Decomposing Direct and Indirect Biases in Linear Models under Demographic Parity Constraint

Bertille Tierny[1,2], Arthur Charpentier[3], François Hu[1]

[1]Milliman France, R&D Department, AI Lab
[2]ENSAE - Institut Polytechnique de Paris
[3]Université du Québec à Montréal
bertille.tierny@milliman.com, charpentier.arthur@uqam.ca, francois.hu@milliman.com

Abstract

Linear models are widely used in high-stakes decision-making due to their simplicity and interpretability. Yet when fairness constraints such as demographic parity are introduced, their effects on model coefficients, and thus on how predictive bias is distributed across features, remain opaque. Existing approaches on linear models often rely on strong and unrealistic assumptions, or overlook the explicit role of the sensitive attribute, limiting their practical utility for fairness assessment. We extend the work of (Chzhen and Schreuder 2022) and (Fukuchi and Sakuma 2023) by proposing a post-processing framework that can be applied on top of any linear model to decompose the resulting bias into direct (sensitive-attribute) and indirect (correlated-features) components. Our method analytically characterizes how demographic parity reshapes each model coefficient, including those of both sensitive and non-sensitive features. This enables a transparent, feature-level interpretation of fairness interventions and reveals how bias may persist or shift through correlated variables. Our framework requires no retraining and provides actionable insights for model auditing and mitigation. Experiments on both synthetic and real-world datasets demonstrate that our method captures fairness dynamics missed by prior work, offering a practical and interpretable tool for responsible deployment of linear models.

Code — https://github.com/bias-mitigator/interpretable.git

1 Introduction

Linear models remain a foundational tool in statistical learning due to their interpretability, scalability, and simplicity (Hastie et al. 2009). They are widely used in high-stakes domains such as credit scoring, hiring, insurance, and healthcare, where algorithmic decisions have significant consequences and fairness considerations are critical (Obermeyer et al. 2019; Barocas, Hardt, and Narayanan 2023). In these settings, linear models may inadvertently encode or amplify unfair biases. These biases can arise *directly*, through the explicit use of sensitive attributes such as race or gender, or *indirectly*, through features correlated with those attributes (Hajian and Domingo-Ferrer 2012; Nabi and Shpitser 2018; Tang, Zhang, and Zhang 2023). Fairness in machine learning has been extensively studied, with various formal definitions and mitigation strategies proposed (Del Barrio, Gordaliza, and Loubes 2020; Mehrabi et al. 2021; Pessach and Shmueli

2022). One of the most common criteria is *Demographic Parity* (DP), which requires that the predictions be statistically independent of sensitive attributes. Although many methods aim to enforce DP in classification settings (Agarwal et al. 2018; Gaucher, Schreuder, and Chzhen 2023; Hu, Ratz, and Charpentier 2024), few provide systematic tools to quantify and separate the sources of unfairness, especially in linear models. In particular, existing approaches, such as (Chzhen and Schreuder 2022; Fukuchi and Sakuma 2023), do not provide a systematic decomposition of bias stemming from the sensitive feature versus that induced by correlated non-sensitive features. This lack of decomposition is especially problematic in linear models, which, despite their transparency, are not well understood in terms of how fairness constraints affect individual model coefficients. As a result, practitioners often lack insight into how these constraints redistribute predictive contributions across features or whether indirect biases persist even after sensitive variables are removed.

1.1 Main Contributions

We propose a framework for learning fair linear models, designed to identify and mitigate both indirect and direct biases in linear models. Specifically:

- We introduce a linear modeling framework aligned with standard practices and derive a closed-form solution for the optimal fair regressor. To our knowledge, this is the first solution that remains linear under group-wise feature standardization. In practice, it can be applied on top of any linear model (penalized, with or without intercept) making it broadly compatible and easily deployable.

- Building on this optimal solution, we disentangle the contributions of sensitive and non-sensitive features to fairness violations (see Fig. 1) while providing clear guidance on how to adjust coefficients toward fairness.

- We illustrate the effectiveness of our approach on both synthetic and real-world datasets, demonstrating its ability to produce fair linear models while offering interpretability of both direct and indirect biases.

This work advances the understanding of fairness in linear models and contributes to the broader literature by providing tools to dissect and interpret bias at the feature level. For
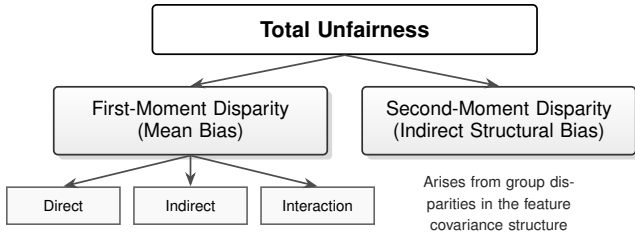
Figure 1: Conceptual decomposition of the total unfairness measure. The unfairness splits into two bias sources: disparities in the **mean** of predictions (First-Moment) and disparities in the **variance** of predictions (Second-Moment).

clarity of presentation, all proofs are provided in the supplementary materials.

1.2 Related Work

The study of fairness constraints in linear regression, particularly under DP, is relatively recent. Most existing methods either focus on model-level fairness objectives or rely on restrictive assumptions that limit their applicability in practice.

(Chzhen and Schreuder 2022) propose a minimax solution for linear regression under DP, deriving a closed-form intercept correction. However, their formulation is based on a strong assumption: the sensitive feature is independent of the other covariates. Therefore, they are omitting completely the indirect biases. This assumption rarely holds in real-world data and significantly restricts both the predictive accuracy of the model and the relevance of its fairness guarantees.

(Fukuchi and Sakuma 2023) extend this line of work by adjusting both intercept and non-sensitive feature coefficients. Although this allows more flexibility, their framework still omits an explicit treatment of the sensitive feature's contribution, which limits bias diagnostics. Moreover, their solution also still builds on simplifying assumptions that may distort the fairness-performance trade-off.

In contrast, our approach explicitly characterizes the effect of DP constraints on all model components, including the sensitive feature. This enables a fine-grained decomposition of direct and indirect biases and provides clearer insights into how fairness interventions affect both predictive behavior and feature-level fairness contributions.

	Direct (Mean)	Indirect (Mean)	Interaction	Indirect (Structural)
[CS22]	✓		✓	
[FS23]	✓	✓	✓	
ours	✓	✓	✓	✓

Table 1: Comparison of bias mitigation methods across linear models proposed by [CS22] (Chzhen and Schreuder 2022), [FS23] (Fukuchi and Sakuma 2023), and our approach. Checkmarks indicate addressed biases.

1.3 Outline of the Paper

The remainder of this article is structured as follows: Section 2, introduces the problem setup and the key metrics used throughout the article. Section 3 reviews the limitations of existing fair linear models. Section 4 presents our main contribution: a general framework for learning optimal fair linear models. This is followed in Section 5 by a decomposition of unfairness into direct and indirect biases. Finally, Section 6 details the practical implementation of our methodology and Section 7 presents numerical results comparing our method to state-of-the-art baselines.

2 Problem Formulation

Let $(\boldsymbol{X}, S, Y)$ be a random triplet, where $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^d$ is a non-sensitive feature vector, $Y \in \mathcal{Y} \subset \mathbb{R}$ is the target variable, and $S \in \mathcal{S} = [M]$ is a discrete sensitive attribute where $[M] := \{1, \ldots, M\}$. We define $p_s = \mathbb{P}(S = s)$ for all $s \in [M]$. Additional notations are provided in Appendix A. Our goal is to find a predictor $f : \mathcal{X} \times \mathcal{S} \to \mathcal{Y}$ from a set $\mathcal{F}$ that balances predictive utility with fairness. We denote by ν_f the distribution of $f(\boldsymbol{X}, S)$, and by $\nu_{f|s}$ its distribution given $S = s$. We make the following standard assumption.

Assumption 1. *For $f \in \mathcal{F}$, measures $(\nu_{f|s})_{s \in [M]}$ are non atomic with finite second moments.*

We evaluate any predictor f along three key and potentially competing dimensions: predictive risk, fairness, and goodness-of-fit. Each is formally defined below.

2.1 Measuring Risk

We measure the predictive performance of a predictor using the classical quadratic risk, defined as:

$$\mathcal{R}(f) = \mathbb{E}\left[(f(\boldsymbol{X}, S) - Y)^2\right].$$

This risk is uniquely minimized by the Bayes optimal predictor $f^*(\boldsymbol{X}, S) = \mathbb{E}[Y \mid \boldsymbol{X}, S]$, recognizing that fairness constraints entail a trade-off with this optimal benchmark.

2.2 Measuring Unfairness

Our work is grounded in the concept of Demographic Parity, which exists in both a weak and a strong form. In particular, a predictor f satisfies *Weak* DP if its expectation is independent of the sensitive attribute. That is,

$$\mathbb{E}[f(\boldsymbol{X}, S) \mid S = s] = \mathbb{E}[f(\boldsymbol{X}, S)], \quad \text{for all } s \in [M],$$

ensuring fairness at the level of the first moment (the mean).

Definition 2 ((Strong) Demographic Parity). *A predictor f satisfies Strong DP if its entire output distribution is independent of the sensitive attribute. That is,*

$$\nu_{f|s} = \nu_f \quad \text{for all } s \in [M].$$

This is a much stricter criterion, requiring equivalence of all statistical moments.

Unfairness Measure We quantify unfairness through the lens of Strong DP, using Wasserstein-2 ($\mathcal{W}_2$) to measure distributional dissimilarities. For further details, we refer the reader to (Santambrogio 2015). Specifically, the unfairness of a predictor f is defined as the weighted sum of $\mathcal{W}_2$ distance between the group-conditional distributions $(\nu_{f|s})_{s \in [M]}$ and their common barycenter:

$$\mathcal{U}(f) = \min_{\nu \in \mathcal{P}_2(\mathbb{R})} \sum_{s=1}^{M} p_s \mathcal{W}_2^2(\nu_{f|s}, \nu) . \tag{1}$$

A predictor f is said to be exactly fair, that is, $\mathcal{U}(f) = 0$ *iff* the predictor satisfies Strong DP. Thus, it provides a measure of how far a model is from achieving exact fairness.

2.3 Measuring Goodness-of-fit

Evaluating fair regression models requires more than assessing overall risk and unfairness. A key consideration is the group-conditional adequacy of the model. The classical coefficient of determination defined as $R^2(f) = \mathrm{Var}(f(\boldsymbol{X}, S))/\mathrm{Var}(Y)$ is a standard metric for explained variance, particularly in linear settings. While it provides a familiar baseline, R^2 can obscure performance disparities and fails to capture group-specific *goodness-of-fit*. For example, a linear model may approximate one group well but fit another poorly, a limitation not revealed by R^2.

Group-Weighted Coefficient of Determination (GWR^2). To diagnose this critical issue, we use the *Group-Weighted R^2* (GWR^2). This metric is the average of the R^2 computed independently within each sensitive group, providing a direct measure of how well a model fits the data, on average, for all populations under consideration. For a predictor f, the definition is:

$$GWR^2(f) := \sum_{s \in \mathcal{S}} p_s R_s^2(f) \ ,$$

where,

$$R_s^2 = 1 - \frac{\mathrm{Var}(Y - f(\boldsymbol{X}, s) \mid S = s)}{\mathrm{Var}(Y \mid S = s)} \ ,$$

The strength of this metric is theoretically grounded in our analysis of the gap between GWR^2 and the global R^2 (we refer to Appendix E for further details). Divergence between these two metrics indicates model failure to capture group-specific structures. Thus, GWR^2 is a necessary diagnostic to signal structural mismatch that global metrics can obscure.

3 Limitations of Existing Fair Linear Models

The existing literature on fair linear regression provides foundational solutions but often relies on simplifying assumptions about the data-generating process. We review two key works that represent the progression from handling direct bias to incorporating some forms of indirect bias.

Mitigating Direct Bias. (Chzhen and Schreuder 2022) consider a hypothesis where unfairness arises solely from a group-dependent intercept term:

$$Y = \langle \boldsymbol{X}, \boldsymbol{\beta}_{CS22} \rangle + \beta_{0,CS22}^{(s)} + \zeta, \quad \text{where } \zeta \sim \mathcal{N}(0,1), \ (2)$$

with the key assumption that features are independent of the sensitive group, i.e., $\boldsymbol{X} \perp\!\!\!\perp S$. In this setting, the associated Bayes optimal predictor is $\langle \boldsymbol{X}, \boldsymbol{\beta}_{CS22} \rangle + \beta_{0,CS22}^{(s)}$. The independence assumption eliminates all sources of indirect bias by construction, isolating direct bias as the only source of unfairness. Therefore, achieving fairness is straightforward.

