depending on the notion of unfair regression exploited and ϵ is a small tunable quantity of unfair regression allowed on the training data so to perform better on the population. By tuning $\mathcal{F}$ and $\epsilon \in [0, 1]$, e.g., with the approach proposed in Section 3.3.1, we obtain f^*_{new}.

Problem (3.14) is of course non-convex and non-differentiable. By properly designing $\mathcal{F}$ and relaxing $\mathtt{L}$ and $\mathtt{UR}^{\diamond}_{f^*_{\text{old}}}$ [170; 176] we can derive the Unfair-Regression-Free version of all the main algorithms available in the literature. In the next

section we will specify Problem (3.14) to the case of Support Vector Machines.

3.3.3 Unfair-Regression-Free Support Vector Machines

Let us consider the Support Vector Machines for binary classification [169] where $f(x) = \sum_{i=1}^{n} \alpha_i \kappa(x_i, x) + \alpha_0$ with $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_n]^T \in \mathbb{R}^n$, $\alpha_0 \in \mathbb{R}$, and κ is a kernel function, $\mathcal{F} = \{f : f(x) = \sum_{i=1}^{n} \alpha_i \kappa(x_i, x) + \alpha_0, \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \kappa(x_i, x_j) \leq W\}$ with $W \in [0, \infty)$, and where the desired ℓ to be minimized, i.e., the misclassification error $\ell_H(f(x), y) = [yf(x) \leq 0]$, is relaxed into a convex version, i.e., the hinge loss $\ell_\xi(f(x), y) = \max[0, 1 - yf(x)]$. By adopting the same approach for the URFERM of Problem (3.14) we obtain

$$
\begin{aligned}
\min_{\boldsymbol{\alpha}, \alpha_0} \quad & \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \xi_i, \\
\text{s.t.} \quad & \xi_i = \max\left[0, 1 - y_i f_i\right], \ i \in \{1, \cdots, n\}, \\
& f_i = \textstyle\sum_{j=1}^{n} \alpha_j \kappa(x_j, x_i) + \alpha_0, \ i \in \{1, \cdots, n\}, \\
& \textstyle\sum_{j=1}^{n} \alpha_i \alpha_j \kappa(x_i, x_j) \leq W, \\
& \left| \tfrac{1}{n^{\text{old},a}} \textstyle\sum_{i \in \mathcal{J}^{\text{old},a}} \xi_i - \tfrac{1}{n^{\text{old},b}} \textstyle\sum_{i \in \mathcal{J}^{\text{old},b}} \xi_i \right| \leq \epsilon,
\end{aligned}
\tag{3.15}
$$

where $\mathcal{J}^{\text{old},\circ} = \{i : i \in \{1, \cdots, n\}, f^*_{\text{old}}(x_i) = y_i, s_i = \circ\}$ and $n^{\text{old},\circ} = |\mathcal{J}^{\text{old},a}|$ with $\circ \in \mathcal{S}$ since we leveraged $\texttt{UR}^{\text{DP}}_{f^*_{\text{old}}}(f^*_{\text{new}})$ as notion of unfair regression[1].

Note that trivially using the SVM approximations in URFERM leads to the last constraint for unfair regression, which leads to a non-convex relaxation. Nevertheless, the objective of the constraint is to make the rate of negative flips equally distributed on the two subgroups. This can be only achieved by (hopefully) decreasing or increasing the negative flips in the two groups. So, without

[1]The case of $\texttt{UR}^{\text{EO}}_{f^*_{\text{old}}}(f^*_{\text{new}})$ is not explicitly reported as analogous to the one derived for $\texttt{UR}^{\text{DP}}_{f^*_{\text{old}}}$.

losing the idea of the constraint we can better relax it as follows

$$\begin{aligned} \min_{\boldsymbol{\alpha},\alpha_0} \quad & \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\xi_i, \\ \text{s.t.} \quad & \xi_i = \max\left[0, 1-y_i f_i\right],\ i \in \{1,\cdots,n\}, \\ & f_i = \textstyle\sum_{j=1}^{n}\alpha_j\kappa(x_j,x_i)+\alpha_0,\ i \in \{1,\cdots,n\}, \\ & \textstyle\sum_{j=1}^{n}\alpha_i\alpha_j\kappa(x_i,x_j) \leq W, \\ & \tfrac{1}{n^{\text{old},a}}\textstyle\sum_{i\in\mathcal{J}^{\text{old},a}}\xi_i \leq \epsilon \vee \tfrac{1}{n^{\text{old},b}}\textstyle\sum_{i\in\mathcal{J}^{\text{old},b}}\xi_i \leq \epsilon, \end{aligned} \tag{3.16}$$

meaning we force the decrease of negative flips in one of the two groups and we leave the ability to increase them in the other. Note that Problem (3.16) is actually composed of two convex problems where we have to decide which constraint to activate (the ones in the logical or) and we can see it as an hyperparameter to tune during the tuning phase of Section 3.3.1.

Finally, note that Problem (3.16) is hard to solve since it is formulated as an Ivanov regularization problem [177], so we can formulate it as follows

$$\begin{aligned} \min_{\boldsymbol{\alpha},\alpha_0} \quad & \boldsymbol{v}^1\boldsymbol{\xi} + \lambda_1\boldsymbol{\alpha}Q\boldsymbol{\alpha} + \lambda_2(\boldsymbol{v}^{\text{old},a}\boldsymbol{\xi} \vee \boldsymbol{v}^{\text{old},b}\boldsymbol{\xi}) \\ \text{s.t.} \quad & \xi_i = \max\left[0, 1-y_i f_i\right],\ i \in \{1,\cdots,n\}, \\ & f_i = \textstyle\sum_{j=1}^{n}\alpha_j\kappa(x_j,x_i)+\alpha_0,\ i \in \{1,\cdots,n\}, \end{aligned} \tag{3.17}$$

with $\boldsymbol{\xi} = [\xi_1,\cdots,\xi_n]$, $v_i^1 = 1\ \forall i \in \{1,\cdots,n\}$, $v_i^{\text{old},a} = 1$ if $i \in \mathcal{J}^{\text{old},a}$ otherwise $v_i^{\text{old},a} = 0$, $v_i^{\text{old},a} = 1$ if $i \in \mathcal{J}^{\text{old},a}$ otherwise $v_i^{\text{old},a} = 0$, and $Q_{i,j} = \kappa(x_i,x_j)$ $\forall i,j \in \{1,\cdots,n\}$.

Note that Problem (3.17) is a simple Support Vector Machine with different weights on each training point and whose hyperaparameters are: λ_1, λ_2, the kernel and its hyperparameters, and the choice in the logical or.

Name	Ref.	# samples	# features	Sensitive Attribute
German	UCI	1000	20	Foreign
Arrhythmia	UCI	452	279	Gender
Compas	ProPublica	6172	10	Ethnicity
Adult	UCI	32561	12	Gender

Table 3.1: The datasets used, the related statistics, and the sensitive features involved. Gender considers the two groups as male and female; ethnicity considers the ethnic groups as white and other; foreign considers being foreign or not.

3.4 Results

In this section, we will test how the methodology presented in Section 3.3 actually perform on real-world datasets. For this purpose, we considered 4 publicly available datasets (from the standard databases UCI[1] and ProPublica[2]), reported in more detail in Table 3.1. We considered the following models: Naive Linear SVM [169], Naive Gaussian-Kernel SVM [169], Linear SVM with Mitigation (i) (Section 3.3.1), Gaussian-Kernel SVM with Mitigation (i) (Section 3.3.1), Linear SVM with Mitigation (i) plus (ii) (Sections 3.3.1 and 3.3.3), and Gaussian-Kernel SVM with Mitigation (i) plus (ii) (Sections 3.3.1 and 3.3.3). For Naive Linear SVM we tuned the regulatization hyperparameter in $10^{\{-4.0,-3.75,\cdots,3.00\}}$, for Naive Gaussian-Kernel SVM we tuned the regulatization hyperparameter in $10^{\{-4.0,-3.75,\cdots,3.00\}}$ and the Gaussian kernel hyperparameter $10^{\{-4.0,-3.75,\cdots,3.00\}}$, for Mitigation (i) we set $\delta = 0.1$, and for Mitigation (ii) we tuned $\lambda_1, \lambda_2 \in 10^{\{-4.0,-3.75,\cdots,3.00\}}$ using the 5-fold cross validation [170]. 25% of the data are kept fixed as test set and the experiments are repeated 30 times to ensure statistical significance of the results.

We considered two updating scenarios:

- Data Update: in this scenario we consider the case when we have an initial set of data, we train the model f^*_{old}, then new data become available and

[1] `https://archive.ics.uci.edu/`

[2] `https://www.propublica.org/datastore/`

Name	# samples for f^*_{old}	# samples for f^*_{new}
German	150	750
Arrhythmia	70	340
Compas	231	3240
Adult	489	4884

Table 3.2: Data Update Scenario: size of the training set for f^*_{old} and size of the training set for f^*_{new}.

we retrain the same model f^*_{new} leveraging the additional data. Table 3.2 reports the size of the training set for f^*_{old} and the size of the training set for f^*_{new};

- Data plus Model Update: this scenario is the same of the one of Data Update except for the fact that model used for generating f^*_{new} is not the same as the one used to generate f^*_{old} but it is a more capable one. In our test we used Linear SVM for f^*_{old} and Gaussian-Kernel SVM for f^*_{new}.

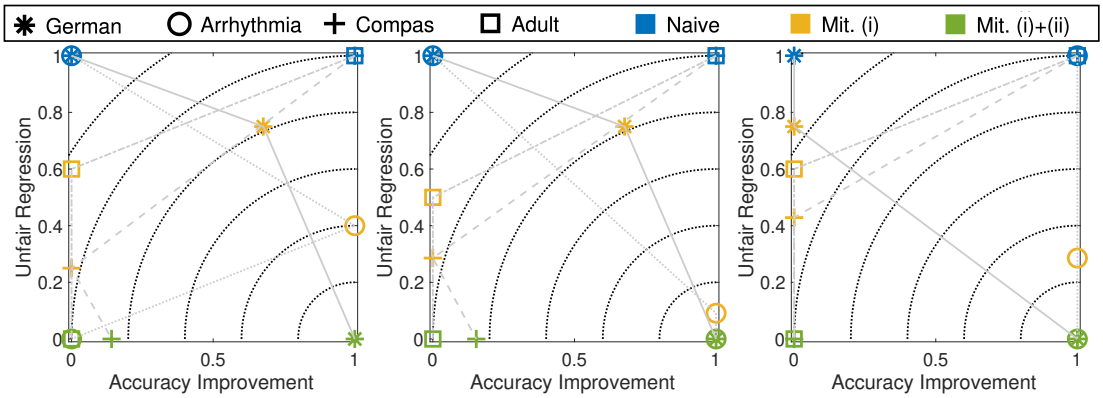
Table 3.3 reports, for the considered scenarios (Data Update and Data plus Model Update), the Accuracy (ACC), the Balanced Accuracy (B-ACC)[1], and the UR (measured both with DP and EO) of f^*_{old}, Naive f^*_{new}, f_{new} with Mitigation (i), and f_{new} with Mitigation (i) plus (ii).

From Table 3.3 it is possible to observe that, as expected from what discussed in Section 3.3, when we update the model without taking into account the UR, the accuracy and the balanced accuracy increase but the UR may be very high. When we introduce our mitigations, generally accuracy and balanced accuracy increase less (but still significantly) with respect to not introducing any mitigation, but UR systematically — and often dramatically — decreases, both on DP and EO. In particular, mitigation (i) plus (ii) usually strongly improves on simple mitigation (i) on the UR phenomenon, without significantly affecting, on average, the accuracy and balanced accuracy. Interestingly, however, sometimes

[1]B-ACC is the maximized metric during the cross validation.

Table 3.3: Accuracy (ACC), Balanced Accuracy (B-ACC), and UR (measured both with DP and EO) of f^*_{old}, Naive f^*_{new}, f_{new} with Mitigation (i), and f_{new} with Mitigation (i) plus (ii) for the considered scenarios (Data Update and Data plus Model Update).

Dataset	f^*_{old}		f_{new} - Naive			f_{new} - Mitigation (i)			f_{new} - Mitigation (i) plus (ii)		
	ACC	B-ACC	ACC	B-ACC	UR	ACC	B-ACC	UR	ACC	B-ACC	UR
Linear SVM - Data Update - $\texttt{UR}^{\text{DP}}_{f^*_{\text{old}}}(f^*_{\text{new}})$											
German	0.69±0.03	0.62±0.04	0.74±0.02	0.63±0.03	0.12±0.04	0.74±0.02	0.65±0.03	0.11±0.04	0.73±0.02	0.66±0.03	0.08±0.03
Arrhythmia	0.67±0.04	0.65±0.04	0.74±0.04	0.74±0.04	0.14±0.04	0.76±0.04	0.75±0.04	0.11±0.03	0.75±0.04	0.74±0.04	0.09±0.03
Compas	0.57±0.03	0.53±0.03	0.74±0.01	0.75±0.01	0.13±0.04	0.60±0.05	0.57±0.05	0.07±0.05	0.64±0.05	0.59±0.05	0.05±0.03
Adult	0.76±0.00	0.51±0.00	0.81±0.01	0.60±0.01	0.006±0.01	0.78±0.01	0.54±0.01	0.004±0.00	0.78±0.01	0.54±0.01	0.001±0.00
Linear SVM - Data Update - $\texttt{UR}^{\text{EO}}_{f^*_{\text{old}}}(f^*_{\text{new}})$											
German	0.69±0.03	0.62±0.04	0.74±0.02	0.63±0.03	0.45±0.15	0.74±0.02	0.64±0.03	0.43±0.15	0.70±0.03	0.68±0.03	0.22±0.15
Arrhythmia	0.67±0.04	0.65±0.04	0.74±0.04	0.74±0.04	0.10±0.04	0.71±0.04	0.71±0.04	0.09±0.03	0.72±0.04	0.73±0.04	0.04±0.03
Compas	0.60±0.03	0.56±0.03	0.75±0.01	0.75±0.01	0.15±0.03	0.67±0.08	0.64±0.09	0.10±0.05	0.67±0.05	0.63±0.05	0.06±0.05
Adult	0.76±0.00	0.51±0.00	0.81±0.01	0.60±0.01	0.006±0.01	0.78±0.01	0.54±0.01	0.005±0.00	0.78±0.01	0.54±0.01	0.002±0.00
Gaussian-Kernel SVM - Data Update - $\texttt{UR}^{\text{DP}}_{f^*_{\text{old}}}(f^*_{\text{new}})$											
German	0.69±0.03	0.62±0.04	0.74±0.05	0.63±0.03	0.12±0.04	0.74±0.05	0.65±0.03	0.11±0.04	0.73±0.02	0.66±0.03	0.08±0.03
Arrhythmia	0.68±0.04	0.65±0.04	0.73±0.06	0.73±0.08	0.33±0.04	0.76±0.04	0.75±0.04	0.13±0.03	0.75±0.04	0.75±0.04	0.11±0.03
Compas	0.60±0.03	0.55±0.03	0.75±0.01	0.76±0.01	0.10±0.03	0.65±0.05	0.60±0.05	0.05±0.05	0.66±0.05	0.62±0.05	0.03±0.03
Adult	0.76±0.00	0.51±0.00	0.81±0.01	0.61±0.01	0.007±0.01	0.78±0.01	0.55±0.01	0.004±0.00	0.78±0.01	0.55±0.01	0.001±0.00
Gaussian-Kernel SVM - Data Update - $\texttt{UR}^{\text{EO}}_{f^*_{\text{old}}}(f^*_{\text{new}})$											
German	0.69±0.03	0.62±0.04	0.74±0.05	0.63±0.03	0.45±0.15	0.74±0.05	0.65±0.03	0.42±0.15	0.73±0.02	0.66±0.03	0.23±0.15
Arrhythmia	0.68±0.04	0.65±0.04	0.73±0.06	0.73±0.08	0.10±0.04	0.76±0.04	0.75±0.04	0.08±0.03	0.72±0.04	0.72±0.04	0.04±0.03
Compas	0.60±0.03	0.55±0.03	0.75±0.01	0.76±0.01	0.13±0.03	0.65±0.05	0.60±0.05	0.06±0.05	0.66±0.05	0.62±0.05	0.05±0.03
Adult	0.76±0.00	0.51±0.00	0.81±0.01	0.61±0.01	0.0003±0.00	0.77±0.01	0.53±0.01	0.0002±0.00	0.78±0.01	0.55±0.01	0.0001±0.00
SVM - Data plus Model Update - $\texttt{UR}^{\text{DP}}_{f^*_{\text{old}}}(f^*_{\text{new}})$											
German	0.69±0.03	0.62±0.04	0.74±0.02	0.63±0.03	0.12±0.04	0.74±0.02	0.63±0.03	0.11±0.04	0.74±0.02	0.66±0.03	0.08±0.04
Arrhythmia	0.67±0.04	0.65±0.04	0.73±0.06	0.73±0.08	0.31±0.10	0.72±0.04	0.73±0.04	0.21±0.10	0.72±0.04	0.73±0.04	0.17±0.20
Compas	0.57±0.03	0.53±0.03	0.75±0.01	0.76±0.01	0.12±0.03	0.65±0.05	0.62±0.04	0.08±0.02	0.66±0.05	0.62±0.03	0.05±0.02
Adult	0.76±0.00	0.51±0.00	0.81±0.01	0.61±0.01	0.007±0.01	0.78±0.01	0.54±0.01	0.005±0.00	0.78±0.01	0.54±0.01	0.002±0.00
SVM - Data plus Model Update - $\texttt{UR}^{\text{EO}}_{f^*_{\text{old}}}(f^*_{\text{new}})$											
German	0.69±0.03	0.62±0.04	0.74±0.02	0.63±0.03	0.53±0.18	0.73±0.02	0.63±0.03	0.51±0.20	0.70±0.02	0.69±0.03	0.24±0.15
Arrhythmia	0.67±0.04	0.65±0.04	0.68±0.06	0.70±0.08	0.02±0.10	0.68±0.04	0.66±0.04	0.01±0.10	0.68±0.04	0.66±0.04	0.001±0.05
Compas	0.57±0.03	0.53±0.03	0.75±0.01	0.76±0.01	0.15±0.03	0.65±0.05	0.62±0.04	0.13±0.02	0.66±0.05	0.62±0.03	0.09±0.01
Adult	0.76±0.00	0.51±0.00	0.81±0.01	0.61±0.01	0.001±0.00	0.77±0.01	0.53±0.01	0.0005±0.00	0.78±0.01	0.54±0.01	0.0002±0.00



(a) Linear SVM - Data Update - $\texttt{UR}^{\text{DP}}_{f^*_{\text{old}}}(f^*_{\text{new}})$.

