On Convexity and Bounds of Fairness-aware Classification

Yongkai Wu
University of Arkansas
yw009@uark.edu

Lu Zhang
University of Arkansas
lz006@uark.edu

Xintao Wu
University of Arkansas
xintaowu@uark.edu

ABSTRACT

In this paper, we study the fairness-aware classification problem by formulating it as a constrained optimization problem. Several limitations exist in previous works due to the lack of a theoretical framework for guiding the formulation. We propose a general fairness-aware framework to address previous limitations. Our framework provides: (1) various fairness metrics that can be incorporated into classic classification models as constraints; (2) the convex constrained optimization problem that can be solved efficiently; and (3) the lower and upper bounds of real-world fairness measures that are established using surrogate functions, providing a fairness guarantee for constrained classifiers. Within the framework, we propose a constraint-free criterion under which any learned classifier is guaranteed to be fair in terms of the specified fairness metric. If the constraint-free criterion fails to satisfy, we further develop the method based on the bounds for constructing fair classifiers. The experiments using real-world datasets demonstrate our theoretical results and show the effectiveness of the proposed framework.

CCS CONCEPTS

• **Theory of computation → Convex optimization**; • **Computing methodologies → Supervised learning by classification**; • **Applied computing →** *Law, social and behavioral sciences.*

KEYWORDS

Fairness-aware machine learning; classification;constrained optimization; algorithmic bias

ACM Reference Format:
Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. On Convexity and Bounds of Fairness-aware Classification. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313723

1 INTRODUCTION

Fairness-aware classification is receiving increasing attention in the machine learning fields. Since the classification models seek to maximize the predictive accuracy, individuals may get unwanted digital bias when the models are deployed for making predictions. As fairness becomes a more and more important requirement in machine learning, it is imperative to ensure that the learned classification models can strike a balance between accurate and fair predictions. Previous works on this topic can be mainly categorized into two groups: the in-processing methods which incorporate the fairness constraints into the classic classification models (e.g., [5, 7, 8, 17, 18]), and the pre/post-processing methods which modify the training data and/or derive fair predictions based on the potentially unfair predictions made by the classifier (e.g., [4, 6, 20–22]). In this work, we focus on the in-processing methods.

Very recently, several works have been proposed for formulating the fairness-aware classification as constrained optimization problems [5, 7, 8, 10, 16–18]. Generally, they aim to minimize a loss function subject to certain fairness constraints, e.g., demographic parity (i.e., the difference of the positive predictions between the sensitive group and non-sensitive group) is less than some threshold. However, most quantitative fairness metrics such as demographic parity [13], mistreatment parity [17], etc., are non-convex due to the use of the indicator function, thus making the optimization problem intractable. A widely-used strategy to achieve convexity in optimization is to adopt surrogate functions for both loss function and constraints. In [18], the authors applied the linear surrogate functions to non-convex risk difference as the decision boundary fairness for margin-based classifiers. Similarly in [5], a convex constraint is derived from the risk difference. One challenge is that, when surrogate functions are used to convert non-convex functions to convex functions, estimation errors must exist due to the difference between the surrogate function and the original non-convex function. Thus, achieving the fairness constraints represented by surrogate functions does not necessarily guarantee achieving the real fairness criterion. Hence, how to achieve fairness-aware classification via constrained optimization still remains an open problem.

In this paper, we propose a general framework for fairness-aware classification which addresses the gap incurred by the estimation errors due to the surrogate function. The framework can formulate various commonly-used fairness metrics (risk difference [12], risk ratio [12], equal odds [6], etc.) as convex constraints that are then directly incorporated into classic classification models. Within the framework, we first present a constraint-free criterion (derived from the training data) which ensures that any classifier learned from the data will guarantee to be fair in terms of the specified fairness metric. Thus, when the criterion is satisfied, there is no need to add any fairness constraint into optimization for learning fair classifiers. When the criterion is not satisfied, we need to learn fair classifiers by solving the constrained optimization problems. To connect the surrogated fairness constraints to the original non-convex fairness metric, we further derive the lower and upper bounds of the real fairness measure based on the surrogate function, and develop the refined fairness constraints. This means that, if the refined constraints are satisfied, then it is guaranteed that the real fairness measure is also bounded within the given interval. The bounds work for any surrogate function that is convex and differentiable at zero with the derivative larger than zero. In the experiments, we evaluate our method and compare with existing works using the real-world datasets. The results demonstrate the correctness of the

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313723

constraint-free criterion and the superiority of our method over existing ones in terms of achieving fairness and retaining prediction accuracy.

2 FAIRNESS-AWARE CLASSIFICATION

In this section we present our fairness-aware classification framework. We first introduce the unconstrained optimization formulation for the classic classification models as proposed in [2], and then present our constrained optimization formulation for fairness-aware classification. Throughout the paper, we use the vector $\mathbf{X} \in \mathcal{X}$ to denote the features used in classification, and $Y \in \mathcal{Y} = \{-1, 1\}$ to denote the binary label. We denote the sensitive attribute by S, assuming that it is associated with two values: sensitive group s^- and non-sensitive group s^+. The training data $\mathbb{D} = \{(\mathbf{x}_i, s_i, y_i)\}_{i=1}^N$ is a sample drawn from a unknown but fixed distribution.

