Mean Parity Fair Regression in RKHS

Shaokui Wei
Shenzhen Research Institute of Big Data
The Chinese University of Hong Kong, Shenzhen

Jiayin Liu
School of Management and Economics
The Chinese University of Hong Kong, Shenzhen

Bing Li
Department of Statistics
Pennsylvania State University

Hongyuan Zha
School of Data Science
The Chinese University of Hong Kong, Shenzhen

Abstract

We study the fair regression problem under the notion of Mean Parity (MP) fairness, which requires the conditional mean of the learned function output to be constant with respect to the sensitive attributes. We address this problem by leveraging reproducing kernel Hilbert space (RKHS) to construct the functional space whose members are guaranteed to satisfy the fairness constraints. The proposed functional space suggests a closed-form solution for the fair regression problem that is naturally compatible with multiple sensitive attributes. Furthermore, by formulating the fairness-accuracy tradeoff as a relaxed fair regression problem, we derive a corresponding regression function that can be implemented efficiently and provides interpretable tradeoffs. More importantly, under some mild assumptions, the proposed method can be applied to regression problems with a covariance-based notion of fairness. Experimental results on benchmark datasets show the proposed methods achieve competitive and even superior performance compared with several state-of-the-art methods.

1 INTRODUCTION

As Machine Learning (ML) algorithms have been increasingly applied to solve real-world problems, such as employment (Kodiyan, 2019), finance (Anshari et al., 2021), and healthcare (Gupta and Mohammad, 2017), the biases exhibited by ML are attracting attention from both industry and academia. Algorithmic fairness has therefore emerged as a new frontier for ML, of which the critical challenge is to design algorithms satisfying fairness constraints, thus mitigating or eliminating the potential discrimination on the basis of legally protected (sensitive) attributes such as race or gender. In recent years, substantial efforts on notions and algorithms of fairness in ML have generally centered on classification problems (Agarwal et al., 2018; Calders and Verwer, 2010; Huang and Vishnoi, 2019; Jiang et al., 2020; Zafar et al., 2019), while the problems of fair regression have received much less attention.

In this paper, we focus on the general regression problem in reproducing kernel Hilbert spaces (RKHS) and propose a novel approach for fair regression by constructing the space of functions that satisfy the constraints of fairness. Specifically, we consider the unfairness in the mean responses across different groups. Such unfairness exists broadly in many real-life problems including wage/payment gap (Oettinger, 1996; Barroso and Brown, 2021), employment inequality (Center, 2016) and educational inequality (Darling-Hammond, 1998; Baker et al., 2014). To mitigate such unfairness, we adopt the Mean Parity (MP) fairness, a notion of group fairness aiming to achieve "equality on average", i.e., the average response of a regression function to the different groups is the same.

By establishing the connection between the covariance operator and MP fairness, we show that the MP-fair functional space can be characterized by a set of orthonormal bases and derive a closed-form solution that minimizes the mean squared error (MSE). Under some mild assumptions, the proposed method can also be applied to regression problems subject to fairness criterion that urges the outcome of the regression function to be uncorrelated with the sensitive attributes. In addition, the proposed method is naturally compatible with multiple sensitive attributes and can be extended to a broad range of loss functions for regression using optimization techniques, e.g., gradient descent.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

As it has been empirically observed that the fair model may suffer from a reduction in accuracy (Berk et al., 2017; Tan et al., 2020), we further generalize our method to consider the tradeoff between fairness and accuracy. By formulating the fairness-accuracy tradeoff as a relaxed fair regression problem, we derive a closed-form solution which is a simple combination of the optimal fair solution and the optimal least-squares solution, controlled by a single parameter. The proposed relaxed solution allows users to quantify and control the cost of fairness in terms of MSE and enjoys good interpretability. Finally, we evaluate our methods on three real datasets and one synthetic dataset. The experimental results demonstrate that our solution can eliminate the discrimination in train data and effectively enforce fairness in test data. Also, experiments on the fairness-accuracy tradeoff show that our method performs on par with other approaches and provides precise control over MSE and fairness levels.

Paper organization. The rest of the paper is organized as follows. Section 2 introduces notations and the formulation of our problem. In Section 3, we study the characterization of fair functional space in RKHS and provide a functional solution to the fair regression problem, after which we discuss the tradeoff between fairness and accuracy. Section 4 presents some empirical evaluations of our methods. We discuss some related works in Section 5 and end with some conclusions and future directions in Section 6. The proofs, derivations, implementation details, and some additional experiments are left in Appendix.

2 PRELIMINARIES

2.1 Notations

We first introduce some important notations and a more comprehensive table of notations can be found in Appendix F. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We consider the random variables X and Y defined on measurable spaces $(\Omega_X, \mathcal{F}_X)$ and $(\Omega_Y, \mathcal{F}_Y)$ where Ω_Y is a subset of $\mathbb{R}$ and $\mathcal{F}_Y$ is the Borel σ-filed on Ω_Y. Let $\Omega_S = \{s^{(j)}\}_{j=1}^k$ be a finite set of k elements from which a random variable S takes values. We set X, S and Y to be the random variables for non-sensitive attributes, sensitive attributes and label/response respectively. In addition, we assume that $\mathbb{P}(s) > 0$ for all $s \in \Omega_S$.

Let $\kappa_X : \Omega_X \times \Omega_X \to \mathbb{R}$ be a universal kernel and $\kappa_S : \Omega_S \times \Omega_S \to \mathbb{R}$ be a discrete kernel. We use $\mathcal{H}_X$ to represent the RKHS generated by κ_X and denote its feature map by $\phi_X : \Omega_X \to \mathcal{H}_X$, i.e., $\phi_X(x) = \kappa_X(\cdot, x)$. Similarly, let $\mathcal{H}_S$ be the RKHS generated by κ_S with feature map ϕ_S. Let $\mathcal{H}_{XS}$ be the RKHS generated by the kernel κ_{XS} defined on $\Omega_{XS} \times \Omega_{XS}$ where $\Omega_{XS} = \Omega_X \times \Omega_S$. Then, each member of $\mathcal{H}_{XS}$ is a function $g(x, s)$ where $x \in \Omega_X, s \in \Omega_S$, and we denote the feature map of $\mathcal{H}_{XS}$ by ϕ_{XS}. By the reproducing property of $\mathcal{H}_{XS}$, evaluating a function $g \in \mathcal{H}_{XS}$ at (x, s) can be written as

$$g(x, s) = \langle \phi_{XS}(x, s), g \rangle_{\mathcal{H}_{XS}},$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}_{XS}}$ is the inner product in $\mathcal{H}_{XS}$. For a space $M \subseteq \mathcal{H}_{XS}$, we denote its orthogonal complement in $\mathcal{H}_{XS}$ by $M^\perp$ such that $\mathcal{H}_{XS} = M \oplus M^\perp$. Moreover, let $\perp\!\!\!\perp$ represent the independence between random variables.

2.2 Notions of fairness

Our goal is to find the optimal fair regression function in $\mathcal{H}_{XS}$ that minimizes the mean squared error while maintaining fairness. For a function $g \in \mathcal{H}_{XS}$, we consider the Mean Parity [1] fairness, as defined below:

Definition 1 (Mean Parity)**.** *The subset $\mathcal{G}_{MP}$ of $\mathcal{H}_{XS}$ defined by*

$$\mathcal{G}_{MP} = \{g \in \mathcal{H}_{XS} : \mathbb{E}[g(X, S)|S] = \mathbb{E}[g(X, S)]\}$$

is called the Mean Parity fair (MP-fair) class of functions.

The above definition says that a function g is MP fair if the expectation of $g(X, S)$ conditioning on S is constant across all sensitive groups.

Besides MP-fairness, there are several other ways of defining fairness. Here, we highlight the connection and distinction between MP fairness and the other two notions of fairness.

Demographic Parity (DP) fairness. A popular requirement for fairness is $g(X, S) \perp\!\!\!\perp S$, i.e., the distribution of $g(X, S)$ conditioning on S is the same, and the class of such functions is called Demographic Parity fair class (Feldman et al., 2015).

To establish the relationship between MP fairness and DP fairness, we provide the following proposition:

Proposition 1. *Assume that the DP disparity (DPD) and the MP disparity (MPD) of function g are measured by*

$$DPD(g) = \sum_{s \in \Omega_S} \mathcal{W}_1(g(X, S)|S = s, g(X, S))$$

$$MPD(g) = \sum_{s \in \Omega_S} |\mathbb{E}(g(X, S)|S = s) - \mathbb{E}(g(X, S))|$$

where $\mathcal{W}_1$ is the 1-Wasserstein distance (Frohmader and Volkmer, 2021) .

Then,

$$MPD(g) \leq DPD(g).$$

[1] also known as Mean Difference (Calders et al., 2013; Žliobaitė, 2017), Mean Distance (Komiyama and Shimao, 2017) or Discrimination Score (Calders and Verwer, 2010; Zemel et al., 2013; Raff et al., 2018).

Therefore, MP disparity is the lower bound of DP disparity and achieving MP fairness is necessary to achieve DP fairness. Moreover, for binary classification problem with a binary S, MP fairness is equivalent to DP fairness.

Covariance based (CB) fairness. Another widely used condition for fairness is the Covariance based (CB) fairness (Komiyama et al., 2018; Mary et al., 2019; Pérez-Suay et al., 2017; Scutari et al., 2021), which requires the output of g to be uncorrelated with the sensitive attribute, i.e., $\mathrm{Cov}(g(X,S),S)=0$.

By the definition of covariance, we can conclude that MP fairness implies that $g(X,S)$ is uncorrelated with S. Thus, an MP-fair regression function is always CB-fair. Moreover, MP fairness is equivalent to CB fairness under some assumptions, which will be discussed later.

2.3 Problem formulation

Now, we introduce the formulation for the MP-fair regression problem. Consider the general regression model

$$Y = g(X,S) + \epsilon,$$

where $g \in \mathcal{H}_{XS}$ and X, S are independent of the centered random noise $\epsilon \in \mathbb{R}$ and $\mathbb{E}(Y^2) \leq \infty$.

Then, we focus on the least-squares MP-fair regression task formulated as a constrained optimization problem

$$\begin{aligned} \min_{g} \quad & \mathbb{E}\left(Y - g(X,S)\right)^2 \\ \text{s.t.} \quad & g \in \mathcal{G}_{MP}. \end{aligned} \tag{1}$$

3 FAIR REGRESSION UNDER MEAN PARITY

In this section, we discuss how to solve Problem 1. To do so, we first develop a theory to characterize the MP-fair class $\mathcal{G}_{MP}$ within $\mathcal{H}_{XS}$. After that, we derive a closed-form solution by employing a projection operator P from $\mathcal{H}_{XS}$ onto $\mathcal{G}_{MP}$ and introduce a formulation to control the fairness-accuracy tradeoff. At last, we discuss the performance guarantees of the derived solution and how to solve the MP-fair regression problem with other loss functions.

3.1 Characterization of MP-fair function space

We begin by introducing some concepts. For Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$ and a linear operator $A: \mathcal{H}_1 \to \mathcal{H}_2$, we define the kernel of A by $\ker(A) = \{f \in \mathcal{H}_1 : Af = 0_{\mathcal{H}_2}\}$ where $0_{\mathcal{H}_2}$ is the zero function in $\mathcal{H}_2$. Let $\mathrm{ran}(A)$ represent the set $\{Af : f \in \mathcal{H}_1\}$, which is the range of A. Let μ_{XS} be the kernel mean embedding of (X,S) in $\mathcal{H}_{XS}$, which is defined as $\mu_{XS} = \mathbb{E}\left[\kappa_{XS}\left(\cdot,(X,S)\right)\right]$. Similarly, let $\mu_S = \mathbb{E}\left[\kappa_S(\cdot,S)\right]$ be the kernel mean embedding of S in $\mathcal{H}_S$. Then, we define the covariance operator between S and (X,S) as

$$\Sigma_{S(XS)} = \mathbb{E}\left[(\phi_S(S) - \mu_S) \otimes (\phi_{XS}(X,S) - \mu_{XS})\right],$$

where $\otimes$ represents the outer product in RKHS.

To characterize $\mathcal{G}_{MP}$, we present the following assumption.

Assumption 1. *Assume that the following system of equations*

$$\sum_{j=1}^{k} \eta_j(\phi_S(s^{(j)}) - \mu_S) = 0_{\mathcal{H}_S}, \quad \sum_{j=1}^{k} \eta_j = 0 \tag{2}$$

has exactly one solution, i.e., $\eta_j = 0$ for all $j \in \{1,\ldots,k\}$.

Note that the choice of κ_S can be independent of κ_{XS} and κ_X. Since the cardinality of Ω_S is finite, Assumption 1 is quite mild. A typical choice for κ_S is to have linearly independent features $\{\phi_S(s^{(j)})\}_{j=1}^k$. For example, a polynomial kernel with degree $k-1$ would satisfy Assumption 1 for $\Omega_S \subset \mathbb{R}$. The proof is given in Appendix C.4.

Then, the following theorem provides insight into the characterization of $\mathcal{G}_{MP}$.

Theorem 1. *Under Assumption 1, $\mathcal{G}_{MP}$ is the kernel of the operator $\Sigma_{S(XS)}$, that is,*

$$\mathcal{G}_{MP} = \ker(\Sigma_{S(XS)}).$$

Based on Theorem 1, $\mathcal{G}_{MP}$ can be found using the relation

$$\ker(\Sigma_{S(XS)}) = \mathrm{ran}(\Sigma_{(XS)S})^{\perp},$$

where $\mathrm{ran}(\Sigma_{(XS)S})$ can be characterized by the generalized eigenvalue problem (Hoegaerts et al., 2005; Schölkopf et al., 1998; Yuan and Cai, 2010).

Since $\mathcal{H}_S$ has finite dimension, $\Sigma_{(XS)S}$ is a finite rank operator. Let us say its rank is $m \leq k$. Let $A: \mathcal{H}_S \to \mathcal{H}_S$ be any positive definite linear operator. Then, the first m eigenfunctions of $\Sigma_{(XS)S} A \Sigma_{S(XS)}$, say, $\theta_1,\ldots,\theta_m$, span $\mathrm{ran}(\Sigma_{(XS)S})$, that is,

$$\ker(\Sigma_{S(XS)}) = \mathrm{span}(\{\theta_1,\ldots,\theta_m\})^{\perp}.$$

Thus, $\mathcal{G}_{MP}$ can be characterized by a set of eigenfunctions $\{\theta_1,\ldots,\theta_m\}$ which allows us to construct an orthogonal projection operator P from $\mathcal{H}_{XS}$ to $\mathcal{G}_{MP}$.

Denote a set of orthonormal bases of $\mathrm{ran}(\Sigma_{(XS)S})$ by $\{\theta_1',\cdots,\theta_m'\}$. Given a function $g \in \mathcal{H}_{XS}$, the orthogonal projection operator from $\mathcal{H}_{XS}$ onto $\mathrm{ran}(\Sigma_{(XS)S})^{\perp}$ eliminates the components of g in $\mathrm{ran}(\Sigma_{(XS)S})$. Thus, we can construct the following orthogonal projection operator

$$P = I - \sum_{j=1}^{m} \theta_j' \otimes \theta_j',$$

where $I: \mathcal{H}_{XS} \to \mathcal{H}_{XS}$ is the identity operator.

For simplicity, the detailed process to estimate P from a given dataset is left to Appendix B.

Remark. Consider the general CB fairness that seeks to remove the correlation between the sensitive feature map used for prediction and the predicted value. By the observation that $\text{Cov}(\phi_S(S), g(X,S)) = \Sigma_{S(XS)}g$, $\ker(\Sigma_{S(XS)})$ is the space whose members are CB fair under the assumption that κ_{XS} is composed of κ_S and κ_X, e.g., $\kappa_{XS} = \kappa_X + \kappa_S$. Therefore, the results in the rest of this paper, except the interpretation of tradeoffs, can also be applied to fair regression with CB constraints. The detailed discussion is left to Appendix C.1. In particular, if both the above assumption and Assumption 1 are satisfied in $\mathcal{H}_{XS}$, MP fairness is equivalent to CB fairness.

