Learning Fair Representations for Kernel Models

Zilong Tan
Carnegie Mellon University
zilongt@cs.cmu.edu

Samuel Yeom
Carnegie Mellon University
syeom@cs.cmu.edu

Matt Fredrikson
Carnegie Mellon University
mfredrik@cs.cmu.edu

Ameet Talwalkar
Carnegie Mellon University
& Determined AI
talwalkar@cmu.edu

Abstract

Fair representations are a powerful tool for establishing criteria like statistical parity, proxy non-discrimination, and equality of opportunity in learned models. Existing techniques for learning these representations are typically *model-agnostic*, as they preprocess the original data such that the output satisfies some fairness criterion, and can be used with arbitrary learning methods. In contrast, we demonstrate the promise of learning a *model-aware* fair representation, focusing on kernel-based models. We leverage the classical Sufficient Dimension Reduction (SDR) framework to construct representations as subspaces of the reproducing kernel Hilbert space (RKHS), whose member functions are guaranteed to satisfy fairness. Our method supports several fairness criteria, continuous and discrete data, and multiple protected attributes. We further show how to calibrate the accuracy tradeoff by characterizing it in terms of the principal angles between subspaces of the RKHS. Finally, we apply our approach to obtain the first Fair Gaussian Process (FGP) prior for fair Bayesian learning, and show that it is competitive with, and in some cases outperforms, state-of-the-art methods on real data.

1 Introduction

Fairness has emerged as a key issue in machine learning as it is increasingly used in areas like hiring [Dastin, 2018], healthcare [Gupta and Mohammad, 2017], and criminal justice [Equivant, 2019]. In particular, models' predictions should not lead to decisions that discriminate on the basis of a legally protected attribute, such as race or gender. Among the proposals to address this issue, a growing body of work focuses on learning fair representations of data for downstream modeling [Calmon et al., 2017, del Barrio et al., 2018, Feldman et al., 2015, Johndrow and Lum, 2019, Kamiran and Calders, 2012]. Most of these approaches are *model agnostic*, which provides flexibility when working with the learned representations, but comes at the cost of potentially suboptimal results in terms of both fairness and accuracy.

In this work, we present a new approach for fair representation learning that takes into account the target hypothesis class of models that will be learned from the representation. Specifically, we show how to leverage information about the reproducing kernel Hilbert space (RKHS) to learn a fair representation for kernel-based models with provable fairness and accuracy guarantees.

Our approach builds on the classic Sufficient Dimension Reduction (SDR) framework [Li, 1991, Cook and Weisberg, 1991, Cook, 1998, Fukumizu et al., 2004, 2009, Wu et al., 2009, Cook and Forzani, 2009] which is used to compute a low-dimensional projection of the feature vector X that captures all information related to the response Y. Our key insight is that we can instead perform SDR with respect to the protected attributes S, and then take the orthogonal complement of the resulting projection to obtain a *fair subspace* of the RKHS that captures information in X unrelated to S. We show that functions in the fair subspace will be independent of S under mild conditions (§ 2.2), and we leverage this fact to prove that our approach

can guarantee several popular definitions of fairness, namely statistical parity [Feldman et al., 2015], proxy nondiscrimination [Datta et al., 2017], equality of opportunity [Hardt et al., 2016], and equalized odds [Hardt et al., 2016]. Moreover, our approach is compatible with both classification and regression, as well as in settings where there are multiple, possibly continuous protected attributes.

Because a fair model might have a lower than desired accuracy in practice, we further generalize our approach to consider this trade-off. In particular, we apply SDR to compute a *predictive subspace* of the RKHS that captures sufficient information in the feature vector X related to the response Y. We then define a third *model subspace* of the RKHS, which is bounded between the fair and predictive subspaces by a specified principal angle [Stewart and Sun, 1990, Golub and Van Loan, 2013]. In contrast to recent regularization- and constraint-based trade-offs [Zemel et al., 2013, Madras et al., 2018, Louizos et al., 2016, Edwards and Storkey, 2016, Song et al., 2019], we provide precise characterizations of how the specified angle affects the fairness and accuracy of any model in this subspace.

Finally, we apply our method to obtain, to the best of our knowledge, the first Fair Gaussian Process (FGP) prior for constructing fair models in the Bayesian setting. Sample paths of the FGP will be functions in the chosen model subspace, and hence satisfy the specified fairness conditions. We identify the covariance kernel of the FGP that corresponds to the chosen model subspace by using the duality between a Gaussian process and its RKHS [Wahba, 1990, Pillai et al., 2007, Tan and Mukherjee, 2018]. Our experiments show that the FGP achieves both rigorous fairness properties and improved accuracy compared to prior methods.

Related Work Much of the prior work on fair representation learning optimizes only for statistical parity [Zemel et al., 2013, del Barrio et al., 2018, Feldman et al., 2015, Johndrow and Lum, 2019, Louizos et al., 2016, Komiyama and Shimao, 2017, Komiyama et al., 2018] or individual fairness [Calmon et al., 2017]. Learned Adversarially Fair and Transferable Representations (LAFTR) [Madras et al., 2018] provides additional support for equality of opportunity and equalized odds by taking into account the model loss while learning the fair representation. The authors prove bounds for statistical parity and equalized odds. It should be noted that their bounds depend on the training data as well as an optimal adversary which may not be available in the non-convex settings. Our approach supports a broader set of fairness criteria (see § 2.1), and we show the generalization performance in terms of both fairness as well as accuracy.

Provably fair kernel learning has been recently studied in Donini et al. [2018] and was generalized to support regression in Oneto et al. [2019]. Both approaches primarily target equality of opportunity in the setting of a single protected attribute. As previously noted, we address a more general setting. Komiyama et al. [2018] also study fair kernel methods, but they only remove linear correlation between the input features X and the protected attribute S; by contrast, we make no such assumption.

Several instances of work on learning rules with explicit fairness constraints or objectives [Kamiran et al., 2010, Zliobaite, 2015, Zafar et al., 2017b] have conducted empirical studies on the fairness-accuracy trade-off, reporting that classifiers trained in this way outperform those trained on model-agnostic fair representations. The fair representations in our work are not model-agnostic, and our performance is competitive if not better in some cases than that of bespoke learning methods. Menon and Williamson [2018] provide a theoretical analysis of the trade-off, providing information-theoretic bounds on accuracy in terms of the correlation between the target and protected attributes, and a regularization parameter analogous to the principal angle between subspaces used to set the tradeoff in our work. In contrast, we provide insights on how the tradeoff impacts generalization performance.

Another approach to fair classification uses randomized post-processing of the classifier's predictions to ensure group fairness criteria. Hardt et al. [2016] propose such a procedure for ensuring equalized odds on binary classifiers. Woodworth et al. [2017] argue that this approach can be suboptimal, and propose an alternative scheme: first learn a classifier with constraints to approximate fairness, and subsequently post-process its predictions to reduce discrimination. These approaches are orthogonal to fair representation learning, and do not consider either regression or multiple protected attributes.

2 Using SDR to Formulate Fairness Desiderata

In this section, we review the basics of the sufficient dimension reduction (SDR) subspace and then demonstrate its implications for several popular fairness definitions. The key idea is to compute a fair subspace as the complement of the SDR subspace with respect to S. We first show in § 2.2 that the fair subspace satisfies the fair representation properties proposed in [Zemel et al., 2013]. Then, we use the fair subspace to enforce several existing fairness criteria in § 2.3.

We begin by introducing some notation. We write $\mathcal{X}$ for the feature space and $\mathcal{S}$ for the space of protected attributes. In addition, $X \in \mathcal{X}$, $S \in \mathcal{S}$, and $Y \in \mathbb{R}$ denote the random variables for the feature vector, protected attributes, and label/response, respectively. We first consider the simple case where $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{S} = \mathbb{R}$, including both the categorical and continuous cases.