2.1 Classification Problem

The learning goal of classification is to find a classifier: $f : \mathcal{X} \mapsto \mathcal{Y}$ that minimizes the average of the classification loss (a.k.a the empirical loss), given by

$$\mathbb{L}(f) = \mathbb{E}_{\mathbf{X},Y}[\mathbb{1}_{f(\mathbf{x})\neq y}], \tag{1}$$

where $\mathbb{1}_{[\cdot]}$ is an indicator function. The classification problem can then be formulated as an optimization problem:

$$\min_{f\in\mathcal{F}} \mathbb{L}(f) = \min_{f\in\mathcal{F}} \mathbb{E}_{\mathbf{X},Y}[\mathbb{1}_{f(\mathbf{x})\neq y}].$$

Directly solving this optimization problem is intractable since the objective function is non-convex [2]. For efficient computation, another predictive function h is adopted which is performed in real number domain $\mathcal{R}$, i.e., $h : \mathcal{X} \mapsto \mathcal{R}$. By letting $f = \text{sign}(h)$, the empirical loss can be reformulated as

$$\begin{aligned}\mathbb{L}(f) = \mathbb{L}(h) &= \mathbb{E}_{\mathbf{X},Y}\Big[\mathbb{1}_{\text{sign}\left(h(\mathbf{x})\right)\neq y}\Big] \\ &= \mathbb{E}_{\mathbf{X}}\Big[Pr(Y=1|\mathbf{x})\mathbb{1}_{h(\mathbf{x})<0} + Pr(Y=-1|\mathbf{x})\mathbb{1}_{h(\mathbf{x})>0}\Big].\end{aligned} \tag{2}$$

If we replace the indicator function (a.k.a 0-1 loss function) with a convex surrogate function ϕ, the empirical loss can be rewritten as

$$\mathbb{L}_\phi(h) = \mathbb{E}_{\mathbf{X}}\Big[Pr(Y=1|\mathbf{x})\phi\big(h(\mathbf{x})\big) + \big(1 - Pr(Y=1|\mathbf{x})\big)\phi\big(-h(\mathbf{x})\big)\Big],$$

which is known as the ϕ-loss, and the optimization problem is reformulated as $\min_{h\in\mathcal{H}} \mathbb{L}_\phi(h)$. In the past decades, a number of surrogate loss functions have been proposed and well studied, such as the hinges loss, the square loss, the logistic loss, the exponential loss, etc..

2.2 Fairness-aware Classification Problem

The fairness-aware classification aims to find a classifier that minimizes the empirical loss while satisfying certain fairness constraints. Several fairness notions or definitions are proposed in the literature, such as demographic parity [13], mistreatment parity [17], etc..

Demographic parity is the most widely-used fairness notion in the fairness-aware learning field. It requires the decision made by the classifier is independent to the sensitive attribute, such as sex or race. Usually, demographic parity is quantified with regard to risk difference [12], i.e., the difference of the positive predictions between the sensitive group and non-sensitive group. For example, in the context of hiring, risk difference can be given by the probability difference of being predicted to be hired between male applicants and female applicants. Using the same language as that in the previous subsection, the risk difference produced by a classifier f is expressed as

$$\mathbb{RD}(f) = \mathbb{E}_{\mathbf{X}|S=s^+}[\mathbb{1}_{f(\mathbf{x})=1}] - \mathbb{E}_{\mathbf{X}|S=s^-}[\mathbb{1}_{f(\mathbf{x})=1}]. \tag{3}$$

As a quantitative metric, we say that classifier f is considered as fair if $|\mathbb{RD}(f)| \leq \tau$, where τ is the user-defined threshold. For instance, the 1975 British legislation for sex discrimination sets $\tau = 0.05$. By directly incorporating the risk difference into the optimization problem, we formulate the fair classification problem as follows.

Problem Formulation 1. *The goal of the fairness-aware classification is to find a classifier f that minimizes the loss $\mathbb{L}(f)$ while satisfying fairness constraint $|\mathbb{RD}(f)| \leq \tau$. It can be approached by solving the following constrained optimization problem*

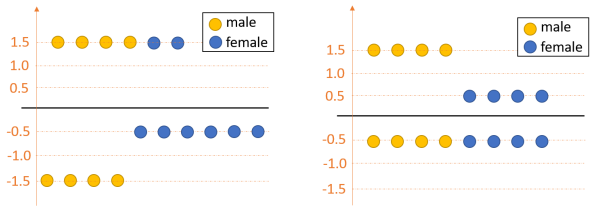
$$\begin{aligned}\min_{f\in\mathcal{F}} \quad & \mathbb{L}(f) \\ \textit{subject to} \quad & \mathbb{RD}(f) \leq \tau, \quad -\mathbb{RD}(f) \leq \tau,\end{aligned} \tag{4}$$

where $\mathbb{L}(f)$ and $\mathbb{RD}(f)$ are defined in Eq. (1) *and Eq.* (3).

Obviously, solving the above problem is computationally intractable, since both $\mathbb{L}(f)$ and $\mathbb{RD}(f)$ contain indicator functions.

The real-value function $h(\mathbf{x})$ and the surrogate functions have been proposed in the recent works [1, 8, 18, 19]. For example, Zafar et al. [19] have proposed the decision boundary covariance to quantify the fairness and serve as constraints, which is equivalent to applying the linear surrogate functions to Problem Formulation 1. They have set the constraint thresholds as c and $-c$, which specify the threshold for the covariance. However, solving the optimization problem with surrogated constraints does not necessarily result in a fair classifier in terms of the original non-convex fairness requirements, e.g., $-\tau \leq \mathbb{RD}(f) \leq \tau$. In fact, there is no any fairness guarantee on the produced classifier. We use an example to show this. Consider two margin-based classifiers where the surrogate functions are linear functions of the distance from the data point to the decision boundary. Therefore, the risk difference is computed by counting the number of data points above and below the decision boundary, and the surrogated risk difference (a.k.a the decision boundary covariance) is computed by measuring the average signed distance from the data points to the decision boundary. In the dataset shown in Figure 1a, we obtain that the surrogated risk difference is 0 but the real risk difference is 0.25. This means that a classier obtained by solving the constrained optimization problem actually can be very unfair. In the dataset shown in Figure 1b, the risk difference is 0 but the surrogated risk difference is 0.5, meaning that some fair classifiers cannot be obtained by solving the optimization problem with surrogated constraints.