3.2 Optimal fair regression function

To find the optimal fair regression function, we introduce the optimality condition for Problem 1.

Lemma 1. *A function $g^*_{\mathcal{G}}$ is an optimal solution for Problem 1 if and only if*

$$\mathbb{E}(Yg(X,S)) = \mathbb{E}(g(X,S)g^*_{\mathcal{G}}(X,S)) \quad \forall g \in \mathcal{G}_{MP}.$$

By introducing the uncentralized covariance operator $\tilde{\Sigma}_{(XS)(XS)} = \mathbb{E}\left[(\phi_{XS}(X,S) \otimes (\phi_{XS}(X,S))\right]$ and a function $h = \mathbb{E}(\phi_{XS}(X,S)Y)$, Lemma 1 tells us that $g^*_{\mathcal{G}}$ is an optimal solution for Problem 1 if and only if

$$\langle h, g\rangle_{\mathcal{H}_{XS}} = \langle \tilde{\Sigma}_{(XS)(XS)}g^*_{\mathcal{G}}, g\rangle_{\mathcal{H}_{XS}} \quad \forall g \in \mathcal{G}_{MP}.$$

Given an orthogonal projection operator P from $\mathcal{H}_{XS}$ to $\mathcal{G}_{MP}$, a key insight is that $g^*_{\mathcal{G}} = Pg^*_{\mathcal{H}}$ where $g^*_{\mathcal{H}}$ can be obtained by solving the following problem

$$\langle Ph, g\rangle_{\mathcal{H}_{XS}} = \langle P\tilde{\Sigma}_{(XS)(XS)}Pg^*_{\mathcal{H}}, g\rangle_{\mathcal{H}_{XS}} \quad \forall g \in \mathcal{H}_{XS},$$

So, we reach the Proposition 2.

Proposition 2. *The optimal MP-fair regression function to Problem 1 is*

$$g^*_{\mathcal{G}} = P[P\tilde{\Sigma}_{(XS)(XS)}P]^{\dagger}Ph, \tag{3}$$

where $(\cdot)^{\dagger}$ is the Moore-Penrose Inverse of an operator (Groetsch, 1977; Wang et al., 2018).

Note that if P is an identity operator, the solution 3 reduces to $g^*_{\mathcal{G}} = [\tilde{\Sigma}_{(XS)(XS)}]^{\dagger}h$, which is the least-squares regression function in $\mathcal{H}_{XS}$.

3.3 Tradeoff between accuracy and fairness

There are multiple ways of relaxing the MP-fair constraint to control the accuracy-fairness tradeoff. One group of relaxed constraints is imposed on the overall unfairness, e.g., $\text{MPD}(g) \leq \beta$ or $\|\Sigma_{S(XS)}g\|_{\mathcal{H}_S} \leq \beta$ for some positive real number β, but such constraints ignore the unfairness for individual group, which weakens their interpretability. Another group of relaxed constraints is imposed on each sensitive group, from which we employ the following relaxed constraint

$$\mathbb{E}(g(X,S)|S) - \mathbb{E}(g(X,S)) = \alpha\left[\mathbb{E}(g^*(X,S)|S) - \mathbb{E}(g^*(X,S))\right] \tag{4}$$

where g^* is the least-squares regression function in $\mathcal{H}_{XS}$ and $\alpha \in [0,1]$ is a scalar to control the level of unfairness. A larger α results in a higher level of unfairness and g is MP-fair if $\alpha = 0$. Thus, the constraint 4 allows us to scale the unfairness of the least-squares regression function for each group by a scalar α and therefore provides good interpretability. More importantly, we will show that constraint 4 provides precise control of the accuracy-fairness tradeoff later.

To move forward, we present the immediate corollary from Theorem 1.

Corollary 1. *Given $g_1, g_2 \in \mathcal{H}_{XS}$, under Assumption 1, $\mathbb{E}(g_1(X,S)|S) - \mathbb{E}(g_1(X,S)) = \mathbb{E}(g_2(X,S)|S) - \mathbb{E}(g_2(X,S))$ if and only if $\Sigma_{S(XS)}g_1 = \Sigma_{S(XS)}g_2$.*

By Corollary 1, it suffices to consider the following relaxed fair regression problem

$$\begin{aligned} \min \quad & \mathbb{E}(Y - g(X,S))^2 \\ \text{s.t.} \quad & \Sigma_{S(XS)}g = \alpha\Sigma_{S(XS)}g^*. \end{aligned} \tag{5}$$

As $\mathcal{H}_{XS} = \mathcal{G}_{MP} \oplus \mathcal{G}^{\perp}_{MP}$, a function $g \in \mathcal{H}_{XS}$ can be written as $g = g_{MP} + g_{MP^{\perp}}$ where $g_{MP} \in \mathcal{G}_{MP}$ and $g_{MP^{\perp}} \in \mathcal{G}^{\perp}_{MP}$. Then, the following proposition is the key to solving the Problem 5.

Proposition 3. *A function $g \in \mathcal{H}_{XS}$ satisfies $\Sigma_{S(XS)}g = \Sigma_{S(XS)}g^*$ if and only if $g_{MP^{\perp}} = g^*_{MP^{\perp}}$.*

By Proposition 3, the optimal solution g^{α} for Problem 5 is of the form $g^{\alpha} = g^{\alpha}_{MP} + \alpha g^*_{MP^{\perp}}$, where g^{α}_{MP} is the optimal solution to the following fair regression problem

$$\min_{g \in \mathcal{G}_{MP}} \mathbb{E}\left(Y - \alpha g^*_{MP^{\perp}}(X,S) - g(X,S)\right)^2. \tag{6}$$

Solving Problem 6 gives the following proposition.

Proposition 4. *The optimal solution of Problem 5 is $g^{\alpha} = (1-\alpha)g^*_{\mathcal{G}} + \alpha g^*$*

Let $L(g) = \mathbb{E}((Y - g(X,S))^2)$. By Proposition 4, the following equations allow us to precisely control the tradeoff between fairness and accuracy

$$\begin{aligned} L(g^{\alpha}) &= (1-\alpha)^2 L(g^*_{\mathcal{G}}) + (1-(1-\alpha)^2)L(g^*) \\ \text{MPD}(g^{\alpha}) &= \alpha\text{MPD}(g^*). \end{aligned}$$

Remark. The detailed derivation for this subsection can be found in Appendix C.2. When $\text{MPD}(g^*) > 0$, the above equations indicate that $L(g^{\alpha})$ is a quadratic function of $\text{MPD}(g^{\alpha})$.

3.4 Performance guarantee

Besides the explicit expression, the optimal regression function $g^*_{\mathcal{G}}$ also enjoys a theoretical performance guarantee with respect to MSE.

Proposition 5. *Under Assumption 1, the MSE of $g^*_{\mathcal{G}}$ is bounded by*

$$L(g^*_{\mathcal{G}}) \leq L(g^*) + \langle \tilde{\Sigma}_{(XS)(XS)} g^*_{MP^\perp}, g^*_{MP^\perp} \rangle_{\mathcal{H}_{XS}},$$

where g^* is the optimal regression function in $\mathcal{H}_{XS}$.

In Proposition 5, the inequality can be obtained by introducing a non-optimal fair regression function Pg^*. Note that $g = 0_{\mathcal{H}_{XS}}$ is always a fair regression, so we can claim that $L(g^*_{\mathcal{G}}) \leq \mathbb{E}(Y^2)$. Since the term $\langle \tilde{\Sigma}_{(XS)(XS)} g^*_{MP^\perp}, g^*_{MP^\perp} \rangle_{\mathcal{H}_{XS}}$ measures the violation of fairness constraints by g^*, Proposition 5 shows that the MSE of fair regression function is bounded and the upper bound is related to the unfairness level of g^*.

3.5 Extension

So far we only consider the fair regression with squared loss function. However, the proposed method can also be applied to other differentiable loss functions in practice. Given a differentiable loss function ℓ and the training dataset $\mathcal{D} = \{x_i, s_i, y_i\}_{i=1}^n$, we consider the following fair regression problem

$$\hat{g}^*_{\mathcal{G}} = \arg\min_{g \in \mathcal{G}_{MP}} \sum_i \ell(y_i, g(x_i, s_i)). \tag{7}$$

By the Representer theorem (Schölkopf et al., 2001), the above problem is to find $\boldsymbol{w}^*_{\mathcal{G}}$ subject to $\Phi_{XS} \boldsymbol{w}^*_{\mathcal{G}} \in \mathcal{G}_{MP}$ that minimizes the following objective function

$$J(\boldsymbol{w}) = \sum_{i=1}^n \ell\left(y_i, \langle \phi_{XS}(x_i, s_i), \Phi_{XS} \boldsymbol{w} \rangle_{\mathcal{H}_{XS}}\right),$$

where Φ_{XS} is the feature matrix of the training data. Given an estimated projection operator $\hat{P}$, we can first find

$$\boldsymbol{w}_{\mathcal{H}} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^n \ell(y_i, \langle \phi_{XS}(x_i, s_i), \hat{P}\Phi_{XS} \boldsymbol{w} \rangle_{\mathcal{H}_{XS}})$$

by optimization techniques, e.g., gradient descent.

Then, the solution to Problem 7 is

$$\hat{g}^*_{\mathcal{G}} = \hat{P}\Phi_{XS} \boldsymbol{w}_{\mathcal{H}}.$$

4 EXPERIMENTS

We adapt the experiment settings in Agarwal et al. (2019) to evaluate the proposed method on simulated and real-world datasets. The datasets are summarized below:

Synthetic dataset has n data points $\{(x_i, s_i, y_i)\}_{i=1}^n$ with d-dimension non-sensitive attributes and e-dimension sensitive attributes. Specifically, we first generate $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_{d\times d})$, $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}_{d+e}, \boldsymbol{I}_{(d+e)\times(d+e)})$ and $\epsilon_i \sim \mathcal{N}(0, \rho^2_{noise})$. Then, $\boldsymbol{s}_i$ is sampled uniformly at random from $\{0.1, -0.1\}^e$. Next, we set $y_i = [\boldsymbol{x}_i, \boldsymbol{s}_i]^T \boldsymbol{w} + \epsilon_i$ for linear regression and $y_i = \sin([\boldsymbol{x}_i, \boldsymbol{s}_i]^T \boldsymbol{w}) + \epsilon_i$ for nonlinear regression (kernel regression case).

Adult dataset (Kohavi et al., 1996) has 48,842 samples with 14 attributes. We aim to predict the probability that an individual's income exceeds \$50k per year while we keep gender as the sensitive attribute. Our experiments evaluate all methods on a subset of the Adult dataset with 2,000 random samples.

Law School dataset (Wightman, 1998) refers to the Law School Admissions Council's National Longitudinal Bar Passage Study with 20,649 samples. We aim is to predict a student's GPA (normalized to [0, 1]) while we keep race as the sensitive attribute. We convert the original race attributes to a single binary attribute, i.e., white or non-white. Our experiments evaluate all methods on a subset of the Law School dataset with 2,000 random samples.

Communities & Crime (C&C) dataset (Redmond and Baveja, 2002) combines socio-economic, law enforcement, and crime data about communities in the US with 1,994 samples. We aim to predict the number of violent crimes per 100,000 population (normalized to [0, 1]) while we keep race as the sensitive attribute (whether the majority population of the community is white).

In all experiments, we measure the loss of function g by the empirical MSE and the MP disparity by the sum of absolute mean difference (SMD) which is the empirical estimation of MPD(g) as defined below

$$\text{SMD}(g) = \sum_{j=1}^k \left| \frac{\sum_{i=1}^n g(x_i, s_i) \mathbb{I}(s_i = s^{(j)})}{\sum_{i=1}^n \mathbb{I}(s_i = s^{(j)})} - \frac{\sum_{i=1}^n g(x_i, s_i)}{n} \right|,$$

where $\mathbb{I}(\cdot)$ is the indicator function.

For all datasets, we split the data into two parts, i.e., 80% for training and 20% for testing. We discuss the experiments on MP fairness in this section and postpone experiments on CB fairness, DP fairness and regression with other loss functions to Appendix E. The code is available at https://github.com/shawkui/MP_Fair_Regression.

4.1 Regression with single binary sensitive attribute

We first consider regression with single binary sensitive attribute. We claim that MP fairness is equivalent to CB fairness in this setting with proof in Appendix A.7, which allows us to compare the proposed method against the state-of-the-art (SOTA) CB-fair algorithms for regression. Specifically, we compare our method with the ordinary least squares method (OLS), Fair Penalty Regression method

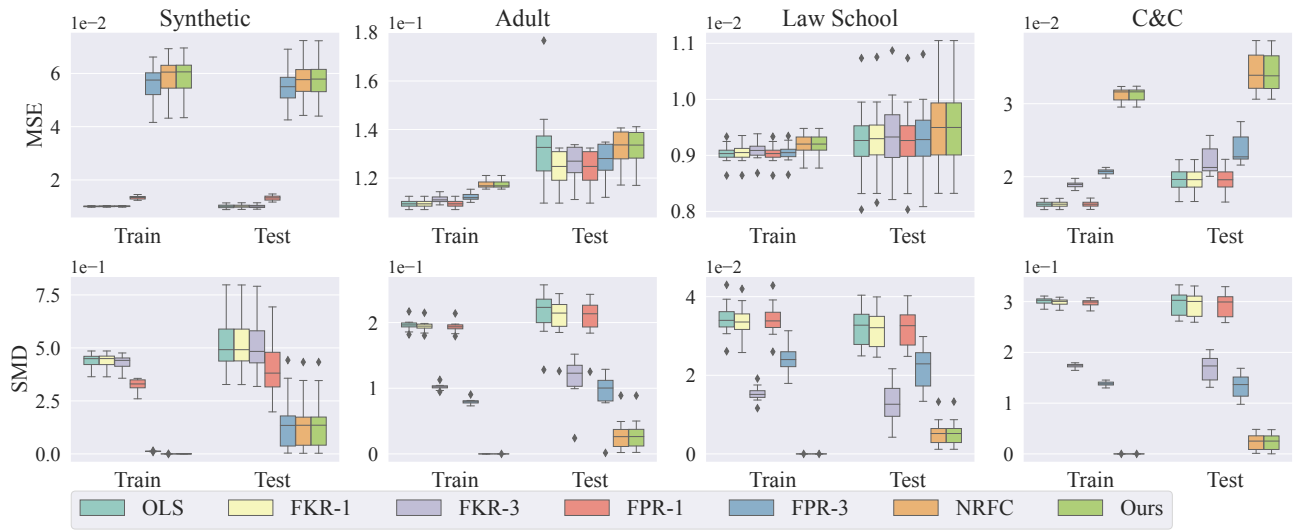

Figure 1: Results of linear regression with single binary sensitive attribute. Figures in the first row show the MSE of different methods, whereas the figures in the second row show the SMD of different methods. The legends FKR-1 and FKR-3 stand for FKR method with regularizer coefficients 10 and 1,000 respectively. Similarly, FPR-1 and FPR-3 stand for FPR method with regularizer coefficients 10 and 1,000 respectively. We also show the experiment results for kernel regression in Appendix E.1.

(FPR), Fair Kernel Learning method (FKR, Pérez-Suay et al. (2017)), and Nonconvex Regression with Fairness Constraints method (NRFC, Komiyama et al. (2018)) in terms of MSE and SMD, where FKR and NRFC are the SOTA algorithms designed for CB fairness. For regularization-based methods, i.e., FPR and FKR, we evaluate them twice with regularization coefficients 10 (FPR-1, FKR-1) and $1,000$ (FPR-3, FKR-3) respectively. More details of the baselines and experiment settings can be found in Appendix D.1.

The experiment results are summarized in Figure 1, from which we see that the proposed method can consistently enforce the MP-fair constraint, and its performance is superior to regularization-based methods and competitive with NRFC. Notably, our method can completely remove the algorithmic discrimination on conditional mean for train data. Supplemental Figure 5 shows our method achieves a smaller MSE than NRFC in kernel regression when both of them reach MP-fairness in train data.

4.2 Tradeoff between fairness and accuracy

We now test the proposed method in Section 3.3 on controlling the accuracy-fairness tradeoff, following the setting in Section 4.1.