2.1 Background on SDR

An m-dimensional vector space is called an *SDR subspace* of $\mathcal{X}$ with respect to S if the projection of X onto the subspace captures the statistical dependency of S on X. Let $\boldsymbol{B}^\top X$ be the projection of X onto the SDR subspace. Then, the SDR condition is given in (1), where $f_S : \mathbb{R}^m \times \mathbb{R}^l \to \mathbb{R}$ is some arbitrary unknown function with $m \leq p$, and the random variable ϵ_S is independent of X.

$$S = f_S\left(\boldsymbol{B}^\top X, \epsilon_S\right) \quad \text{or equivalently} \quad X \perp\!\!\!\perp S \mid \boldsymbol{B}^\top X \tag{1}$$

If $\boldsymbol{B}$ is square non-singular, then (1) is satisfied since there is a one-to-one correspondence between $\boldsymbol{B}^\top X$ and X. The goal is thus to recover a subspace with the lowest dimension. Under mild conditions, this recovery is guaranteed without requiring the knowledge of f_S [Li, 1991, Hall and Li, 1993].

The SDR method also applies to more general settings, including 1) the case where S is multivariate, which entails joint conditions $S_i = f_{S_i}\left(\boldsymbol{B}^\top X, \epsilon_{S_i}\right)$ [Coudret et al., 2014]; and 2) the kernel variant where X and $\boldsymbol{B}_i$ in (1) are respectively replaced by the representer $\kappa(\cdot, X)$ and a function in the RKHS, and the projection $\boldsymbol{B}^\top X$ in (1) is replaced by the vector of inner products $\left(\langle \boldsymbol{B}_1, \kappa(\cdot, X)\rangle, \ldots, \langle \boldsymbol{B}_m, \kappa(\cdot, X)\rangle\right)^\top$ (see e.g., [Ferré and Yao, 2003, Wu et al., 2013] and § 3). Finally, we remark that condition (1) is fairly general and enjoys wide adoption in single and multiple index models [Lin et al., 2017], ridge functions [Pinkus, 2015], projection pursuit regression [Friedman and Stuetzle, 1981], as well as additive models [Hastie and Tibshirani, 1986].

2.2 SDR-Induced Fair Representations

We now illustrate how the SDR framework (1) can be used to obtain a fair representation that satisfies the design goals elaborated in [Zemel et al., 2013]. Specifically, we compute the fair subspace of $\mathbb{R}^p$ from the SDR subspace specified by $\boldsymbol{B}$. Recall that the fair subspace is chosen to capture the "residual" information in X that is complement to $\boldsymbol{B}^\top X$. The projection onto the desired fair subspace is given by $X' \coloneqq \boldsymbol{C}^\top X$, where $\boldsymbol{C}$ is defined as in Proposition 1, which itself follows immediately from Brilliger's lemma [Brillinger, 1977, 1983] with a property of elliptically contoured distributions [Cambanis et al., 1981, Corollary 5].

Proposition 1. *Suppose that* $\mathbb{E}|S| < \infty$ *and* $\mathbb{E}|X_i S| < \infty$ *for* $i = 1, \cdots, p$. *Let the columns of* $\boldsymbol{C}$ *form a basis of the nullspace of* $\operatorname{Var}(X)\boldsymbol{B}$. *If condition* (1) *holds and* X *follows an elliptically contoured distribution, then* $\operatorname{Cov}\left(\boldsymbol{C}^\top X, S\right) = \boldsymbol{0}$.

In the case where X' and S are jointly multivariate normal, the lack of correlation guaranteed by Proposition 1 implies that X' is independent of S. We remark that the multivariate normal requirement of the pair (X', S) is reasonable for high-dimensional X, as most low-dimensional projections of high-dimensional data are nearly normal under mild conditions [Diaconis and Freedman, 1984, Hall and Li, 1993]. Moreover, the high-dimensional condition holds for kernel models where input data is mapped to potentially infinite-dimensional feature space, as we will discuss in § 3.

Now it can be easily verified that X' given by Proposition 1 satisfies the properties of the fair representation proposed in [Zemel et al., 2013]. First, X' is independent of S, as measured on the training data. Second, X' retains information in X that is not in S since the SDR method chooses the lowest-rank $\boldsymbol{B}$. Third, subject to the first two conditions, the model trained using features X' achieves good accuracy. This goal is met by extending the SDR framework (1) to the RKHS setting, which supports a rich class of models. In addition to these properties on training data, the proposed approach enjoys consistency guarantees in terms of generalizing these properties to test data [Hsing and Carroll, 1992, Ferré and Yao, 2003].

2.3 Subspace Formulations of Fairness

In this section, we formulate several common fairness criteria in terms of the statistical independence $X' \perp\!\!\!\perp S$, where X' is the projection of X onto the fair subspace described in § 2.2. In the following, h denotes the model, and $\widehat{Y} = h(X')$ denotes the model output.

Statistical Parity *Statistical parity* (SP), also called *demographic parity*, is one of the simplest notions of fairness and requires model predictions to be independent of the protected attributes, i.e., $\widehat{Y} \perp\!\!\!\perp S$. Since $\widehat{Y}$ is a function of X', this immediately follows from $X' \perp\!\!\!\perp S$.

Proxy Nondiscrimination *Proxy nondiscrimination* [Datta et al., 2017, Yeom et al., 2018] goes further than statistical parity in that it considers all components of the model rather than just its output. A component $\boldsymbol{c}$ of a linear model $h(X) = \boldsymbol{\beta}^\top X$ is the linear combination $\widehat{Y}_{\boldsymbol{c}} \coloneqq \sum_{i=1}^{p} c_i \beta_i X_i$ for $c_i \in [0, 1]$. Let $\widehat{Y}_{\boldsymbol{c}}$ denote the output of component $\boldsymbol{c}$, the strictest version of proxy nondiscrimination requires $\widehat{Y}_{\boldsymbol{c}} \perp\!\!\!\perp S$ for all components of the model. This follows from $X' \perp\!\!\!\perp S$ since $\widehat{Y}_{\boldsymbol{c}}$ is a function of X' for any $\boldsymbol{c}$.

Equalized Odds, Equality of Opportunity In the binary classification setting where $Y \in \{0, 1\}$, the equalized odds (EO) condition [Hardt et al., 2016] is defined as the conditional independence $\widehat{Y} \perp\!\!\!\perp S \mid Y$. Compared to statistical parity, one advantage of equalized odds is that it admits the perfect model $\widehat{Y} = Y$. In comparison, equality of opportunity (EOP) [Hardt et al., 2016] is a relaxation of equalized odds, requiring only that $\widehat{Y} \perp\!\!\!\perp S \mid Y{=}1$. To attain EOP, we can apply (1) to only the individuals with $Y{=}1$, and use the resulting X' as input features. Similarly, we can achieve EO by restricting (1) to individuals with $Y{=}1$ to obtain $\boldsymbol{B}_{Y=1}$ and to individuals with $Y{=}0$ to obtain $\boldsymbol{B}_{Y=0}$. Then, we compute X' by taking the union of SDR subspaces $[\boldsymbol{B}_{Y=1} \;\; \boldsymbol{B}_{Y=0}]$ as the $\boldsymbol{B}$ in Proposition 1.

Unsupported Criteria We conclude the discussion by pointing out that our approach does not support accuracy parity [Zafar et al., 2017a], which requires $\mathbb{1}(\widehat{Y} = Y) \perp\!\!\!\perp S$, or the calibration condition $Y \perp\!\!\!\perp S \mid \widehat{Y}$ [Chouldechova, 2017]. This is because without further assumptions on the model, e.g., consistency, a fair representation alone cannot preclude a constant model, i.e., set $\widehat{Y}$ to be a constant, the resulting relationship between Y and S does not hold in general. Thus, it could be interesting future work to further identify additional conditions on the model needed to support the other fairness definitions.

