

Achieving Fairness-Utility Trade-off through Synthetic Data

Bei Jiang

Department of Mathematical and Statistical Sciences, University of Alberta

Alberta Machine Intelligence Institute (Amii)

2026 IUSEP

Algorithms can make life-changing decisions

- Learning algorithms have been increasingly used to improve decision making, e.g., job employment, criminal justice, loan approval and student admission.
- ▶ Judges across the United States use a computer algorithm called COMPAS to help them decide if a defendant can be released on bail before trials.



We have problems.

The New York Times

Algorithms Learn Our Workplace Biases. Can They Help Us Unlearn Them?

With the help of intelligent machines, humans can be nudged to make choices that make workplaces fairer for everyone.

Give this article



- ▶ Around 2015, Amazon built a machine-learning system to review resumes and rate candidates, using prior applicant data.
- ▶ The algorithm discriminated against female candidates.
- ▶ Amazon abandoned the tool in 2017.

What went wrong in Amazon's algorithms?

- ▶ The developed algorithm did not “malfunction” in a technical sense; it did exactly what it was intended for, learning patterns from the training data.
- ▶ It was trained on historical hiring data in which men dominated technical roles.
- ▶ The algorithm picked up the pattern that female candidates were less preferable.
- ▶ In other words, **training learning algorithms on unfair data was the true culprit behind the unfair outcomes.**

Unfairness from under-representation

- ▶ Each panel shows the same feature space for a binary classification task.
- ▶ The Majority and Minority groups exhibit distinct decision boundaries.
- ▶ When data are pooled, **the learned global classifier is dominated by the Majority group**, performing poorly for the Minority group.

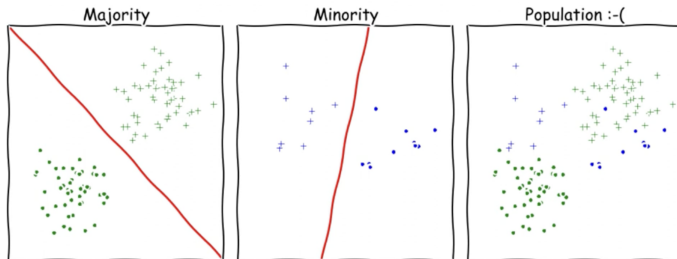


Illustration by Moritz Hardt (<https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de>)

Root causes of algorithmic fairness:

- ▶ **Data bias:** Training data reflect existing societal or historical inequalities (our focus).
- ▶ **Sampling bias:** minority or marginalized groups under-represented in the samples.

Other types of unfairness include **deployment bias**: human interactions and feedback loops that alter the data distribution and reinforce existing biases over time.

- ▶ $(\underbrace{\text{feature}}_X, \underbrace{\text{sensitive attribute}}_S, \underbrace{\text{label}}_Y) \sim \mathbb{P} \text{ on } \mathbb{R}^p \times [K] \times \mathbb{R}.$
- ▶ Observations: $(X_1, S_1, Y_1), \dots, (X_n, S_n, Y_n).$
- ▶ **Goal:** build a prediction model $f(X_i, S_i)$ that minimizes

$$\operatorname{argmin}_f \sum_i \mathcal{L}(Y_i - f(X_i, S_i)) + \mathcal{R}(f)$$

under some fairness constraint (next slide)

- ▶ $\mathcal{L}$ is a loss function
- ▶ $\mathcal{R}(f)$ is a possible penalty/regularization term

- ▶ **Group-based fairness:** seeks equality of certain performance measures across subgroups defined by a sensitive attribute S . Typical examples include:
 - ▶ **Demographic Parity (DP):** The prediction is independent of S , i.e. $\hat{Y} \perp S$. In the binary case,

$$P(\hat{Y} = 1 \mid S = 0) = P(\hat{Y} = 1 \mid S = 1).$$

- ▶ **Equalized Odds:** Predictions are independent of S *conditional* on the true label Y :

$$P(\hat{Y} = 1 \mid Y = y, S = 0) = P(\hat{Y} = 1 \mid Y = y, S = 1), \quad y \in \{0, 1\}.$$

- ▶ **Equal Opportunity:** A relaxed version of Equalized Odds, requiring equality only for truly positive cases:

$$P(\hat{Y} = 1 \mid Y = 1, S = 0) = P(\hat{Y} = 1 \mid Y = 1, S = 1).$$

All these criteria can be extended to **conditional** versions by additionally conditioning on X , e.g., conditional DP requires $\hat{Y} \perp S \mid X$.

