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 satisfying fairness goals such as statistical parity and equality of opportunity in learned models. Existing techniques for learning these representations are typically *model-agnostic*, as they pre-process 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 a given fairness criterion. Our method supports several fairness criteria, continuous and discrete data, and multiple protected attributes. We also characterize the fairness-accuracy trade-off with a parameter that relates to 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 the learned models are increasingly used in areas such as hiring (Dastin, 2018), healthcare (Gupta and Mohammad, 2017), and criminal justice (Equivant, 2019). In particular, the 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 novel approach for fair representation learning that takes into account the target hypothesis space 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 sufficient dimension reduction (SDR) framework (Fukumizu et al., 2009, Li, 1991, Wu et al., 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 (§ 2.1), 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

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

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 (Golub and Van Loan, 2013, Stewart and Sun, 1990). In contrast to recent regularization- and constraint-based trade-offs (Edwards and Storkey, 2016, Louizos et al., 2016, Madras et al., 2018, Song et al., 2019, Zemel et al., 2013), 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 (Pillai et al., 2007, Tan and Mukherjee, 2018, Wahba, 1990). Our experiments show that the FGP achieves both rigorous fairness properties and improved accuracy compared to prior methods.

Additional Related Work Much of the prior work on fair representation learning optimizes only for statistical parity (del Barrio et al., 2018, Feldman et al., 2015, Johndrow and Lum, 2019, Komiyama and Shimao, 2017, Komiyama et al., 2018, Louizos et al., 2016, Oneto et al., 2019a, Zemel et al., 2013) 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, but it should be noted that these bounds depend on the optimal adversary, which may not be available in non-convex settings. Our approach supports a broader set of fairness criteria (see § 2.2), and we characterize the generalization performance in terms of both fairness and accuracy.

Provably fair kernel learning has been recently studied by Donini et al. (2018) and Oneto et al. (2019b). 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 can remove general statistical associations between X and S.

Bayesian formulations of fairness have been studied by Foulds et al. (2018) and Dimitrakakis et al. (2019), who take into account the uncertainty in model parameters. Dimitrakakis et al. (2019) propose Bayesian versions of the balance (Kleinberg et al., 2017) and calibration criteria (Chouldechova, 2017) based on decision-theoretic risk formulations. Foulds et al. (2018) introduce a differential fairness criterion inspired by the definition of differential privacy in the setting of multiple protected attributes. Unlike these works, we do not propose new fairness criteria, instead focusing on several widely used criteria as described in § 2.2.

Much previous research on learning rules with explicit fairness constraints or objectives (Kamiran et al., 2010, Zafar et al., 2017b, Zliobaite, 2015) includes empirical studies on the fairness-accuracy trade-off, reporting that classifiers trained in this way outperform those trained on model-agnostic fair representations. Our proposed fair representations are not model-agnostic, and our performance is competitive if not better in some cases than that of those 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, as well as a regularization parameter analogous to the principal angle between subspaces used to set the trade-off in our work. In contrast, we provide insight into how the trade-off 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

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 write accordingly $\{(\boldsymbol{x}_i, s_i, y_i) \in \mathcal{X} \times \mathcal{S} \times \mathbb{R}\}_{i=1}^n$ for the training data of n examples.

The fair subspace is chosen as a vector space such that the projection of X onto the fair subspace does not contain information about S while retaining the residual information of X. We then use the projection as the fair representation. In the next subsection, we present an SDR-based method, i.e., Proposition 1, for finding a fair subspace. In § 2.2, we show that the fair subspace based representation satisfies several existing fairness criteria.

2.1 SDR-Induced Fair Representations

First consider the simple case where $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{S} = \mathbb{R}$, including both the categorical and continuous cases. The goal is to obtain a basis of the fair subspace of dimension $d \leq p$ represented by the columns of $\boldsymbol{C} \in \mathbb{R}^{p \times d}$. A salient challenge in finding the fair subspace arises as the link between S and X is unknown. To address this issue, the key insight we use is that $\boldsymbol{C}$ can be obtained from an *SDR subspace* of X with respect to S. Next, we briefly recall the definition of the SDR subspace as well as its assumption (Cook and Forzani, 2009, Li, 1991). Then, we provide the construction of $\boldsymbol{C}$ in Proposition 1.