3 Algorithms

We now present an algorithm to compute the fair subspace as well as the model subspace that attains a desired fairness-accuracy trade-off. We theoretically analyze the algorithm, and provide generalization bounds in Theorem 2 and (10) for the deviation between an optimal fair or predictive model in the RKHS and the model obtained from the model subspace on unseen data. We do not prove the consistency of our SDR estimators—convergence of the SDR subspace estimates to the population quantity. We instead rely on prior consistency results for SDR [Hsing and Carroll, 1992, Ferré and Yao, 2003, Wu et al., 2013].

We first formalize the problem. Let $\mathcal{X}$ be a separable metric space and $\mathcal{H}_\kappa$ be an RKHS of functions $f : \mathcal{X} \to \mathbb{R}$ with trace-class kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, i.e., κ has a finite sum of eigenvalues. A standard learning

problem under Tikhonov regularization is given by [Cucker and Smale, 2002, Hofmann et al., 2008]:

$$\min_{f \in \mathcal{H}_\kappa} L\left(f, \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n\right) + R\left(\|f\|_{\mathcal{H}_\kappa}\right), \tag{2}$$

where L is a convex loss and R is a monotonically increasing regularization function. While $\mathcal{H}_\kappa$ is infinite-dimensional, the well-known representer theorem [Wahba, 1990, Schölkopf et al., 2001] states that the solution $f_\star$ for (2) is in a data-dependent finite-dimensional subspace of $\mathcal{H}_\kappa$:

$$\mathcal{H}_{\kappa,n} \coloneqq \left\{ \sum_{i=1}^n a_i \kappa\left(\cdot, \boldsymbol{x}_i\right) \;\middle|\; \{a_i\}_{i=1}^n \subset \mathbb{R} \right\}. \tag{3}$$

Our goal is to further obtain a subset of functions in $\mathcal{H}_{\kappa,n}$ that meet the fairness desiderata described in § 2.3. This subset will be the fair subspace of $\mathcal{H}_{\kappa,n}$ under the SDR framework (1). In the RKHS setting, we use a functional version of the SDR condition (1) [Ferré and Yao, 2003, Wu et al., 2013]. Specifically, X in (1) is replaced by the representer function $\kappa(\cdot, X)$, and $\boldsymbol{B}_i$ will be functions in $\mathcal{H}_\kappa$ expressed as $\boldsymbol{B}_i = \sum_{j=1}^n W_{ji} \kappa(\cdot, \boldsymbol{x}_j)$ for some $\boldsymbol{W} \in \mathbb{R}^{n \times m}$ by the representer theorem and Riesz representation theorem. The projection $\boldsymbol{B}^\top X$ in (1) becomes accordingly $\left(\kappa(X, \boldsymbol{X})\boldsymbol{W}_1, \cdots, \kappa(X, \boldsymbol{X})\boldsymbol{W}_m\right)^\top$ with $\kappa(X, \boldsymbol{X}) \coloneqq \left(\kappa(X, \boldsymbol{x}_1), \ldots, \kappa(X, \boldsymbol{x}_n)\right)$. Similarly, we also adapt Proposition 1 to the RKHS setting with $\boldsymbol{C}_i$ replaced by $\sum_{j=1}^n Q_{ji} \kappa(\cdot, \boldsymbol{x}_j)$ for some $\boldsymbol{Q} \in \mathbb{R}^{n \times r}$. This yields the corresponding fair subspace $\operatorname{span}\left\{\sum_{i=1}^n Q_{i1}\kappa(\cdot, \boldsymbol{x}_i), \ldots, \sum_{i=1}^n Q_{ir}\kappa(\cdot, \boldsymbol{x}_i)\right\}$. Next, we show how to determine $\boldsymbol{W}$ and $\boldsymbol{Q}$ using the SDR method and Proposition 1.

3.1 Learning the Fair and Predictive Subspaces

We estimate the predictive subspace $\mathcal{G}$ and the fair subspace $\mathcal{F}$ of $\mathcal{H}_\kappa$ using the SDR estimator for RKHS [Tan and Mukherjee, 2018]. For $\mathcal{G}$, we compute the SDR subspace with respect to Y; for $\mathcal{F}$, we compute the SDR subspace with respect to S and then use Proposition 1 to obtain $\boldsymbol{Q}$. Note that the SDR subspace of the infinite-dimensional RKHS $\mathcal{H}_\kappa$ will be a subspace of the finite-dimensional $\mathcal{H}_{\kappa,n}$ under Tikhonov regularization [Wu et al., 2013], a fact we will later use to show generalization properties in Theorem 2 and (10).

Since $\mathcal{G}$ and $\mathcal{F}$ are subspaces of $\mathcal{H}_{\kappa,n}$, we write $\mathcal{G} = \operatorname{span}\{\boldsymbol{\phi}\boldsymbol{A}_1, \cdots, \boldsymbol{\phi}\boldsymbol{A}_d\}$ and $\mathcal{F} = \operatorname{span}\{\boldsymbol{\phi}\boldsymbol{Q}_1, \cdots, \boldsymbol{\phi}\boldsymbol{Q}_r\}$ with $r = n - m$ and $\boldsymbol{\phi} \coloneqq \left(\kappa(\cdot, \boldsymbol{x}_1), \cdots, \kappa(\cdot, \boldsymbol{x}_n)\right)$. Our goal is thus to compute $\boldsymbol{A}$ and $\boldsymbol{Q}$, the latter of which requires the SDR subspace specified by $\boldsymbol{W}$. Next we briefly review the estimation of $\boldsymbol{A}$; $\boldsymbol{W}$ is obtained similarly with respect to S. In the sequel, and without loss of generality, we assume $d + m \leq n$.

Estimating the SDR Subspace The estimate of $\boldsymbol{A}$ is given by the eigenvectors of the following generalized eigenvalue decomposition [Tan and Mukherjee, 2018]:

$$\boldsymbol{\Gamma}_n \boldsymbol{K} \boldsymbol{A}_i = \tau_i \left(\boldsymbol{\Delta} + n\eta \boldsymbol{I}_n\right) \boldsymbol{A}_i, \tag{4}$$

where $\boldsymbol{\Gamma}_n \coloneqq \boldsymbol{I}_n - \boldsymbol{1}_n \boldsymbol{1}_n^\top / n$, $\boldsymbol{K}$ represents the kernel matrix with $K_{ij} \coloneqq \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\eta > 0$ is a regularization parameter, and $\boldsymbol{\Delta}$ is a matrix discussed shortly. For simplicity, (4) assumes that the data tuples $(\boldsymbol{x}_i, \boldsymbol{s}_i, y_i)$ are sorted by y_i, either in ascending or descending order. To obtain $\boldsymbol{\Delta}$, one first partitions the data into slices $\{(\boldsymbol{x}_1, \boldsymbol{s}_1, y_1), \cdots, (\boldsymbol{x}_{n_1}, \boldsymbol{s}_{n_1}, y_{n_1})\}$, $\{(\boldsymbol{x}_{n_1+1}, \boldsymbol{s}_{n_1+1}, y_{n_1+1}), \cdots, (\boldsymbol{x}_{n_1+n_2}, \boldsymbol{s}_{n_1+n_2}, y_{n_1+n_2})\}$, and so forth, where n_i denotes the size of the i-th slice. Then, set $\boldsymbol{\Delta} = \operatorname{diag}\left(\boldsymbol{\Gamma}_{n_i}\right) \boldsymbol{K}$ where $\operatorname{diag}\left(\boldsymbol{\Gamma}_{n_i}\right)$ is the block-diagonal matrix with diagonal blocks $\boldsymbol{\Gamma}_{n_i}$. Clearly, the overall computational complexity for estimating $\boldsymbol{A}$ is $O\left(n^2 d\right)$.

