Nonconvex Optimization for Regression with Fairness Constraints

Junpei Komiyama [1] **Akiko Takeda** [1 2] **Junya Honda** [1 2] **Hajime Shimao** [3]

Abstract

The unfairness of a regressor is evaluated by measuring the correlation between the estimator and the sensitive attribute (e.g., race, gender, age), and the coefficient of determination (CoD) is a natural extension of the correlation coefficient when more than one sensitive attribute exists. As is well known, there is a trade-off between fairness and accuracy of a regressor, which implies that a perfectly fair optimizer does not always yield a useful prediction. Taking this into consideration, we optimize the accuracy of the estimation subject to a user-defined level of fairness. However, a fairness level as a constraint induces a nonconvexity of the feasible region, which disables the use of an off-the-shelf convex optimizer. Despite such nonconvexity, we show that an exact solution is available by using tools of global optimization theory. Unlike most of existing fairness-aware machine learning methods, our method allows us to deal with numeric and multiple sensitive attributes.

1. Introduction

Algorithmic decision-making process now affects many aspects of our lives. Emails are spam-filtered by classifiers, images are automatically tagged and sorted, and news articles are clustered and ranked. These days, even decisions regarding individual people are being made algorithmically. For example, computer-generated credit scores are popular in many countries, and job interviewees are sometimes evaluated by assessment algorithms. However, a potential loss of transparency, accountability, and fairness arises when decision making is conducted on the basis of past data. If a dataset indicates that specific groups based on sensitive attributes (e.g., gender, race, and religion) are of higher risk in defaulting on loans, direct application of machine learning algorithm would highly likely result in loan applicants on those groups being rejected.

This could be viewed as an algorithmic version of *disparate treatment*[1], where decisions are made on the basis of these sensitive attributes. However, removing sensitive attributes from the dataset is not a sufficient solution as it has a *disparate impact*: In 1970s, the U.S. Supreme Court ruled that the hiring decision at the center of the Griggs v. Duke Power Co. case[2] was illegal because it disadvantaged an application of an applicant of certain race, even though the decision was not explicitly determined on the basis of race. Duke Power Co. was subsequently forced to stop using test scores and diplomas, which are highly correlated with race, in its hiring decisions. In this paper, we consider fair machine learning algorithms that remove disparate impact that arises from the correlation between the sensitive and non-sensitive attributes.

Most of existing fairness-aware machine learning algorithms are for classification. While such classifiers are naturally applied in decision making, regressors provide more useful information in some of the human-related tasks. For example, in the case of criminal records (Calders et al., 2013; Angwin et al., 2016), assessing the risk of reoffending of each criminal is reasonable. In hiring decisions, an employer would naturally consider the productivity of a job applicant. Moreover, in recommendation tasks, the preference of items are usually represented as numeric values.

Taking above into consideration, we study a fair regressor. By definition, a fair algorithm tries to treat several groups equally, and thus it sacrifices some accuracy that could be achieved if it had treated these groups unequally. Therefore, the challenge lies in balancing the regression accuracy and fairness. As discussed in Zafar et al. (2017a), depending on each business necessity, a user of an algorithm can justify some degree of disparate impact to increase the predictive power of the algorithm. Such a degree of unfair impact should be strictly controlled.

A natural interest is how to define fairness of algorithm. In

[1]The University of Tokyo, Tokyo, Japan. [2]RIKEN AIP, Tokyo, Japan. [3]Santa Fe Institute, New Mexico, United States. Correspondence to: Junpei Komiyama <junpei@komiyama.info>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[1]The U.S. Civil Rights Act, July 2, 1964.

[2]Case: 401 U.S. 424, March 8, 1971.

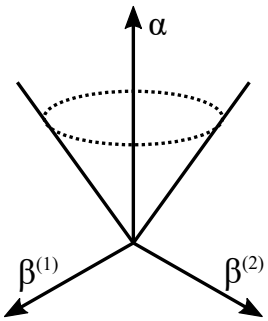


Figure 1. An illustrative example on the feasible region. A linear regression is defined as $\hat{y} = \boldsymbol{s}^\top \boldsymbol{\alpha} + \boldsymbol{u}^\top \boldsymbol{\beta}$, where $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are the corresponding coefficients of attributes. In this example, let the feature dimensions of s and $\boldsymbol{u}$ (and the corresponding coefficients α and $\boldsymbol{\beta}$) be 1 and 2, respectively. Moreover, $\beta^{(l)}$ denotes the l-th component of $\boldsymbol{\beta}$. As detailed in the later section, the upper-half of feasible region of an optimization with the CoD constraint is $\{(\alpha, \boldsymbol{\beta}) : 0 \leq \alpha \leq \sqrt{(\beta^{(1)})^2 + (\beta^{(2)})^2}\}$ after diagonalization and normalization. The region is outside (not inside!) the "ice-cream cone", which cannot be divided into a finite union of convex regions.

this paper, we consider a coefficient of determination (CoD) of the sensitive attributes as a constraint. Let $\boldsymbol{s}$ be the sensitive attributes, and $\boldsymbol{x}$ be the non-sensitive attributes. In general, $\boldsymbol{x}$ is highly correlated with $\boldsymbol{s}$, and we construct $\boldsymbol{u}$ from $\boldsymbol{x}$ by removing its correlation with $\boldsymbol{s}$. Let y be the target variable to predict, and $\hat{y} = \hat{y}(\boldsymbol{s}, \boldsymbol{u})$ be its estimator. CoD is defined as the proportion of the variance of the estimator $\hat{y}$ that is predictable from $\boldsymbol{s}$. In fact, CoD defined in such a way is a natural extension the correlation coefficient to multiple sensitive attributes (Section 2.2).

While CoD is a natural measure of the predictive power, no literature on a fair estimator with CoD as a constraint exists presumably due to its inherent nonconvexity: Figure 1 shows that the feasible region of linear regressors is nonconvex even in the case of single s. As a result, off-the-shelf tools for convex optimization, such as gradient methods, do not give a global solution.

In the context of fairness-aware machine learning, strictly complying with the fairness constraint is of primal importance. However, obtaining an exact solution in nonconvex optimizations is generally hard: For example, even one negative eigenvalue makes a quadratic programming NP-hard (Pardalos & Vavasis, 1991). Fortunately, the optimization under CoD constraint can be solved exactly unlike most of these nonconvex optimizations: We propose two optimization methods by utilizing tools of global optimization theory. The first one is based on a Lagrangian dual that boils down the problem into a semidefinite programming (SDP). Although the Lagrangian dual is efficiently computed and yields an exact optimal value in the optimization, recovering an optimal solution in this problem is not always possible due to a relaxed solution space. To address this issue, we show another optimization method that converts the original nonconvex quadratically constrained quadratic program (QCQP) into a convex QCQP, which yields an exact solution of the problem.