(b) Gaussian-Kernel SVM- Data Update - $\texttt{UR}^{\text{DP}}_{f^*_{\text{old}}}(f^*_{\text{new}})$.

(c) SVM - Data plus Model Update - $\texttt{UR}^{\text{DP}}_{f^*_{\text{old}}}(f^*_{\text{new}})$.

Figure 3.1: Normalized (dataset-wise) percentage of improvement in balanced accuracy (the larger the better) and UR (the smaller the better) for each dataset and updating strategy (only for the UR on DP because of space constraints) taken from Table 3.3. The gray lines connect the results of the different updating strategies on the same dataset.

the mitigations do not simply improve the UR but also the (balanced) accuracy. To improve the readability of Table 3.3, Figure 3.1 reports, only for the UR on DP because of space constraints, for each dataset and updating strategy, the normalized (dataset-wise) percentage of improvement in balanced accuracy (the larger the better) and UR (the smaller the better). From the graph it appears evident the ability of the proposed mitigation strategy to strongly improve the UR without impacting too much on the accuracy, or actually improving it in some cases.

3.5 Conclusions

In this study, we have explored the interaction between regression in Machine Learning model updates and Algorithmic Fairness. In particular, we introduced the concept of unfair regression, a phenomenon where Machine Learning model,

while improving average performance, may lead to the introduction of pointwise decrease in performance in a disproportionate way among the subgroups defined by a sensitive attributes. This represents a new form of hidden debt in the use of Machine Learning models, increasing both the technical and ethical debt. Our investigation confirmed the existence of unfair regression through empirical analysis on real-world datasets. To address these challenges, we proposed and evaluated two mitigation strategies, specified to the case of Support Vector Machines. The first strategy involves modifying the learning algorithm while the second one focuses on adjusting the performance tuning phase. Our experimental results demonstrated that both approaches are effective in mitigating the effects of unfair regression. Future research will focus on extending the experimental session to a larger variety of models and datasets, improving and specifying the mitigation strategies for different Machine Learning algorithms and supporting the proposal with consistent results on both accuracy and unfair regression metrics.

Acknowledgments

This work was performed in collaboration with Anna Pallarès López, Eduard Martín Jiménez, Davide Anguita, Fabio Roli, and Luca Oneto. Furthermore, it is partially supported by (i) project ELSA – European Lighthouse on Secure and Safe AI funded by the European Union's Horizon Europe under the grant agreement No. 101070617, (ii) EU - NGEU National Sustainable Mobility Center (CN00000023) Italian Ministry of University and Research Decree n. 1033—17/06/2022 (Spoke 10), (iii) project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU, and (iv) project FAIR (PE00000013) under the NRRP MUR program funded by the EU - NGEU.

Chapter 4

Towards Robust Fair Empirical Risk Optimization: a Theoretical Perspective

Summary

In this work, we study the interaction between fairness and robustness, deriving a new optimization framework, Robust Fair Empirical Risk Optimization (RFERM), for building ML models aware of the *robustness bias*. Our contribution is twofold. We (i) investigate the theoretical consistency of the new paradigm introduced by RFERM, and (ii) implement RFERM with three renowned categories of machine learning algorithms, specifically linear, kernelized, and deep models. The work lands the foundations for future empirical experiments validating these results on real-world data.

4.1 Introduction

Among the attributes associated with trustworthiness, fairness and robustness stand out as two of the most critical for building responsible and effective ML systems. Fairness ensures that algorithmic decisions do not systematically disadvantage particular individuals or demographic groups, thereby promoting equity. Robustness, in turn, guarantees that models perform reliably under a range of conditions, including adversarial manipulation, noisy inputs, or shifts in data distribution.

A robust model delivers consistent results, strengthening both its practical utility and the trust users place in its functionality. However, the challenges become more complex when fairness and robustness are considered together. It is well established that ML models can be deceived by adversaries who introduce subtle perturbations to the input data, often causing misclassifications [16]. Recent studies have shown that these misclassifications can disproportionately affect certain historically discriminated groups, giving rise to what has been termed robustness bias [158].

Robustness bias can stem from several factors. Firstly, the data used to train models can be biased towards specific demographic groups. If certain groups are underrepresented in the training data, the models may lack confidence when predicting for these groups [158]. This issue, often referred to as sample bias, is a specific example of the broader category of dataset biases [20]. Other types of dataset biases include prejudice bias, where the data collection or analysis is influenced by discriminatory views, and group attribution bias, where inaccurate conclusions are drawn about a demographic group based on the data of a few individuals. All these biases can contribute to increasing the model's robustness bias [178]. Another factor is the complexity of ML models, such as Deep Neural Networks, which often have intricate decision boundaries that can facilitate

the targeting of specific subgroups by an attacker [158]. Ultimately, an adversary could simply design their attack to be particularly effective against targeted discriminated subgroups.

In this study, we address the challenge of mitigating the robustness bias of ML models — i.e., reducing the effectiveness of adversarial attacks targeting disadvantaged subgroups within the population. Our approach involves establishing a specific threat model, which is the case of white-box attacks under an L_2 ball constraint. In this scenario, the attacker has full knowledge of and access to the model, and their attacks are designed to maximize model loss within a sphere centred on the original samples.

We employ a recently introduced fairness notion, $\epsilon-$Fairness Revised [166], which can extend the traditional fairness definitions found in the literature. In this work, this notion is modeled to address robustness bias under the specified threat model — in Section 4.3, we also examine its theoretical consistency.

Moreover, we revisit the traditional optimization framework presented by Empirical Risk Minimization (ERM) [170; 179], Robust Empirical Risk Minimization (RERM) [121], and Fair Empirical Risk Minimization (FERM) [26] with the ultimate goal of merging these frameworks. To accomplish this, we introduce a new optimization paradigm, Robust Fair Empirical Risk Minimization (RFERM), which is well-suited for addressing the issue of robustness bias. Finally, we formulate three concrete problems that leverage linear, kernelized, and deep models.

The rest of the paper is structured as follows. Section 4.2 introduces the notation used in this work while presenting the context and the different optimization paradigms as well. Section 4.3 analyzes the theoretical consistency of our newly introduced notion for tackling robustness bias. Section 4.4 implements the theoretical notion by deriving three concrete optimization frameworks ex-

ploiting linear, kernelized, and deep models. Finally, Section 4.5 concludes the work.

4.2 Preliminaries

Consider the dataset $\mathcal{D} = \{(x_1, s_1, y_1), \ldots, (x_n, s_n, y_n)\}$ to be a sequence of n samples identically and independently drawn from a joint probability distribution μ over $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space[1], $\mathcal{S} = \{a, b\}$ is the space of the sensitive attributes[2] and $\mathcal{Y} = \{-1, +1\}$ is the output space[3]. Note that the input space $\mathcal{X}$ may also directly include $\mathcal{S}$, depending on whether the legal requirements allow the sensitive attributes to be used as predictors [65]. Let us consider the space of hypotheses $\mathcal{H}$ from which we want to select a specific hypothesis $h : \mathcal{X} \mapsto \widehat{\mathcal{Y}}$ which approximates $\mathbb{P}\{y|x\}$. Ideally, in classical ML, we want to find the best approximating model $h \in \mathcal{H}$ that minimises the true risk $\mathtt{L}_\ell(h) = \mathbb{E}_{x,y}\{\ell(h(x), y)\}$, where $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \mapsto \mathbb{R}^+$ is the loss function. Since μ is unknown, $\mathtt{L}_\ell(h)$ cannot be computed and we have to resort to one of its empirical estimators. In our case, we consider the empirical risk $\widehat{\mathtt{L}}_\ell(h) = \widehat{\mathbb{E}}_{x,y}\{\ell(h(x), y)\}$, where $\widehat{\mathbb{E}}$ denotes the empirical expectation computed over $\mathcal{D}$. Thus, our oracle for finding the best hypothesis h^* is

$$h^* = \arg\min_{h \in \mathcal{H}} \ \mathtt{L}_\ell(h), \tag{4.1}$$

while its empirical estimator $\widehat{h}$ is

$$\widehat{h} = \arg\min_{h \in \mathcal{H}} \ \widehat{\mathtt{L}}_\ell(h). \tag{4.2}$$

[1]Which can be unstructured (i.e., tabular data) or structured (e.g., images or graphs).

[2]For the sake of notation simplicity we restrict to a binary-valued sensitive attribute (e.g., male/female or dark/light skinned color) but the approach can be easily generalized to any space.

[3]For the sake of notation simplicity we restrict to the binary classification but the results can be easily generalized to the whole supervised learning setting.

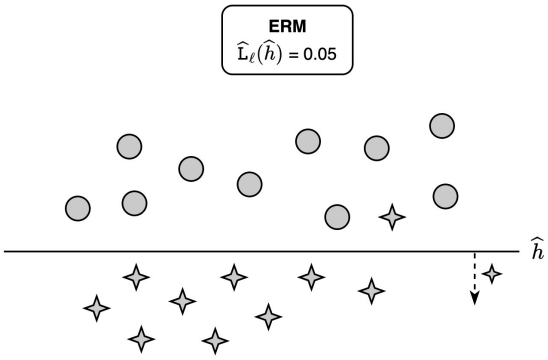


Figure 4.1: ERM: the positively labelled samples are represented with stars, the negatively labelled ones with circles, $\widehat{h}$ is the black line, and the dashed arrow indicates the semi-space labelled positively. In this example, ℓ is the misclassification loss.

The latter approach is known as the Empirical Risk Minimization (ERM) [170; 179] framework which often results in overfitting the training data. To mitigate this issue one needs to tune $\mathcal{H}$ in the so-called model selection phase [175]. Figure 4.1 depicts a simple case of ERM with linear models.

Let us now consider the presence of an adversary under the following threat model. The adversary can perturb the input x according to a possible set of manipulation $\mathcal{B} : \mathcal{X} \mapsto \mathcal{X}$ such that $\mathcal{B}(x) \subseteq \mathcal{X}$ and $x \in \mathcal{B}(x)$. The purpose of the adversary is to induce a hypothesis $h \in \mathcal{H}$ into mistakes, namely to increase its (empirical and/or true) error [16]. Formally, the adversary performs the following attack

$$\widetilde{x}_\ell = \arg \max_{x \in \mathcal{B}(x)} \ \ell(h(x), y), \tag{4.3}$$

namely the adversary searches for the modification of the input that maximizes the loss of the hypothesis h. To mitigate the effect of the adversary, we need to modify Problems (4.1) and (4.2). For this purpose let us define the following loss

function

$$\tilde{\ell}(h(x), y) = \max_{x \in \mathcal{B}(x)} \ell(h(x), y). \tag{4.4}$$

This loss measures the worst error committed by the hypothesis on one sample under the adversarial attacks defined in Eq. (4.3). Then, minimizing this loss actually maximizes the robustness of the model to the adversary. As a result, the counterparts of Problems (4.1) and (4.2) able to mitigate the adversary effect are the following ones

$$h^*_{\mathrm{R}} = \arg\min_{h \in \mathcal{H}} \ \mathtt{L}_{\tilde{\ell}}(h), \tag{4.5}$$

$$\widehat{h}_{\mathrm{R}} = \arg\min_{h \in \mathcal{H}} \ \widehat{\mathtt{L}}_{\tilde{\ell}}(h). \tag{4.6}$$

The latter problem, namely the empirical one, is called Adversarial Training [121] or, in optimization terms, Robust Empirical Risk Minimization (RERM). Figure 4.2 depicts a comparison between ERM and RERM with linear models.

Let us now consider the case where our learned model is audited on its fairness, testing whether it behaves differently when applied to elements in the dataset having $s = a$ or $s = b$. Generally, fairness can be measured with both individual and group measures [180]. Individual fairness checks whether the pointwise decision may change by changing the group membership, while group fairness checks whether statistical performances on the population cohorts are significantly different. In this work, we will deal with this second category and the most common group fairness metrics can be generalized as follows

$$\mathtt{F}^{\star}_{\ell,\diamond}(h) = \mathbb{E}_{\circ \in \star} \left\{ \left| \mathtt{L}^{a,\circ}_{\ell,\diamond}(h) - \mathtt{L}^{b,\circ}_{\ell,\diamond}(h) \right| \right\}, \tag{4.7}$$

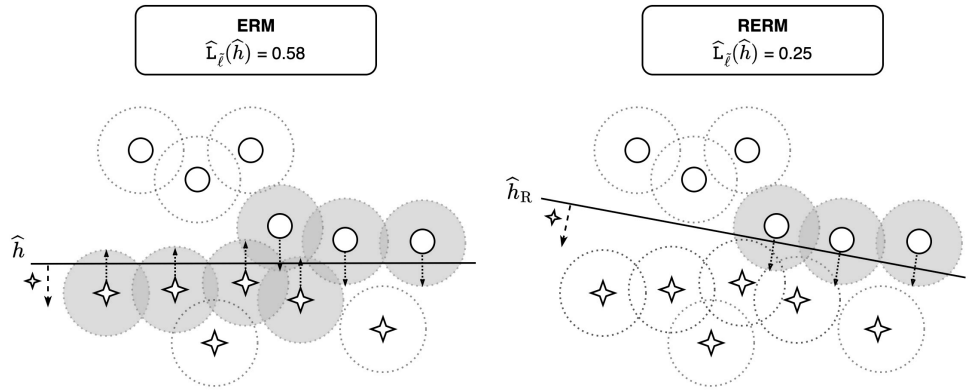


Figure 4.2: RERM: the positively labelled samples are represented with stars, the negatively labelled ones with circles, $\widehat{h}$ is the black line in the left figure, $\widehat{h}_{\mathrm{R}}$ is the black line in the right figure, dashed arrows indicate the semi-space labelled positively, and grey areas indicate the set of manipulations for each sample under a bounded L_2 constraint. In this example, ℓ is the misclassification loss.

where

$$\mathtt{L}^{\cdot,\circ}_{\ell,\diamond}(h) = \mathbb{E}_{x,s,y}\left\{\ell(h(x),\diamond) \mid s = \cdot, y \in \circ\right\}. \tag{4.8}$$