The use of the surrogate function inevitably produces estimation errors and leads to the mismatch between the surrogated constraints and the original non-convex fairness constraints. Some intuitive techniques have been introduced to tune the threshold of the surrogated constraints for learning fair classifiers. For example, Zafar et al. [19] have proposed to build an unconstrained classifier and



(a) A classifier that meets the surrogated RD constraint makes unfair predictions.

(b) A classifier that doesn't meet the surrogated RD constraint makes fair predictions.

Figure 1: Two classifiers and their predictions.

consider its risk difference as the initial threshold, say c^*, then they heuristically select a factor $m \in [0, 1]$ and let the threshold $c = m \times c^*$. However, the relationship between the threshold c of the surrogated constraints and the hard threshold τ of the original metrics is unclear hence users have to repeatedly conduct experiments on the datasets.

3 CONVEX FAIRNESS CLASSIFICATION FRAMEWORK

In this section, we propose a general framework for fairness-aware classification which addresses the gap incurred by the estimation error due to the use of the surrogate function. Our framework can formulate various fairness metrics (e.g., risk difference, risk ration, equal odds, etc.) as convex constraints and incorporate them into classic classification model. In the following sections, we present our framework based on the risk difference. In the appendix, we show how our framework can be easily extended to other fairness metrics, e.g., risk ratio, equalized odds.

We first present a constraint-free criterion that is derived from the data. This criterion ensures that any classifier learned from the data are fair in terms of the specified fairness metric. Then when this criterion is satisfied, there is no need to incorporate any fairness constraints for learning fair classification. When this criterion is not met, we formulate the fairness-aware classification task as a convex optimization problem. To fill the gap between the surrogated constraints and the real fairness metrics, we derive the upper and lower bounds for the real fairness metrics and further develop refined convex constraints. If the refined constraints are satisfied, it is guaranteed that the original non-convex fairness requirements are satisfied, e.g., $-\tau \leq \mathbb{RD}(f) \leq \tau$.

3.1 Constraint-free Criterion

We propose a constraint-free criterion to determine whether the fairness constraints are necessary. As discussed in Section 2.1, the unconstrained classification problem is well studied and users can safely apply the classic methods for building a fair classifier.

We first define two special classifiers f_{max} and f_{min} which obtain the maximal and the minimal risk differences respectively.

Definition 1. *The maximal risk difference classifier $f_{\max}$ and the minimal risk difference classifier $f_{\min}$ are defined as:*

$$f_{\max}(\mathbf{x}) = \begin{cases} 1 & \text{if } \eta(\mathbf{x}) \geq p, \\ -1 & \text{otherwise,} \end{cases} \qquad f_{\min}(\mathbf{x}) = \begin{cases} -1 & \text{if } \eta(\mathbf{x}) \geq p, \\ 1 & \text{otherwise,} \end{cases}$$

where we denote $P(S = s^+|\mathbf{x})$ by $\eta(\mathbf{x})$ and $P(S = s^+)$ by p.

These two classifiers provide the maximum and minimum of risk difference among all classifiers f out of the model space $\mathcal{F}$:

Theorem 1. *For any classifier f, it always holds that $\mathbb{RD}^- \leq \mathbb{RD}(f) \leq \mathbb{RD}^+$, where $\mathbb{RD}^- = \mathbb{RD}(f_{\min})$ and $\mathbb{RD}^+ = \mathbb{RD}(f_{\max})$.*

The proof of Theorem 1 is included in the manuscript[1] on the arXiv. From Theorem 1, we directly obtain Corollary 2.

Corollary 2. *Given the threshold τ, for a training data if we have $\mathbb{RD}^+ \leq \tau$ and $\mathbb{RD}^- \geq -\tau$, then any classifier learned from this dataset is fair in terms of risk difference.*

Given a dataset, we can always build two classifiers f_{max} and f_{min}, then compute $\mathbb{RD}^+$ and $\mathbb{RD}^-$. If Corollary 2 is satisfied, users can safely apply any classification models to build classifiers without any fairness concern.

3.2 Convex Fairness-aware Classification

When the constraint-free criterion is not satisfied, it is required to incorporate fairness constraints when learning classifiers, e.g., solving Problem Formulation 1. To this end, we adopt two different surrogate functions for converting the original problem into a convex optimization. We firstly adopt a real-value predictive function h and let $f = \text{sign}(h)$, then rewrite the risk difference as

$$\begin{aligned} &\mathbb{RD}(f) = \mathbb{RD}(h) \\ &= \mathbb{E}_{\mathbf{X}|S=s^+}\left[\mathbb{1}\left[\text{sign}\left(h(\mathbf{x})\right) = 1\right]\right] - \mathbb{E}_{\mathbf{X}|S=s^-}\left[\mathbb{1}\left[\text{sign}\left(h(\mathbf{x})\right) = 1\right]\right] \\ &= \mathbb{E}_{\mathbf{X}|S=s^+}[\mathbb{1}_{h(\mathbf{x})>0}] + \mathbb{E}_{\mathbf{X}|S=s^-}[\mathbb{1}_{h(\mathbf{x})<0}] - 1. \end{aligned}$$

It follows that

$$\begin{aligned} \mathbb{RD}(f) &= \mathbb{E}_{\mathbf{X}}\left[\frac{P(S=s^+|\mathbf{x})}{P(S=s^+)}\mathbb{1}_{h(\mathbf{x})>0} + \frac{P(S=s^-|\mathbf{x})}{P(S=s^-)}\mathbb{1}_{h(\mathbf{x})<0} - 1\right] \\ &= \mathbb{E}_{\mathbf{X}}\left[\frac{\eta(\mathbf{x})}{p}\mathbb{1}_{h(\mathbf{x})>0} + \frac{1-\eta(\mathbf{x})}{1-p}\mathbb{1}_{h(\mathbf{x})<0} - 1\right], \end{aligned} \tag{5}$$

where we denote $P(S = s^+|\mathbf{x})$ by $\eta(\mathbf{x})$ and $P(S = s^+)$ by p for simplicity, thus $P(S = s^-|\mathbf{x}) = 1 - \eta(\mathbf{x})$ and $P(S = s^-) = 1 - p$.