Another relevant problem is to decide the dimension of $\mathcal{G}$ as well as $\mathcal{F}$, i.e., the value of m and d. When y_i (resp. the entries of $\boldsymbol{s}_i$) is categorical with N categories, at most $N-1$ linearly independent directions are needed. Thus, one can set d (resp. m) to $N-1$. In the continuous case, one can use the methods proposed in [Li, 1991, Schott, 1994]. For example, Li [1991] introduced an eigenvalue-based sequential test which can be equivalently written as Theorem 1 [Tan and Mukherjee, 2018]. The conditions of Theorem 1 essentially state that Y depends on X through a latent Gaussian process.

Theorem 1. *Let $\{\boldsymbol{z}_i\}_{i=1}^{\infty}$ be a dense subset of $\mathcal{X}$ and let $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_n$ be the eigenvalues of (4). If there exists a function $g : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ and Gaussian random variables α_i such that $Y = g\left(\sum_{i=1}^{\infty} \alpha_i \kappa\left(X, \boldsymbol{z}_i\right), \epsilon\right)$ for some random variable ϵ independent of $\{\alpha_i\}_{i=1}^{\infty}$. Then, the true SDR subspace dimension $d_\star \geq O_P\left(\frac{n}{n-1}\right)$.*

Estimating the Fair Subspace Given $\boldsymbol{W}$, we invoke Proposition 1 to compute the fair subspace $\mathcal{F}$ which is parameterized by $\boldsymbol{Q}$. In the RKHS setting, the covariance matrix $\mathrm{Var}\left(X\right)$ in Proposition 1 is replaced by the covariance operator $\mathrm{Var}\left(\kappa\left(\cdot, X\right)\right)$ on $\mathcal{H}_\kappa$ whose empirical estimator is written $\mathrm{Var}\left(\kappa\left(\cdot, X\right)\right) \coloneqq \mathbb{E}_X\left[\left(\kappa\left(\cdot, X\right) - \mathbb{E}_X \kappa\left(\cdot, X\right)\right) \otimes \left(\kappa\left(\cdot, X\right) - \mathbb{E}_X \kappa\left(\cdot, X\right)\right)\right] \approx \frac{1}{n} \boldsymbol{\phi}\boldsymbol{\Gamma}_n \otimes \boldsymbol{\phi}\boldsymbol{\Gamma}_n$. Proposition 1 states that for all $i \in [m]$ and $j \in [r]$, $\boldsymbol{Q}$ satisfies $\left\langle \boldsymbol{\phi}\boldsymbol{W}_i, \left(\frac{1}{n}\boldsymbol{\phi}\boldsymbol{\Gamma}_n \otimes \boldsymbol{\phi}\boldsymbol{\Gamma}_n\right) \boldsymbol{\phi}\boldsymbol{Q}_j \right\rangle = \frac{1}{n} \boldsymbol{W}_i^\top \boldsymbol{K}\boldsymbol{\Gamma}_n \boldsymbol{K}\boldsymbol{Q}_j = \boldsymbol{0}$. Thus, the columns of $\boldsymbol{Q}$ are given by a basis of the nullspace of $\boldsymbol{W}^\top \boldsymbol{K}\boldsymbol{\Gamma}_n \boldsymbol{K}$.

A subtlety in estimating the fair subspace for EO and EOP, as described in § 2.3, is that only a subset of the training data with certain value of Y is used. In this case, $\boldsymbol{W}$ and $\boldsymbol{Q}$ both will have a reduced number of rows. This can be easily handled as the fair subspace is still a subspace of $\mathcal{H}_{\kappa,n}$. The issue is addressed in the pseudo-code provided later in Algorithm 1.

3.2 Controlling the Trade-off between Accuracy and Fairness

We now describe a fairness-accuracy trade-off specified by the maximum principal angle between two subspaces of $\mathcal{H}_{\kappa,n}$. Recall that the i-th principal angle θ_i between $\mathcal{F}$ and $\mathcal{G}$ is defined as [Stewart and Sun, 1990, Golub and Van Loan, 2013]:

$$\cos\theta_i \coloneqq \max_{\substack{f_i \in \mathcal{F}, \|f_i\| \leq 1 \\ \forall j < i : \langle f_i, f_j \rangle = 0}} \max_{\substack{g_i \in \mathcal{G}, \|g_i\| \leq 1 \\ \forall j < i : \langle g_i, g_j \rangle = 0}} \langle f_i, g_i \rangle .$$

If the largest principal angle $\max_i \theta_i$ equals 0, $\mathcal{F}$ and $\mathcal{G}$ coincide. Based on this idea, we consider constructing the hypothesis class of the model as a subspace $\mathcal{M}$ of $\mathcal{H}_{\kappa,n}$ such that the largest principal angle between $\mathcal{M}$ and $\mathcal{F}$ is small. Intuitively, functions in $\mathcal{M}$ would be approximately fair.

More formally, our goal is to enforce the distance between $\mathcal{M}$ and $\mathcal{F}$ as measured by the largest principal angle to be no greater than a given threshold. This is equivalent to requiring the cosine of the largest principal angle to be no less than a parameter $0 \leq \epsilon \leq 1$ specified by the user. Recall that the cosines of principal angles are the singular values of the projection of an orthonormal basis of one subspace onto an orthonormal basis of the other [Golub and Van Loan, 2013]. Thus, a direct method for finding an $\mathcal{M}$ that satisfies the principal angle constraint is by reversing the well-known Wedin's bound for the perturbation of singular subspaces [Wedin, 1972]. However, a limitation that inherits from the bound is the dependency on the eigengap. We instead consider a simple construction of $\mathcal{M}$ given by:

$$\mathcal{M} \coloneqq \mathrm{span}\left\{a_i e_i + b_i u_i\right\}_{i=1}^{d}$$

for some orthonormal set of functions $\{e_i\}_{i=1}^{r}$ in $\mathcal{F}$ and orthonormal set of functions $\{u_j\}_{j=1}^{d}$ in $\mathcal{G}$. By careful choices of a_i, b_j, as well as the orthonormal sets, we show that the above hypothesis class satisfies the principal angle constraint as well as several desirable properties.

First, we compute an orthonormal basis for $\mathcal{F}$ and $\mathcal{G}$ by performing the eigenvalue decompositions $\boldsymbol{Q}^\top \boldsymbol{K}\boldsymbol{Q}\boldsymbol{M} = \boldsymbol{M}\boldsymbol{\Lambda}$ and $\boldsymbol{A}^\top \boldsymbol{K}\boldsymbol{A}\boldsymbol{T} = \boldsymbol{T}\boldsymbol{\Omega}$. The columns of $\boldsymbol{M}$ and $\boldsymbol{T}$ are eigenvectors, while $\boldsymbol{\Lambda}$ and $\boldsymbol{\Omega}$ are diagonal containing the corresponding eigenvalues, i.e., $\lambda_i = \Lambda_{ii}$ and $\omega_i \coloneqq \Omega_{ii}$. It is easy to see that $\mathcal{F}$ and $\mathcal{G}$ have the following orthonormal bases:

$$\mathcal{F} = \mathrm{span}\left\{\lambda_i^{-1/2} \boldsymbol{\phi}\boldsymbol{Q}\boldsymbol{M}_i\right\}_{i=1}^{r}, \qquad \mathcal{G} \coloneqq \mathrm{span}\left\{\omega_i^{-1/2} \boldsymbol{\phi}\boldsymbol{A}\boldsymbol{T}_i\right\}_{i=1}^{d}. \tag{5}$$

Using the orthonormal bases (5), Theorem 2 gives the hypothesis space $\mathcal{M}$ for the model which is bounded between the fair RKHS $\mathcal{F}$ and the predictive RKHS $\mathcal{G}$ through ϵ specifying the cosine of the largest principal angle between $\mathcal{M}$ and $\mathcal{F}$.

