# Understanding Fairness and Prediction Error through Subspace Decomposition and Influence Analysis

**Enze Shi, Pankaj Bhagwat, Zhixian Yang, Linglong Kong, Bei Jiang***

Department of Mathematical and Statistical Science
University of Alberta
{eshi,pbhagwat,zhixian,lkong,bei1}@ualberta.ca

## Abstract

Machine learning models have achieved widespread success but often inherit and amplify historical biases, resulting in unfair outcomes. Traditional fairness methods typically impose constraints at the prediction level, without addressing underlying biases in data representations. In this work, we propose a principled framework that adjusts data representations to balance predictive utility and fairness. Using sufficient dimension reduction, we decompose the feature space into target-relevant, sensitive, and shared components, and control the fairness–utility trade-off by selectively removing sensitive information. We provide a theoretical analysis of how prediction error and fairness gaps evolve as shared subspaces are added, and employ influence functions to quantify their effects on the asymptotic behavior of parameter estimates. Experiments on both synthetic and real-world datasets validate our theoretical insights and show that the proposed method effectively improves fairness while preserving predictive performance.

## 1 Introduction

Machine learning (ML) models have achieved remarkable success across a wide range of high-stakes applications, including finance Hardt et al. [2016], Liu et al. [2018], healthcare Potash et al. [2015], Rudin and Ustun [2018], and criminal justice Van Dijck [2022], Billi et al. [2023]. Despite these advances, growing evidence highlights that ML systems often inherit and reinforce historical biases, leading to unfair outcomes Tolan et al. [2019], Mehrabi et al. [2021]. Biases in data collection Liang et al. [2020], Pagano et al. [2023] and disparities in group representation De-Arteaga et al. [2019], Dablain et al. [2024] can manifest in model predictions, ultimately amplifying social inequities Bolukbasi et al. [2016], Hassani [2021], Hu et al. [2024], Ding et al. [2024].

To address fairness concerns, researchers have introduced a range of formal definitions and algorithmic interventions. Early work focused on ensuring statistical criteria such as Demographic Parity Kamishima et al. [2012], Jiang et al. [2020], and more recent developments extend fairness guarantees to multiple sensitive attributes Tian et al. [2024], Chen et al. [2024] or local prediction regions Jin et al. [2024]. Many approaches operationalize fairness by adding constraints or regularization terms to learning objectives Hardt et al. [2016], Li et al. [2023]. However, this strategy typically treats fairness as an external correction layered on top of predictive modeling, without addressing the root causes of bias encoded within data representations themselves.

This work proposes a new perspective: instead of adjusting model predictions post hoc, we seek to understand and manage the trade-off between utility and fairness at the level of data representation. Specifically, we focus on how to construct representations that balance predictive accuracy for the

*Corresponding Author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

target variable and independence from sensitive attributes. This viewpoint naturally connects fairness with the framework of sufficient dimension reduction (SDR) Cook and Li [2002], Adragni and Cook [2009], where the goal is to project high-dimensional features onto lower-dimensional subspaces that preserve essential information.

We consider a structured setup in which both the prediction target $Y$ and the sensitive attribute $Z$ admit low-dimensional sufficient reductions with respect to the covariates $X$. By analyzing the subspaces associated with $Y$ and $Z$, we separate two types of directions: (1) directions informative about $Y$ but orthogonal to $Z$ and (2) directions jointly informative about both $Y$ and $Z$. This decomposition enables a principled subspace selection: start with $Z$-orthogonal directions for fairness, then add shared ones to boost accuracy as needed.

To quantify the fairness–utility trade-off, we develop a theoretical framework showing how prediction error and group-wise disparities evolve as shared directions are added. While incorporating $Z$-related directions improves accuracy, it reintroduces unfairness, which is captured through variance decomposition and gap metrics. We further apply influence functions to analyze the impact of partially fair representations on parameter estimation and prediction. The contributions of this paper are summarized as follows:

1. We propose a general framework for fairness-aware learning by directly manipulating data representations. It enables controlled removal of sensitive information without relying on a specific fairness definition.
2. We theoretically characterize how prediction risk and fairness gaps evolve as shared information is gradually introduced, and use influence functions to analyze the effect of partially fair representations on estimation and prediction.
3. We validate our approach through experiments on synthetic and real-world datasets, demonstrating that subspace elimination achieves trade-offs between fairness and predictive performance.

The structure of this paper is as follows. Section 3 introduces the motivation and problem setup, along with theoretical analysis of the utility–fairness trade-off under subspace elimination. Section 4 details the estimation procedure. Section 5 analyzes the asymptotic behavior of the estimators and predictors using influence functions. Section 6 presents simulation and real-world experiments that demonstrate the effectiveness of our approach.

## 2 Related Work

A major approach to fairness in machine learning focuses on enforcing fairness during model training, known as in-processing Wang et al. [2022a], Berk et al. [2023], Caton and Haas [2024]. These methods incorporate fairness constraints Zafar et al. [2019], Agarwal et al. [2019] into the optimization objective to satisfy predefined criteria such as Demographic Parity Dwork et al. [2012] or Equality of Opportunity Hardt et al. [2016], Shen et al. [2022], and have been applied across various paradigms including contrastive learning Wang et al. [2022b], Zhang et al. [2022], adversarial training Han et al. [2021], domain adaptation Wang et al. [2020], and balanced supervised learning Han et al. [2022]. However, fairness guarantees obtained this way are often task-specific and may not generalize across downstream applications. A complementary direction focuses on achieving fairness at the representation level by removing sensitive information from feature embeddings. Notable methods such as Iterative Null-space Projection (INLP) Ravfogel et al. [2020, 2023] and Relaxed Linear Adversarial Concept Erasure (RLACE) Ravfogel et al. [2022] aim to systematically eliminate sensitive information. More recently, sufficient dimension reduction (SDR) techniques have been adapted for debiasing Shi et al. [2024], identifying and removing subspaces associated with sensitive attributes while offering theoretical guarantees, with an extension of non-linear SDR via deep neural networks as proposed in Shi et al. [2025].

## 3 Methodology

### 3.1 Motivation

We study the problem of removing sensitive information from vector representations while preserving task-relevant content. Let $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ be random variables where $\boldsymbol{X} \in \mathbb{R}^p$ denotes the representation,

$\boldsymbol{Y} \in \mathbb{R}^K$ is the prediction target, and $\boldsymbol{Z} \in \mathbb{R}^d$ is the sensitive attribute. Our goal is to learn a transformation $h : \mathbb{R}^p \to \mathbb{R}^p$ such that the transformed representation $h(\boldsymbol{X})$ satisfies: (1) Minimal dependence on the sensitive attribute $\boldsymbol{Z}$ and (2) Sufficient information for predicting the target variable $\boldsymbol{Y}$. A direct approach involves constructing a linear transformation $h(\boldsymbol{X}) = P\boldsymbol{X}$, where $P \in \mathbb{R}^{p\times p}$ is a projection matrix. This induces a decomposition of the input space as $\boldsymbol{X} = P\boldsymbol{X} + (I - P)\boldsymbol{X}$, with $P$ projecting onto a subspace $\mathcal{S}_1 \subseteq \mathbb{R}^p$ and $I - P$ projecting onto its orthogonal complement $\mathcal{S}_2 = \mathcal{S}_1^\perp$. The representation space is thus decomposed as $\mathbb{R}^p = \mathcal{S}_1 \oplus \mathcal{S}_2$. For effective debiasing, $\mathcal{S}_1$ should minimize information about $\boldsymbol{Z}$, while $\mathcal{S}_2$ captures the $\boldsymbol{Z}$-relevant components to be removed.

This formulation is introduced by Shi et al. [2024], who adopt the sufficient dimension reduction (SDR) framework to identify and eliminate sensitive information in representations. Given $\boldsymbol{X}$ and $\boldsymbol{Z}$, their goal is to find a matrix $B_z \in \mathbb{R}^{p\times r}$ with orthonormal columns such that

$$\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{X} \mid B_z^\top \boldsymbol{X}, \tag{1}$$

ensuring that $B_z^\top \boldsymbol{X}$ contains all the information in $\boldsymbol{X}$ relevant to predicting $\boldsymbol{Z}$. The column space of $B_z$, known as the SDR subspace of $\boldsymbol{X}$ with respect to $\boldsymbol{Z}$, see Cook and Li [2002], Adragni and Cook [2009], thus serves as a natural candidate for $\mathcal{S}_2$ with $\dim(\mathcal{S}_2) = r$. Its orthogonal complement, spanned by $P = I - B_z B_z^\top$, defines $\mathcal{S}_1$ with $\dim(\mathcal{S}_1) = p - r$. Then $h(\boldsymbol{X}) = P\boldsymbol{X}$ removes information associated with $\boldsymbol{Z}$, yielding a fair projection.

While the sufficient projection method in Shi et al. [2024] performs well across downstream tasks, it primarily focuses on removing $\boldsymbol{Z}$-related information without explicitly preserving information relevant to $\boldsymbol{Y}$. This can lead to substantial utility loss when $\boldsymbol{Y}$ and $\boldsymbol{Z}$ share overlapping subspaces. To address this, we propose a finer decomposition of the representation space $\mathbb{R}^p$, separating directions informative about $\boldsymbol{Y}$ but orthogonal to $\boldsymbol{Z}$, directions shared by both, and directions unrelated to either. This more granular perspective allows for better control of the fairness–utility trade-off by retaining task-relevant features while minimizing bias from sensitive attributes.