An m-dimensional vector space is called an *SDR subspace* of X with respect to S if the projection of X onto the subspace captures the statistical dependency of S on X. The proposed SDR-induced fair representation relies on a kernel version of the following SDR assumption.

Assumption 1. *There exists a function $f_S : \mathbb{R}^m \times \mathbb{R}^l \to \mathbb{R}$ and a matrix $\boldsymbol{B} \in \mathbb{R}^{p \times m}$ such that*

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

where ϵ_S is a random variable independent of X.

The column span of $\boldsymbol{B}$ is known as the SDR subspace. It is worth pointing out that the condition (1) is *always* satisfied since there is a one-to-one correspondence between $\boldsymbol{B}^\top X$ and X whenever $\boldsymbol{B}$ is square non-singular. The condition (1) states that the projection $\boldsymbol{B}^\top X$ captures all information in X about S or more. The goal is thus to recover an SDR subspace with the lowest dimension. Under mild conditions on X, this recovery is guaranteed without requiring the knowledge of f_S (Hall and Li, 1993, Li, 1991).

The SDR-induced fair representation is given by the projection onto the fair subspace $X' \coloneqq \boldsymbol{C}^\top X$, where $\boldsymbol{C}$ is defined as in Proposition 1. Note that the proposition does not make assumptions about the underlying link function f_S. Its proof adapts techniques used in Brillinger, 1983 as well as the properties of elliptically contoured distributions, and is provided in supplementary material.

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 $\mathrm{Var}(X)\boldsymbol{B}$. If condition (1) holds and X follows an elliptically contoured distribution, then $\mathrm{Cov}\left(\boldsymbol{C}^\top X, S\right) = \boldsymbol{0}$.*

The class of elliptically contoured distributions contains the normal distribution. In the case where X' and S are jointly multivariate normal, the lack of correlation guaranteed by Proposition 1 implies $X' \perp\!\!\!\perp S$. 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.

Extensions to Multivariate S and RKHS Our approach incorporates two generalizations of the linear SDR condition (1). First, the condition (1) does not apply to the setting with multiple protected attributes. This is handled by replacing (1) with the joint conditions $S_i = f_{S_i}\left(\boldsymbol{B}^\top X, \epsilon_{S_i}\right)$ for each protected attribute S_i (Coudret et al., 2014). Second, S can depend on nonlinear structures of X, and hence the linear condition (1) may not yield a low-rank $\boldsymbol{B}$. In that case, the resulting fair subspace $\boldsymbol{C}^\top X$ may be too low-dimensional for accurately predicting Y. To address this issue, we use the RKHS counterpart of condition (1).

Denote by $\mathcal{H}_\kappa$ an RKHS of functions $f : \mathcal{X} \mapsto \mathbb{R}$ generated by kernel $\kappa(\cdot, \cdot)$. In the RKHS setting with training points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, X is replaced by the feature function $\kappa(\cdot, X)$, and $\boldsymbol{B}_i$ will instead 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 (Schölkopf et al., 2001). Thus, condition (1) in the RKHS setting reads

$$S = f_S\left(\kappa(X, \boldsymbol{X})\boldsymbol{W}_1, \cdots, \kappa(X, \boldsymbol{X})\boldsymbol{W}_m, \epsilon_S\right) \tag{2}$$

with $\kappa(X, \boldsymbol{X}) \coloneqq (\kappa(X, \boldsymbol{x}_1), \ldots, \kappa(X, \boldsymbol{x}_n))$. 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

$$\mathrm{span}\left\{\sum_{j=1}^n Q_{j1}\kappa(\cdot, \boldsymbol{x}_j), \ldots, \sum_{j=1}^n Q_{jr}\kappa(\cdot, \boldsymbol{x}_j)\right\}, \tag{3}$$

and the fair representation $X' = \kappa(X, \boldsymbol{X})\boldsymbol{Q}$ of X.

Kernel Misspecification In the RKHS setting, Proposition 1 states that the fair representation X' does not contain information about S. However, this does not necessarily imply that X' is predictive in terms of Y. For example, one can trivially satisfy (2) by letting $\boldsymbol{W}$ be the identity matrix. Then, the corresponding fair representation will be the constant 0, which is trivially independent of X. Thus, in order for the fair representation to be predictive, $\kappa(\cdot,\cdot)$ and $\boldsymbol{W}$ should be chosen such that (2) holds for a small m.