It is intuitive that the indicator function in above formula can be replaced with the surrogate function. The challenge here is, two constraints $\mathbb{RD}(f) \leq \tau$ and $-\mathbb{RD}(f) \leq \tau$ are opposite to each other. Thus, replacing all indicator functions with a single surrogate function will result in a convex-concave problem, where only heuristic solutions for finding the local optima are known to exist. Therefore, we adopt two surrogate functions, a convex one $\kappa(\cdot)$ and a concave one $\delta(\cdot)$, each of which replaces the indicator function for one constraint. As a result, the formulated constrained optimization problem is convex and can be efficiently solved. We call the risk difference represented by $\kappa(\cdot)$ and $\delta(\cdot)$ as the κ, δ-risk difference, denoted by $\mathbb{RD}_\kappa(h)$ and $\mathbb{RD}_\delta(h)$. Almost all commonly-used surrogate functions can be adopted for $\kappa(\cdot)$ and $\delta(\cdot)$, by performing some shift or flip. Curves of some examples for $\kappa(\cdot)$ and $\delta(\cdot)$ are shown in Figure 2.

As a result, we obtain the following convex optimization formulation for learning fair classifiers.

[1] https://arxiv.org/abs/1809.04737

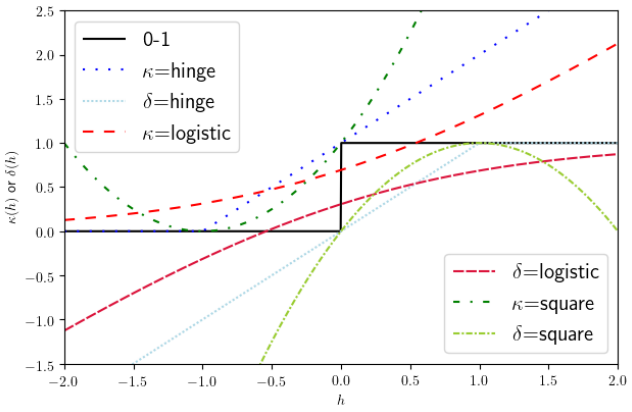


Figure 2: Curves of examples for $\kappa(\cdot)$ and $\delta(\cdot)$.

Problem Formulation 2. *The fairness-aware classification is converted into a convex optimization problem. The optimal solution h^* can be obtained by solving*

$$\begin{aligned} \min_{h\in\mathcal{H}} \quad & \mathbb{L}_\phi(h) \\ \textit{subject to} \quad & \mathbb{RD}_\kappa(h) \le c_1, \quad -\mathbb{RD}_\delta(h) \le c_2, \end{aligned}$$

where $\kappa(\cdot)$ is a convex surrogate function, $\delta(\cdot)$ is a concave surrogate function, c_1, c_2 are the thresholds of the κ, δ-risk difference, and

$$\mathbb{L}_\phi(h) = \mathbb{E}_{\mathbf{X}}\Big[Pr(Y=1|\mathbf{x})\phi\Big(h(\mathbf{x})\Big) + \Big(1-Pr(Y=1|\mathbf{x})\Big)\phi\Big(-h(\mathbf{x})\Big)\Big],$$

$$\mathbb{RD}_\kappa(h) = \mathbb{E}_{\mathbf{X}}\Big[\frac{\eta(\mathbf{x})}{p}\kappa\Big(h(\mathbf{x})\Big) + \frac{1-\eta(\mathbf{x})}{1-p}\kappa\Big(-h(\mathbf{x})\Big) - 1\Big],$$

$$\mathbb{RD}_\delta(h) = \mathbb{E}_{\mathbf{X}}\Big[\frac{\eta(\mathbf{x})}{p}\delta\Big(h(\mathbf{x})\Big) + \frac{1-\eta(\mathbf{x})}{1-p}\delta\Big(-h(\mathbf{x})\Big) - 1\Big].$$

After obtaining h^*, we build the fair classifier by letting $f^* = \text{sign}(h^*)$ and f^* is the final fair classifier. We emphasize that in Problem Formulation 2, the constraint thresholds are rewritten as c_1 and c_2 due to the difference between the surrogated constraints and the original non-convex constraints.

3.3 Refined Fairness-aware Classification

In this section, we develop the upper and lower bounds of the risk difference $\mathbb{RD}(h)$ with the κ, δ-risk difference $\mathbb{RD}_\kappa(h)$ and $\mathbb{RD}_\kappa(h)$. Based on the bounds, we present the method to derive c_1, c_2 for $\mathbb{RD}_\kappa(h), \mathbb{RD}_\delta(h)$, which provides a fairness guarantee that the solution $f^* = \text{sign}(h^*)$ to Problem Formulation 2 satisfies the fairness requirements, e.g., $-\tau \le \mathbb{RD}(f^*) \le \tau$. The method works for various types of surrogate functions (e.g., hinge, square, logistic, exponential, etc.).

We begin with defining the conditional risk difference $C^\eta\Big(h(\mathbf{x})\Big)$ for a specific subpopulation $\mathbf{x}$:

$$C^\eta\Big(h(\mathbf{x})\Big) = \frac{\eta(\mathbf{x})}{p}\mathbb{1}_{h(\mathbf{x})>0} + \frac{1-\eta(\mathbf{x})}{1-p}\mathbb{1}_{h(\mathbf{x})<0} - 1,$$

where η is the abbreviation of $\eta(\mathbf{x})$.

