General Fair Empirical Risk Minimization

Luca Oneto
University of Genoa
luca.oneto@unige.it

Michele Donini
Amazon Web Services
donini@amazon.com

Massimiliano Pontil
Istituto Italiano di Teconologia
University College London
massimiliano.pontil@iit.it

***Abstract*—We tackle the problem of algorithmic fairness, where the goal is to avoid the unfairly influence of sensitive information, in the general context of regression with possible continuous sensitive attributes. We extend the framework of fair empirical risk minimization of [1] to this general scenario, covering in this way the whole standard supervised learning setting. Our generalized fairness measure reduces to well known notions of fairness available in literature. We derive learning guarantees for our method, that imply in particular its statistical consistency, both in terms of the risk and the fairness measure. We then specialize our approach to kernel methods and propose a convex fair estimator in that setting. We test the estimator on a commonly used benchmark dataset (Communities and Crime) and on a new dataset collected at the University of Genoa[1], containing the information of the academic career of five thousand students. The latter dataset provides a challenging real case scenario of unfair behaviour of standard regression methods that benefits from our methodology. The experimental results show that our estimator is effective at mitigating the trade-off between accuracy and fairness requirements.**

***Index Terms*—Machine Learning, Algorithmic Fairness, Regression, Kernel Methods**

I. Introduction

The problem of designing learning methods that do not use sensitive information in a discriminatory way (e.g. knowledge about the ethnic group of an individual, sex, age) is receiving increasing attention, due to its fundamental importance in real-life scenarios, see e.g. [2]–[21] and references therein. In this paper we follow a recent line of work [1], [7], [9]–[13], [22]–[26] in which the fairness constraint is directly taken into account during the learning procedure. An important departure from previous work that we take in this paper is to consider the possibility that the sensitive feature and/or the output (response variable) we wish to predict take real values.

The importance of being able to solve regression tasks and possibly dealing with continuous sensitive features can be highlighted by the following example. At the University of `Anonymous`, automatic systems are needed to predict students' performance for the purpose of improving the teaching quality and the students' support systems. In this case, the response variable is the course mark and the sensitive features can be both categorical (e.g. sex or ethnic group) or continuous (e.g. age or financial status).

Common notions of fairness that have been used in the setting of classification with categorical sensitive features is that of Equal Opportunity or Equalized Edds [4]. They aim to balance decisions of a classifier among the different sensitive groups and label sets. We show how these notions can be extended to the general supervised learning setting (regression and classification) with general sensitive features (categorical and continuous). We observe that these novel fairness constraints can be incorporated within the Empirical Risk Minimization (ERM) framework. Our method and analysis build up and extend the Fair ERM (FERM) framework developed in [1]. As the fairness measures used here are more general than those employed in that work, we name our approach General FERM (G-FERM). We show that G-FERM is supported by consistency guarantees both in terms of risk and fairness measure. Specifically, we derive both risk and fairness bounds, which support the statistically consistency of G-FERM. We give a concrete instance of G-FERM in the setting of kernel methods, leading to a form of constrained regularized empirical risk minimization, in which the fairness constraint is obtained by composting the ℓ_1 norm with a linear transformation.

Contributions. First, we present new generalized notions of fairness that encompass well studied notions used for classification and regression with categorical and numerical sensitive feature. Second, we study statistical bounds for G-FERM that imply consistency properties both in terms of fairness measure and risk of the selected model. As a third contribution, we instantiate G-FERM in the setting of kernel methods, leading to an efficient convex estimator. We test this estimator on a commonly used benchmark dataset (Communities and Crime) and on a new dataset collected at University of `Anonymous`, containing the information of the academic career of five thousand students. The latter dataset provides a challenging real case scenario of unfair behaviour of standard methods for regression that is solvable by using our methodology. The experimental results show that our estimator is effective at mitigating the trade-off between accuracy and fairness requirements.

Paper Organization. In Section II we discuss previous work on fairness, with a particular focus on regression and or continuous sensitive features. In Section III we introduce our notion of fairness which leads us to the G-FERM and study its statistical properties. In Section IV we give the kernel-based G-FERM estimator and in Section V report on numerical experiments on two real datasets. Finally in Section VI we draws conclusions and comment on future research directions.

[1]The data and the research are related to the project DROP@UNIGE of the University of Genoa.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

II. Related works

In the context of fairness, most of the papers in literature address the problem of binary classification task with categorical (or even binary) sensitive features [4], [7]; a broad review on classification with categorical sensitive feature is provided in [1]. This task is indeed very important, because it is strictly related to the possibility of having access to specific benefits (e.g. loans) without being discriminated due to gender or ethnic characteristics. On the other hand, the set of problems solvable by using these methods is limited and not comprehensive of all the real-world case scenarios.

Focusing on the works able to handle regression tasks, we can divide them by the type of problems they are able to solve and the notion of fairness they exploit. As we will see, with very few exceptions – e.g. [27] – most of the methods in literature are not able to deal with both classification and regression task and with both numerical and categorical sensitive features with an unified approach supported by theoretical consistency results. In fact, they introduce task oriented notions of fairness and/or do address the statistical consistency of their method with respect to the risk and the fairness measure employed.

The largest family of methods tackle regression problems with (single) categorical or binary sensitive feature [13], [28]–[30]. For example, in [13], a convex approach for regression is proposed, where the authors use a specific definition of fairness in order to have models which treat similar examples in a similar way, in the sense of the predicted outcome. The authors tackle the problem by introducing a new convex regularizer and by imposing this notion on different regression tasks. Another example is [29], where the authors use an adapted version of Demographic Parity [31] for classification, in the context of regression.

Reducing the regression problem to have only categorical sensitive features is a serious limitation. In this sense, few interesting papers present regression methods able to deal with continuous sensitive attributes [12], [27], [32]. Differently to our approach, the authors impose other definitions of fairness (e.g. Disparate Impact [7] or even ad-hoc brand new definitions). Moreover, it is important to note that these methods do not naturally extend to the case of not-continuous sensitive attributes.