Furthermore, we show that our optimization framework is extended to capture non-linearity by proposing the kernel extension of our framework that is also exactly solvable. As a result, our framework allows us to remove disparate impact that is non-linear to a numeric sensitive attribute (e.g., an unfair deal for young and old people that favors the people in between).

The proposed method is empirically evaluated by four real-world datasets. Unlike most methods, our method is capable of considering the possibly non-linear interaction of numeric sensitive attributes with the target variable. As we consider nonconvexity that naturally arises in measuring a correlation between $\boldsymbol{s}$ and y, we think this result is a first step that ties the study of nonconvex optimization in the context of fairness-aware machine learning.

1.1. Related Work

Most of the tasks in fairness-aware machine learning and data mining fields are divided into two categories (Ruggieri et al., 2010): The former is to discover unfairness (Adebayo & Kagal, 2016; Adler et al., 2018), whereas the latter is to prevent unfair treatments. Classification, regression, and other tasks such as recommendations (Kamishima et al., 2012b; 2016), voting (Bredereck et al., 2018), data summarization(Celis et al., 2018), dimensional reduction (Pérez-Suay et al., 2017), and representational learning (Bolukbasi et al., 2016) are categorized into the latter one. As the goal of this paper is to build a fair regressor, this paper is also categorized into the latter.

Most of the existing papers in the latter category (Kamiran & Calders, 2010; Zliobaite et al., 2011; Kamishima et al., 2012a; Ristanoski et al., 2013; Fish et al., 2015; Hardt et al., 2016; Goh et al., 2016; Zafar et al., 2017a) deal with classification tasks, and thus cannot directly deal with regression tasks. Note that there are several papers that take pre-processing strategy, which makes the data into fair representation so that we can put them into off-the-shelf machine learning algorithms. Among this approach, Zemel et al. (2013) segregated the data by mapping them into finite sets. Feldman et al. (2015) merged distribution of datapoints with binary sensitive attribute $s = 1$ and $s = 0$ into a single distribution. Calmon et al. (2017) characterized a class of convex data preprocessing related to non-discrimination. While the methods in these papers are general enough to deal with regression tasks, this approach treats the algorithm as a black box and could potentially

reduce the predictive power the algorithm by excessively reducing the information in the original dataset.

A few papers considered fairness in regression problems. Fukuchi et al. (2013) considered a generative model that is neutral to a finite set of viewpoints. Calders et al. (2013) introduced a propensity score based approach that enables us to divide people into several clusters on the basis of explainable attributes. Kamishima et al. (2012a) introduced a regularizer that encourages fairness. Berk et al. (2017) considered a convex framework where fairness is imposed by the regularizer, and Pérez-Suay et al. (2017) introduced a regularizer inspired by the Hilbert-Schmidt Independence Criteria (Gretton et al., 2005). Unlike the existing approaches, our method (i) is capable of not only discrete sensitive attributes (e.g., gender, races) but also numeric sensitive attributes such as ages and (ii) enables strict control of fairness by posing the fairness as an explicit constraint.

2. Problem Setup

Each d-dimensional vector in this paper is a column vector and is identified as a $d \times 1$ matrix. Let n be the number of datapoints. The i-th datapoint is comprised of a tuple $(\boldsymbol{s}_i, \boldsymbol{x}_i, y_i)$, where $\boldsymbol{s}_i \in \mathbb{R}^{d_s}$ is the sensitive attributes of d_s dimensions that require special care (e.g., gender, race, and age), $\boldsymbol{x}_i \in \mathbb{R}^{d_x}$ is the normal (non-sensitive) attributes of d_x dimensions, and $y_i \in \mathbb{R}$ is the target attribute to predict. Given a training dataset of $\{(\boldsymbol{s}_i, \boldsymbol{x}_i, y_i)\}_{i=1}^n$, a fairness-aware algorithm outputs $\hat{y}(\boldsymbol{s}, \boldsymbol{x})$, which is an estimator of y that complies with the fairness criteria. We also denote $\boldsymbol{y} = (y_1, y_2, \dots, y_n)^\top \in \mathbb{R}^{n\times 1}, \mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n\times d_x}, \mathbf{S} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \dots, \boldsymbol{s}_n)^\top \in \mathbb{R}^{n\times d_s}$ to denote a sequence of n datapoints. We assume that each feature in $\boldsymbol{s}$ and $\boldsymbol{x}$, and y is zero-mean. If not, we can always remove their (empirical) means.

2.1. Preprocessing and An Asymptotically Fair Regressor

In practice, $\boldsymbol{s}$ has a strong predictive power and highly correlated with $\boldsymbol{x}$ (e.g., gender is highly correlated with occupation), and thus using $\boldsymbol{x}$ in estimating y leads to a disparate impact. Such a correlation is removed by conducting a regression as follows: Namely,

$$\begin{aligned}\hat{\mathbf{B}} &= (\mathbf{S}^\top\mathbf{S})^{-1}\mathbf{S}^\top\mathbf{X} \quad \in \mathbb{R}^{d_s\times d_x}\\ \mathbf{U} &= \mathbf{X} - \hat{\mathbf{B}}^\top\mathbf{S} \quad \in \mathbb{R}^{n\times d_x}\end{aligned} \tag{2.1}$$

and we define $\boldsymbol{u}_i$ as the i-th datapoint of $\mathbf{U}$. The value $\hat{\mathbf{B}}^\top\mathbf{S}$ is a part of $\boldsymbol{x}$ that is explainable by $\boldsymbol{s}$. The following theorem states the learnability of the linear relation between $\boldsymbol{x}$ and $\boldsymbol{s}$.

Theorem 1. (Asymptotic fairness of a preprocesssed regressor) *Assume a linear relation between $\boldsymbol{s}$ and $\boldsymbol{x}$ such that*

$$\boldsymbol{x}_i = \mathbf{B}^\top\boldsymbol{s}_i + \epsilon_i,$$

where ϵ_i is a zero mean noise $\mathbb{E}[\epsilon_i|\mathbf{X}] = 0$. Then, $\hat{\mathbf{B}} \to \mathbf{B}$ in probability.

The proof of Theorem 1 directly follows from the fact that each of l-th column of $\mathbf{B}$ is an ordinary linear regression (OLS) from $\boldsymbol{s}$ to the l-th column of $\boldsymbol{x}$ and standard asymptotic normality of OLS (e.g., Theorem 5.1 in Wooldridge, 2013). Note that, since the linear regression is a parametric model, one can easily see that the correlation between $\boldsymbol{s}$ and $\boldsymbol{u}$ is $O(1/\sqrt{n})$ while $(\boldsymbol{s}, \boldsymbol{u})$ has the same information as $(\boldsymbol{s}, \boldsymbol{x})$.

The discussion above implies that, if we devise a linear regressor $\hat{y} = \hat{y}(\boldsymbol{u})$, the regressor is asymptotically fair in the sense that $\mathrm{Cov}(\hat{y}, \boldsymbol{s})$ approaches zero as $n \to \infty$. Although such a regressor maximizes fairness, it sacrifices the predictive power that stems from $\boldsymbol{s}$. Instead, the rest of this paper maximizes the predictive power of $\hat{y}$ subject to a user-defined level of fairness. Given asymptotically fair features $\boldsymbol{u}$, the next section defines the coefficient of determination, which measures the explainable power of $\boldsymbol{s}$ over $\hat{y}$.