Note that different baselines adopt different metrics and notions for such tradeoff and we only evaluate them in terms of MSE and SMD. For this purpose, we test the regularization-based methods with fairness regularizer coefficients from 0 to 10^6 while for NRFC and the proposed method, we evaluate them with the fairness level parameters from 0 to 1.

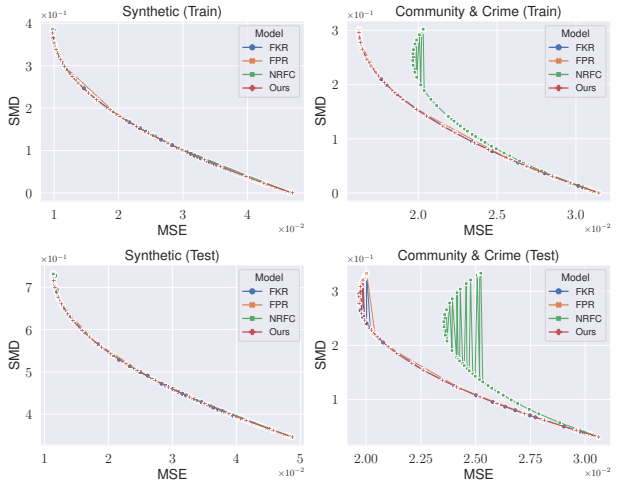

Figure 2: Results of the fairness-accuracy tradeoff. The first row presents the experiment results for train data whereas the second row shows the experiment results for test data.

The curves of the fairness-accuracy tradeoff are shown in Figure 2. As discussed in Section 3.3, the MSE climbs when stricter fairness constraints are imposed. In Figure 2, the curve of our method coincides with the curves of FKR and FPR, and performs better than the curve of NRFC. When SMD is approaching 0, all methods receive almost the same MSE while NRFC has a higher MSE than other methods on the Communities & Crime dataset when weaker fairness constraints are imposed. A similar pattern can be found

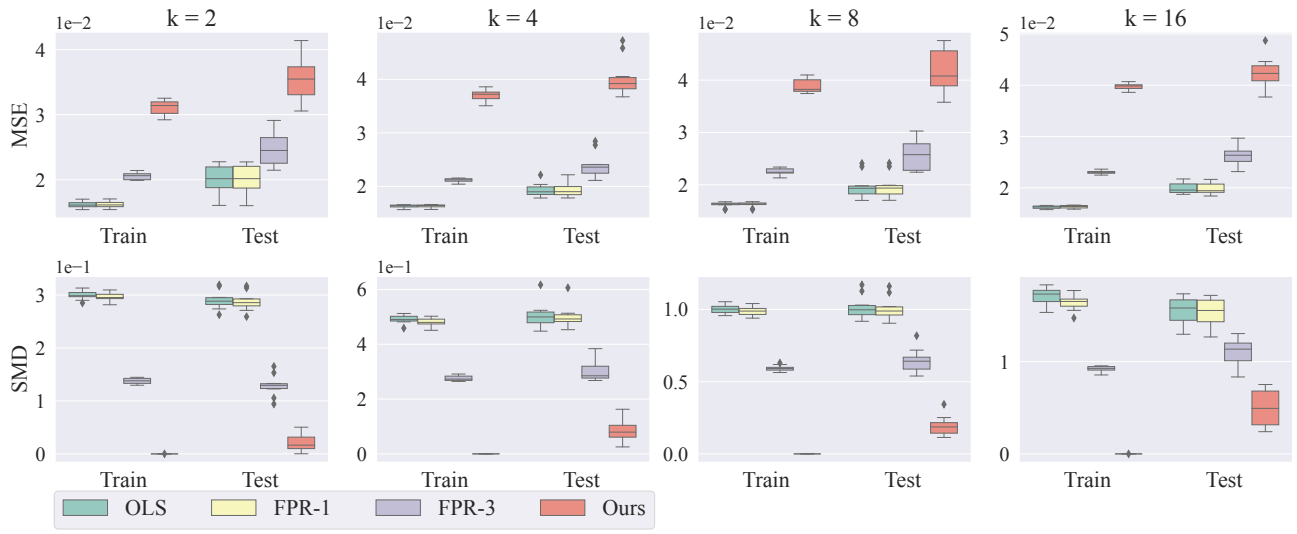

Figure 3: Results of linear regression on Communities & Crime dataset with multiple binary sensitive attributes. Figures in the first row show the MSE of different methods whereas the figures in the second row show the SMD of different methods. The experiment on kernel regression shows similar results in Appendix E.1.

in supplemental Figure 7 but NRFC and FKR achieve a slightly smaller test MSE sometimes. Although the curves are similar, our method enjoys better explainability and much lower complexity. Unlike other methods which need to solve the regression problem for each level of fairness, our method only solves the regression problem twice and produces a precise tradeoff between fairness and accuracy.

4.3 Regression with multiple sensitive attributes

As aforementioned, our method can be naturally generalized to regression with multiple sensitive attributes as long as κ_S satisfies Assumption 1. In this experiment, we set κ_S to be a polynomial kernel and choose multiple binary sensitive attributes on the Communities & Crime dataset. The number of sensitive groups is $k = 2^r$ where r is the number of binary sensitive attributes.

In this case, we consider only two baselines: FPR and the OLS since MP-fairness may be not equivalent to CB fairness. Figure 3 depicts the MSE and SMD for different numbers of sensitive attributes, from which we can see that our method can enforce fairness with different numbers of sensitive attributes.

4.4 Distribution of MP-fair response

In this section, we visualize the distribution of response Y and the predicted response $\hat{Y}$ produced by our method on MP-fair regression problem to demonstrate the effect of MP-fairness. Specifically, we consider linear regression with single binary sensitive attribute $S \in \{0, 1\}$. Note that to test our method on an extreme case, the synthetic dataset is generated following the linear regression setting with sensitive attribute drawn from $\{-10, 10\}$ uniformly so that the distributions of response in two groups are significantly different.

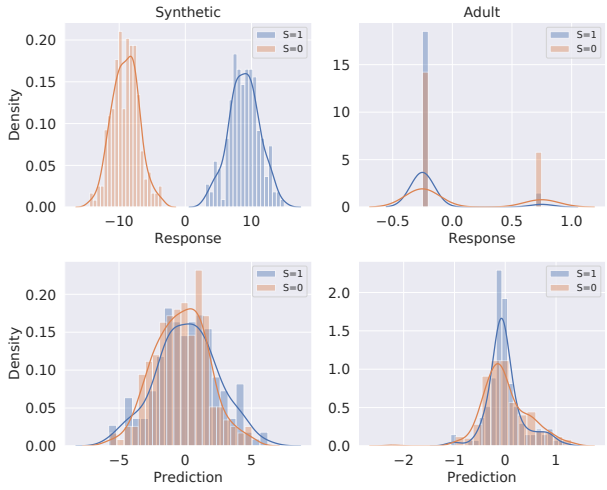

Figure 4: Visualization of centralized response distribution. Both the normalized histograms (bins) and the estimated density (curves) are reported. Figures in the first row show the conditional distribution of response in the test dataset while the figures in the second row show the corresponding conditional distribution of the MP-fair predicted response.

The results of the Synthetic test dataset and the Adult test dataset are summarized in Figure 4, from which we can see that the distribution of $\hat{Y}$ conditioning on the sensitive attribute S are similar to each other. This observation agrees with the experiment results in Appendix E.4 which says

that enforcing MP fairness can significantly reduce the DP disparity.

5 RELATED WORK

Fair regression. Most prior work on fair regression approximates the optimal fair regression function by data preprocessing or regularizers. Inspired by the two-stage least-squares method used in economics, Komiyama and Shimao (2017) propose a two-stage algorithm for linear regression that aims to remove the correlation in the dataset, and extend their work to control the level of fairness by employing a nonconvex optimization method (Komiyama et al., 2018). To provide a general framework for fair regression, Berk et al. (2017) introduce a family of fairness regularizers for linear regression problems which enjoy convexity and permit fast optimization. Similarly, Steinberg et al. (2020) and Mary et al. (2019) propose to measure the fairness using mutual information and Renyi maximum correlation coefficient respectively and incorporate the proposed criterion into regularized risk minimization framework. Recently, Scutari et al. (2021) propose a framework for estimating regression models subject to a user-defined level of fairness by introducing a ridge penalty for unfairness. Unlike those works, this paper focuses on the explicit solution to the MP-fair regression problem with both interpretability and theoretical performance guarantees.

Several works are seeking the explicit solution to the fair regression problem. Calders et al. (2013) consider the fair linear regression problem with MP-constraints and provide a closed-form solution using the method of Lagrange multipliers. Based on the connection between least-squares fair regression under Demographic Parity and optimal transport theory, Chzhen et al. (2020) and Gouic et al. (2020) recently establish the general form of the optimal DP fair regression function and propose a post-processing algorithm that transforms a base estimator of the regression function into a nearly fair one using random smoothing. In the work of Chzhen and Schreuder (2022), the authors consider learning regression function satisfying α-relative DP fair constraint and propose a framework that continuously interpolates between two extreme cases, which is similar to our fairness-accuracy tradeoff method. Other approaches to fair regression include optimization-based methods (Oneto et al., 2020), reduction-based methods (Agarwal et al., 2018), and adversary-based methods (Chi et al., 2021) under some notions of fairness. Unlike them, we focus on MP-fair regression problem in RKHS and derive a closed-form solution by the characterization of fair functional space, which can be extended to covariance-based fairness and other loss functions.

Kernel methods for algorithmic fairness. In recent years, kernel methods have drawn increasing attention from the algorithmic fairness community, which can be roughly categorized into two classes. The first class of work aims to employ the kernel method as a regularizer for fairness. Pérez-Suay et al. (2017) present the fair kernel ridge regression formulation by incorporating the kernel Hilbert Schmidt independence criterion (KHSIC) as the regularizer on the dependence between the predictor and the sensitive attribute. Similarly, Kim and Gittens (2021) propose to learn fair low-rank tensor decompositions by regularizing the Canonical Polyadic Decomposition factorization with the KHSIC. Cho et al. (2020) develop a kernel density estimation (KDE) methodology for classification problems to quantify the fairness measure as a differentiable function and incorporate it as a regularizer. Another class of work aims to learn fair representation by leveraging kernel models. In Grünewälder and Khaleghi (2021), the authors study the relaxed Maximum Mean Discrepancy (MMD) criterion and propose to generate new features that are minimally dependent on the sensitive features while closely approximating the non-sensitive ones. In Okray et al. (2019), the authors consider fair regression with binary sensitive attributes and propose to learn fair feature embeddings in kernel space by minimizing the mean discrepancy between the protected group and the unprotected group. In Tan et al. (2020), the authors leverage the classical sufficient dimension reduction (SDR) framework to construct fair representations as subspaces of the RKHS under some criterion. Our method differs from those methods from two perspectives: we root in constructing the fair function space and aim to find the explicit solution to the MP fair regression problem.

6 CONCLUSION

In this paper, we have proposed a novel approach for regression under Mean Parity fairness which is appealing both theoretically and practically. By characterizing the space of fair regression functions, we derive a closed-form solution to the fair regression problem which has a simple implementation in practice. The proposed fair function space can also be applied to regression under covariance-based fairness and other loss functions. In addition, our method allows users to control the fairness-accuracy tradeoff systemically and offers a simple interpretation. Experimental results suggest that our approach is promising for applications and improves fairness with multiple sensitive attributes.

Limitations and future work. One important direction of future work, and a current challenge is the scalability of the proposed algorithm which is also a common limitation of kernel methods. We remark that many approaches have been proposed to reduce the computational cost of kernel-based algorithms by low-rank matrix approximation (El Alaoui and Mahoney, 2014; Kumar et al., 2009) or random projection (Cesa-Bianchi et al., 2015), which can also benefit our method. Another valuable direction is to apply our method to other kernel-based models such as Support Vector Ma-

chine (Noble, 2006) and Generalized Linear model (Nelder and Wedderburn, 1972). Other directions of interest include studying the generalization problem of fair algorithms and the characterization of the fair function space for more notions of fairness.

References

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

A. Agarwal, M. Dudík, and Z. S. Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129. PMLR, 2019.

M. Anshari, M. N. Almunawar, M. Masri, and M. Hrdy. Financial technology with AI-enabled and ethical challenges. *Society*, pages 1–7, 2021.

B. D. Baker, D. G. Sciarra, and D. Farrie. Is school funding fair? a national report card. *Education Law Center*, 2014.

A. Barroso and A. Brown. Gender pay gap in us held steady in 2020. *Pew Research Center*, 2021.

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.

T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang. Controlling attribute effect in linear regression. In *2013 IEEE 13th International Conference on Data Mining*, pages 71–80. IEEE, 2013.

P. R. Center. On views of race and inequality, blacks and whites are worlds apart. *Social and Demographic Trends*, 2016.

N. Cesa-Bianchi, Y. Mansour, and O. Shamir. On the complexity of learning with kernels. In *Conference on Learning Theory*, pages 297–325. PMLR, 2015.

J. Chi, Y. Tian, G. J. Gordon, and H. Zhao. Understanding and mitigating accuracy disparity in regression. In *International Conference on Machine Learning*, pages 1866–1876. PMLR, 2021.

J. Cho, G. Hwang, and C. Suh. A fair classifier using kernel density estimation. *Advances in Neural Information Processing Systems*, 33:15088–15099, 2020.

E. Chzhen and N. Schreuder. A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4):2416–2442, 2022.

E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair regression with wasserstein barycenters. *arXiv preprint arXiv:2006.07286*, 2020.

L. Darling-Hammond. Unequal opportunity: Race and education. *The Brookings Review*, 16(2):28–32, 1998.

A. El Alaoui and M. W. Mahoney. Fast randomized kernel methods with statistical guarantees. *stat*, 1050:2, 2014.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

A. Frohmader and H. Volkmer. 1-wasserstein distance on the standard simplex. *Algebraic Statistics*, 12(1):43–56, 2021.

T. L. Gouic, J.-M. Loubes, and P. Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020.

A. Gretton. Introduction to rkhs, and some simple kernel algorithms. *Adv. Top. Mach. Learn. Lecture Conducted from University College London*, 16:5–3, 2013.

C. W. Groetsch. *Generalized Inverses of Linear Operators: Representation and Approximation*. Dekker, 1977.

S. Grünewälder and A. Khaleghi. Oblivious data for fairness with kernels. *Journal of Machine Learning Research*, 22 (208):1–36, 2021.

M. Gupta and Q. Mohammad. Advances in AI and ML are reshaping healthcare. *SAP News Center. Available online at: https://techcrunch. com/2017/03/16/advances-in-ai-and-ml-are-reshaping-healthcare/(Accessed Jun 20, 2018)*, 2017.

L. Hoegaerts, J. A. Suykens, J. Vandewalle, and B. De Moor. Subset based least squares subspace regression in rkhs. *Neurocomputing*, 63:293–323, 2005.

L. Huang and N. Vishnoi. Stable and fair classification. In *International Conference on Machine Learning*, pages 2879–2890. PMLR, 2019.

R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pages 862–872. PMLR, 2020.

H. Kadri, E. Duflos, P. Preux, S. Canu, and M. Davy. Nonlinear functional regression: a functional rkhs approach. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 374–380. JMLR Workshop and Conference Proceedings, 2010.

K. Kim and A. Gittens. Learning fair canonical polyadical decompositions using a kernel independence criterion. *arXiv preprint arXiv:2104.13504*, 2021.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

A. A. Kodiyan. An overview of ethical issues in using ai systems in hiring with a case study of amazon's ai based hiring tool. *Researchgate Preprint*, 2019.

R. Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, pages 202–207, 1996.

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

J. Komiyama, A. Takeda, J. Honda, and H. Shimao. Nonconvex optimization for regression with fairness constraints. In *International Conference on Machine Learning*, pages 2737–2746. PMLR, 2018.

S. Kumar, M. Mohri, and A. Talwalkar. Sampling techniques for the nystrom method. In *Artificial intelligence and statistics*, pages 304–311. PMLR, 2009.

J. Mary, C. Calauzenes, and N. El Karoui. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pages 4382–4391. PMLR, 2019.