Considering a larger spectrum of possible methodologies, it is possible to find in literature other methods able to solve regression tasks by imposing some concept of fairness. [33] and [34] tackle the regression problem exploiting the causal machine learning framework. These methods can handle potentially both continuous and categorical sensitive features. The authors' analysis considers only the case of categorical ones, leaving the evolution to continuous sensitive attributes as possible future works. Another interesting idea, presented in [35], is to study the fairness as a property of the metric of the feature space. The authors introduce a new definition of metric-related fairness allowing them to solve a regression problem with categorical and continuous sensitive attributes. Finally, learning fair pre-processing rules is another possible way to obtain a regression model that is fair. In fact, for example in [17], the fair representation of the data can be used in synergy with any classic regression method, in order to generate a fair regression model.

III. Learning with Fairness Constraints

In this section, we introduce our framework for learning under fairness constraints. We first recall some notation used throughout this work in Section III-A. We then present the proposed fairness measures in Section III-B, which lead us to consider in Section III-C a generalized version of the FERM approach [1]. Finally in Section III-D we discuss the statistical properties of our method.

A. Setting

Let $\mathcal{D}=\{(\boldsymbol{x}_1, s_1, y_1),\ \ldots,\ (\boldsymbol{x}_n, s_n, y_n)\}$ be a training set formed by n samples drawn independently from an unknown probability distribution μ over $\mathcal{X}\times\mathcal{S}\times\mathcal{Y}$, where $\mathcal{X}$ is the input space, $\mathcal{S}$ is the space of the sensitive attribute and $\mathcal{Y}$ is the output space. Both $\mathcal{S}$ and $\mathcal{Y}$ may be finite or continuous; if $\mathcal{Y}$ is a finite set of labels we are dealing with the classification setting and if $\mathcal{Y}\subseteq\mathbb{R}$ we are dealing with the regression setting.

Let K and Q be positive integers and define the sets

$$\mathcal{Y}_K=\{t_1,\cdots,t_{K+1}\}\subset\mathbb{R},\quad \mathcal{S}_Q=\{\sigma_1,\cdots,\sigma_{Q+1}\}\subset\mathbb{R},$$

where $t_1<t_2<\cdots<t_{K+1}$, and $\sigma_1<\sigma_2<\cdots<\sigma_{Q+1}$. The sets $\mathcal{Y}_K$ and $\mathcal{S}_Q$ are prescribed by the user: the discretization process is driven by the application at hand and points in the same interval are regarded as indistinguishable. For example, it does not make sense to state that a group of students at the University of `Anonymous` is mistreated because the average grades are distant by less than 5% of the mark range. We also define, for every $1\leq k\leq K$ and $1\leq q\leq Q$, the subsets of training points

$$\mathcal{D}_{k,q}=\{(\boldsymbol{x}_i, s_i, y_i) : 1\leq i\leq n, y\in[t_k, t_{k+1}), s\in[\sigma_q, \sigma_{q+1})\}$$

and let $n_{k,q}=|\mathcal{D}_{k,q}|$.

We consider a function (or model) f chosen from a set $\mathcal{F}$ of possible ones. The functional form of the model may explicitly depend on the sensitive feature (i.e. $f:\mathcal{X}\times\mathcal{S}\to\mathbb{R}$) or not (i.e. $f:\mathcal{X}\to\mathbb{R}$) based on specific legal requirements in the application at hand [26], [36]. For this reason we will indicate $f:\mathcal{Z}\to\mathbb{R}$ where $\mathcal{Z}$ may contain the sensitive feature (i.e. $\mathcal{Z}=\mathcal{X}\times\mathcal{S}$) or not (i.e. $\mathcal{Z}=\mathcal{X}$). The error (risk) of f is measured by a prescribed loss function $\ell:\mathbb{R}\times\mathcal{Y}\to\mathbb{R}$. The risk of a model $L(f)$, together with its empirical counterpart $\hat{L}(f)$, are defined respectively as

$$L(f)=\mathbb{E}\left[\ell(f(\boldsymbol{z}), y)\right],$$

and

$$\hat{L}(f)=\tfrac{1}{n}\textstyle\sum_{(\boldsymbol{z},y)\in\mathcal{D}}\ell(f(\boldsymbol{z}), y).$$

When necessary we will indicate with a subscript the particular loss function used and the associated risk, i.e. $L_p(f) = \mathbb{E}\left[\ell_p(f(\boldsymbol{z}), y)\right]$.

The purpose of a learning procedure is to find a model that minimizes the risk. Since the probability measure μ is usually unknown, the risk cannot be computed, however we can compute the empirical risk and a natural learning strategy, called Empirical Risk Minimization (ERM), is then to minimize the empirical risk within a prescribed set of functions, see e.g. [37].

B. ϵ-Loss General Fair

In the literature different definitions of fairness of a classifier or real-valued function exist as described in Section II. It is important to stress that there is not yet a consensus about which definition should be employed to evaluate algorithmic fairness. Moreover, most of the current fairness definitions are not able to deal with regression problems (or with continuous sensitive attributes), losing their meaning or being even not definable. In this work we proposes a general notion of fairness able to deal with both classification and regression and with both categorical and numerical sensitive features and which generalizes previously known notions of fairness.

Definition 1: A model f is ϵ-general fair (ϵ-GF) with $\epsilon \in [0,1]$ if satisfies the following condition

$$\tfrac{1}{KQ^2}\sum_{k=1}^{K}\sum_{p,q=1}^{Q}\left|P^{k,p}(f)-P^{k,q}(f)\right|\leq\epsilon$$

where, for every $1\leq k\leq K$ and $1\leq q\leq Q$, we have defined the conditional probabilities

$$P^{k,q}(f)=\mathbb{P}\Big\{f(\boldsymbol{z})\in[t_k,t_{k+1})\Big|y\in[t_k,t_{k+1}),s\in[\sigma_q,\sigma_{q+1})\Big\}.$$

This definition says that a model is fair if its predictions are equally distributed independently of the value of the sensitive attribute. It can be further generalized as follows.