This definition has been introduced in the work by Franco et al. [166] from which this contribution stems. If we set $\ell(h(x), y) = [yh(x) \leq 0]$[1] (Hard loss) we get Demographic Parity (DP) [181] when $\diamond = +1$ and $\star = \{\pm 1\}$, the Equal Opportunity (EOp) [35] when $\diamond = +1$ and $\star = \{+1\}$, and the Equalized Odds (EOd) [35] when $\diamond = +1$ and $\star = \{\{+1\}, \{-1\}\}$. Let us also define

$$n^{\cdot,\circ} = \left|\{(x, s, y) \mid (x, s, y) \in \mathcal{D}, s = \cdot, y \in \circ\}\right|. \tag{4.9}$$

As before, $\mathtt{F}^{\star}_{\ell,\diamond}(h)$ cannot be computed since it requires the knowledge of the underlying generating distribution μ, thus the only quantity that we can compute

[1]Note that we opt for the indicator function with the Iverson bracket notation.

is its empirical estimator

$$\widehat{\mathtt{F}}^{\star}_{\ell,\diamond}(h) = \mathbb{E}_{\circ\in\star}\left\{\left|\widehat{\mathtt{L}}^{a,\circ}_{\ell,\diamond}(h) - \widehat{\mathtt{L}}^{b,\circ}_{\ell,\diamond}(h)\right|\right\}, \tag{4.10}$$

$$\widehat{\mathtt{L}}^{\cdot,\circ}_{\ell,\diamond}(h) = \widehat{\mathbb{E}}_{x,s,y}\left\{\ell(h(x),\diamond) \mid s = \cdot, y \in \circ\right\}. \tag{4.11}$$

To mitigate the unfairness, we need to modify Problems (4.1) and (4.2) as follows

$$h^*_{\mathrm{F}} = \arg\min_{h\in\mathcal{H}} \ \mathtt{L}_\ell(h), \quad \text{s.t. } \mathtt{F}^{\star}_{\ell,\diamond}(h) \leq \epsilon, \tag{4.12}$$

$$\widehat{h}_{\mathrm{F}} = \arg\min_{h\in\mathcal{H}} \ \widehat{\mathtt{L}}_\ell(h), \quad \text{s.t. } \widehat{\mathtt{F}}^{\star}_{\ell,\diamond}(h) \leq \widehat{\epsilon}, \tag{4.13}$$

where Problems (4.12) and (4.13) are the counterparts of Problems (4.1) and (4.2) under fairness constraints and ϵ and $\widehat{\epsilon}$ are the true and empirical levels of unfairness that we are willing to bear[1]. Moreover, $\epsilon = 0$ may not be the best thing to do in practical situations because of overfitting [26]. Problem (4.13) is usually referred to as Fair Empirical Risk Minimization (FERM) [26]. Figure 4.3 depicts a comparison between ERM and FERM with linear models. Recent works have shown that a possible adversary, the noise in the data, and/or the intrinsic error of the model may lead to a robustness bias in particular applications [158; 159]. This new measure of unfairness, based on the previous notions, can be defined as follows

$$\mathtt{F}^{\star}_{\tilde{\ell},\diamond}(h) = \mathbb{E}_{\circ\in\star}\left\{\left|\mathtt{L}^{a,\circ}_{\tilde{\ell},\diamond}(h) - \mathtt{L}^{b,\circ}_{\tilde{\ell},\diamond}(h)\right|\right\}, \tag{4.14}$$

$$\mathtt{L}^{\cdot,\circ}_{\tilde{\ell},\diamond}(h) = \mathbb{E}_{x,s,y}\left\{\tilde{\ell}(h(x),\diamond) \mid s = \cdot, y \in \circ\right\}, \tag{4.15}$$

[1]Note that there is a natural trade-off between fairness and accuracy that, depending on the application, may lead to different ϵ and $\widehat{\epsilon}$ [182].

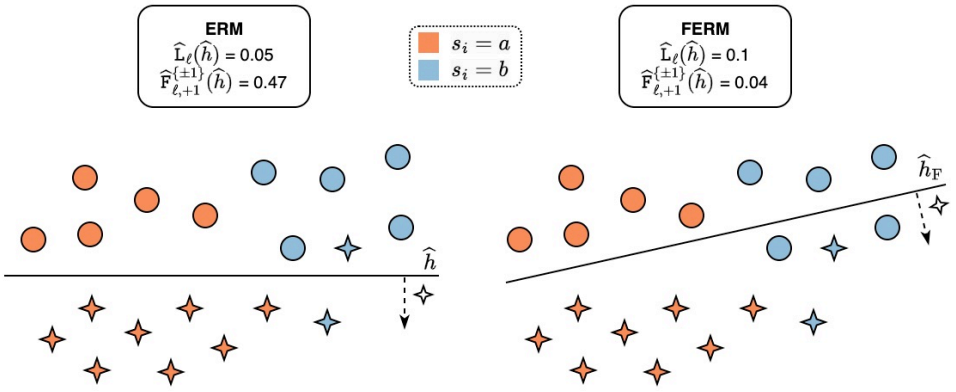


Figure 4.3: FERM: the positively labelled samples are represented with stars, the negatively labelled ones with circles, the samples belonging to the first subgroups are coloured in orange, while the samples belonging to the second subgroup are coloured in light blue, $\widehat{h}$ is the black line in the left figure, $\widehat{h}_{\mathrm{F}}$ is the black line in the right figure, and dashed arrows indicate the semi-space positively labelled. In this example, ℓ is the misclassification loss and $\widehat{\mathrm{F}}$ is a measure of DP.

together with its empirical counterpart

$$\widehat{\mathrm{F}}^{\star}_{\tilde{\ell},\diamond}(h) = \mathbb{E}_{\circ\in\star}\left\{\left|\widehat{\mathrm{L}}^{a,\circ}_{\tilde{\ell},\diamond}(h) - \widehat{\mathrm{L}}^{b,\circ}_{\tilde{\ell},\diamond}(h)\right|\right\}, \tag{4.16}$$

$$\widehat{\mathrm{L}}^{\cdot,\circ}_{\tilde{\ell},\diamond}(h) = \widehat{\mathbb{E}}_{x,s,y}\left\{\tilde{\ell}(h(x),\diamond) \mid s = \cdot, y \in \circ\right\}. \tag{4.17}$$

As before, note that by setting $\diamond = 1$ and $\ell(h(x), y) = [yh(x) \leq 0]$ we can reduce to the robust extension of classical fairness notions. If we set $\ell(h(x), y) = [yh(x) \leq 0]$ (Hard loss) we get Robust Demographic Parity (RDP) [181] when $\diamond = +1$ and $\star = \{\pm 1\}$, the Robust Equal Opportunity (REOp) [35] when $\diamond = +1$ and $\star = \{+1\}$, and the Robust Equalized Odds (REOd) [35] when $\diamond = +1$ and $\star = \{\{+1\}, \{-1\}\}$.

To mitigate the adversary and unfairness simultaneously, we need to modify

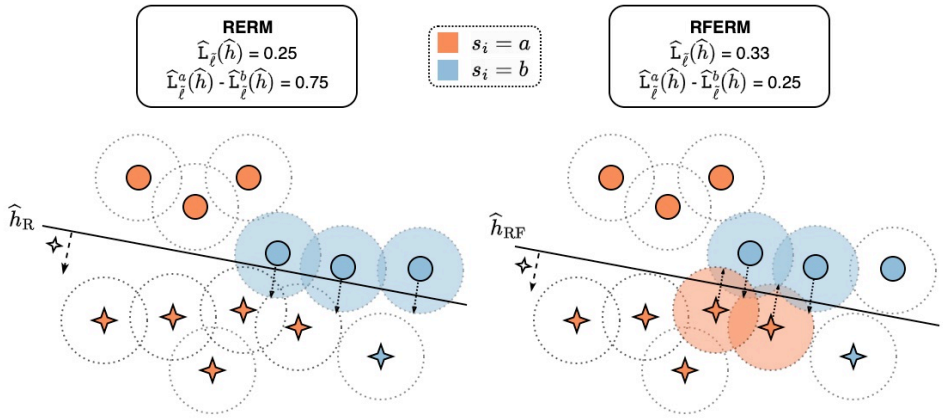


Figure 4.4: RFERM: the positively labelled samples are represented with stars, the negatively labelled ones with circles, the samples belonging to the first subgroups are coloured in orange, while the samples belonging to the second subgroups are coloured in orange light blue, $\widehat{h}_{\mathrm{R}}$ is the black line in the left figure, $\widehat{h}_{\mathrm{RF}}$ is the black line in the right figure, dashed arrow indicates the semi-space positively labelled, and transparent orange and light blue indicate the set of manipulations for each sample of each subgroup under a bounded L_2 constraint.

Problems (4.1) and (4.2) as follows

$$h^*_{\mathrm{RF}} = \arg\min_{h\in\mathcal{H}} \ \mathtt{L}_{\tilde{\ell}}(h), \quad \text{s.t. } \mathtt{F}^{\star}_{\tilde{\ell},\diamond}(h) \leq \epsilon, \tag{4.18}$$

$$\widehat{h}_{\mathrm{RF}} = \arg\min_{h\in\mathcal{H}} \ \widehat{\mathtt{L}}_{\tilde{\ell}}(h), \quad \text{s.t. } \widehat{\mathtt{F}}^{\star}_{\tilde{\ell},\diamond}(h) \leq \widehat{\epsilon}, \tag{4.19}$$

where Problems (4.18) and (4.19) are the counterparts of Problems (4.1) and (4.2) under robustness bias constraints. Problem (4.19) is known as Fair Adversarial Training [183] or, alternatively for an optimization setting, Robust Fair Empirical Risk Minimization (RFERM). Figure 4.4 depicts a comparison between RERM and RFERM with linear models.

Section 4.3 will deal with the problem of proving for the first time that, in some sense that will be properly defined later, $\widehat{h}_{\mathrm{RF}}$ is a consistent estimator of h^*_{RF}.

Section 4.4 will show that RFERM may result in a non-convex, non-differentiable,

constrained optimization problem. For this reason, we will consider two scenarios. In the first one, shallow models (Kernel methods [169]) will be considered and we will derive a convex relaxation with theoretical guaranties that is able to approximate RFERM well. In the second one, we will consider deep models [171] and we will derive a differentiable relaxation of RFERM.

4.3 Robustness Bias: Theory

In this section, similarly to what has been done for ERM [170; 179], RERM [184; 185], and FERM [26], we want to prove the consistency of Problem (4.19) (RFERM) for a particular value of $\widehat{\epsilon}$. In our context, consistency means that $\widehat{h}_{\text{RF}}$ converges, as n increases, to h^*_{RF} both in terms of $\mathtt{L}_\ell(\cdot)$ and $\mathtt{F}^\star_{\ell,\diamond}(\cdot)$. More formally we want to bind the following quantities

$$\mathtt{L}_{\tilde{\ell}}(\widehat{h}_{\text{RF}}) - \mathtt{L}_{\tilde{\ell}}(h^*_{\text{RF}}), \tag{4.20}$$

$$\mathtt{F}^\star_{\tilde{\ell},\diamond}(\widehat{h}_{\text{RF}}) - \mathtt{F}^\star_{\tilde{\ell},\diamond}(h^*_{\text{RF}}) \leq \mathtt{F}^\star_{\tilde{\ell},\diamond}(\widehat{h}_{\text{RF}}) - \epsilon, \tag{4.21}$$

with and high probability bound that decreases with n at least as $1/\sqrt{n}$.

In order to be able to prove the consistency of Problem (4.19) we need to make some (minor) assumptions. The first one is that the

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}} \left|\mathtt{L}_\ell(h) - \widehat{\mathtt{L}}_\ell(h)\right| \leq \mathtt{U}_\ell(\delta, n, \mathcal{H})\right\} \geq 1-\delta, \tag{4.22}$$

where $\mathtt{U}_\ell(\delta, n, \mathcal{H})$ goes to zero as n grows to infinity if the class $\mathcal{H}$ is learnable with respect to the loss [170]. For example, if $\mathcal{H}$ is a compact subset of linear separators in a Reproducing Kernel Hilbert Space (RKHS), and the loss is Lipschitz in its first argument, then $\mathtt{U}_\ell(\delta, n, \mathcal{H})$ can be obtained via Rademacher bounds [186]. In this case $\mathtt{U}_\ell(\delta, n, \mathcal{F})$ goes to zero at least as $1/\sqrt{n}$ as n grows to infinity. Note that

this assumption is the same behind the prove of the consistency of RERM [184; 185] and FERM [26]. For example, for RERM [184; 185], it has been proved that if $\mathcal{H}$ is learnable with respect to the loss ℓ it exists $\mathtt{U}_{\tilde{\ell}}(\delta, n, \mathcal{H})$ such that [184; 185]

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}}\left|\mathtt{L}_{\tilde{\ell}}(h) - \widehat{\mathtt{L}}_{\tilde{\ell}}(h)\right| \leq \mathtt{U}_{\tilde{\ell}}(\delta, n, \mathcal{H})\right\} \geq 1-\delta. \tag{4.23}$$

At this point, we can present the main result of this section, namely, we can prove the following theorem which proves the consistency of $\widehat{h}_{\mathrm{RF}}$ with respect to h^*_{RF}, namely $\widehat{h}_{\mathrm{RF}}$ tends to h^*_{RF} as n grows both in terms of robustness and robustness bias.

Theorem 1. *Let $\mathcal{H}$ be a learnable set of functions with respect to the loss function ℓ and let h^*_{RF} and $\widehat{h}_{RF}$ be the models defined in Problems* (4.18) *and* (4.19) *with*

$$\hat{\epsilon} = \epsilon + \sum_{\circ\in\star}\sum_{s\in\mathcal{S}} \mathtt{U}_{\tilde{\ell}}\left(\delta, n^{s,\circ}, \mathcal{H}\right). \tag{4.24}$$

Then, with probability at least $1-(2+4|\star|)\delta$, it holds simultaneously that

$$\mathtt{L}_{\tilde{\ell}}(\widehat{h}_{RF}) - \mathtt{L}_{\tilde{\ell}}(h^*_{RF}) \leq 2\mathtt{U}_{\tilde{\ell}}\left(\delta, n, \mathcal{H}\right), \tag{4.25}$$

$$\mathtt{F}^{\star}_{\tilde{\ell},\diamond}(\widehat{h}_{RF}) \leq \epsilon + 2\sum_{\circ\in\star}\sum_{s\in\mathcal{S}} \mathtt{U}_{\tilde{\ell}}\left(\delta, n^{s,\circ}, \mathcal{H}\right). \tag{4.26}$$

Proof. We first leverage the learnability of $\mathcal{H}$, and then Eqns. (4.22) and (4.23), state, with probability at least $1-2|\star|\delta$,

$$\begin{aligned}
&\sup_{h\in\mathcal{H}}\left|\mathtt{F}^{\star}_{\tilde{\ell},\diamond}(h) - \widehat{\mathtt{F}}^{\star}_{\tilde{\ell},\diamond}(h)\right| \\
&= \sup_{h\in\mathcal{H}}\left|\mathbb{E}_{\circ\in\star}\left\{\left|\mathtt{L}^{a,\circ}_{\tilde{\ell},\diamond}(h) - \mathtt{L}^{b,\circ}_{\tilde{\ell},\diamond}(h)\right|\right\} - \mathbb{E}_{\circ\in\star}\left\{\left|\widehat{\mathtt{L}}^{a,\circ}_{\tilde{\ell},\diamond}(h) - \widehat{\mathtt{L}}^{b,\circ}_{\tilde{\ell},\diamond}(h)\right|\right\}\right| \\
&\leq \sum_{\circ\in\star}\sum_{s\in\mathcal{S}} \mathtt{U}_{\tilde{\ell}}\left(\delta, n^{s,\circ}, \mathcal{H}\right).
\end{aligned} \tag{4.27}$$

This inequality allows also us to state that, with probability at least $1 - 2|\star|\delta$

$$\left\{h \mid h \in \mathcal{H}, \mathtt{F}^{\star}_{\tilde{\ell},\diamond}(h) \leq \epsilon\right\} \subseteq \left\{h \mid h \in \mathcal{H}, \widehat{\mathtt{F}}^{\star}_{\tilde{\ell},\diamond}(h) \leq \epsilon + \sum_{\circ\in\star}\sum_{s\in\mathcal{S}} \mathtt{U}_{\tilde{\ell}}(\delta, n^{s,\circ}, \mathcal{H})\right\}. \tag{4.28}$$