### 3.2 Problem Setup

We adopt a SDR framework for both the target variable $\boldsymbol{Y}$ and the sensitive attribute $\boldsymbol{Z}$, modeled as

$$\boldsymbol{Y} = f(\beta_1^\top \boldsymbol{X}, \beta_2^\top \boldsymbol{X}, \ldots, \beta_q^\top \boldsymbol{X}, \varepsilon_Y), \tag{2}$$

$$\boldsymbol{Z} = g(\psi_1^\top \boldsymbol{X}, \psi_2^\top \boldsymbol{X}, \ldots, \psi_r^\top \boldsymbol{X}, \varepsilon_Z), \tag{3}$$

for some measurable functions $f$ and $g$, where $\{\beta_k\}_{k=1}^q \subset \mathbb{R}^p$ and $\{\psi_j\}_{j=1}^r \subset \mathbb{R}^p$ are orthonormal direction vectors, and $\varepsilon_Y$, $\varepsilon_Z$ are noise terms independent of $\boldsymbol{X}$. The SDR assumption implies that all information relevant for predicting $\boldsymbol{Y}$ and $\boldsymbol{Z}$ is captured by the low-dimensional projections $\{\beta_k^\top \boldsymbol{X}\}_{k=1}^q$ and $\{\psi_j^\top \boldsymbol{X}\}_{j=1}^r$, respectively. Equivalently, the models in (2) and (3) imply the conditional independence statements:

$$\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X} \mid M_Y \boldsymbol{X}, \quad \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{X} \mid M_Z \boldsymbol{X}, \tag{4}$$

where $M_Y,\ M_Z \in \mathbb{R}^{p\times p}$ are matrices with rank $q$ and $r$. The subspaces spanned by these matrices are referred to as the central subspaces, defined by

$$\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}} = \mathrm{Span}(M_Y) = \mathrm{Span}\{\beta_1, \ldots, \beta_q\}, \quad \mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}} = \mathrm{Span}(M_Z) = \mathrm{Span}\{\psi_1, \ldots, \psi_r\}.$$

Assume that the central subspaces $\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}}$ and $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$ intersect in a subspace

$$\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}} \cap \mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}} = \mathrm{Span}\{\phi_1, \ldots, \phi_s\},$$

where $s \le \min\{q, r\}$. When $s = 0$, the subspaces intersect only at the origin. In the special case where $\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}} \subset \mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}^\perp$, the two subspaces are completely separable. In this setting, removing all information associated with $\boldsymbol{Z}$ does not affect the information relevant for predicting $\boldsymbol{Y}$, and thus fairness can be achieved without sacrificing utility.

In practice, the subspaces $\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}}$ and $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$ often overlap, with a nontrivial intersection ($s > 0$). In such cases, removing all $\boldsymbol{Z}$-related components may also eliminate valuable information for predicting $\boldsymbol{Y}$. To address this, we decompose $\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}}$ into two orthogonal parts: one shared with $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$ and one independent of it. This decomposition enables a principled approach to balancing fairness and utility by selectively retaining target-relevant features uncorrelated with the sensitive attribute.

Without loss of generality, we assume that the shared basis vectors satisfy $\phi_i = \beta_{q-s+i} = \psi_{r-s+i}$ for $i = 1, \ldots, s$. Define the projection matrix onto $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$ as $P_z = \sum_{j=1}^{r} \psi_j \psi_j^\top$ and $Q_z = I_p - P_z$, where $Q_z$ projects onto the orthogonal complement of the sensitive subspace. Let $B = (\beta_1, \ldots, \beta_q) \in \mathbb{R}^{p \times q}$ and $\Phi = (\phi_1, \ldots, \phi_s) \in \mathbb{R}^{p \times s}$. Since $\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}} \cap \mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}} = \mathrm{Span}(\Phi)$, the uncorrelated component $\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}} \cap \mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}^\perp$ can be identified by finding a matrix $\widetilde{B}$ such that

$$\boldsymbol{Y} \perp\!\!\!\perp Q_z \boldsymbol{X} \mid \widetilde{B}^\top \boldsymbol{X}, \tag{5}$$

**Theorem 3.1.** *Let $\widetilde{B}$ be an orthonormal matrix satisfying condition* (5)*, then* $\mathrm{Span}(\widetilde{B}) \subseteq \mathrm{Span}(Q_z B) = \mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}} \cap \mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}^\perp$.

Theorem 3.1 provides a direct link between the central subspace of $\boldsymbol{Y}$ and its component orthogonal to the sensitive subspace $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$. Building on this decomposition, we define a sequence of partially fair projection matrices $\{P^{(m)}\}_{m=0}^{s}$ as

$$P^{(m)} = \widetilde{B}\widetilde{B}^\top + \Phi_m \Phi_m^\top, \quad \text{where} \quad \Phi_m = (\phi_1, \ldots, \phi_m).$$

Each $P^{(m)}$ projects $\boldsymbol{X}$ onto a subspace that retains $\widetilde{B}$-based directions uncorrelated with $\boldsymbol{Z}$, along with $m$ of the $s$ shared directions between $\boldsymbol{Y}$ and $\boldsymbol{Z}$. Let $\boldsymbol{\Xi}^{(m)} = P^{(m)}\boldsymbol{X}$ denote the resulting representation. When $m = 0$, the representation is entirely uncorrelated with $\boldsymbol{Z}$; when $m = s$, the representation spans the full central subspace of $\boldsymbol{Y}$, preserving complete utility but potentially reintroducing bias.

To study the predictive behavior under this fairness–utility trade-off, we define the Bayes optimal predictor using the partially fair representation:

$$\tilde{f}^{(m)}(\boldsymbol{\Xi}^{(m)}) = \mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{\Xi}^{(m)}].$$

This formulation allows gradual control over the balance between fairness and accuracy by varying $m$, i.e., the number of shared components included. In the following, we analyze the theoretical properties of $\tilde{f}^{(m)}$, characterizing how prediction risk and fairness evolve with subspace selection.

### 3.3 Utility and Fairness Trade-off

Let $\boldsymbol{\Xi}^{(m)} = P^{(m)}\boldsymbol{X}$ denote the reduced representation. The following result characterizes the prediction error of the Bayes optimal predictor based on the reduced representation.

**Theorem 3.2.** *Let $\tilde{f}^{(m)}(\boldsymbol{\Xi}^{(m)}) = \mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{\Xi}^{(m)}]$ denote the Bayes predictor using the partially fair representation $\boldsymbol{\Xi}^{(m)}$, and let $f^*(\boldsymbol{X}) = \mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ denote the Bayes optimal predictor using the original representation. Then, the expected squared prediction error satisfies*

$$\mathbb{E}[(\boldsymbol{Y} - \tilde{f}^{(m)}(\boldsymbol{\Xi}^{(m)}))^2] = \underbrace{\mathbb{E}[\mathrm{Var}(f^*(\boldsymbol{X}) \mid \boldsymbol{\Xi}^{(m)})]}_{\textit{Approximation error}} + \underbrace{\mathbb{E}[\varepsilon_Y^2]}_{\textit{Irreducible noise}} := \Delta(m) + \sigma_Y^2.$$

*Moreover, the approximation error $\Delta(m)$ is non-increasing in m, i.e.,*

$$\Delta(m+1) \leq \Delta(m) \quad \textit{for all } m \in \{0, \ldots, s-1\}.$$

This decomposition follows from the orthogonality principle in $L^2$ space, which ensures that the conditional expectation minimizes mean squared error. The term $\mathbb{E}[\mathrm{Var}(f^*(\boldsymbol{X}) \mid \boldsymbol{\Xi}^{(m)})]$ quantifies the loss in predictive information due to reducing the representation to $\boldsymbol{\Xi}^{(m)}$. As more shared directions are included, the reduced representation becomes increasingly informative, and the error $\Delta(m)$ decreases. When $m = s$, the full central subspace for $\boldsymbol{Y}$ is recovered, yielding $\Delta(s) = 0$.

Quantifying unfairness theoretically is inherently challenging, as it often stems from disparities in prediction outcomes across sensitive subpopulations. Instead of relying on specific fairness metrics like TPR or demographic parity (DP) gaps, we adopt a distributional perspective by measuring the statistical dependence between the predictor and the sensitive attribute using *distance covariance* (dCov) Székely et al. [2007], which quantifies the discrepancy between joint and marginal characteristic functions. The following result illustrates how the reduced representation $\boldsymbol{\Xi}^{(m)}$ mitigates unfairness by weakening the dependency between the predictor and the sensitive attribute. While the result is stated for binary $\boldsymbol{Z}$, it naturally extends to multivariate cases with $\boldsymbol{Z} \in \mathbb{R}^d$.

**Theorem 3.3.** *Let $Z \in \{0,1\}$ be a binary sensitive attribute with $p = \mathbb{P}(Z = 1)$, and let $\tilde{f}^{(m)} = \mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{\Xi}^{(m)}]$ be the Bayes predictor using the reduced SDR representation $\boldsymbol{\Xi}^{(m)}$. Then the squared population distance covariance between $\tilde{f}^{(m)}$ and $Z$ satisfies*

$$\mathrm{dCov}^2(\tilde{f}^{(m)}, Z) = 2p(1-p)\left(\mathbb{E}[|\tilde{f}_1^{(m)} - \tilde{f}_0^{(m)}|] - \mathbb{E}[|\tilde{f}^{(m)} - \tilde{f}^{(m)\prime}|]\right),$$

*where $\tilde{f}_z^{(m)} \sim \tilde{f}^{(m)} \mid Z = z$ for $z \in \{0,1\}$, and $\tilde{f}^{(m)\prime}$ is an independent copy of $\tilde{f}^{(m)}$. Specifically, when $m = 0$, we have $\mathrm{dCov}^2(\tilde{f}^{(m)}, Z) = 0$ and thereby $\tilde{f}^{(m)} \perp\!\!\!\perp Z$.*

This expression reveals that distance covariance is determined by the discrepancy between within-group and between-group variations in predictions. When the reduced representation $\boldsymbol{\Xi}^{(m)}$ sufficiently removes dependence on $Z$, the term $\mathbb{E}[|\tilde{f}_1^{(m)} - \tilde{f}_0^{(m)}|]$ approaches the population-level variation $\mathbb{E}[|\tilde{f}^{(m)} - \tilde{f}^{(m)\prime}|]$, leading to a smaller dCov and improved fairness.