Definition 2: For every $1\leq k\leq K$ let ℓ_k be a loss function. For every $1\leq k\leq K,\ 1\leq q\leq Q$, define the conditional risks

$$L_k^{k,q}(f)=\mathbb{E}\big[\ell_k(f(\boldsymbol{z}),y)|y\in[t_k,t_{k+1}),s\in[\sigma_q,\sigma_{q+1})\big].$$

We say that a function f is ϵ-loss general fair (ϵ-LGF) with $\epsilon\in[0,1]$ if it satisfies the following condition

$$\tfrac{1}{KQ^2}\sum_{k=1}^{K}\sum_{p,q=1}^{Q}\left|L_k^{k,p}(f)-L_k^{k,q}(f)\right|\leq\epsilon.$$

This definition says that a model is fair if its errors, relative to the loss function, are approximately equally distributed independently of the value of the sensitive attribute. Definition 2 includes Definition 1 when we choose $\ell_k(\hat{y},y)=\mathbb{1}\{\hat{y}\notin[t_k,t_{k+1})\}$, for $1\leq k\leq K$. Moreover, it is possible to link Definition 2 to other fairness measures used before in the literature.

Remark 1: If we choose $\epsilon=0$, $\mathcal{Y}=\{-1,+1\}$, $\mathcal{S}=\{0,1\}$, $\mathcal{Y}_K=\{-1.5,\ 0,\ +1.5\}$, $\mathcal{S}_Q=\{-0.5,\ 0.5,\ 1.5\}$ and, for every $1\leq k\leq K$, let ℓ_k be the 0-1-loss, that is $\ell_k(y,y)=\mathbb{1}\{y\hat{y}\leq 0\}$, then Definition 2 reduces to the notion of Equalized Odds [1], [4]. On the other hand, in the same setting, if we let, for every k, ℓ_k be the linear loss, $\ell_k(\hat{y},y)=(1-y\hat{y})/2$, then we recover other notions of fairness introduced in [26]. When $\epsilon=0$, $\mathcal{Y}\subseteq\mathbb{R}$, $\mathcal{S}=\{0,1\}$, $\mathcal{Y}_K=\{-\infty,\infty\}$, $\mathcal{S}_Q=\{-0.5,0.5,1.5\}$ then Definition 2 reduces to the notion of Mean Distance introduced in [28] and also exploited in [27]. Finally, in the same setting, if $\mathcal{S}\subseteq\mathbb{R}$ in [27] it is proposed to use the correlation coefficient which is equivalent to setting $\mathcal{S}_Q=\mathcal{S}$ in Definition 2.

C. General Fair Empirical Risk Minimization

In this paper, we aim at minimizing the risk subject to a fairness constraint. Specifically, we consider the problem

$$\min_{f\in\mathcal{F}}\left\{L(f):\sum_{k=1}^{K}\sum_{p,q=1}^{Q}\left|L_k^{k,p}(f)-L_k^{k,q}(f)\right|\leq\epsilon\right\},\qquad(1)$$

where $\epsilon\in[0,1]$ is the amount of unfairness that we are willing to bear. Since the measure μ is unknown we replace the deterministic quantities with their empirical counterparts. That is, we replace Problem (1) with

$$\min_{f\in\mathcal{F}}\left\{\hat{L}(f):\sum_{k=1}^{K}\sum_{p,q=1}^{Q}\left|\hat{L}_k^{k,p}(f)-\hat{L}_k^{k,q}(f)\right|\leq\hat{\epsilon}\right\},\qquad(2)$$

where $\hat{\epsilon}\in[0,1]$, and, for every $k\in\{1,\cdots,K\}$ and every $q\in\{1,\cdots,Q\}$ we defined the empirical conditional risks

$$\hat{L}_k^{k,q}(f)=\tfrac{1}{n_{k,q}}\textstyle\sum_{(\boldsymbol{z},y)\in\mathcal{D}_{k,q}}\ell_k(f(\boldsymbol{z}),y).$$

We will refer to Problem (2) as G-FERM since it generalizes the FERM approach introduced in [1].

D. Statistical Analysis

Let f^* be a solution of Problem (1), and let $\hat{f}$ a solution of Problem (2). In this section we will show that these solutions are linked one to another. In particular, if the parameter $\hat{\epsilon}$ is chosen appropriately, we will show that, in a certain sense, the estimator $\hat{f}$ is consistent. Our analysis extends the reasoning in [1] to the more general setting presented here.

For this purpose, we require that for any data distribution, it holds with probability at least $1-\delta$ with respect to the draw of a dataset that

$$\sup_{f\in\mathcal{F}}\left|L(f)-\hat{L}(f)\right|\leq B(\delta,n,\mathcal{F})\qquad(3)$$

where $B(\delta,n,\mathcal{F})$ goes to zero as n grows to infinity, that is the class $\mathcal{F}$ is learnable with respect to the loss [37]. Moreover $B(\delta,n,\mathcal{F})$ is usually an exponential bound which means that $B(\delta,n,\mathcal{F})$ grows logarithmically with respect to the inverse of δ.

Remark 2: If $\mathcal{F}$ is a compact subset of linear separators in a reproducing kernel Hilbert space, and the loss is Lipschitz in its first argument, then $B(\delta,n,\mathcal{F})$ can be obtained via Rademacher bounds [38]. In this case $B(\delta,n,\mathcal{F})$ goes to zero at least as $\sqrt{1/n}$ as n grows and decreases with δ as $\sqrt{\ln(1/\delta)}$.

We are now ready to state the first result of this section.

Theorem 1: Let $\mathcal{F}$ be a learnable set of functions with respect to the loss function $\ell:\mathbb{R}\times\mathcal{Y}\to\mathbb{R}$, let f^* be a solution of Problem (1) and let $\hat{f}$ be a solution of Problem (2) with

$$\hat{\epsilon}=\epsilon+\textstyle\sum_{k=1}^{K}\sum_{q,q'=1}^{Q}\sum_{p\in\{q,q'\}}B(\delta,n_{k,p},\mathcal{F}).$$

With probability at least $1-\delta$ it holds simultaneously that

$$L(\hat{f}) - L(f^*) \leq 2B\left(\tfrac{\delta}{(4KQ^2+2)}, n, \mathcal{F}\right),$$
$$\sum_{k=1}^{K}\sum_{p,q=1}^{Q}\left|L_k^{k,p}(f) - L_k^{k,q}(f)\right| \leq \epsilon + 2\sum_{k=1}^{K}\sum_{q,q'=1}^{Q}\sum_{p\in\{q,q'\}} B\left(\tfrac{\delta}{(4KQ^2+2)}, n_{k,p}, \mathcal{F}\right).$$