Theorem 2. *Suppose that* $\boldsymbol{\Lambda}^{-1/2}\boldsymbol{M}^\top\boldsymbol{Q}^\top\boldsymbol{K}\boldsymbol{A}\boldsymbol{T}\boldsymbol{\Omega}^{-1/2} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ *is the thin singular value decomposition with singular values* $\sigma_i \coloneqq \Sigma_{ii}$*, and let the hypothesis class of the model be*

$$\mathcal{M} = \operatorname{span}\left\{\boldsymbol{\phi}\left[\left(\gamma_i - \rho_i\sigma_i\right)\boldsymbol{Q}\boldsymbol{M}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}_i + \rho_i\boldsymbol{A}\boldsymbol{T}\boldsymbol{\Omega}^{-1/2}\boldsymbol{V}_i\right]\right\}_{i=1}^{d} \tag{6}$$

with $\gamma_i \coloneqq \max\{\sigma_i, \epsilon\}$*, and* $\rho_i \coloneqq \sqrt{\frac{1-\gamma_i^2}{1-\sigma_i^2}}$ *if* $\sigma_i < 1$ *and* $\rho_i \coloneqq 0$ *if* $\sigma_i = 1$ *for* $i = 1, 2, \cdots, d$*. Denote by* $\sigma_{min} \coloneqq \min_i \sigma_i$ *and let* $\mathcal{P}_\mathcal{F}$*,* $\mathcal{P}_\mathcal{G}$*, and* $\mathcal{P}_\mathcal{M}$ *be the orthogonal projection operators onto the fair RKHS* $\mathcal{F}$*, predictive RKHS* $\mathcal{G}$*, and the model RKHS* $\mathcal{M}$*, respectively. Then, the following operator norms hold:*

$$\|\mathcal{P}_\mathcal{F} - \mathcal{P}_\mathcal{M}\| = \sqrt{1 - \max\{\epsilon^2, \sigma_{min}^2\}} \tag{7}$$

$$\|\mathcal{P}_\mathcal{G} - \mathcal{P}_\mathcal{M}\| = \max\left\{0, \epsilon\sqrt{1-\sigma_{min}^2} - \sigma_{min}\sqrt{1-\epsilon^2}\right\}. \tag{8}$$

Proof of Theorem 2. We first show that the basis (6) of the classifier hypothesis RKHS is orthonormal, and then compute the canonical angles using the basis (6) and an orthonormal basis of $\mathcal{F}$. Denote by $\xi_i \coloneqq \left(\gamma_i - \rho_i\sigma_i\right)\boldsymbol{Q}\boldsymbol{M}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}_i + \rho_i\boldsymbol{A}\boldsymbol{T}\boldsymbol{\Omega}^{-1/2}\boldsymbol{V}_i$ the i-th basis function in (6), we have

$$\begin{aligned}
\langle\xi_i, \xi_j\rangle &= \left(\gamma_i - \rho_i\sigma_i\right)\left(\gamma_j - \rho_j\sigma_j\right)\boldsymbol{U}_i^\top\boldsymbol{U}_j + \rho_i\rho_j\boldsymbol{V}_i^\top\boldsymbol{V}_j + 2\rho_i\sigma_i\left(\gamma_i - \rho_i\sigma_i\right)\mathbb{1}_{i=j} \\
&= \left[\left(\gamma_i - \rho_i\sigma_i\right)^2 + \rho_i^2 + 2\rho_i\sigma_i\left(\gamma_i - \rho_i\sigma_i\right)\right]\mathbb{1}_{i=j} \\
&= \mathbb{1}_{i=j},
\end{aligned}$$

where the first equality follows from the orthonormal basis (5). This shows that (6) is an orthonormal basis. Using the orthonormal basis $\left\{\psi_i \coloneqq \boldsymbol{\phi}\boldsymbol{Q}\boldsymbol{M}\boldsymbol{\Lambda}_i^{-1/2}\right\}_{i=1}^{r}$ of $\mathcal{F}$, we can use the SVD to compute the canonical angles (see e.g., Algorithm 6.4.3 in Golub and Van Loan, 2013) as

$$\left[\xi_1, \cdots, \xi_d\right]^\top\left[\psi_1, \cdots, \psi_r\right] = \operatorname{diag}\left(\gamma_i - \rho_i\sigma_i\right)\boldsymbol{U}^\top + \operatorname{diag}\left(\rho_i\right)\boldsymbol{\Sigma}\boldsymbol{U}^\top = \boldsymbol{I}_d\operatorname{diag}\left(\gamma_i\right)\boldsymbol{U}^\top. \tag{9}$$

Here, $\operatorname{diag}\left(d_i\right)$ denotes the diagonal matrix with diagonal elements d_i. Note that the last term in (9) is the (thin) SVD, and the singular values γ_i are the canonical angles between $\mathcal{M}$ and $\mathcal{F}$. Finally, we relate the canonical angles to the operator norm in (7). Recall that the orthogonal projector can be expressed as the tensor product $\mathcal{P}_\mathcal{F} = \sum_{i=1}^r \psi_i \otimes \psi_i$, and $\mathcal{P}_\mathcal{F} h = \sum_{i=1}^r \langle h, \psi_i\rangle\,\psi_i$. We have

$$\begin{aligned}
\|\mathcal{P}_\mathcal{F} - \mathcal{P}_\mathcal{M}\| &= \|\left(\mathcal{P}_\mathcal{F} + \mathcal{P}_{\mathcal{F}_\perp}\right)\left(\mathcal{P}_\mathcal{F} - \mathcal{P}_\mathcal{M}\right)\left(\mathcal{P}_\mathcal{M} + \mathcal{P}_{\mathcal{M}_\perp}\right)\| \\
&= \|\mathcal{P}_\mathcal{F}\mathcal{P}_{\mathcal{M}_\perp} - \mathcal{P}_{\mathcal{F}_\perp}\mathcal{P}_\mathcal{M}\| \\
&= \sup_{h\in\mathcal{H}_{\kappa,n}:\,\|h\|\le 1}\left(\|\mathcal{P}_\mathcal{F}\mathcal{P}_{\mathcal{M}_\perp}h\| + \|\mathcal{P}_{\mathcal{F}_\perp}\mathcal{P}_\mathcal{M}h\|\right),
\end{aligned}$$

where $\mathcal{F}_\perp$ and $\mathcal{M}_\perp$ represent respectively the orthogonal complements of $\mathcal{F}$ and $\mathcal{M}$. It can be shown that $\mathcal{P}_\mathcal{F}\mathcal{P}_{\mathcal{M}_\perp}$ and $\mathcal{P}_\mathcal{F}\mathcal{P}_{\mathcal{M}_\perp}$ have the same nonzero singular values which are the sines of the principal angles between $\mathcal{F}$ and $\mathcal{M}$ (see e.g., p.249 of Stewart, 2001). From (9), these principal angles are $\arccos\left(\gamma_i\right)$. Thus, we obtain $\|\mathcal{P}_\mathcal{F} - \mathcal{P}_\mathcal{M}\| = \sqrt{1 - \min_i \gamma_i^2}$. To obtain (8), one can simply apply the trigonometric identity of sines yielding

$$\|\mathcal{P}_\mathcal{G} - \mathcal{P}_\mathcal{M}\| = \gamma_{\min}\sqrt{1-\sigma_{\min}^2} - \sigma_{\min}\sqrt{1-\gamma_{\min}^2} = \max\left\{0, \epsilon\sqrt{1-\sigma_{\min}^2} - \sigma_{\min}\sqrt{1-\epsilon^2}\right\},$$

where we denote by $\gamma_{\min} \coloneqq \min_i \gamma_i$. ∎

For the case that $\epsilon = 1$, the basis of (6) are linear combinations of the orthonormal basis of $\mathcal{F}$ in (5), and hence $\mathcal{M} \subset \mathcal{F}$. Additionally, it can be verified that $\|\mathcal{P}_\mathcal{F} - \mathcal{P}_\mathcal{M}\| = 0$ from (7), and (8) becomes