We note that a misspecified kernel affects the predictive power, but not fairness, of the fair representation as long as (2) holds. Kernel misspecification can be detected in practice because the resulting model based on X' would give a large prediction error. The problem of finding appropriate $\boldsymbol{W}$ and m will be discussed in § 3.1.

2.2 Fairness as Statistical Independence

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 § 2.1. Let $h(\cdot)$ denote the model and let $\widehat{Y} = h(X')$ be the model output. This paper focuses on the following fairness criteria:

- *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., $\widehat{Y} \perp\!\!\!\perp S$. This follows from $X' \perp\!\!\!\perp S$ since $\widehat{Y}$ is a function of X'.

- *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. For example, a component $\boldsymbol{c}$ of a linear model $h(X) = \boldsymbol{\beta}^\top X$ has the output $\widehat{Y}_{\boldsymbol{c}} \coloneqq \sum_{i=1}^{p} c_i \beta_i X_i$ for $c_i \in [0,1]$. The strictest version of proxy nondiscrimination requires $\widehat{Y}_{\boldsymbol{c}} \perp\!\!\!\perp S$ for all component $\boldsymbol{c}$. This follows from $X' \perp\!\!\!\perp S$ since $\widehat{Y}_{\boldsymbol{c}}$ is a function of X'.

- *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 $\widehat{Y} = Y$. 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.

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, a fair representation alone cannot preclude a constant model. If $\widehat{Y}$ is a constant, to satisfy accuracy parity and calibration we would need some independence condition between Y and S, which does not in general hold. Thus, it could be interesting future work to further identify additional conditions on the model needed to support the other fairness definitions.

3 Computing the Hypothesis Space

We now describe how to compute the fair subspace as well as the model subspace in which every function attains a desired fairness-accuracy trade-off. The construction is presented analytically in (8), and we provide generalization bounds in Theorem 1 and (11) for the deviation between an optimal fair or predictive model in the RKHS and the model obtained from the model subspace on unseen data. The pseudo-code and all proofs are provided as supplementary material.

Problem Setup Recall that a kernel-based learning problem is typically formulated as the following optimization under Tikhonov regularization (Cucker and Smale, 2002, Hofmann et al., 2008):

$$\min_{f \in \mathcal{H}_\kappa} L\left(f, \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n\right) + R\left(\|f\|_{\mathcal{H}_\kappa}\right), \qquad (4)$$

where L is a convex loss, R is a monotonically increasing regularization function, and $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is the set of training points. While $\mathcal{H}_\kappa$ is infinite dimensional, the well-known representer theorem (Schölkopf et al., 2001, Wahba, 1990) states that the solution $f_\star$ for (4) is in a data-dependent finite-dimensional subspace of $\mathcal{H}_\kappa$:

$$\mathcal{H}_{\kappa,n} \coloneqq \left\{ \sum_{i=1}^{n} a_i \kappa(\cdot, \boldsymbol{x}_i) \;\middle|\; \{a_i\}_{i=1}^n \subset \mathbb{R} \right\}. \qquad (5)$$

Our goal is to obtain a subset of functions in $\mathcal{H}_{\kappa,n}$ that meets the fairness criteria described in § 2.2. This subset will be the fair subspace (3) of $\mathcal{H}_{\kappa,n}$.

3.1 Learning the Fair and Predictive Subspaces

Both the predictive subspace $\mathcal{G}$ and the fair subspace $\mathcal{F}$ of $\mathcal{H}_\kappa$ can be estimated using a standard SDR estimator for the RKHS. Specifically, the predictive subspace $\mathcal{G}$ is given by the SDR subspace with respect to Y. Similarly, the fair subspace $\mathcal{F}$ given by (3) is obtained by first computing the SDR subspace with respect to S and then using Proposition 1 to obtain $\boldsymbol{Q}$.

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 $\boldsymbol{\phi} \coloneqq (\kappa(\cdot, \boldsymbol{x}_1), \cdots, \kappa(\cdot, \boldsymbol{x}_n))$. 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 following, we assume without loss of generality that $d + m \leq n$.