Proof 1: We first use Eq. (3) to conclude that, with probability at least $1-2KQ^2\delta$,

$$\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{K}\sum_{p,q=1}^{Q}\left|L_k^{k,p}(f)-L_k^{k,q}(f)\right|-\left|\hat{L}_k^{k,p}(f)-\hat{L}_k^{k,q}(f)\right|\right| \leq \sum_{k=1}^{K}\sum_{q,q'=1}^{Q}\sum_{p\in\{q,q'\}} B(\delta, n_{k,p}, \mathcal{F}). \quad (4)$$

This inequality in turn implies that, with probability at least $1-2KQ^2\delta$, it holds that

$$\left\{f : f\in\mathcal{F}, \sum_{k=1}^{K}\sum_{p,q=1}^{Q}\left|L_k^{k,p}(f)-L_k^{k,q}(f)\right| \leq \epsilon\right\} \subseteq \left\{f : f\in\mathcal{F}, \sum_{k=1}^{K}\sum_{p,q=1}^{Q}\left|\hat{L}_k^{k,p}(f)-\hat{L}_k^{k,q}(f)\right| \leq \hat{\epsilon}\right\}. \quad (5)$$

Now, in order to prove the first statement of the theorem, let us decompose the excess risk as

$$L(\hat{f})-L(f^*)=L(\hat{f})-\hat{L}(\hat{f})+\hat{L}(\hat{f})-\hat{L}(f^*)+\hat{L}(f^*)-L(f^*).$$

The inclusion property of Eq. (5) implies that $\hat{L}(\hat{f})-\hat{L}(f^*) \leq 0$ with probability at least $1-2KQ^2\delta$. Consequently with probability at least $1-2KQ^2\delta$ it holds that

$$L(\hat{f}) - L(f^*) \leq L(\hat{f}) - \hat{L}(\hat{f}) + \hat{L}(f^*) - L(f^*).$$

The first statement now follows by Eq. (3). As for the second statement, its proof consists in exploiting the results of Eqns. (4) and (5) together with a union bound.

A consequence of the first statement of Theorem 1 is that as n tends to infinity $L(\hat{f})$ tends to a value which is not larger than $L(f^*)$, that is, G-FERM is consistent with respect to the risk of the selected model. The second statement of Theorem 1, instead, implies that as n tends to infinity we have that $\hat{f}$ tends to be ϵ-fair. In other words, G-FERM is consistent with respect to the fairness of the selected model.

Remark 3: Since $K, Q \leq n$ the bound in Theorem 1 behaves as $\sqrt{\ln(1/\delta)/n}$ in the same setting of Remark 2 which is optimal [37].

Thanks to Theorem 1 we can state that f^* is close to $\hat{f}$ both in term of its risk and its fairness. Nevertheless, our final goal is to find an f_h^* which solves the following problem

$$\min_{f\in\mathcal{F}}\left\{L(f): \sum_{k=1}^{K}\sum_{p,q=1}^{Q}\left|P^{k,p}(f)-P^{k,q}(f)\right| \leq \epsilon\right\}. \quad (6)$$

Note that, the quantities in Problem (6) cannot be computed since the underline data generating distribution is unknown. Moreover, the objective function and the fairness constraint of Problem (6) are non convex.

Theorem 1 allow us to solve the first issue since we can safely search for a solution $\hat{f}_h$ of the empirical counterpart of Problem (6), which is given by

$$\min_{f\in\mathcal{F}}\left\{\hat{L}(f): \sum_{k=1}^{K}\sum_{p,q=1}^{Q}\left|\hat{P}^{k,p}(f)-\hat{P}^{k,q}(f)\right| \leq \hat{\epsilon}\right\} \quad (7)$$

where

$$\hat{P}^{k,q}(f)=\frac{1}{n_{k,q}}\sum_{(\boldsymbol{z},y)\in\mathcal{D}_{k,q}} \mathbb{1}\left\{f(\boldsymbol{z})\in[t_k, t_{k+1})\right\}. \quad (8)$$

Unfortunately, Problem (7) is still a difficult non-convex non-smooth problem, and for this reason it is more convenient to solve a convex relaxation. That is, we replace the possible non-convex loss function in the risk with its convex upper bound ℓ_c (e.g. the square loss $\ell_c=(y-f(\boldsymbol{z}))^2$) and the losses ℓ_k, $1\leq k\leq K$, in the constraint with a relaxation (e.g. the linear loss $\ell_l(\hat{y}, y)=\hat{y}-y$) which allows to make the constraint convex. In this way, we look for a solution $\hat{f}_c$ of the convex G-FERM problem

$$\min_{f\in\mathcal{F}}\left\{\hat{L}_c(f): \sum_{k=1}^{K}\sum_{p,q=1}^{Q}\left|\hat{L}_l^{k,p}(f)-\hat{L}_l^{k,q}(f)\right| \leq \hat{\epsilon}\right\}. \quad (9)$$

Note that this approximation of the fairness constraint correspond to matching the first order moment [1].

The questions that arise here are whether $\hat{f}_c$ is close to $\hat{f}_h$, how much, and under which assumptions. The following proposition sheds some lights on these issues.

Proposition 1: If ℓ_c is a convex upper bound of the loss exploited to compute the risk then $\hat{L}_h(f) \leq \hat{L}_c(f)$. Moreover, if for $f : \mathcal{X} \to \mathbb{R}$ and for ℓ_l

$$\sum_{k=1}^{K}\sum_{p,q=1}^{Q}\left|\hat{P}^{k,p}(f)-\hat{P}^{k,q}(f)\right| - \left|\hat{L}_l^{k,p}(f)-\hat{L}_l^{k,q}(f)\right| \leq \hat{\Delta}$$

with $\hat{\Delta}$ small, then also the fairness is well approximated.

The first statement of Proposition 1 tells us that exploiting the quality in approximating the risk depend on the quality of the convex approximation. The second statement of Proposition 1, instead, tells us that if $\hat{\Delta}$ is small then the linear loss based fairness is close to the GF. This condition is quite natural, empirically verifiable, and it has been exploited in previous work [1], [39]. Moreover, in Section V we present experiments showing that $\hat{\Delta}$ is small.

The bound in Proposition 1 may be tighten by using different non-linear approximations of the GF. However, the linear approximation proposed in this work gives a convex problem, and as we shall see in Section V, works well in practice.

In summary, the combination of Theorem 1 and Proposition 1 provides conditions under which a solution $\hat{f}_c$ of Problem (2), which is convex, is close, *both in terms of risk*

and fairness measure, to a solution f_h^* of Problem (6), which is our final goal.