Algorithm 1: $\boldsymbol{E} = \texttt{MBasis}\left(\boldsymbol{K}, \boldsymbol{y}, \boldsymbol{S}, m, d, \epsilon\right)$ — Compute the basis $\boldsymbol{\phi E}$ for $\mathcal{M}$

Initialize $\boldsymbol{W} = [\,]$, n with the number of rows of $\boldsymbol{K}$, as well as indices $\texttt{pos} = (\boldsymbol{y} == 1)$ and $\texttt{neg} = (\boldsymbol{y} \neq 1)$.
foreach *column* $\boldsymbol{s}$ *of* $\boldsymbol{S}$ **do**
  if *EqualizedOdds* **or** *EqualityOfOpportunity* **then**
    Set $\boldsymbol{B} = \boldsymbol{0}_{n\times m}$ and update $\boldsymbol{B}\left(\texttt{pos}, :\right) = \texttt{SDR}\left(\boldsymbol{K}\left(\texttt{pos}, \texttt{pos}\right), \boldsymbol{s}, m\right)$.
    Append basis $\boldsymbol{B}$ to $\boldsymbol{W}$: $\boldsymbol{W} = [\boldsymbol{W}\ \ \boldsymbol{B}]$.
    if *EqualizedOdds* **then**
      Set $\boldsymbol{B} = \boldsymbol{0}_{n\times m}$, then $\boldsymbol{B}\left(\texttt{neg}, :\right) = \texttt{SDR}\left(\boldsymbol{K}\left(\texttt{neg}, \texttt{neg}\right), \boldsymbol{s}, m\right)$.
      Let $\boldsymbol{W} = [\boldsymbol{W}\ \ \boldsymbol{B}]$.
    end
  else
    Compute $\boldsymbol{B} = \texttt{SDR}\left(\boldsymbol{K}, \boldsymbol{s}, m\right)$, and update $\boldsymbol{W} = [\boldsymbol{W}\ \ \boldsymbol{B}]$.
  end
end
Predictive Subspace: Compute the predictive subspace as the SDR subspace $\boldsymbol{A} = \texttt{SDR}\left(\boldsymbol{K}, \boldsymbol{y}, d\right)$.
Fair Subspace: Obtain $\boldsymbol{K}'$ by subtracting the mean of each column of $\boldsymbol{K}$. Let $\tilde{\boldsymbol{K}} = \boldsymbol{K}'^{\top}\boldsymbol{K}'$, and use QR decomposition to compute $\boldsymbol{Q}$ as the nullspace basis of the column space of $\tilde{\boldsymbol{K}}\boldsymbol{W}$.
Perform the eigenvalue decompositions to obtain Equation (5), and then use Theorem 2 to compute $\boldsymbol{E}$.

Algorithm 2: $\boldsymbol{W} = \texttt{SDR}\left(\boldsymbol{K}, \boldsymbol{s}, m\right)$ — Compute the SDR subspace $\boldsymbol{\phi W}$

Sort $\boldsymbol{s}$ such that $s\left(\texttt{idx}\right)$ is non-decreasing. Let $\texttt{invIdx}$ be the inverse of $\texttt{idx}$ satisfying $\texttt{idx}\left(\texttt{invIdx}\right) = 1:n$, where n is the number of rows of $\boldsymbol{K}$.
Slice $\boldsymbol{s}$ approximately evenly as described in § 3.1 such that entries with the same value are in the same partition. Denote by n_i the size of partition i.
Initialize $\eta = 10^{-4}$, i.e., a small constant. Let $\boldsymbol{K}' \coloneqq \boldsymbol{K}\left(\texttt{idx}, \texttt{idx}\right)$, and solve $\boldsymbol{\Gamma}_n\boldsymbol{K}'\boldsymbol{A}_i = \tau_i\left[\operatorname{diag}\left(\boldsymbol{\Gamma}_{n_i}\right)\boldsymbol{K}' + n\eta\boldsymbol{I}_n\right]\boldsymbol{A}_i$ for $\boldsymbol{A}$.
Return $\boldsymbol{W} = \boldsymbol{A}\left(\texttt{invIdx}, :\right)$.

$\|\mathcal{P}_{\mathcal{G}} - \mathcal{P}_{\mathcal{M}}\| = \sqrt{1-\sigma_{\min}^2}$ which is the sine of the largest principal angle between $\mathcal{F}$ and $\mathcal{G}$ as desired. Similarly, we have $\mathcal{M} = \mathcal{G}$ by setting $\epsilon = 0$.

The key utility of Theorem 2 involves bounding the difference between the model obtained using $\mathcal{M}$ and an ideal fair (or predictive) model. For instance, let $f_{\text{fair}} \in \mathcal{H}_\kappa$ denote the ideal fair model and let $\delta_{\boldsymbol{x}} f \coloneqq f\left(\boldsymbol{x}\right)$ be the evaluation functional, then for any $\boldsymbol{x} \in \mathcal{X}$:

$$\begin{aligned}\left|\left(\mathcal{P}_{\mathcal{M}} f_{\text{fair}}\right)\left(\boldsymbol{x}\right) - f_{\text{fair}}\left(\boldsymbol{x}\right)\right| &= \left\|\delta_{\boldsymbol{x}}\left(\mathcal{P}_{\mathcal{M}} f_{\text{fair}}\right) - \delta_{\boldsymbol{x}} f_{\text{fair}}\right\| \leq \left\|\delta_{\boldsymbol{x}}\right\| \left\|\mathcal{P}_{\mathcal{M}} f_{\text{fair}} - f_{\text{fair}}\right\| \\ &= \left\|\delta_{\boldsymbol{x}}\right\| \left\|\mathcal{P}_{\mathcal{M}} f_{\text{fair}} - \mathcal{P}_{\mathcal{F}} f_{\text{fair}} + \mathcal{P}_{\mathcal{F}} f_{\text{fair}} - f_{\text{fair}}\right\| \\ &\leq \left\|\delta_{\boldsymbol{x}}\right\| \left(\left\|\mathcal{P}_{\mathcal{M}} - \mathcal{P}_{\mathcal{F}}\right\| \left\|f_{\text{fair}}\right\| + \left\|\mathcal{P}_{\mathcal{F}} f_{\text{fair}} - f_{\text{fair}}\right\|\right).\end{aligned} \tag{10}$$

While $\mathcal{P}_{\mathcal{F}}$ is based on training data, the last norm in (10) converges to zero from the consistency of $\mathcal{F}$ and $\mathcal{G}$ [Hsing and Carroll, 1992, Ferré and Yao, 2003, Wu et al., 2013]. Together with (7), (10) sheds light on the impact of ϵ on the generalization of the fairness criteria; similar arguments can be made for an ideal predictive model.

Algorithm 1 gives the Matlab-style pseudo-code for our approach which can handle multiple protected attributes. This algorithm use the SDR procedure described in Algorithm 2 to compute the desired model representation with a specified trade-off ϵ.

4 Application to Fair Gaussian Processes

In this section, we apply our approach to Gaussian Processes (GPs), providing a prior over functions in the model subspace $\mathcal{M}$. We call this GP the FGP. The FGP can then be used to develop a rich class of fair models in the standard Bayesian setting [Rasmussen and Williams, 2006].

The covariance kernel is the critical component of a GP that characterizes the class of functions the GP can realize. The key to deriving the FGP is to identify the covariance kernel such that sample paths (realizations) of the FGP will be functions in $\mathcal{M}$. We construct the FGP using the integral representation of GPs [Pillai et al., 2007, Tan and Mukherjee, 2018] which ensures that the sample paths are in a given RKHS defined by a reproducing kernel $\kappa_{\mathcal{M}}$.