In order to prove the first statement of the theorem, i.e., Eq. (4.25), let us note that

$$\begin{aligned}\mathtt{L}_{\tilde{\ell}}(\widehat{h}_{\mathrm{RF}}) - \mathtt{L}_{\tilde{\ell}}(h^*_{\mathrm{RF}}) =& \mathtt{L}_{\tilde{\ell}}(\widehat{h}_{\mathrm{RF}}) - \widehat{\mathtt{L}}_{\tilde{\ell}}(\widehat{h}_{\mathrm{RF}}) \\ &+ \widehat{\mathtt{L}}_{\tilde{\ell}}(\widehat{h}_{\mathrm{RF}}) - \widehat{\mathtt{L}}_{\tilde{\ell}}(h^*_{\mathrm{RF}}) \\ &+ \widehat{\mathtt{L}}_{\tilde{\ell}}(h^*_{\mathrm{RF}}) - \mathtt{L}_{\tilde{\ell}}(h^*_{\mathrm{RF}}).\end{aligned} \tag{4.29}$$

Eq. (4.28) implies that $\widehat{\mathtt{L}}_{\tilde{\ell}}(\widehat{h}_{\mathrm{RF}}) - \widehat{\mathtt{L}}_{\tilde{\ell}}(h^*_{\mathrm{RF}}) \leq 0$ with probability at least $1 - 2|\star|\delta$ and consequently with probability at least $1 - 2|\star|\delta$ it holds that

$$\mathtt{L}_{\tilde{\ell}}(\widehat{h}_{\mathrm{RF}}) - \mathtt{L}_{\tilde{\ell}}(h^*_{\mathrm{RF}}) = \mathtt{L}_{\tilde{\ell}}(\widehat{h}_{\mathrm{RF}}) - \widehat{\mathtt{L}}_{\tilde{\ell}}(\widehat{h}_{\mathrm{RF}}) + \widehat{\mathtt{L}}_{\tilde{\ell}}(h^*_{\mathrm{RF}}) - \mathtt{L}_{\tilde{\ell}}(h^*_{\mathrm{RF}}). \tag{4.30}$$

The statement of Eq. (4.25) now comes from the Eq. (4.23). For the statement of Eq. (4.26), its proof consists in exploiting the results of Eqns. (4.27) and (4.28) together with a union bound. □

A consequence of the first statement of Theorem 1 is that as n tends to infinity $\mathtt{L}_{\tilde{\ell}}(\widehat{h}_{\mathrm{RF}})$ tends to a value which is not larger than $\mathtt{L}_{\tilde{\ell}}(h^*_{\mathrm{RF}})$, that is, RFERM is consistent with respect to the robustness of the selected model. The second statement of Theorem 1, instead, implies that as n tends to infinity we have that $\widehat{h}_{\mathrm{RF}}$ tends to have ϵ robustness bias. In other words, RFERM is consistent with respect to the fairness of the selected model.

Thanks to Theorem 1 we can state that $\widehat{h}_{\mathrm{RF}}$ is close to h^*_{RF} both in terms of

its robustness and its robustness bias.

4.4 Robustness Bias: Practice

The learning frameworks described in Section 4.2 (i.e., ERM, RERM, FERM, and RFERM) result in optimization problems (i.e., Problems (4.1), (4.5), (4.12), and (4.18) respectively) that may be non-convex, or, even worse, non-differentiable, depending on the choices of the hypothesis space $\mathcal{H}$, the loss function ℓ, and the manipulations space $\mathcal{B}(x)$. For example, exploiting the Hard loss function in the fairness definition of Eq. (4.7) allows the reduction to well-known fairness notions that are non-convex and non-differentiable making the FERM optimization problem np-hard [166]. Another example is RERM which is inherently defined as a min-max optimization problem which is, in general, non-convex even if we use convex loss functions [185].

In this section, we will discuss how to properly define a fully convex, or partially convex, or, at least, differentiable relaxation of the ERM, RERM, FERM, and RFERM without distorting their original nature while maintaining some theoretical guarantees. In particular, we will follow a three-step approach.

In the first step, we will start with the case when $\mathcal{H}$ is a class of linear functions, then we continue relaxing the loss functions ℓ to get convex functionals, and finally, we will consider a manipulations space $\mathcal{B}(x)$ that allows to solve the inner max of RERM and RFERM problem in close form reducing ERM, RERM, FERM, and RFERM to a single convex optimization problem or a to series of convex optimization problems. During this process we will show how classical convex relaxation of the loss functions may work for some problems like ERM and FERM but not for RERM and RFERM. This step is fundamental and instrumental for the second and third ones shading some light on the complexity

of our work and the solution to make it simple or at least tractable.

In the second step, we will consider $\mathcal{H}$ as a compact subset of linear separators in a Reproducing Kernel Hilbert Space, we will leverage the convex relaxation of the loss ℓ discussed in the first step deriving, in a straightforward manner, a non-linear version of the ERM and FERM that are still single convex optimization problems. For RERM and RFERM, we will leverage the same manipulations space $\mathcal{B}(x)$ noting that the non-linearity makes deriving a single convex formulation impossible. Nevertheless, the min-max will result in a bi-level optimization problem where one problem, the minimization, will be still convex, while the other one, the maximization, will be solved through gradient descent resulting in the so-called adversarial training.

Finally, we will consider the deep-learning scenario where $\mathcal{H}$ is induced by a particular choice of the functional form of the network, the optimizer, the regularizers, etc., and leveraging the results obtained in first two steps, we will show that ERM, RERM, FERM, and RFERM can be formulated as an optimization problem with a single differentiable objective (min-max in case of RERM and RFERM) by introducing specific regularizers, in the cost to be optimized during the training phase of the network, based on the problem under exam.

4.4.1 First Step: Linear Models

Let us consider the case where $\mathcal{X} \subseteq \mathbb{R}^d$ and let $h(x) = w^\top x + b$ with $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ be the class of linear models.

In this setting, Problem (4.2), namely ERM, searches for the model in this class which minimizes the classification error, meaning that we rely on the Hard loss function $\ell_{\text{Hard}}(h(x), y) = [yh(x) \leq 0]$ (see Figure 4.5). This produces a non-convex and non-differentiable optimization problem. For this reason, state of the art approach is to relax ℓ_{Hard} with the Hinge loss function $\ell_{\text{Hinge}}(h(x), y) =$

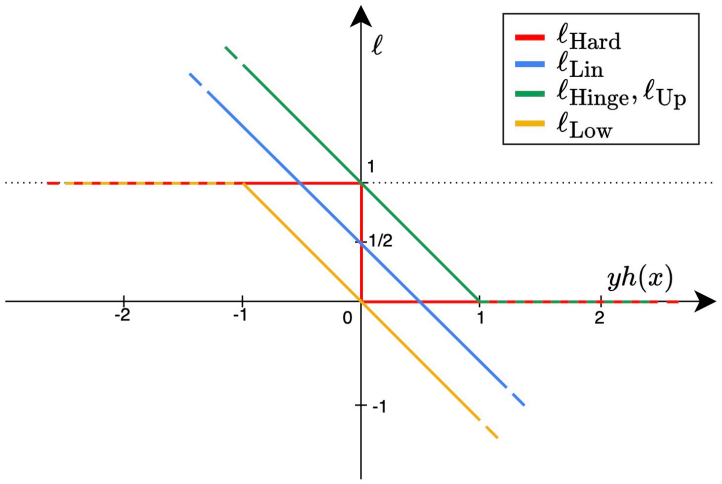


Figure 4.5: Graphical representation of the losses exploited in this work.

$\max[0, 1 - yh(x)]$ (see Figure 4.5), namely its best convex upper bound and approximation [173]. In this setting, let $\mathcal{H}$ be the the class of linear model with $\|w\|_2 \leq W$ and $W \in [0, \infty)$. Then, the convex relaxation of Problem (4.2) can be obtained as

$$\begin{aligned} \min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad & \sum_{i=1}^{n} \max[0, 1 - y_i(w^\top x_i + b))] \\ \text{s.t.} \quad & \|w\|_2^2 \leq W^2, \end{aligned} \tag{4.31}$$

where W is a hyperparameter to be tuned. Problem (4.31) is hard to solve but fully equivalent, for a specific value of $\lambda \in [0, \infty)$ to

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \sum_{i=1}^{n} \max[0, 1 - y_i(w^\top x_i + b))] + \lambda \|w\|_2^2, \tag{4.32}$$

which is the Support Vector Machine [179] algorithm.

For what concerns Problem (4.6), namely RERM, we need to decide what

manipulation to use on top of ERM. Let us consider the case of $\mathcal{B}(x) = \{\widetilde{x} \in \mathbb{R}^d \mid \|\widetilde{x} - x\|_2 \leq \rho\}$ with $\rho \in [0, \infty)$. In this setting the counterpart of Problem (4.32) for RERM is

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \sum_{i=1}^{n} \max_{\widetilde{x} \in \mathbb{R}^d : \|\widetilde{x} - x_i\|_2 \leq \rho} \max[0, 1 - y_i(w^\top \widetilde{x} + b))] + \lambda \|w\|_2^2, \tag{4.33}$$

where λ needs to be tuned while ρ depends on the application (threat model). Note that the model robustness should be evaluated for ranging values of ρ. Having formulated $\mathcal{H}$ as the space of linear functions, the solution to the inner max problem can be found in closed form as follows

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \sum_{i=1}^{n} \max[0, 1 - y_i(w^\top x_i + b) + \|w\|\rho)] + \lambda \|w\|_2^2. \tag{4.34}$$

This problem can be easily solved with a series of convex optimization problems like the following one

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \sum_{i=1}^{n} \max[0, 1 - y_i(w^\top x_i + b) + W\rho)] + \lambda \|w\|_2^2, \tag{4.35}$$

line searching $W \in [0, \infty)$ so to find the solution w^* at minimal but with $\|w^*\| = W$.

For what concerns Problem (4.13), namely FERM, we can apply the same approach proposed with ERM but we have to deal with the constraint $\widehat{\mathtt{F}}^{\star}_{\ell,\diamond}(h) \leq \widehat{\epsilon}$, which might be difficult to make convex. In the literature [26], the first proposal was to relax it by using a linear loss defined as $\ell_{\text{Lin}}(h(x), y) = \frac{1 - yh(x)}{2}$ (see

Figure 4.5) obtaining the following form

$$\widehat{\mathtt{F}}^{\star}_{\ell_{\mathrm{Lin}},\diamond}(h) = \frac{1}{2}\mathbb{E}_{\circ\in\star}\left\{\left|\widehat{\mathbb{E}}_{x,s,y}\left\{\diamond w^\top x \mid s=a, y\in\circ\right\} - \widehat{\mathbb{E}}_{x,s,y}\left\{\diamond w^\top x \mid s=b, y\in\circ\right\}\right|\right\} \leq \widehat{\epsilon}, \tag{4.36}$$

which is a convex constraint. Note that, if we relax the DP (that is, $\diamond = +1$, $\star = \{\pm 1\}$) we get

$$\begin{aligned}
\widehat{\mathtt{F}}^{\{\pm1\}}_{\ell_{\mathrm{Lin}},+1}(h) &= \frac{1}{2}\left|\widehat{\mathbb{E}}_{x,s,y}\left\{w^\top x \mid s=a\right\} - \widehat{\mathbb{E}}_{x,s,y}\left\{w^\top x \mid s=b\right\}\right| \\
&= \frac{1}{2}\left|w^\top\left(\widehat{\mathbb{E}}_{x,s,y}\left\{x \mid s=a\right\} - \widehat{\mathbb{E}}_{x,s,y}\left\{x \mid s=b\right\}\right)\right| \\
&= \frac{1}{2}\left|w^\top\left(\frac{1}{|n^{a,\{\pm1\}}|}\sum_{i=1:s_i=a}^{n} x_i - \frac{1}{|n^{b,\{\pm1\}}|}\sum_{i=1:s_i=b}^{n} x_i\right)\right| \\
&= \frac{1}{2}\left|w^\top c\right| \leq \widehat{\epsilon},
\end{aligned} \tag{4.37}$$

which are two simple linear constraints. If $\widehat{\epsilon} = 0$ we have a single linear constraint that can be mapped in a simple data pre-processing method. With this approach, the counterpart of Problem (4.32) for FERM is

$$\begin{aligned}
\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \quad & \sum_{i=1}^{n} \max[0, 1 - y_i(w^\top x_i + b))] + \lambda\|w\|_2^2, \\
\text{s.t.} \quad & \frac{1}{2}\frac{1}{|\star|}\sum_{\circ\in\star}\left|w^\top\left(\frac{1}{|n^{a,\circ}|}\sum_{i=1:s_i=a,y_i\in\circ}^{n} \diamond x_i - \frac{1}{|n^{b,\circ}|}\sum_{i=1:s_i=b,y_i\in\circ}^{n} \diamond x_i\right)\right| \leq \widehat{\epsilon},
\end{aligned} \tag{4.38}$$

which is a convex problem. The problem of the convex relaxation of fairness constraint with ℓ_{Lin} is its limited theoretical groundness. For this reason, recently,

another approximation has been proposed. In particular, let us note that

$$\ell_{\text{Hard}}(h(x), y) \leq \ell_{\text{Up}}(h(x), y) = \ell_{\text{Hinge}}(h(x), y), \tag{4.39}$$

$$\ell_{\text{Hard}}(h(x), y) \geq \ell_{\text{Low}}(h(x), y) = \min[1, 1 - yh(x)]. \tag{4.40}$$

Then

$$\begin{aligned} \widehat{\mathrm{F}}^{\star}_{\ell_{\text{Hard}},\diamond}(h) &= \mathbb{E}_{\circ\in\star}\left\{\left|\widehat{\mathrm{L}}^{a,\circ}_{\ell_{\text{Hard}},\diamond}(h) - \widehat{\mathrm{L}}^{b,\circ}_{\ell_{\text{Hard}},\diamond}(h)\right|\right\} \\ &\leq \mathbb{E}_{\circ\in\star}\left\{\begin{cases} \widehat{\mathrm{L}}^{a,\circ}_{\ell_{\text{Up}},\diamond}(h) - \widehat{\mathrm{L}}^{b,\circ}_{\ell_{\text{Low}},\diamond}(h) & \text{if } \widehat{\mathrm{L}}^{a,\circ}_{\ell_{\text{Hard}},\diamond}(h) \geq \widehat{\mathrm{L}}^{b,\circ}_{\ell_{\text{Hard}},\diamond}(h), \\ \widehat{\mathrm{L}}^{b,\circ}_{\ell_{\text{Up}},\diamond}(h) - \widehat{\mathrm{L}}^{a,\circ}_{\ell_{\text{Low}},\diamond}(h) & \text{otherwise} \end{cases}\right\}. \end{aligned} \tag{4.41}$$

Note that this approximation results in two convex constraints where the condition can be verified after the problem is solved with just one of the two constraints active at the time. Note also that most of the fairness metrics (e.g., DP, EOp, and EOd) are expressed in terms of $\widehat{\mathrm{F}}^{\star}_{\ell_{\text{Hard}},\diamond}(h)$ meaning that this relaxation is the exact counterpart of the relaxation with the Hinge loss of the Hard loss in the ERM. With this new relaxation of the fairness constraint, the counterpart of Problem (4.32) for FERM is

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \quad \sum_{i=1}^{n} \max[0, 1 - y_i(w^\top x_i + b))] + \lambda\|w\|_2^2, \tag{4.42}$$

$$\text{s.t.} \quad \frac{1}{|\star|}\sum_{\circ\in\star}\begin{cases} \frac{1}{|n^{a,\circ}|}\sum_{i=1:s_i=a,y_i\in\circ}^{n}\max[0,1-\diamond w^\top x_i] \\ \quad -\frac{1}{|n^{b,\circ}|}\sum_{i=1:s_i=b,y_i\in\circ}^{n}\min[1,-\diamond\, w^\top x_i] \\ \quad \text{if } \widehat{\mathrm{L}}^{a,\circ}_{\ell_{\text{Hard}},\diamond}(h) \geq \widehat{\mathrm{L}}^{b,\circ}_{\ell_{\text{Hard}},\diamond}(h), \\ \frac{1}{|n^{b,\circ}|}\sum_{i=1:s_i=b,y_i\in\circ}^{n}\max[0,1-\diamond w^\top x_i] \\ \quad -\frac{1}{|n^{a,\circ}|}\sum_{i=1:s_i=a,y_i\in\circ}^{n}\min[1,-\diamond\, w^\top x_i] \\ \quad \text{otherwise} \end{cases} \leq \widehat{\epsilon},$$

which can be solved with a series of convex problems.