2.2. Coefficient of Determination

The coefficient of determination (CoD) is widely used to measure the predictive power of features to a target variable. Here, our interest lies in measuring the contribution of $\boldsymbol{s}$ to the estimator $\hat{y}$ of the target variable y. Namely, let

$$\hat{y} = \boldsymbol{s}^\top\boldsymbol{\alpha} + \boldsymbol{u}^\top\boldsymbol{\beta}$$

be the estimator of y. Given $\boldsymbol{s}$ and $\boldsymbol{u}$ are zero-mean and not correlated to each other, the best estimator of $\hat{y}$ by using only non-sensitive features $\boldsymbol{u}$ is $\bar{y} = \boldsymbol{u}^\top\boldsymbol{\beta}$ in view of the mean squared error. The variance of $\hat{y}$ is $\boldsymbol{\alpha}^\top\mathbf{V}_s\boldsymbol{\alpha} + \boldsymbol{\beta}^\top\mathbf{V}_u\boldsymbol{\beta}$, where $\mathbf{V}_s \in \mathbb{R}^{d_s\times d_s}$ and $\mathbf{V}_u \in \mathbb{R}^{d_x\times d_x}$ are the covariances of $\boldsymbol{s}$ and $\boldsymbol{u}$, respectively. Morevoer, the variance of $\hat{y} - \bar{y}$ is $\boldsymbol{\alpha}^\top\mathbf{V}_s\boldsymbol{\alpha}$. The CoD (or the R-squared) of sensitive attribute $\boldsymbol{s}$ over $\hat{y}$ is defined as

$$R^2 = \frac{\mathrm{Var}(\hat{y} - \bar{y})}{\mathrm{Var}(\hat{y})} = \frac{\boldsymbol{\alpha}^\top\mathbf{V}_s\boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top\mathbf{V}_s\boldsymbol{\alpha} + \boldsymbol{\beta}^\top\mathbf{V}_u\boldsymbol{\beta}}.$$

CoD and the correlation coefficient: Let there be only one sensitive attribute (i.e. $s, \alpha \in \mathbb{R}$ and $\mathbf{V}_s = \mathrm{Var}(s)$). In this case, CoD matches the correlation coefficient: The correlation coefficient $\rho(\hat{y}, s)$ is transformed as

$$\begin{aligned}\rho(\hat{y}, s) &= \frac{\mathrm{Cov}(\hat{y}, s)}{\sqrt{\mathrm{Var}(\hat{y})\mathrm{Var}(s)}}\\ &= \frac{\boldsymbol{\alpha}\mathrm{Var}(s)}{\sqrt{\boldsymbol{\alpha}^\top\mathbf{V}_s\boldsymbol{\alpha} + \boldsymbol{\beta}^\top\mathbf{V}_u\boldsymbol{\beta}}\sqrt{\mathrm{Var}(s)}}\\ &= \frac{\boldsymbol{\alpha}\sqrt{\mathbf{V}_s}}{\sqrt{\boldsymbol{\alpha}^\top\mathbf{V}_s\boldsymbol{\alpha} + \boldsymbol{\beta}^\top\mathbf{V}_u\boldsymbol{\beta}}} = R.\end{aligned} \tag{2.2}$$

Note also that, the mean difference (MD) (Calders et al., 2013) is very similar to the correlation coefficient with binary s (See Appendix D). In summary, CoD is a multi-attribute generalization of the correlation coefficient for vector s in the least square regression.

2.3. Least Square Regression with Coefficient of Determination Constraints

In this paper, we consider the least square regression with CoD constraint. Namely,

$$\begin{aligned} \min \ & \mathbb{E}[(y-\hat{y})^2] \\ \text{s.t.} \ & R^2 \leq \epsilon, \end{aligned} \tag{2.3}$$

where $\epsilon \in [0,1]$ is a user-defined value that determines how fair the estimator is. The value $\epsilon = 0$ corresponds to a fully fair regressor, whereas $\epsilon = 1$ corresponds to a completely fairness-ignorant regressor that solely maximizes the predictive power.

3. Optimization

The optimization problem in Eqn. (2.3) is equivalently written as:

$$\begin{aligned} \min \ & \boldsymbol{\alpha}^\top \mathbf{V}_s \boldsymbol{\alpha} + \boldsymbol{\beta}^\top \mathbf{V}_u \boldsymbol{\beta} - 2(\mathbb{E}[y\boldsymbol{s}^\top \boldsymbol{\alpha}] + \mathbb{E}[y\boldsymbol{u}^\top \boldsymbol{\beta}]) \\ \text{s.t.} \ & (1-\epsilon)\boldsymbol{\alpha}^\top \mathbf{V}_s \boldsymbol{\alpha} - \epsilon \boldsymbol{\beta}^\top V_{\boldsymbol{u}} \boldsymbol{\beta} \leq 0, \end{aligned} \tag{3.1}$$

where $\mathbf{V}_s, \mathbf{V}_u$, and the expectations are taken with the true data generating distribution. Given limited number of training datapoints, we replace them with the empirical analogues. That is, the (l,m)-entry of $\mathbf{V}_s$ is $(1/n)\sum_{i=1}^n s_i^{(l)} s_i^{(m)}$, where we assume that $\boldsymbol{s}$ is normalized to be zero-mean. Let $\boldsymbol{q}_s = \mathbb{E}_n[y\boldsymbol{s}] \in \mathbb{R}^{d_s}$, and $\boldsymbol{q}_u = \mathbb{E}_n[y\boldsymbol{u}] \in \mathbb{R}^{d_x}$, where $\mathbb{E}_n$ is a sample mean such as $\mathbb{E}_n[y\boldsymbol{s}] = (1/n)\sum_{i=1}^n y_i \boldsymbol{s}_i$. Then, the optimization problem (3.1) is explicitly written as:

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \ & \begin{bmatrix}\boldsymbol{\alpha}^\top & \boldsymbol{\beta}^\top\end{bmatrix} \begin{bmatrix}\mathbf{V}_s & 0 \\ 0 & \mathbf{V}_u\end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} - 2 \begin{bmatrix}\boldsymbol{q}_s^\top & \boldsymbol{q}_u^\top\end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} \\ \text{s.t.} \ & \begin{bmatrix}\boldsymbol{\alpha}^\top & \boldsymbol{\beta}^\top\end{bmatrix} \begin{bmatrix}(1-\epsilon)\mathbf{V}_s & 0 \\ 0 & -\epsilon\mathbf{V}_u\end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} \leq 0, \end{aligned} \tag{3.2}$$

where we use 0 to denote a matrix block of zeros.