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

$$\boldsymbol{\Gamma}_n \boldsymbol{K} \boldsymbol{A}_i = \tau_i \left(\boldsymbol{\Delta} + n\eta \boldsymbol{I}_n\right) \boldsymbol{A}_i, \tag{6}$$

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 to be discussed shortly. For simplicity, (6) 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 $\boldsymbol{\Delta} = \operatorname{diag}(\boldsymbol{\Gamma}_{n_i}) \boldsymbol{K}$ where $\operatorname{diag}(\boldsymbol{\Gamma}_{n_i})$ is the block-diagonal matrix with diagonal blocks $\boldsymbol{\Gamma}_{n_i}$. The overall computational complexity for estimating $\boldsymbol{A}$ is $O\left(n^2 d\right)$.

Another relevant problem is to decide the dimensions of $\mathcal{G}$ and $\mathcal{F}$, i.e., the values 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. This gives an upper bound for d (resp. m). In both the categorical and continuous cases, one can use the methods proposed by Li (1991) and Schott (1994). For example, Li (1991) introduced an eigenvalue-based sequential test for the true SDR subspace dimension $d_\star$. Based on the test, Tan and Mukherjee (2018) provided a lower bound for the eigenvalue $\tau_{d_\star}$. We use the lower bound to select the SDR dimension d. In particular, we choose the largest dimension d such that τ_d satisfies the lower bound.

Estimating the Fair Subspace Given $\boldsymbol{W}$, we invoke Proposition 1 to compute the fair subspace $\mathcal{F}$ which is described by $\boldsymbol{Q}$. In the RKHS setting, the covariance matrix $\operatorname{Var}(X)$ in Proposition 1 is replaced by the covariance operator $\operatorname{Var}(\kappa(\cdot, X))$ on $\mathcal{H}_\kappa$. Let $\otimes$ denote the tensor product and let $\mu_X \coloneqq \mathbb{E}_X[\kappa(\cdot, X)]$, the empirical estimator of $\operatorname{Var}(\kappa(\cdot, X))$, be written $\mathbb{E}_X[(\kappa(\cdot, X) - \mu_X) \otimes (\kappa(\cdot, X) - \mu_X)] \approx \frac{1}{n} \boldsymbol{\phi}\boldsymbol{\Gamma}_n \otimes \boldsymbol{\phi}\boldsymbol{\Gamma}_n$. 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 § 2.2, 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 is not an issue as the fair subspace is still a subspace of $\mathcal{H}_{\kappa,n}$, and the pseudo-code in supplementary material shows how to handle this case.

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 (Golub and Van Loan, 2013, Stewart and Sun, 1990):

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

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 then 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). 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. Therefore, we instead consider a simple construction of $\mathcal{M}$ given by:

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

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 $\mathcal{G}$. With 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:

$$\begin{aligned} \mathcal{F} &\coloneqq \operatorname{span}\left\{\lambda_i^{-1/2}\boldsymbol{\phi}\boldsymbol{Q}\boldsymbol{M}_i\right\}_{i=1}^r \\ \mathcal{G} &\coloneqq \operatorname{span}\left\{\omega_i^{-1/2}\boldsymbol{\phi}\boldsymbol{A}\boldsymbol{T}_i\right\}_{i=1}^d. \end{aligned} \tag{7}$$

Using the orthonormal bases (7), Theorem 1 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 1. *Let* $\boldsymbol{\Lambda}^{-1/2}\boldsymbol{M}^\top\boldsymbol{Q}^\top\boldsymbol{K}\boldsymbol{A}\boldsymbol{T}\boldsymbol{\Omega}^{-1/2} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ *be the thin singular value decomposition with singular values* $\sigma_i \coloneqq \Sigma_{ii}$*, and let the hypothesis class of the model be*

$$\begin{aligned} \mathcal{M} = \operatorname{span}\Big\{\boldsymbol{\phi}\Big[&(\gamma_i - \rho_i\sigma_i)\,\boldsymbol{Q}\boldsymbol{M}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}_i + \\ &\rho_i\boldsymbol{A}\boldsymbol{T}\boldsymbol{\Omega}^{-1/2}\boldsymbol{V}_i\Big]\Big\}_{i=1}^d \end{aligned} \tag{8}$$

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:*

$$\|\mathcal{P}_\mathcal{F} - \mathcal{P}_\mathcal{M}\| = \sqrt{1 - \max\{\epsilon^2, \sigma_{min}^2\}} \tag{9}$$

$$\|\mathcal{P}_\mathcal{G} - \mathcal{P}_\mathcal{M}\| = \max\left\{0, \epsilon\sqrt{1-\sigma_{min}^2} - \sigma_{min}\sqrt{1-\epsilon^2}\right\}. \tag{10}$$