# 4 Subspace Estimation and Algorithm Implementation

## 4.1 Estimation of Projections

We describe the procedure to estimate the sufficient directions for predicting $\boldsymbol{Y}$. The first step is to estimate the intersection subspace $\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}} \cap \mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$, spanned by $\{\phi_1, \ldots, \phi_s\}$. This is equivalent to estimating a matrix $\Phi \in \mathbb{R}^{p \times s}$ such that

$$\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \Phi^\top \boldsymbol{X}, \tag{6}$$

which corresponds to dimension reduction with respect to the interaction between response variables as proposed in Luo [2022].

Let $\Sigma$ denote the covariance matrix of $\boldsymbol{X}$. Note that $M_Y$ and $M_Z$ are the symmetric candidate matrices from SDR methods that satisfy (4), with estimates $\widehat{M}_Y$ and $\widehat{M}_Z$. Define the cross-matrix $M_{Y,Z} = M_Y \Sigma M_Z$, and let $s = \mathrm{rank}(M_{Y,Z})$. Luo [2022] show that $\Phi$ satisfies (6) if and only if

$$M_Y \Sigma\, P_{\Sigma,\Phi} M_Z = M_{Y,Z}, \tag{7}$$

where $P_{\Sigma,B} = B(B^\top \Sigma B)^{-1} B^\top \Sigma$. The following theorem characterize the estimation of intersection subspace as a generalized eigenvalue decomposition problem.

**Theorem 4.1.** *Suppose $\mathrm{Span}\{\Phi\} \subseteq \mathbb{R}^p$ is an $s$-dimensional subspace. Then $\mathrm{Span}\{\Phi\}$ is given by the span of the leading $s$ eigenvectors of the following generalized eigenvalue decomposition problem: $M_Y \Sigma M_Z \nu = \lambda \Sigma \nu$. The candidate symmetric matrix for* (6) *is $M_{Y,Z} = M_Y\, \Sigma\, M_Z$.*

Then the estimation procedure of $\widehat{\Phi}$ is as follows. First, construct $\widehat{M}_Y$ and $\widehat{M}_Z$ using any exhaustive inverse regression method (e.g., SIR Li [1991], SAVE Cook and Weisberg [1991], or directional regression Li and Wang [2007]), and compute the sample covariance matrix $\widehat{\Sigma}$. Next, form $\widehat{M}_{Y,Z} = \widehat{M}_Y \widehat{\Sigma} \widehat{M}_Z$ and estimate the dimension $\hat{s} = \mathrm{rank}(\widehat{M}_{Y,Z})$ using the ladle estimator [Luo and Li, 2016].

Once $\widehat{\Phi}$ is estimated, the remaining directions relevant for predicting $\boldsymbol{Y}$ can be obtained via projection. Rather than estimating $\beta_1, \ldots, \beta_{q-s}$ directly, we estimate their projections onto the orthogonal complement of $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$. Define the estimated projection matrix $\widehat{Q}_z = I_p - \sum_{j=1}^{\hat{r}} \hat{\psi}_j \hat{\psi}_j^\top$, then we apply any SDR method to $(\boldsymbol{Y}, \widehat{Q}_z \boldsymbol{X})$ to obtain $\widehat{B}_{Y,Q_z}$ such that $\boldsymbol{Y} \perp\!\!\!\perp \widehat{P}_z \boldsymbol{X} \mid \widehat{B}_{Y,Q_z}^\top \boldsymbol{X}$. This matrix approximates the projected directions $\beta_1^\top P_z, \ldots, \beta_{q-s}^\top P_z$. Together, the estimated shared directions $\widehat{\Phi}$ and the unshared components $\widehat{B}_{Y,P_z}$ form a sufficient projection $\widehat{P}^{(m)}$. The consistency and $n^{-1/2}$ convergence rates of the estimated directions and projections are well-established in the SDR literature; we omit these details for brevity. The complete procedure is summarized in Algorithm 1.

*Remark* 4.2. The computational cost of obtaining the fair projection matrix primarily arises from estimating the candidate matrices $M_Y$, $M_Z$ and $M_{Y,Z}$, each typically constructed as a weighted covariance matrix. These computations scale linearly with the sample size $n$. In addition, the procedure includes an eigen-decomposition step for $p \times p$ matrices, and the rank estimation step scales linearly with dimension $p$.

**Algorithm 1** Estimation of Sufficient Projections

1: **Input:** Data $(X_i, Y_i, Z_i)_{i=1}^n$; SDR method for candidate matrix construction
2: **Output:** Estimated sufficient projections $\widehat{P}^{(m)}$, $m = 0, \dots, \hat{s}$.
3: Compute $\widehat{M}_Y$, $\widehat{M}_Z$ using SDR method; compute $\widehat{\Sigma} = \mathrm{cov}(X)$.
4: Form $\widehat{M}_{Y,Z} = \widehat{M}_Y \widehat{\Sigma} \widehat{M}_Z$; apply ladle estimator to estimate rank $\hat{s}$.
5: Obtain estimator $\widehat{\Phi} \in \mathbb{R}^{p \times \hat{s}}$ and get projection $\widehat{Q}_z$.
6: Apply SDR to $(Y, \widehat{Q}_z X)$ and obtain $\widehat{B}_{Y,Q_z}$ with rank $\hat{d}_{Y,Q_z}$.
7: Obtain $\widehat{P}^{(m)} = \widehat{B}_{Y,P_z} \widehat{B}_{Y,P_z}^\top + \widehat{\Phi}_m \widehat{\Phi}_m^\top$ for $m = 0, \dots, \hat{s}$

### 4.2 Algorithm Implementation

The sufficient variables $\boldsymbol{\Xi}^{(m)} = \widehat{P}^{(m)} \boldsymbol{X}$ can then be used to fit regression or classification models for downstream tasks. Note that the columns of $\widehat{\Phi}$ are ordered by the eigenvalues of $\widehat{M}_{Y,Z}$, reflect their predictive power for both $\boldsymbol{Y}$ and $\boldsymbol{Z}$. By gradually adding these columns in the projection, we can incrementally build models with varying levels of sensitive information included.

This setup allows users to manually control the amount of sensitive information retained when predicting $\boldsymbol{Y}$. The post-SDR training procedure is summarized in Algorithm 2, using a classification task with accuracy (Acc) as the utility metric and demographic parity (DP) as the fairness metric. The optimal feature set and fair model are chosen to achieve at least 95% of the validation accuracy from the full dataset while minimizing unfairness.

*Remark* 4.3. Algorithm 2 presents one example of post-SDR training. In practice, DP and Acc can be replaced by any fairness and utility metrics, and the 95% threshold generalized to any $\tau \in (0, 1)$. The shared dimension $s$ acts as a tuning parameter analogous to a regularization coefficient, offering two advantages: (1) $s$ is selected from a finite set, avoiding continuous grid search (one can step every 2–3 dimensions or use bisection); and (2) each added component remains interpretable, enabling clear insight into the fairness–performance trade-off.

**Algorithm 2** Sequential Fair Projection: Post-SDR Fair Model Training

1: **Input:** Sufficient variables $\boldsymbol{\Xi}^{(m)} = \widehat{P}^{(m)} \boldsymbol{X}$, training and validation datasets
2: **Output:** Fair model $\mathcal{M}_{\text{fair}}$ and selected feature set $\boldsymbol{\Xi}^{(m^*)}$
3: **for** $i = 0$ to $s$ **do**
4: Train model $\mathcal{M}_{\text{fair}}^{(i)}$ based on $\boldsymbol{\Xi}^{(i)}$ and record $\mathrm{Acc}^{(i)}$ and $\mathrm{DP}^{(i)}$ on validation set.
5: **end for**
6: Let $s^* = \arg\min_i \{\mathrm{DP}^{(i)} \mid \mathrm{Acc}^{(i)} > 95\% \, \mathrm{Acc}^{(orig)}\}$, where $\mathrm{Acc}^{(orig)}$ is the accuracy trained by original datasets $\boldsymbol{X}$.
7: **return** Selected model $\mathcal{M}_{\text{fair}} = \mathcal{M}_{\text{fair}}^{m^*}$ and corresponding feature set $\boldsymbol{\Xi}^{(m^*)}$.

## 5 Theoretical Analysis

In this section, we analyze the impact of using sufficient variables in the post-SDR training procedure, with a focus on how they influence parameter estimation and expected errors. We employ influence functions to trace a model's prediction through the learning algorithm and back to its input features.

### 5.1 Influence Functions

Let $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \sim F_0$ be the joint law and we denote the empirical distribution on $n$ samples of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ as $F_n$. Define the SDR functional $M_Y(F)$ and $M_Z(F)$ for estimating central subspaces $\mathcal{S}_{Y|X}$ and $\mathcal{S}_{Z|X}$ respectively, and let $\Sigma(F) = \mathrm{Var}(X)$ be the functional for the covariance matrix. Therefore, the intersection subspace estimation for $\Phi$ corresponds to reduction functional $M_{Y,Z}(F) = M_Y(F)\Sigma(F)M_Z(F)$.

In the following, an asterisk on a symbol always indicates the influence function of a statistical functional represented by that symbol. For a sample point $S = (x, y, z)$ drawn from $F_0$, let $\delta_S$ be the

Dirac measure at $S$. The influence function of the functional $R$ is defined as

$$R^*(S) = \frac{\partial}{\partial \varepsilon} R[(1-\varepsilon)F_0 + \varepsilon\delta_S] \,|_{\varepsilon=0} \;.$$

For notation simplicity, we abbreviate $R^*(S)$ by $R^*$. In the following, an asterisk on a symbol always indicates the influence function of a statistical functional represented by that symbol. For example, we denote the influence functions of functionals $M_Y(F)$, $M_Z(F)$, $M_{Y,Z}(F)$ and $\Sigma(F)$ as $M_Y^*$, $M_Z^*$, $M_{Y,Z}^*$ and $\Sigma^*$, respectively. For notation simplicity, we omit the Using the product rule for Gateaux derivatives, we obtain the influence function of $M_{Y,Z}^*$.