For ease of discussion, we assume the following condition:

Assumption 1. (Regularity condition) *Covariance matrices $\mathbf{V}_s$ and $\mathbf{V}_u$ are full rank.*

We may expect that Assumption 1 always holds because we may remove some of the redundant features if the assumption is violated. Note that Assumption 1 implies the existence of an interior solution: $\{(\boldsymbol{\alpha},\boldsymbol{\beta}) : (1-\epsilon)\boldsymbol{\alpha}^\top \mathbf{V}_s \boldsymbol{\alpha} - \epsilon\boldsymbol{\beta}^\top \mathbf{V}_u \boldsymbol{\beta} < 0\} \neq \emptyset$. The optimization problem (3.2) is nonconvex due to the negative definiteness of the lower right block $-\epsilon\mathbf{V}_u$ of the quadratic constraints. In the rest of this section, we propose two methods for solving this problem. The first one solves the Lagrangian dual problem, which boils down to a semidefinite programming (SDP). Unfortunately, solving SDP does not always give the solution of the original problem. The second method exploits the structure of the quadratically constrained quadratic programs (QCQP) and makes it convex. From this optimization we can recover the solution of the original problem unlike the SDP-based method. Note that, both methods give the exact optimal objective value to the target optimization problem as shown later in Sections 3.1 and 3.2.

3.1. Lagrangian Dual and SDP-based Optimization

The Lagrangian dual problem of (3.1) is written as

$$\max_{\xi \geq 0} \ \phi(\xi), \tag{3.3}$$

where $\phi(\xi)$ is the optimal function defined as

$$\begin{aligned} \phi(\xi) = \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \ & \begin{bmatrix}\boldsymbol{\alpha}^\top & \boldsymbol{\beta}^\top\end{bmatrix} \begin{bmatrix}\mathbf{V}_s & 0 \\ 0 & \mathbf{V}_u\end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} - 2 \begin{bmatrix}\boldsymbol{q}_s^\top & \boldsymbol{q}_u^\top\end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} \\ & + \xi \begin{bmatrix}\boldsymbol{\alpha}^\top & \boldsymbol{\beta}^\top\end{bmatrix} \begin{bmatrix}\mathbf{V}_s & 0 \\ 0 & -\epsilon\mathbf{V}_u\end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix}. \end{aligned}$$

The biggest advantage in considering the Lagrangian dual lies in the convexity of the optimal value function $\phi(\xi)$ even though the original problem is nonconvex. Although a Lagrangian dual has a duality gap in general, the following theorem, which is well-known in the context of the control theory, assures the inexistence of the duality gap.

Theorem 2. (No duality gap, Theorem 1 in Sturm & Zhang 2003) *Under Assumption 1, the original optimization* (3.2) *and its Lagrangian dual* (3.3) *gives the same optimal value.*

Moreover, a standard discussion on the Schur complement (details are in Appendix A) boils the problem down to the following equivalent optimization

$$\begin{aligned} \max_{\gamma,\xi} \ & \gamma \\ \text{s.t.} \ & \begin{bmatrix}0 & -\boldsymbol{q}_s^\top & -\boldsymbol{q}_u^\top \\ -\boldsymbol{q}_s & \mathbf{V}_s & 0 \\ -\boldsymbol{q}_u & 0 & \mathbf{V}_u\end{bmatrix} - \gamma \begin{bmatrix}1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0\end{bmatrix} \\ & + \xi \begin{bmatrix}0 & 0 & 0 \\ 0 & (1-\epsilon)\mathbf{V}_s & 0 \\ 0 & 0 & -\epsilon\mathbf{V}_u\end{bmatrix} \succeq 0, \ \xi \geq 0, \end{aligned} \tag{3.4}$$

where $\succeq 0$ denotes the positive-definiteness of a matrix. Solving (3.4) only yields the Lagrange coefficient that is

not very useful. Instead, we solve the following dual problem of (3.4) defined as the optimization over a matrix $\mathbf{A} \in \mathbb{R}^{(1+d_s+d_x)\times(1+d_s+d_x)}$:

$$\begin{aligned}\max_{\mathbf{A}\succeq 0} & \begin{bmatrix} 0 & -\boldsymbol{q}_s^\top & -\boldsymbol{q}_u^\top \\ -\boldsymbol{q}_s & \mathbf{V}_s & 0 \\ -\boldsymbol{q}_u & 0 & \mathbf{V}_u \end{bmatrix} \cdot \mathbf{A} \\ \text{s.t. } \lambda & \begin{bmatrix} 0 & 0 & 0 \\ 0 & (1-\epsilon)\mathbf{V}_s & 0 \\ 0 & 0 & -\epsilon\mathbf{V}_u \end{bmatrix} \cdot \mathbf{A} \le 0, \\ & \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \cdot \mathbf{A} = 1, \end{aligned} \tag{3.5}$$

where $\mathbf{A} \cdot \mathbf{B} := \sum_{i,j} A_{i,j} B_{i,j}$ is the element-wise inner product between matrices. Note that Assumption 1 implies the existence of the interior of the feasible region, which leads to the Slater condition of (3.4). As is well-known, the Slater condition suffices for a large class of (possibly nonconvex) optimizations including our ones to have no duality gap, thus leads the following theorem.

Theorem 3. *Under Assumption 1, optimization* (3.4) *and its dual* (3.5) *gives the same objective value.*

In summary, the original optimization problem (2.3) boils down to solving (3.5), which is a semidefinite optimization that off-the-shelf solvers can deal with.

3.2. Convex QCQP Optimization

Although solving the dual of SDP in (3.5) yields the exact objective value, it does not always yield an exact solution of the original problem. If the solution is rank-one, decomposing the solution of SDP into $\mathbf{A} = \theta\theta^\top$ recovers the desired solution of $\boldsymbol{\alpha}, \boldsymbol{\beta}$. Moreover, how $\mathbf{A}$ is close to rank-one can be verified by conducting the singular value decomposition (SVD) to $\mathbf{A}$ and checking whether or not the second and subsequent eigenvalues are sufficiently small or not. Even if the solution is not exactly rank-one, one can still consider the first eigenvalue and the corresponding eigenvector as an approximated solution by using SVD[3]. However, such a solution possibly violates the constraint of the original problem, and recovering a solution that complies with the constraint is hard when $\mathbf{A}$ is not rank-one. Note that the interior-point method, which is used in solving SDP, tends to find an interior point that is not rank-one solution. Taking the above discussion into consideration, we also propose another optimization method.