For the case where $\epsilon = 1$, the basis of (8) are linear combinations of the orthonormal basis of $\mathcal{F}$ in (7), and hence $\mathcal{M} \subset \mathcal{F}$. Additionally, it can be verified that $\|\mathcal{P}_\mathcal{F} - \mathcal{P}_\mathcal{M}\| = 0$ from (9), and (10) becomes $\|\mathcal{P}_\mathcal{G} - \mathcal{P}_\mathcal{M}\| = \sqrt{1-\sigma_{\min}^2}$ which is the sine of the largest principal angle between $\mathcal{F}$ and $\mathcal{G}$ as desired. Similarly, if we set $\epsilon = 0$ we get $\mathcal{M} = \mathcal{G}$.

The key utility of Theorem 1 involves bounding the difference between the model obtained using $\mathcal{M}$ and an optimal fair (or predictive) model. In particular, Equation (11) shows that $\mathcal{M}$ contains a function which approximates the optimal fair model $f_{\text{fair}} \in \mathcal{H}_\kappa$. Let $\delta_{\boldsymbol{x}} f \coloneqq f(\boldsymbol{x})$ be the evaluation functional. Then, for any $\boldsymbol{x} \in \mathcal{X}$:

$$\begin{aligned} &|(\mathcal{P}_\mathcal{M} f_{\text{fair}})(\boldsymbol{x}) - f_{\text{fair}}(\boldsymbol{x})| \\ &= \|\delta_{\boldsymbol{x}}(\mathcal{P}_\mathcal{M} f_{\text{fair}}) - \delta_{\boldsymbol{x}} f_{\text{fair}}\| \\ &\le \|\delta_{\boldsymbol{x}}\|\,\|\mathcal{P}_\mathcal{M} f_{\text{fair}} - f_{\text{fair}}\| \\ &= \|\delta_{\boldsymbol{x}}\|\,\|\mathcal{P}_\mathcal{M} f_{\text{fair}} - \mathcal{P}_\mathcal{F} f_{\text{fair}} + \mathcal{P}_\mathcal{F} f_{\text{fair}} - f_{\text{fair}}\| \\ &\le \|\delta_{\boldsymbol{x}}\|\left(\|\mathcal{P}_\mathcal{M} - \mathcal{P}_\mathcal{F}\|\,\|f_{\text{fair}}\| + \|\mathcal{P}_\mathcal{F} f_{\text{fair}} - f_{\text{fair}}\|\right). \end{aligned} \tag{11}$$

Here, $\|\delta_{\boldsymbol{x}}\|$ and $\|f_{\text{fair}}\|$ are bounded from the property of the RKHS, and the last norm in (11) converges in probability to zero at rate $O_P\left(n^{-1/4}\right)$ by the consistency of $\mathcal{F}$ and $\mathcal{G}$ (Wu et al., 2013). Together with (9), (11) sheds light on the impact of ϵ on the generalization of the fairness criteria; similar arguments can be made for an optimal predictive model.

4 Application to Fair GPs

In this section, we demonstrate the utility of our approach by constructing a fair Gaussian process (FGP) which can be used to develop a class of fair models under the Bayesian framework (Rasmussen and Williams, 2006). In particular, the FGP specifies a prior over functions in $\mathcal{M}$ that satisfy the fairness criteria discussed in § 2.2.

Recall that a GP $\{f(x) : x \in \mathcal{X}\}$ is specified by a mean function and a covariance function. The covariance function characterizes the class of functions, i.e., the curves of $f(\cdot)$, the GP can realize. The FGP is a GP equipped with a covariance function that ensures that any sample path $f(x)$ is fair. To obtain the covariance function, we use the integral representation of GPs (Itô, 1954):

$$f(\boldsymbol{x}) = \int_\mathcal{X} \kappa(\boldsymbol{x}, \boldsymbol{z})\,\nu(\boldsymbol{z})\,d\mu(\boldsymbol{z}), \tag{12}$$

where $\nu : \mathcal{X} \times \Omega \mapsto \mathbb{R}$ is another GP on a probability space $(\Omega, \mathcal{F}, P)$ and μ is the measure on $\mathcal{X}$. Clearly, $f(\cdot)$ given by (12) is a GP. Without loss of generality, we assume the GP has mean zero, then the covariance function $\operatorname{Cov}(f(\boldsymbol{x}), f(\boldsymbol{z}))$ is written