Lemma 3 (Adapted from (Chzhen and Schreuder 2022)). *Given the equation in Eq. (2), the optimal DP-fair predictor is obtained by averaging out the group-specific intercepts:*

$$f_{CS22}(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \boldsymbol{\beta}_{CS22} \rangle + \sum_{s \in [M]} p_s \beta_{0,CS22}^{(s)}.$$

Mitigating Indirect Mean Bias. (Fukuchi and Sakuma 2023) relax the feature independence assumption, allowing for group-dependent feature means and slopes:

$$Y = \langle \boldsymbol{X}, \boldsymbol{\beta}_{FS23}^{(S)} \rangle + \zeta, \quad \text{where } \zeta \sim \mathcal{N}(0,1), \quad (3)$$

where $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}^{(s)}, \sigma_X^2 I)$. This structure introduces an indirect bias that results from the differing feature means $\boldsymbol{\mu}^{(s)}$. However, it maintains a restrictive assumption of homoscedastic, uncorrelated features across groups.

Lemma 4 (Adapted from (Fukuchi and Sakuma 2023), Lemma 1). *Given the model in Eq. (3), the optimal DP-fair predictor is:*

$$f_{FS23}(\boldsymbol{x}, s) = \|\boldsymbol{\beta}_{FS23}^{(.)}\| \langle \tilde{\boldsymbol{\beta}}_{FS23}^{(s)}, \boldsymbol{x} - \boldsymbol{\mu}^{(s)} \rangle \\ + \sum_{s' \in [M]} p_{s'} \langle \boldsymbol{\beta}_{FS23}^{(s')}, \boldsymbol{\mu}^{(s')} \rangle,$$

with

$$\|\boldsymbol{\beta}_{FS23}^{(.)}\| = \sum_{s \in [M]} p_s \|\boldsymbol{\beta}_{FS23}^{(s)}\| \quad and \quad \tilde{\boldsymbol{\beta}}_{FS23}^{(s)} = \frac{\boldsymbol{\beta}_{FS23}^{(s)}}{\|\boldsymbol{\beta}_{FS23}^{(s)}\|}.$$

Limitations of Prior Work While these works represent important progress, they rely on restrictive assumptions about the data covariance structure. In particular, they do not address heteroscedasticity, where the feature covariance matrix $\Sigma^{(s)}$ varies across groups. As a result, it overlooks *indirect structural bias* from distributional disparities, highlighting the need for a more general approach.

4 A General Framework for Optimal Fair Regression

We introduce a linear model framework that captures all key sources of bias, enabling us to derive the optimal fair predictor for more complex, group-dependent data structures.

4.1 The General Model

We consider a setting where the outcome Y is generated by:

$$Y = \langle \boldsymbol{X}, \boldsymbol{\beta}^* \rangle + \gamma^* S + \beta_0^* + \zeta, \quad (4)$$

where the features $\boldsymbol{X} \mid S = s \sim \mathcal{N}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)})$ are group-dependent, and the noise $\zeta \sim \mathcal{N}(0,1)$ is independent of S and $\boldsymbol{X}$. This model captures direct bias (γ^*), indirect mean bias ($\boldsymbol{\mu}^{(s)}$), and indirect structural bias ($\boldsymbol{\Sigma}^{(s)}$).

Our goal is to find the optimal predictor within the class of linear models, $\mathcal{F}_{\text{linear}}$, that minimizes the quadratic risk $\mathcal{R}$ subject to Strong DP. Given $(\boldsymbol{x}, s) \in \mathcal{X} \times \mathcal{S}$, the Bayes optimal predictor is $f^*(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \boldsymbol{\beta}^* \rangle + \gamma^* s + \beta_0^*$.

4.2 The Optimal Risk-Fairness Trade-off

We seek to find the predictor that optimally navigates the trade-off between minimizing risk and ensuring fairness. To formalize this, we adopt the ε-*Relative Fairness Improvement* (ε-RI) constraint from (Chzhen and Schreuder 2022). A predictor f_ε satisfies this constraint if its unfairness is bounded by an ε-fraction of the Bayes-optimal predictor:

$$\mathcal{U}(f_\varepsilon) \leq \varepsilon \cdot \mathcal{U}(f^*) \ .$$

A key result, applicable to our framework, is that the predictor achieving the optimal risk-fairness trade-off under this constraint, *i.e.*, verifying $f_\varepsilon^* \in \arg\min\{\mathcal{R}(f) : \mathcal{U}(f) \leq \varepsilon \cdot \mathcal{U}(f^*)\}$, is a linear interpolation of the Bayes predictor f^* and the optimal fair predictor f_{DP}^*:

$$f_\varepsilon^* = (1 - \sqrt{\varepsilon})f_{DP}^* + \sqrt{\varepsilon}f^* \ .$$

Our main result is to derive the explicit closed-form expression for f_ε^* within our Gaussian linear model framework.

4.3 Characterizing the Optimal Fair Predictor

To state our main result, we first define the group-conditional mean and standard deviation of the Bayes optimal score:

- Group-conditional mean:

$$\mu_{f^*}^{(s)} := \mathbb{E}[f^*(\boldsymbol{X}, S) \mid S = s] = \langle \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^* \rangle + \gamma^* s + \beta_0^*.$$

- Group-conditional variance:

$$(\sigma_{f^*}^{(s)})^2 := \mathrm{Var}(f^*(\boldsymbol{X}, S) \mid S = s) = (\boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma}^{(s)} \boldsymbol{\beta}^*.$$

We also define their population-level averages, weighted by the group prior probabilities p_s:

$$\bar{\mu}_{f^*} = \sum_{s' \in [M]} p_{s'} \mu_{f^*}^{(s')} \quad \text{and} \quad \bar{\sigma}_{f^*} = \sum_{s' \in [M]} p_{s'} \sigma_{f^*}^{(s')} \ .$$

Proposition 5 (Optimal ε-Fair Predictor). *For the model in Eq. (4), the unique predictor f_ε^* that satisfies the ε-RI constraint and minimizes the quadratic risk is given by:*

$$f_\varepsilon^*(\boldsymbol{x}, s) = \sigma_\varepsilon^{(s)} \left(\frac{\langle \boldsymbol{x} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^* \rangle}{\sigma_{f^*}^{(s)}} \right) + \mu_\varepsilon^{(s)} \ , \qquad (5)$$

where the mean and std are convex combinations of the group-specific and population-averaged statistics:

$$\mu_\varepsilon^{(s)} = (1 - \sqrt{\varepsilon})\bar{\mu}_{f^*} + \sqrt{\varepsilon}\mu_{f^*}^{(s)}$$
$$\sigma_\varepsilon^{(s)} = (1 - \sqrt{\varepsilon})\bar{\sigma}_{f^*} + \sqrt{\varepsilon}\sigma_{f^*}^{(s)} \ .$$

The optimal exactly-fair predictor f_{DP}^* is recovered at $\varepsilon = 0$, and the Bayes optimal predictor f^* is recovered at $\varepsilon = 1$.

4.4 Interpreting the Fairness Mechanism

The structure of f_ε^* reveals a clear and tunable mechanism for enforcing fairness, which can be understood from two complementary perspectives.

Perspective 1: Tunable Standardization and Averaging. This perspective views fairness as the controlled shift of group-dependent moments toward global average moments.

1. **Group-wise Standardization:** within each group s, the term $\langle \boldsymbol{x} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^* \rangle / \sigma_{f^*}^{(s)}$ creates a standardized score (zero mean and unit variance). This procedure simultaneously removes indirect mean and structural biases.

2. **Controlled Re-scaling and Shifting:** This standardized score is then re-scaled by $\sigma_\varepsilon^{(s)}$ and shifted by $\mu_\varepsilon^{(s)}$. These coefficients are a direct interpolation between the group-specific moments $(\mu_{f^*}^{(s)}, \sigma_{f^*}^{(s)})$ and the global averages $(\bar{\mu}_{f^*}, \bar{\sigma}_{f^*})$. The parameter ε directly control this trade-off: at $\varepsilon = 0$, the predictor uses only global averages, eliminating all bias; at $\varepsilon = 1$, it uses only group-specific values, retaining all original bias for maximum accuracy.

Perspective 2: A Group-Conditional Fair Model. Alternatively, we can express the predictor as a linear model,

$$f_\varepsilon^*(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \boldsymbol{\beta}_\varepsilon^{(s)} \rangle + \beta_{0,\varepsilon}^{(s)} \ ,$$

to see how fairness is encoded into the parameters of the model. By rearranging the terms from Proposition 5, we find the effective slope and intercept for each group are:

$$\boldsymbol{\beta}_\varepsilon^{(s)} = \left(\frac{\sigma_\varepsilon^{(s)}}{\sigma_{f^*}^{(s)}} \right) \boldsymbol{\beta}^* \quad \text{and} \quad \beta_{0,\varepsilon}^{(s)} = \mu_\varepsilon^{(s)} - \left(\frac{\sigma_\varepsilon^{(s)}}{\sigma_{f^*}^{(s)}} \right) \langle \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^* \rangle \ .$$

This view highlights that fairness is achieved by constructing a group-aware model with parameters systematically adjusted to counteract group-specific biases. The scaling factor $\sigma_\varepsilon^{(s)}/\sigma_{f^*}^{(s)}$ compensates for the structural bias, while the intercept $\beta_{0,\varepsilon}^{(s)}$ corrects for the mean-based biases.

5 Decomposition of direct and indirect biases through the unfairness

In this section, we develop a comprehensive framework for understanding unfairness in linear regression.

5.1 Prediction-level Decomposition of Unfairness

We begin by decomposing our unfairness measure $\mathcal{U}(f)$ for any predictor within the class of linear models, $\mathcal{F}_{\text{linear}}$.

Proposition 6 (Linear Model Bias Decomposition). *For any predictor $f \in \mathcal{F}_{\text{linear}}$ with coefficients $(\boldsymbol{\beta}, \gamma, \beta_0)$, its total unfairness $\mathcal{U}(f)$ decomposes into First-Moment Disparity (FMD) and Second-Moment Disparity (SMD):*

$$\mathcal{U}(f) = \underbrace{\mathrm{Var}(\mathbb{E}[f|S])}_{\text{FMD}} + \underbrace{\mathrm{Var}(\sqrt{\mathrm{Var}(f|S)})}_{\text{SMD}}. \qquad (6)$$

These components further decompose into four bias sources:

$$\mathcal{U}(f) = \underbrace{\gamma^2 \mathrm{Var}(S)}_{\text{Direct Mean}} + \underbrace{\mathrm{Var}(\langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle)}_{\text{Indirect Mean}} + \underbrace{2\gamma \mathrm{Cov}(S, \langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle)}_{\text{Interaction}}$$
$$+ \underbrace{\mathrm{Var}\left(\sqrt{\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{(s)} \boldsymbol{\beta}} \right)}_{\text{Indirect Structural}}. \qquad (7)$$

This decomposition formalizes the conditions required to achieve Strong DP, showing that fairness in this stronger sense necessitates mitigating bias at two distinct levels:

- The **First-Moment Disparity** $\mathrm{Var}(\mathbb{E}[f \mid S])$ captures unfairness in average predictions. It arises from direct dependence on the sensitive attribute (Direct Mean Bias, related to Weak DP) or from correlations between group membership and feature means (Indirect Mean Bias).

- The **Second-Moment Disparity** $\mathrm{Var}(\sqrt{\mathrm{Var}(f \mid S)})$ captures a more subtle form of unfairness (Indirect Structural Bias) where predictive certainty differs across groups due to variations in feature covariance $\boldsymbol{\Sigma}^{(s)}$.

This decomposition reveals that a model can satisfy Weak DP (without FMD) while remaining unfair under Strong DP. The following corollary demonstrates a key advantage of our optimal ε-fair predictor:

Corollary 7 (Residual Unfairness of our method). *The total unfairness of our predictor f_ε^*, (see Prop. 5), is exactly:*

$$\mathcal{U}(f_\varepsilon^*) = \varepsilon \cdot \mathrm{Var}(\mathbb{E}[f^* \mid S]) + \varepsilon \cdot \mathrm{Var}(\sqrt{\mathrm{Var}(f^* \mid S)}) \ .$$

This corollary highlights a direct, analytical link between a single control parameter (ε) and the total amount of multi-source unfairness, a property not available in prior models.

5.2 Feature-level Decomposition of Unfairness via Approximation

While the prediction-level decomposition quantifies total unfairness, practical intervention requires attributing this unfairness to individual features. A fully additive decomposition is challenging due to the nonlinearity introduced by the square root in the structural bias. To enable interpretability, we apply a first-order Taylor expansion to linearize this term, yielding a tractable and accurate additive approximation.

The Additive Case: Uncorrelated Features. We consider a simplified setting where features are mutually uncorrelated within each group ($\Sigma^{(s)}$ are diagonal matrices). In this case, the total indirect unfairness of any linear model decomposes into a sum of marginal contributions from each feature.

