

SUPPLEMENTARY MATERIAL

A Data Generation Processes in Section 3.2

For our simulation study, first, we define the data generation process. The protected attribute $A \in \{0, 1\}$ is considered in simple binary form. $n, d \in \mathbb{N}$ are our sample size and feature dimension, respectively. This general framework is considered unless otherwise specified.

A.1 Linear Bias

Let's define $\mu_{\text{gap}} \geq 0$, to introduce mean shift in the first feature, and the coefficient $\gamma \in \mathbb{R}$ to quantify the effect of A on Y .

Formally, let

$$m = (1, 0, \dots, 0)^\top \in \mathbb{R}^d, \quad \mu_0 = 0 \in \mathbb{R}^d, \quad \mu_1 = \mu_{\text{gap}} m.$$

For each observation $i = 1, \dots, n$, we independently draw

$$A_i \sim \text{Bernoulli}(0.5), \quad Z_i \sim \mathcal{N}(0, I_d), \quad \varepsilon_i \sim \mathcal{N}(0, 1),$$

and the feature vector is defined as

$$X_i = \mu_{A_i} + Z_i,$$

such that

$$X_i \mid (A_i = a) \sim \mathcal{N}(\mu_a, I_d).$$

A score is then specified as

$$Z_i^\star = m^\top X_i + \gamma A_i + \varepsilon_i,$$

from which the binary response is obtained as

$$Y_i = \mathbf{1}\{Z_i^\star > 0\}.$$

Accordingly, the conditional probability of a positive outcome satisfies

$$\Pr(Y_i = 1 \mid X_i = x, A_i = a) = \Phi(m^\top x + \gamma a),$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

A.2 Nonlinear Bias

Let's define $a_{\text{scale}} > 0$, and $\sigma_{\text{proxy}} > 0$. The logistic link function is denoted by $\sigma(z) = \frac{1}{1+e^{-z}}$. For each observation $i = 1, \dots, n$, we independently draw,

$$A_i \sim \text{Bernoulli}(0.5), \quad U_i \sim \mathcal{N}(0, 1), \quad \varepsilon_{2i}, \varepsilon_{3i} \sim \mathcal{N}(0, 1), \quad \eta_i \sim \mathcal{N}(0, \sigma_{\text{proxy}}^2).$$

We define the features as follows,

$$X_{1i} \sim \mathcal{N}(1.5 a_{\text{scale}} A_i + 0.5 U_i, 1)$$

$$X_{2i} = \sin(X_{1i}) + 0.3 a_{\text{scale}} A_i + \varepsilon_{2i}$$

$$X_{3i} = \log(|X_{1i}| + 1) (2A_i - 1) a_{\text{scale}} + 0.2U_i + \varepsilon_{3i}$$

A proxy variable is then defined as,

$$\text{proxy}_i = 0.4X_{1i} - 0.2X_{2i} + 0.6X_{3i} + \eta_i.$$

so, the feature vector is,

$$X_i = (X_{1i}, X_{2i}, X_{3i}, \text{proxy}_i)^\top \in \mathbb{R}^4.$$

We define a latent score,

$$Z_i = 1.2X_{1i} - X_{2i} + 0.8X_{3i} + 0.7 \text{proxy}_i + 0.9 a_{\text{scale}} A_i + 0.5U_i,$$

and the binary outcome is generated as

$$Y_i \mid (X_i, A_i, U_i) \sim \text{Bernoulli}(\sigma(Z_i)).$$

A.3 Noisy Outcome Process

This data generation process makes the observed Y noisy based on the classes of A .

Let $n, d \in \mathbb{N}$, $\mu_{\text{gap}} \geq 0$, and $\gamma \in \mathbb{R}$ be as previously defined. Let's define $q_0, q_1 \in [0, 1]$, the probabilities of corrupting the observed response for individuals with $A = 0$ and $A = 1$, respectively.

We generate clean features and outcomes from the linear bias process in **A.1**:

$$X_i \in \mathbb{R}^d, \quad Y_i^{\text{clean}} \in \{0, 1\}, \quad A_i \in \{0, 1\}.$$

Y is then altered with probability q_a , conditional on the class of A :

$$\Pr(Y_i^{\text{obs}} \neq Y_i^{\text{clean}} \mid A_i = a) = q_a.$$

When $q_0 \neq q_1$, one subgroup experiences a higher level of output corruption.