For what concerns Problem (4.19), namely RFERM, we need to combine the

approach developed for RERM and FERM. This procedure is not trivial because of the fairness constraint. In fact, if we relax the Fairness constraint $\widehat{\mathtt{F}}^{\star}_{\widetilde{\ell},\diamond}(h) \leq \widehat{\epsilon}$ with the linear loss as in FERM and we use the same $\mathcal{B}(x)$ that we exploited in RERM we obtain

$$\begin{aligned}
\widehat{\mathtt{F}}^{\star}_{\widetilde{\ell}_{\text{Lin}},\diamond}(h) &= \frac{1}{2}\mathbb{E}_{\circ\in\star}\left\{\left|\widehat{\mathbb{E}}_{x,s,y}\left\{\max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho}\left(1-\diamond w^\top\widetilde{x}\right)\ \middle|\ s=a, y\in\circ\right\}\right.\right. \\
&\qquad\left.\left.-\widehat{\mathbb{E}}_{x,s,y}\left\{\max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho}\left(1-\diamond w^\top\widetilde{x}\right)\ \middle|\ s=b, y\in\circ\right\}\right|\right\} \\
&= \frac{1}{2}\mathbb{E}_{\circ\in\star}\left\{\left|\widehat{\mathbb{E}}_{x,s,y}\left\{1-\diamond w^\top x_i+\|w\|\rho\ \middle|\ s=a, y\in\circ\right\}\right.\right. \\
&\qquad\left.\left.-\widehat{\mathbb{E}}_{x,s,y}\left\{1-\diamond w^\top x_i+\|w\|\rho\ \middle|\ s=b, y\in\circ\right\}\right|\right\} \\
&= \frac{1}{2}\mathbb{E}_{\circ\in\star}\left\{\left|\widehat{\mathbb{E}}_{x,s,y}\left\{\diamond w^\top x_i\ \middle|\ s=a, y\in\circ\right\}\right.\right. \\
&\qquad\left.\left.-\widehat{\mathbb{E}}_{x,s,y}\left\{\diamond w^\top x_i\ \middle|\ s=b, y\in\circ\right\}\right|\right\} = \widehat{\mathtt{F}}^{\star}_{\ell_{\text{Lin}},\diamond}(h), \qquad (4.43)
\end{aligned}$$

namely the robust fairness metric degenerates in the classical fairness metric. This means that the linear loss is a too-raw approximation that leads to degenerate solutions: no robustness in the fairness metrics. This result is another confirmation that the use of the linear loss is not theoretically grounded. For this reason, the only alternative we have is to use the second relaxation we exploited in FERM. Specifically, we can note that

$$
\begin{aligned}
&\widehat{\mathtt{F}}^{\star}_{\tilde{\ell}_{\mathrm{Hard}},\diamond}(h) \\
&= \mathbb{E}_{\circ\in\star}\left\{\left|\widehat{\mathtt{L}}^{a,\circ}_{\tilde{\ell}_{\mathrm{Hard}},\diamond}(h) - \widehat{\mathtt{L}}^{b,\circ}_{\tilde{\ell}_{\mathrm{Hard}},\diamond}(h)\right|\right\} \\
&\leq \mathbb{E}_{\circ\in\star}\left\{\begin{cases}\widehat{\mathtt{L}}^{a,\circ}_{\tilde{\ell}_{\mathrm{Up}},\diamond}(h) - \widehat{\mathtt{L}}^{b,\circ}_{\tilde{\ell}_{\mathrm{Low}},\diamond}(h) & \text{if } \widehat{\mathtt{L}}^{a,\circ}_{\tilde{\ell}_{\mathrm{Hard}},\diamond}(h) \geq \widehat{\mathtt{L}}^{b,\circ}_{\tilde{\ell}_{\mathrm{Hard}},\diamond}(h), \\ \widehat{\mathtt{L}}^{b,\circ}_{\tilde{\ell}_{\mathrm{Up}},\diamond}(h) - \widehat{\mathtt{L}}^{a,\circ}_{\tilde{\ell}_{\mathrm{Low}},\diamond}(h) & \text{otherwise}\end{cases}\right\} \\
&= \frac{1}{|\star|}\sum_{\circ\in\star}\begin{cases}\frac{1}{|n^{a,\circ}|}\sum_{i=1:s_i=a,y_i\in\circ}^{n}\max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho}\max[0,1-\diamond w^\top x_i] \\ \quad -\frac{1}{|n^{b,\circ}|}\sum_{i=1:s_i=b,y_i\in\circ}^{n}\max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho}\min[1,-\diamond w^\top x_i] \\ \quad \text{if } \widehat{\mathtt{L}}^{a,\circ}_{\tilde{\ell}_{\mathrm{Hard}},\diamond}(h) \geq \widehat{\mathtt{L}}^{b,\circ}_{\tilde{\ell}_{\mathrm{Hard}},\diamond}(h), \\ \frac{1}{|n^{b,\circ}|}\sum_{i=1:s_i=b,y_i\in\circ}^{n}\max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho}\max[0,1-\diamond w^\top x_i] \\ \quad -\frac{1}{|n^{a,\circ}|}\sum_{i=1:s_i=a,y_i\in\circ}^{n}\max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho}\min[1,-\diamond w^\top x_i] \\ \quad \text{otherwise}\end{cases} \\
&= \frac{1}{|\star|}\sum_{\circ\in\star}\begin{cases}\frac{1}{|n^{a,\circ}|}\sum_{i=1:s_i=a,y_i\in\circ}^{n}\max[0,1-\diamond w^\top x_i+\|w\|\rho] \\ \quad -\frac{1}{|n^{b,\circ}|}\sum_{i=1:s_i=b,y_i\in\circ}^{n}\min[1,-\diamond w^\top x_i+\|w\|\rho] \\ \quad \text{if } \widehat{\mathtt{L}}^{a,\circ}_{\ell_{\mathrm{Hard}},\diamond}(h) \geq \widehat{\mathtt{L}}^{b,\circ}_{\ell_{\mathrm{Hard}},\diamond}(h), \\ \frac{1}{|n^{b,\circ}|}\sum_{i=1:s_i=b,y_i\in\circ}^{n}\max[0,1-\diamond w^\top x_i+\|w\|\rho] \\ \quad -\frac{1}{|n^{a,\circ}|}\sum_{i=1:s_i=a,y_i\in\circ}^{n}\min[1,-\diamond w^\top x_i+\|w\|\rho] \\ \quad \text{otherwise}\end{cases}.
\end{aligned}
\tag{4.44}
$$

In this case, the $\widehat{\mathtt{F}}^{\star}_{\tilde{\ell}_{\mathrm{Hard}},\diamond}(h)$ is again sensitive to the perturbation. Note also that computing $\widehat{\mathtt{L}}^{\cdot,\circ}_{\tilde{\ell}_{\mathrm{Hard}},\diamond}(h)$ is not hard

$$
\begin{aligned}
\widehat{\mathtt{L}}^{\cdot,\circ}_{\tilde{\ell}_{\mathrm{Hard}},\diamond}(h) &= \frac{1}{n^{\cdot,\circ}}\sum_{i=1:s_i=\cdot,y_i\in\circ}^{n}\max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho}[y_i w^\top\widetilde{x}\leq 0] \\
&= \frac{1}{n^{\cdot,\circ}}\sum_{i=1:s_i=\cdot,y_i\in\circ}^{n}[y_i w^\top x_i - \|w\|\rho\leq 0].
\end{aligned}
\tag{4.45}
$$

Thanks to this result, we can obtain the counterpart of Problem (4.32), for

RFERM exploiting the same step adopted for FERM and RERM

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \quad \sum_{i=1}^{n} \max[0, 1 - y_i(w^\top x_i + b) + \|w\|\rho] + \lambda\|w\|_2^2, \tag{4.46}$$

$$\text{s.t.} \quad \widehat{\epsilon} \geq \frac{1}{|\star|} \sum_{\circ\in\star} \begin{cases} \frac{1}{|n^{a,\circ}|} \sum_{i=1:s_i=a,y_i\in\circ}^{n} \max[0, 1 - \diamond w^\top x_i + \|w\|\rho] \\ \quad - \frac{1}{|n^{b,\circ}|} \sum_{i=1:s_i=b,y_i\in\circ}^{n} \min[1, -\diamond w^\top x_i + \|w\|\rho], \\ \quad \text{if } \widehat{\mathrm{L}}^{a,\circ}_{\tilde{\ell}_{\mathrm{Hard}},\diamond}(h) \geq \widehat{\mathrm{L}}^{b,\circ}_{\tilde{\ell}_{\mathrm{Hard}},\diamond}(h), \\ \frac{1}{|n^{b,\circ}|} \sum_{i=1:s_i=b,y_i\in\circ}^{n} \max[0, 1 - \diamond w^\top x_i + \|w\|\rho] \\ \quad - \frac{1}{|n^{a,\circ}|} \sum_{i=1:s_i=a,y_i\in\circ}^{n} \min[1, -\diamond w^\top x_i + \|w\|\rho], \\ \quad \text{otherwise} \end{cases}.$$

This problem can be easily solved with a series of convex optimization problems like the following one

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \quad \sum_{i=1}^{n} \max[0, 1 - y_i(w^\top x_i + b) + W\rho] + \lambda\|w\|_2^2, \tag{4.47}$$

$$\text{s.t.} \quad \widehat{\epsilon} \geq \frac{1}{|\star|} \sum_{\circ\in\star} \begin{cases} \frac{1}{|n^{a,\circ}|} \sum_{i=1:s_i=a,y_i\in\circ}^{n} \max[0, 1 - \diamond w^\top x_i + W\rho] \\ \quad - \frac{1}{|n^{b,\circ}|} \sum_{i=1:s_i=b,y_i\in\circ}^{n} \min[1, -\diamond w^\top x_i + W\rho], \\ \quad \text{if } \widehat{\mathrm{L}}^{a,\circ}_{\tilde{\ell}_{\mathrm{Hard}},\diamond}(h) \geq \widehat{\mathrm{L}}^{b,\circ}_{\tilde{\ell}_{\mathrm{Hard}},\diamond}(h), \\ \frac{1}{|n^{b,\circ}|} \sum_{i=1:s_i=b,y_i\in\circ}^{n} \max[0, 1 - \diamond w^\top x_i + W\rho] \\ \quad - \frac{1}{|n^{a,\circ}|} \sum_{i=1:s_i=a,y_i\in\circ}^{n} \min[1, -\diamond w^\top x_i + W\rho], \\ \quad \text{Otherwise} \end{cases},$$

line searching $W \in [0, \infty)$ so to find the solution w^* at minimal but with $\|w^*\| = W$.

As a final remark, note that the RFERM optimization problem of Eq. (4.47), degenerates into the FERM optimization problem of Eq. (4.42) for $\rho = 0$, into the RERM optimization problem of Eq. (4.35) for $\widehat{\epsilon} = \infty$, and into the ERM optimization problem of Eq. (4.32) for $\rho = 0$ and $\widehat{\epsilon} = \infty$.

4.4.2 Second Step: Kernel Methods

Let us now consider $\mathcal{H}$ as the compact subset of linear separators in a Reproducing Kernel Hilbert Space (RKHS). We let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel and let $\phi : \mathcal{X} \to \mathbb{H}$ be an induced feature mapping such that $\kappa(x, x') = \phi(x)^\top \phi(x')$, for all $x, x' \in \mathcal{X}$, where $\mathbb{H}$ is the Hilbert space of square summable sequences. Functions in the RKHS can be parameterized as $h(x) = w^\top \phi(x) + b$ with $w \in \mathbb{H}$ and $b \in \mathbb{R}$.

Let us consider then the ERM, the RERM, the FERM, and RFERM described in the case of linear models (Section 4.4.1 Problems (4.32), (4.35), (4.42), and (4.47)). Note that, as described in Section 4.4.1, we can consider just Problem (4.47) since it reduces to all the others for a particular value of ρ and $\widehat{\epsilon}$. The only difference is that, in the case of kernel methods, we have that

$$\begin{aligned}
\tilde{\ell}_{\text{Hard}}(h(x), y) &= \max_{\widetilde{x} \in \mathbb{R}^d : \|\widetilde{x} - x_i\|_2 \leq \rho} \ell_{\text{Hard}}(h(\widetilde{x}), y) \\
&= \max_{\widetilde{x} \in \mathbb{R}^d : \|\widetilde{x} - x_i\|_2 \leq \rho} [y(w^\top \phi(\widetilde{x}) + b) \leq 0] \qquad (4.48) \\
\tilde{\ell}_{\text{Hinge}}(h(x), y) &= \tilde{\ell}_{\text{Up}}(h(x), y) = \max_{\widetilde{x} \in \mathbb{R}^d : \|\widetilde{x} - x_i\|_2 \leq \rho} \ell_{\text{Hinge}}(h(\widetilde{x}), y) \\
&= \max_{\widetilde{x} \in \mathbb{R}^d : \|\widetilde{x} - x_i\|_2 \leq \rho} \max[0, 1 - y(w^\top \phi(\widetilde{x}) + b)], \qquad (4.49) \\
\tilde{\ell}_{\text{Low}}(h(x), y) &= \max_{\widetilde{x} \in \mathbb{R}^d : \|\widetilde{x} - x_i\|_2 \leq \rho} \ell_{\text{Low}}(h(\widetilde{x}), y) \\
&= \max_{\widetilde{x} \in \mathbb{R}^d : \|\widetilde{x} - x_i\|_2 \leq \rho} \min[1, -y(w^\top \phi(\widetilde{x}) + b)], \qquad (4.50)
\end{aligned}$$

have no closed-form solutions. The only solution here is to perform the optimization via gradient descent on $\widetilde{x}$ [16]. In fact, following the same procedure adopted

in Section 4.4.1 to obtain Problem (4.47), it is possible to get

$$\min_{w\in\mathbb{H},b\in\mathbb{R}} \quad \sum_{i=1}^{n} \max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho} \max[0, 1-y_i(w^\top\phi(\widetilde{x})+b)] + \lambda\|w\|_2^2, \tag{4.51}$$

$$\text{s.t.} \quad \frac{1}{|\star|}\sum_{\circ\in\star} \begin{cases} \frac{1}{|n^{a,\circ}|}\sum_{i=1:s_i=a,y_i\in\circ}^{n} \max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho} \max[0,1-\diamond w^\top\widetilde{x}] \\ \quad -\frac{1}{|n^{b,\circ}|}\sum_{i=1:s_i=b,y_i\in\circ}^{n} \max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho} \min[1,-\diamond w^\top\widetilde{x}], \\ \quad \text{if } \widehat{\mathtt{L}}^{a,\circ}_{\tilde{\ell}_{\text{Hard}},\diamond}(h) \geq \widehat{\mathtt{L}}^{b,\circ}_{\tilde{\ell}_{\text{Hard}},\diamond}(h), \\ \frac{1}{|n^{b,\circ}|}\sum_{i=1:s_i=b,y_i\in\circ}^{n} \max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho} \max[0,1-\diamond w^\top\widetilde{x}] \\ \quad -\frac{1}{|n^{a,\circ}|}\sum_{i=1:s_i=a,y_i\in\circ}^{n} \max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho} \min[1,-\diamond w^\top\widetilde{x}], \\ \quad \text{otherwise} \end{cases} \leq \widehat{\epsilon},$$

Exploiting the representer theorem [187] it is possible to reformulate kernelized Problem (4.51) by minimizing over $\alpha \in \mathbb{R}^n$ simply making the following now-classical substitutions

$$w = \sum_{i=1}^{n} \alpha_i \phi(x_i) \tag{4.52}$$

and noting that

$$f(x) = w^\top x + b = \sum_{i=1}^{n} \alpha_i \phi(x_i)^\top \phi(x) + b = \sum_{i=1}^{n} \alpha_i \kappa(x_i, x) + b, \tag{4.53}$$

$$\|w\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \phi(x_i)^\top\phi(x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \kappa(x_i, x_j). \tag{4.54}$$

Consequently, the kerneled version of Problem (4.47) (RFERM) is then

$$\min_{\alpha\in\mathbb{R}^n, b\in\mathbb{R}} \quad \sum_{i=1}^{n} \max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho} \max\left[0, 1-y_i\left(\sum_{j=1}^{n}\alpha_j\kappa(x_j,\widetilde{x})+b\right)\right] + \lambda\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa(x_i,x_j), \tag{4.55}$$

$$\text{s.t.} \quad \widehat{\epsilon} \geq \frac{1}{|\star|}\sum_{\circ\in\star}\begin{cases} \frac{1}{|n^{a,\circ}|}\sum_{i=1:s_i=a,y_i\in\circ}^{n}\max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho}\max\left[0,1-\diamond\sum_{j=1}^{n}\alpha_j\kappa(x_j,\widetilde{x})\right] \\ \quad -\frac{1}{|n^{b,\circ}|}\sum_{i=1:s_i=b,y_i\in\circ}^{n}\max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho}\min\left[1,-\diamond\sum_{j=1}^{n}\alpha_j\kappa(x_j,\widetilde{x})\right], \\ \quad \text{if } \widehat{\mathrm{L}}^{a,\circ}_{\tilde{\ell}_{\text{Hard}},\diamond}(h) \geq \widehat{\mathrm{L}}^{b,\circ}_{\tilde{\ell}_{\text{Hard}},\diamond}(h), \\ \frac{1}{|n^{b,\circ}|}\sum_{i=1:s_i=b,y_i\in\circ}^{n}\max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho}\max\left[0,1-\diamond\sum_{j=1}^{n}\alpha_j\kappa(x_j,\widetilde{x})\right] \\ \quad -\frac{1}{|n^{a,\circ}|}\sum_{i=1:s_i=a,y_i\in\circ}^{n}\max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho}\min\left[1,-\diamond\sum_{j=1}^{n}\alpha_j\kappa(x_j,\widetilde{x})\right], \\ \quad \text{otherwise} \end{cases},$$

This problem is the kernelized version of Problem (4.47) of Section 4.4.1.