The original problem (3.2) is nonconvex QCQP and easily converted into the following equivalent optimization:

[3]Conducting SVD yields a primal eigenvalue λ and the corresponding eigenvector $\boldsymbol{v} \in \mathbb{R}^{1+d_s+d_x}$. The solution $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is the last $d_s + d_x$ dimension of $-\sqrt{\lambda}\boldsymbol{v}$.

$$\begin{aligned}\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\gamma} & \\ \text{s.t. } & \begin{bmatrix}\boldsymbol{\alpha}^\top \boldsymbol{\beta}^\top\end{bmatrix} \begin{bmatrix} \mathbf{V}_s & 0 \\ 0 & \mathbf{V}_u \end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} - 2 \begin{bmatrix}\boldsymbol{q}_s^\top & \boldsymbol{q}_u^\top\end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} - \gamma \le 0, \\ & \begin{bmatrix}\boldsymbol{\alpha}^\top \boldsymbol{\beta}^\top\end{bmatrix} \begin{bmatrix} (1-\epsilon)\mathbf{V}_s & 0 \\ 0 & -\epsilon\mathbf{V}_u \end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} \le 0. \end{aligned} \tag{3.6}$$

The following theorem, which is derived in the context of global optimization (Yamada & Takeda, 2018), converts the nonconvex QCQP into a convex QCQP:

Theorem 4. (Reduction to a convex problem) *Assume that there exist at least one* $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ *such that* $(1-\epsilon)\boldsymbol{\alpha}^\top \mathbf{V}_s \boldsymbol{\alpha} - \epsilon \boldsymbol{\beta}^\top \mathbf{V}_u \boldsymbol{\beta} < 0$. *Then, the feasible region of the following relaxed problem is the convex hull of the feasible region of* (3.6)*:*

$$\begin{aligned}\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\gamma} & \\ \textit{s.t. } & \begin{bmatrix}\boldsymbol{\alpha}^\top \boldsymbol{\beta}^\top\end{bmatrix} \begin{bmatrix} \mathbf{V}_s & 0 \\ 0 & \mathbf{V}_u \end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} - 2 \begin{bmatrix}\boldsymbol{q}_s^\top & \boldsymbol{q}_u^\top\end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} - \gamma \le 0, \\ & \begin{bmatrix}\boldsymbol{\alpha}^\top \boldsymbol{\beta}^\top\end{bmatrix} \begin{bmatrix} \frac{1}{\epsilon}\mathbf{V}_s & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} - 2 \begin{bmatrix}\boldsymbol{q}_s^\top & \boldsymbol{q}_u^\top\end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} - \gamma \le 0. \end{aligned} \tag{3.7}$$

Proof sketch of Theorem 4. Note that the second constraint in (3.7) is a linear combination of the two constraints of (3.6), which implies that the feasible region of (3.7) includes the feasible region of (3.6). For each feasible point $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \gamma_0)$ of (3.7) that is infeasible in (3.6), we explicitly construct two points that lie in the feasible region of (3.6) such that $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \gamma_0)$ is a linear combination of the two points. The formal proof follows directly from Theorem 2 in Yamada & Takeda (2018) by putting $t = \gamma, \mathbf{x} = (\boldsymbol{\alpha}, \boldsymbol{\beta}), \underline{\sigma} = 0, \overline{\sigma} = 1/\epsilon$. □

Note that that Assumption 1 implies the conditions required in Theorem 4. The linearity of the objective, combined with Theorem 4 states that solving (3.7) yields an optimal solution of (3.6).

In summary, Theorem 4 allows us to relax the constraint so that the new feasible region is convex without compromising its objective value. As a result, (3.7), which is a convex QCQP, is computed efficiently by off-the-shelf optimizers.

3.3. Computational Complexity

The proposed optimization runs in time $O(n)$: Building $\mathbf{U}$ requires $O((d_s^2 n + d_s^3) d_x)$ time because for each feature in x we train a linear regressor from $\boldsymbol{s}$ to $\boldsymbol{x}$ that yields each feature of $\boldsymbol{u}$. Moreover, optimization in SDP and convex QCQP requires empirical variance $\mathbf{V}_s$ and $\mathbf{V}_u$ that are computed in $O((d_x^2 + d_s^2)n)$. The sizes of matrices in SDP and QCQP are $O((d_x + d_s) \times (d_x + d_s))$, which is

constant to n. One can check that the required memory is $O((d_x + d_s)^2 + (d_x + d_s)n)$.

Note that, the interior point method is known as a polynomial-time method for finding solutions for SDPs and convex QCQPs with arbitrary precision. The complexity of SDP and convex 2-QCQP are $O((d_s + d_x)^{3.5})$ and $O((d_s + d_x)^3)$, respectively (see Section 6.6.1 in Ben-Tal & Nemirovskiaei 2001). In practice, our simulation in Appendix B shows SDP and convex QCQP appear to scale more similarly to $O((d_s + d_x)^2)$ around $10 \leq d_s + d_x \leq 3,000$, which is not very surprising because many instances scale better than the worst-case.

3.4. Regularization

It is straightforward to add a regularizer into our optimization problem described in Section 3: That is, we can incorporate regularization term $(\lambda/n)(\boldsymbol{\alpha}^T\boldsymbol{\alpha} + \boldsymbol{\beta}^\top\boldsymbol{\beta})$, where a larger $\lambda > 0$ induces a stronger regularization toward smaller parameters. The regularizer increases the diagonal entries as $\mathbf{V}_s + (\lambda/n)I_s$ and $\mathbf{V}_u + (\lambda/n)I_{\boldsymbol{u}}$, and does not change positive definite property of $\mathbf{V}_s$ and $\mathbf{V}_u$.

3.5. Approximated Kernelization

The kernelized least squared regression with fairness constraint is formalized as follows. Let $Z_s(\boldsymbol{s})$ (resp. $Z_u(\boldsymbol{u})$) be the functions that map $\boldsymbol{s}$ (resp. $\boldsymbol{u}$) into high-dimensional spaces, and $K_s(\boldsymbol{s}_i, \boldsymbol{s}_j) = Z_s^\top(\boldsymbol{s}_i)Z_s(\boldsymbol{s}_j) \in \mathbb{R}$ (resp. $K_u(\boldsymbol{u}_i, \boldsymbol{u}_j) = Z_u^\top(\boldsymbol{u}_i)Z_u(\boldsymbol{u}_j) \in \mathbb{R}$) be the corresponding positive-definite kernel functions. The representer theorem implies that the estimator $\hat{y}_i$ of datapoint i is written as a linear combination of the kernel functions as

$$\hat{y}_i = \sum_{j=1}^{n} c_{j,s}K_s(\boldsymbol{s}_i, \boldsymbol{s}_j) + c_{j,u}K_u(\boldsymbol{u}_i, \boldsymbol{u}_j),$$

where $c_{j,s}, c_{j,u} \in \mathbb{R}$ are the weight parameters associated with each datapoint j. With an abuse of notation, let $\mathbf{K}_s$ and $\mathbf{K}_u$ be the corresponding $n \times n$ matrices, and $\boldsymbol{c}_s, \boldsymbol{c}_u$ be corresponding size-n vectors. Let

$$\begin{aligned} S_s &= \boldsymbol{c}_s^\top \mathbf{K}_s^2 \boldsymbol{c}_s, \\ S_u &= \boldsymbol{c}_u^\top \mathbf{K}_u^2 \boldsymbol{c}_u, \\ s_s &= \boldsymbol{c}_s^\top \mathbf{K}_s \boldsymbol{y}, \\ s_u &= \boldsymbol{c}_u^\top \mathbf{K}_u \boldsymbol{y}, \end{aligned} \tag{3.8}$$

then the corresponding optimization is:

$$\begin{aligned} \min\ & S_s + S_u - 2s_s - 2s_{\boldsymbol{u}} \\ \text{s.t.}\ & (1-\epsilon)S_s - \epsilon S_u \leq 0. \end{aligned} \tag{3.9}$$

Unfortunately, the following two issues make the optimization in (3.9) impractical: (i) As is customary with kernel learning, solving (3.9) is computationally prohibiting with large n because the corresponding optimizations involve $O(n \times n)$ matrices. Moreover, (ii) removing the correlation between $\boldsymbol{s}$ and $\boldsymbol{u}$ on the (possibly infinite) representation space $Z_s(\boldsymbol{s})$ and $Z_u(\boldsymbol{u})$ is highly non-trivial.