IV. G-FERM with Kernel Methods

In this section, we specify the G-FERM framework to the case that the underlying space of models is a reproducing kernel Hilbert space (RKHS) [40], [41].

We let $\kappa{:}\mathcal{Z}\times\mathcal{Z}\to\mathbb{R}$ be a positive definite kernel and let $\boldsymbol{\phi}{:}\mathcal{Z}\to\mathbb{H}$ be an induced feature mapping such that $\kappa(\boldsymbol{z},\boldsymbol{z}')=\langle\boldsymbol{\phi}(\boldsymbol{z}),\boldsymbol{\phi}(\boldsymbol{z}')\rangle$, for all $\boldsymbol{z},\boldsymbol{z}'\in\mathcal{Z}$, where $\mathbb{H}$ is the Hilbert space of square summable sequences. Functions in the RKHS can be parametrized as

$$f(\boldsymbol{z}) = \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{z})\rangle, \quad \boldsymbol{z}\in\mathcal{Z}, \tag{10}$$

for some vector of parameters $\boldsymbol{w}\in\mathbb{H}$. In practice a bias term (threshold) can be added to f but to ease our presentation we do not include it here.

We propose to solve Problem (9) in the case that $\mathcal{F}$ is a ball in the RKHS and employ a convex loss function $\ell_c(y,\hat{y})$ to measure the empirical error. Standard choices are the square loss in the case of regression or the hinge loss in the case of binary classification. They are defined, for every $y,\hat{y}\in\mathbb{R}$, as $(y-\hat{y})^2$ and $\max(0,1-y\hat{y})$, respectively. As for the fairness constraint we use the linear loss function ℓ_l which implies the constraint to be convex. Then, we introduce the mean of the feature vectors associated with the training points restricted by the discretization of the sensitive feature and real outputs, namely

$$\boldsymbol{u}_{k,q} = \tfrac{1}{N_{k,q}}\textstyle\sum_{(\boldsymbol{z},y)\in\mathcal{D}_{k,q}}\boldsymbol{\phi}(\boldsymbol{z}). \tag{11}$$

Using Eq. (10) the constraint in Problem (9) becomes

$$\textstyle\sum_{k=1}^{K}\sum_{p,q=1}^{Q}|\langle \boldsymbol{w}, \boldsymbol{u}_{k,p}-\boldsymbol{u}_{k,q}\rangle| \le \hat{\epsilon} \tag{12}$$

which can be written with more compact notation as $\|A^T w\|_1\le\hat{\epsilon}$, where $A:\mathbb{H}\to\mathbb{R}^{KQ^2}$ is the linear operator mapping a vector w to the vector $\langle \boldsymbol{w}, \boldsymbol{u}_{k,p}-\boldsymbol{u}_{k,q}\rangle$. With this notation, the fairness constraint can be interpret as the composition of $\hat{\epsilon}$ ball of the ℓ_1 norm with a linear transformation A.

In practice, we solve the following Tikhonov regularization problem

$$\begin{aligned}\min_{\boldsymbol{w}\in\mathbb{H}} \quad & \sum_{i=1}^{n}\ell_c(y_i,\langle \boldsymbol{w},\boldsymbol{\phi}(\boldsymbol{z}_i)\rangle)+\lambda\|\boldsymbol{w}\|^2 \\ \text{s.t.} \quad & \|A^\top w\|_1\le\hat{\epsilon},\end{aligned} \tag{13}$$

where λ is a positive parameter. Note that, if $\hat{\epsilon}{=}0$ the constraint reduces to the linear constraint $A^\top w{=}0$.

Problem (13) can be kernelized by observing that, thanks to the Representer Theorem [40]

$$\boldsymbol{w} = \textstyle\sum_{i=1}^{n}\alpha_i\boldsymbol{\phi}(\boldsymbol{z}_i). \tag{14}$$

The dual of Problem (13) may be derived using Fenchel duality, see e.g. [42, Theorem 3.3.5]. We postpone the discussion to future work since in our experiments we employed an off-the-shelf convex optimization solver[2].

[2]https://www.ibm.com/analytics/cplex-optimizer

Finally, we note that in the case when $\boldsymbol{\phi}$ is the identity mapping (i.e. κ is the linear kernel on $\mathbb{R}^d$) and $\hat{\epsilon}{=}0$ then the fairness constraint of Problem (13) can be implicitly enforced by making a change of representation [1].

V. Experiments

In this section we present a set of experiments to test the performance of the proposed method, both in terms of error and fairness. We will study both the cases with categorical and continuous sensitive feature in the context of the regression (continuous label). The classification task, as special case of our proposed framework, has been already studied in [1]. For this purpose, we selected two metrics to compare our method with the other baselines. Concerning the error we collected the Mean Absolute Percentage Error (MAPE), that is equal to $\hat{L}(f)$ on the test set when $\ell(f(\boldsymbol{z}),y)=100\frac{|y-f(\boldsymbol{z})|}{|y|}$. For what concerns the fairness of the model we will exploit the Differences of GF (DGF), see Definition 1, that is the following quantity, still estimated on the test set as

$$\text{DGF}(f)=\textstyle\sum_{k=1}^{K}\sum_{p,q=1}^{Q}\left|\hat{P}^{k,p}(f)-\hat{P}^{k,q}(f)\right|$$

where the expression of $\hat{P}^{k,p}(f)$ is given in Eq. (8).

A set of four different algorithms is considered, with two different types of validation procedures. The algorithms are divided in two groups: linear and non-linear kernels. Concerning the linear methods, the baseline is regularized least squares (RLS), where we solve Problem (13) with no fairness constraint and a linear kernel. Fair RLS is our method in this category, that solves Problem (13) with a linear kernel including the fairness constraint. A kernel version of the same methods is KRLS, that solves Problem (13) with no fairness constraint and a Gaussian kernel, i.e. $\kappa(\boldsymbol{z},\boldsymbol{z}')=e^{-\gamma\|\boldsymbol{z}-\boldsymbol{z}'\|^2}$. In comparison, our proposed algorithm is Fair KRLS, where we tackle Problem (13) with the fairness constraint and a Gaussian kernel.