Proposition 8 (Additive Feature-Level Decomposition). *Given $f \in \mathcal{F}_{linear}$ with coefficients $(\boldsymbol{\beta}, \gamma)$, let its indirect unfairness be $\mathcal{U}_{\mathrm{indirect}}(f) = \mathcal{U}(f) - \gamma^2 \mathrm{Var}(S)$. If all $\Sigma^{(s)}$ are diagonal, then this unfairness can be approximated by an additive sum:*

$$\mathcal{U}_{\mathrm{indirect}}(f) \approx \sum_{j=1}^{d} \mathcal{U}_j^{approx}(f),$$

with the approximate main contribution from feature X_j is:

$$\mathcal{U}_j^{approx}(f) = \underbrace{(\beta_j)^2 \mathrm{Var}(\mu_j^{(S)})}_{Mean} + \frac{1}{4\bar{V}} \underbrace{(\beta_j)^4 \mathrm{Var}((\sigma_j^{(S)})^2)}_{Structural} + \underbrace{2\gamma \beta_j \mathrm{Cov}(S, \mu_j^{(S)})}_{Interaction},$$

where $\mu_j^{(s)} = \mathbb{E}[X_j | S = s]$ and $(\sigma_j^{(s)})^2 = \mathrm{Var}(X_j | S = s)$. Here, $\bar{V} = \mathbb{E}[\mathrm{Var}(f|S)]$ is the average conditional score variance.

This proposition attributes model unfairness to individual features via three pathways: (1) mean disparity, (2) variance disparity (structural bias), and (3) interaction with direct bias. The term $1/(4\bar{V})$ indicates that structural bias diminishes as predictive variance increases.

The General Case: Interactional Unfairness. When features are correlated, the decomposition becomes more complex due to cross-terms capturing *interactional unfairness* (see Appendix). This includes: (1) the compounding of mean biases through correlated feature means, and (2) a deeper structural effect, which we term *Covariance Disparity*, driven by group-level differences in feature correlations.

This analysis provides both practical and comprehensive insight. The additive decomposition highlights features with primary unfairness, while the general case reveals how feature correlations amplify or mitigate these effects.

6 Practical Implementation and Estimation

To apply our framework in practice, the optimal fair predictor must be estimated from finite data, since the population parameters $(\boldsymbol{\beta}^*, \gamma^*, \boldsymbol{\mu}^{(s)}, \Sigma^{(s)})$ are unknown.

The Plug-in Estimator The plug-in estimator $\hat{f}_\varepsilon$ of f_ε^* is constructed by replacing all quantities in Prop. 5 with their empirical estimates.

1. **Estimate Model Parameters**. We estimate the base model parameters $(\hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\beta}_0)$. Our framework is agnostic to the fitting procedure; any standard method, such as OLS or penalized version (Ridge, Lasso), is applicable.

2. **Estimating Group Statistics**. For each s, we compute the standard estimates for the group proportions $\hat{p}_s$, feature means $\hat{\boldsymbol{\mu}}^{(s)}$, and feature covariance matrices $\hat{\Sigma}^{(s)}$.

3. **Assemble the Fair Predictor.** Finally, these empirical components are used to construct the plug-in versions of the conditional score moments $(\hat{\mu}_f^{(s)}, \hat{\sigma}_f^{(s)})$ and their population averages $(\hat{\bar{\mu}}_f, \hat{\bar{\sigma}}_f)$. These are then combined according to Prop. 5 to form the final estimator.

Evaluation Metrics We evaluate all models on a held-out test set using empirical estimators of our three key metrics. For both the Risk and GWR^2, we consider their empirical counterparts, denoted $\hat{\mathcal{R}}$ (mean squared error) and $\widehat{GWR^2}$, respectively, where:

$$\widehat{GWR^2}(f) = \sum_{s \in [M]} \hat{p}_s \left(1 - \frac{\widehat{\mathrm{Var}}(Y - f \mid S = s)}{\widehat{\mathrm{Var}}(Y \mid S = s)} \right).$$

We quantify the unfairness using the Kolmogorov-Smirnov (KS) test, as it is model-agnostic and does not rely on structural assumptions.

$$\hat{\mathcal{U}}_{\mathrm{KS}}(f) = \max_{s_j, s_k \in [M]} D_{\mathrm{KS}}(\hat{F}_{f|s_j}, \hat{F}_{f|s_k}).$$

Here, $\hat{F}_{f|s}$ is the empirical CDF of scores for group s.

7 Numerical Experiments

We run experiments on synthetic and real-world data to: (1) validate our bias decomposition framework and (2) demonstrate that our tunable predictor, $\hat{f}_\varepsilon$, effectively traces the optimal risk-fairness frontier, outperforming prior methods.

7.1 Application on Synthetic Data

We generated synthetic triplets $(\boldsymbol{X}, S, Y)$ where we can precisely control each source of bias. The sensitive attribute $S \in \{1, 2\}$ is drawn from a Bernoulli distribution. The data-generating process is governed by four control parameters $T := (T_y, T_{\mathrm{mean}}, T_{\mathrm{std}}, T_{\mathrm{corr}})$ that map directly to our bias decomposition: T_y sets the **direct bias** coefficient γ^*; T_{mean} introduces **indirect mean bias** by creating differences between feature means $\boldsymbol{\mu}^{(s)}$; and T_{std} and T_{corr} introduce **indirect structural bias** by creating group-specific differences in the variances and correlations within the covariance matrix $\Sigma^{(s)}$. When a parameter is set to zero, the corresponding source of bias is absent. We provide full implementation details on the simulation of $(\boldsymbol{X}, S, Y)$ in the Appendix F.1.
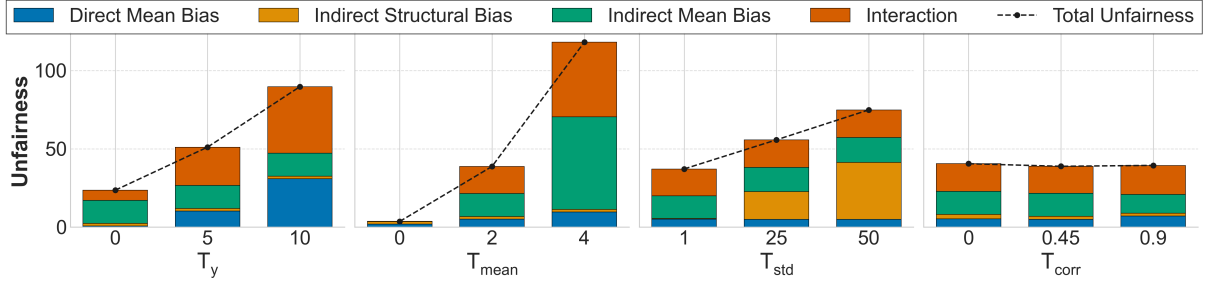
Figure 2: Bias decomposition (see Prop. 6) of the linear model on synthetic data using by default $T = (3, 2, 3, 0.7)$.
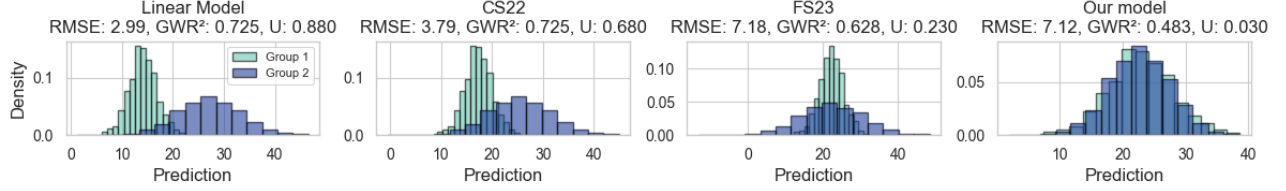

Figure 3: Comparison of group-conditioned model output distribution on synthetic data using $T = (10, 2, 2, 0.7)$.
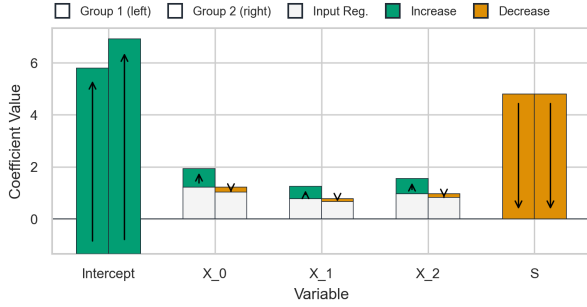

Figure 4: Coefficients adjustments for fairness, shown for a sample of features on synthetic data with $T = (3, 2, 3, .7)$.
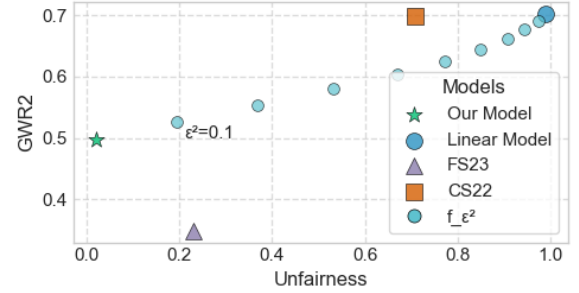

Figure 5: Analysis of Approximate fairness model on synthetic data with $T = (10, 2, 3, 0.7)$.

Experimentation scheme Default parameters are set as follows : $d = 5$, $\tau = 0.6$. Given the vector $T = (T_y, T_{\text{mean}}, T_{\text{std}}, T_{\text{corr}})$, we create synthetic dataset of $n = 20,000$ samples and split it into training (50%), testing (25%), and unlabeled (25%) subsets. As a base model, we chose the linear regression of Y on X and S, using standard parameters from `scikit-learn` in Python. The coefficients of this linear regression are used as an input to build of fair linear model.

Validating the Bias Decomposition Our framework provides a direct way to diagnose the sources of unfairness. Fig. 2 applies our decomposition to a linear model trained on synthetic data. The results empirically validate our theory: increasing the direct bias parameter (T_y) primarily inflates the Direct Mean and Interaction terms, while increasing the indirect parameters ($T_{\text{mean}}, T_{\text{std}}$) maps clearly to the Indirect Mean and Indirect Structural bias components, respectively. This confirms that our decomposition is a practical tool for identifying the root causes of unfairness in linear models.

Fairness Mitigation and Robustness to Bias Shifts. In complex scenarios with full bias interactions $(3, 2, 2, 0.7)$, our model uniquely preserves remediation capabilities (Fig. 3). The remediation operates through three visible mechanisms (Fig. 4): (1) direct bias elimination via sensitive attribute coefficient nullification, (2) structural bias remediation through feature coefficient adjustments, and (3) intercept compensation for all bias corrections. We refer to Appendix F for additional analyses.

We also test robustness under bias shifts by varying $T_y, T_{\text{mean}}, T_{\text{corr}}, T_{\text{std}}$. As shown in Fig. 6, our method and CS22 remain stable in both performance and fairness as direct bias (T_y) increases, while FS23 deteriorates.

Tracing the Optimal Risk–Fairness Frontier. Under the ε-RI constraint, the parameter ε provides continuous control over the desired fairness level. Fig. 5 shows that our method consistently dominates in this bias scenario: it either achieves higher accuracy than the baselines at a given

Model	CRIME			LAW			GOSSIS		
	GWR^2	RMSE	Unfairness	GWR^2	RMSE	Unfairness	GWR^2	RMSE	Unfairness
Base Model Unaware	$.45 \pm .05$	0.15 ± 0.01	0.55 ± 0.04	$.15 \pm .01$	$0.37 \pm .00$	$.13 \pm .01$	$.69 \pm .01$	10.3 ± 0.1	$.14 \pm .01$
Base Model	$.46 \pm .05$	0.15 ± 0.01	0.61 ± 0.04	$.15 \pm .01$	$0.37 \pm .00$	$.43 \pm .02$	$.69 \pm .01$	10.3 ± 0.1	$.15 \pm .01$
CS22	$.46 \pm .05$	0.15 ± 0.01	0.54 ± 0.04	$.15 \pm .01$	$0.37 \pm .00$	$.08 \pm .01$	$.69 \pm .01$	10.3 ± 0.1	$.14 \pm .01$
FS23	$.35 \pm .09$	0.19 ± 0.01	0.20 ± 0.05	$.08 \pm .05$	$0.39 \pm .01$	$.15 \pm .05$	$.51 \pm .40$	12.5 ± 3.3	$.13 \pm .07$
Our model	$.38 \pm .07$	0.19 ± 0.01	**0.12 ± 0.04**	$.15 \pm .01$	$0.37 \pm .00$	**$.07 \pm .02$**	$.69 \pm .01$	10.4 ± 0.1	**$.03 \pm .01$**

Table 2: Comparison of model performances across all datasets. Results are presented as mean $\pm$ standard deviation over 50 runs. Bold cells indicate the lowest unfairness.
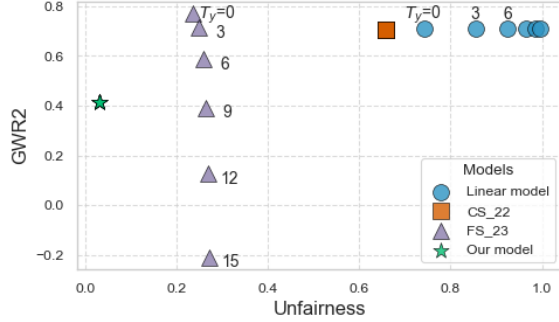

Figure 6: Analysis of Model performance *w.r.t.* direct bias shifts (T_y) on synthetic data using $T = (*, 2, 2, 0.7)$.