$$\int_\mathcal{X}\int_\mathcal{X} \kappa_\nu(\boldsymbol{s}, \boldsymbol{t})\,\kappa(\boldsymbol{x}, \boldsymbol{s})\,\kappa(\boldsymbol{z}, \boldsymbol{t})\,d\mu(\boldsymbol{s})\,d\mu(\boldsymbol{t}),$$

where $\kappa_\nu(\cdot,\cdot)$ is the covariance function of the GP $\nu(\cdot)$. A key property of (12) we use to construct the FGP is that functions in the form of (12) are contained in the RKHS generated by kernel $\kappa(\cdot,\cdot)$ (Pillai et al., 2007,

Tan and Mukherjee, 2018). By replacing $\kappa(\cdot,\cdot)$ in (12) with the reproducing kernel of $\mathcal{M}$, we obtain the desired FGP which inherits the fairness as well as accuracy guarantees of $\mathcal{M}$. It is worth noting that the representation $\mathcal{M}$ in (8) can be computed independent of data. This can be done using a likelihood for SDR subspaces (Cook and Forzani, 2009).

In practice, we use a sample version of the FGP for improved computational efficiency. Consider the sample average of (12) given by

$$f_n(\boldsymbol{x}) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \nu(\boldsymbol{x}_i)\,\kappa_{\mathcal{M}}(\boldsymbol{x},\boldsymbol{x}_i), \tag{13}$$

which converges in probability to (12) at rate $O_p\left(n^{-1/2}\right)$ by the central limit theorem for Hilbert spaces (Ledoux and Talagrand, 1991). The covariance function of the sample FGP has two parameters, namely the covariance function $\kappa_\nu(\cdot,\cdot)$ of $\nu(\cdot)$ and the reproducing kernel $\kappa_{\mathcal{M}}(\cdot,\cdot)$ of the RKHS $\mathcal{M}$.

Finally, we give a reparameterization to simplify the sample FGP. Let $\boldsymbol{\phi}\boldsymbol{E}_i$ represent the i-th basis function of (8), and denote by $\boldsymbol{\Pi}(\boldsymbol{z}) \coloneqq \left(\kappa(\boldsymbol{z},\boldsymbol{X}) - \mathbf{1}\mathbf{1}_n^\top \boldsymbol{K}/n\right)\boldsymbol{E}$, where $\boldsymbol{K}$ is the kernel matrix of κ as defined in (6). The sample FGP can be rewritten as

$$f_n(\cdot) \sim \mathcal{GP}\left(\mathbf{0}, \boldsymbol{\Pi}(\cdot)\,\boldsymbol{\Lambda}\boldsymbol{\Pi}(\cdot)^\top\right), \tag{14}$$

where the covariance function has only a single hyperparameter, a positive definite matrix $\boldsymbol{\Lambda}$. Now, it is straightforward to choose $\boldsymbol{\Lambda}$ as to maximizes the marginal likelihood while training the FGP (14).

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, including the processed datasets, code, as well as configurations, used in prior work (Donini et al., 2018, Komiyama et al., 2018) to compare the proposed FGP[1] 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, for the baseline methods we use the code published online by their respective authors, and for the GPs we use a linear mean and a radial basis covariance.

5.1 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 trade-off ϵ. 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 the accuracy of our approach tends to converge to that of standard GP, which is expected as setting the trade-off $\epsilon = 0$ in (8) 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.

5.2 Fair Regression with Multiple Protected Attributes

We consider a regression setting with two protected attributes. For this evaluation, we use three real datasets with continuous target values, namely the UCI Communities and Crime (Redmond and Baveja, 2002), the National Longitudinal Survey of Youth (NLSY) (Bureau of Labor Statistics, 2014) and the Law School

[1]The datasets and our Matlab implementation of the FGP are available at `https://github.com/ZilongTan/fgp`.