J. A. Nelder and R. W. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

W. S. Noble. What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567, 2006.

G. S. Oettinger. Statistical discrimination and the early career evolution of the black-white wage gap. *Journal of Labor Economics*, 14(1):52–78, 1996.

A. Okray, H. Hu, and C. Lan. Fair kernel regression via fair feature embedding in kernel space. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1417–1421. IEEE, 2019.

L. Oneto, M. Donini, and M. Pontil. General fair empirical risk minimization. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

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*, pages 339–355. Springer, 2017.

E. Raff, J. Sylvester, and S. Mills. Fair forests: Regularized tree induction to minimize model bias. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 243–250, 2018.

M. Redmond and A. Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer, 2001.

M. Scutari, F. Panero, and M. Proissl. Achieving fairness with a simple ridge penalty. *arXiv preprint arXiv:2105.13817*, 2021.

D. Steinberg, A. Reid, S. O'Callaghan, F. Lattimore, L. McCalman, and T. Caetano. Fast fair regression via efficient approximations of mutual information. *arXiv preprint arXiv:2002.06200*, 2020.

Z. Tan, S. Yeom, M. Fredrikson, and A. Talwalkar. Learning fair representations for kernel models. In *International Conference on Artificial Intelligence and Statistics*, pages 155–166. PMLR, 2020.

G. Wang, Y. Wei, and S. Qiao. Moore-penrose inverse of linear operators. In *Generalized Inverses: Theory and Computations*, pages 317–338. Springer, 2018.

M. Welling. Kernel ridge regression. *Max Welling's Classnotes in Machine Learning*, pages 1–3, 2013.

L. F. Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

M. Yuan and T. T. Cai. A reproducing kernel hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444, 2010.

M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1):2737–2778, 2019.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/zemel13.html.

I. Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31 (4):1060–1089, 2017.

A PROOFS

A.1 Proof of Proposition 1

Proof. By the fact that $\mathbb{E}(X) = \int_0^\infty (1 - F(x))dx - \int_{-\infty}^0 F(x)dx$ where $F(X)$ is the cumulative distribution functions (CDF) of X, we have

$$\begin{aligned}
\text{MPD}(g) &= \sum_{s\in\Omega_S} |\mathbb{E}(g(X,S)|S=s) - \mathbb{E}(g(X,S))| \\
&= \sum_{s\in\Omega_S} \left| \int_{\mathbb{R}} (F_{g(X,S)|S=s}(t) - F_{g(X,S)}(t))dt \right| \\
\text{DPD}(g) &= \sum_{s\in\Omega_S} \mathcal{W}_1(g(X,S)|S=s, g(X,S)) \\
&= \sum_{s\in\Omega_S} \left(\int_{\mathbb{R}} \left| (F_{g(X,S)|S=s}(t) - F_{g(X,S)}(t)) \right| dt \right)
\end{aligned}$$

where $F_{g(X,S)|S=s}$ and $F_{g(X,S)}$ are the CDF of $g(X,S)|S=s$ and $g(X,S)$ respectively.

By the Triangle inequality, we have

$$\left| \int_{\mathbb{R}} (F_{g(X,S)|S=s}(t) - F_{g(X,S)}(t))dt \right| \leq \int_{\mathbb{R}} \left| (F_{g(X,S)|S=s}(t) - F_{g(X,S)}(t)) \right| dt \quad \forall s \in \Omega_S. \tag{8}$$

So,

$$\text{MPD}(g) \leq \text{DPD}(g)$$

□

A.2 Proof of Theorem 1

Proof. By the definition of $\ker(\Sigma_{S(XS)})$, a function g is in $\ker(\Sigma_{S(XS)})$ if and only if

$$\Sigma_{S(XS)} g = 0_{\mathcal{H}_S}.$$

For a function $g \in \mathcal{H}_{XS}$, notice that

$$\begin{aligned}
\Sigma_{S(XS)} g &= \mathbb{E}_{XS}\left[(\phi_S(S) - \mu_s) \otimes (\phi_{XS}(X,S) - \mu_{XS})g\right] \\
&= \mathbb{E}_{XS}\left[\langle \phi_{XS}(X,S) - \mu_{XS}, g\rangle_{\mathcal{H}_{XS}} (\phi_S(S) - \mu_s)\right] \quad \text{(By the definition of } \otimes) \\
&= \mathbb{E}_{XS}\left[(g(X,S) - \mathbb{E}_{XS}(g(X,S)))(\phi_S(S) - \mu_s)\right] \quad \text{(By the reproducing property)} \\
&= \mathbb{E}_S\left[(\mathbb{E}_X(g(X,S)|S) - \mathbb{E}_{XS}(g(X,S)))\,(\phi_S(S) - \mu_s)\right] \\
&= \sum_{j=1}^k \mathbb{P}(S = s^{(j)}) \left(\mathbb{E}_X(g(X,S)|S=s^{(j)}) - \mathbb{E}_{XS}(g(X,S))\right)(\phi_S(s^{(j)}) - \mu_s)
\end{aligned}$$

and

$$\sum_{j=1}^k \mathbb{P}(S = s^{(j)}) \left(\mathbb{E}_X(g(X,S)|S=s^{(j)}) - \mathbb{E}_{XS}(g(X,S))\right) = 0$$

where we use the reproducing property of $\mathcal{H}_{XS}$ and the definition that $(a \otimes b)c = \langle b, c\rangle_{\mathcal{H}} a$ for $b, c \in \mathcal{H}$ (Gretton, 2013).

Note that we assume $\mathbb{P}(S = s^{(j)}) > 0$ for all $s^{(j)} \in \Omega_S$ since the sensitive attributes with zero probability don't influence the fairness in practice. Then, under the Assumption 1 that the system of equations

$$\sum_{j=1}^k \eta_j(\phi_S(s^{(j)}) - \mu_S) = 0_{\mathcal{H}_S}, \quad \sum_{j=1}^k \eta_j = 0 \tag{9}$$

has unique solution $\eta_j = 0$ for all $j \in \{1, \ldots, k\}$, we can conclude that

$$\Sigma_{S(XS)} g = 0_{\mathcal{H}_S} \quad \forall g \in \mathcal{G}_{MP}$$
$$\mathbb{E}_X(g(X,S)|S) - \mathbb{E}_{XS}(g(X,S)) = 0 \quad \forall g \in \ker(\Sigma_{S(XS)}).$$

So, $\mathcal{G}_{MP} = \ker(\Sigma_{S(XS)})$. □

A.3 Proof of Lemma 1

Proof. Recall that Problem 1 considers the following objective

$$\mathbb{E}(Y - g(X,S))^2,$$

which is equivalent to

$$\mathbb{E}(Y^2) - 2\mathbb{E}(Y g(X,S)) + \mathbb{E}(g(X,S))^2.$$

Denote the optimal regression function of Problem 1 by $g^*_{\mathcal{G}}$. Let Δ be an arbitrary function in $\mathcal{G}_{MP}$, then $g' = g^*_{\mathcal{G}} + \Delta$ is a function in $\mathcal{G}_{MP}$ and

$$\begin{aligned}\mathbb{E}((Y - g'(X,S))^2 &= \mathbb{E}(Y^2) - 2\mathbb{E}(Y g'(X,S)) + \mathbb{E}(g'(X,S))^2 \\ &= \mathbb{E}(Y - g^*_{\mathcal{G}}(X,S))^2 - 2\mathbb{E}(Y\Delta(X,S)) \\ &\quad + 2\mathbb{E}(g^*_{\mathcal{G}}(X,S)\Delta(X,S)) + \mathbb{E}(\Delta(X,S))^2.\end{aligned}$$

Note that $g^*_{\mathcal{G}}$ is an optimal solution if and only if

$$\mathbb{E}((Y - g'(X,S))^2 \geq \mathbb{E}(Y - g^*_{\mathcal{G}}(X,S))^2$$

which is equivalent to

$$-2\mathbb{E}(Y\Delta(X,S)) + 2\mathbb{E}(g^*_{\mathcal{G}}(X,S)\Delta(X,S)) + \mathbb{E}(\Delta(X,S))^2 \geq 0 \quad \forall \Delta \in \mathcal{G}_{MP}.$$

The above inequality holds if and only if

$$-2\mathbb{E}(Y\Delta(X,S)) + 2\mathbb{E}(g^*_{\mathcal{G}}(X,S)\Delta(X,S)) = 0 \quad \forall \Delta \in \mathcal{G}_{MP},$$

which is equivalent to

$$\mathbb{E}(Y\Delta(X,S)) = \mathbb{E}(g^*_{\mathcal{G}}(X,S)\Delta(X,S)) \quad \forall \Delta \in \mathcal{G}_{MP},$$

otherwise, scaling Δ by a proper scalar yields a contradiction.

□

A.4 Proof of Corollary 1

Proof. Given $g_1, g_2 \in \mathcal{H}_{XS}$, $\mathbb{E}(g_1(X,S)|S) - \mathbb{E}(g_1(X,S)) = \mathbb{E}(g_2(X,S)|S) - \mathbb{E}(g_2(X,S))$ indicates that $g_1 - g_2 \in \mathcal{G}_{MP}$, that is, under Assumption 1,

$$\Sigma_{S(XS)}(g_1 - g_2) = 0_{\mathcal{H}_S}.$$

Rewriting the above equation gives

$$\Sigma_{S(XS)} g_1 = \Sigma_{S(XS)} g_2.$$

□

A.5 Proof of Proposition 3

Proof. A function $g \in \mathcal{H}_{XS}$ satisfies $\Sigma_{S(XS)} g = \Sigma_{S(XS)} g^*$ if and only if $\Sigma_{S(XS)}(g - g^*) = 0_{\mathcal{H}_S}$. By the definition of $\mathcal{G}_{MP}$, we have

$$\begin{aligned}\Sigma_{S(XS)}(g - g^*) &= \Sigma_{S(XS)}(g_{MP} - g^*_{MP}) + \Sigma_{S(XS)}(g_{MP^\perp} - g^*_{MP^\perp}) \\ &= \Sigma_{S(XS)}(g_{MP^\perp} - g^*_{MP^\perp}) \\ &= 0_{\mathcal{H}_S}.\end{aligned}$$

As $g_{MP^\perp} - g^*_{MP^\perp} \in \mathcal{G}^\perp_{MP}$, the above equation holds if and only if $g_{MP^\perp} = g^*_{MP^\perp}$. □

A.6 Proof of Proposition 5

Proof. To bound the MSE of $g^*_{\mathcal{G}}$, we introduce a sub-optimal fair regression function $g' = Pg^*$ where g^* is the optimal regression function in $\mathcal{H}_{XS}$. Then, the reduction of MSE are

$$\begin{aligned}
L(g^*_{\mathcal{G}}) - L(g^*) &\leq L(g') - L(g^*) \\
&= \mathbb{E}(Y - g'(X,S))^2 - \mathbb{E}(Y - g^*(X,S))^2 \\
&= -2\mathbb{E}(Yg'(X,S)) + \mathbb{E}(g'(X,S))^2 - \left(-2\mathbb{E}(Yg^*(X,S)) + \mathbb{E}(g^*(X,S))^2\right) \\
&= \mathbb{E}(g'(X,S))^2 - 2\mathbb{E}(Yg'(X,S)) + \mathbb{E}(g^*(X,S))^2 \quad \text{(By the optimal condition of } g^*\text{)} \\
&= \mathbb{E}(g'(X,S))^2 - 2\mathbb{E}(g^*(X,S)g'(X,S)) + \mathbb{E}(g^*(X,S))^2 \quad \text{(By the optimal condition of } g^*\text{)} \\
&= \mathbb{E}(g'(X,S) - g^*(X,S))^2 \\
&= \mathbb{E}(g^*_{MP^\perp}(X,S))^2 \quad (\text{By } g^* = Pg^* + g^*_{MP^\perp}) \\
&= \langle \tilde{\Sigma}_{(XS)(XS)} g^*_{MP^\perp}, g^*_{MP^\perp} \rangle_{\mathcal{H}_{XS}}.
\end{aligned}$$

Therefore, under Assumption 1, the MSE of $g^*_{\mathcal{G}}$ is bounded by

$$L(g^*_{\mathcal{G}}) \leq L(g^*) + \langle \tilde{\Sigma}_{(XS)(XS)} g^*_{MP^\perp}, g^*_{MP^\perp} \rangle_{\mathcal{H}_{XS}}. \tag{10}$$

□

A.7 Proof in Example 1

In this section, we prove that when S is a binary random variable, $\kappa_S(s_i, s_j) = s_i s_j$ satisfies Assumption 1.

Proof. Without loss of generality, we assume that $S \in \{0, 1\}$. Since the following system of equations

$$\begin{aligned}
\eta_1(0 - \mathbb{P}(S=1)) + \eta_2(1 - \mathbb{P}(S=1)) &= 0 \\
\eta_1 + \eta_2 &= 0,
\end{aligned}$$

has a unique solution $\eta_1 = \eta_2 = 0$, Assumption 1 is satisfied. □

B IMPLEMENTATION

In this section, we focus on the estimation of optimal regression function by solving the empirical approximation of Problem 1. Specifically, given the training dataset $\mathcal{D} = \{x_i, s_i, y_i\}_{i=1}^n$, we seek the solution to the following regularized fair regression problem (Kadri et al., 2010; Hoegaerts et al., 2005),

$$\min_{g \in \mathcal{G}_{MP}} \frac{1}{n} \sum_{i=1}^n (y_i - g(x_i, s_i))^2 + \frac{\lambda}{n} \|g\|_{\mathcal{H}_{XS}}, \tag{11}$$

where $\lambda \geq 0$ is a real number (regularization coefficient) to control the tradeoff between approximating properties and the smoothness of g. Note that when $\lambda = 0$, Problem 11 is the estimation of Problem 1, but it may be ill-posed depending on $\mathcal{H}_{XS}$.

To solve Problem 11, we first show the empirical estimation of $\Sigma_{(XS)S}$ and how to estimate the eigenfunctions of $\Sigma_{(XS)S} A \Sigma_{S(XS)}$, which allows us to construct an orthogonal projection operator. After that, we derive the closed-form solution for Problem 11 which is the empirical estimation of the optimal fair regression function 3 when $\lambda = 0$.

B.1 Empirical estimation of MP-fair function space

Recall that the feature maps of $\mathcal{H}_{XS}$ and $\mathcal{H}_S$ are ϕ_{XS} and ϕ_S respectively. Let us define $\bar{\phi}_{XS}(x_i, s_i) = \phi_{XS}(x_i, s_i) - \frac{1}{n}\sum_{j=1}^n \phi_{XS}(x_j, s_j)$ and $\bar{\phi}_S(s_i) = \phi_S(s_i) - \frac{1}{n}\sum_{j=1}^n \phi_S(s_j)$. Then, the empirical estimation of $\Sigma_{(XS)S}$ is

$$\hat{\Sigma}_{(XS)S} = \frac{1}{n} \sum_{i=1}^n \bar{\phi}_{XS}(x_i, s_i) \otimes \bar{\phi}_S(s_i).$$

To simplify the derivation, we set A to be the identity operator and focus on $\hat{\Sigma}_{(XS)S}\hat{\Sigma}_{S(XS)}$.

Define the feature matrix Φ_{XS} and Gram matrix $\boldsymbol{K}_{XS}$ as

$$\Phi_{XS} = [\phi_{XS}(x_1, s_1), \ldots, \phi_{XS}(x_n, s_n)]^T$$
$$\boldsymbol{K}_{XS} = \Phi_{XS}^T \Phi_{XS},$$

such that the ith column of Φ_{XS} is $\phi_{XS}(x_i, s_i)$ and the (i, j) entry of $\boldsymbol{K}_{XS}$ is $\kappa_{XS}((x_i, s_i), (x_j, s_j))$. Similarly, we denote the feature matrix and Gram matrix of S by Φ_S and $\boldsymbol{K}_S$ respectively.

For simplicity, we assume that $\{\bar{\phi}_{XS}(x_i, s_i)\}$ is a set of linearly independent feature maps which ensures that an eigenfunction $\hat{\theta}_l$ is uniquely determined by a set of scalars. In case where $\{\bar{\phi}_{XS}(x_i, s_i)\}$ are not linearly independent e.g., duplicated data samples, the following process can still be applied since we can get orthonormal bases by removing the duplicated eigenfunctions.