- ▶ **2. Individual Fairness:** Individuals with similar features should receive similar outcomes.
- ▶ **3. Counterfactual fairness:** require that an individual's prediction would remain the same if, counterfactually, their sensitive attribute were changed.

Important caveat: Many of these criteria are *mutually incompatible*.

Please refer to tutorial by Barocas et al. (2022) at <https://fairmlbook.org>.

Goal: modify training data to remove bias with the goal that any downstream prediction model f , trained on **debiased** data achieve desired fairness.

- ▶ **Relabeling:** Flip labels for top M disadvantaged individuals so that group-level fairness is achieved *at the sample level*.

However, there is no guarantee that the underlying distribution for modified labels $\tilde{Y}$ truly satisfies $\tilde{Y} \perp S$.

References: Kamiran & Calders (2012).

- **Fair representation learning:** learn an encoder f in the hope

$$Z = f(X, S) \perp S$$

then any predictor g built upon it satisfies $\hat{Y} = g(Z) \perp S$.

- **Implementation:** solve f and g such that they maximize predictive utility while minimizing dependence on S :

$$\min_{f, g} \mathcal{L}_{\text{task}}(Y, g(Z)) + \lambda \mathcal{R}_{\text{fair}}(S, Z),$$

However, such methods are heuristic:

- **No global convergence guarantee:** imposing fairness constraints (often non-convex) can lead to instability.
- **No theoretical guarantee that $Z \perp S$:** fairness is evaluated empirically on the training sample.
- **No theoretical guarantee on the fairness–utility trade-off** controlled by the tuning parameter λ .

- **GAN-based fair synthetic data:** learn a conditional generator G in the hope

$$\tilde{Y} = G(X, S, U), \quad U \sim P_U,$$

$$P(\tilde{Y} \mid X, S) \approx P(Y \mid X, S), \quad \text{while } \tilde{Y} \text{ satisfies a fairness target.}$$

then the synthetic labels can be used in place of the observed label:

$$(X, S, Y) \longrightarrow (X, S, \tilde{Y}).$$

- **Implementation:** solve G and discriminator D such that they maximize conditional fidelity while minimizing unfair dependence on S :

$$\min_G \max_D \mathcal{L}_{\text{cGAN}}((X, S, Y), (X, S, \tilde{Y})) + \lambda \mathcal{R}_{\text{fair}}(S, \tilde{Y}, X),$$

where $\mathcal{R}_{\text{fair}}$ may encode, for example, demographic parity, equalized odds, or an adversarial penalty discouraging prediction of S from $\tilde{Y}$.

We need better methods.

Limitations of existing ML methods:

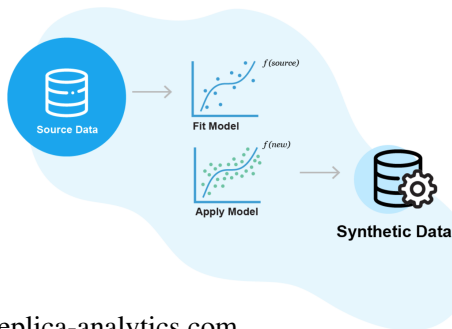
- ▶ **Relabeling approaches:** Intuitive but lack theoretical guarantees at the population level.
- ▶ **Fair representation and GAN-based approaches:** Frequently rely on black-box generative models that are unstable and no guarantee on achieving convergence, fairness or fairness-utility trade-off.

Goal: Given $\mathcal{D} = \{Y_i, X_i, S_i\}_{i=1}^n$, we generate a fair version $\hat{\mathcal{D}} = \{\hat{Y}_i, X_i, S_i\}_{i=1}^n$ by replacing Y_i with synthetic labels $\hat{Y}_i$, sampled from a data-augmented predictive distribution.