Then, according to Eq. (5), we have $\mathbb{RD}(f) = \mathbb{E}_{\mathbf{X}}[C^\eta\Big(h(\mathbf{x})\Big)]$. When surrogate function $\kappa(\cdot)$ (resp. $\delta(\cdot)$) is adopted, we similarly define the conditional κ-risk difference

$$C^\eta_\kappa\Big(h(\mathbf{x})\Big) = \frac{\eta(\mathbf{x})}{p}\kappa\Big(h(\mathbf{x})\Big) + \frac{1-\eta(\mathbf{x})}{1-p}\kappa\Big(-h(\mathbf{x})\Big) - 1,$$

and we have $\mathbb{RD}_\kappa(h) = \mathbb{E}_{\mathbf{X}}\Big[C^\eta_\kappa\Big(h(\mathbf{x})\Big)\Big]$.

Note that the values of $C^\eta\Big(h(\mathbf{x})\Big)$ and $C^\eta_\kappa\Big(h(\mathbf{x})\Big)$ depend on $\eta(\mathbf{x})$ and $h(\mathbf{x})$, which are determined by the subpopulation of the data specified by $\mathbf{x}$, as well as predictive function h. In order to study the general situations for any specific subpopulation and any possible predictive function, we denote $h(\mathbf{x})$ as α and define the generic conditional risk difference $C^\eta(\alpha)$ and the generic conditional κ-risk difference $C^\eta_\kappa(\alpha)$:

$$C^\eta(\alpha) = \frac{\eta}{p}\mathbb{1}_{\alpha>0} + \frac{1-\eta}{1-p}\mathbb{1}_{\alpha<0} - 1, \quad C^\eta_\kappa(\alpha) = \frac{\eta}{p}\kappa(\alpha) + \frac{1-\eta}{1-p}\kappa(-\alpha) - 1,$$

for any $\eta \in [0,1]$ and $\alpha \in \mathcal{R}$. Then, the minimal conditional risk difference $H^-(\eta)$ and the minimal conditional κ-risk difference $H^-_\kappa(\eta)$ for any arbitrary subpopulation and any possible predictive function are given by

$$H^-(\eta) = \min_{\alpha\in R} C^\eta(\alpha) = \min_{\alpha\in R}\Big[\frac{\eta}{p}\mathbb{1}_{\alpha>0} + \frac{1-\eta}{1-p}\mathbb{1}_{\alpha<0} - 1\Big],$$

$$H^-_\kappa(\eta) = \min_{\alpha\in R} C^\eta_\kappa(\alpha) = \min_{\alpha\in R}\Big[\frac{\eta}{p}\kappa(\alpha) + \frac{1-\eta}{1-p}\kappa(-\alpha) - 1\Big]. \quad (6)$$

It is straightforward that the minimal risk difference $\mathbb{RD}^-$ is equivalent to the expectation of $H^-(\eta(\mathbf{x}))$ since for any possible $\mathbf{x}$, $H^-(\eta(\mathbf{x}))$ provides the minimal conditional risk difference. Similarly, the minimal κ-risk difference achieved by any predictive function (denoted by $\mathbb{RD}^-_\kappa$) is the expectation of $H^-_\kappa(\eta(\mathbf{x}))$, as given by

$$\mathbb{RD}^-_\kappa = \mathbb{E}_{\mathbf{X}}\Big[H^-_\kappa(\eta(\mathbf{x}))\Big].$$

Finally, we define the minimal conditional κ-risk difference within interval α s.t. $\alpha(\eta - p) \ge 0$:

$$H^\circ_\kappa(\eta) = \min_{\alpha:\alpha(\eta-p)\ge 0} C^\eta_\kappa(\alpha). \quad (7)$$

We similarly define $H^+(\eta)$ the maximal conditional risk difference, $H^+_\delta(\eta)$ the maximal conditional δ-risk difference, $\mathbb{RD}^+_\delta$ the maximal δ-risk difference, as well as $H^\circ_\delta(\eta)$ the minimal conditional δ-risk difference within interval α s.t. $\alpha(\eta - p) \ge 0$.

Now, we are able to present our results, which are given in Theorem 3 and Corollary 4. The proof can be found in the manuscript[2] on the arXiv.

Theorem 3. *If $\kappa(\cdot)$ is convex and differentiable at zero with $\kappa'(0) > 0$, $\delta(\cdot)$ is concave and differentiable at zero with $\delta'(0) > 0$, then for any predictive function h, we have*

$$\begin{aligned} \psi_\kappa(\mathbb{RD}(h) - \mathbb{RD}^-) &\le \mathbb{RD}_\kappa(h) - \mathbb{RD}^-_\kappa, \\ \psi_\delta(\mathbb{RD}^+ - \mathbb{RD}(h)) &\le \mathbb{RD}^+_\delta - \mathbb{RD}_\delta(h), \end{aligned} \quad (8)$$

where

$$\psi_\kappa(\mu) = H^\circ_\kappa\Big(p(1-p)\mu + p\Big) - H^-_\kappa\Big(p(1-p)\mu + p\Big),$$

$$\psi_\delta(\mu) = H^+_\delta\Big(p(1-p)\mu + p\Big) - H^\circ_\delta\Big(p(1-p)\mu + p\Big).$$

[2] https://arxiv.org/abs/1809.04737

In Theorem 3, $\psi_\kappa(\mu)$ and $\psi_\delta(\mu)$ are directly derived from the surrogate function κ and δ. Some commonly-used surrogate functions κ, δ and their corresponding ψ_κ, ψ_δ functions are listed in Table 1. The inequalities in Theorem 3 bound the difference between $\mathbb{RD}(h)$ and $\mathbb{RD}^+, \mathbb{RD}^-$ by the differences $\mathbb{RD}_\kappa(h) - \mathbb{RD}_\kappa^-$ and $\mathbb{RD}_\delta^+ - \mathbb{RD}_\delta(h)$. Since $\mathbb{RD}^-, \mathbb{RD}^+, \mathbb{RD}_\kappa^-, \mathbb{RD}_\delta^+$ can be computed from the dataset, we connect the original non-convex constraints and the surrogated convex constraints.