By the observation that $\mathrm{ran}(\hat{\Sigma}_{(XS)S}\hat{\Sigma}_{S(XS)})$ is a subspace of $\mathrm{span}(\{\phi_{XS}(x_i, s_i)\}_{i=1}^n)$, the jth eigenfunction of $\hat{\Sigma}_{(XS)S}\hat{\Sigma}_{S(XS)}$ can be written as

$$\hat{\theta}_j = \Phi_{XS}\boldsymbol{a}_j \quad \text{or} \quad \hat{\theta}_j = \bar{\Phi}_{XS}\bar{\boldsymbol{a}}_j \quad \forall j \in \{1, \ldots, m\},$$

where $\boldsymbol{a}_j, \bar{\boldsymbol{a}}_j \in \mathbb{R}^n$ are vectors of coefficients and $\bar{\Phi}_{XS} = \Phi_{XS}\boldsymbol{H}$ for $\boldsymbol{H} = \boldsymbol{I}_{n\times n} - \frac{1}{n}\boldsymbol{1}_{n\times n}$ (Schölkopf et al., 1998).

The generalized eigenvalue ψ_j corresponding to $\hat{\theta}_j$ satisfies

$$\psi_j\hat{\theta}_j = \hat{\Sigma}_{(XS)S}\hat{\Sigma}_{S(XS)}\hat{\theta}_j.$$

Writing the above equation as a matrix form yields

$$\psi_j\bar{\boldsymbol{a}}_j = \frac{1}{n^2}\bar{\boldsymbol{K}}_S\bar{\boldsymbol{K}}_{XS}\bar{\boldsymbol{a}}_j,$$

where $\bar{\boldsymbol{K}}_S = \boldsymbol{H}\boldsymbol{K}_S\boldsymbol{H}$ and $\bar{\boldsymbol{K}}_{XS} = \boldsymbol{H}\boldsymbol{K}_{XS}\boldsymbol{H}$.

Thus, $\bar{\boldsymbol{a}}_j$ is the eigenvector of the matrix $\frac{1}{n^2}\bar{\boldsymbol{K}}_S\bar{\boldsymbol{K}}_{XS}$ and $\boldsymbol{a}_j = \boldsymbol{H}\bar{\boldsymbol{a}}_j$. Since $\hat{\Sigma}_{(XS)S}\hat{\Sigma}_{S(XS)}$ is self-adjoint, the first m eigenfunctions are orthogonal. So, we can normalize the eigenfunctions to construct a set of orthonormal bases of $\mathrm{ran}(\hat{\Sigma}_{(XS)S})$. A more detailed derivation can be found in Appendix C.3.

B.2 Construction of projection operator

With some abuse of notation, we denote a set of orthonormal bases of $\mathrm{ran}(\Sigma_{(XS)S})$ by $\{\theta_1, \cdots, \theta_m\}$ and its estimation by $\{\hat{\theta}_1, \cdots, \hat{\theta}_m\}$ where $\hat{\theta}_j = \Phi_{XS}\boldsymbol{a}_j$ to avoid complicated symbols. Given a function $g \in \mathcal{H}_{XS}$, the orthogonal projection operator from $\mathcal{H}_{XS}$ onto $\mathrm{ran}(\Sigma_{(XS)S})^\perp$ eliminates the components of g in $\mathrm{ran}(\Sigma_{(XS)S})$. Thus, we can construct the following orthogonal projection operator

$$P = I - \sum_{j=1}^m \theta_j \otimes \theta_j,$$

where $I : \mathcal{H}_{XS} \to \mathcal{H}_{XS}$ is the identity operator.

So, the estimation of P can be written as

$$\hat{P} = I - \sum_{j=1}^m \hat{\theta}_j \otimes \hat{\theta}_j \tag{12}$$

Note that given $g = \Phi_{XS}\boldsymbol{c}$, the projection of g on $\mathcal{G}_{MP}$ is

$$\hat{P}g = g - \sum_{j=1}^m \langle g, \hat{\theta}_j\rangle_{\mathcal{H}_{XS}}\hat{\theta}_j = \Phi_{XS}\boldsymbol{P}\boldsymbol{c},$$

where $\boldsymbol{P} = (\boldsymbol{I}_{n\times n} - \sum_{j=1}^m \boldsymbol{a}_m\boldsymbol{a}_m^T\boldsymbol{K}_{XS})$.

B.3 Estimation of fair regression function

Given an orthogonal projection operator estimation $\hat{P}$, the optimal solution to Problem 11 is $\hat{g}^*_{\mathcal{G}} = \hat{P}\hat{g}^*_{\mathcal{H}}$ where $\hat{g}^*_{\mathcal{H}}$ can be obtained by solving the following problem

$$\min_{g \in \mathcal{H}_{XS}} \frac{1}{n}\left(y_i - \langle \phi_{XS}(x_i, s_i), \hat{P}g\rangle\right)^2 + \frac{\lambda}{n}\left\|\hat{P}g\right\|_{\mathcal{H}_{XS}}.$$

By the Representer theorem (Schölkopf et al., 2001), $\hat{g}_{\mathcal{H}}$ is of the form $\hat{g}_{\mathcal{H}} = \Phi_{XS}\boldsymbol{w}_{\mathcal{H}}$ for $\boldsymbol{w}_{\mathcal{H}} \in \mathbb{R}^n$. So, it suffices to minimize the following objective function

$$\begin{aligned} J(\boldsymbol{w}) =& \boldsymbol{w}^T\boldsymbol{P}^T\boldsymbol{K}_{XS}\boldsymbol{K}_{XS}\boldsymbol{P}\boldsymbol{w} - 2\boldsymbol{Y}^T\boldsymbol{K}_{XS}\boldsymbol{P}\boldsymbol{w} \\ &+ \boldsymbol{Y}^T\boldsymbol{Y} + \lambda\boldsymbol{w}^T\boldsymbol{P}^T\boldsymbol{K}_{XS}\boldsymbol{P}\boldsymbol{w}, \end{aligned}$$

where $\boldsymbol{Y} = [y_1, \ldots, y_n]$ is a vector in $\mathbb{R}^n$.

Since $J(\boldsymbol{w})$ is convex, it has a minimizer. Setting $\frac{\partial J}{\partial \boldsymbol{w}}$ to zero yields

$$\boldsymbol{w}_{\mathcal{H}} = (\boldsymbol{P}^T\boldsymbol{K}_{XS}\boldsymbol{K}_{XS}\boldsymbol{P} + \lambda\boldsymbol{P}^T\boldsymbol{K}_{XS}\boldsymbol{P})^\dagger\boldsymbol{P}^T\boldsymbol{K}_{XS}\boldsymbol{Y}.$$

So, the optimal fair regression function is $\hat{g}^*_{\mathcal{G}} = \Phi_{XS}\boldsymbol{w}^*_{\mathcal{G}}$ where

$$\boldsymbol{w}^*_{\mathcal{G}} = \boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{K}_{XS}\boldsymbol{K}_{XS}\boldsymbol{P} + \lambda\boldsymbol{P}^T\boldsymbol{K}_{XS}\boldsymbol{P})^\dagger\boldsymbol{P}^T\boldsymbol{K}_{XS}\boldsymbol{Y}.$$

Example: fair linear regression. Consider the fair linear regression problem with single binary sensitive attribute. The kernels are

$$\begin{aligned} \kappa_S(s_i, s_j) &= s_i s_j \\ \kappa_X(x_i, x_j) &= x_i^T x_j \\ \kappa_{XS}((x_i, s_i), (x_j, s_j)) &= \kappa_S(s_i, s_j) + \kappa_X(x_i, x_j). \end{aligned}$$

We prove that the above setting satisfies Assumption 1 in Appendix A.7, which implies that MP fairness is equivalent to CB fairness in this example. Let $\lambda = 0$. The optimal fair regression function is

$$\hat{g}^*_{\mathcal{G}} = \Phi_{XS}\boldsymbol{P}(\boldsymbol{K}_{XS}\boldsymbol{P})^\dagger\boldsymbol{Y}$$

and the fitted value of $\boldsymbol{Y}$ is

$$\hat{\boldsymbol{Y}} = \boldsymbol{K}_{XS}\boldsymbol{P}(\boldsymbol{K}_{XS}\boldsymbol{P})^\dagger\boldsymbol{Y}.$$

C DERIVATIONS AND DISCUSSIONS

C.1 Relation between MP fairness and CB fairness

In this section, we discuss general CB fairness and its relation to MP fairness. We first provide the following assumption

Assumption 2. *Assume the κ_{XS} is composed of κ_S and κ_X.*

which is the assumption in ordinary kernelized regression problem where S and X are mapped to $\phi_S(S)$ and $\phi_X(X)$ respectively. As discussed in the work of Komiyama et al. (2018) and Pérez-Suay et al. (2017), the general CB fairness seeks to remove the correlation between S and $g(X, S)$ on the (possibly infinite) representation space. Specifically, the CB fairness requires that the regression function $g \in \mathcal{H}_{XS}$ achieves $\mathrm{Cov}(\phi_S(S), g(X, S)) = 0_{\mathcal{H}_S}$ under Assumption 2. By the definition of $\mathrm{Cov}(\phi_S(S), g(X, S))$, we have

$$\begin{aligned} \mathrm{Cov}(\phi_S(S), g(X, S)) &= \mathbb{E}\left[(g(X, S) - \mathbb{E}(g(X, S)))(\phi_S(S) - \mu_S)\right] \\ &= \mathbb{E}\left[\langle \phi_{XS}(X, S) - \mu_{XS}, g\rangle_{\mathcal{H}_{XS}}(\phi_S(S) - \mu_S)\right] \\ &= \mathbb{E}\left[(\phi_S(S) - \mu_S) \otimes (\phi_{XS}(X, S) - \mu_{XS})\rangle\right] g \\ &= \Sigma_{S(XS)}g, \end{aligned} \tag{13}$$

where we use the reproducing property of $\mathcal{H}_{XS}$ and the definition that $(a \otimes b)c = \langle b, c\rangle_{\mathcal{H}} a$ for $b, c \in \mathcal{H}$ (Gretton, 2013) to derive this result.

Equation 13 claims that a function $g \in \mathcal{H}_{XS}$ is CB-fair if and only if g is in $\ker(\Sigma_{S(XS)})$. Since the proposed method solves the fair regression problem by the characterization of $\ker(\Sigma_{S(XS)})$, it can also be applied to CB fairness under Assumption 2. In particular, if both Assumption 1 and Assumption 2 are satisfied, MP fairness is equivalent to CB fairness.

C.2 Derivation of equations in Section 3.3

Solving Problem 6 gives

$$\begin{aligned} g^{\alpha}_{MP} &= P[P\tilde{\Sigma}_{(XS)(XS)}P]^{\dagger}P(h - \alpha\tilde{\Sigma}_{(XS,XS)}g^{*}_{MP^{\perp}}) \\ &= g^{*}_{\mathcal{G}} - \alpha P[P\tilde{\Sigma}_{(XS)(XS)}P]^{\dagger}P\tilde{\Sigma}_{(XS,XS)}g^{*}_{MP^{\perp}}. \end{aligned}$$

As $g^{*}_{MP^{\perp}} = (I - P)g^{*}$ where I is the identity operator, we have

$$\begin{aligned} &P[P\tilde{\Sigma}_{(XS)(XS)}P]^{\dagger}P\tilde{\Sigma}_{(XS,XS)}g^{*}_{MP^{\perp}} \\ &\qquad = P[P\tilde{\Sigma}_{(XS)(XS)}P]^{\dagger}P\tilde{\Sigma}_{(XS,XS)}g^{*} - P[P\tilde{\Sigma}_{(XS)(XS)}P]^{\dagger}P\tilde{\Sigma}_{(XS,XS)}Pg^{*} \end{aligned}$$

where the first term equals to $g^{*}_{\mathcal{G}}$ by the property that $\langle g, \tilde{\Sigma}_{(XS,XS)}g^{*}\rangle_{\mathcal{H}_{XS}} = \langle g, h\rangle_{\mathcal{H}_{XS}}$ for all $g \in \mathcal{H}_{XS}$, and the second term equals to Pg^{*} since it's the optimal solution of $\min_{g\in\mathcal{G}_{MP}} \mathbb{E}(Pg^{*} - g)^2$.

Therefore, we get

$$g^{\alpha}_{MP} = (1 - \alpha)g^{*}_{\mathcal{G}} + \alpha Pg^{*}.$$

Alternatively, we can show the above equation using the fact that $g^{\alpha} = g^{*}$ when $\alpha = 1$.

Thus, the optimal solution to Problem 5 is $g^{\alpha} = (1 - \alpha)g^{*}_{\mathcal{G}} + \alpha g^{*}$.

Now, we turn to the MSE of g^{α}. We have

$$\begin{aligned} L(g^{\alpha}) &= \mathbb{E}(Y - g^{\alpha}(X, S))^2 \\ &= \mathbb{E}(Y - (1 - \alpha)g^{*}_{\mathcal{G}}(X, S) - \alpha g^{*}(X, S))^2 \\ &= \mathbb{E}((1 - \alpha)(Y - g^{*}_{\mathcal{G}}(X, S)) + \alpha(Y - g^{*}(X, S)))^2 \\ &= \mathbb{E}((1 - \alpha)(Y - g^{*}_{\mathcal{G}}(X, S)))^2 + \mathbb{E}(\alpha(Y - g^{*}(X, S)))^2 \\ &\quad + 2\mathbb{E}((1 - \alpha)(Y - g^{*}_{\mathcal{G}}(X, S))(\alpha(Y - g^{*}(X, S)))) \\ &= \mathbb{E}((1 - \alpha)(Y - g^{*}_{\mathcal{G}}(X, S)))^2 + \mathbb{E}(\alpha(Y - g^{*}(X, S)))^2 \\ &\quad + 2\alpha(1 - \alpha)\mathbb{E}((Y - g^{*}_{\mathcal{G}}(X, S))(Y - g^{*}(X, S))) \\ &= (1 - \alpha)^2\mathbb{E}(Y - g^{*}_{\mathcal{G}}(X, S)))^2 + (1 - (1 - \alpha)^2)\mathbb{E}((Y - g^{*}(X, S)))^2 \\ &= (1 - \alpha)^2 L(g^{*}_{\mathcal{G}}) + (1 - (1 - \alpha)^2)L(g^{*}) \\ &= \alpha^2(L(g^{*}_{\mathcal{G}}) - L(g^{*})) - 2\alpha(L(g^{*}_{\mathcal{G}}) - L(g^{*})) + L(g^{*}_{\mathcal{G}}), \\ &= (1 - \alpha)^2 L(g^{*}_{\mathcal{G}}) + (1 - (1 - \alpha)^2)L(g^{*}) \\ \mathrm{MPD}(g^{\alpha}) &= \alpha\mathrm{MPD}(g^{*}). \end{aligned} \tag{14}$$

since $\mathbb{E}(Y(Y - g^{*}(X, S))) = \mathbb{E}(Y - g^{*}(X, S))^2$ and $\mathbb{E}(g^{*}_{\mathcal{G}}(X, S)(Y - g^{*}(X, S))) = 0$.

C.3 Estimating eigenfunctions and orthonormal bases

Now we provide detailed derivation about finding the eigenfunctions of $\hat{\Sigma}_{(XS)S}A\hat{\Sigma}_{S(XS)}$.