- ▶ Our framework provides the **first provable fairness-utility trade-off guarantee**, controlled by a **user-specified parameter** $\alpha \in [0, 1]$.
 - ▶ **Faithfulness guarantee:** For any fixed α , the Wasserstein distance $W_2(\mu_{\hat{Y}|X,S}, \mu_{Y|X,S})$ is bounded by a quantity that depends on α . In particular, $\alpha \rightarrow 1, \hat{Y}|X, S \xrightarrow{P} Y|X, S$.
 - ▶ **Bias reduction guarantee:** $\mathcal{UF}(\mathcal{P}_{\hat{\mathcal{D}}}) \leq \alpha \mathcal{UF}(\mathcal{P}_{\mathcal{D}})$.

What is synthetic data?

- ▶ The idea is not new: Rubin (1993) proposed the first model-based synthetic method using multiple imputation (MI):
 - ▶ replace the original data by *predictive* values drawn from the posterior predictive distributions
 - ▶ use combining rules to analyze the released multiple synthetic datasets.
 - ▶ Later, Wang and Reiter (2012) extended it using a frequentist plug-in approach, generating synthetic data from the fitted model using parameter estimates instead of posterior draws.



source: <https://www.replica-analytics.com>

Naive idea: First learn a fair representation $W = f(X, S)$, and then generate synthetic labels via

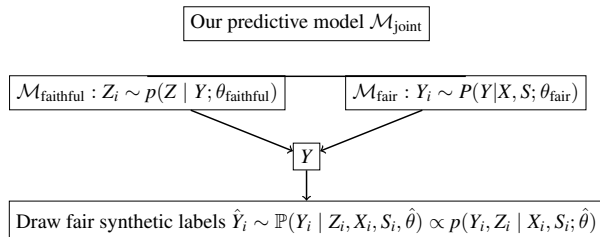
$$\hat{Y} \sim P(Y \mid W).$$

Limitations of this approach:

- ▶ Although W may satisfy a fairness criterion, it can discard important information about Y that is not fully captured by (X, S) or is removed during debiasing.
- ▶ The resulting $\hat{Y}$ may be tailored to a specific prediction task and not preserve all for *all* downstream analyses.

Our new data-augmented predictive framework

- Our new framework generalizes Rubin's original framework:



- **Data augmentation step:** we generate Z based on Y through $\mathcal{M}_{\text{faithful}}$ and the tuning parameter in θ_{faithful} (Details will be given later).
- **Fit the joint model** for $\mathcal{D}$ and $\{Z_i\}_{i=1}^n$, where

$$\underbrace{p(Y_i, Z_i | X_i, S_i; \theta)}_{\mathcal{M}_{\text{joint}}} \propto \underbrace{p(Y | X, S; \theta_{\text{fair}})}_{\mathcal{M}_{\text{fair}}} \underbrace{p(Z | Y; \theta_{\text{faithful}})}_{\mathcal{M}_{\text{faithful}}}$$

Synthetic labels $\hat{Y}_i$ are sampled from the predictive distribution:

$$\hat{Y}_i \sim p(Y_i | Z_i, X_i, S_i; \hat{\theta}) \propto \underbrace{p(Y | X, S; \hat{\theta}_{\text{fair}})}_{\mathcal{M}_{\text{fair}}} \underbrace{p(Z | Y; \hat{\theta}_{\text{faithful}})}_{\mathcal{M}_{\text{faithful}}}$$

The resulting data $\hat{\mathcal{D}} = \{(\hat{Y}_i, X_i, S_i)\}_{i=1}^n$ achieves controlled fairness-faithfulness tradeoff.

- ▶ Both $\mathcal{M}_{\text{faithful}}$ and $\mathcal{M}_{\text{fair}}$ contribute to the generation synthetic labels, with their relative influence governed by the tuning parameter in θ_{faithful} .
 - ▶ $\mathcal{M}_{\text{fair}}$ enforces fairness by imposing independence between Y and S .
 - ▶ $\mathcal{M}_{\text{faithful}}$ preserves faithfulness by generating Z that carries information about Y .
- ▶ Consequently, θ_{faithful} controls the balance between the two competing goals: preserving faithfulness to the original data and enforcing the desired fairness constraints.