We reformulate Theorem 3 and explicitly give the upper and lower bounds of $\mathbb{RD}(h)$ in Corollary 4.

Corollary 4. *For any predictive function h, let classifier $f = sign(h)$, if $\kappa(\cdot)$ is convex and differentiable at zero with $\kappa'(0) > 0$, $\delta(\cdot)$ is concave and differentiable at zero with $\delta'(0) > 0$, then risk difference $\mathbb{RD}(f)$ is bounded by following inequalities:*

$$\mathbb{RD}(f) \le \mathbb{RD}^- + {\psi_\kappa}^{-1}\Big(\mathbb{RD}_\kappa(h) - \mathbb{RD}_\kappa^-\Big),$$
$$\mathbb{RD}(f) \ge \mathbb{RD}^+ - {\psi_\delta}^{-1}\Big(\mathbb{RD}_\delta^+ - \mathbb{RD}_\delta(h)\Big).$$

Based on the upper and lower bounds of $\mathbb{RD}(f)$, we can derive the thresholds c_1, c_2 for the surrogated constraints in Problem Formulation 2. For example, if we aim to obtain a classifier f such that $-\tau \le \mathbb{RD}(f) \le \tau$, we only require the upper bound of $\mathbb{RD}(f)$ is smaller than τ and the lower bound is larger than $-\tau$. That is:

$$\mathbb{RD}^- + {\psi_\kappa}^{-1}\Big(\mathbb{RD}_\kappa(h) - \mathbb{RD}_\kappa^-\Big) \le \tau,$$
$$\mathbb{RD}^+ - {\psi_\delta}^{-1}\Big(\mathbb{RD}_\delta^+ - \mathbb{RD}_\delta(h) \ge -\tau.$$

Thus, we obtain the refined constraints and if the refined constraints are satisfied, the original risk difference requirements are guaranteed to be satisfied.

We modify Problem Formulation 2 to obtain Problem Formulation 3 with refined fairness constraints which guarantee the real non-convex fairness requirement.

Problem Formulation 3. *A classifier $f^* = sign(h^*)$ that achieves fairness guarantee $-\tau \le \mathbb{RD}(f) \le \tau$ can be obtained by solving the following constrained optimization*

$$\min_{h \in \mathcal{H}} \quad \mathbb{L}_\phi(h) \tag{9}$$
$$\textit{subject to} \quad \mathbb{RD}_\kappa(h) \le \psi_\kappa(\tau - \mathbb{RD}^-) + \mathbb{RD}_\kappa^-,$$
$$-\mathbb{RD}_\delta(h) \le \psi_\delta(-\tau + \mathbb{RD}^+) + \mathbb{RD}_\delta^+.$$

Note that the right-hand sides of above two inequalities are constants for a given dataset. Therefore, the constrained optimization problem is still convex. We can optimally solve this problem and the solution $f^* = sign(h^*)$ is guaranteed to satisfy $-\tau \le \mathbb{RD}(f^*) \le \tau$.

Table 1: Some common surrogate functions for κ-δ and the corresponding $\psi_\kappa(\mu)$ and $\psi_\delta(\mu)$.

Name of κ-δ	$\kappa(\alpha)$ for $\alpha \in \mathbf{R}$	$\delta(\alpha)$ for $\alpha \in \mathbf{R}$	$\psi_\kappa(\mu)$ or $\psi_\delta(\mu)$ for $\mu \in (0, 1/p]$
Hinge	$\max\{\alpha+1, 0\}$	$\min\{\alpha, 1\}$	μ
Square	$(\alpha+1)^2$	$1-(1-\alpha)^2$	μ^2
Exponential	$\exp(\alpha)$	$1-\exp(-\alpha)$	$(\sqrt{(1-p)\mu+1} - \sqrt{1-p\mu})^2$

4 EXPERIMENTS

4.1 Experimental Setup

Dataset. In the experiments we use two datasets: Adult and Dutch. The Adult dataset [9] contains a total of 48,842 instances, each of which is characterized by 14 attributes (e.g., `sex`, `age`, `work_class`, `education`, `income`, etc.). We consider `sex` as the sensitive attribute with two values, `male` and `female`. Then, we consider `income` as the class label. The Dutch dataset [23] contains a total of 60,420 instances, each of which is characterized by 12 attributes. Similarly, we use `sex` as the sensitive attribute, and `occupation` as the class label.

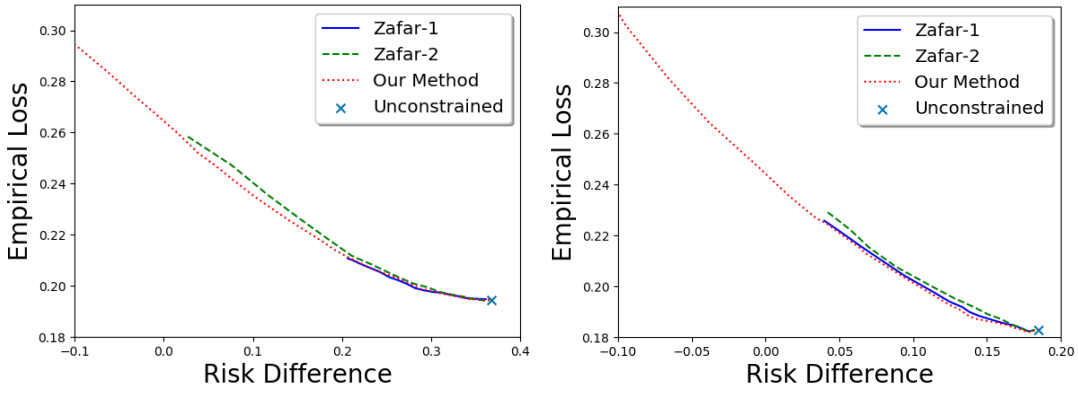
Baseline. We compare our method with two related works, referred to as Zafar-1 [18] and Zafar-2 [17], both of which formulate the fairness-aware classification problem as constrained optimization problems. In [18], the authors quantify fairness using the covariance between the users' sensitive attribute and the signed distance from the feature vectors to the decision boundary. The fairness constraint is formulated as covariance $\le m \times c^*$, where c^* is the measured fairness of the unconstrained optimal classifier and m is a multiplication factor $\in [0, 1]$. In [17], the fairness is quantified similarly with the distance function being replaced with a convex non-linear function. As a result, the obtained problem is a convex-concave optimization problem. In the experiments, we adopt the Disciplined Convex-Concave Programming (DCCP) [15] as proposed in [17] for solving the convex-concave optimization problem. For our method and Zafar-1, the convex optimization problem is solved using CVXPY [3]. The datasets and implementation are available at http://tiny.cc/fair-classification.