Note that in the case of ERM ($\rho = 0$ and $\widehat{\epsilon} = \infty$) and FERM ($\rho = 0$) Problem (4.55) is a simple convex problem that can be solved analogously to the ones of Problems (4.32) and (4.42) respectively.

Instead, for RERM ($\widehat{\epsilon} = \infty$) and RFERM, we have a min-max optimization problem. In particular, the min is $\min_{\alpha\in\mathbb{R}^n,b\in\mathbb{R}}\cdots$ while the max is the series of $\max_{\widetilde{x}\in\mathbb{R}^d:\|\widetilde{x}-x_i\|_2\leq\rho}\cdots$. This problem is, of course, non-convex. Nevertheless, the standard solution to solve this problem is to perform adversarial training meaning:

1. We fix all the $\widetilde{x}$ to x_i in Problem (4.55) (i.e., we do not address the max);
2. We optimize them following Problem (4.55) which is now convex, obtaining $\widetilde{\alpha}^*$ and b^* (we can use the same approaches we adopt in Problems (4.35) and (4.47));
3. We fix $\widetilde{\alpha} = \widetilde{\alpha}^*$ and $b = b^*$ and we optimized all the $\widetilde{x}$ through gradient descent (it can be done for all differentiable κ) obtaining a series of $\widetilde{x}_i^*$ (this can be done with single or multiple gradient steps but also with other

optimization algorithms);

4. We put all the $(\widetilde{x}_i^*, y_i)$ found at previous step in $\mathcal{D}$;
5. We optimize them Problem (4.55) which is now convex obtaining $\widetilde{\alpha}^*$ and b^* with the new $\mathcal{D}$ enriched at the previous step;
6. Optionally, we repeat Step 3-5 multiple times.

Finally note that the check of the condition $\widehat{\mathtt{L}}^{a,\circ}_{\tilde{\ell}_{\text{Hard}},\diamond}(h) \geq \widehat{\mathtt{L}}^{b,\circ}_{\tilde{\ell}_{\text{Hard}},\diamond}(h)$ require again to solve a non-convex problem that is usually solved via gradient descent (see Eq. (4.48)).

4.4.3 Third Step: Neural Networks

As third and final step, we will discuss how to apply our approach in the field of (Deep) Neural Networks, namely a sort of Kernel Method but where the projection ϕ is not fixed nor defined by the kernel function κ but it is learned directly from the data. This case is quite important especially when data are non-tabular (e.g., sequences, graphs, and images). Because of the generality of this section, we need to opt for a more high-level notation. In particular, we need to define a model functional form f_θ characterized by a set of parameters θ. Then, based on $\mathcal{D}$, θ is selected as follows

$$\min_{\theta} \quad \widehat{\mathtt{C}}(\theta), \tag{4.56}$$

where $\widehat{\mathtt{C}}(\theta)$ is a cost function containing at least an utility term $\widehat{\mathtt{U}}(\theta)$ (e.g., cross-entropy) possibly plus one or more regularizers $\widehat{\mathtt{R}}(\theta)$ (e.g., p-norm of the weights) [188]. Regularizers, can be also implicitly enforced via model functional form: choosing a particular type layers (e.g., convolutions, attentions, and dropout) or via layer combination (e.g., residual net, and inception net). Implicit regularizers can be also more complex like overparameterization [189; 190], or via optimization algo-

rithm used to solve Problem (4.56) (e.g., early stopping in gradient decent [188]). This approach corresponds to ERM in Sections 4.4.1 and 4.4.2. To extend this formalization to RERM we need to note that $\widehat{\mathtt{U}}(\theta)$ is mostly a sum of point-wise utility, so Problem (4.56) can be reformulated as

$$\min_{\theta} \quad \sum_{i=1}^{n} \widehat{\mathtt{U}}_i(\theta, x_i, y_i) + \widehat{\mathtt{R}}(\theta), \tag{4.57}$$

where $\widehat{\mathtt{U}}_i(\theta, x_i, y_i)$ is the point-wise utility term on the i^{th} sample of $\mathcal{D}$. Then, in order to extend Problem (4.57) to RERM we then just need to

$$\min_{\theta} \quad \sum_{i=1}^{n} \max_{\widetilde{x} \in \mathbb{B}(x_i)} \widehat{\mathtt{U}}(\theta, \widetilde{x}, y_i) + \widehat{\mathtt{R}}(\theta). \tag{4.58}$$

Problem (4.58), the RERM formalization in the case of (Deep) Neural Networks, is usually solved via adversarial training, namely we solve $\min_\theta$ fixing $\widetilde{x} = x_i$, and we found θ^* then we optimize $\max_{\widetilde{x} \in \mathbb{B}(x_i)}$ (usually via gradient descent) and we include all these data in $\mathcal{D}$, then we perform again $\min_\theta$, possibly repeating these step analogously to what has been explained for Problem (4.55). This approach corresponds to RERM in Sections 4.4.1 and 4.4.2.

To extend Problems (4.56) and (4.58), namely ERM and RERM in (Deep) Neural Networks, to FERM and RFERM the step is quite simple and follows directly from Sections 4.4.1 and 4.4.2. FERM and RFERM of Sections 4.4.1 and 4.4.2 are constrained version of the optimization problem of ERM and RERM. Unfortunately, solving constrained optimization problems in (Deep) Neural Networks is not easy nor computationally efficient. Consequently, the idea is that, using the partial Lagrangian relaxation, the constraints can be saw as regularizers.

In particular, form FERM the constraint of Problem (4.51) can be saw as

$\widehat{\mathtt{F}}(\theta) \leq \widehat{\epsilon}$ with

$$\widehat{\mathtt{F}}(\theta) = \frac{1}{|\star|} \sum_{\circ \in \star} \begin{cases} \frac{1}{|n^{a,\circ}|} \sum_{i=1:s_i=a,y_i \in \circ}^{n} \max\left[0, 1 - \diamond f_\theta(x_i)\right] \\ \quad - \frac{1}{|n^{b,\circ}|} \sum_{i=1:s_i=b,y_i \in \circ}^{n} \min\left[1, - \diamond f_\theta(x_i)\right] \\ \quad \text{if } \widehat{\mathtt{L}}^{a,\circ}_{\tilde{\ell}_{\text{Hard}},\diamond}(h) \geq \widehat{\mathtt{L}}^{b,\circ}_{\tilde{\ell}_{\text{Hard}},\diamond}(h), \\ \frac{1}{|n^{b,\circ}|} \sum_{i=1:s_i=b,y_i \in \circ}^{n} \max\left[0, 1 - \diamond f_\theta(x_i)\right] \\ \quad - \frac{1}{|n^{a,\circ}|} \sum_{i=1:s_i=a,y_i \in \circ}^{n} \min\left[1, - \diamond f_\theta(x_i)\right] \\ \quad \text{otherwise} \end{cases}, \tag{4.59}$$

and then the FERM version of Problems (4.56) becomes

$$\min_{\theta} \quad \widehat{\mathtt{C}}(\theta) + \lambda \widehat{\mathtt{F}}(\theta), \tag{4.60}$$

a differentiable problem, if ERM was differentiable, since we added a convex and differentiable term and where $\lambda \in [0, \infty)$ is an hyperparameter to be tuned.

Form RERM the constraint of Problem (4.51) can be saw as $\widehat{\widetilde{\mathtt{F}}}(\theta) \leq \widehat{\epsilon}$ with

$$\widehat{\widetilde{\mathtt{F}}}(\theta) = \frac{1}{|\star|} \sum_{\circ \in \star} \begin{cases} \frac{1}{|n^{a,\circ}|} \sum_{i=1:s_i=a,y_i \in \circ}^{n} \max_{\widetilde{x} \in \mathbb{B}(x_i)} \max\left[0, 1 - \diamond f_\theta(\widetilde{x})\right] \\ \quad - \frac{1}{|n^{b,\circ}|} \sum_{i=1:s_i=b,y_i \in \circ}^{n} \max_{\widetilde{x} \in \mathbb{B}(x_i)} \min\left[1, - \diamond f_\theta(\widetilde{x})\right] \\ \quad \text{if } \widehat{\mathtt{L}}^{a,\circ}_{\tilde{\ell}_{\text{Hard}},\diamond}(h) \geq \widehat{\mathtt{L}}^{b,\circ}_{\tilde{\ell}_{\text{Hard}},\diamond}(h), \\ \frac{1}{|n^{b,\circ}|} \sum_{i=1:s_i=b,y_i \in \circ}^{n} \max_{\widetilde{x} \in \mathbb{B}(x_i)} \max\left[0, 1 - \diamond f_\theta(\widetilde{x})\right)] \\ \quad - \frac{1}{|n^{a,\circ}|} \sum_{i=1:s_i=a,y_i \in \circ}^{n} \max_{\widetilde{x} \in \mathbb{B}(x_i)} \min\left[1, - \diamond f_\theta(\widetilde{x})\right] \\ \quad \text{otherwise} \end{cases}, \tag{4.61}$$

where, if we use and L2 ball as space of perturbations we get

$$\widehat{\widetilde{\mathtt{F}}}(\theta) = \frac{1}{|\star|} \sum_{\circ \in \star} \begin{cases} \frac{1}{|n^{a,\circ}|} \sum_{i=1:s_i=a,y_i \in \circ}^{n} \max_{\widetilde{x} \in \mathbb{R}^d : \|\widetilde{x} - x_i\|_2 \leq \rho} \max\left[0, 1 - \diamond f_\theta(\widetilde{x})\right] \\ \quad - \frac{1}{|n^{b,\circ}|} \sum_{i=1:s_i=b,y_i \in \circ}^{n} \max_{\widetilde{x} \in \mathbb{R}^d : \|\widetilde{x} - x_i\|_2 \leq \rho} \min\left[1, - \diamond f_\theta(\widetilde{x})\right] \\ \quad \text{if } \widehat{\mathtt{L}}^{a,\circ}_{\tilde{\ell}_{\text{Hard}},\diamond}(h) \geq \widehat{\mathtt{L}}^{b,\circ}_{\tilde{\ell}_{\text{Hard}},\diamond}(h), \\ \frac{1}{|n^{b,\circ}|} \sum_{i=1:s_i=b,y_i \in \circ}^{n} \max_{\widetilde{x} \in \mathbb{R}^d : \|\widetilde{x} - x_i\|_2 \leq \rho} \max\left[0, 1 - \diamond f_\theta(\widetilde{x})\right)] \\ \quad - \frac{1}{|n^{a,\circ}|} \sum_{i=1:s_i=a,y_i \in \circ}^{n} \max_{\widetilde{x} \in \mathbb{R}^d : \|\widetilde{x} - x_i\|_2 \leq \rho} \min\left[1, - \diamond f_\theta(\widetilde{x})\right] \\ \quad \text{otherwise} \end{cases}. \tag{4.62}$$

Then the RFERM version of Problems (4.58) becomes

$$\min_{\theta} \quad \sum_{i=1}^{n} \max_{\widetilde{x}\in\mathbb{B}(x_i)} \widehat{\mathtt{U}}(\theta, \widetilde{x}, y_i) + \widehat{\mathtt{R}}(\theta) + \lambda\widehat{\widetilde{\mathtt{F}}}(\theta), \tag{4.63}$$

that can be solved exactly like Problems (4.58).

To sum up, $\widehat{\mathtt{F}}(\theta)$ and $\widehat{\widetilde{\mathtt{F}}}(\theta)$ are just theoretically grounded regularizer to incorporate into the learning phase to make any algorithm fair or robustly fair respectively.

4.5 Conclusions

In this work, we introduced Robust Fair Empirical Risk Minimization (RFERM), a novel optimization paradigm that jointly addresses adversarial robustness and algorithmic fairness. Building on the foundations of ERM, RERM, and FERM, RFERM provides a principled approach to mitigating robustness bias — the phenomenon whereby adversarial perturbations disproportionately harm specific demographic subgroups.

From a theoretical perspective, we established the consistency of RFERM, proving that the empirical solution converges to its population counterpart in terms of both robustness and fairness guarantees. These results extend the statistical learning theory of fairness-constrained learning to adversarial settings, providing rigorous generalization bounds for robustly fair models.

On the algorithmic side, we proposed tractable relaxations of the RFERM problem for linear, kernelized, and deep learning models. For linear models, we derived convex approximations enabling efficient optimization. For kernel methods, we developed a bilevel optimization scheme based on adversarial training in RKHS, and for neural networks we showed how fairness and robust-fairness

constraints can be reformulated as differentiable regularizers suitable for gradient-based training.

Overall, this work bridges theoretical guarantees with practical implementations, laying the groundwork for the empirical development and test on real-world data.

Acknowledgments

This work includes the fundamental contribution of Danilo Franco and Luca Oneto.

Furthermore, this research is partially supported by (i) project ELSA – European Lighthouse on Secure and Safe AI funded by the European Union's Horizon Europe under the grant agreement No. 101070617, (ii) EU - NGEU National Sustainable Mobility Center (CN00000023) Italian Ministry of University and Research Decree n. 1033—17/06/2022 (Spoke 10), (iii) project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU, and (iv) project FAIR (PE00000013) under the NRRP MUR program funded by the EU - NGEU.

Chapter 5

Conclusions

This thesis set out to investigate the interdependence of fairness with two key dimensions of trustworthy ML: robustness and regression. The overarching aim was to advance both the theoretical understanding and the methodological treatment of fairness when considered not in isolation, but as part of a broader ecosystem of interacting properties that shape the reliability, equity, and societal impact of AI technologies.

The first contribution of this work was the identification and characterization of unfair regression, a previously unstudied phenomenon that arises when model updates, despite improving average performance, disproportionately degrade outcomes for specific demographic subgroups. To address this issue, new fairness metrics tailored to regression settings were introduced, alongside two mitigation strategies — one acting at the algorithmic level and the other at the hyperparameter tuning phase. Experimental analyses on real-world datasets confirmed both the existence of unfair regression and the effectiveness of the proposed mitigations, thereby establishing a foundation for future explorations of fairness in dynamic, evolving machine learning systems.

The second contribution concerned the interplay between fairness and robustness. To this end, the thesis introduced Robust Fair Empirical Risk Minimization

(RFERM), a novel optimization framework explicitly designed to address the already known problem of robustness bias. Theoretical results established the consistency of RFERM and extended the learning theory of fairness-constrained algorithms to adversarial contexts. Furthermore, tractable formulations were proposed for linear, kernel-based, and deep learning models, showing how fairness constraints can be embedded into practical training pipelines. This theoretical and methodological groundwork opens promising avenues for the development of models that are simultaneously fair, robust, and practically deployable.

Taken together, these contributions highlight the importance of treating fairness as an inherently relational property that cannot be disentangled from other desiderata of trustworthy AI. By documenting novel phenomena, proposing targeted mitigation strategies, and formulating principled optimization frameworks, this thesis advances the broader agenda of aligning machine learning systems with societal expectations of equity, accountability, and resilience.

Looking forward, several directions remain open. Extending the study of unfair regression to a wider spectrum of learning algorithms and real-world domains will be essential to assess its prevalence and refine mitigation techniques. Similarly, further empirical validation of RFERM will strengthen its role as a practical tool. Finally, integrating fairness-aware robustness methods with complementary principles such as transparency, interpretability, and privacy constitutes an exciting and necessary frontier in the ongoing pursuit of trustworthy AI.