**Lemma 5.1.** *Suppose $M_Y(F)$, $M_Z(F)$ and $\Sigma(F)$ are Hadamard differentiable. Then, the influence function of the reduction functional $M_{Y,Z}(F)$ is*

$$M_{Y,Z}^* = M_Y^*\,\Sigma(F)\,M_Z(F) \;+\; M_Y(F)\,\Sigma^*\,M_Z(F) \;+\; M_Y(F)\,\Sigma(F)\,M_Z^*.$$

### 5.2 Asymptotic Normality of Estimators

Let $f(X;\theta)$ be the differentiable predictive function parametrized by $\theta \in \Theta$ and let $L(x, y; \theta)$ be the differentiable loss function with respect to $\theta$, where we fold in any regularization terms into $L$. For notation simplicity, we denote $L(x, y; \theta)$ by $L(S; \theta)$. Then the population and empirical risk minimizer of the parameter is given by

$$\widetilde{\theta} = \theta(F_0) \triangleq \arg\min_{\theta\in\Theta} \mathbb{E}[L(S,\theta)], \quad \widehat{\theta}_n = \theta(F_n) \triangleq \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} L(S_i,\theta)$$

Similarly, we define the parameters estimated by sufficient variables $S^{(m)} = (P^{(m)}x, y, z)$ as

$$\widetilde{\theta}^{(m)} = \theta^{(m)}(F_0, P^{(m)}(F_0)) \triangleq \arg\min_{\theta\in\Theta} \mathbb{E}[L(S^{(m)},\theta)]$$

$$\widehat{\theta}_n^{(m)} = \theta^{(m)}(F_n, P^{(m)}(F_n)) \triangleq \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} L(\widehat{S}_i^{(m)},\theta),$$

where $P^{(m)}(F_n) = \widehat{P}^{(m)}$ and $\widehat{S}^{(m)} = (\widehat{P}^{(m)}x, y, z)$. We refer to $\widehat{\theta}_n$ as the original estimator and $\widehat{\theta}_n^{(m)}$ as the fair estimator. The entire post-SDR learning procedure can be viewed as a composition of three mappings:

$$F_n \xrightarrow{\text{reduction}} P^{(m)}(F_n) \xrightarrow{\text{estimation}} \theta^{(m)}\big(F_n,\, P^{(m)}(F_n)\big) \xrightarrow{\text{prediction}} f\big(\,\cdot\,;\, \theta^{(m)}\big(F_n, P^{(m)}(F_n)\big)\big).$$

The composition mapping allows us to derive the asymptotic normality of both the estimators and predictors after applying fair projections, as stated in the following theorems.

**Theorem 5.2.** *Suppose all the statistical functionals presented above are Hadamard differentiable, and their influence functions has zero expectation and finite variance, then we have the following asymptotic normality for the estimator $\widehat{\theta}_n^{(m)}$*

$$\sqrt{n}(\widehat{\theta}_n^{(m)} - \widetilde{\theta}^{(m)}) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \mathrm{Var}\big[-H_{\widetilde{\theta}^{(m)}}^{-1} G^{(m)} + D^{(m)}\,\mathrm{vec}(P^{(m)*})\big]\right), \tag{8}$$

*where $H_{\widetilde{\theta}^{(m)}}^{-1} = \mathbb{E}[\nabla_\theta^2 L(S^{(m)}, \widetilde{\theta}^{(m)})]$ is the expected Hessian of loss function, $G^{(m)} = \nabla_\theta L(S^{(m)}, \widetilde{\theta}^{(m)})$ is the gradient, $D^{(m)} = (\partial\,\theta(F_0, P^{(m)}(F_0))/\partial\,\mathrm{vec}(P^{(m)}))^\top$, $\mathrm{vec}(\cdot)$ is vectorization of the matrix, and $P^{(m)*}$ denotes the influence function of $P^{(m)}(F)$, whose explicit form depends on the choice of SDR method and is provided in the Appendix.*

**Corollary 5.3.** *Under the same assumptions stated in Theorem 5.2, we have*

$$\mathbb{E}(\|\widehat{\theta}_n^{(m)} - \widehat{\theta}_n\|^2) \le \|\widetilde{\theta}^{(m)} - \widetilde{\theta}\|^2 + \frac{1}{n}\,\mathrm{Tr}\left(\mathrm{Var}\big[H_{\widetilde{\theta}}^{-1}G - H_{\widetilde{\theta}^{(m)}}^{-1} G^{(m)} + D^{(m)}\,\mathrm{vec}(P^{(m)*})\big]\right), \tag{9}$$

*where $H_{\widetilde{\theta}}^{-1} = \mathbb{E}[\nabla_\theta^2 L(S, \widetilde{\theta})]$ and $G = \nabla_\theta L(S, \widetilde{\theta})$.*

**Theorem 5.4.** *Let $f(x;\theta)$ be the predictor evaluated at covariate $x$. Under the same assumptions stated in Theorem 5.2, we have the following asymptotic normality for the predictor*

$$\sqrt{n}\left(f(x;\widehat{\theta}_n^{(m)}) - f(x;\widetilde{\theta}^{(m)})\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, g^\top\,\mathrm{Var}\big[-H_{\widetilde{\theta}^{(m)}}^{-1} G^{(m)} + D^{(m)}\,\mathrm{vec}(P^{(m)*})\big]g\right), \tag{10}$$

*where $g = \nabla_\theta f(x;\widetilde{\theta}^{(m)}))$.*

Theorems 5.2 and 5.4 provide feasible inference procedures for the fair estimators corresponding to each $m$, which are useful for subsequent statistical inference or hypothesis testing. Corollary 5.3 follows directly from Theorem 5.2 by comparing the asymptotic distributions of the fair and original estimators. It shows that the expected distance between the fair and original estimators is upper bounded by the distance between their respective true parameters and the sampling variability.

Moreover, the projection matrix $P^{(m)}$ plays a crucial role in shaping the asymptotic variance of the fair estimator. By restricting the estimation to a subspace spanned by $P^{(m)}$, it reduces the dimensionality of the parameter space, leading to smaller asymptotic variance. However, smaller $m$ also implies that important predictive directions are truncated and inflate the $\|\widetilde{\theta}^{(m)} - \widetilde{\theta}\|^2$ term. Thus, $P^{(m)}$ provides a natural mechanism to balance variance reduction and bias introduction.

# 6 Experiments

## 6.1 Simulation Studies

In this section, we use simulation results to justify the behavior of using the sufficient variables under the fairness setting. Consider the multivariate linear regression, we simulate a dataset $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, where the sensitive attribute $\boldsymbol{Z} \in \{0, 1\}$ is binary.

Let $A$ be a randomly generated matrix with columns are orthonormal. Denote $A_Y \in \mathbb{R}^{p\times q}$ is the first $q$ columns of $A$ and $A_Z \in \mathbb{R}^{p\times r}$ is the $q-s$ to $q-s+r$ columns of A, which means $A_Y$ and $A_Z$ has $s$ shared columns. Define the latent variables $\boldsymbol{U}_Y = \boldsymbol{X} A_Y A_Y^\top$ and $\boldsymbol{U}_Z = \boldsymbol{X} A_Z A_Z^\top$. Let $\boldsymbol{X} \in \mathbb{R}^p$ drawn from a standard multivariate normal distribution: $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, I_p)$. The response $\boldsymbol{Y} \in \mathbb{R}^K$ is generated from

$$\boldsymbol{Y} = \boldsymbol{U}_Y \boldsymbol{\theta} + \varepsilon_Y, \quad \varepsilon_Y \sim \mathcal{N}(0, 0.5^2 I_K),$$

where the coefficients $\boldsymbol{\theta} \in \mathbb{R}^{p\times K}$ are randomly generated with each entry samples from $\mathcal{N}(1, 1)$. The sensitive variables $\boldsymbol{Z}$ is generated based on the following latent score

$$\boldsymbol{Z} = 1 \text{ if } \xi \geq 1, \text{ and } \boldsymbol{Z} = 0 \text{ otherwise}; \quad \text{where } \xi = \frac{1}{p}\sum_{j=1}^{p} \tanh([\boldsymbol{U}_Z]_j) + \varepsilon_Z \text{ and } \varepsilon_Z \sim \mathcal{N}(0, 1).$$

Finally, we introduce a distributional shift between groups by applying a non-linear transformation to the $\boldsymbol{X}$ samples based only on the span of $A_Z$. Specifically, for all samples $\boldsymbol{X}$ with $\boldsymbol{Z} = 1$,

$$\boldsymbol{X} \leftarrow \boldsymbol{X} + 0.5 \cdot \exp\left(\boldsymbol{X} A_Z A_Z^\top\right).$$

We generate synthetic data with parameters $n = 5000$, $p = 10$, $K = 5$, $q = 8$, $r = 8$, and $s = 6$. The dataset is randomly split into 4000 training samples and 1000 testing samples. We fit a multivariate ordinary least squares (OLS) model on both the original features $\boldsymbol{X}$ and the projected representations $P^{(m)}\boldsymbol{X}$ to estimate the model parameters $\widehat{\boldsymbol{\theta}}$.

To assess performance, we repeat the entire process over 30 independent replications. For each trial, we compute the root mean squared error (RMSE) on the test set overall and separately for the subgroups $\boldsymbol{Z} = 0$ and $\boldsymbol{Z} = 1$, along with the RMSE gap between groups. We also report the parameter distance $\|\widehat{\boldsymbol{\theta}}_n^{(m)} - \widehat{\boldsymbol{\theta}}_n\|$, where $\widehat{\boldsymbol{\theta}}_n^{(m)}$ is the OLS estimator obtained from the projected data $P^{(m)}\boldsymbol{X}$, and $\widehat{\boldsymbol{\theta}}_n$ is the baseline estimator from the original data. Results are averaged over the 30 trials and summarized in the left panel of Figure 1. We also project $\boldsymbol{X}$ onto the direction that best discriminates the sensitive attribute $\boldsymbol{Z}$ using Linear Discriminant Analysis (LDA), in order to visualize the distributional discrepancy between the two sensitive groups.