4.2 Constraint-free Criterion of Ensuring Fairness

To demonstrate the sufficiency criterion of learning fair classifiers, we build the maximal/minimal risk difference classifiers $f_{\min}, f_{\max}$ for both Adult and Dutch datasets, and measure the risk differences they produce, i.e., $\mathbb{RD}^-$ and $\mathbb{RD}^+$. The results are shown in the first two rows in Table 2. As can be seen, in both datasets we have large maximal and minimal risk differences. In order to evaluate a situation with small a risk difference, we also create a variant of Adult, referred to as Adult*, where all attributes are binarized and the sensitive attribute `sex` is shuffled to incur a small risk difference. Then, we build a number of classifiers including Linear Regression (LR), Support Vector Machine (SVM) with linear kernel, Decision Tree (DT), and Naive Bayes (NB), using the three datasets as the training data with with 5-fold cross-validation. After that, their risk differences are quantified on the testing data, as shown in the last four rows in Table 2. We can see that all values are within $\mathbb{RD}^-, \mathbb{RD}^+$ which are consistent with our constraint-free criterion.

4.3 Learning Fair Classifiers

We build our fair classifiers on both Adult and Dutch datasets by solving the optimization problem defined in Problem Formulation 2. For surrogate functions, we use the logistic function for $\phi(\cdot)$, and the hinge function for $\kappa(\cdot)$ and $\delta(\cdot)$. We also compare our methods with Zafar-1 and Zafar-2. The results are shown in Figure 3, which



(a) Adult **(b) Dutch**

Figure 3: Comparison of fair classifiers.

Table 2: $\mathbb{RD}^+$, $\mathbb{RD}^-$ and risk differences of Linear Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and Naive Bayes (NB).

$\mathbb{RD}(\cdot)$	Adult	Dutch	Adult*
$\mathbb{RD}^+$	0.967	0.516	0.046
$\mathbb{RD}^-$	-0.967	-0.516	-0.046
LR	0.371	0.185	0.000
SVM	0.434	0.156	0.001
DT	0.316	0.184	0.001
NB	0.447	0.144	0.001

depict the relationship between the obtained risk difference and empirical loss. For our method, different risk differences are obtained by adjusting relax terms c_1 and c_2, while for Zafar-1 and Zafar-2 different risk differences are obtained by adjusting the multiplication factor m. As can be seen, our method can achieve much smaller risk difference than Zafar-1 and Zafar-2. This may be because Zafar-1 linear functions to formulate the fairness constraints, which may incur large estimation errors; while Zafar-2 formulates a convex-concave optimization problem, where only the local optima can be reached. For the same reason, we can observe that our method produces better empirical loss than Zafar-2 given any same risk difference.

5 CONCLUSIONS

In this paper, we studied the fairness-aware classification problem and formulated it as the constrained optimization problem. We proposed a general framework which addresses all limitations of previous works in terms of: (1) various fairness metrics can be incorporated into classic classification models as constraints; (2) the formulated constrained optimization problem is convex and can be solved efficiently; and (3) the lower and upper bounds of real fairness measures are established using surrogate functions, which provide a fairness guarantee for our framework. Within the framework, we proposed a constraint-free criterion under which the learned classifier is guaranteed to be fair in terms of the specified fairness metric, as well as developed the method for learning fair classifiers if the constraint-free criterion fails to satisfy. The results demonstrate the correctness of the constraint-free criterion and the superiority of our method over existing ones in terms of achieving fairness and retaining prediction accuracy.

ACKNOWLEDGMENTS

This work was supported in part by NSF 1646654.

A OTHER FAIRNESS NOTIONS

Risk ratio is a common fairness notion [11, 14]. It also requires the decision is independent with the protected attribute. Different with the risk difference, the unfairness is quantified by the ratio of the positive decisions between the non-protected group and the protected group. Let's formalize the risk ratio $\mathbb{RR}(h)$ of classifier h:

$$\mathbb{RR}(h) = \frac{\mathbb{E}_{\mathbf{X}|S=s^+}\left[\mathbb{1}_{h(\mathbf{x})>0}\right]}{\mathbb{E}_{\mathbf{X}|S=s^-}\left[\mathbb{1}_{h(\mathbf{x})>0}\right]}.$$

The fairness constraints with regards to risk ratio could be expressed as

$$\mathbb{RR}(h) = \frac{\mathbb{E}_{\mathbf{X}|S=s^+}\left[\mathbb{1}_{h(\mathbf{x})>0}\right]}{\mathbb{E}_{\mathbf{X}|S=s^-}\left[\mathbb{1}_{h(\mathbf{x})>0}\right]} \leq \tau.$$

Similar to Eq. (5), we express the constraints as

$$\mathbb{E}_{\mathbf{X}}\left[\frac{\eta}{p}\mathbb{1}_{h(\mathbf{x})>0} + \tau\frac{1-\eta(\mathbf{x})}{1-p}\mathbb{1}_{h(\mathbf{x})>0}\right] - \tau \leq 0. \tag{10}$$

Equalized odds and equalized opportunity are proposed by Hardt et al. [6]. Equalized odds requires the protected attribute and the predicted label are independent conditional on the truth label. To quantify the strength of equalized odds, we simply propose the prediction difference between two groups conditional on the truth label. So, the equalized odds is

$$\mathbb{EO}(h) = \mathbb{E}_{\mathbf{X}|S=s^+,Y}[\mathbb{1}_{h(\mathbf{x})>0}] - \mathbb{E}_{\mathbf{X}|S=s^-,Y}[\mathbb{1}_{h(\mathbf{x})>0}].$$

Similarly, a classifier h is considered as fair with regard to equalized odds if $\mathbb{EO}(h) \leq \tau$.