To address the issues above, an approximated kernel representation methods apply: Nyström methods (Rasmussen & Williams, 2006) and the random Fourier features (Rahimi & Recht, 2008) provide us a finite representation of $K_s(\boldsymbol{s}_i, \boldsymbol{s}_j) = Z_s^\top(\boldsymbol{s}_i)Z_s(\boldsymbol{s}_j)$ and $K_u(\boldsymbol{u}_i, \boldsymbol{u}_j) = Z_u^\top(\boldsymbol{u}_i)Z_u(\boldsymbol{u}_j)$. With these representation we no longer use the gram matrices K_s and K_u: We solve the original optimization (3.2) with converted features $Z_s^\top(\boldsymbol{s}_i)$ and $Z_u^\top(\boldsymbol{u}_i)$. Moreover, to remove the correlation between $\boldsymbol{s}$ and $\boldsymbol{u}$, we first map $\boldsymbol{x}$ into the representation space $Z_u(\boldsymbol{x})$, and then remove correlation between $Z_s(\boldsymbol{s})$ and $Z_u(\boldsymbol{x})$ by applying linear regression on the finite representation space. Assuming that the dimension of Z_s (resp. Z_u) is p_s (resp. p_u), the new optimization involves matrices of $O((p_s + p_u) \times (p_s + p_u))$. Therefore, if we choose p_s, p_u as constants with respect to n, the optimizations scale with a large dataset.

Table 1. Statistics of the datasets. The value $d = d_s + d_x$ is the number of attributes (after expanding categorical attributes into dummies), and n is the number of datapoints. $s^{(1)}$ and $s^{(2)}$ are the first and second sensitive attributes of each dataset, respectively. We consider the C&C datasets, which is a classification dataset, as a regression with $y \in \{-1, 1\}$.

datasets	d	n	$s^{(1)}$	$s^{(2)}$
C&C	102	1,994	race	origin
Compas	12	5,855	gender	race
NLSY	22	7,244	gender	age
LSAC	19	20,798	race	age

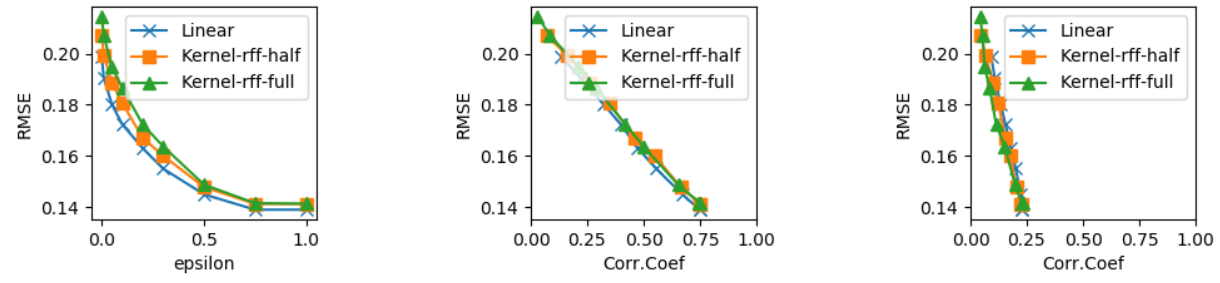
4. Experiment

This section verifies the performance of the proposed method in four real-world datasets.

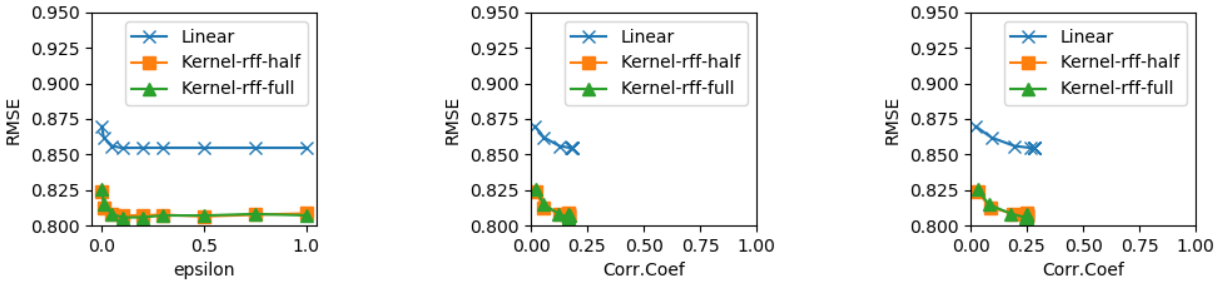
Computation environment: The simulation here was conducted by using modern Xeon-core PC servers[4]. Preliminary experiment (Appendix B) revealed that the running times of SDP and convex QCQP optimization are more or less the same. Taking the fact that the convex QCQP always yields an exact solution into consideration, we adapted it in the subsequent simulations. We solved the convex QCQP optimization by using the Gurobi optimizer[5].

[4]The source code used in the simulation is available at https://github.com/jkomiyama/fairregresion.