The simulation results clearly support our theoretical findings. Specifically, as the shared dimension increases, more directions in $\Phi$ are used to predict the target variable, incorporating additional sensitive and predictive information. This leads to a reduction in the MSE of the predictor and the parameter distance to the original estimator, but an increase in the MSE gap between sensitive groups. Both the distribution discrepancy and the Wasserstein distance also increase, indicating growing divergence between groups as more sensitive information is included.

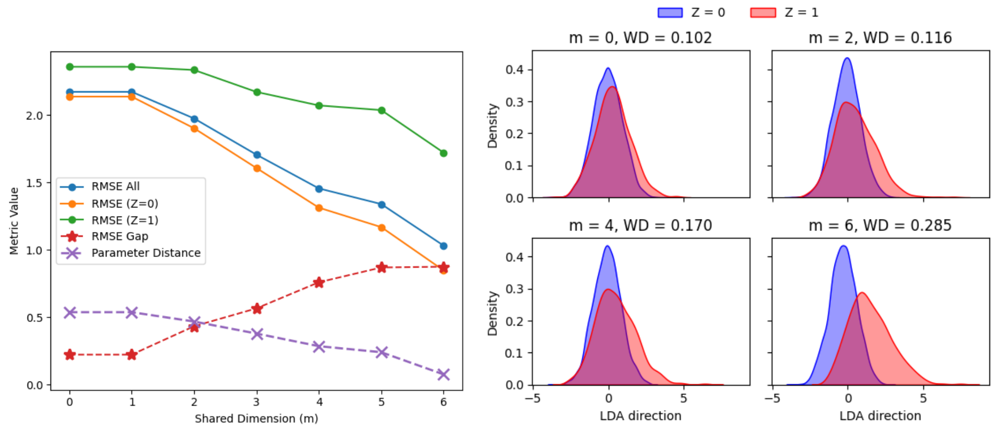


Figure 1: Left panel: Trends of RMSE and parameter distance as the number of shared dimensions increases, averaged over 30 replications. Right panel: Distributional discrepancy between sensitive groups visualized via LDA, along with the average Wasserstein distance (WD) across all $p$ dimensions between the original and projected data as the shared dimension increases in one replication.

## 6.2 Real Data Applications

We evaluate our proposed **Sequential Fair Projection (SFP)** method on two tabular datasets: **Adult** Kohavi [1996] and **Bank** Moro et al. [2014]. The **Adult** dataset contains personal data from over 40K individuals, with the task of predicting whether annual income exceeds $50K; gender is used as the sensitive attribute. The **Bank** dataset originates from Portuguese marketing campaigns, where the goal is to predict whether a client will subscribe to a deposit; age (over/under 25) is treated as the sensitive attribute. The data is standardized during preprocessing and split into training, validation, and test sets with a ratio of 70%:10%:20%.

We compare **SFP** against the following baselines: **Logistic Regression (LR)**, **AdvDebias** Zhang et al. [2018], **FairMixup** Chuang and Mroueh [2021], **DRAlign** Li et al. [2023], **DiffMCDP** Jin et al. [2024], **INLP** Ravfogel et al. [2020], **RLACE** Ravfogel et al. [2022], and **SUP** Shi et al. [2024]. Fairness is evaluated using TPR gap, DP, and MCDP, and utility is evaluated using accuracy. The detailed experimental settings are shown in Appendix B. All methods are repeated 20 times, and we report the average performance with standard deviations. The results are presented in Table 1.

Table 1: Performance metrics on the Adult and Bank datasets over 20 replications. Optimal results are in **bold**, and sub-optimal results are underlined.

**Adult Dataset**

| Method | Accuracy ↑ | DP ↓ | TPR ↓ | MCDP ↓ |
|---|---|---|---|---|
| LR | 84.90 | 17.37 | 6.87 | 35.12 |
| AdvDebias | 76.49 (0.65) | 12.45 (1.56) | 5.27 (0.53) | 29.32 (2.31) |
| FairMixup | 74.48 (0.43) | 3.93 (1.34) | 5.11 (0.34) | 24.91 (1.58) |
| DRAlign | 75.01 (0.41) | 7.58 (1.03) | 4.78 (0.29) | 22.04 (1.22) |
| DiffMCDP | 73.93 (0.31) | 5.92 (1.25) | 5.33 (0.46) | 11.50 (1.09) |
| INLP | 68.27 (0.57) | 4.16 (0.34) | 4.79 (0.51) | 8.58 (1.24) |
| RLACE | 72.79 (0.83) | 5.24 (0.56) | 4.56 (0.31) | **7.45 (0.86)** |
| SUP | 70.53 (0.36) | 4.33 (0.27) | **4.37 (0.29)** | 8.16 (1.25) |
| **SFP (Ours)** | **76.88 (0.47)** | **3.78 (0.15)** | 4.43 (0.26) | 10.52 (0.72) |

**Bank Dataset**

| Method | Accuracy ↑ | DP ↓ | TPR ↓ | MCDP ↓ |
|---|---|---|---|---|
| LR | 91.17 | 6.72 | 2.62 | 26.28 |
| AdvDebias | 60.56 (1.00) | 2.01 (1.36) | 2.20 (0.13) | 17.55 (1.39) |
| FairMixup | 59.48 (1.96) | 1.07 (0.25) | 2.21 (0.24) | 13.18 (2.94) |
| DRAlign | 59.12 (1.56) | 1.16 (0.41) | **1.96 (0.21)** | 13.61 (1.23) |
| DiffMCDP | 60.01 (1.83) | 1.19 (0.39) | 2.18 (0.13) | 11.00 (0.97) |
| INLP | 70.13 (0.86) | 0.93 (0.21) | 2.38 (0.31) | 10.44 (0.63) |
| RLACE | 72.51 (0.53) | 1.02 (0.35) | 2.26 (0.27) | **8.98 (0.38)** |
| SUP | 70.82 (0.47) | 0.96 (0.16) | 2.51 (0.22) | 10.63 (0.42) |
| **SFP (Ours)** | **89.66 (0.27)** | **0.51 (0.26)** | 2.33 (0.27) | 10.57 (0.34) |

Our results show that SFP consistently strikes an effective balance between predictive accuracy and fairness. On both the Adult and Bank datasets, SFP achieves competitive or superior accuracy compared to existing fairness-aware methods while substantially reducing group-level disparities. On the Adult dataset, SFP attains the highest accuracy (76.88%) and the lowest DP gap, demonstrating strong control over statistical bias. It also performs well in terms of TPR gap and MCDP, with only minor trade-offs relative to sub-optimal methods. On the Bank dataset, SFP also achieves the highest accuracy (89.66%) and the lowest DP gap (0.51), indicating minimal disparity. Although RLACE slightly outperforms in MCDP, SFP maintains competitive fairness across all metrics. These findings

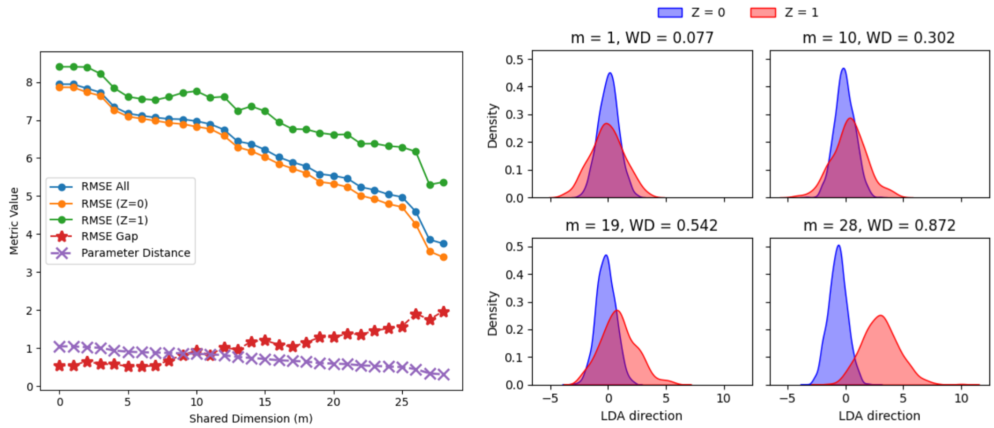


Figure 2: Trends of RMSE, parameter distance and distributional discrepancy as the number of shared dimensions increases when the linear SDR assumption is violated.

confirm that SFP is a practical and flexible framework for fairness-aware learning, enabling users to suppress sensitive information while preserving task-relevant predictive power.

### 6.3 Model Misspecification

In practice, the linear SDR assumption could be violated and the estimated central subspaces will fail to capture the full conditional independence structure. Therefore, conditional independence tests can be applied to verify whether the estimated projections preserve sufficiency. When the linear SDR assumption is strongly violated, the estimated matrices $M_Y$ and $M_Z$ may lose their low-rank structure. This indicates that no low-dimensional linear subspace can capture the dependence between $(X, Y)$ or $(X, Z)$. In such cases, one practical remedy is to retain a subset of directions corresponding to the leading eigenvalues and $M_{Y,Z}$ and construct a fair projection by eliminating leading directions.

To examine the trade-off between utility and fairness of SFP under model misspecification, we conduct additional experiments based on the simulation setup described in Section 6.1, with an added nonlinear term $\|\boldsymbol{X}\|_2^2/p$ in the generating processes of both $\boldsymbol{Y}$ and $\xi$. This modification violates the linear SDR assumption. We set $p = 30$, $K = 5$, and $q = r = s = 30$, making it impossible to recover a low-rank representation through SDR. After applying SFP, the estimated shared dimension is $\widehat{s} = 28$, indicating that the rank of $\widehat{M}_{Y,Z}$ is 28. We then sequentially add the directions obtained from $\widehat{M}_{Y,Z}$ and report the resulting trend in Figure 2.

The overall RMSE and RMSE gap using the original representation are 3.18 and 3.06, respectively. As shown in Figure 2, there is a clear trend that, as the number of shared dimensions used to construct the projection matrix increases, both the overall MSE and the parameter distance decrease, indicating that the projected representation gradually approaches the information contained in the original features. Meanwhile, the RMSE gap and the distributional discrepancy increase as more sensitive information is included. This demonstrates that SFP can still capture the trade-off between utility and fairness even when the linear SDR assumption is violated.