Figure 7: Analysis of coefficient shifts from the linear model to our fair model on the CRIME dataset.

unfairness level, or ensures lower unfairness at a fixed accuracy, effectively tracing the optimal risk-fairness trade-off.

7.2 Results on Real-World Data

We use three standard and diverse fairness benchmarks.

GOSSIS (Raffa et al. 2022): contains medical information from over 130,000 patients admitted to intensive care units. The task consists in predicting the vital variable *h1_diaspb_max* using ethnicity as a protected attribute.

CRIME (Redmond and Baveja 2002): includes demographic, economic and crime data about US communities with 1994 samples. We predict the number of violent crimes per 10^5 population. As (Calders et al. 2013), we constructed a sensitive attribute based on Black population percentages.

LAW: corresponds to the Law School Admissions Councils National Longitudinal Bar Passage Study. The regression task consists in predicting students' GPA (normalized to the range $[0, 1]$) using race as the protected attribute.

Comparison w.r.t state-of-the-art. Experimental results (Table 2) demonstrate that our model effectively reduces unfairness while maintaining acceptable predictive performance. The Linear Model Unaware confirms that omitting the sensitive attribute fails to eliminate discriminatory patterns inherent in the data and does not prevent discriminatory outcomes. On the LAW dataset, unfairness is notably reduced between the Linear Model and the Linear Model Unaware, which explains the competitive performance of CS22 in this case, as this method is effective at mitigating direct bias; nevertheless, our model achieves even lower unfairness. Compared to the best-performing baselines, our approach achieves substantial unfairness reductions of 43% on
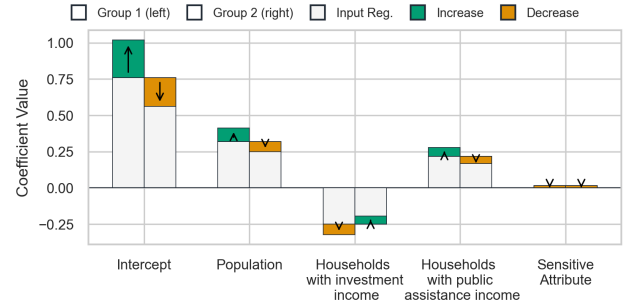
CRIME, 13% on LAW, and 74% on GOSSIS datasets. This improvement comes with a performance trade-off, showing approximately 15-20% reduction in GWR^2 on CRIME while preserving competitive accuracy elsewhere.

Feature-level interpretation on CRIME Dataset While the direct bias is nullified (Fig. 7), the model mitigates indirect biases through group-specific coefficient adjustments.

Conclusion

We propose a closed-form solution for fair linear regression that enables exact control over the risk-fairness trade-off via the optimal predictor f_ε^*. Building upon this Gaussian framework, we introduce a novel decomposition of unfairness into direct and indirect components, highlighting four distinct sources, including the previously overlooked **Indirect Structural Bias** arising from disparities in predictive variance.

Our results demonstrate that mean-based fairness alone is insufficient. By explicitly accounting for structural disparities, our method ensures fairness in both average predictions and predictive certainty across groups. The decomposition, along with the Group-Weighted R^2, provides actionable tools for diagnosing unfairness in linear models. While grounded in Gaussian assumptions, our approach shows strong empirical robustness on real-world data. Future work may extend these insights to non-linear models and broader fairness notions.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classifi-

cation. In *International conference on machine learning*, 60–69. PMLR.

Agueh, M.; and Carlier, G. 2011. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2): 904–924.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.

Calders, T.; Karim, A.; Kamiran, F.; Ali, W.; and Zhang, X. 2013. Controlling attribute effect in linear regression. In *2013 IEEE 13th international conference on data mining*, 71–80. IEEE.

Chzhen, E.; Denis, C.; Hebiri, M.; Oneto, L.; and Pontil, M. 2020. Fair Regression with Wasserstein Barycenters. In *Advances in Neural Information Processing Systems*.

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

Del Barrio, E.; Gordaliza, P.; and Loubes, J.-M. 2020. Review of mathematical frameworks for fairness in machine learning. *arXiv preprint arXiv:2005.13755*.

Fréchet, M. 1957. Sur la distance de deux lois de probabilité. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*, 244: 689–692.

Fukuchi, K.; and Sakuma, J. 2023. Demographic parity constrained minimax optimal regression under linear model. *Advances in Neural Information Processing Systems*, 36: 8653–8689.

Gaucher, S.; Schreuder, N.; and Chzhen, E. 2023. Fair learning with Wasserstein barycenters for non-decomposable performance measures. In *International Conference on Artificial Intelligence and Statistics*, 2436–2459. PMLR.

Hajian, S.; and Domingo-Ferrer, J. 2012. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, 25(7): 1445–1459.

Hastie, T.; Tibshirani, R.; Friedman, J. H.; and Friedman, J. H. 2009. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hu, F.; Ratz, P.; and Charpentier, A. 2024. A sequentially fair mechanism for multiple sensitive attributes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12502–12510.

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

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Nabi, R.; and Shpitser, I. 2018. Fair inference on outcomes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Pessach, D.; and Shmueli, E. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3): 1–44.

Raffa, J. D.; Johnson, A. E. W.; O'Brien, Z.; Pollard, T. J.; Mark, R. G.; Celi, L. A.; Pilcher, D.; and Badawi, O. 2022. The Global Open Source Severity of Illness Score (GOSSIS). *Critical Care Medicine*, 50(7): 1040–1050.

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

Santambrogio, F. 2015. *Optimal transport for applied mathematicians*. Springer.

Tang, Z.; Zhang, J.; and Zhang, K. 2023. What-is and how-to for fairness in machine learning: A survey, reflection, and perspective. *ACM Computing Surveys*, 55(13s): 1–37.

A Notations and Background

This section introduces the mathematical notations and fundamental concepts that serve as a foundation for the proofs and derivations that follow.

A.1 Notation

Let f be a function and $(\boldsymbol{X}, S) \in \mathcal{X} \times \mathcal{S} \subset \mathbb{R}^d \times \mathbb{N}$ a random pair, where d is a positive integer and S denotes a discrete sensitive attribute. Let $\mathcal{V}$ be the space of probability measures on a target space $\mathcal{Y} \subset \mathbb{R}$. We denote by $\nu_f \in \mathcal{V}$ the distribution of $f(\boldsymbol{X}, S)$, and by $\nu_{f|s} \in \mathcal{V}$ its conditional distribution given $S = s$. Let $F_{f|s}(u) := \mathbb{P}(f(\boldsymbol{X}, S) \leq u \mid S = s)$ be the cumulative distribution function (CDF) of $\nu_{f|s}$, and define the corresponding quantile function as $Q_{f|s}(v) := \inf \{u \in \mathbb{R} : F_{f|s}(u) \geq v\}$.

A.2 Background on Wasserstein distance and barycenter

This section reminds the concepts of Wasserstein distance and barycenter from one-dimensional optimal transport theory (Santambrogio 2015).

Definition 9 (Wasserstein $\mathcal{W}_2$ Distance, see (Santambrogio 2015)). *For probability measures ν_0, ν_1 on $\mathbb{R}$, the squared 2–Wasserstein distance is*

$$\mathcal{W}_2(\nu_0, \nu_1)^2 = \int_0^1 \left| Q_0(u) - Q_1(u) \right|^2 \mathrm{d}u.$$

Let $\{\nu_s : s \in \mathcal{S}\}$ be a finite family of measures on $\mathbb{R}$ with associated weights $\{p_s : s \in \mathcal{S}\}$.

Definition 10 (Wasserstein Barycenter). *The Wasserstein barycenter ν^* is the unique minimizer*

$$\nu^* = \arg \min_{\nu \in \mathcal{P}_2(\mathbb{R})} \sum_{s \in \mathcal{S}} p_s \, \mathcal{W}_2(\nu_s, \nu)^2,$$

where $\mathcal{P}_2(\mathbb{R})$ denotes the set of probability measures on $\mathbb{R}$ with finite second moment.

B Proof of Proposition 5

This section provides a detailed proof for the closed-form expression of the optimal ε-DP predictor, f_ε^*, as stated in Proposition 5. Our derivation relies on the following foundational result concerning the characterization of the optimal exactly fair predictor, denoted f_{DP}^*.

Theorem 11 ((Chzhen et al. 2020; Le Gouic, Loubes, and Rigollet 2020)). *Let f^* be a predictor whose conditional distributions, $\nu_{f^*|s} := \mathrm{law}(f^*|S = s)$, have densities for all $s \in [M]$. The optimal predictor f_{DP}^* that minimizes the risk $\mathcal{R}(f)$ while satisfying the Strong DP constraint is given by the composition of the average quantile function and the conditional CDF of f^*:*

$$f_{DP}^*(\boldsymbol{x}, s) = \left(\sum_{s' \in [M]} p_{s'} F_{f^*|s'}^{-1} \right) \circ F_{f^*|s}(f^*(\boldsymbol{x}, s)),$$

where $F_{f^*|s}$ is the CDF of f^* conditional on $S = s$, and $F_{f^*|s'}^{-1}$ is its corresponding quantile function (inverse CDF).

B.1 Proof of the optimality of f_{DP}^* in our setting

To apply the aforementioned theorem, we proceed in three steps: we first derive the conditional CDF of our Bayes predictor f^*, then its inverse, and finally, we assemble the expression for f_{DP}^*.

Deriving the Conditional CDF. Given our model, the Bayes predictor $f^*(\boldsymbol{X}, S)$ conditional on $S = s$ is a linear transformation of a Gaussian random variable, and is therefore itself Gaussian. Its conditional mean and variance are:

$$\mu_{f^*}^{(s)} = \mathbb{E}[f^*(\boldsymbol{X}, S) \mid S = s] = \langle \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta^*} \rangle + \gamma^* s + \beta_0^*,$$

$$(\sigma_{f^*}^{(s)})^2 = \mathrm{Var}(f^*(\boldsymbol{X}, S) \mid S = s) = (\boldsymbol{\beta^*})^\top \boldsymbol{\Sigma^{(s)}} \boldsymbol{\beta^*}.$$

The conditional CDF, $F_{f^*|s}(t) = \mathbb{P}(f^*(\boldsymbol{X}, S) \leq t | S = s)$, is found by standardizing the variable:

$$F_{f^*|s}(t) = \mathbb{P}\left(\frac{f^*(\boldsymbol{X}, S) - \mu_{f^*}^{(s)}}{\sigma_{f^*}^{(s)}} \leq \frac{t - \mu_{f^*}^{(s)}}{\sigma_{f^*}^{(s)}} \middle| S = s \right)$$

$$= \Phi\left(\frac{t - \mu_{f^*}^{(s)}}{\sigma_{f^*}^{(s)}} \right),$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution.