In conclusion, this thesis underscores that fairness in machine learning is not a solitary constraint but a central element of a broader constellation of values that collectively determine the ethical and societal trajectory of AI systems. By investigating fairness at its intersections with robustness and regression, this work contributes to the conceptual and practical foundations required to build AI technologies that are not only effective, but also equitable and just.

References

[1] D. Ueda, T. Kakinuma, S. Fujita, K. Kamagata, Y. Fushimi, R. Ito, Y. Matsui, T. Nozaki, T. Nakaura, N. Fujima, *et al.*, "Fairness of artificial intelligence in healthcare: review and recommendations," *Japanese journal of radiology*, vol. 42, no. 1, pp. 3–15, 2024. 1, 6

[2] J. W. Anderson and S. Visweswaran, "Algorithmic individual fairness and healthcare: A scoping review," *JAMIA open*, vol. 8, no. 1, p. ooae149, 2025. 1, 6

[3] T. Bono, K. Croxson, and A. Giles, "Algorithmic fairness in credit scoring," *Oxford Review of Economic Policy*, vol. 37, no. 3, pp. 585–617, 2021. 1, 6

[4] N. Kozodoi, J. Jacob, and S. Lessmann, "Fairness in credit scoring: Assessment, implementation and profit implications," *European Journal of Operational Research*, vol. 297, no. 3, pp. 1083–1094, 2022. 1, 6

[5] A. Fabris, N. Baranowska, M. J. Dennis, D. Graus, P. Hacker, J. Saldivar, F. Zuiderveen Borgesius, and A. J. Biega, "Fairness and bias in algorithmic hiring: a multidisciplinary survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 1, pp. 1–54, 2025. 1, 6

[6] C. Rigotti and E. Fosch-Villaronga, "Fairness, AI & recruitment," *Computer Law & Security Review*, vol. 53, p. 105966, 2024. 1, 6

[7] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, *et al.*, "Practical lessons from predicting clicks on ads at facebook," in *Proceedings of the eighth international workshop on data mining for online advertising*, pp. 1–9, 2014. 1

[8] C. Perlich, B. Dalessandro, T. Raeder, O. Stitelman, and F. Provost, "Machine learning for targeted display advertising: Transfer learning in action," *Machine learning*, vol. 95, no. 1, pp. 103–127, 2014. 1

[9] V. Eubanks, *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018. 1

[10] A. Chouldechova, D. Benavides-Prado, O. Fialko, and R. Vaithianathan, "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions," in *Conference on fairness, accountability and transparency*, pp. 134–148, PMLR, 2018. 1

[11] L. Oneto, S. Ridella, and D. Anguita, "Towards algorithms and models that we can trust: A theoretical perspective," *Neurocomputing*, vol. 592, p. 127798, 2024. 1, 3, 23, 31, 36

[12] A. W. Flores, K. Bechtel, and C. T. Lowenkamp, "False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks," *Fed. Probation*, vol. 80, p. 38, 2016. 2

[13] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias." `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`. Accessed: 2025-08-13. 2

[14] X. Li and T. Zhang, "An exploration on artificial intelligence application: From security, privacy and ethic perspective," in *2017 IEEE 2nd interna-*

tional conference on cloud computing and big data analysis (ICCCBDA), pp. 416–420, IEEE, 2017. 2

[15] P. Rosati, P. Deeney, M. Cummins, L. Van der Werff, and T. Lynn, "Social media and stock price reaction to data breach announcements: Evidence from us listed companies," *Research in International Business and Finance*, vol. 47, pp. 458–469, 2019. 2

[16] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2154–2156, 2018. 2, 15, 16, 46, 49, 67

[17] A. Demontis, M. Melis, B. Biggio, D. Maiorca, D. Arp, K. Rieck, I. Corona, G. Giacinto, and F. Roli, "Yes, machine learning can be more secure! A case study on android malware detection," *IEEE transactions on dependable and secure computing*, vol. 16, no. 4, pp. 711–724, 2017. 2, 16

[18] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017. 2

[19] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017. 2

[20] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021. 2, 7, 8, 31, 46

[21] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," in *Ethics of data and analytics*, 2022. 3, 7

[22] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Transactions on information forensics and security*, vol. 7, no. 6, pp. 1789–1801, 2012. 3

[23] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?," in *CHI conference on human factors in computing systems*, 2019. 3

[24] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Conference on fairness, accountability, and transparency*, 2019. 3, 7

[25] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," *arXiv preprint arXiv:1810.08810*, 2018. 3

[26] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil, "Empirical risk minimization under fairness constraints," in *Neural information processing systems*, 2018. 3, 36, 47, 52, 55, 56, 61

[27] S. Thiebes, S. Lins, and A. Sunyaev, "Trustworthy artificial intelligence," *Electronic Markets*, vol. 31, no. 2, pp. 447–464, 2021. 3

[28] F. Ezzeddine, "Privacy implications of explainable AI in data-driven systems," *arXiv preprint arXiv:2406.15789*, 2024. 3

[29] H. Xu, X. Liu, Y. Li, A. Jain, and J. Tang, "To be robust or to be fair: Towards fairness in adversarial training," in *International conference on machine learning*, pp. 11492–11501, PMLR, 2021. 3, 24

[30] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, "Trustworthy AI: From principles to practices," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–46, 2023. 3, 23

[31] S. Barocas and A. D. Selbst, "Big data's disparate impact," in *California Law Review*, vol. 104, pp. 671–732, 2016. 6

[32] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science advances*, vol. 4, no. 1, p. eaao5580, 2018. 6

[33] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017. 6, 9

[34] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," in *Ethics of data and analytics*, pp. 254–264, Auerbach Publications, 2022. 7

[35] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, vol. 29, 2016. 7, 9, 11, 34, 51, 53

[36] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum, "Algorithmic fairness: Choices, assumptions, and definitions," *Annual Review of Statistics and Its Application*, vol. 8, pp. 141–163, 2021. 7

[37] H. Suresh and J. V. Guttag, "A framework for understanding unintended consequences of machine learning," *arXiv preprint arXiv:1901.10002*, 2019. 7, 8

[38] K. A. Clarke, "The phantom menace: Omitted variable bias in econometric research," *Conflict management and peace science*, vol. 22, no. 4, pp. 341–352, 2005. 7

[39] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? Comparing data from twitter's streaming api with twitter's firehose," in *Proceedings of the 7th international AAAI conference on weblogs and social media*, 2013. 7

[40] D. Danks and A. J. London, "Algorithmic bias in autonomous systems.," in *IJCAI*, pp. 4691–4697, 2017. 8

[41] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, pp. 77–91, 2018. 8, 15

[42] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, "Social data: Biases, methodological pitfalls, and ethical boundaries," *Frontiers in big data*, vol. 2, p. 13, 2019. 8

[43] T. Wang and D. Wang, "Why amazon's ratings might mislead you: The story of herding effects," *Big data*, vol. 2, no. 4, pp. 196–204, 2014. 8

[44] H. J. Miller, J. Thebault-Spieker, S. Chang, I. Johnson, L. Terveen, and B. Hecht, ""blissfully happy" or "ready to fight": Varying interpretations of emoji," in *Tenth International AAAI Conference on Web and Social Media*, 2016. 8

[45] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 59–68, ACM, 2019. 9

[46] R. T. Rabonato and L. Berton, "A systematic review of fairness in machine learning," *AI Ethics*, vol. 5, no. 3, pp. 1943–1954, 2025. 9

[47] N. A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, and Y. Liu, "How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 99–106, ACM, 2019. 9

[48] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 259–268, Association for Computing Machinery, 2015. 9, 10

[49] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010. 9, 34

[50] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Comput. Surv.*, vol. 55, no. 3, 2022. 9, 10, 29

[51] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012. 9

[52] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2017. 9

[53] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012. 10

[54] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii, "Fair clustering through fairlets," *Advances in neural information processing systems*, vol. 30, 2017. 10

[55] S. Samadi, U. Tantipongpipat, J. Morgenstern, M. Singh, and S. Vempala, "The price of fair PCA: One extra dimension," in *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 10999–11010, Curran Associates Inc., 2018. 10

[56] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50, Springer, 2012. 10

[57] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017. 10

[58] L. Oneto, M. Donini, and M. Pontil, "General fair empirical risk minimization," in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2020. 11

[59] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," in *2010 IEEE international conference on data mining*, pp. 869–874, IEEE, 2010. 11

[60] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*, pp. 325–333, 2013. 11

[61] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," *arXiv preprint arXiv:1511.00830*, 2015. 11

[62] N. Quadrianto and V. Sharmanska, "Recycling privileged learning and distribution matching for fairness," *Advances in neural information processing systems*, vol. 30, 2017. 11

[63] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806, ACM, 2017. 11

[64] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (S. A. Friedler and C. Wilson, eds.), vol. 81 of *Proceedings of Machine Learning Research*, pp. 107–118, PMLR, 2018. 11

[65] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson, "Decoupled classifiers for group-fair and efficient machine learning," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (S. A. Friedler and C. Wilson, eds.), vol. 81 of *Proceedings of Machine Learning Research*, pp. 119–133, PMLR, 2018. 11, 48

[66] O. Parraga, M. D. More, C. M. Oliveira, N. S. Gavenski, L. S. Kupssinskü, A. Medronha, L. V. Moura, G. S. Simões, and R. C. Barros, "Fairness in deep learning: A survey on vision and language research," *ACM Computing Surveys*, vol. 57, no. 6, pp. 1–40, 2025. 12, 14

[67] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022. 12

[68] K. Choi, A. Grover, T. Singh, R. Shu, and S. Ermon, "Fair generative modeling via weak supervision," in *International Conference on Machine Learning*, pp. 1887–1898, PMLR, 2020. 12

[69] J. Cho, A. Zala, and M. Bansal, "Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3043–3054, 2023. 12

[70] S. Gambs and R. C. Ngueveu, "Fair mapping," *arXiv preprint arXiv:2209.00617*, 2022. 12

[71] M. Ngxande, J.-R. Tapamo, and M. Burke, "Bias remediation in driver drowsiness detection systems using generative adversarial networks," *IEEE Access*, vol. 8, pp. 55592–55601, 2020. 12

[72] S. Rickman, "Evaluating gender bias in large language models in long-term care," *BMC Med Inform Decis Mak*, vol. 25, 2025. 13

[73] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," *Advances in neural information processing systems*, vol. 29, 2016. 13

[74] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017. 13

[75] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, "On measuring social biases in sentence encoders," *arXiv preprint arXiv:1903.10561*, 2019. 13

[76] S. Bordia and S. R. Bowman, "Identifying and reducing gender bias in word-level language models," *arXiv preprint arXiv:1904.03035*, 2019. 13

[77] K. Webster, X. Wang, I. Tenney, A. Beutel, E. Pitler, E. Pavlick, J. Chen, E. Chi, and S. Petrov, "Measuring and reducing gendered correlations in pre-trained models," *arXiv preprint arXiv:2010.06032*, 2020. 13

[78] R. Qian, C. Ross, J. Fernandes, E. Smith, D. Kiela, and A. Williams, "Perturbation augmentation for fairer NLP," *arXiv preprint arXiv:2205.12586*, 2022. 13

[79] S. E. Davis, C. Dorn, D. J. Park, and M. E. Matheny, "Emerging algorithmic bias: fairness drift as the next dimension of model maintenance and sustainability," *Journal of the American Medical Informatics Association*, vol. 32, no. 5, pp. 845–854, 2025. 14, 24

[80] H. R. Kirk, Y. Jun, F. Volpin, H. Iqbal, E. Benussi, F. Dreyer, A. Shtedritski, and Y. Asano, "Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models," *Advances in neural information processing systems*, vol. 34, pp. 2611–2624, 2021. 15

[81] W. Guo and A. Caliskan, "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 122–133, 2021. 15

[82] M. O'Brien, "Tech industry tried reducing AI's pervasive bias. Now Trump wants to end its 'woke AI' efforts." `https://apnews.com/article/artificial-intelligence-dei-trump-algorithmic-bias-woke-ai-8302e12dd74df69a1adc6565710f033d`. Accessed: 2025-08-13. 15

[83] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, "Adversarial classification," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 99–108, 2004. 15, 16

[84] D. Lowd and C. Meek, "Adversarial learning," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 641–647, 2005. 15

[85] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?," in *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, pp. 16–25, 2006. 15, 16

[86] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013. 15

[87] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International conference on machine learning*, pp. 1885–1894, PMLR, 2017. 15

[88] S. Mei and X. Zhu, "Using machine teaching to identify optimal training-set attacks on machine learners," in *Proceedings of the aaai conference on artificial intelligence*, vol. 29, 2015. 15

[89] L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, and F. Roli, "Towards poisoning of deep learning algorithms with back-gradient optimization," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 27–38, 2017. 15

[90] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402, Springer, 2013. 16

[91] O. Dekel and O. Shamir, "Learning to classify with missing and corrupted features," in *Proceedings of the 25th international conference on Machine learning*, pp. 216–223, 2008. 16

[92] A. Globerson and S. Roweis, "Nightmare at test time: Robust learning by feature deletion," in *Proceedings of the 23rd international conference on Machine learning*, pp. 353–360, 2006. 16

[93] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine learning*, vol. 81, no. 2, pp. 121–148, 2010. 16

[94] B. Biggio, G. Fumera, and F. Roli, "Security evaluation of pattern classifiers under attack," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 4, pp. 984–996, 2013. 16

[95] N. Šrndić and P. Laskov, "Practical evasion of a learning-based classifier: A case study," in *2014 IEEE symposium on security and privacy*, pp. 197–211, IEEE, 2014. 16

[96] M. Brückner, C. Kanzow, and T. Scheffer, "Static prediction games for adversarial learning problems," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2617–2654, 2012. 16

[97] Q. Zheng, X. Zou, Y. Dong, Y. Cen, D. Yin, J. Xu, Y. Yang, and J. Tang, "Graph robustness benchmark: Benchmarking the adversarial robustness of graph machine learning," *arXiv preprint arXiv:2111.04314*, 2021. 16

[98] N. Drenkow, N. Sani, I. Shpitser, and M. Unberath, "A systematic review of robustness in deep learning for computer vision: Mind the gap?," *arXiv preprint arXiv:2112.00639*, 2021. 17

[99] J. Wang, Z. Yin, P. Hu, A. Liu, R. Tao, H. Qin, X. Liu, and D. Tao, "Defensive patches for robust recognition in the physical world," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2456–2465, 2022. 17, 18

[100] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019. 17

[101] V. Shankar, A. Dave, R. Roelofs, D. Ramanan, B. Recht, and L. Schmidt, "Do image classifiers generalize across time?," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9661–9669, 2021. 17

[102] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *International conference on machine learning*, pp. 5637–5664, PMLR, 2021. 17

[103] A. Dittadi, S. Papa, M. De Vita, B. Schölkopf, O. Winther, and F. Locatello, "Generalization and robustness implications in object-centric learning," *arXiv preprint arXiv:2107.00637*, 2021. 17

[104] T. Gokhale, S. Mishra, M. Luo, B. S. Sachdeva, and C. Baral, "Generalized but not robust? Comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness," *arXiv preprint arXiv:2203.07653*, 2022. 17

[105] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky, "Human uncertainty makes classification more robust," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9617–9626, 2019. 17

[106] H. Chen, J. He, K. Narasimhan, and D. Chen, "Can rationalization improve robustness?," *arXiv preprint arXiv:2204.11790*, 2022. 18, 19

[107] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*, pp. 39–57, Ieee, 2017. 18

[108] S.-T. Chen, C. Cornelius, J. Martin, and D. H. Chau, "Shapeshifter: Robust physical adversarial attack on faster R-CNN object detector," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 52–68, Springer, 2018. 18

[109] W. Woods, J. Chen, and C. Teuscher, "Adversarial explanations for understanding image classification decisions and improved neural network robustness," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 508–516, 2019. 18

[110] A. N. Bhagoji, D. Cullina, C. Sitawarin, and P. Mittal, "Enhancing robustness of machine learning systems via data transformations," in *2018 52nd Annual conference on information sciences and systems (CISS)*, pp. 1–5, IEEE, 2018. 18

[111] C.-H. Chang, G. A. Adam, and A. Goldenberg, "Towards robust classification model by counterfactual and invariant data generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15212–15221, 2021. 18