## 7 Discussion

We propose a principled and model-agnostic framework for fairness-aware learning through subspace decomposition of data representations. Our method manipulates the representation space by selectively removing shared information between the target and sensitive attributes, which enables flexible control over the fairness-utility trade-off. We provide theoretical guarantees on how prediction error and fairness metrics evolve as more sensitive information is incorporated, and further apply influence function analysis to characterize the impact of partially fair representations on estimator behavior. Empirical results on both synthetic and real-world datasets validate our theoretical insights, demonstrating that the proposed SFP method achieves good performance.

## Acknowledgements

Bei Jiang and Linglong Kong were partially supported by grants from the Canada CIFAR AI Chairs program, the Alberta Machine Intelligence Institute (AMII), and Natural Sciences and Engineering Council of Canada (NSERC), and Linglong Kong was also partially supported by grants from the Canada Research Chair program from NSERC.

## References

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

Eric Potash, Joe Brew, Alexander Loewi, Subhabrata Majumdar, Andrew Reece, Joe Walsh, Eric Rozier, Emile Jorgenson, Raed Mansour, and Rayid Ghani. Predictive modeling for public health: Preventing childhood lead poisoning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2039–2047, 2015.

Cynthia Rudin and Berk Ustun. Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, 48(5):449–466, 2018.

Gijs Van Dijck. Predicting recidivism risk meets ai act. *European Journal on Criminal Policy and Research*, 28(3):407–423, 2022.

Marco Billi, Thiago Raulino Dal Pont, Isabela Cristina Sabo, Francesca Lagioia, Giovanni Sartor, and Aires José Rover. Supervised learning, explanation and interpretation from pretrial detention decisions by italian and brazilian supreme courts. In *International Conference on Conceptual Modeling*, pages 131–140. Springer, 2023.

Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. In *Proceedings of the seventeenth international conference on artificial intelligence and law*, pages 83–92, 2019.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Tiago P Pagano, Rafael B Loureiro, Fernanda VN Lisboa, Rodrigo M Peixoto, Guilherme AS Guimarães, Gustavo OR Cruz, Maira M Araujo, Lucas L Santos, Marco AS Cruz, Ewerton LS Oliveira, et al. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, 7(1):15, 2023.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.

Damien Dablain, Bartosz Krawczyk, and Nitesh Chawla. Towards a holistic view of bias in machine learning: bridging algorithmic fairness and imbalanced learning. *Discover Data*, 2(1):4, 2024.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Bertrand K Hassani. Societal bias reinforcement through machine learning: a credit scoring perspective. *AI and Ethics*, 1(3):239–247, 2021.

Yang Hu, Nicole Denier, Lei Ding, Monideepa Tarafdar, Alla Konnikov, Karen D Hughes, Shenggang Hu, Bran Knowles, Enze Shi, Jabir Alshehabi Al-Ani, et al. Language in job advertisements and the reproduction of labor force gender and racial segregation. *PNAS nexus*, 3(12):pgae526, 2024.

Lei Ding, Yang Hu, Nicole Denier, Enze Shi, Junxi Zhang, Qirui Hu, Karen Hughes, Linglong Kong, and Bei Jiang. Probing social bias in labor market text generation by chatgpt: A masked language model approach. *Advances in Neural Information Processing Systems*, 37:139912–139937, 2024.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 35–50. Springer, 2012.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *Uncertainty in artificial intelligence*, pages 862–872. PMLR, 2020.

Huan Tian, Bo Liu, Tianqing Zhu, Wanlei Zhou, and S Yu Philip. Multifair: Model fairness with multiple sensitive attributes. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. Fairness improvement with multiple protected attributes: How far are we? In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13, 2024.

Jinqiu Jin, Haoxuan Li, and Fuli Feng. On the maximal local disparity of fairness-aware classifiers. In *Forty-first International Conference on Machine Learning*, 2024.

Tianlin Li, Qing Guo, Aishan Liu, Mengnan Du, Zhiming Li, and Yang Liu. Fairer: fairness as decision rationale alignment. In *International Conference on Machine Learning*, pages 19471–19489. PMLR, 2023.

R Dennis Cook and Bing Li. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474, 2002.

Kofi P Adragni and R Dennis Cook. Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405, 2009.

Xiaomeng Wang, Yishi Zhang, and Ruilin Zhu. A brief review on algorithmic fairness. *Management System Engineering*, 1(1):7, 2022a.

Richard A Berk, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Fair risk algorithms. *Annual Review of Statistics and Its Application*, 10(1):165–187, 2023.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, 2024.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20 (75):1–42, 2019.

Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129. PMLR, 2019.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. Optimising equal opportunity fairness in model training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4073–4084, 2022.

Ruijia Wang, Xiao Wang, Chuan Shi, and Le Song. Uncovering the structural fairness in graph contrastive learning. *Advances in neural information processing systems*, 35:32465–32473, 2022b.

Fengda Zhang, Kun Kuang, Long Chen, Yuxuan Liu, Chao Wu, and Jun Xiao. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *The Eleventh International Conference on Learning Representations*, 2022.

Xudong Han, Timothy Baldwin, and Trevor Cohn. Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2760–2765, 2021.

Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020.

Xudong Han, Timothy Baldwin, and Trevor Cohn. Balancing out bias: Achieving fairness through balanced training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11335–11350, 2022.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, 2020.

Shauli Ravfogel, Yoav Goldberg, and Ryan Cotterell. Log-linear guardedness and its implications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9413–9431, 2023.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pages 18400–18421. PMLR, 2022.

Enze Shi, Lei Ding, Linglong Kong, and Bei Jiang. Debiasing with sufficient projection: A general theoretical framework for vector representations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5960–5975, 2024.

Enze Shi, Linglong Kong, and Bei Jiang. Deep fair learning: A unified framework for fine-tuning representations with sufficient networks. *arXiv preprint arXiv:2504.06470*, 2025.

Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769, 2007.

Wei Luo. On efficient dimension reduction with respect to the interaction between two response variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):269–294, 2022.

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

R Dennis Cook and Sanford Weisberg. Discussion of sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):328–332, 1991.

Bing Li and Shaoli Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007.

Wei Luo and Bing Li. Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 103(4):875–887, 2016.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 202–207, 1996.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2021.

Li-Xing Zhu and Kai-Tai Fang. Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics*, 24(3):1053–1068, 1996.

Kyongwon Kim, Bing Li, Zhou Yu, and Lexin Li. On post dimension reduction statistical inference. *Annals of Statistics*, 48(3):1567–1592, 2020.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All the claims made in the abstract and introduction are established in Methodology and Theoretical Analysis sections and demonstrated on both simulations and real datasets.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section D of the Appendix

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are clearly stated and all the proofs are provided in Section C of Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the technical details of our experimental setup. Furthermore, we provide the code necessary to reproduce our results in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All datasets used are publicly available, and we have uploaded code in the supplementary materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide all the technical details of our experimental setup.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All the results are reported with mean and standard deviation for multiple replicates.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We conduct all our experiments on an Ubuntu Server with CPU AMD Ryzen Threadripper PRO 3995WX 64-Cores Processor and 256G RAM.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impact in Section E of Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the methods, codes and datasets used in the paper are properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper propose new algorithm and does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM used only for editing and grammatical refinements.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

# A Influence Function of Projections

In this section, we give explicit form of the influence function of projection $P^{(m)}$, which can used to derive the asymptotic normality of the parameter estimators.

Let $\{(\lambda_{\phi,i}, \phi_i)\}_{i=1}^{s}$ be the eigenvalues and associated eigenvectors of $M_{Y,Z}$, $\{(\lambda_{\beta,i}, \beta_i)\}_{i=1}^{q}$ be the eigenvalues and associated eigenvectors of $M_Y$ and $\{(\lambda_{\psi,i}, \psi_i)\}_{i=1}^{r}$ be the eigenvalues and associated eigenvectors of $M_Z$.

Then, by Zhu and Fang [1996], the influence function of the directions estimated by SDR techniques can be written as, for $i = 1, \dots, s$,

$$\phi_i^* = \sum_{j=1,\, j\neq i}^{s} \frac{\phi_j \phi_j^\top M_{Y,Z}^* \phi_i}{\lambda_{\phi,i} - \lambda_{\phi,j}} \tag{11}$$

And for $k = 1, \dots, q - s$

$$(Q_z \beta_k)^* = Q_z \beta_k^* + Q_z^* \beta_k = Q_z \beta_k^* + \beta_k \left( I_p - \sum_{j=1}^{r} (\psi_j^* \psi_j^\top + \psi_j \psi_j^{*\top}) \right) \tag{12}$$

where

$$\beta_k^* = \sum_{\ell=1,\, \ell\neq k}^{q} \frac{\beta_\ell \beta_\ell^\top M_Y^* \beta_k}{\lambda_{\beta,k} - \lambda_{\beta,\ell}} \quad \text{and} \quad \psi_j^* = \sum_{\ell=1,\, \ell\neq j}^{r} \frac{\psi_\ell \psi_\ell^\top M_Z^* \psi_j}{\lambda_{\psi,j} - \lambda_{\psi,\ell}}.$$

Then we have

$$P^{(m)*} = (Q_z\beta_k)^* (Q_z\beta_k)^\top + (Q_z\beta_k)(Q_z\beta_k)^{*\top} + \Phi_m^* \Phi_m^\top + \Phi_m \Phi_m^{*\top},$$

where $\Phi_m^* = (\phi_1^*, \dots, \phi_m^*)$.

The influence functions for the candidate matrices $M_Y$ and $M_Z$ estimated via SDR techniques have been well studied. For details, we refer readers to Section 4 of Kim et al. [2020] and omit the derivations here.