Deriving the Conditional Quantile Function. The quantile function, $F_{f^*|s}^{-1}(p)$, is obtained by inverting the CDF expression. For a probability $p \in (0, 1)$:

$$p = \Phi\left(\frac{F_{f^*|s}^{-1}(p) - \mu_{f^*}^{(s)}}{\sigma_{f^*}^{(s)}} \right)$$

$$\implies \Phi^{-1}(p) = \frac{F_{f^*|s}^{-1}(p) - \mu_{f^*}^{(s)}}{\sigma_{f^*}^{(s)}}$$

$$\implies F_{f^*|s}^{-1}(p) = \sigma_{f^*}^{(s)} \Phi^{-1}(p) + \mu_{f^*}^{(s)}.$$

Assembling the Optimal DP Predictor. We now substitute these forms into the expression from Theorem 11. First, let's evaluate the inner part of the composition for a given input $(\boldsymbol{x}, s)$:

$$F_{f^*|s}(f^*(\boldsymbol{x}, s)) = \Phi\left(\frac{f^*(\boldsymbol{x}, s) - \mu_{f^*}^{(s)}}{\sigma_{f^*}^{(s)}} \right).$$

Now, we apply the averaged quantile function to this result. Let $p = F_{f^*|s}(f^*(\boldsymbol{x}, s))$. The composition is:

$$f_{DP}^*(\boldsymbol{x}, s) = \sum_{s' \in [M]} p_{s'} F_{f^*|s'}^{-1}(p)$$

$$= \sum_{s' \in [M]} p_{s'} \left(\sigma_{f^*}^{(s')} \Phi^{-1}(p) + \mu_{f^*}^{(s')} \right)$$

$$= \left(\sum_{s' \in [M]} p_{s'} \sigma_{f^*}^{(s')} \right) \Phi^{-1}(p) + \left(\sum_{s' \in [M]} p_{s'} \mu_{f^*}^{(s')} \right).$$

Recognizing the definitions of the population-averaged moments, $\bar{\sigma}_{f^*} = \sum p_{s'}\sigma_{f^*}^{(s')}$ and $\bar{\mu}_{f^*} = \sum p_{s'}\mu_{f^*}^{(s')}$, we have:

$$f_{DP}^*(\boldsymbol{x}, s) = \bar{\sigma}_{f^*}\Phi^{-1}(p) + \bar{\mu}_{f^*}.$$

Finally, we substitute back the expression for p, where the Φ^{-1} and Φ functions cancel:

$$f_{DP}^*(\boldsymbol{x}, s) = \bar{\sigma}_{f^*}\Phi^{-1}\left(\Phi\left(\frac{f^*(\boldsymbol{x}, s) - \mu_{f^*}^{(s)}}{\sigma_{f^*}^{(s)}}\right)\right) + \bar{\mu}_{f^*}$$

$$= \bar{\sigma}_{f^*}\left(\frac{f^*(\boldsymbol{x}, s) - \mu_{f^*}^{(s)}}{\sigma_{f^*}^{(s)}}\right) + \bar{\mu}_{f^*}.$$

Using the definitions $f^*(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \boldsymbol{\beta}^* \rangle + \gamma^* s + \beta_0^*$ and $\mu_{f^*}^{(s)} = \langle \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^* \rangle + \gamma^* s + \beta_0^*$, their difference is simply $f^*(\boldsymbol{x}, s) - \mu_{f^*}^{(s)} = \langle \boldsymbol{x} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^* \rangle$. This yields the final expression:

$$f_{DP}^*(\boldsymbol{x}, s) = \bar{\sigma}_{f^*}\left(\frac{\langle \boldsymbol{x} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^* \rangle}{\sigma_{f^*}^{(s)}}\right) + \bar{\mu}_{f^*},$$

which proves the result stated in Proposition 5 for exact fairness. In the following, we prove the optimality of our expression in Approximate fairness.

B.2 Proof of the optimality of f_ε^* in our setting

Proof Sketch. Our proof relies on the known general solution for the optimal ε-RI predictor from (Chzhen and Schreuder 2022), which is the linear interpolation described in the following proposition.

Proposition 12 (adapted from (Chzhen and Schreuder 2022)). *Assuming that $\{\nu_s^*\}_{s\in[K]}$ are non-atomic with finite second moments, then for all $\alpha \in [0, 1]$ and all $(\boldsymbol{x}, s) \in \mathbb{R}^p \times [K]$, the closed form solution of the minimization problem*

$$\arg\min\{\mathcal{R}(f) : \mathcal{U}(f) \le \varepsilon \times \mathcal{U}(f^*)\} \quad (8)$$

is given by:

$$f_\varepsilon^*(\boldsymbol{x}, s) = \sqrt{\varepsilon}f^*(\boldsymbol{x}, s)$$
$$+ (1 - \sqrt{\varepsilon})\sum_{s'=1}^{K} p_{s'}F_{\nu_{s'}^*}^{-1} \circ F_{\nu_s^*} \circ f^*(\boldsymbol{x}, s)$$
$$= \sqrt{\varepsilon}f^*(\boldsymbol{x}, s) + (1 - \sqrt{\varepsilon})f_{DP}^*(\boldsymbol{x}, s)$$

where $f_{DP}^*(\boldsymbol{x}, s)$ represents the demographic parity optimal predictor.

We will then substitute the explicit expressions for our Bayes predictor f^* and optimal DP predictor f_{DP}^* and show that the resulting expression simplifies to the form stated in the proposition 5.

Expressing Predictors in a Common Form. To simplify the algebra, we first express both f^* and f_{DP}^* in terms of a common standardized score. Let the group-standardized score for an input $\boldsymbol{x}$ be:

$$z^{(s)}(\boldsymbol{x}) = \frac{\langle \boldsymbol{x} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^* \rangle}{\sigma_{f^*}^{(s)}}.$$

By construction, the random variable $z^{(s)}(\boldsymbol{X})$ has a conditional mean of 0 and a conditional standard deviation of 1 for group s.

Using the results from the proof in B.1, we can write both predictors as linear transformations of this standardized score:

- The Bayes predictor f^* can be rewritten as:

$$f^*(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \boldsymbol{\beta}^* \rangle + \gamma^* s + \beta_0^*$$
$$= \langle \boldsymbol{x} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^* \rangle + \langle \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^* \rangle + \gamma^* s + \beta_0^*$$
$$= \sigma_{f^*}^{(s)} \cdot z^{(s)}(\boldsymbol{x}) + \mu_{f^*}^{(s)}.$$

- The optimal DP predictor f_{DP}^* from proof in B.1 is:

$$f_{DP}^*(\boldsymbol{x}, s) = \bar{\sigma}_{f^*} \cdot z^{(s)}(\boldsymbol{x}) + \bar{\mu}_{f^*}.$$

Substituting and Simplifying the Interpolation. Now we substitute these two expressions back into the interpolation formula:

$$f_\varepsilon^*(\boldsymbol{x}, s) = (1 - \sqrt{\varepsilon})\left(\bar{\sigma}_{f^*} \cdot z^{(s)}(\boldsymbol{x}) + \bar{\mu}_{f^*}\right)$$
$$+ \sqrt{\varepsilon}\left(\sigma_{f^*}^{(s)} \cdot z^{(s)}(\boldsymbol{x}) + \mu_{f^*}^{(s)}\right).$$

We can now collect the terms multiplying the standardized score $z^{(s)}(\boldsymbol{x})$ and the constant terms separately:

$$f_\varepsilon^*(\boldsymbol{x}, s) = \left[(1 - \sqrt{\varepsilon})\bar{\sigma}_{f^*} + \sqrt{\varepsilon}\sigma_{f^*}^{(s)}\right] \cdot z^{(s)}(\boldsymbol{x})$$
$$+ \left[(1 - \sqrt{\varepsilon})\bar{\mu}_{f^*} + \sqrt{\varepsilon}\mu_{f^*}^{(s)}\right].$$

Identifying the Final Form. We recognize the expressions in the brackets as the definitions of the effective standard deviation $\sigma_\varepsilon^{(s)}$ and mean $\mu_\varepsilon^{(s)}$ from the proposition statement:

$$\sigma_\varepsilon^{(s)} = (1 - \sqrt{\varepsilon})\bar{\sigma}_{f^*} + \sqrt{\varepsilon}\sigma_{f^*}^{(s)}$$
$$\mu_\varepsilon^{(s)} = (1 - \sqrt{\varepsilon})\bar{\mu}_{f^*} + \sqrt{\varepsilon}\mu_{f^*}^{(s)}.$$

Substituting these definitions back, we obtain:

$$f_\varepsilon^*(\boldsymbol{x}, s) = \sigma_\varepsilon^{(s)} \cdot z^{(s)}(\boldsymbol{x}) + \mu_\varepsilon^{(s)}.$$

Finally, substituting the definition of $z^{(s)}(\boldsymbol{x})$ gives the exact expression from the proposition:

$$f_\varepsilon^*(\boldsymbol{x}, s) = \sigma_\varepsilon^{(s)}\left(\frac{\langle \boldsymbol{x} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^* \rangle}{\sigma_{f^*}^{(s)}}\right) + \mu_\varepsilon^{(s)}.$$

This completes the proof.

C Proof of Proposition 6 and Corollary 7

This section provides a detailed proof for the characterization of the Bias decomposition. We begin by stating the foundational results from optimal transport theory that underpin our analysis.

C.1 Preliminaries for the computation of the unfairness in our framework

Our unfairness measure is based on the Wasserstein-2 distance. Its computation for Gaussian distributions relies on two key results.

First, the squared Wasserstein-2 distance between two one-dimensional Gaussian distributions has a simple closed form.

Lemma 13 (From (Fréchet 1957)). *For any means $\mu_1, \mu_2 \in \mathbb{R}$ and standard deviations $\sigma_1, \sigma_2 \in \mathbb{R}^+$, the squared $\mathcal{W}_2$ distance between the corresponding Gaussian distributions is:*

$$\mathcal{W}_2^2(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2.$$

Second, the Wasserstein barycenter of a set of Gaussian distributions is also a Gaussian, whose moments are the weighted averages of the input moments.

Lemma 14 (From (Agueh and Carlier 2011)). *Let $(\mathcal{N}(\mu_s, \sigma_s^2))_{s \in [M]}$ be a set of Gaussian distributions and let $(p_s)_{s \in [M]}$ be a probability vector. The unique Wasserstein barycenter that solves the minimization problem in Eq.* (1) *is the Gaussian distribution $\mathcal{N}(\bar{\mu}, \bar{\sigma}^2)$, where:*

$$\bar{\mu} = \sum_{s=1}^{M} p_s \mu_s \quad and \quad \bar{\sigma} = \sum_{s=1}^{M} p_s \sigma_s.$$

C.2 Proof of Proposition 6

Proof. The proof proceeds in three steps. First, we simplify the definition of $\mathcal{U}(f)$ under our Gaussian assumption. Second, we decompose the first-moment term. Third, we decompose the second-moment term.

Step 1: From Wasserstein Distance to a Variance Decomposition. The unfairness measure is defined as:

$$\mathcal{U}(f) = \min_{\nu \in \mathcal{P}_2(\mathbb{R})} \sum_{s=1}^{M} p_s W_2^2(\nu_{f|s}, \nu).$$

Under our modeling assumption, the conditional distribution $\nu_{f|s}$ is Gaussian with mean $\mu_f^{(s)} = \mathbb{E}[f|S = s]$ and standard deviation $\sigma_f^{(s)} = \sqrt{\text{Var}(f \mid S = s)}$. Using Lemma 14, the Wasserstein barycenter ν^* is also a Gaussian, say $\mathcal{N}(\mu_\nu, \sigma_\nu^2)$. Substituting this into the definition of $\mathcal{U}(f)$ gives:

$$\mathcal{U}(f) = \min_{\mu_\nu, \sigma_\nu} \sum_{s=1}^{M} p_s \left[(\mu_f^{(s)} - \mu_\nu)^2 + (\sigma_f^{(s)} - \sigma_\nu)^2 \right].$$

This optimization problem decouples. The optimal μ_ν^* is the population average mean, $\mathbb{E}_S[\mu_f^{(S)}]$, and the optimal σ_ν^* is the population average standard deviation, $\mathbb{E}_S[\sigma_f^{(S)}]$. Plugging these back and using Lemma 13 gives the exact decomposition:

$$\mathcal{U}(f) = \sum_{s=1}^{M} p_s (\mu_f^{(s)} - \mathbb{E}[\mu_f^{(S)}])^2 + \sum_{s=1}^{M} p_s (\sigma_f^{(s)} - \mathbb{E}[\sigma_f^{(S)}])^2$$
$$= \text{Var}(\mathbb{E}[f \mid S]) + \text{Var}(\sqrt{\text{Var}(f \mid S)}).$$

Step 2: Decomposing the First-Moment Disparity. The FMD, $\text{Var}(\mathbb{E}[f \mid S])$, for $f(\boldsymbol{X}, S) = \langle \boldsymbol{X}, \boldsymbol{\beta} \rangle + \gamma S + \beta_0$ is derived as:

$$\text{Var}(\mathbb{E}[f \mid S])$$
$$= \underbrace{\gamma^2 \text{Var}(S)}_{\text{Direct Mean Bias}} + \underbrace{\text{Var}(\langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle)}_{\text{Indirect Mean Bias}} + \underbrace{2\gamma \text{Cov}(S, \langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle)}_{\text{Interaction}}.$$

Step 3: Decomposing the Second-Moment Disparity. The SMD is $\text{Var}(\sqrt{\text{Var}(f \mid S)})$. The conditional variance of f is $\text{Var}(f \mid S = s) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{(s)} \boldsymbol{\beta}$. Therefore, the conditional standard deviation is $\sigma_f^{(s)} = \sqrt{\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{(s)} \boldsymbol{\beta}}$. The SMD is the variance of this quantity:

$$\text{Var}(\sqrt{\text{Var}(f \mid S)}) = \text{Var}\left(\sqrt{\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{(S)} \boldsymbol{\beta}} \right).$$

This term corresponds to the **Indirect Structural Bias**.

Combining the results from Step 2 and Step 3 gives the full four-term decomposition stated in the proposition. $\square$

In the following, we provide the detailed derivations for the unfairness of the Bayes predictor (f^*) and our optimal DP predictor (f_{DP}^*), applying the general result from Appendix C.1 that $\mathcal{U}(f) = \text{Var}(\mathbb{E}[f|S]) + \text{Var}(\sqrt{\text{Var}(f|S)})$.