Setting:

- ▶ Observed data: $\mathcal{D} = \{(Y_i, X_i, S_i)\}_{i=1}^n$.
- ▶ Labels Y_i are **continuous** and may be biased / unfair.
- ▶ Goal: generate fair synthetic labels $\hat{Y}_i$ while achieving desired fairness-faithfulness tradeoff.

Fair model $\mathcal{M}_{\text{fair}}$:

$$\mathcal{M}_{\text{fair}} : \quad Y_i = f(X_i, S_i; \beta) + \varepsilon_i, \quad \varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \eta^2), \quad (1)$$

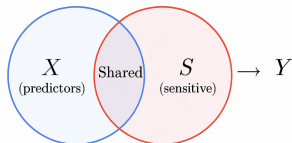
- ▶ $f(\cdot)$: any **fair predictor** chosen so that Y_i under $\mathcal{M}_{\text{fair}}$ satisfies conditional demographic parity $Y \perp S | X$.
- ▶ Simplest choice: **constant mean model**, $f(X_i, S_i; \beta) \equiv \beta_0$.

Our earlier attempt via SDR

Key idea: Use *sufficient dimension reduction* (SDR) to decompose feature space into orthogonal subspaces: $\mathbb{R}^p = \mathcal{S}_{S|X} \oplus \mathcal{S}_{S|X}^\perp$,

$$X = \underbrace{P_{\mathcal{S}_{S|X}} X}_{\text{part of } X \text{ related to } S} + \underbrace{P_{\mathcal{S}_{S|X}^\perp} X}_{\text{part of } X \text{ unrelated to } S}$$

- ▶ $\mathcal{S}_{S|X}$: representing the shared subspace; it contains info about S .
- ▶ $\mathcal{S}_{S|X}^\perp$: orthogonal subspace; information in X that is irrelevant to S .



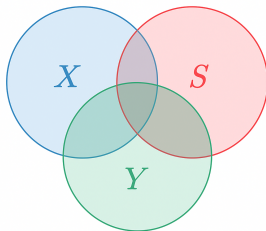
- ▶ We estimate an orthonormal $B \in \mathbb{R}^{p \times r}$, typically $r < p$ such that $\mathcal{S}_{S|X} = \text{span}(B)$ and $S \perp X | B^\top X$.
- ▶ In other words, $(I - BB^\top) X$ encodes information in X that is independent of S .

Reference: Shi et al. (2024), *NAACL 2024*, “Debiasing with sufficient projection: A general theoretical framework for vector representations.”

Our earlier attempt on fairness-utility tradeoff

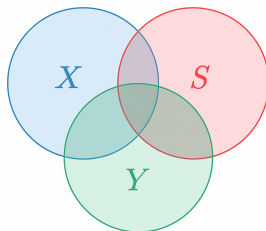
Main idea:

- ▶ Extend the **sufficient projection framework (Shi et al., 2024)** by decomposing X into three orthogonal components: $\mathbb{R}^p = \mathcal{S}_{Y|X} \oplus \mathcal{S}_{S|X} \oplus \mathcal{S}_R$, where
 - ▶ $\mathcal{S}_{Y|X}$: subspace in X predictive of Y ;
 - ▶ $\mathcal{S}_{S|X}$: subspace in X correlated with sensitive attribute S ;
 - ▶ $\mathcal{S}_R$: residual subspace (irrelevant to both).



Reference: Shi et al. (2025), *NeurIPS 2025*, “Understanding Fairness and Prediction Error through Subspace Decomposition and Influence Analysis.”