We follow two different types of possible validation procedures[3]. The first one is standard, and we call it Naive Validation (Naive). In particular, we performed a nested 10-fold cross validation (CV) to select the best hyperparameters and to test the final model. This procedure is repeated 30 times, and we reported the average performance on the test set alongside its standard deviation. A second validation procedure, called Novel Validation Procedure (NVP) as in [1], is slightly different and more focused on finding the best fair model among the ones with low error. Also in this case, as general structure, we performed a nested 10-fold CV to test the final model. For the inner part of the nested CV, we employ a two steps procedure. In the first step, the 10-fold CV error for each of the combination of the hyperparameters is computed. In the second step, we shortlist all the hyperparameters' combinations with error close to the best one (in our case, above 90% of the best MAPE). Finally, from this list, we select the hyperparameters with the lowest DGF.

[3]Hyperparameters range: $\lambda\in\{10^{-4.0},10^{-3.5},\cdots,10^{+4.0}\}$ and $\gamma\in\{10^{-4},10^{-3},\cdots,10^{+4}\}$.

For the sake of completeness, all the experiments have been performed both having and not having the sensitive feature in the model's functional form, i.e. the sensitive feature is available (or not available) at test time.

A. Datasets

For the purpose of testing the proposed proposed methodology we employed two different datasets for regression.

The first one is a classic benchmark dataset for fairness called Communities and Crime dataset[4] (CRIME). CRIME combines socioeconomic data and crime rate data on communities in the United States. In the case of categorical sensitive feature, following [28], we made a binary attribute s as to the percentage of black population, which yielded 970 instances of $s{=}1$ with a mean crime rate 0.35 and 1024 instances of $s{=}0$ with a mean crime rate 0.13. In this case $\mathcal{S}_Q{=}\{-0.5, 0.5, 1.5\}$. Concerning the experiments with continuous sensitive feature we maintain the real value of the percentage of black population, avoiding the binarization step of it and then we consider $Q{=}5$ and a uniform set $\mathcal{S}_Q$ over $[0, 1]$, i.e. $\mathcal{S}_Q{=}\{0.0, 0.2, \ldots, 0.8, 1.0\}$.

The second dataset we propose is new and it has been collected at the University of `Anonymous` (UNIV). This dataset is a proprietary and highly sensitive dataset containing all the data about the past and present students enrolled at the UNIV. In this study we take into consideration students who enrolled, in the academic year (a.y.) 2017-2018. The dataset contains 5000 instances, each one described by 35 attributes (both numeric and categorical) about ethnicity, gender, financial status, and previous school experience. The scope is to predict the average grades and the end of the first semester. In the case of categorical sensitive feature, we consider as sensitive feature the gender ($s{=}1$ female and $s{=}0$ male) and consequently $\mathcal{S}_Q{=}\{-0.5, 0.5, 1.5\}$. In the context of continuous sensitive attribute, we select as sensitive feature the income of the student, with $Q{=}5$ following the official separation in five bins from the tuition system of the University of `Anonymous` (details at link `link anonymous`).

B. Results and Discussion

Results for regression tasks with categorical sensitive feature are presented in Table I, where MAPE and DGF are shown for the different datasets (CRIME and UNIV), algorithms (RLS and KRLS), validation procedure (Naive and NVM), with and without the fairness constraints, and availability of the sensitive feature at test time.

For both datasets, it is clear the advantage of using our method (see also Figure 1) in order to obtain more fair models (i.e. lower DGF) at the expenses of a slightly higher error (i.e. higher MAPE). Moreover, having the sensitive feature at test time increases model accuracy (i.e. lower MAPE) and reduces the fairness measure (i.e. higher DGF). The improvement is stronger in the kernel case, and where the original unfairness of the standard method is higher.

[4]http://archive.ics.uci.edu/ml/datasets/communities+and+crime

TABLE I
RESULTS WITH $\hat{\epsilon} = 0$ AND $K = 10$.

Method	CRIME MAPE	CRIME DGF	UNIV MAPE	UNIV DGF
Sensitive Feature $\notin$ the model's functional form.				
Naive RLS	9.1±0.5	0.19±0.06	21.2±1.8	0.29±0.08
NVM RLS	10.2±0.8	0.16±0.05	23.4±1.9	0.23±0.09
NVM Fair RLS	10.5±1.0	0.11±0.04	24.2±1.9	0.15±0.09
Naive KRLS	8.7±0.4	0.18±0.05	12.2±0.8	0.19±0.05
NVM KRLS	8.9±0.7	0.17±0.05	13.7±1.1	0.12±0.05
NVM Fair KRLS	9.0±0.7	0.11±0.04	14.1±1.2	0.06±0.03
Sensitive Feature $\in$ the model's functional form.				
Naive RLS	9.1±0.6	0.20±0.05	19.7±1.7	0.33±0.11
NVM RLS	9.5±0.6	0.18±0.05	21.9±1.9	0.28±0.09
NVM Fair RLS	9.5±0.7	0.12±0.03	21.8±1.8	0.19±0.10
Naive KRLS	8.5±0.6	0.19±0.04	11.5±0.8	0.21±0.06
NVM KRLS	8.6±0.6	0.18±0.05	12.6±0.9	0.13±0.05
NVM Fair KRLS	8.7±0.7	0.12±0.04	12.9±0.9	0.07±0.03

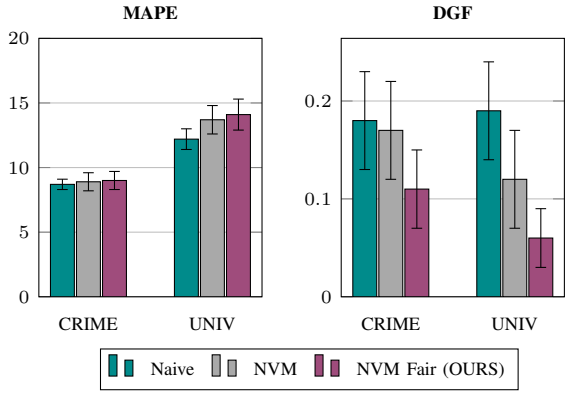

Fig. 1. Graphical representation of Table I for the non-linear version of the methods, when the sensitive feature is not in the model's functional form.

TABLE II
$\hat{\Delta}$ WITH $\hat{\epsilon} = 0$ AND $K = 10$.