Unfairness of the Bayes Predictor, $\mathcal{U}(f^*)$. We apply the general unfairness formula to the Bayes predictor $f^* \in \mathcal{F}_{\text{linear}}$, which has coefficients $(\boldsymbol{\beta}^*, \gamma^*, \beta_0^*)$. The total unfairness naturally gives the FMD and SMD decomposition with:

1. **First-Moment Disparity:** This is the variance of the conditional mean, $\text{Var}(\mu_{f^*}^{(S)})$:

$$\text{Var}(\mu_{f^*}^{(S)}) =$$
$$\text{Var}(\langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta}^* \rangle) + \text{Var}(\gamma^* S) + 2\text{Cov}(\langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta}^* \rangle, \gamma^* S).$$

This corresponds exactly to the sum of the Indirect Mean Bias, Direct Mean Bias, and Interaction terms for the Bayes predictor.

2. **Second-Moment Disparity:** This is the variance of the conditional standard deviation, $\text{Var}(\sigma_{f^*}^{(S)})$:

$$\text{Var}(\sigma_{f^*}^{(S)}) = \text{Var}\left(\sqrt{(\boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma}^{(S)} \boldsymbol{\beta}^*} \right).$$

This is the Indirect Structural Bias for the Bayes predictor.

Unfairness of the Optimal DP Predictor, $\mathcal{U}(f_{DP}^*)$. We now prove that our optimal fair predictor, f_{DP}^*, has zero unfairness. We start with the definition of f_{DP}^* from Proposition 5 with $\varepsilon = 0$:

$$f_{DP}^*(\boldsymbol{x}, s) := \bar{\sigma}_{f^*} \left(\frac{\langle \boldsymbol{x} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^* \rangle}{\sigma_{f^*}^{(s)}} \right) + \bar{\mu}_{f^*}.$$

To compute its unfairness, we must find its conditional moments given $S = s$.

1. **Conditional Mean of f_{DP}^*:**

$$\mathbb{E}[f_{DP}^*(\boldsymbol{X}, S)|S = s]$$

$$= \mathbb{E}\left[\bar{\sigma}_{f^*}\left(\frac{\langle \boldsymbol{X} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^*\rangle}{\sigma_{f^*}^{(s)}}\right) + \bar{\mu}_{f^*}\,\middle|\, S = s\right]$$

$$= \frac{\bar{\sigma}_{f^*}}{\sigma_{f^*}^{(s)}}\mathbb{E}[\langle \boldsymbol{X} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^*\rangle|S = s] + \bar{\mu}_{f^*}$$

$$= \frac{\bar{\sigma}_{f^*}}{\sigma_{f^*}^{(s)}}\langle \mathbb{E}[\boldsymbol{X}|S = s] - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^*\rangle + \bar{\mu}_{f^*}$$

$$= \frac{\bar{\sigma}_{f^*}}{\sigma_{f^*}^{(s)}}\langle \boldsymbol{\mu}^{(s)} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^*\rangle + \bar{\mu}_{f^*}$$

$$= 0 + \bar{\mu}_{f^*}.$$

The conditional mean, $\mathbb{E}[f_{DP}^*|S = s]$, is equal to the constant $\bar{\mu}_{f^*}$ for all groups s.

2. **Conditional Variance of f_{DP}^*:**

$$\text{Var}(f_{DP}^*(\boldsymbol{X}, S)|S = s)$$

$$= \text{Var}\left(\frac{\bar{\sigma}_{f^*}}{\sigma_{f^*}^{(s)}}\langle \boldsymbol{X} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^*\rangle + \bar{\mu}_{f^*}\,\middle|\, S = s\right)$$

$$= \left(\frac{\bar{\sigma}_{f^*}}{\sigma_{f^*}^{(s)}}\right)^2 \text{Var}(\langle \boldsymbol{X}, \boldsymbol{\beta}^*\rangle|S = s)$$

$$= \left(\frac{\bar{\sigma}_{f^*}}{\sigma_{f^*}^{(s)}}\right)^2 ((\sigma_{f^*}^{(s)})^2)$$

$$= (\bar{\sigma}_{f^*})^2.$$

The conditional variance, $\text{Var}(f_{DP}^*|S = s)$, is equal to the constant $(\bar{\sigma}_{f^*})^2$ for all groups s. The conditional standard deviation is therefore also constant: $\sqrt{\text{Var}(f_{DP}^*|S = s)} = \bar{\sigma}_{f^*}$.

Since both the conditional mean and the conditional standard deviation of f_{DP}^* are constant across all groups s, their variance with respect to S is zero. Therefore, the total unfairness is $\mathcal{U}(f_{DP}^*) = 0 + 0 = 0$.

C.3 Proof of the Corollary 7: Unfairness of ε-RI predictor

Proof. The corollary states that the total unfairness of the optimal ε-fair predictor, $\mathcal{U}(f_{\varepsilon}^*)$, is exactly ε times the unfairness of the original Bayes predictor, $\mathcal{U}(f^*)$.

We recall the definition of the optimal ε-RI predictor as the geodesic interpolation between the optimal fair predictor (f_{DP}^*) and the Bayes predictor (f^*):

$$f_{\varepsilon}^* = (1 - \sqrt{\varepsilon})f_{DP}^* + \sqrt{\varepsilon}f^*.$$

We will compute the total unfairness of f_{ε}^* using the exact Wasserstein decomposition derived in the proof of Proposition 6:

$$\mathcal{U}(f_{\varepsilon}^*) = \text{Var}(\mathbb{E}[f_{\varepsilon}^*|S]) + \text{Var}(\sqrt{\text{Var}(f_{\varepsilon}^*|S)}).$$

Let us compute each of the two variance terms separately.

We first compute the conditional expectation of f_{ε}^* given $S = s$. By linearity of expectation:

$$\mathbb{E}[f_{\varepsilon}^*|S = s] = \mathbb{E}[(1 - \sqrt{\varepsilon})f_{DP}^* + \sqrt{\varepsilon}f^*|S = s]$$

$$= (1 - \sqrt{\varepsilon})\mathbb{E}[f_{DP}^*|S = s] + \sqrt{\varepsilon}\mathbb{E}[f^*|S = s].$$

From the analysis of f_{DP}^* in the proof of our main Proposition 5, we know that its conditional mean is the constant population average, $\mathbb{E}[f_{DP}^*|S = s] = \bar{\mu}_{f^*}$. The conditional mean of the Bayes predictor is simply $\mu_{f^*}^{(s)}$. Substituting these in gives:

$$\mathbb{E}[f_{\varepsilon}^*|S = s] = (1 - \sqrt{\varepsilon})\bar{\mu}_{f^*} + \sqrt{\varepsilon}\mu_{f^*}^{(s)}.$$

Now, we compute the variance of this expression with respect to the random variable S. Since $\bar{\mu}_{f^*}$ is a constant, it does not contribute to the variance:

$$\text{Var}(\mathbb{E}[f_{\varepsilon}^*|S]) = \text{Var}((1 - \sqrt{\varepsilon})\bar{\mu}_{f^*} + \sqrt{\varepsilon}\mu_{f^*}^{(S)})$$

$$= \text{Var}(\sqrt{\varepsilon}\mu_{f^*}^{(S)})$$

$$= \varepsilon \cdot \text{Var}(\mathbb{E}[f^*|S]).$$

Decomposing the Second-Moment Disparity of f_{ε}^*. Next, we compute the conditional variance, $\text{Var}(f_{\varepsilon}^*|S = s)$. As shown in the proof of Proposition 5, the predictor f_{ε}^* can be expressed in terms of a standardized score, $z^{(s)}(\boldsymbol{x}) = \langle \boldsymbol{x} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^*\rangle/\sigma_{f^*}^{(s)}$, which has a conditional variance of 1. The predictor is:

$$f_{\varepsilon}^*(\boldsymbol{x}, s) = \sigma_{\varepsilon}^{(s)} \cdot z^{(s)}(\boldsymbol{x}) + \mu_{\varepsilon}^{(s)}.$$

Since $\sigma_{\varepsilon}^{(s)}$ and $\mu_{\varepsilon}^{(s)}$ are constant given $S = s$, the conditional variance is:

$$\text{Var}(f_{\varepsilon}^*|S = s) = \text{Var}(\sigma_{\varepsilon}^{(s)} \cdot z^{(s)}(\boldsymbol{X})|S = s)$$

$$= (\sigma_{\varepsilon}^{(s)})^2 \text{Var}(z^{(s)}(\boldsymbol{X})|S = s)$$

$$= (\sigma_{\varepsilon}^{(s)})^2.$$

The conditional standard deviation is therefore simply $\sigma_{\varepsilon}^{(s)}$:

$$\sqrt{\text{Var}(f_{\varepsilon}^*|S = s)} = \sigma_{\varepsilon}^{(s)} = (1 - \sqrt{\varepsilon})\bar{\sigma}_{f^*} + \sqrt{\varepsilon}\sigma_{f^*}^{(s)}.$$

Since $\bar{\sigma}_{f^*}$ is a constant, it does not contribute to the variance:

$$\text{Var}(\sqrt{\text{Var}(f_{\varepsilon}^*|S)}) = \text{Var}((1 - \sqrt{\varepsilon})\bar{\sigma}_{f^*} + \sqrt{\varepsilon}\sigma_{f^*}^{(S)})$$

$$= \text{Var}(\sqrt{\varepsilon}\sigma_{f^*}^{(S)})$$

$$= (\sqrt{\varepsilon})^2 \text{Var}(\sigma_{f^*}^{(S)})$$

$$= \varepsilon \cdot \text{Var}(\sqrt{\text{Var}(f^*|S)}).$$

Assembling the Final Result. Finally, we sum the two decomposed disparity terms:

$$\mathcal{U}(f_{\varepsilon}^*) = \text{Var}(\mathbb{E}[f_{\varepsilon}^*|S]) + \text{Var}(\sqrt{\text{Var}(f_{\varepsilon}^*|S)})$$

$$= \varepsilon \cdot \text{Var}(\mathbb{E}[f^*|S]) + \varepsilon \cdot \text{Var}(\sqrt{\text{Var}(f^*|S)})$$

$$= \varepsilon \cdot \left(\text{Var}(\mathbb{E}[f^*|S]) + \text{Var}(\sqrt{\text{Var}(f^*|S)})\right)$$

$$= \varepsilon \cdot \mathcal{U}(f^*).$$

This completes the proof.

$\square$

D Proof of Proposition 8 : Additive Feature-Level Decomposition

This appendix provides the full derivation for the additive approximation of the feature-level decomposition of indirect unfairness, as stated in Proposition 8.

D.1 The General Decomposition with Interactional Terms

Our goal is to decompose the indirect unfairness of a predictor $f \in \mathcal{F}_{\text{linear}}$ with coefficients $(\boldsymbol{\beta}, \gamma)$. We begin with the exact expression for indirect unfairness:

$$\mathcal{U}_{\text{indirect}}(f) = \text{Var}(\langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle)$$
$$+ 2\gamma \text{Cov}(S, \langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle)$$
$$+ \text{Var}\left(\sqrt{\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{(S)} \boldsymbol{\beta}}\right).$$

Even under the simplifying assumption of uncorrelated features (diagonal $\Sigma^{(s)}$), a full decomposition reveals the presence of cross-terms that capture the statistical relationships between the group-level properties of different features.

Decomposition of Mean-Based Terms. The mean-based terms decompose as follows:

$$\text{Var}(\langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle) = \sum_{j=1}^{d} (\beta_j)^2 \text{Var}(\mu_j^{(S)})$$
$$+ 2 \sum_{j<k} \beta_j \beta_k \text{Cov}(\mu_j^{(S)}, \mu_k^{(S)}),$$

$$2\gamma \text{Cov}(S, \langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle) = \sum_{j=1}^{d} 2\gamma \beta_j \text{Cov}(S, \mu_j^{(S)}).$$

The term $\text{Cov}(\mu_j^{(S)}, \mu_k^{(S)})$ represents the **compounding of mean biases**.

Decomposition of the Structural Term (via Linearization). The structural bias term, $\text{Var}(\sqrt{\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{(S)} \boldsymbol{\beta}})$, is nonlinear. To decompose it, we first apply a first-order Taylor expansion. Let $V(S) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{(S)} \boldsymbol{\beta}$ be the conditional score variance. We linearize the function $g(v) = \sqrt{v}$ around $\bar{V} = \mathbb{E}[V(S)]$:

$$\sqrt{V(S)} \approx \sqrt{\bar{V}} + \frac{1}{2\sqrt{\bar{V}}}(V(S) - \bar{V}).$$

Taking the variance of this linear approximation yields:

$$\text{Var}(\sqrt{V(S)}) \approx \frac{1}{4\bar{V}} \text{Var}(V(S)).$$

Under the diagonal $\Sigma^{(s)}$ assumption,

$$V(S) = \sum_{j=1}^{d} \beta_j^2 (\sigma_j^{(S)})^2 \ .$$

The variance of this sum is:

$$\text{Var}(V(S))$$
$$= \sum_{j=1}^{d} \beta_j^4 \text{Var}((\sigma_j^{(S)})^2) + 2 \sum_{j<k} \beta_j^2 \beta_k^2 \text{Cov}((\sigma_j^{(S)})^2, (\sigma_k^{(S)})^2).$$

The term $\text{Cov}((\sigma_j^{(S)})^2, (\sigma_k^{(S)})^2)$ represents the **compounding of structural biases**.