Without loss of generality, consider a zero-mean GP $\{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}$ on a probability space $(\Omega, \mathcal{F}, P)$, that is, the function value $f(\boldsymbol{x})$ is a Gaussian random variable. The integral representation of f is given by the Bochner integral

$$f(\cdot) = \int_{\mathcal{X}} \kappa_{\mathcal{M}}(\cdot, \boldsymbol{z})\, \nu(\boldsymbol{z})\, d\mu(\boldsymbol{z}), \tag{11}$$

where μ denotes the measure on $\mathcal{X}$, and $\nu : \mathcal{X} \times \Omega \mapsto \mathbb{R}$ is another GP on $(\Omega, \mathcal{F}, P)$ whose covariance needs to be estimated. It has been shown that sample paths of the GP (11) are contained in the RKHS with kernel $\kappa_{\mathcal{M}}$ [Pillai et al., 2007, Tan and Mukherjee, 2018]. By letting $\kappa_{\mathcal{M}}$ in (11) be the reproducing kernel of $\mathcal{M}$, the resulting GP is the desired FGP which inherits the fairness as well as accuracy guarantees of $\mathcal{M}$.

In practice, we use the sample variant of (11) expressed as $f_n(\boldsymbol{x}) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \nu(\boldsymbol{x}_i)\, \kappa_{\mathcal{M}}(\boldsymbol{x}, \boldsymbol{x}_i)$, which converges to (11) at a rate $O_P\left(n^{-1/2}\right)$ by the central limit theorem for Hilbert spaces [Ledoux and Talagrand, 1991]. Now we further rewrite the sample FGP in terms of the kernel κ of $\mathcal{H}_\kappa$. Specifically, let $\boldsymbol{\phi}\boldsymbol{E}_i$ represent the i-th basis function of (6), and denote by $\boldsymbol{\Pi}(\boldsymbol{z}) \coloneqq \left(\kappa(\boldsymbol{z}, \boldsymbol{X}) - \mathbf{1}_n^\top \boldsymbol{K}/n\right)\boldsymbol{E}$, where $\boldsymbol{K}$ is the kernel matrix of κ as defined in (4). Then, the sample FGP can be written as shown in (12).

$$f_n(\cdot) \sim \mathcal{GP}\left(\mathbf{0}, \boldsymbol{\Pi}(\cdot)\, \boldsymbol{\Sigma}_\beta \boldsymbol{\Pi}(\cdot)^\top\right). \tag{12}$$

Here, the sample FGP formulation is an instance of the integral GP, and $\boldsymbol{\Sigma}_\beta$ is a parameter of the covariance kernel which can be estimated efficiently [Tan and Mukherjee, 2018].

5 Experiments

We present experiments on five real datasets to: *(1)* demonstrate the efficacy of our approach in mitigating discrimination while maintaining prediction accuracy; *(2)* characterize the empirical behavior of the algorithms developed in § 3; and *(3)* highlight the ability of our method to handle multiple, possibly continuous, protected attributes.

We adapt the experimental setup and configurations used in prior work [Donini et al., 2018, Komiyama et al., 2018] to compare the proposed FGP against several approaches: a standard GP trained on an adversarially-fair representation [Madras et al., 2018] (LAFTR-GP), fairness-constrained ERM [Donini et al., 2018], both the linear (Linear-FERM) and nonlinear (FERM) variants, and non-convex fair regression [Komiyama et al., 2018] (NCFR) which supports settings with multiple protected attributes. We also report the results of a standard GP with no fairness objective.

We measure fairness conditions empirically using the absolute correlation coefficient, as it can be generalized to the regression setting. Specifically, we compute the population $\left|\mathrm{Corr}\left(\widehat{Y}, S\right)\right|$ as the SP risk score, $\left|\mathrm{Corr}\left(\widehat{Y}, S\right)\right|$ on individuals with $Y = 1$ for EOP, and EO is given by the maximum absolute correlation on individuals with $Y = 1$ and $Y = 0$. All scores are calculated on holdout test data. For the experiments, we do not consider proxy non-discrimination as it relies on certain model structures, and is not comparable to our chosen baselines. Finally, we used the online code for the baseline methods, and equip GPs with a linear mean and a radial basis covariance. The datasets and our Matlab implementation of the FGP are available at `https://github.com/ZilongTan/fgp`.

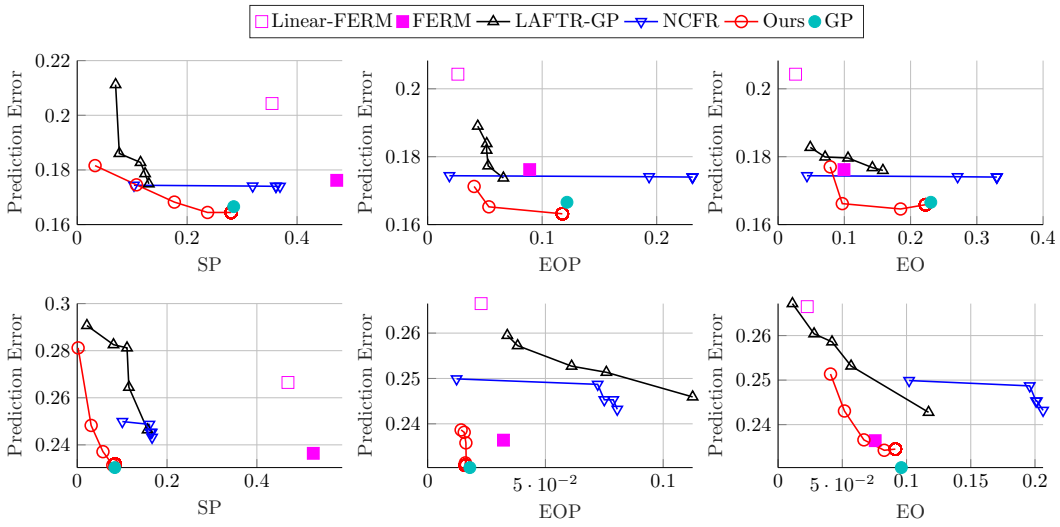

Figure 1: Comparing the accuracy-fairness trade-offs on the Adult (first row) and Compas (second row) datasets. The prediction error denotes the misclassification rate.

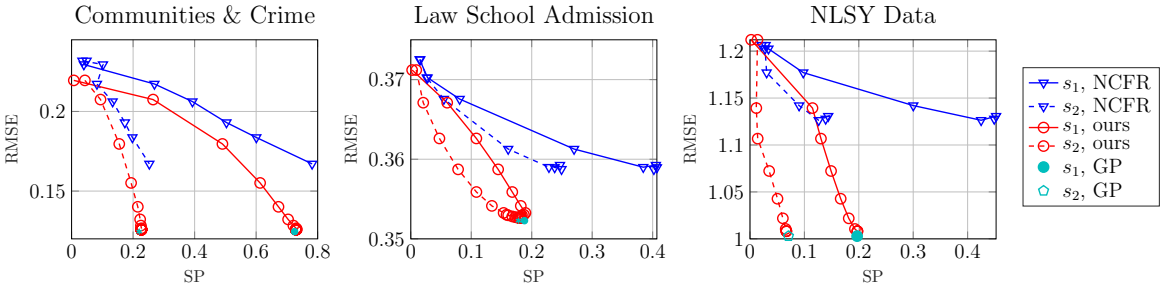

Figure 2: Regression results with two protected attributes s_1 and s_2.

Fair Classification A primary goal of our approach is to enforce a specified fairness criterion with minimal loss in accuracy. We evaluate how each of the fairness conditions are satisfied empirically on two standard datasets: the Adult income dataset [Lichman, 2013], and the Compas recidivism risk score data [Angwin et al., 2016]. We illustrate how the fairness score and prediction error react to various choices of the fairness-accuracy tradeoff. For both datasets, we use gender as the single protected attribute.