Let's reformulate the equalized odds constraints:

$$\begin{aligned}\mathbb{EO}(h) &= \mathbb{E}_{\mathbf{X}|S=s^+,Y}[\mathbb{1}_{h(\mathbf{x})>0}] + \mathbb{E}_{\mathbf{X}|S=s^-,Y}[\mathbb{1}_{h(\mathbf{x})<0}] - 1 \\ &= \mathbb{E}_{\mathbf{X}|Y}\left[\frac{P(S=s^+|\mathbf{x},y)}{P(S=s^+|y)}\mathbb{1}_{h(\mathbf{x})>0} + \frac{1-P(S=s^+|\mathbf{x},y)}{1-P(S=s^+|y)}\mathbb{1}_{h(\mathbf{x})<0}\right] - 1 \leq \tau.\end{aligned} \tag{11}$$

Equalized opportunity is a relaxation of equalized odds where only the positive group ($Y = 1$) is taken into account:

$$\mathbb{EOP}(h) = \mathbb{E}_{\mathbf{X}|Y=1}\Big[\frac{P(S = s^+|\mathbf{x}, Y = 1)}{P(S = s^+|Y = 1)}\mathbb{1}_{h(\mathbf{x})>0} + \frac{1 - P(S = s^+|\mathbf{x}, Y = 1)}{1 - P(S = s^+|Y = 1)}\mathbb{1}_{h(\mathbf{x})<0}\Big] - 1 \leq \tau. \quad (12)$$

By simply replacing the indicator functions with surrogate functions, we can readily extend our framework to the constraints (10), (11), (12) with regard to the three notions. Our criterion and bounds are also extensible to the three notions.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, and John Langford. 2017. A Reductions Approach to Fair Classification. In *Conference on Fairness, Accountability, and Transparency in Machine Learning*.

[2] Peter L Bartlett, Michael I Jordan, and Jon D. McAuliffe. 2006. Convexity, Classification, and Risk Bounds. *J. Amer. Statist. Assoc.* 101, 473 (mar 2006), 138–156. https://doi.org/10.1198/016214505000000907

[3] Steven Diamond and Stephen Boyd. 2016. CVXPY: A Python-Embedded Modeling Language for Convex Optimization. *Journal of Machine Learning Research* 17, 83 (2016), 1–5.

[4] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *KDD '15*. ACM Press. https://doi.org/10.1145/2783258.2783311

[5] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. 2016. Satisfying real-world goals with dataset constraints. In *NIPS'16*. 2415–2423.

[6] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[7] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware Learning through Regularization Approach. In *ICDMW'11*. IEEE, 643–650. https://doi.org/10.1109/ICDMW.2011.83

[8] Aditya Krishna and Robert C Williamson. 2018. The Cost of Fairness in Binary Classification. *Proceedings of Machine Learning Research* 81 (2018), 1–12. http://proceedings.mlr.press/v81/menon18a.html

[9] M Lichman. 2013. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[10] Mahbod Olfat and Anil Aswani. 2018. Spectral Algorithms for Computing Fair Support Vector Machines. In *International Conference on Artificial Intelligence and Statistics*. 1933–1942.

[11] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2009. Measuring Discrimination in Socially-Sensitive Decision Records. In *Proceedings of the 2009 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 581–592. https://doi.org/10.1137/1.9781611972795.50

[12] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2012. A study of top-k measures for discrimination discovery. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing - SAC '12*. ACM Press, New York, New York, USA, 126. https://doi.org/10.1145/2245276.2245303

[13] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *KDD 08*. ACM Press, New York, New York, USA, 560. https://doi.org/10.1145/1401890.1401959

[14] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29, 05 (nov 2014), 582–638. https://doi.org/10.1017/S0269888913000039

[15] Xinyue Shen, Steven Diamond, Yuantao Gu, and Stephen Boyd. 2016. Disciplined Convex-Concave Programming. Cdc (2016), 1009–1014. https://doi.org/10.1109/CDC.2016.7798400

[16] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning Non-Discriminatory Predictors. In *Conference on Learning Theory*. 1920–1953.

[17] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact. In *WWW '17*. ACM Press, New York, NY, USA, 1171–1180. https://doi.org/10.1145/3038912.3052660

[18] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Artificial Intelligence and Statistics*. Fort Lauderdale, Florida, USA.

[19] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P Gummadi, and Adrian Weller. 2017. From Parity to Preference-based Notions of Fairness in Classification. (jun 2017).

[20] Lu Zhang and Xintao Wu. 2017. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics* 4, 1 (aug 2017), 1–16. https://doi.org/10.1007/s41060-017-0058-x

[21] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *IJCAI '17*. California, 3929–3935. https://doi.org/10.24963/ijcai.2017/549

[22] Lu Zhang, Yongkai Wu, and Xintao Wu. 2018. Achieving Non-Discrimination in Prediction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. 3097–3103. https://doi.org/10.24963/ijcai.2018/430

[23] Indre Zliobaite, Faisal Kamiran, and Toon Calders. 2011. Handling Conditional Discrimination. In *2011 IEEE 11th International Conference on Data Mining*. IEEE, 992–1001. https://doi.org/10.1109/ICDM.2011.72