Method	CRIME $\hat{\Delta}$	UNIV $\hat{\Delta}$
Sensitive Feature not included in the model's functional form.		
NVM Fair RLS	0.03	0.02
NVM Fair KRLS	0.03	0.03
Sensitive Feature included in the model's functional form.		
NVM Fair RLS	0.04	0.03
NVM Fair KRLS	0.03	0.03

An important question concerns the sensitivity of our method with respect to the parameter $\hat{\epsilon}$ (acceptable unfairness) and the number of bins K. Tables III and IV reports this analysis. We repeated the same experimental procedure of Table I for both datasets (CRIME and UNIV), and algorithms (RLS and KRLS), and possible availability of the sensitive feature at test time, when the fairness constraint is active and with the NVM. We let $\hat{\epsilon}$ range in $\{0, 0.005, 0.001\}$ with fixed $K{=}10$, and also let K range in $\{5, 10, 20\}$ maintaining $\hat{\epsilon}{=}0$. The results confirm our theoretical insights. Making $\hat{\epsilon}$ larger induces lower MAPE and larger DGF, confirming the trade-off between error and fairness. Considering K, we have that

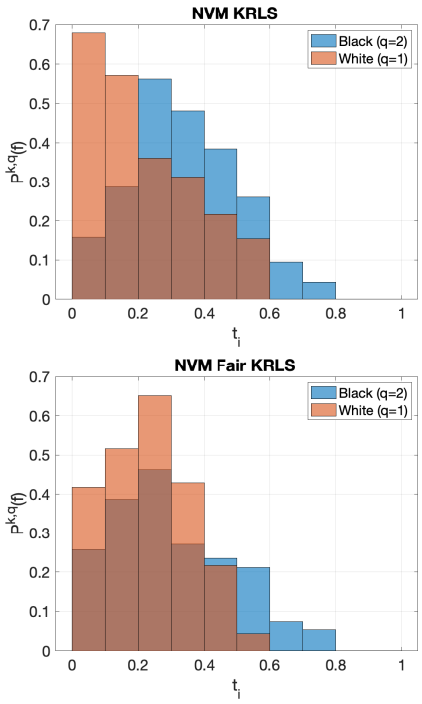

Fig. 2. Two overlapped (White q=1 Black q=2) histograms of $\mathbb{P}^{k,q}$ for the CRIME dataset with NVM KRLS and NVM Fair KRLS when the sensitive feature not included in the function form of the model.

TABLE III
RESULTS VARYING $\hat{\epsilon}$ WITH $K = 10$

Method	$\hat{\epsilon} = 0$ MAPE	$\hat{\epsilon} = 0$ DGF	$\hat{\epsilon} = 0.005$ MAPE	$\hat{\epsilon} = 0.005$ DGF	$\hat{\epsilon} = 0.01$ MAPE	$\hat{\epsilon} = 0.01$ DGF
CRIME						
Sensitive Feature $\notin$ the model's functional form.						
NVM Fair RLS	10.5	0.11	10.3	0.14	10.2	0.16
NVM Fair KRLS	9.0	0.11	8.9	0.14	8.9	0.17
Sensitive Feature $\in$ the model's functional form.						
NVM Fair RLS	9.5	0.12	9.5	0.16	9.5	0.18
NVM Fair KRLS	8.7	0.12	8.6	0.17	8.6	0.18
UNIV						
Sensitive Feature $\notin$ the model's functional form.						
NVM Fair RLS	24.2	0.15	23.7	0.19	23.4	0.23
NVM Fair KRLS	14.1	0.06	13.9	0.09	13.7	0.12
Sensitive Feature $\in$ the model's functional form.						
NVM Fair RLS	21.8	0.19	21.8	0.24	21.9	0.28
NVM Fair KRLS	12.9	0.07	12.7	0.09	12.6	0.13

larger values of K corresponds to impose a higher number of constraints, something that impacts negatively the MAPE value (i.e. the higher K, the higher MAPE). On the other hand, increasing the value of K makes the final model more fair, with a lower DGF.

Figure 2 shows the different behaviours of the standard non-linear regression models (without fairness constraints, i.e. NVM KRLS) and our method (NVM Fair KRLS) over the CRIME dataset, specifically when the sensitive feature is not part of the model's functional form. In particular, we reported the different element in the summation which composes the DGF: $P^{k,q}(f)$ for White (q=1) and Black (q=2). Our method,

TABLE IV
RESULTS VARYING K WITH $\hat{\epsilon} = 0$

Dataset / Method	$K = 5$ MAPE	$K = 5$ DGF	$K = 10$ MAPE	$K = 10$ DGF	$K = 20$ MAPE	$K = 20$ DGF
CRIME						
Sensitive Feature $\notin$ the model's functional form.						
NVM Fair RLS	10.4	0.13	10.5	0.11	15.5	0.05
NVM Fair KRLS	9.0	0.14	9.0	0.11	14.8	0.04
Sensitive Feature $\in$ the model's functional form.						
NVM Fair RLS	9.5	0.16	9.5	0.12	13.8	0.05
NVM Fair KRLS	8.7	0.15	8.7	0.12	13.7	0.04
UNIV						
Sensitive Feature $\notin$ the model's functional form.						
NVM Fair RLS	23.6	.019	24.2	0.15	35.7	0.06
NVM Fair KRLS	13.7	.010	14.1	0.06	22.4	0.03
Sensitive Feature $\in$ the model's functional form.						
NVM Fair RLS	21.8	0.25	21.8	0.19	33.9	0.09
NVM Fair KRLS	12.8	0.11	12.9	0.07	21.8	0.03

TABLE V
RESULTS WITH $\hat{\epsilon} = 0$, $K = 10$ AND $Q = 5$.

Method	CRIME MAPE	CRIME DGF	UNIV MAPE	UNIV DGF
Sensitive Feature $\notin$ the model's functional form.				
NVM KRLS	8.9±0.7	0.17±0.05	15.9±1.3	0.16±0.06
NVM Fair KRLS	10.5±0.8	0.05±0.02	17.8±1.4	0.04±0.02
Sensitive Feature $\in$ the model's functional form.				
NVM KRLS	8.6±0.6	0.18±0.05	14.5±1.3	0.19±0.07
NVM Fair KRLS	10.1±0.8	0.06±0.03	16.2±1.4	0.05±0.02

(bottom plot) obtains two probability distributions among the two different groups that are more similar with respect the baseline (top plot).