[112] D. D. Thang and T. Matsui, "Image transformation can make neural networks more robust against adversarial examples," *arXiv preprint arXiv:1901.03037*, 2019. 18

[113] S. W. Akhtar, S. Rehman, M. Akhtar, M. A. Khan, F. Riaz, Q. Chaudry, and R. Young, "Improving the robustness of neural networks using k-support norm based adversarial training," *IEEE Access*, vol. 4, pp. 9501–9511, 2016. 18

[114] X. Chen, C. Xie, M. Tan, L. Zhang, C.-J. Hsieh, and B. Gong, "Robust and accurate object detection via adversarial learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16622–16631, 2021. 18

[115] K. Sun, Z. Zhu, and Z. Lin, "Enhancing the robustness of deep neural networks by boundary conditional GAN," *arXiv preprint arXiv:1902.11029*, 2019. 18, 20

[116] H. Wang and C.-N. Yu, "A direct approach to robust deep learning using adversarial networks," *arXiv preprint arXiv:1905.09591*, 2019. 18

[117] Z. Deng, L. Zhang, A. Ghorbani, and J. Zou, "Improving adversarial robustness via unlabeled out-of-domain data," in *International Conference on Artificial Intelligence and Statistics*, pp. 2845–2853, PMLR, 2021. 18

[118] M. Abdelaty, S. Scott-Hayward, R. Doriguzzi-Corin, and D. Siracusa, "GADoT: GAN-based adversarial training for robust DDoS attack detection," in *2021 IEEE Conference on Communications and Network Security (CNS)*, pp. 119–127, IEEE, 2021. 18

[119] P. Vaishnavi, T. Cong, K. Eykholt, A. Prakash, and A. Rahmati, "Can attention masks improve adversarial robustness?," in *International Work-*

shop on Engineering Dependable and Secure Machine Learning Systems, pp. 14–22, Springer, 2020. 18

[120] S. Gupta, P. Dube, and A. Verma, "Improving the affordability of robustness training for DNNs," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 780–781, 2020. 18, 19

[121] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017. 18, 19, 47, 50

[122] D. Stutz, M. Hein, and B. Schiele, "Confidence-calibrated adversarial training: Generalizing to unseen attacks," in *International conference on machine learning*, pp. 9155–9166, PMLR, 2020. 18

[123] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, "Improving adversarial robustness requires revisiting misclassified examples," in *International conference on learning representations*, 2019. 18

[124] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "A self-supervised approach for adversarial robustness," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 262–271, 2020. 19

[125] J.-B. Alayrac, J. Uesato, P.-S. Huang, A. Fawzi, R. Stanforth, and P. Kohli, "Are labels required for improving adversarial robustness?," *Advances in Neural Information Processing Systems*, vol. 32, 2019. 19

[126] M. Terzi, G. A. Susto, and P. Chaudhari, "Directional adversarial training for cost sensitive deep learning classification applications," *Engineering Applications of Artificial Intelligence*, vol. 91, p. 103550, 2020. 19

[127] C. Devaguptapu, D. Agarwal, G. Mittal, P. Gopalani, and V. N. Balasubramanian, "On adversarial robustness: A neural architecture search perspective," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 152–161, 2021. 19

[128] J. Jin, A. Dundar, and E. Culurciello, "Robust convolutional neural networks under adversarial noise," *arXiv preprint arXiv:1511.06306*, 2015. 19

[129] M. Momeny, A. M. Latif, M. A. Sarram, R. Sheikhpour, and Y. D. Zhang, "A noise robust convolutional neural network for image classification," *Results in Engineering*, vol. 10, p. 100225, 2021. 19

[130] C. Xu, X. Li, and M. Yang, "An orthogonal classifier for improving the adversarial robustness of neural networks," *Information Sciences*, vol. 591, pp. 251–262, 2022. 19

[131] Z. Li, S. Liu, X. Yu, K. Bhavya, J. Cao, J. D. Diffenderfer, P.-T. Bremer, and V. Pascucci, "Understanding robustness lottery: A geometric visual comparative analysis of neural network pruning approaches," *IEEE transactions on visualization and computer graphics*, 2024. 19

[132] D. Pruthi, B. Dhingra, and Z. C. Lipton, "Combating adversarial misspellings with robust word recognition," *arXiv preprint arXiv:1905.11268*, 2019. 20

[133] S. S. Roy, S. I. Hossain, M. Akhand, and K. Murase, "A robust system for noisy image classification combining denoising autoencoder and convolutional neural network," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 1, pp. 224–235, 2018. 20

[134] D. Ye, C. Chen, C. Liu, H. Wang, and S. Jiang, "Detection defense against

adversarial attacks with saliency map," *International Journal of Intelligent Systems*, vol. 37, no. 12, pp. 10193–10210, 2022. 20

[135] A. Zhou, B. Li, and H. Wang, "Robust prompt optimization for defending language models against jailbreaking attacks," *Advances in Neural Information Processing Systems*, vol. 37, pp. 40184–40211, 2024. 20

[136] D. A. Alber, Z. Yang, A. Alyakin, E. Yang, S. Rai, A. A. Valliani, J. Zhang, G. R. Rosenbaum, A. K. Amend-Thomas, D. B. Kurland, *et al.*, "Medical large language models are vulnerable to data-poisoning attacks," *Nature Medicine*, vol. 31, no. 2, pp. 618–626, 2025. 20

[137] T. Shen, R. Jin, Y. Huang, C. Liu, W. Dong, Z. Guo, X. Wu, Y. Liu, and D. Xiong, "Large language model alignment: A survey," *arXiv preprint arXiv:2309.15025*, 2023. 20

[138] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM computing surveys*, vol. 55, no. 12, pp. 1–38, 2023. 20

[139] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu, *et al.*, "A survey on in-context learning," *arXiv preprint arXiv:2301.00234*, 2022. 20

[140] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022. 20

[141] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," *arXiv preprint arXiv:2104.07567*, 2021. 20

[142] D. Li, B. Hu, Q. Chen, T. Xu, J. Tao, and Y. Zhang, "Unifying model explainability and robustness for joint text classification and rationale extraction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 10947–10955, 2022. 21

[143] A. Kortylewski, Q. Liu, A. Wang, Y. Sun, and A. Yuille, "Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion," *International Journal of Computer Vision*, vol. 129, no. 3, pp. 736–760, 2021. 21

[144] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J. F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," in *Neural Information Processing Systems*, 2015. 21, 29

[145] S. Yan, Y. Xiong, K. Kundu, S. Yang, S. Deng, M. Wang, W. Xia, and S. Soatto, "Positive-congruent training: Towards regression-free model updates," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 21, 22, 29, 32, 33, 35

[146] D. Angioni, L. Demetrio, M. Pintor, L. Oneto, D. Anguita, B. Biggio, and F. Roli, "Robustness-congruent adversarial training for secure machine learning model updates," *arXiv preprint arXiv:2402.17390*, 2024. 21, 32

[147] M. Srivastava, B. Nushi, E. Kamar, S. Shah, and E. Horvitz, "An empirical analysis of backward compatibility in machine learning systems," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 3272—-3280, Association for Computing Machinery, 2020. 21, 23

[148] D. Ghiani, D. Angioni, A. Sotgiu, M. Pintor, B. Biggio, *et al.*, "Understanding regression in continual learning for malware detection," in *CEUR WORKSHOP PROCEEDINGS*, vol. 3962, CEUR, 2025. 22

[149] Y. Xie, Y.-A. Lai, Y. Xiong, Y. Zhang, and S. Soatto, "Regression bugs are in your model! Measuring, reducing and analyzing regressions in NLP model updates," *arXiv preprint arXiv:2105.03048*, 2021. 22

[150] Y. Zhao, Y. Shen, Y. Xiong, S. Yang, W. Xia, Z. Tu, B. Schiele, and S. Soatto, "ELODI: Ensemble logit difference inhibition for positive-congruent training," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 7529–7541, 2024. 22

[151] R. Schumann, E. Mansimov, Y.-A. Lai, N. Pappas, X. Gao, and Y. Zhang, "Backward compatibility during data updates by weight interpolation," *arXiv preprint arXiv:2301.10546*, 2023. 22

[152] R. Benkert, M. Prabhushankar, and G. AlRegib, "Targeting negative flips in active learning using validation sets," in *2024 IEEE International Conference on Big Data (BigData)*, pp. 820–829, 2024. 22

[153] F. Träuble, J. von Kügelgen, M. Kleindessner, F. Locatello, B. Schölkopf, and P. Gehler, "Backward-compatible prediction updates: A probabilistic approach," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 116–128, Curran Associates, Inc., 2021. 22

[154] R. Matsuno and K. Sakuma, "A robust backward compatibility metric for model retraining," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4190–4194, 2023. 23

[155] A. Tocchetti, L. Corti, A. Balayn, M. Yurrita, P. Lippmann, M. Brambilla, and J. Yang, "AI robustness: a human-centered perspective on technological challenges and opportunities," *ACM Computing Surveys*, vol. 57, no. 6, pp. 1–38, 2025. 23

[156] A. Rezaei, A. Liu, O. Memarrast, and B. D. Ziebart, "Robust fairness under covariate shift," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 9419–9427, 2021. 24

[157] M. Yurochkin, A. Bower, and Y. Sun, "Training individually fair ML models with sensitive subspace robustness," *arXiv preprint arXiv:1907.00020*, 2019. 24

[158] V. Nanda, S. Dooley, S. Singla, S. Feizi, and J. P. Dickerson, "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. 24, 46, 47, 52

[159] D. Franco, L. Oneto, and D. Anguita, "Mitigating robustness bias: Theoretical results and empirical evidences," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2023. 24, 52

[160] H. Chang, T. D. Nguyen, S. K. Murakonda, E. Kazemi, and R. Shokri, "On adversarial bias and the robustness of fair machine learning," *arXiv preprint arXiv:2006.08669*, 2020. 24

[161] P. Benz, C. Zhang, A. Karjauv, and I. S. Kweon, "Robustness may be at odds with fairness: An empirical study on class-wise accuracy," in *NeurIPS 2020 Workshop on pre-registration in machine learning*, pp. 325–342, PMLR, 2021. 24

[162] X. Ma, Z. Wang, and W. Liu, "On the tradeoff between robustness and fairness," *Advances in Neural Information Processing Systems*, vol. 35, pp. 26230–26241, 2022. 24

[163] Y. Pruksachatkun, S. Krishna, J. Dhamala, R. Gupta, and K.-W. Chang, "Does robustness improve fairness? Approaching fairness with word substitution robustness methods for text classification," *arXiv preprint arXiv:2106.10826*, 2021. 24

[164] M. Bertran, N. Martinez, A. Oesterling, and G. Sapiro, "Distributionally robust group backwards compatibility," *arXiv preprint arXiv:2112.10290*, 2021. 24

[165] I. Buselli, A. P. López, E. M. Jiménez, D. Anguita, F. Roli, and L. Oneto, "Mitigating unfair regression in machine learning model updates," in *2024 International Conference on Machine Learning and Applications (ICMLA)*, pp. 1878–1884, IEEE, 2024. 26

[166] D. Franco, L. Oneto, and D. Anguita, "Fair empirical risk minimization revised," in *International Work-Conference on Artificial Neural Networks*, 2023. 27, 47, 51, 58

[167] W. Cunningham, "The WyCash portfolio management system," *ACM Sigplan Oops Messenger*, vol. 4, no. 2, pp. 29–30, 1992. 29

[168] C. Petrozzino, "Who pays for ethical debt in AI?," *AI and Ethics*, vol. 1, no. 3, pp. 205–208, 2021. 29

[169] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge university press, 2004. 30, 38, 40, 55

[170] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014. 31, 32, 37, 40, 47, 49, 55

[171] C. C. Aggarwal, *Neural networks and deep learning: a textbook*. Spinger, 2023. 31, 55

[172] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014. 31

[173] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri, "Are loss functions all the same?," *Neural computation*, vol. 16, no. 5, pp. 1063–1076, 2004. 31, 60

[174] G. Brown and R. Ali, "Bias/variance is not the same as approximation/estimation," *Transactions on Machine Learning Research*, pp. 2835–8856, 2024. 31, 32

[175] L. Oneto, *Model selection and error estimation in a nutshell*. Springer, 2020. 32, 49

[176] C. M. Bishop and H. Bishop, *Deep learning: Foundations and concepts*. Springer, 2024. 37

[177] L. Oneto, S. Ridella, and D. Anguita, "Tikhonov, Ivanov and Morozov regularization for support vector machine learning," *Machine Learning*, vol. 103, pp. 103–136, 2016. 39

[178] S. Pundhir, U. Ghose, and V. Kumari, "LegitANN: Neural network model

with unbiased robustness," in *International Conference on Communication and Artificial Intelligence*, 2021. 46

[179] V. N. Vapnik, *Statistical Learning Theory*. Wiley New York, 1998. 47, 49, 55, 60

[180] S. Verma and J. Rubin, "Fairness definitions explained," in *International Workshop on Software Fairness*, 2018. 50

[181] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *IEEE international conference on data mining workshops*, 2009. 51, 53

[182] M. L. Wick, J. Tristan, and S. Panda, "Unlocking fairness: A trade-off revisited," in *Advances in neural information processing systems*, 2019. 52

[183] J. Chai and X. Wang, "To be robust and to be fair: Aligning fairness with robustness," *arXiv preprint arXiv:2304.00061*, 2023. 54

[184] D. Yin, R. Kannan, and P. L. Bartlett, "Rademacher complexity for adversarially robust generalization," in *International conference on machine learning*, 2019. 55, 56

[185] L. Oneto, S. Ridella, and D. Anguita, "The benefits of adversarial defense in generalization," *Neurocomputing*, vol. 505, pp. 125–141, 2022. 55, 56, 58

[186] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002. 55

[187] F. Dinuzzo and B. Schölkopf, "The representer theorem for hilbert spaces: a necessary and sufficient condition," *Advances in neural information processing systems*, vol. 25, 2012. 68

[188] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016. 70, 71

[189] M. Belkin, "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation," *Acta Numerica*, vol. 30, pp. 203–248, 2021. 70

[190] L. Oneto, S. Ridella, and D. Anguita, "Do we really need a new theory to understand over-parameterization?," *Neurocomputing*, vol. 543, p. 126227, 2023. 70

[191] S. Caton and C. Haas, "Fairness in machine learning: A survey," *ACM Comput. Surv.*, vol. 56, no. 7, 2024.

[192] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and machine learning*. 2019. `http://fairmlbook.org`.

[193] J. D. Angrist and J.-S. Pischke, *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

[194] P. J. Bickel, E. A. Hammel, and J. W. O'Connell, "Sex bias in graduate admissions: Data from Berkeley," *Science*, vol. 187, no. 4175, pp. 398–404, 1975.

[195] J. Hessel and L. Lee, "Cats and captions vs. creators and the clock: Comparing multimodal content to context in predicting relative popularity," in *Proceedings of the 26th international conference on World Wide Web*, pp. 927–936, 2017.

[196] B. Friedman and H. Nissenbaum, "Bias in computer systems," in *ACM Transactions on Information Systems (TOIS)*, vol. 14, pp. 330–347, 1996.

[197] M. J. Salganik, P. S. Dodds, and D. J. Watts, "Experimental study of inequality and unpredictability in an artificial cultural market," *Science*, vol. 311, no. 5762, pp. 854–856, 2006.

[198] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is BERT really robust? A strong baseline for natural language attack on text classification and entailment," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 8018–8025, 2020.

[199] Y. Qian, H. Hu, and T. Tan, "Data augmentation using generative adversarial networks for robust speech recognition," *Speech Communication*, vol. 114, pp. 1–9, 2019.

[200] K.-P. Ning, L. Tao, S. Chen, and S.-J. Huang, "Improving model robustness by adaptively correcting perturbation levels with active queries," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, pp. 9161–9169, 2021.

[201] X. Gao, R. K. Saha, M. R. Prasad, and A. Roychoudhury, "Fuzz testing based data augmentation to improve robustness of deep neural networks," in *Proceedings of the acm/ieee 42nd international conference on software engineering*, pp. 1147–1158, 2020.

[202] Z. Wang and A. Culotta, "Robustness to spurious correlations in text classification via automatically generated counterfactuals," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 14024–14031, 2021.

[203] M. Zhang, S. Levine, and C. Finn, "MEMO: Test time robustness via adaptation and augmentation," *Advances in neural information processing systems*, vol. 35, pp. 38629–38642, 2022.