Our earlier attempt on fairness-utility tradeoff



- ▶ Focus on $\mathcal{S}_{Y|X} \cap \mathcal{S}_{S|X} = \text{Span}\{\phi_1, \dots, \phi_s\}$, which encodes the indirect biases from S to Y through X .
- ▶ Fairness-utility tradeoff can be achieved by **partially fair representation**
 $Z^{(m)} = P^{(m)}X$

$$P^{(m)} = \tilde{B}\tilde{B}^\top + \Phi_m\Phi_m^\top,$$

where $\tilde{B}$ spans $\mathcal{S}_{Y|X} \cap \mathcal{S}_{S|X}^\perp$ (predictive of Y but independent of S), and $\Phi_m = (\phi_1, \dots, \phi_m)$ includes the first m ($m \leq s$) shared directions.

Faithful model $\mathcal{M}_{\text{faithful}}$:

$$\mathcal{M}_{\text{faithful}} : \quad Z_{im} = Y_i + e_{im}, \quad e_{im} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2), \quad m \in [M], \quad (2)$$

- For each Y_i , generate M noisy copies:

$$Z_i = (Z_{i1}, \dots, Z_{iM})^\top.$$

- Z_i retains information about Y_i , with **faithfulness controlled by** the number of copies M and noise variance σ^2 .
- More copies ($M \uparrow$) and/or less noise ($\sigma^2 \downarrow$) $\Rightarrow Z_i$ is more informative about Y_i .

Tuning:

- In practice, fix one of $\{M, \sigma^2\}$ and tune the other.
- Choice guided by user-specified bias reduction parameter $\alpha = \frac{\eta^2}{\eta^2 + \lambda}$, with $\lambda = \frac{\sigma^2}{M}$.

Given estimates $(\hat{\beta}, \hat{\eta}^2, \hat{\sigma}^2)$, we generate synthetic labels from

$$\hat{Y}_i \mid Z_i, X_i, S_i \sim \mathcal{N}(\tilde{\mu}_i, \tilde{\tau}_i^2), \quad (3)$$

where

$$\tilde{\tau}_i^2 = \left(\frac{M}{\hat{\sigma}^2} + \frac{1}{\hat{\eta}^2} \right)^{-1},$$
$$\tilde{\mu}_i = \tilde{\tau}_i^2 \left(\frac{M}{\hat{\sigma}^2} \bar{Z}_i + \frac{1}{\hat{\eta}^2} f(X_i, S_i; \hat{\beta}) \right) \text{ with } \bar{Z}_i = \frac{1}{M} \sum_{m=1}^M Z_{im}.$$

- ▶ $\tilde{\mu}_i$ is a weighted average of:
 - ▶ the fair prediction $f(X_i, S_i; \hat{\beta})$ from $\mathcal{M}_{\text{fair}}$;
 - ▶ the auxiliary mean $\bar{Z}_i$ from $\mathcal{M}_{\text{faithful}}$.
- ▶ M and $\hat{\sigma}^2$ control their relative influence.

► **Fairness guarantee:**

$$\mathcal{UF}(\mathcal{P}_{\hat{\mathcal{D}}|X=x}) \leq \alpha \mathcal{UF}(\mathcal{P}_{\mathcal{D}|X=x}),$$

where $\alpha \in [0, 1]$ is a user-specified bias reduction factor; it determines the extent to which unfairness is mitigated in the synthetic data distribution.

► **Faithfulness guarantee:**

$$W_2(\mu_{\hat{Y}|X,S}, \mu_{Y|X,S}) \leq \sqrt{(1 - \alpha)^2 \mathbb{E} \left[(Y - f(X, S, \beta))^2 | X, S \right] + (1 - \alpha^2) \eta^2}$$

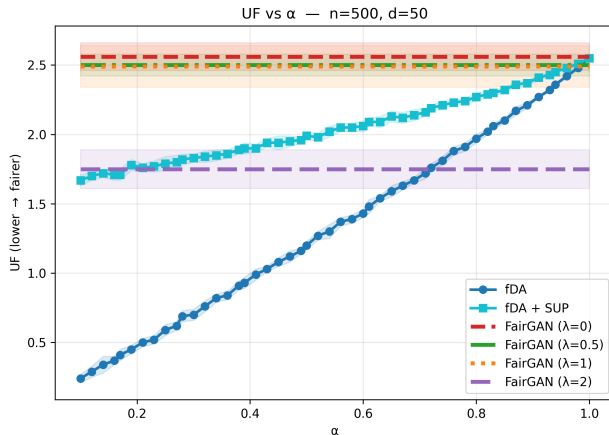
Particularly, when $M \rightarrow \infty$, or $\sigma^2 \rightarrow 0$, we have

$$\hat{Y}|X, S \xrightarrow{P} Y|X, S, \text{ and consequently, } \hat{Y}|X, S \xrightarrow{d} Y|X, S,$$

for any choice of $\mathcal{M}_{\text{fair}}$.