Recall that

$$\hat{\Sigma}_{(XS)S} = \frac{1}{n}\sum_{i=1}^{n} \bar{\phi}_{XS}(x_i, s_i) \otimes \bar{\phi}_{S}(s_i).$$

Let A be the identity operator, and we get

$$\begin{aligned}\hat{\Sigma}_{(XS)S}A\hat{\Sigma}_{S(XS)} &= \left(\frac{1}{n}\sum_{i=1}^{n}\bar{\phi}_{XS}(x_i,s_i)\otimes\bar{\phi}_S(s_i)\right)\left(\frac{1}{n}\sum_{i=1}^{n}\bar{\phi}_S(s_i)\otimes\bar{\phi}_{XS}(x_i,s_i)\right)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{\phi}_{XS}(x_i,s_i)\otimes\bar{\phi}_S(s_i)\bar{\phi}_S(s_j)\otimes\bar{\phi}_{XS}(x_j,s_j)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle\bar{\phi}_S(s_i),\bar{\phi}_S(s_j)\rangle_{\mathcal{H}_S}\bar{\phi}_{XS}(x_i,s_i)\otimes\bar{\phi}_{XS}(x_j,s_j)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{\kappa}_S(s_i,s_j)\bar{\phi}_{XS}(x_i,s_i)\otimes\bar{\phi}_{XS}(x_j,s_j).\end{aligned}$$

For simplicity, we assume that $\{\bar{\phi}_{XS}(x_i,s_i)\}$ is a set of independent feature maps which ensures that $\hat{\theta}_l$ is uniquely determined by a set of scalars. In case where $\{\bar{\phi}_{XS}(x_i,s_i)\}$ are not independent e.g., duplicated data samples, the following process can still be applied since we can get orthonormal bases by removing the duplicated eigen functions.

Since $\hat{\Sigma}_{(XS)S}\hat{\Sigma}_{S(XS)}$ is in $\mathrm{ran}(\{\bar{\phi}_{XS}(x_i,s_i)\}_{i=1}^n)$, the l^{th} eigenfunction of $\hat{\Sigma}_{(XS)S}\hat{\Sigma}_{S(XS)}$ can be written as

$$\hat{\theta}_l = \sum_{i=1}^{n}\bar{a}_{l_i}\bar{\phi}_{XS}(x_i,s_i).$$

By the definition of eigenfunction, we get

$$\psi_l\hat{\theta}_l = \hat{\Sigma}_{(XS)S}\hat{\Sigma}_{S(XS)}\hat{\theta}_l. \tag{15}$$

Observe that

$$\langle\bar{\phi}_{XS}(x_i,s_i),\sum_{j=1}^{n}\bar{a}_{l_j}\bar{\phi}_{XS}(x_j,s_j)\rangle_{\mathcal{H}_{XS}} = \sum_{j=1}^{n}\bar{a}_{l_j}\bar{\kappa}_{XS}\left((x_i,s_i),(x_j,s_j)\right),$$

where $\bar{\kappa}_{XS}\left((x_i,s_i),(x_j,s_j)\right)$ is the (i,j) entry of the matrix $\bar{\boldsymbol{K}}_{XS} = \boldsymbol{H}\boldsymbol{K}_{XS}\boldsymbol{H}$ with Gram matrix $\boldsymbol{K}_{XS}$ and $\boldsymbol{H} = \boldsymbol{I}_{n\times n} - n^{-1}\mathbf{1}_{n\times n}$. Thus, we get

$$\hat{\Sigma}_{(XS)S}\hat{\Sigma}_{S(XS)}\hat{\theta}_l = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{l_j}\bar{\kappa}_S(s_i,s_j)\bar{\phi}_{XS}(x_i,s_i),$$

where $\beta_{l_j} = \sum_{r=1}^{n}\bar{a}_{l_r}\bar{\kappa}_{XS}\left((x_j,s_j),(x_r,s_r)\right)$.

By Equation 15, it suffices to solve

$$\lambda_l\bar{a}_{l_i} = \frac{1}{n^2}\sum_{j=1}^{n}\bar{\kappa}_S(s_i,s_j)\sum_{r=1}^{n}\bar{a}_{l_r}\bar{\kappa}_{XS}\left((x_j,s_j)(x_r,s_r)\right).$$

Writing the above equation as a matrix equation yields

$$\psi_l\bar{\boldsymbol{a}}_l = \frac{1}{n^2}\bar{\boldsymbol{K}}_S\bar{\boldsymbol{K}}_{XS}\bar{\boldsymbol{a}}_l,$$

where $\boldsymbol{a}_l = [\bar{a}_{l_1},\ldots,\bar{a}_{l_n}]$ is a column vector in $\mathbb{R}^n$.

Thus, $\boldsymbol{a}_l$ is the eigenvector of the matrix $\frac{1}{n^2}\bar{K}_S\bar{K}_{XS}$. Let

$$\boldsymbol{a}_l = \boldsymbol{H}\bar{\boldsymbol{a}}_l.$$

The l^{th} eigenvector can be rewritten as

$$\hat{\theta}_l = \sum_{i=1}^{n}a_{l_i}\phi_{XS}(x_i,s_i).$$

Since $\hat{\Sigma}_{(XS)S}\hat{\Sigma}_{S(XS)}$ is self-adjoint, the first m eigenfunctions are orthogonal. So, we can normalize the eigenfunctions to construct a set of orthonormal bases of $\mathrm{ran}(\hat{\Sigma}_{(XS)S})$.

C.4 Choice of kernel

For Mean Parity Fair Regression, the choice of κ_S is independent of κ_{XS} and κ_X as long as ϕ_S satisfies Assumption 1. Here we show that a polynomial kernel with degree $k-1$ would satisfy Assumption 1 for $\Omega_S \subseteq \mathbb{R}$, i.e., S is a scalar variable.

Consider a polynomial kernel with degree of $k-1$, i.e., $\kappa_S(s_1, s_2) = (1 + s_1 s_2)^{k-1}$. The feature map $\phi_S(s)$ is

$$\phi_S(s) = [c_0, c_1 s, c_2 s^2, \cdots, c_{k-1} s^{k-1}]$$

where $c_i = \sqrt{\binom{k-1}{i}}$ according to the binomial theorem.

Now we show that $\{\phi_S(s^{(j)})\}_{j=1}^k$ is a set of linearly independent feature maps by showing the following problem has no non-zero solution

$$\sum_{j=1}^{k} w_j \phi_S(s^{(j)}) = 0_{\mathcal{H}_S} \tag{16}$$

Equation 16 is equivalent to

$$\begin{bmatrix} c_0 & c_0 & c_0 & \cdots & c_0 \\ c_1(s^{(1)})^1 & c_1(s^{(2)})^1 & c_1(s^{(3)})^1 & \cdots & c_1(s^{(k)})^1 \\ c_2(s^{(1)})^2 & c_2(s^{(2)})^2 & c_2(s^{(3)})^2 & \cdots & c_2(s^{(k)})^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{k-1}(s^{(1)})^{k-1} & c_{k-1}(s^{(2)})^{k-1} & c_{k-1}(s^{(3)})^{k-1} & \cdots & c_{k-1}(s^{(k)})^{k-1} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

We can simplify the above problem to

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ (s^{(1)})^1 & (s^{(2)})^1 & (s^{(3)})^1 & \cdots & (s^{(k)})^1 \\ (s^{(1)})^2 & (s^{(2)})^2 & (s^{(3)})^2 & \cdots & (s^{(k)})^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ (s^{(1)})^{k-1} & (s^{(2)})^{k-1} & (s^{(3)})^{k-1} & \cdots & (s^{(k)})^{k-1} \end{bmatrix}}_{\boldsymbol{V}} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Since the matrix $\boldsymbol{V}$ is a Vandermonde Matrix, it has determinant $\det(\boldsymbol{V}) = \prod_{1 \le i < j \le k} (s^{(j)} - s^{(i)}) \neq 0$. Therefore, the above problem has no non-zero solution and $\{\phi_S(s^{(j)})\}_{j=1}^k$ is a set of linearly independent features. Thus, a polynomial kernel with degree $k-1$ would satisfy Assumption 1 for scalar-valued S.

For sensitive attributes with non-scalar value, a modified polynomial kernel that first maps S to scalar value and then computes the features by the standard polynomial kernel can be well adopted.

D EXPERIMENTS DETAILS

D.1 Baselines

The details of the baselines used in the experiments are summarized below:

- Constant Prediction: a regression function with a constant outcome that minimizes the MSE. It achieves MP, DP and CB fairness.
- Ordinary Least Squares: the standard linear regression model without regularizers.
- Kernel Ridge Regression: the standard kernel regression (Welling, 2013) method with regularizers.
- Fair Penalty Regression: a regression model with MP-fair regularizers. Derivation can be found in Appendix D.3.

- Fair Kernel Learning (Pérez-Suay et al., 2017): a regularizer-based method aims to eliminate the covariance between the predicted value and the sensitive attributes. The implementation is borrowed from https://isp.uv.es/soft_regression.html.

- Nonconvex Regression with Fairness Constraints (Komiyama et al., 2018): a nonconvex optimization method aims to control the correlation between the predicted value and the sensitive attributes. Note that the optimization process is applied only when the target CB disparity is set to be larger than 0, otherwise, NRFC is reduced to a data preprocessing method. We adapt the official implementation from https://github.com/jkomiyama/fairregresion.

- Reduction Based Algorithm (Agarwal et al., 2019): a reduction based method aims to achieve DP fairness for a randomized predictor using discretization. We adapt the official implementation from https://github.com/steven7woo/fair_regression_reduction.

D.2 Experiment settings

The detailed settings in each experiment are summarized below:

- **Data preprocessing.** For all experiments, both response values in train data and test data are centralized using the mean of training response values.

- **Linear regression.** For synthetic dataset, we choose $n = 2,000$, $d = 5$ and $e = 1$ for regression with single sensitive attribute. The variance of noise ρ^2_{noise} is set to be 0.1. For the proposed method and FPR, we set the kernel of sensitive attributes as the polynomial kernel. All methods focus on the unregularized least-squares problem, i.e, $\lambda = 0$. We test FKR with fairness regularizer coefficients 10 and $1,000$ which are represented by FKR-1 and FKR-3 respectively. Similarly, We test FPR with coefficients of fairness regularizer 10 and $1,000$ which are represented by FPR-1 and FPR-3 respectively. Note that for NRFC, it defaults to fit linear regression with intercept. So, when evaluating other methods, we add a column of ones to X to match the setting of NRFC. Other settings for the hyper-parameters in the baselines follow the default settings of their corresponding papers. We run each method 10 times.

- **Kernel regression.** For the proposed method, we set κ_S to be polynomial kernel while all other kernels are set to be Radial Basis Function (RBF) Kernel with $\gamma = 0.1$. We focus on the regularized least-squares problem with $\lambda = 1$. Other settings for the hyper-parameters in the baselines are the same as the settings in the linear regression experiment.

- **Tradeoff.** The proposed method is evaluated with $\alpha = [0, 1/50, 2/50, \ldots, 1]$. For FKR and FPR, we alter the coefficient of fairness regularizer from 0 to 10^6. Moreover, we run NRFC with ζ, the parameter for the level of fairness from 0 to 1. Note that except α, all other parameters need to be tuned carefully since the relation between fairness and accuracy is hard to interpret (sometimes a small change in the fairness parameter will make a dramatic change to the loss while sometimes the change is negligible). In particularly, the values of ζ concentrate in $[0, 0.1]$ and even $[0, 0.01]$. For the regularizer coefficient of FKR and FPR, the values concentrate in $[10^2, 10^4]$. To make the figures clear, we plot a subset of experiment results in Figure 2 and Figure 7 by subsampling $1/5$ of the results uniformly.

- **Multiple sensitive attributes.** For regression with multiple sensitive attributes on the Communities and Crime dataset, we choose race, medIncome, householdsize and medFamInc as the sensitive attributes sequentially. For medIncome, householdsize and medFamInc, we convert them to binary attributes by whether their values are larger than 0.5.

D.3 Fair penalty regression

In this section, we derive an FPR model for MP fairness using the framework of Pérez-Suay et al. (2017) which is used as a baseline in our experiment. For the FPR model, the key point is to find function $Q(g(X, S), S)$ which measures the level of MP-fairness of a regression function. Notice that a function $g \in \mathcal{H}_{XS}$ satisfies MP fairness if and only if its projection onto $\mathcal{G}_{MP}$ is itself, i.e., $g - Pg = 0_{\mathcal{H}_{XS}}$. So, we set

$$Q(g(X, S), S) = \|g - Pg\|_{\mathcal{H}_{XS}} .$$

Given the training dataset $\mathcal{D} = \{x_i, s_i, y_i\}_{i=1}^n$, we seek the solution of the following regularized optimization problem,

$$\min_{g \in \mathcal{G}_{MP}} \frac{1}{n} \sum_{i=1}^{n} (y_i - g(x_i, s_i))^2 + \frac{\lambda}{n} \|g\|_{\mathcal{H}_{XS}} + \frac{\zeta}{n} \|g - Pg\|_{\mathcal{H}_{XS}} . \tag{17}$$

where $\zeta \geq 0$ is the parameter to control the level of fairness.

By the Representer theorem, the optimal solution g^* is of the form $\Phi_{XS}\boldsymbol{w}$. So, we need to solve the problem

$$\min_{g \in \mathcal{G}_{MP}} \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{Y}^T\boldsymbol{K}_{XS}\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{K}_{XS}\boldsymbol{K}_{XS}\boldsymbol{w} + \lambda\boldsymbol{w}^T\boldsymbol{K}_{XS}\boldsymbol{w} + \zeta\boldsymbol{w}^T\boldsymbol{K}_{XS}\boldsymbol{A}^T\boldsymbol{K}_{XS}\boldsymbol{A}\boldsymbol{K}_{XS}\boldsymbol{w}, \tag{18}$$

where $\boldsymbol{A} = \sum_{j=1}^{m} \boldsymbol{a}_j\boldsymbol{a}_j^T$. Since the above problem is convex, its has a solution

$$\begin{aligned}\boldsymbol{w} &= \left(\boldsymbol{K}_{XS}\boldsymbol{K}_{XS} + \lambda\boldsymbol{K}_{XS} + \zeta\boldsymbol{K}_{XS}\boldsymbol{A}^T\boldsymbol{K}_{XS}\boldsymbol{A}\boldsymbol{K}_{XS}\right)^{\dagger}\boldsymbol{K}_{XS}\boldsymbol{Y} \\ &= \left(\boldsymbol{K}_{XS}\boldsymbol{K}_{XS} + \lambda\boldsymbol{K}_{XS} + \zeta\boldsymbol{K}_{XS}\boldsymbol{A}\boldsymbol{K}_{XS}\right)^{\dagger}\boldsymbol{K}_{XS}\boldsymbol{Y}.\end{aligned}$$

E ADDITIONAL EXPERIMENT RESULTS

E.1 Supplementary results for Section 4

In this section, we provide the supplementary experiment results for Section 4. In Figure 5, we compare different baselines in the kernel regression setting for single binary sensitive attribute. Figure 6 describes the performance of KRR, FPR and the proposed method in the setting of kernel regression for multiple sensitive attributes. In Figure 7, we summarize the experiment results for the fairness-accuracy tradeoff on different datasets.

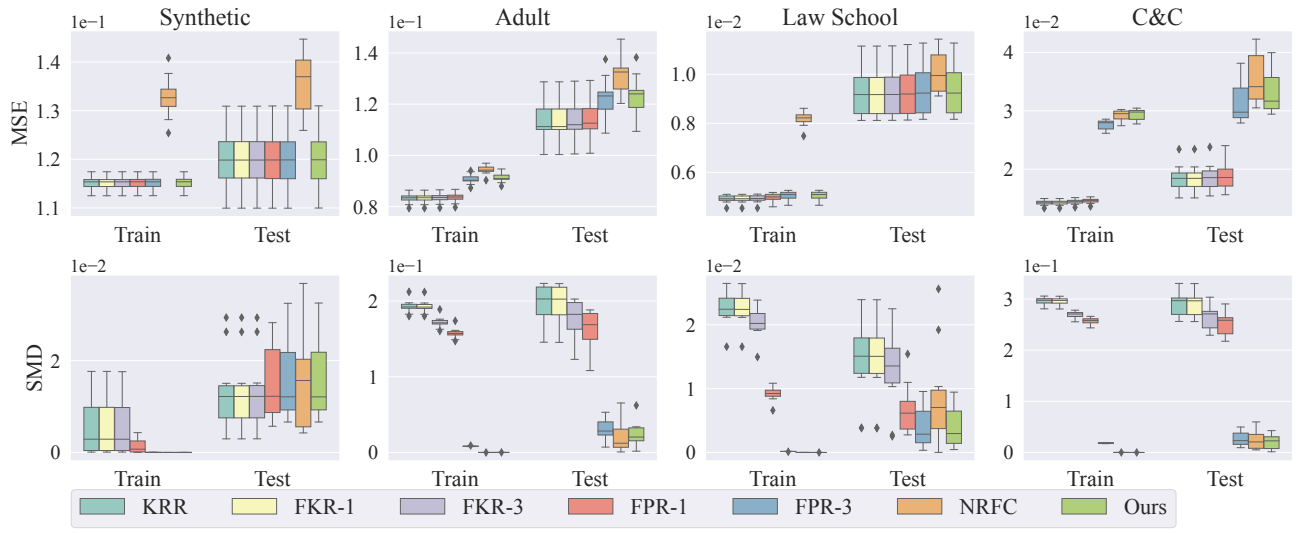

Figure 5: Results of kernel regression for all datasets with one binary sensitive attribute. Figures in the first row show the MSE of different methods whereas the figures in the second row show the SMD of different methods. The methods FKR-1 and FKR-3 stand for FKR method with regularizer coefficients 10 and 1000 respectively. And FPR-1 and FPR-3 stand for FPR method with regularizer coefficients 10 and 1000 respectively.