We collected in Table II the $\hat{\Delta}$ values (see Proposition 1), for both datasets, for both NVM Fair RLS and NVM Fair KRLS, with and without the sensitive feature in the model's functional form. As it can be noted, the value $\hat{\Delta}$ remains small and, consequently, our method provides a good convex approximation of the original non-convex optimization problem of Eq. (7) in practice.

As a final experiment, we empirically demonstrate that it is possible to generate fair models with continuous sensitive features. Table V reports the results for NVM KRLS and NVM Fair KRLS for both datasets with and without the sensitive feature in the functional form of the model. The obtained MAPE and DGF confirm the results described above in the case of categorical sensitive attributes, empirically demonstrating that our methodology is able to tackle the regression tasks having categorical and continuous sensitive feature.

VI. CONCLUSION AND FUTURE WORK

In this work, we studied the problem of enhancing supervised learning with fairness requirements. We presented a framework based on empirical risk minimization under a novel and generalized fairness constraint. Contrarily to the previous methods, our approach can handle both regression and classification problems and both continuous or categorical

sensitive attributes. Furthermore we observed that our approach generalizes and reduces to known approaches available in literature. We addressed the statistical properties of the method and considered a convex relaxation of the fairness constraint, which can be linked to the non-convex constraint by means of a data dependent bound. We instantiated this approach in the setting of kernel methods, for which the convex fairness constraint can be efficiently implemented both implicitly and explicitly. Finally, we provided experimental results on two real-world datasets that indicate the effectiveness of our approach in comparison with some baselines which either do not impose the fairness constraint or impose the constraint during the validation procedure. Future work will be devoted to extend the range of applicability of our method and to study tighter bounds under specialized conditions.

ACKNOWLEDGEMENTS

This work was supported in part by both SAP SE and Amazon Web Services.

REFERENCES

[1] M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil, "Empirical risk minimization under fairness constraints," in *Advances in Neural Information Processing Systems*, 2018.

[2] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Advances in Neural Information Processing Systems*, 2017.

[3] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, "Data decisions and theoretical implications when adversarially learning fair representations," in *Conference on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

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

[5] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[6] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro, "Learning non-discriminatory predictors," in *Computational Learning Theory*, 2017.

[7] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *International Conference on World Wide Web*, 2017.

[8] M. B. Zafar, I. Valera, M. Rodriguez, K. Gummadi, and A. Weller, "From parity to preference-based notions of fairness in classification," in *Advances in Neural Information Processing Systems*, 2017.

[9] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *International Conference on Artificial Intelligence and Statistics*, 2017.

[10] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *International Conference on Data Mining Workshops*, 2011.

[11] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," *arXiv preprint arXiv:1711.05144*, 2017.

[12] A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz-Marí, L. Gómez-Chova, and G. Camps-Valls, "Fair kernel learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2017.

[13] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A convex framework for fair regression," *arXiv preprint arXiv:1706.02409*, 2017.

[14] J. Adebayo and L. Kagal, "Iterative orthogonal feature projection for diagnosing bias in black-box models," in *Conference on Fairness, Accountability, and Transparency in Machine Learning*, 2016.

[15] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, 2017.

[16] F. Kamiran and T. Calders, "Classifying without discriminating," in *International Conference on Computer, Control and Communication*, 2009.

[17] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013.

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

[19] ——, "Classification with no discrimination by preferential sampling," in *Machine Learning Conference*, 2010.

[20] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa, "Wasserstein fair classification," in *Uncertainty in Artificial Intelligence*, 2019.

[21] S. Chiappa, "Path-specific counterfactual fairness," in *AAAI Conference on Artificial Intelligence*, 2019.

[22] A. Agarwal, A. Beygelzimer, M. Dudík, and J. Langford, "A reductions approach to fair classification," in *Conference on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

[23] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *Conference on Fairness, Accountability and Transparency*, 2018.

[24] Y. Bechavod and K. Ligett, "Penalizing unfairness in binary classification," *arXiv preprint arXiv:1707.00044v3*, 2018.

[25] D. Alabi, N. Immorlica, and A. T. Kalai, "When optimizing nonlinear objectives is no harder than linear objectives," *arXiv preprint arXiv:1804.04503*, 2018.

[26] C. Dwork, N. Immorlica, A. T. Kalai, and M. D. M. Leiserson, "Decoupled classifiers for group-fair and efficient machine learning," in *Conference on Fairness, Accountability and Transparency*, 2018.

[27] J. Komiyama and H. Shimao, "Two-stage algorithm for fairness-aware machine learning," *arXiv preprint arXiv:1710.04924*, 2017.

[28] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang, "Controlling attribute effect in linear regression," in *IEEE International Conference on Data Mining*, 2013.

[29] J. Fitzsimons, A. A. Ali, M. Osborne, and S. Roberts, "Equality constrained decision trees: For the algorithmic enforcement of group fairness," *arXiv preprint arXiv:1810.05041*, 2018.

[30] E. Raff, J. Sylvester, and S. Mills, "Fair forests: Regularized tree induction to minimize model bias," *arXiv preprint arXiv:1712.08197*, 2017.

[31] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Innovations in theoretical computer science conference*, 2012.

[32] J. Komiyama, A. Takeda, J. Honda, and H. Shimao, "Nonconvex optimization for regression with fairness constraints," in *International Conference on Machine Learning*, 2018.

[33] R. Nabi and I. Shpitser, "Fair inference on outcomes," in *AAAI Conference on Artificial Intelligence*, 2018.

[34] R. Nabi, D. Malinsky, and I. Shpitser, "Learning optimal fair policies," *arXiv preprint arXiv:1809.02244*, 2018.

[35] G. Yona and G. Rothblum, "Probably approximately metric-fair learning," in *International Conference on Machine Learning*, 2018.

[36] L. Oneto, M. Donini, A. Elders, and M. Pontil, "Taking advantage of multitask learning for fair classification," in *AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

[37] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[38] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 463–482, 2002.

[39] A. Maurer, "A note on the pac bayesian theorem," *arXiv preprint cs/0411099*, 2004.

[40] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

[41] A. J. Smola and B. Schölkopf, *Learning with Kernels*. MIT Press, 2001.

[42] J. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2010.