Figure 1 compares different methods for achieving each fairness goal (column). First, observe that Linear-FERM and FERM do not meet SP on both datasets. This is because Linear-FERM and FERM only target EOP. Also note that our approach approaches the accuracy of standard GP, which is expected as setting the trade-off $\epsilon = 0$ in (6) yields this model. Some baselines attain the the best fairness, e.g., LAFTR-GP delivers the lowest EO on Compas, at the cost of accuracy. However, overall our approach generally achieves greater accuracy for a given level of fairness.

Fair Regression with Multiple Protected Attributes We consider a regression setting with two protected attributes. For this evaluation, we used three real datasets with continuous target values, including the UCI Communities and Crime dataset [Redmond and Baveja, 2002], the National Longitudinal Survey of Youth (NLSY) dataset [Bureau of Labor Statistics, 2014] and the Law School Admissions Council dataset [Law School Admissions Council]. The protected attribute pairs $\{s_1, s_2\}$ used for these datasets are respectively $\{race, origin\}$, $\{gender, age\}$, and $\{race, age\}$.

Note that NCFR is the only baseline that handles multiple protected attributes. In addition, EO and EOP are defined in the context of binary classification, and thus may not be used in this regression experiment. We use the root mean squared error (RMSE) to measure the prediction error.

Figure 2 depicts the prediction error as well as SP for each protected attribute. As stricter fairness conditions are enforced, the RMSE climbs. Across these datasets, the proposed method achieves consistently improved accuracy. Interestingly, the curves correspond to our approach are generally steeper than the curves of NCFR, suggesting more effective fairness-accuracy trade-offs.

6 Conclusions

We have presented a novel and theoretically principled method for learning fair representations for kernel models, which also enables users to systemically navigate the accuracy-fairness tradeoff. We apply our approach to obtain a Fair Gaussian Process, demonstrating competitive empirical performance on several datasets relative to state-of-the-art methods. Our work hinges on the idea of learning a model-aware representation, along with the key insight that several popular fairness notations can be reformulated as sufficient dimension reduction (SDR) problems. Future work involves supporting additional fairness notions like calibration and accuracy parity through additional model assumptions, developing more scalable algorithms using randomized approximations [Rahimi and Recht, 2008], and generalizing the strategy of learning model-aware representations to other model classes.

Acknowledgements

This work was supported in part by DARPA FA875017C0141, the National Science Foundation grants IIS1705121, IIS1838017, CNS1704845, an Okawa Grant, a Google Faculty Award, an Amazon Web Services Award, a JP Morgan A.I. Research Faculty Award, and a Carnegie Bosch Institute Research Award. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA, the National Science Foundation, or any other funding agency.

References

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. How we analyzed the compas recidivism algorithm, 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

David Brillinger. A generalized linear model with "Gaussian" regressor variables. *A Festschrift for Erich L. Lehmann*, 1983.

David R. Brillinger. The identification of a particular nonlinear time series system. *Biometrika*, 64(3): 509–515, 1977.

Bureau of Labor Statistics. National Longitudinal Survey of Youth 1979 cohort, 1979-2012 (rounds 1-25), 2014. Produced and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH. Available at: https://www.bls.gov/nls/.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

Stamatis Cambanis, Steel Huang, and Gordon Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368–385, 1981.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

R. D. Cook and Liliana Forzani. Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, 104(485):197–208, 2009.

R. D. Cook and S. Weisberg. Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991.

R.D. Cook. *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Probability and statistics section. Wiley, 1998.

Raphaël Coudret, Stéphane Girard, and Jérôme Saracco. A new sliced inverse regression method for multivariate response. *Computational Statistics & Data Analysis*, 77:285–299, 2014.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.

Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 2018.

Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. Proxy non-discrimination in data-driven systems. *arXiv preprint arXiv:1707.08120*, 2017.

Eustasio del Barrio, Fabrice Gamboa, Paula Gordaliza, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. *arXiv preprint arXiv:1806.03195*, 2018.

Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *Annals of Statistics*, 12 (3):793–815, 1984.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems 31*, pages 2791–2801, 2018.

Harrison Edwards and Amos J. Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations*, 2016.

Equivant. Practitioner's guide to COMPAS core. https://bit.ly/2xfADkP, 2019.

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

L. Ferré and A. F. Yao. Functional sliced inverse regression analysis. *Statistics*, 37(6):475–488, 2003.

Jerome H. Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981.

Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Kernel dimension reduction in regression. *Annals of Statistics*, 37(4):1871–1905, 2009.

G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2013. ISBN 9781421407944.

Megh Gupta and Qasim Mohammad. Advances in AI and ML are reshaping healthcare. *TechCrunch*, 2017.

Peter Hall and Ker-Chau Li. On almost linearity of low dimensional projections from high dimensional data. *Annals of Statistics*, 21(2):867–889, 1993.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3323–3331, 2016.

T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1:297–318, 1986.

T. Hofmann, B. Schölkopf, and AJ. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3): 1171–1220, 2008.

Tailen Hsing and Raymond J. Carroll. An asymptotic theory for sliced inverse regression. *Annals of Statistics*, 20(2):1040–1061, 1992.

James E Johndrow and Kristian Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220, 2019.

F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*, pages 869–874, Dec 2010.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

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

Junpei Komiyama, Akiko Takeda, Junya Honda, and Hajime Shimao. Nonconvex optimization for regression with fairness constraints. In *International Conference on Machine Learning*, pages 2742–2751, 2018.

Law School Admissions Council. National Longitudinal Bar Passage Study. Available at: http://www2.law.ucla.edu/sander/Systemic/Data.htm.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. A Series of Modern Surveys in Mathematics Series. Springer, 1991.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Qian Lin, Xinran Li, Dongming Huang, and Jun S. Liu. On the optimality of sliced inverse regression in high dimensions. *arXiv e-prints*, art. arXiv:1701.06009, Jan 2017.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. In *International Conference on Learning Representations*, 2016.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3381–3390, 2018.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, pages 107–118, 23–24 Feb 2018.

Luca Oneto, Michele Donini, and Massimiliano Pontil. General fair empirical risk minimization. *arXiv preprint arXiv:1901.10080*, 2019.

N.S. Pillai, Q. Wu, F. Liang, S. Mukherjee, and R.L. Wolpert. Characterizing the function space for bayesian kernel models. *Journal of Machine Learning Research*, 8:1769–1797, 2007.

A. Pinkus. *Ridge Functions*. Cambridge Tracts in Mathematics. Cambridge University Press, 2015. ISBN 9781107124394.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008.

CE. Rasmussen and CKI. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, 2006.

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

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, pages 416–426, 2001.

James R. Schott. Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, 89(425):141–148, 1994.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pages 2164–2173, 2019.

G.W. Stewart. *Matrix Algorithms Volume 2: Eigensystems*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 2001. ISBN 9780898715033.

G.W. Stewart and Ji-Guang Sun. *Matrix Perturbation Theory*. Computer science and scientific computing. Academic Press, 1990. ISBN 9780126702309.

Zilong Tan and Sayan Mukherjee. Learning integral representations of Gaussian processes. *arXiv preprint arXiv:1802.07528*, 2018.

G. Wahba. Spline models for observational data. *Regional Conference Series in Applied Mathematics*, 59, 1990.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, Mar 1972.

Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1920–1953, 07–10 Jul 2017.

Qiang Wu, Sayan Mukherjee, and Feng Liang. Localized sliced inverse regression. In *Advances in Neural Information Processing Systems*, pages 1785–1792, 2009.

Qiang Wu, Feng Liang, and Sayan Mukherjee. Kernel sliced inverse regression: Regularization and consistency. *Abstract and Applied Analysis*, 2013, 01 2013.

Samuel Yeom, Anupam Datta, and Matt Fredrikson. Hunting for discriminatory proxies in linear regression models. In *Advances in Neural Information Processing Systems*, pages 4568–4578, 2018.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 962–970, 2017b.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, volume 28, pages 325–333, 2013.

I. Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv e-prints*, May 2015.