# B Experiments Details

## B.1 Experiment Setup

Throughout the experiments, we use the MSAVE method to estimate the candidate matrices $M_Y$ and $M_Z$, with the number of slices set to $p + 1$, where $p$ denotes the dimensionality of the input $\boldsymbol{X}$. In the post-SDR training procedure (as described in Algorithm 2), we apply multivariate linear regression for synthetic simulations and logistic regression for real-world datasets. The MCDP metric is used to select the optimal fair model $\mathcal{M}_{\text{fair}}^{(m^*)}$. To estimate the rank of each candidate matrix, we adopt the ladle estimator with 30 bootstrap replications.

For baseline methods, we use the official code released by the respective authors. Specifically, for AdvDebias, FairMixup, DRAlign, and DiffMCDP, we use the implementation provided in Jin et al. [2024]. For INLP, we set the number of iterations to 100. For RLACE, we follow the training hyperparameters used in the original paper. For SUP, we adopt the same number of slices ($p + 1$) as in our method, and select the dimension to remove based on 10-fold cross-validation, choosing the model that achieves the lowest MCDP.

## B.2 Fairness Measurement

We consider a multi-class classification setting where the target label $\boldsymbol{Y} \in \mathbb{R}^K$ is a one-hot encoded vector, i.e., $\boldsymbol{Y}_j = 1$ if the true label corresponds to class $j$ and 0 otherwise. We denote the predicted probability vector as $\hat{\boldsymbol{Y}}$, and the sensitive attribute as a binary variable $\boldsymbol{Z} \in \{0, 1\}$.

**Demographic Parity (DP) Gap.** Demographic Parity requires that the predicted output is independent of the sensitive attribute. For each class $j \in \{1, \dots, K\}$, we define the group-wise expected predicted score as:

$$\mathrm{DP}_{z,j} = \mathbb{E}[\hat{\boldsymbol{Y}}_j \mid \boldsymbol{Z} = z],$$

and the corresponding DP gap for class $j$ as:

$$\mathrm{DP}_{\text{gap},j} = \mathrm{DP}_{1,j} - \mathrm{DP}_{0,j}.$$

The overall DP gap across classes is then aggregated as:

$$\mathrm{DP}_{\text{gap}} = \sqrt{\frac{1}{K-1} \sum_{j=1}^{K-1} (\mathrm{DP}_{\text{gap},j})^2} \times 100\%.$$

When $K = 2$, this definition reduces to the binary case, where $\mathrm{DP}_{\text{gap}} = |\mathrm{DP}_{1,j} - \mathrm{DP}_{0,j}|$.

**True Positive Rate (TPR) Gap.** For each class $j \in \{1, \dots, K\}$, the TPR for group $z \in \{0, 1\}$ is defined as:

$$\mathrm{TPR}_{z,j} = \mathbb{P}(\hat{\boldsymbol{Y}}_j = 1 \mid \boldsymbol{Z} = z, \boldsymbol{Y}_j = 1),$$

and the corresponding TPR gap is:

$$\mathrm{TPR}_{\text{gap},j} = \mathrm{TPR}_{1,j} - \mathrm{TPR}_{0,j}.$$

We aggregate TPR gaps across all classes into a single fairness metric:

$$\mathrm{TPR}_{\text{gap}} = \sqrt{\frac{1}{K} \sum_{j=1}^{K} (\mathrm{TPR}_{\text{gap},j})^2} \times 100\%.$$

**Maximal Cumulative Ratio Disparity along Predictions (MCDP).** For each class $j$, define the group-wise cumulative distribution function:

$$F_{z,j}(y) = \mathbb{P}(\hat{\boldsymbol{Y}}_j \le y \mid \boldsymbol{Z} = z).$$

The MCDP for class $j$ is then the Kolmogorov–Smirnov distance between the distributions of predicted probabilities for the two sensitive groups:

$$\mathrm{MCDP}_j = \max_{y \in [0,1]} \left| F_{1,j}(y) - F_{0,j}(y) \right|.$$

Since the components of $\hat{\boldsymbol{Y}}$ sum to 1 due to the softmax constraint, only $K-1$ of them are linearly independent. Therefore, we define the overall MCDP gap as:

$$\mathrm{MCDP}_{\text{gap}} = \sqrt{\frac{1}{K-1} \sum_{j=1}^{K-1} (\mathrm{MCDP}_j)^2} \times 100\%.$$

When $K = 2$, this definition reduces to the binary case as in Jin et al. [2024].

# C Proof of Theorems

*Proof of Theorem 3.1.* Let $B$ be an orthonormal basis for the central subspace $\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}}$, so that $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}] = \mathbb{E}[\boldsymbol{Y} \mid B^\top \boldsymbol{X}]$. Since $Q_z \boldsymbol{X}$ is a measurable function of $\boldsymbol{X}$, and $\boldsymbol{Y} \perp \boldsymbol{X} \mid B^\top \boldsymbol{X}$, it follows that $\mathbb{E}[\boldsymbol{Y} \mid Q_z \boldsymbol{X}] = \mathbb{E}[\boldsymbol{Y} \mid B^\top Q_z \boldsymbol{X}]$. Noting that $B^\top Q_z \boldsymbol{X} = (Q_z B)^\top \boldsymbol{X}$, we conclude that

$$\mathbb{E}[\boldsymbol{Y} \mid Q_z \boldsymbol{X}] = \mathbb{E}[\boldsymbol{Y} \mid (Q_z B)^\top \boldsymbol{X}],$$

which implies that $Q_z B$ spans a sufficient subspace for $\boldsymbol{Y}$ with respect to $Q_z \boldsymbol{X}$.

By definition, the central subspace for $\boldsymbol{Y} \mid Q_z \boldsymbol{X}$ is the minimal subspace satisfying this conditional independence. Therefore, if $\widetilde{B}$ is any orthonormal matrix such that $\mathbb{E}[\boldsymbol{Y} \mid Q_z \boldsymbol{X}] = \mathbb{E}[\boldsymbol{Y} \mid \widetilde{B}^\top \boldsymbol{X}]$, it must hold that

$$\mathrm{Span}(\widetilde{B}) \subseteq \mathrm{Span}(Q_z B).$$

and $\mathrm{Span}(\widetilde{B}) = \mathrm{Span}(Q_z B)$ if and only if $\mathrm{rank}(\widetilde{B}) = \mathrm{rank}(Q_z B)$ □

*Proof of Theorem 3.2.* We begin by substituting the model into the expected error:

$$\mathbb{E}[(Y-\tilde{f}^{(m)})^2] = \mathbb{E}[(f^*(\boldsymbol{X})+\varepsilon_Y-\tilde{f}^{(m)})^2] = \mathbb{E}\left[(f^*(\boldsymbol{X}) - \tilde{f}^{(m)})^2 + 2(f^*(\boldsymbol{X}) - \tilde{f}^{(m)})\varepsilon_Y + \varepsilon_Y^2\right].$$

Take expectations term-by-term. For the first term, we observe that $\tilde{f}^{(m)} = \mathbb{E}[Y \mid \boldsymbol{\Xi}^{(m)}]$, which is the projection of $Y$ onto a coarser sigma-algebra. Therefore $\tilde{f}^{(m)} = \mathbb{E}[f^*(\boldsymbol{X}) \mid \boldsymbol{\Xi}^{(m)}]$ and

$$\mathbb{E}[(f^*(\boldsymbol{X}) - \tilde{f}^{(m)})^2] = \mathbb{E}[\mathrm{Var}(f^*(\boldsymbol{X}) \mid \boldsymbol{\Xi}^{(m)})].$$

For the second term, since $\mathbb{E}[\varepsilon_Y \mid \boldsymbol{\Xi}^{(m)}] = \mathbb{E}[\mathbb{E}[\varepsilon_Y \mid \boldsymbol{X}] \mid \boldsymbol{\Xi}^{(m)}] = 0$, we have:

$$\mathbb{E}[(f^*(\boldsymbol{X}) - \tilde{f}^{(m)})\varepsilon_Y] = 0.$$

Finally, the third term is simply $\mathbb{E}[\varepsilon_Y^2] = \sigma_Y^2$. Putting everything together gives:

$$\mathbb{E}[(Y - \tilde{f}^{(m)})^2] = \mathbb{E}[\mathrm{Var}(f^*(\boldsymbol{X}) \mid \boldsymbol{\Xi}^{(m)})] + \sigma_Y^2.$$

Note that $\tilde{f}^{(m)}(\boldsymbol{\Xi}^{(m)})$ is the orthogonal projection of $f^*(\boldsymbol{X}) := \mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ onto $\sigma(\boldsymbol{\Xi}^{(m)})$ in $L^2$. Since $\boldsymbol{\Xi}^{(m)} \subset \boldsymbol{\Xi}^{(m+1)}$, the projection becomes finer, and by the Pythagorean identity in $L^2$:

$$\|f^*(\boldsymbol{X}) - \tilde{f}^{(m+1)}\|_{L^2}^2 \leq \|f^*(\boldsymbol{X}) - \tilde{f}^{(m)}\|_{L^2}^2.$$

Hence, the approximation error $\Delta(m) := \mathbb{E}[\mathrm{Var}(f^*(\boldsymbol{X}) \mid \boldsymbol{\Xi}^{(m)})]$ satisfies $\Delta(m+1) \leq \Delta(m)$. □

*Proof of Theorem 3.3.* From the definition of distance covariance for scalar $X$ and binary $Z$, we consider the V-statistic form:

$$\mathrm{dCov}^2(X, Z) = \mathbb{E}[|X - X'| \cdot |Z - Z'|] + \mathbb{E}[|X - X'|] \cdot \mathbb{E}[|Z - Z'|] - 2\mathbb{E}[|X - X'| \cdot |Z - Z''|].$$

Specializing to $Z \in \{0, 1\}$, we know:

$$\mathbb{E}[|Z - Z'|] = \mathbb{E}[|Z - Z''|] = 2p(1 - p),$$

and $|Z - Z'| = 1$ only when $Z \neq Z'$, meaning cross-group expectation. Then:

$$\mathbb{E}[|X - X'| \cdot |Z - Z'|] = 2p(1 - p) \cdot \mathbb{E}[|X_0 - X_1|],$$
$$\mathbb{E}[|X - X'| \cdot |Z - Z''|] = 2p(1 - p) \cdot \mathbb{E}[|X - X'|].$$

So:

$$\mathrm{dCov}^2(X, Z) = 2p(1 - p)\left(\mathbb{E}[|X_0 - X_1|] - \mathbb{E}[|X - X'|]\right).$$

Substituting $X = \tilde{f}^{(m)}$, we obtain the stated identity. Specifically, when $m = 0$, we have $\tilde{f}^{(m)} = \tilde{f}_0^{(m)} = \tilde{f}_1^{(m)}$ and thereby the distance covariance equals zero. □

*Proof of Theorem 4.1.* Following Luo [2022], $\mathrm{Span}\{\Phi\}$ is the unique solution of

$$M_Y\, \Sigma\, P_{\Sigma,\Phi}\, M_Z \;=\; M_Y\, \Sigma\, M_Z, \tag{13}$$

where, for any full rank $B$, $P_{\Sigma,B}$ is the $\Sigma$-orthogonal projector onto $\mathrm{Span}\{B\}$.