D.2 Proof of the Additive Approximation (Proposition 8)

Proof. By combining the full decompositions above, we can express the exact indirect unfairness as a sum of two components: the primary (approximated) contributions from each feature and the interactional unfairness from pairs of features.

$$\mathcal{U}_{\text{indirect}}(f) \approx \sum_{j=1}^{d} \left((\beta_j)^2 \text{Var}(\mu_j^{(S)}) + \right.$$
$$\left. \frac{1}{4\bar{V}} (\beta_j)^4 \text{Var}((\sigma_j^{(S)})^2) + 2\gamma \beta_j \text{Cov}(S, \mu_j^{(S)}) \right)$$
$$+ 2 \sum_{j<k} \left[\beta_j \beta_k \text{Cov}(\mu_j^{(S)}, \mu_k^{(S)}) + \right.$$
$$\left. \frac{1}{4\bar{V}} \beta_j^2 \beta_k^2 \text{Cov}((\sigma_j^{(S)})^2, (\sigma_k^{(S)})^2) \right].$$

We define the first sum as the sum of the 'Approximated' primary feature contributions, $\sum_j \mathcal{U}_j^{\text{approx}}(f)$.

The proposition provides an additive approximation of the total indirect unfairness. This approximation is formally derived by neglecting the second term, the *Interactional Unfairness*. This is an approximation that can be highly accurate under the common scenario where the primary contributions from individual features are significantly larger than the second-order interactional effects. Therefore, we have:

$$\mathcal{U}_{\text{indirect}}(f) \approx \sum_{j=1}^{d} \mathcal{U}_j^{\text{approx}}(f)$$
$$= \sum_{j=1}^{d} \left((\beta_j)^2 \text{Var}(\mu_j^{(S)}) + \frac{1}{4\bar{V}} (\beta_j)^4 \text{Var}((\sigma_j^{(S)})^2) \right.$$
$$\left. + 2\gamma \beta_j \text{Cov}(S, \mu_j^{(S)}) \right).$$

This completes the proof. □

E Auditing Model Adequacy with the GWR^2

We prove the precise statistical conditions under which the global R^2 and our proposed GWR^2 are equivalent, demonstrating that their divergence is a direct consequence of underlying group-level disparities in the data.

Proposition 15 (Characterization of the GWR^2). *Let $f(\boldsymbol{X}, S)$ be a linear model of the form*

$$f(\boldsymbol{X}, S) = \langle \boldsymbol{X}, \boldsymbol{\beta} \rangle + \gamma S + \beta_0 \ .$$

We also define the global coefficient of determination by

$$R_{\text{global}}^2 := 1 - \frac{\text{Var}(Y - f(\boldsymbol{X}, S))}{\text{Var}(Y)} \ .$$

Then, we have $R_{\text{global}}^2 = GWR^2$ iff all three are verified:

*(i) **(Unaware model)** The model does not include the sensitive attribute as a predictor. This means we set $\gamma = 0$.*

*(ii) **(No residual association)** After accounting for $\boldsymbol{X}$, the attribute S has no remaining linear association with the outcome Y. This means the true parameter $\gamma^* = 0$.*

*(iii) **(Feature independence)** The features $\boldsymbol{X}$ are independent of the sensitive attribute S.*

Proof can be found in Appendix E.1.

Decomposition of the R^2 Discrepancy When the strict conditions for equality between the global and group-weighted R^2 metrics are not met, a discrepancy arises. The following proposition provides an exact analytical expression for this discrepancy under the simplifying assumption of homoscedasticity, revealing that the gap is driven by how a *fair-unaware* model accounts for the between-group variance present in the data.

Proposition 16 (Decomposition of the R^2 Gap). *Let $f(\boldsymbol{X})$ be a fair-unaware predictor. For simplicity sake we assume (only in this proposition) that the conditional variances of both the outcome and the residuals are constant across all groups $s \in \mathcal{S}$ (homoscedasticity):*

- *Within-group outcome variance:*

$$\mathrm{Var}(Y \mid S = s) = W_Y, \quad \text{for all } s \in \mathcal{S}$$

- *Within-group residual variance:*

$$\mathrm{Var}(Y - f(\boldsymbol{X}) \mid S = s) = W_R, \quad \text{for all } s \in \mathcal{S}$$

Let $B_Y := \mathrm{Var}(\mathbb{E}[Y \mid S])$ denote the between-group variance of the outcome, and let $B_R := \mathrm{Var}(\mathbb{E}[Y - f(\boldsymbol{X}) \mid S])$ denote the between-group variance of the model's residuals.

Then the difference between the group-weighted and global R-squared metrics is given by:

$$GWR^2 - R^2_{\text{global}} = \frac{W_Y B_R - W_R B_Y}{W_Y(W_Y + B_Y)} \quad (9)$$

Proof can be found in Appendix E.2.

Implication The decomposition in (9) provides a clear interpretation of the discrepancy. Specifically, we show that $GWR^2 > R^2_{\text{global}}$ if and only if $W_Y B_R > W_R B_Y$, which can be rewritten as:

$$\frac{B_R}{W_R} > \frac{B_Y}{W_Y} \quad .$$

This inequality compares the ratio of between-group to within-group variance for the residuals (R) and for the original outcome (Y). When a fair-unaware model f fails to explain the between-group variance (B_Y) that exists in the outcome, this unexplained variance is transferred to the residuals, inflating B_R. This leads to a situation where the model appears to perform worse from a global perspective than it does on average within the groups. The gap between the two R-squared metrics is therefore a direct signal of how well the model accounts for the group-level structural differences present in the data.

E.1   Proof of Proposition 15

Proof. The equality $R^2_{\text{global}} = GWR^2$ holds if and only if their fractional parts are equal:

$$\frac{\mathrm{Var}(Y - f)}{\mathrm{Var}(Y)} = \sum_s p_s \frac{\mathrm{Var}(Y - f(\cdot, s) \mid S = s)}{\mathrm{Var}(Y \mid S = s)} \quad (*)$$

Part 1: ($\Leftarrow$) Assume conditions (i), (ii), and (iii) are all true. Our goal is to show that equation (*) holds.

Step A: Equality of denominators. We first show that $\mathrm{Var}(Y) = \mathrm{Var}(Y \mid S = s)$ for all s. By the Law of Total Variance,

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid S)] + \mathrm{Var}(\mathbb{E}[Y \mid S]) \quad .$$

From Eq. (4) and using condition (ii) ($\gamma^* = 0$), we have $Y = \langle \boldsymbol{X}, \boldsymbol{\beta}^* \rangle + \beta_0^* + \zeta$.

- The conditional variance is

$$\mathrm{Var}(Y \mid S = s) = \mathrm{Var}(\langle \boldsymbol{X}, \boldsymbol{\beta}^* \rangle + \zeta \mid S = s) \quad .$$

By condition (iii), the distribution of $\boldsymbol{X}$ is independent of S, so this variance is a constant, C, for all s. Thus,

$$\mathbb{E}[\mathrm{Var}(Y \mid S)] = C \quad .$$

- The conditional expectation is

$$\mathbb{E}[Y \mid S = s] = \mathbb{E}[\langle \boldsymbol{X}, \boldsymbol{\beta}^* \rangle + \beta_0^* \mid S = s] \quad .$$

By condition (iii), $\mathbb{E}[\boldsymbol{X} \mid S = s]$ is a constant vector, making $\mathbb{E}[Y \mid S = s]$ a constant value for all s.

It follows that $\mathrm{Var}(\mathbb{E}[Y \mid S]) = 0$. Therefore,

$$\mathrm{Var}(Y) = C + 0 = C \quad ,$$

which means $\mathrm{Var}(Y) = \mathrm{Var}(Y \mid S = s)$ for all s.

With identical denominators, equation (*) simplifies to showing that:

$$\mathrm{Var}(Y - f(\boldsymbol{X}))$$
$$= \sum_s p_s \mathrm{Var}(Y - f(\boldsymbol{X}) \mid S = s)$$
$$= \mathbb{E}[\mathrm{Var}(Y - f(\boldsymbol{X}) \mid S)] \quad .$$

Note we used condition (i) to write $f(\boldsymbol{X}, s) = f(\boldsymbol{X})$.

Step B: Equality of numerators. Let the residual be

$$R = Y - f(\boldsymbol{X}) \quad .$$

The Law of Total Variance for R is

$$\mathrm{Var}(R) = \mathbb{E}[\mathrm{Var}(R \mid S)] + \mathrm{Var}(\mathbb{E}[R \mid S]) \quad .$$

The simplified equality from Step A holds if and only if $\mathrm{Var}(\mathbb{E}[R \mid S]) = 0$. This requires $\mathbb{E}[R \mid S = s]$ to be constant for all s.

$$\mathbb{E}[R \mid S = s] = \mathbb{E}[Y \mid S = s] - \mathbb{E}[f(\boldsymbol{X}) \mid S = s]$$

We already showed in Step A that $\mathbb{E}[Y \mid S = s]$ is constant. By condition (iii), the distribution of $\boldsymbol{X}$ is independent of S, so $\mathbb{E}[f(\boldsymbol{X}) \mid S = s]$ must also be constant for any function f. Since both terms are constant, their difference is constant, meaning $\mathrm{Var}(\mathbb{E}[R \mid S]) = 0$. The equality holds.

Part 2: ($\Rightarrow$)   Assume equation (*) holds.

- *Condition (i) must hold.* If f depended on S, the global error variance $\mathrm{Var}(Y - f(\boldsymbol{X}, S))$ would contain a term related to the main effect of S in the model f, whereas the group-level variances $\mathrm{Var}(Y - f(\boldsymbol{X}, s) \mid S = s)$ would not. The functional forms would be different, making a robust equality impossible.
- *Conditions (ii) and (iii) must hold.* If $\gamma^* \neq 0$ or if $\boldsymbol{X}$ is not independent of S, then as shown in Step A,

$$\mathrm{Var}(\mathbb{E}[Y \mid S]) > 0 \ ,$$

and consequently

$$\mathrm{Var}(Y) > \mathbb{E}[\mathrm{Var}(Y \mid S)] \ .$$

This creates a structural mismatch in the denominators of (*). Similarly, the conditional mean of the residual $\mathbb{E}[Y - f(\boldsymbol{X}) \mid S = s]$ would not be constant, creating a mismatch in the numerators where $\mathrm{Var}(Y - f) > \mathbb{E}[\mathrm{Var}(Y - f \mid S)]$. For the equality (*) to hold generally, these structural misalignments must be absent, which requires $\gamma^* = 0$ and $\boldsymbol{X}$ to be independent of S.

This completes the proof. $\square$

E.2   Proof of Proposition 16

Proof. We derive the expressions for GWR^2 and R^2_{global} separately under the stated assumptions.

First, we express the group-weighted R-squared. The R-squared for a specific group s is:

$$R_s^2 = 1 - \frac{\mathrm{Var}(Y - f(\boldsymbol{X}) \mid S = s)}{\mathrm{Var}(Y \mid S = s)} = 1 - \frac{W_R}{W_Y}$$

Since this value is constant for all groups, the weighted average is simply the value itself:

$$GWR^2 = \sum_s p_s \left(1 - \frac{W_R}{W_Y}\right) = 1 - \frac{W_R}{W_Y} \qquad (10)$$

Next, we express the global R-squared. We apply the Law of Total Variance to the denominator, $\mathrm{Var}(Y)$, and the numerator, $\mathrm{Var}(Y - f(\boldsymbol{X}))$.