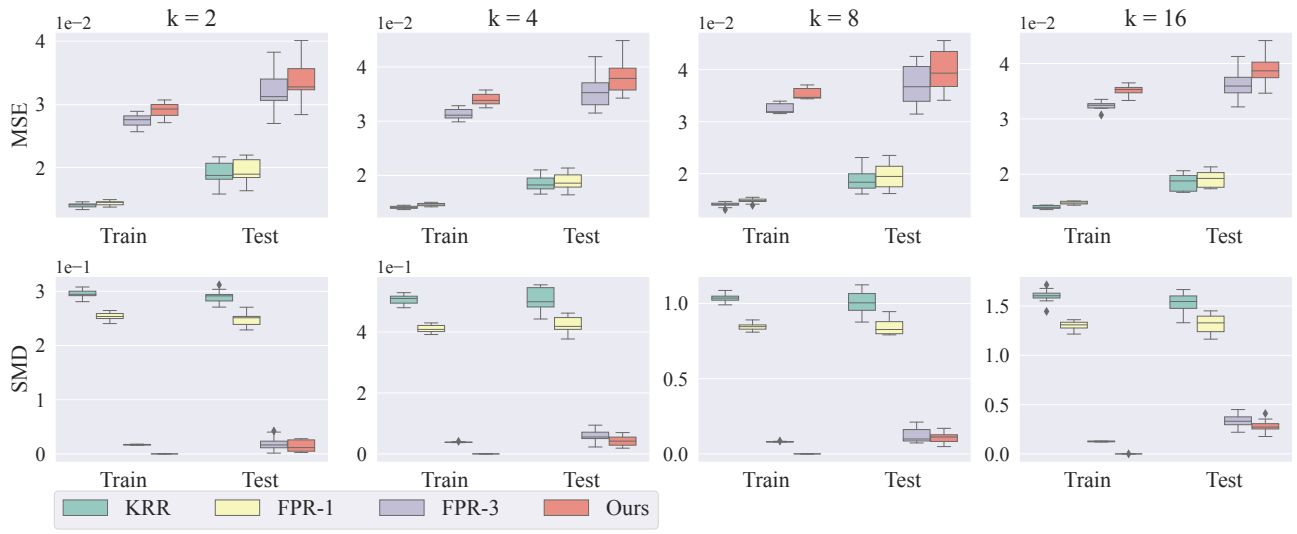

Figure 6: Results of kernel regression on the Communities & Crime dataset with multiple binary sensitive attributes. Figures in the first row show the MSE of different methods whereas the figures in the second row show the SMD of different methods.

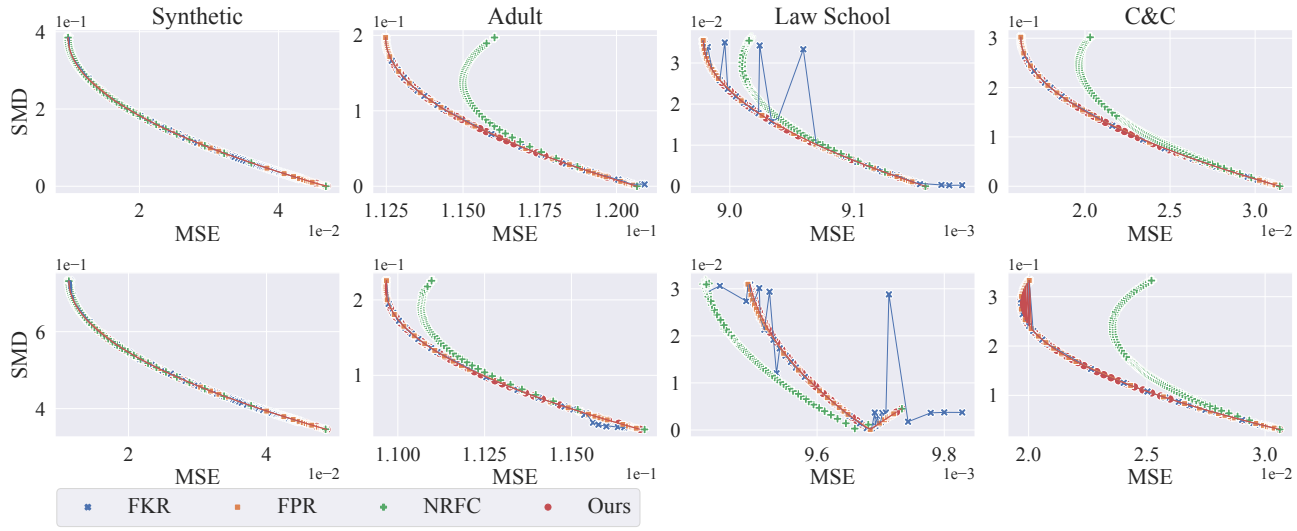

Figure 7: Results of the tradeoff between fairness and accuracy. Figures in the first row show the experiment results on train data whereas the figures in the second row show the tradeoff in test data. In order to compare the results on different datasets, some results in Section 4 are repeated. We remark that both FKR and NRFC suffer from numerical instability with respect to MSE when fairness constraint is removed or strictly imposed which can be seen from experiment results on the Law School dataset.

E.2 Experiments on constant baselines

In this section, we evaluate the baseline with constant prediction equal to the mean of the labels. Since the "Constant Prediction" baseline can achieve perfect MP fairness in both train data and test data, we only compare the MSE of the "Constant Prediction" baseline and our method for simplicity. The experiment results are summarized in Table 1 and Table 2.

The experiment results show that our method significantly outperforms the "Constant Prediction" baseline in all settings, as expected.

We remark that for linear regression on the Synthetic dataset, the MSE of the "Constant Prediction" is about 77× higher

than the MSE of our method since the MSE of "Constant Prediction" baseline is highly dependent on the scale of response.

Method	Metric	Adult	Law School	Communities & Crime	Synthetic
Constant	MSE (Train)	0.1858 ± 0.0016	0.0101 ± 0.0002	0.0544 ± 0.0010	4.5342 ± 0.0450
Constant	MSE (Test)	0.1839 ± 0.0062	0.0103 ± 0.0008	0.0536 ± 0.0039	4.6285 ± 0.1806
Ours	MSE (Train)	0.1175 ± 0.0018	0.0092 ± 0.0002	0.0313 ± 0.0009	0.0585 ± 0.0081
Ours	MSE (Test)	0.1327 ± 0.0072	0.0095 ± 0.0008	0.0344 ± 0.0026	0.0577 ± 0.0081

Table 1: Experiment results on constant baselines for linear regression.

Method	Metric	Adult	Law School	Communities & Crime	Synthetic
Constant	MSE (Train)	0.1858 ± 0.0016	0.0101 ± 0.0002	0.0544 ± 0.0010	0.1474 ± 0.0015
Constant	MSE (Test)	0.1839 ± 0.0062	0.0103 ± 0.0008	0.0536 ± 0.0039	0.1476 ± 0.0061
Ours	MSE (Train)	0.0913 ± 0.0020	0.0050 ± 0.0002	0.0294 ± 0.0010	0.1151 ± 0.0014
Ours	MSE (Test)	0.1232 ± 0.0076	0.0093 ± 0.0010	0.0332 ± 0.0034	0.1202 ± 0.0060

Table 2: Experiment results on constant baselines for kernel regression.

E.3 Experiments on CB fair regression

In this section, we show the experimental results of applying our method to CB fairness. The datasets and experiment settings are the same as in Section 4 except for the choice of κ_S for FPR in the proposed method. In this experiment, we choose κ_S under Assumption 2. We compare the proposed method with baselines in terms of MSE and the Norm of the covariance matrix, i.e.,

$$\textbf{Norm of Cov} = \|\mathrm{Cov}(\phi_S(S), g(X,S))\|_{\mathcal{H}_S}.$$

Figure 8 describes the results for the linear regression case which shows that our method achieves almost the same performance as NRFC. In Figure 9, we can find that the MSE of our methods is much lower than the MSE of NRFC in the train data. However, our method receives higher MSE than NRFC in the test data, which shows an overfitting problem in this setting. A similar trend can be found with respect to the norm of covariance.

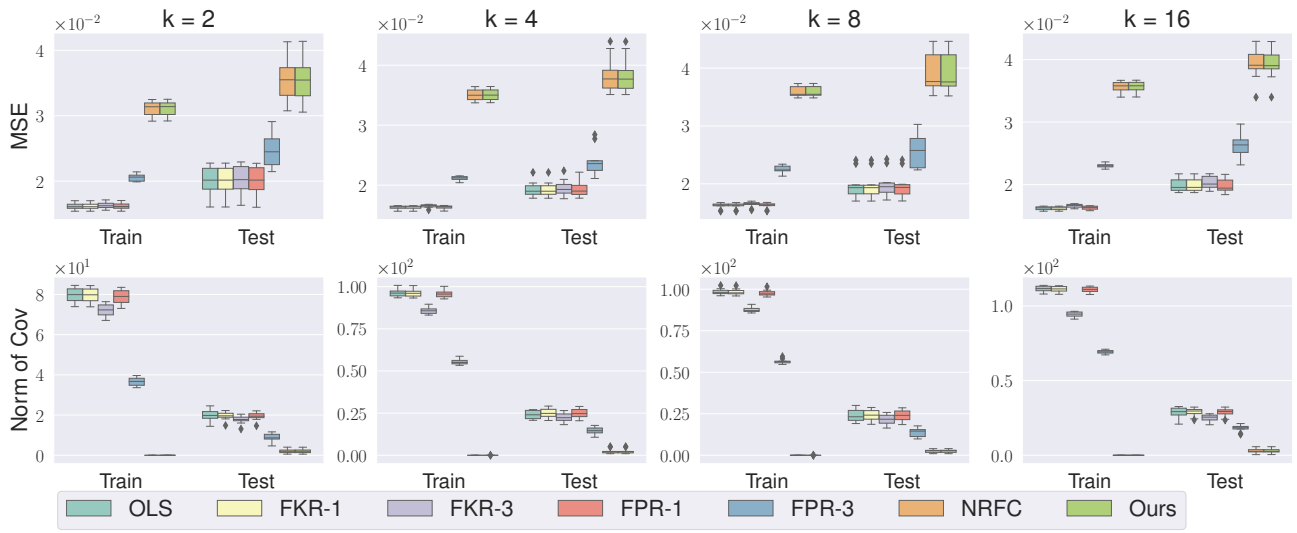

Figure 8: Results of linear regression on the Communities & Crime dataset with multiple binary sensitive attributes. Figures in the first row show the MSE of different methods whereas the figures in the second row show the **Norm of Cov** of different methods.

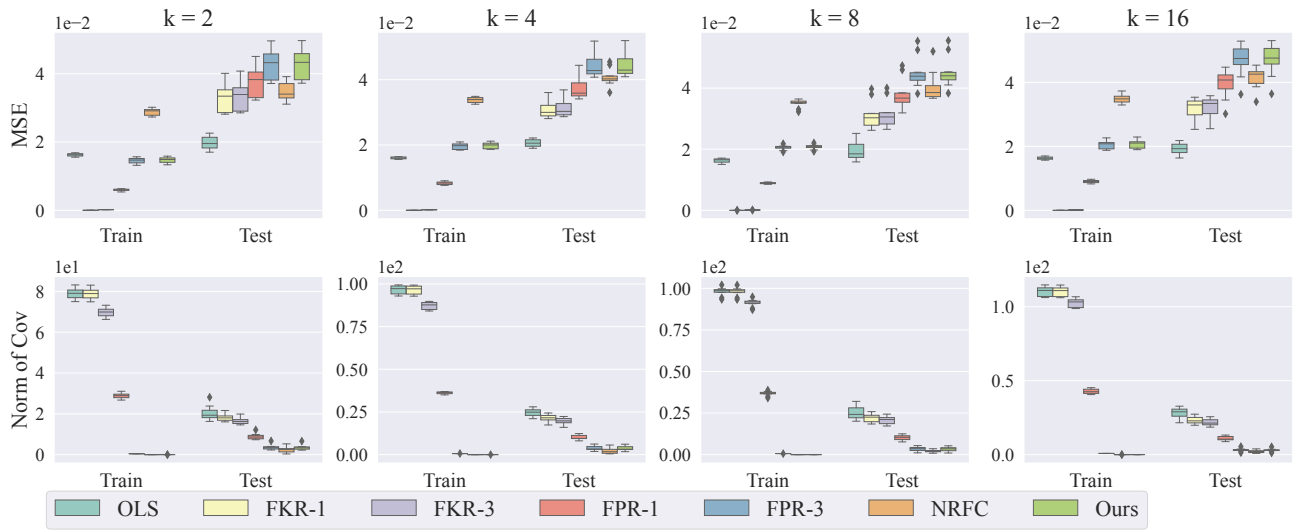

Figure 9: Results of kernel regression on the Communities & Crime dataset with multiple binary sensitive attributes. Figures in the first row show the MSE of different methods whereas the figures in the second row show the **Norm of Cov** of different methods.

E.4 Experiments on DP fairness regression

We also compare the performance of our method with a recent (in-processing) method for DP fairness, i.e., the reduction-based algorithm (RBA, Agarwal et al. (2018)). Note that RBA is designed to produce a DP-fair randomized predictor rather than a simple linear/kernel regression function. We test RBA under the setting of the linear regression with a single binary sensitive attribute, and the experiment results on two benchmark datasets are shown in Figure 10.

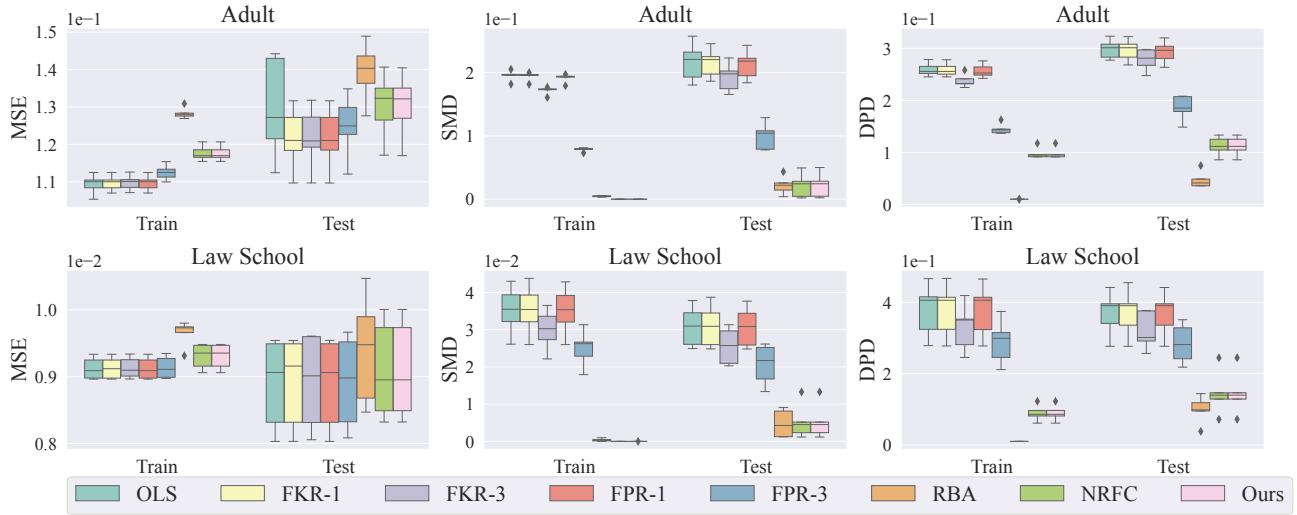

Figure 10: Results under the setting of linear regression with a single binary sensitive attribute.