Observe that

$$P_{\Sigma,\Phi} = \Sigma^{-1/2}\, P_U\, \Sigma^{1/2}, \quad \text{where} \quad U = \Sigma^{1/2}\, \Phi, \quad P_U = U\, (U^\top U)^{-1} U^\top.$$

Substituting into (13) yields

$$M_Y\, \Sigma^{1/2}\, P_U\, \Sigma^{1/2}\, M_Z = M_Y\, \Sigma\, M_Z.$$

Left- and right-multiplying by $\Sigma^{-1/2}$ gives

$$A\, P_U\, B = E, \quad A = \Sigma^{-1/2} M_Y\, \Sigma^{1/2}, \quad B = \Sigma^{1/2} M_Z\, \Sigma^{-1/2}, \quad E = A\, B = \Sigma^{-1/2} M_Y\, \Sigma\, M_Z\, \Sigma^{-1/2}.$$

Since $\mathrm{range}(A^\top) \subseteq \mathrm{range}(U)$ and $\mathrm{range}(B) \subseteq \mathrm{range}(U)$, we have $A\,P_U = A$ and $P_U\,B = B$. It follows that

$$E\,P_U = E.$$

Because $E$ is symmetric, its nonzero-eigenvalue subspace coincides with $\mathrm{range}(U)$. Writing the spectral decomposition

$$E = V\,\Lambda\,V^\top, \quad V^\top V = I,$$

and setting $\Phi = \Sigma^{-1/2} V_{(s)}$ (the $s$ leading eigenvectors), we recover the equivalent generalized eigenproblem

$$(M_Y\,\Sigma\,M_Z)\,\phi = \lambda\,\Sigma\,\phi,$$

whose top $s$ eigenvectors $\{\phi_i\}$ span $\mathrm{Span}\{\Phi\}$. □

*Proof of Theorem 5.2.* It is well-known that, if $\theta^{(m)}(F)$ is Hadamard differentiable, then it satisfies the following expansion

$$\begin{aligned}\theta^{(m)}(F_n, P^{(m)}(F_n)) &= \theta^{(m)}(F_0, P^{(m)}(F_0)) + \mathbb{E}_n[\theta^{(m)*}] + o_p(n^{-1/2}) \\ &= \theta^{(m)}(F_0, P^{(m)}(F_0)) + \frac{1}{n}\sum_{i=1}^{n} \theta^{(m)*}(S_i, P^{(m)}(F_n)) + o_p(n^{-1/2})\end{aligned}$$

By Theorem 1 and Proposition 1 in Kim et al. [2020], since $L(S,\theta)$ is differentiable, we have

$$\begin{aligned}\theta^{(m)*}(S, P^{(m)}(F_n)) &= \theta^{(m)*}(S, P^{(m)}(F_0)) + \left(\frac{\partial\,\theta(F_0, P^{(m)}(F_0))}{\partial\,\mathrm{vec}(P^{(m)})}\right)^\top \mathrm{vec}(P^{(m)*}(S)) \\ &= -H^{-1}_{\widetilde{\theta}^{(m)}} \nabla_\theta L(S^{(m)}, \widetilde{\theta}^{(m)}) + \left(\frac{\partial\,\theta(F_0, P^{(m)}(F_0))}{\partial\,\mathrm{vec}(P^{(m)})}\right)^\top \mathrm{vec}(P^{(m)*}(S)),\end{aligned}$$

where $H^{-1}_{\widetilde{\theta}^{(m)}} = \mathbb{E}[\nabla^2_\theta L(S, \widetilde{\theta}^{(m)})]$. Therefore, we have

$$\begin{aligned}\widehat{\theta}^{(m)}_n &= \widetilde{\theta}^{(m)} + \frac{1}{n}\sum_{i=1}^{n} \theta^{(m)*}(S_i, P^{(m)}(F_n)) + o_p(n^{-1/2}) \\ &= \widetilde{\theta}^{(m)} + \frac{1}{n}\sum_{i=1}^{n}\left[-H^{-1}_{\widetilde{\theta}^{(m)}} \nabla_\theta L(S_i^{(m)}, \widetilde{\theta}^{(m)}) + \left(\frac{\partial\,\theta(F_0, P^{(m)}(F_0))}{\partial\,\mathrm{vec}(P^{(m)})}\right)^\top \mathrm{vec}(P^{(m)*}(S_i))\right] + o_p(n^{-1/2})\end{aligned}$$

Consequently, by the Central Limit Theorem

$$\sqrt{n}(\widehat{\theta}^{(m)}_n - \widetilde{\theta}^{(m)}) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \mathrm{Var}\big[-H^{-1}_{\widetilde{\theta}^{(m)}} \nabla_\theta L(S^{(m)}, \widetilde{\theta}^{(m)}) + D^{(m)}\,\mathrm{vec}(P^{(m)*})\big]\right),$$

where $D^{(m)} = (\partial\,\theta(F_0, P^{(m)}(F_0))/\partial\,\mathrm{vec}(P^{(m)}))^\top$. □

*Proof of Corollary 5.3.* Similar as shown in Theorem 5.2, the original estimator satisfies the following expansion

$$\begin{aligned}\widehat{\theta}_n &= \widetilde{\theta} + \frac{1}{n}\sum_{i=1}^{n} \theta^*(S_i) + o_p(n^{-1/2}) \\ &= \widetilde{\theta} + \frac{1}{n}\sum_{i=1}^{n} H^{-1}_{\widetilde{\theta}} \nabla_\theta L(S_i, \widetilde{\theta}^{(m)}) + o_p(n^{-1/2})\end{aligned}$$

Then we have

$$\begin{aligned}&(\widehat{\theta}^{(m)}_n - \widehat{\theta}_n) - (\widetilde{\theta}^{(m)} - \widetilde{\theta}) \\ =&\frac{1}{n}\sum_{i=1}^{n}\left[H^{-1}_{\widetilde{\theta}} \nabla_\theta L(S_i, \widetilde{\theta}^{(m)}) - H^{-1}_{\widetilde{\theta}^{(m)}} \nabla_\theta L(S_i^{(m)}, \widetilde{\theta}^{(m)}) + \left(\frac{\partial\,\theta(F_0, P^{(m)}(F_0))}{\partial\,\mathrm{vec}(P^{(m)})}\right)^\top \mathrm{vec}(P^{(m)*}(S_i))\right]\end{aligned}$$

Therefore, by Central Limit Theorem, we have

$$\mathbb{E}(\|\widehat{\theta}^{(m)}_n - \widehat{\theta}_n\|^2) \le \|\widetilde{\theta}^{(m)} - \widetilde{\theta}\|^2 + \frac{1}{n}\,\mathrm{Tr}\left(\mathrm{Var}\big[H^{-1}_{\widetilde{\theta}} G - H^{-1}_{\widetilde{\theta}^{(m)}} G^{(m)} + D^{(m)}\,\mathrm{vec}(P^{(m)*})\big]\right),$$

□

*Proof of Theorem 5.4.* Consider the first-order Taylor expansion of $f(x;\theta)$

$$f(x;\widehat{\theta}_n^{(m)}) - f(x;\widetilde{\theta}^{(m)}) = \nabla_\theta f(x;\widetilde{\theta}^{(m)})^\top(\widehat{\theta}_n^{(m)} - \widetilde{\theta}^{(m)}) + o(\|\widehat{\theta}_n^{(m)} - \widetilde{\theta}^{(m)}\|).$$

From Theorem 5.2, we know $\|\widehat{\theta}_n^{(m)} - \widetilde{\theta}^{(m)}\| = O_p(n^{-1/2})$. Then, following the same expansion as shown in the proof of Theorem 5.2, along with Slutsky's theorem, gives the results. □

# D Limitations

The current work builds upon linear SDR, which assumes that both the target label and the sensitive attribute depend on linear projections of the representation. While this is a relatively mild assumption, it limits our ability to capture and mitigate nonlinear dependencies that may be embedded in the representation.

A promising future direction is to extend the framework using nonlinear SDR theory. In this approach, the representation is first mapped into a higher-dimensional feature space using a predefined kernel function, denoted by $\mathrm{Ker}(\boldsymbol{X})$. The goal is then to find a projection matrix $B_z$ such that

$$\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{X} \mid B_z^\top \mathrm{Ker}(\boldsymbol{X}),$$

enforcing conditional independence in the nonlinear space. Importantly, the core derivations and structural components of our method remain applicable in this generalized setting. Developing this extension is a promising direction for future research.

# E Impact Statement

The framework proposed in this paper tackles the fundamental challenge of fairness in machine learning by directly mitigating bias in learned data representations. Departing from post-hoc debiasing methods, we enforce fairness for new representations, enabling scalable and generalizable solutions for sensitive domains such as employment, healthcare, and finance. By balancing fairness and predictive utility through subspace decomposition, SFP helps reduce systemic disparities and promotes the development of trustworthy AI systems. This work underscores the importance of rigorous fairness evaluation and responsible deployment to advance equitable and accountable machine learning technologies.