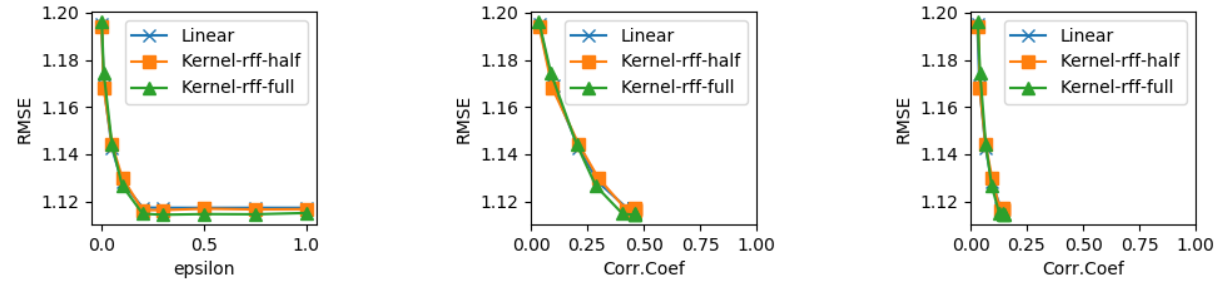
[5]http://www.gurobi.com/



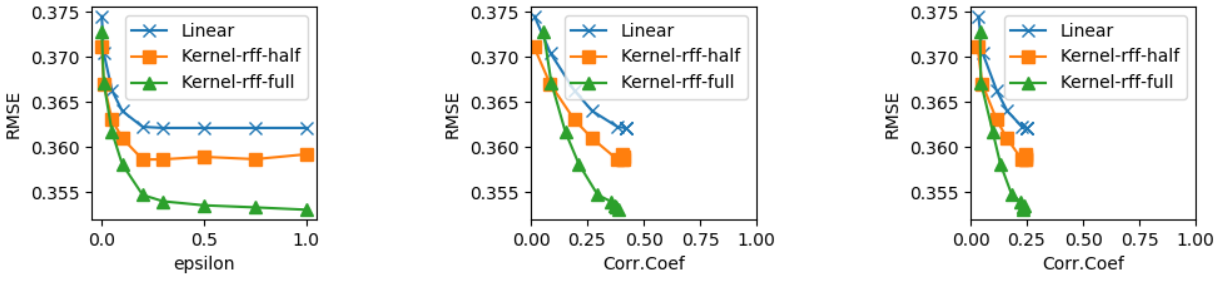
(a) C&C (ϵ-RMSE) (b) C&C (Corr. Race dist.-RMSE) (c) C&C (Corr. Origin dist.-RMSE)



(d) COMPAS (ϵ-RMSE) (e) COMPAS (Corr. Gender-RMSE) (f) COMPAS (Corr. Race-RMSE)



(g) NLSY (ϵ-RMSE) (h) NLSY (Corr. Gender-RMSE) (i) NLSY (Corr. Age-RMSE)



(j) LSAC (ϵ-RMSE) (k) LSAC (Corr. Race-RMSE) (l) LSAC (Corr. Age-RMSE)

Figure 2. Figures in the left column show the relation between RMSE as a function of the user-defined fairness level ϵ (smaller is more fair), whereas the figures in the center (resp. the right) columns show RMSE as a function of the correlation coefficient between $s^{(1)}$ (resp. $s^{(2)}$) and $\hat{y}$. "Linear" is the optimization of Eqn. (3.1) solved by the QCQP method. "Kernel-rff-full" is the the kernelized optimization where the representations $Z_s(\boldsymbol{s})$ and $Z_u(\boldsymbol{u})$ were approximated by using the random Fourier features. We made the dimensions of $Z_s(\boldsymbol{s})$ and $Z_u(\boldsymbol{u})$ ten times larger than that of the original dimension. "Kernel-rff-half" is a midway between "Linear" and "Kernel-rff-full" in which only $\boldsymbol{u}$ was kernelized (i.e., we used $\boldsymbol{s}$ and $Z_u(\boldsymbol{u})$). We also show RMSE as a function of the mean difference (MD) and the area under the curve (AUC) (Calders et al., 2013) in Appendix D, which behaved similarly to the correlation coefficients.

Datasets: The Communities and Crime (C&C) dataset combines socio-economic data and crime rate data on communities in the United States where each datapoint corresponds to a community. The target y is the normalized violent crime rate of each community and $s^{(1)}, s^{(2)}$ are the ratio of African American people and foreign-born people, respectively. The COMPAS dataset (Angwin et al., 2016) is a collection of criminal offenders screened in Florida U.S. during 2013–2014, where $\boldsymbol{x}$ is a demographic and criminal records of offenders and y is whether or not a person recidivated within two years after the screening, and $s^{(1)}, s^{(2)}$ are the gender and race, respectively. The National Longitudinal Survey of Youth (NLSY) dataset[6] involves survey results of the U.S. Bureau of Labor Statistics that is intended to gather information on the labor market activities and other life events of several groups, where y is the income of each person in 1990 and $s^{(1)}, s^{(2)}$ are the gender and age, respectively. The Law School Admissions Council (LSAC) dataset[7] is a survey among students attending law schools in the U.S. in 1991, where y indicates the GPA score of each student, and $s^{(1)}, s^{(2)}$ are the race and the age, respectively. Statistics of the datasets are in Table 1, and further details of the datasets are in Appendix C.

Evaluation settings: We split the data into 5-folds: One was for validation dataset that was used to optimize the hyperparameters, and another was for the test dataset. The resting three folds were the training dataset. All the reported results are the ones of the test dataset averaged over the 5 runs with different choices of the folds. The features $\boldsymbol{u}$ were built from $\boldsymbol{x}$ by de-correlating it from $\boldsymbol{s}$ by using regularized least square regression. The hyperparameters were optimized in validation datasets among $\lambda = \{1.0, 10.0, 100.0\}$ and $\gamma = \{0.1, 1.0, 10.0, 100.0\}$, where γ was the hyper-parameter of the RBF kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\gamma(\boldsymbol{x} - \boldsymbol{y})^2)$.

4.1. Results

Figure 2 shows the results of our simulations. The following summarizes our observation: (i) In all datasets, there was a clear tradeoff between the predictive power (i.e., RMSE) and the degree of fairness measured by ϵ. (ii) The advantage of non-linear representation varied: In C&C and NLSY, the linear method performs as good as the two kernelized methods, whereas in COMPAS and LSAC the kernel methods significantly outperformed the linear. (iii) The correlation coefficient was saturated at some point to which the predictive power of $\boldsymbol{s}$ is fully utilized. (iv) While two kernel methods performed similarily in the first three datasets, "Kernel-rff-full" significantly outperformed "Kernel-rff-half" in the LSAC dataset. In other words, the advantage of the non-linear sensitive attributes $\boldsymbol{s}$ was observed in LSAC: This dataset involved a numeric sensitive attribute (i.e., age) from which the method exploited the non-linear relationship between $\boldsymbol{s}$ and y. (v) The correlation between $s^{(l)}$ and y varied among sensitive features $\{s^{(1)}, s^{(2)}\}$: For example, in the NLSY dataset, gender was more predictive than age, and thus the regressor exploited more on the former attribute than the latter.

[6] https://www.bls.gov/nls/

[7] http://www2.law.ucla.edu/sander/Systemic/Data.htm

5. Conclusion

We have focused on the optimization perspective of fair regression. We considered CoD that is a natural extension of the correlation coefficient into multiple sensitive attributes as a fairness criterion. Although the least square regressor subject to a CoD constraint involves a nonconvex feasible region, it boils down to exactly-solvable convex optimizations. The proposed method has the following aspects: (i) The exact control on the fairness level as a constraint, (ii) a capability of dealing with numeric and multiple $\boldsymbol{s}$ and (iii) an extension that captures non-linear interaction between sensitive and non-sensitive attributes. We consider this result as a first step that controls the nonconvexity that naturally appears in considering fairness related constraints.