In this experiment, we found that enforcing DP fairness helps to improve MP fairness and vice versa. However, as DP is a stronger notion of fairness, a DP-fair regression function has a significantly larger cost of fairness, i.e., a larger loss. Note that all algorithms suffer from distribution shifts in the test data, so both MSE, SMD, and DPD are higher in the testing phase. However, since DP fairness is stronger than MP fairness, RAB can achieve comparable and even lower SMD on the test dataset sometimes, even if our algorithm can eliminate MP unfairness in the train data. This motivates us to investigate the generalization problem for fair algorithms in our future work. We remark that our method is almost $200\times$ faster than RBA in the above experiment.

E.5 Experiments on other loss functions

In this section, we evaluate the proposed method on other loss functions using gradient descent (Fair-GD). Specifically, we set the loss function to be Smooth L1 Loss, a commonly used loss function that is less sensitive to outliers than the MSE as it treats error as square only inside an interval. We evaluate Fair-GD in the setting of linear regression with single binary sensitive attribute and compare Fair-GD with the gradient descent (GD) algorithm to show its effect on enforcing fairness. In this experiment, we use Adam (Kingma and Ba, 2014) as our optimizer with a learning rate 1×10^{-4}. The results are summarized in Figure 11 and Figure 12, from which we can see that Fair-GD enjoys the same convergence rate as GD while consistently enforcing the fairness constraint.

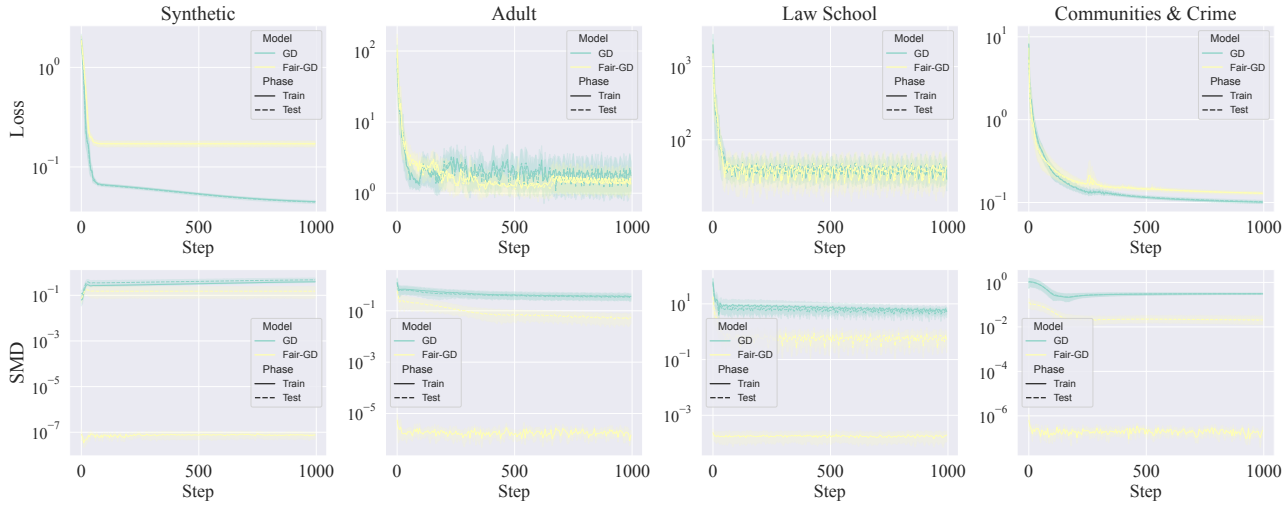

Figure 11: Results of Fair-GD with single binary sensitive attribute using Smooth L1 Loss ($\beta = 0.1$). Figures in the first row show the loss of different methods whereas the figures in the second row show the SMD of different methods.

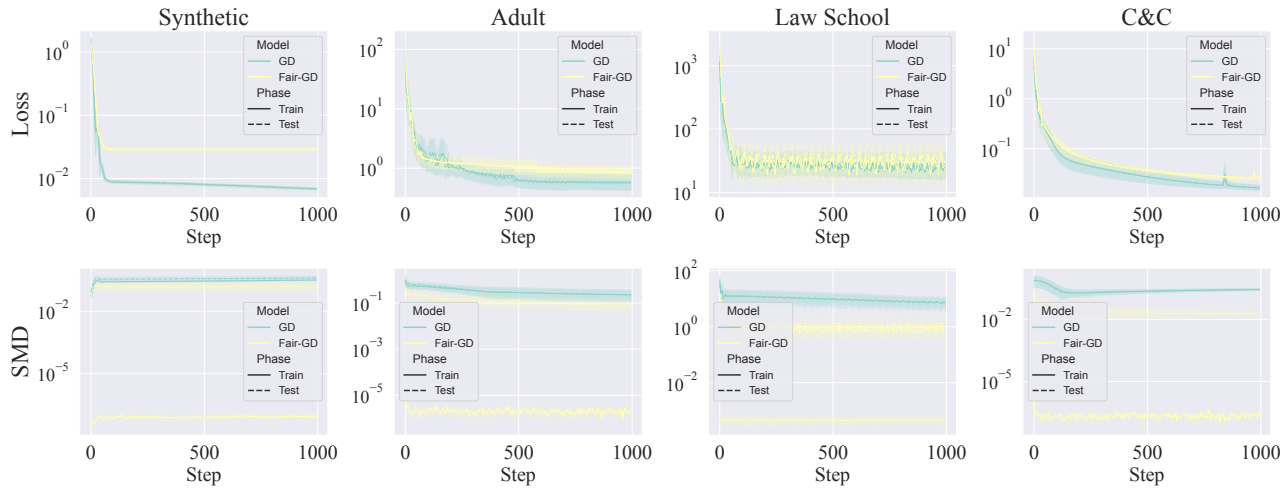

Figure 12: Results of Fair-GD with single binary sensitive attribute using Smooth L1 Loss ($\beta = 1$). Figures in the first row show the loss of different methods whereas the figures in the second row show the SMD of different methods.

E.6 Visualization of distribution

In this section, we provide the visualization of MP-fair response for all datasets as an extension to Figure 4. The results are summarized in Figure 13.

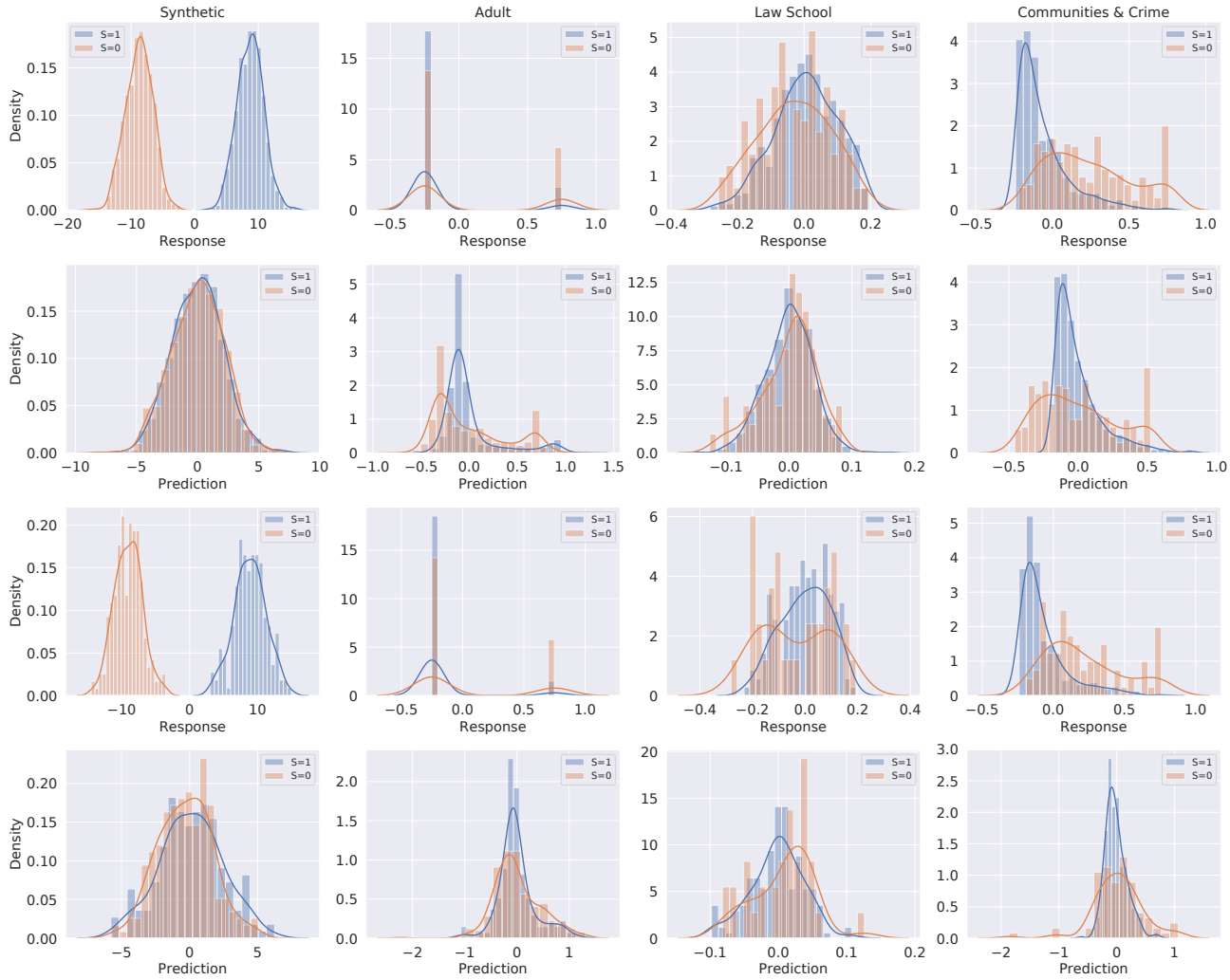

Figure 13: Visualization of response and predicted MP-fair response. Figures in the first row and the third row show the conditional distribution of response variables in the training dataset and test dataset, respectively. Similarly, the figures in the second row and the fourth row show the conditional distribution of the predicted response variables in the training dataset and test dataset, respectively.

E.7 Removing sensitive attributes

In this section, we consider the case of removing the sensitive attributes from the regression function which is a good choice for mitigating unfairness. We remark that such a setting can be regarded as a special case of our general setting. By choosing a kernel κ_{XS} which ignores the input S, i.e., $\kappa_{XS}(X, S) = \kappa_X(X)$, the proposed method can be adapted to fair regression without sensitive attributes. We evaluate the proposed method for regression without inputting sensitive attributes on the linear regression with binary sensitive attribute case and summarize the experiment results in Table 3. Note that we omit the SMD in the training dataset since it is zero in our experiments. The experiment results show that *including the sensitive attribute in regression can help to reduce the MSE while removing the sensitive attribute may help to improve the testing fairness.*

F TABLE OF NOTATIONS

We summarize the notations used throughout this paper in the following table.

	Train		Test			
	MSE w/ S	MSE w/o S	MSE w/ S	MSE w/ S	SMD w/o S	SMD w/o S
Synthetic	**0.0584±0.0081**	2.2994±0.6630	**0.0577±0.0081**	2.4286±0.7560	**0.1508±0.1348**	0.1957±0.1235
Adult	**0.1176±0.0018**	0.1203±0.0022	**0.1327±0.0072**	0.1350±0.0076	**0.0300±0.0243**	0.0355±0.0221
Law School	**0.0092±0.0002**	0.0094±0.0002	**0.0095±0.0008**	0.0097±0.0008	0.0055±0.0034	**0.0042±0.0033**
C&C	**0.0313±0.0009**	0.0375±0.0010	**0.0344±0.0026**	0.0402±0.0028	0.0232±0.0164	**0.0187±0.0087**

Table 3: Experiments on Mean Parity fair linear regression with (w/) and without (w/o) the sensitive attitude S.

Table 4: Table of notations

Notation	Description/Definition
Δ	An arbitrary function in $\mathcal{G}_{MP}$
Σ	The covariance operator
$\hat{\Sigma}$	The empirical estimation of Σ
$\tilde{\Sigma}$	The uncentralized covariance operator
Φ	The feature matrix of the data
Ω	A set from which a random variable is chosen
$\mathcal{F}$	Borel σ-filed on Ω
$\mathbb{E}$	The expectation function
$\mathbb{P}$	The probability function
$\mathbb{R}$	Set of real numbers
$\mathcal{G}_{MP}$	A MP-fair space
A	A linear operator
F	The cumulative distribution function of random variable
J	The generalized objective function for regression problem
L	The mean square loss function
P	The projection operator
$\hat{P}$	The empirical estimation of P
S	Random variable for sensitive attributes
X	Random variable for non-sensitive attributes
Y	Random variable for label/response
$\boldsymbol{H}$	$\boldsymbol{H} = \boldsymbol{I}_{n\times n} - \frac{1}{n}\mathbf{1}_{n\times n}$
$\boldsymbol{K}$	The Gram matrix
$\boldsymbol{P}$	$\boldsymbol{P} = (\boldsymbol{I}_{n\times n} - \sum_{j=1}^{m} \boldsymbol{a}_m \boldsymbol{a}_m^T \boldsymbol{K}_{XS})$
$\boldsymbol{Y}$	The vector of response in dataset
$\hat{\boldsymbol{Y}}$	The predicted value of $\boldsymbol{Y}$
DPD	The DP disparity
MPD	The MP disparity
α	A scalar in $[0, 1]$ to control the accuracy-fairness tradeoff
β	A real number
ϵ	Random noise
ζ	The parameter for fairness penalty term
η	A real number
θ	An (normalized) eigenfunction of $\Sigma_{(XS)S} A \Sigma_{S(XS)}$
κ	Kernel function
λ	The regularization coefficient
μ	A kernel mean embedding
ϕ	Feature map
ψ	Eigenvalue
$\bar{\phi}$	The empirically centralized feature map
$\boldsymbol{a}$	The weight vector for an (normalized) eigenfunction with respect to Φ_{XS}
$\bar{\boldsymbol{a}}$	The weight vector for an (normalized) eigenfunction with respect to $\bar{\Phi}_{XS}$

g, f	Functions
g^*	The least-squares regression function
$g^*_{\mathcal{G}}$	An optimal solution for Problem 1
$\hat{g}^*_{\mathcal{G}}$	The empirical estimation of $g^*_{\mathcal{G}}$
g^{α}	An optimal solution to Problem 5
g_{MP}	The projection of g on $\mathcal{G}_{MP}$
$g_{MP^{\perp}}$	The projection of g on $\mathcal{G}_{MP^{\perp}}$
h	A function defined as $h = \mathbb{E}(\phi_{XS}(X, S)Y)$
k	The cardinality of Ω_S
m	The rank of $\Sigma_{(XS)S}$
n	The number of training samples
s	A realization of S
x	A realization of X
y	A realization of Y
$\boldsymbol{w}$	A weight vector
ℓ	A differentiable loss function
$\ker$	The kernel of a linear operator
ran	The range of a linear operator
$\otimes$	The outer product
$\langle \cdot, \cdot \rangle$	The inner product
$0_{\mathcal{H}}$	The zero function in $\mathcal{H}$
$\dagger$	The Moore-Penrose Inverse of an operator
$\mathbb{I}(\cdot)$	The indicator function
$\perp$	The orthogonal complement of a space
$\perp\!\!\!\perp$	Independence between two random variables