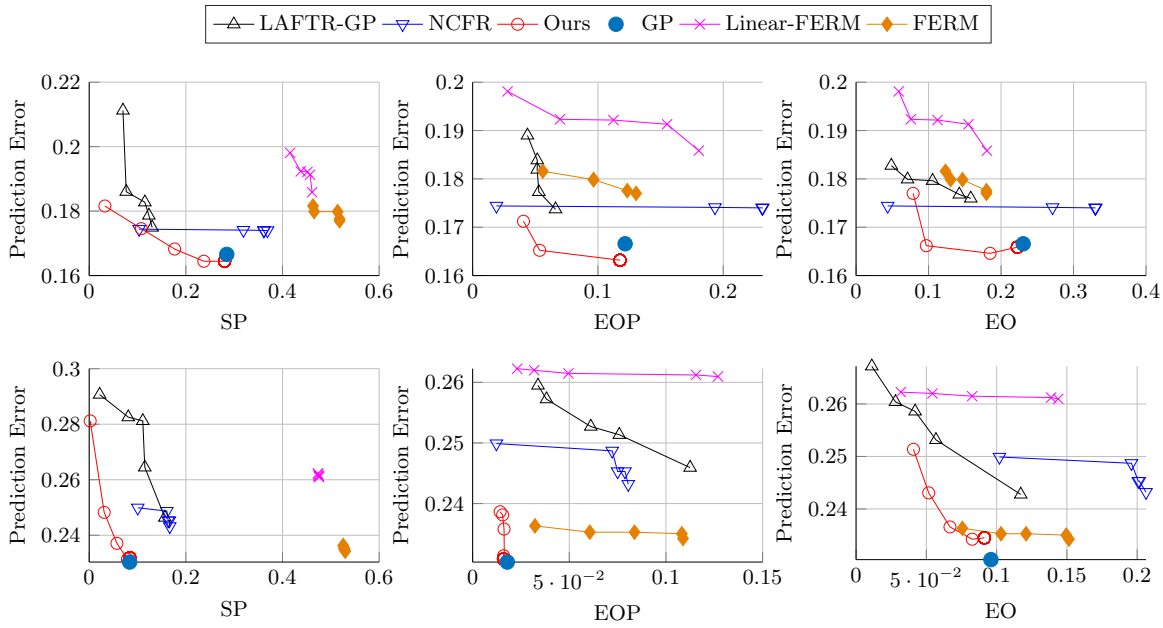

Figure 1: Comparing the accuracy-fairness trade-offs on the Adult (first row) and Compas (second row) datasets. The prediction error denotes the misclassification rate.

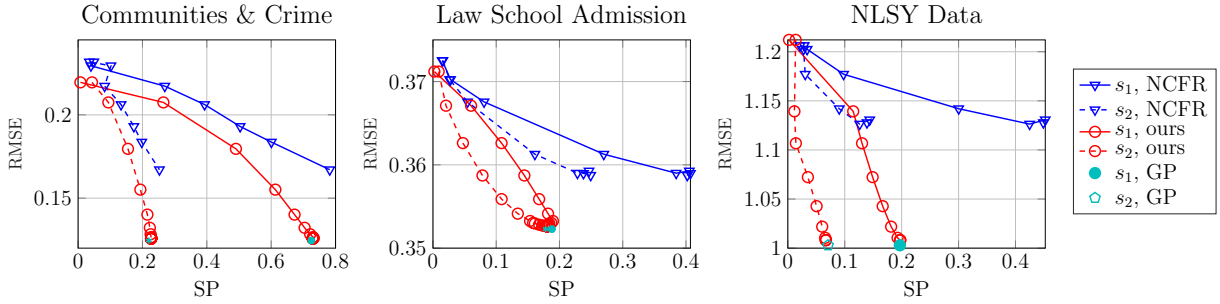

Figure 2: Regression results with two protected attributes s_1 and s_2.

Admissions Council (Law School Admissions Council) datasets. 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, so they are not suitable 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, our approach 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 fairness-accuracy trade-off. 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, 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://tinyurl.com/h326oye`.

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

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.

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.

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.

Christos Dimitrakakis, Yang Liu, David C. Parkes, and Goran Radanovic. Bayesian fairness. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 509–516, 2019.

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.

James Foulds, Rashidul Islam, Kamrun Keya, and Shimei Pan. Bayesian Modeling of Intersectional Fairness: The Variance of Bias. *arXiv e-prints*, Nov 2018.

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

Kiyosi Itô. Stationary random distributions. *Mem. College Sci. Univ. Kyoto Ser. A Math.*, 28(3):209–223, 1954.

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.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science*, pages 43:1–43:23, 2017.

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`.

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, Andreas Maurer, and Massimiliano Pontil. Learning fair and transferable representations. *arXiv preprint arXiv:1906.10673*, 2019a.

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

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.

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 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.