While the prevention of disparate impact is justified by the legal contexts, some alternative criteria of fairness, such as the equality odds condition (Hardt et al., 2016; Zafar et al., 2017b), have been proposed. Considering them as a constraint would be interesting.

Acknowledgements

The authors sincerely thank the anonymous reviewers for their useful comments. The authors are grateful to Toshihiro Kamishima for pointing out related papers. This work was supported in part by JSPS KAKENHI Grant Number 17K12736, 18K17998 and Inamori Foundation Research Grant.

References

Adebayo, Julius and Kagal, Lalana. Iterative orthogonal feature projection for diagnosing bias in black-box models. In *Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2016.

Adler, Philip, Falk, Casey, Friedler, Sorelle A., Nix, Tionney, Rybeck, Gabriel, Scheidegger, Carlos, Smith, Brandon, and Venkatasubramanian, Suresh. Auditing black-box models for indirect influence. *Knowl. Inf. Syst.*, 54 (1):95–122, 2018.

Angwin, Julia, Larson, Jeff, Mattu, Surya, and Kirchner, Lauren. *Machine Bias: Theres software used across the country to predict future criminals.* 2016. URL https://www.propublica.org/.

Ben-Tal, Aharon and Nemirovskiaei, Arkadiaei Semenovich. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001. ISBN 0-89871-491-5.

Berk, Richard, Heidari, Hoda, Jabbari, Shahin, Joseph, Matthew, Kearns, Michael, Morgenstern, Jamie, Neel, Seth, and Roth, Aaron. A convex framework for fair regression. In *Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2017.

Bolukbasi, Tolga, Chang, Kai-Wei, Zou, James Y., Saligrama, Venkatesh, and Kalai, Adam Tauman. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, pp. 4349–4357, 2016.

Bredereck, Robert, Faliszewski, Piotr, Igarashi, Ayumi, Lackner, Martin, and Skowron, Piotr. Multiwinner elections with diversity constraints. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, 2018.

Calders, Toon, Karim, Asim, Kamiran, Faisal, Ali, Wasif, and Zhang, Xiangliang. Controlling attribute effect in linear regression. In *2013 IEEE 13th International Conference on Data Mining*, pp. 71–80, 2013.

Calmon, Flávio, Wei, Dennis, Vinzamuri, Bhanukiran, Ramamurthy, Karthikeyan Natesan, and Varshney, Kush R. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 3995–4004, 2017.

Celis, L. Elisa, Keswani, Vijay, Straszak, Damian, Deshpande, Amit, Kathuria, Tarun, and Vishnoi, Nisheeth K. Fair and diverse dpp-based data summarization. *CoRR*, abs/1802.04023, 2018.

Feldman, Michael, Friedler, Sorelle A., Moeller, John, Scheidegger, Carlos, and Venkatasubramanian, Suresh. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, 2015.

Fish, Benjamin, Kun, Jeremy, and Lelkes, Ádám Dániel. Fair boosting: a case study. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2015.

Fukuchi, Kazuto, Sakuma, Jun, and Kamishima, Toshihiro. Prediction with model-based neutrality. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD*, pp. 499–514, 2013.

Goh, Gabriel, Cotter, Andrew, Gupta, Maya R., and Friedlander, Michael P. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, pp. 2415–2423, 2016.

Gretton, Arthur, Bousquet, Olivier, Smola, Alexander J., and Schölkopf, Bernhard. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory, 16th International Conference, ALT 2005, Singapore, October 8-11, 2005, Proceedings*, pp. 63–77, 2005.

Hardt, Moritz, Price, Eric, and Srebro, Nati. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, pp. 3315–3323, 2016.

Kamiran, Faisal and Calders, T. Classification with no discrimination by preferential sampling. In *The annual machine learning conference of Belgium and The Netherlands (BENELEARN)*, 01 2010.

Kamishima, Toshihiro, Akaho, Shotaro, Asoh, Hideki, and Sakuma, Jun. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge*

Discovery in Databases - European Conference, ECML PKDD, pp. 35–50, 2012a.

Kamishima, Toshihiro, Akaho, Shotaro, Asoh, Hideki, and Sakuma, Jun. Enhancement of the neutrality in recommendation. In *Proceedings of the 2nd Workshop on Human Decision Making in Recommender Systems*, pp. 8–14, 2012b.

Kamishima, Toshihiro, Akaho, Shotaro, Asoh, Hideki, and Sato, Issei. Model-based approaches for independence-enhanced recommendation. In *IEEE International Conference on Data Mining Workshops*, pp. 860–867, 2016.

Pardalos, Panos M. and Vavasis, Stephen A. Quadratic programming with one negative eigenvalue is np-hard. *J. Global Optimization*, 1(1):15–22, 1991.

Pérez-Suay, Adrián, Laparra, Valero, Mateo-Garcia, Gonzalo, Muñoz-Marí, Jordi, Gómez-Chova, Luis, and Camps-Valls, Gustau. Fair kernel learning. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18-22, 2017, Proceedings, Part I*, pp. 339–355, 2017.

Rahimi, Ali and Recht, Benjamin. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems*, pp. 1313–1320, 2008.

Rasmussen, Carl Edward and Williams, Christopher K. I. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006. ISBN 026218253X. URL http://www.worldcat.org/oclc/61285753.

Ristanoski, Goce, Liu, Wei, and Bailey, James. Discrimination aware classification for imbalanced datasets. In *22nd ACM International Conference on Information and Knowledge Management*, pp. 1529–1532, 2013.

Ruggieri, Salvatore, Pedreschi, Dino, and Turini, Franco. Data mining for discrimination discovery. *TKDD*, 4(2): 9:1–9:40, 2010.

Sturm, Jos F. and Zhang, Shuzhong. On cones of nonnegative quadratic functions. *Math. Oper. Res.*, 28(2):246–267, May 2003.

Wooldridge, Jeffrey M. *Introductory Econometrics: A Modern Approach*. South-Western, Cengage Learning, 5th edition, 2013. ISBN 978-1-111-53104-1.

Yamada, Shinji and Takeda, Akiko. Successive lagrangian relaxation algorithm for nonconvex quadratic optimization. *Journal of Global Optimization*, 71(2):313–339, Jun 2018.

Zafar, Muhammad Bilal, Valera, Isabel, Gomez-Rodriguez, Manuel, and Gummadi, Krishna P. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 962–970, 2017a.

Zafar, Muhammad Bilal, Valera, Isabel, Gomez-Rodriguez, Manuel, and Gummadi, Krishna P. 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, 2017b.

Zemel, Richard S., Wu, Yu, Swersky, Kevin, Pitassi, Toniann, and Dwork, Cynthia. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 325–333, 2013.

Zliobaite, Indre, Kamiran, Faisal, and Calders, Toon. Handling conditional discrimination. In *11th IEEE International Conference on Data Mining*, pp. 992–1001, 2011.