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathbb{E}[\mathrm{Var}(Y \mid S)] + \mathrm{Var}(\mathbb{E}[Y \mid S]) \\
&= W_Y + B_Y \\
\mathrm{Var}(Y - f(\boldsymbol{X})) &= \mathbb{E}[\mathrm{Var}(Y - f(\boldsymbol{X}) \mid S)] \\
&\quad + \mathrm{Var}(\mathbb{E}[Y - f(\boldsymbol{X}) \mid S]) \\
&= W_R + B_R
\end{aligned}$$

Substituting these into the definition of global R-squared yields:

$$R^2_{\mathrm{global}} = 1 - \frac{\mathrm{Var}(Y - f(\boldsymbol{X}))}{\mathrm{Var}(Y)} = 1 - \frac{W_R + B_R}{W_Y + B_Y} \qquad (11)$$

Finally, we compute the difference between (10) and (11):

$$\begin{aligned}
GWR^2 - R^2_{\mathrm{global}} &= \left(1 - \frac{W_R}{W_Y}\right) - \left(1 - \frac{W_R + B_R}{W_Y + B_Y}\right) \\
&= \frac{W_R + B_R}{W_Y + B_Y} - \frac{W_R}{W_Y} \\
&= \frac{W_Y(W_R + B_R) - W_R(W_Y + B_Y)}{W_Y(W_Y + B_Y)} \\
&= \frac{W_Y W_R + W_Y B_R - W_R W_Y - W_R B_Y}{W_Y(W_Y + B_Y)} \\
&= \frac{W_Y B_R - W_R B_Y}{W_Y(W_Y + B_Y)}
\end{aligned}$$

This completes the proof. $\square$

F   Numerical Experiments
F.1   Synthetic Dataset

We provide additional details on the construction of the synthetic dataset used in our experiments. The data-generating process is governed by four control parameters $T := (T_y, T_{\mathrm{mean}}, T_{\mathrm{std}}, T_{\mathrm{corr}})$. We generate synthetic triplets $(\boldsymbol{X}, S, Y)$ with the following properties:

- **Sensitive attribute:** $S \in \{1, 2\}$ follows a shifted Bernoulli distribution $\mathcal{B}(q) + 1$ where $q = \mathbb{P}(Z > \tau)$ with $Z \sim \mathcal{N}(0, 1)$ and fixed threshold τ.
- **Features:** $\boldsymbol{X} \in \mathbb{R}^d$ is a multivariate Gaussian random vector whose distribution depends on the sensitive attribute S. The conditional distribution parameters are designed to introduce **indirect bias** through group differences in means, variances and correlation structures.
- **Outcome:** $Y = \sum_{j=1}^d X_j + T_y \cdot S$, where parameter T_y directly controls the **direct bias** and introduces outcome disparity between groups.

Definition of the features X simulations: The conditional distribution of $\boldsymbol{X}$ given $S = s$ is $\mathcal{N}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)})$, for each $s \in \mathcal{S}$, where the parameters are constructed as follows to systematically introduce different types of bias:

1. **Mean vectors:** For each feature X_j, with $j \in \{1, \dots, d\}$:

$$\begin{aligned}
\mu_j^{(1)} &\sim \mathcal{B}(3, p) \quad \text{where } p \sim \mathcal{U}_{[0,1]} \\
\mu_j^{(2)} &= \mu_j^{(1)} + T_{\mathrm{mean}}
\end{aligned}$$

The parameter T_{mean} controls indirect mean bias by systematically shifting the mean of features for group 2 relative to group 1. When $T_{\mathrm{mean}} = 0$, both groups have identical feature means, eliminating this source of bias.

2. **Standard deviations:** For each feature X_j, with $j \in \{1, \dots, d\}$:

$$\begin{aligned}
\sigma_j^{(1)} &\sim \mathcal{U}_{[0,2]} \\
\sigma_j^{(2)} &= \sigma_j^{(1)} + \sqrt{T_{\mathrm{std}}}
\end{aligned}$$

The parameter T_{std} contributes to indirect structural bias by controlling differences in feature variances between groups. When $T_{\mathrm{std}} = 0$, both groups have identical feature standard deviations.

3. **Correlation matrices:** The correlation matrix $\boldsymbol{\rho}^{(s)}$ of $\boldsymbol{X}$ given $S = s$, for each $s \in \mathcal{S}$, is constructed as follows to control the correlation component of structural bias.

   For each group s, we generate a random matrix $\boldsymbol{A}^{(s)} \in \mathbb{R}^{d \times d}$ with i.i.d. entries $A_{ij}^{(s)} \sim \mathcal{N}(0, 1)$.

   - If $T_{corr} = 0$ (Independent features for both groups):
   $$\boldsymbol{\rho}^{(1)} = \boldsymbol{\rho}^{(2)} = \boldsymbol{I}_d \ .$$

   - If $T_{corr} \in (0, 1)$ (Group-Specific Correlations):
   $$\boldsymbol{\rho}^{(s)} = T_{corr} \cdot \boldsymbol{C}^{(s)} + (1 - T_{corr}) \cdot \boldsymbol{I}_d \ .$$

   where $\boldsymbol{C}^{(s)} = \boldsymbol{D}_s^{-\frac{1}{2}} (\boldsymbol{A}^{(s)\top} \boldsymbol{A}^{(s)}) \boldsymbol{D}_s^{-\frac{1}{2}}$ is the normalized correlation matrix derived from $\boldsymbol{A}^{(s)}$, $\boldsymbol{D}_s = \mathrm{diag}(\boldsymbol{A}^{(s)\top} \boldsymbol{A}^{(s)})$ is the diagonal matrix of $\boldsymbol{A}^{(s)\top} \boldsymbol{A}^{(s)}$ and $\boldsymbol{I}_d$ is the identity matrix.

   - If $T_{corr} = 1$ (Identical Correlations for both groups):
   $$\boldsymbol{\rho}^{(1)} = \boldsymbol{\rho}^{(2)} = \boldsymbol{D}_1^{-\frac{1}{2}} (\boldsymbol{A}^{(1)\top} \boldsymbol{A}^{(1)}) \boldsymbol{D}_1^{-\frac{1}{2}} \ .$$

   Both groups share the same correlation structure.

   The parameter $T_{corr} \in (0, 1)$ defines the strength of correlation-based structural bias, while $T_{corr} \in \{0, 1\}$ eliminates correlation-based structural bias.

4. **Covariance matrices:** The final covariance matrix of $\boldsymbol{X}$ given $S = s$, for each $s \in \mathcal{S}$, is computed as:
$$\boldsymbol{\Sigma}^{(s)} = \mathrm{diag}(\boldsymbol{\sigma}^{(s)}) \boldsymbol{\rho}^{(s)} \mathrm{diag}(\boldsymbol{\sigma}^{(s)}) \ ,$$

   where $\mathrm{diag}(\boldsymbol{\sigma}^{(s)})$ is the diagonal matrix containing the standard deviations $\sigma_j^{(s)}$.

Systematic Bias Control. This synthetic data generation process allows us to systematically control and isolate different sources of bias through the parameter vector T:

- **Direct Bias** is controlled by T_y through explicit dependence of Y on S.
- **Indirect Mean Bias** is driven by T_{mean} through systematic differences in $\boldsymbol{\mu}^{(s)}$.
- **Indirect Structural Bias** is controlled jointly by T_{std} and T_{corr} through group-specific differences in $\boldsymbol{\Sigma}^{(s)}$.

F.2 Additional illustrations of coefficients adjustments for fairness in bias scenario

Building on Section 7.1, we further illustrate transparent bias remediation through coefficient adjustments. Using our synthetic dataset, we examine how our model adapts coefficients across three scenarios of increasing complexity, where each scenario adds new bias sources to the previous one: (1) Direct Bias, (2) Direct and Indirect Mean Bias, (3) Direct, Indirect Mean and Structural Bias.

Direct Bias. In presence of direct bias only, our model eliminates unfairness through two adjustments: (1) it nullifies the sensitive attribute coefficient, removing direct dependence on group membership, and (2) it adjusts the intercept equally for both groups to maintain predictive accuracy (Fig. 8).
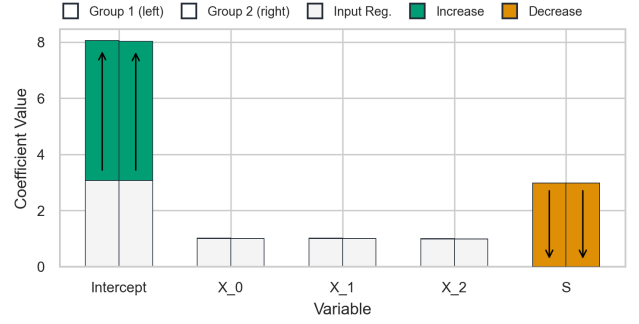

Figure 8: Coefficients adjustments for fairness in a direct bias scenario T = $(3, 0, 0, 0)$, shown on synthetic data for a sample of features.

Direct and Indirect Mean Bias. When indirect mean bias is added to direct bias, the remediation strategy becomes asymmetric. Beyond nullifying the sensitive attribute coefficient, the model compensates for the systematic difference in feature means between groups through unequal intercept adjustments. Group 1, which has lower feature means by construction ($\mu^{(1)} < \mu^{(2)}$ due to $T_{mean} > 0$), receives a larger positive intercept adjustment to offset this disadvantage and achieve fair outcomes (Fig. 9).
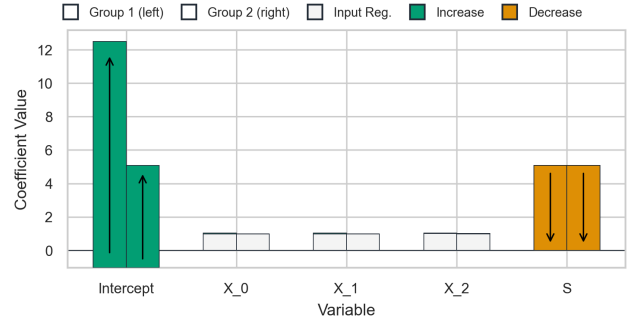

Figure 9: Coefficients adjustments for fairness in a direct and indirect mean bias scenario T = $(3, 2, 0, 0)$, shown on synthetic data for a sample of features.

Direct, Indirect Mean and Structural Bias. Adding structural bias ($T_{std} > 0$, $T_{corr} > 0$) to the previously described biases activates the scaling factor $\frac{\bar{\sigma}_{f^*}}{\sigma_{f^*}^{(s)}}$ of our fair model. As a result, structural bias introduces two additional remediation mechanisms that build upon the existing direct and indirect mean bias corrections: (1) a new mechanism, the adjustment of the slope and (2) the further adjustment of the intercept. To analyze both effects clearly, we examine them in two stages: first considering variance differences alone, then incorporating both variance and correlation differences.

1. **Adjustment of the slope:** The features' coefficients are

now adjusted by the group-dependent scaling factor since

$$\boldsymbol{\beta}_0^{(s)} = \frac{\bar{\sigma}_{f^*}}{\sigma_{f^*}^{(s)}} \boldsymbol{\beta}^*,$$

(as introduced in Perspective 2 with $\varepsilon = 0$, within Section 4.4).

- **Structural Bias through Variance:** Group 2's higher variance leads to coefficient slightly down-scaling, while Group 1's coefficients are up-scaled (Fig. 10). Due to the uniform variance structure (no correlations and constant variance differences across features), all coefficients within each group are adjusted by the same scaling factor, resulting in parallel shifts for all features.
- **Structural Bias through Variance and Correlations:** In a full bias scenario (Fig. 4), all coefficients within each group are adjusted by a slightly different scaling factor due to the group correlations.

2. **Modified adjustment of the intercept:** The intercept adjustment mechanism, observed in the previous scenario, is now modified because the scaling factor affects the mean-based correction. The intercept of our fair model (as introduced in Perspective 2 with $\varepsilon = 0$, within Section 4.4) becomes :

$$\beta_{0,0}^{(s)} = \bar{\mu}_{f^*} - \left(\frac{\bar{\sigma}_{f^*}}{\sigma_{f^*}^{(s)}} \right) \langle \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^* \rangle .$$

- **Structural Bias through Variance:** In Fig. 10, due to Group 2 higher variance, the scaling factor $\frac{\bar{\sigma}_{f^*}}{\sigma_{f^*}^{(s)}}$ reduces the correction applied to its higher means. Thus, Group 2 receives a slightly larger intercept's adjustment than Group 1, reversing the asymmetry observed in the previous scenario with direct and indirect mean bias (Fig. 9).
- **Structural Bias through Variance and Correlations:** Fig. 4 shows that group-specific correlation structures (randomly generated in our synthetic data) further modify the intercept adjustments, with Group 1 receiving additional corrections that reduce the asymmetry between both intercepts.

Thus, unlike previous scenarios where adjustments were purely additive, structural bias remediation works through scaling effects.

This detailed coefficient analysis demonstrates the value of our framework's interpretability: practitioners can precisely diagnose bias sources, understand the specific remediation mechanisms at work, and make informed decisions about fairness interventions.
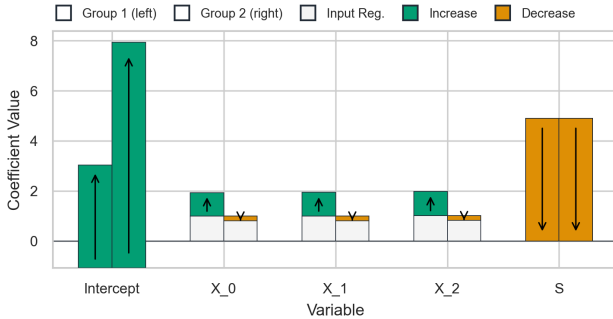

Figure 10: Coefficients adjustments for fairness in a direct and indirect (through mean and variance) bias scenario $T = (3, 2, 3, 0)$, shown on synthetic data for a sample of features.