Simulation (continuous Y , $n = 500$ and $d = 50$)

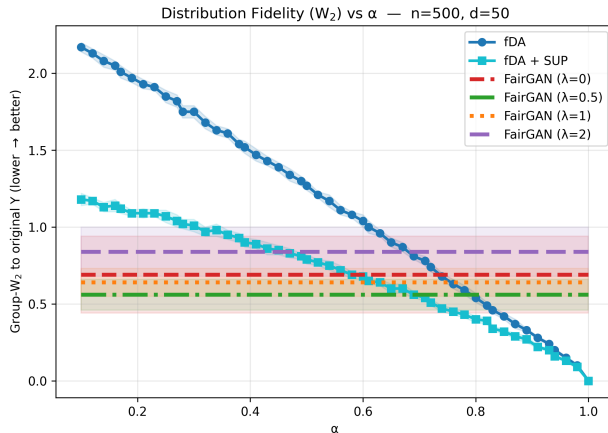
Our **fDA** method, compared with GAN-based approach **FairGAN**: $\lambda = 0$ no fairness constraint.



FairGAN does not provide a stable monotone relationship between λ , unfairness and fidelity.

Simulation (continuous Y , $n = 500$ and $d = 50$)

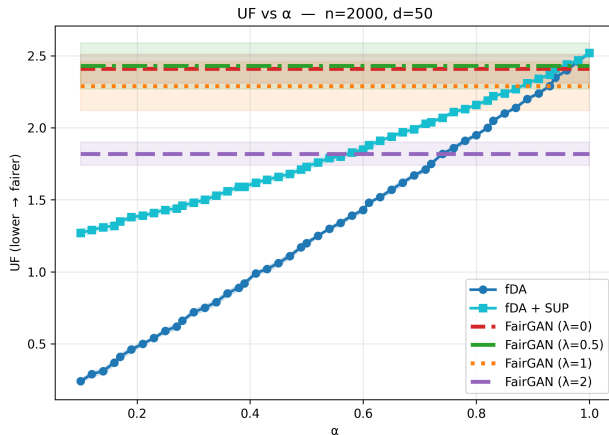
Our **fDA** method, compared with GAN-based approach **FairGAN**: $\lambda = 0$ no fairness constraint.



FairGAN does not provide a stable monotone relationship between λ , unfairness and fidelity.

Simulation (continuous Y , $n = 2000$ and $d = 50$)

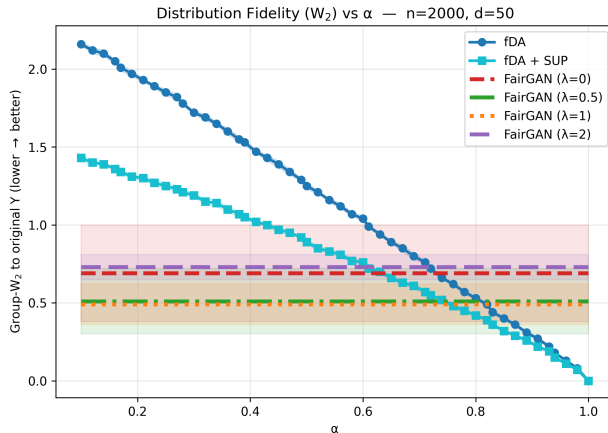
Our **fDA** method, compared with GAN-based approach **FairGAN**: $\lambda = 0$ no fairness constraint.



FairGAN does not provide a stable monotone relationship between λ , unfairness and fidelity.

Simulation (continuous Y , $n = 2000$ and $d = 50$)

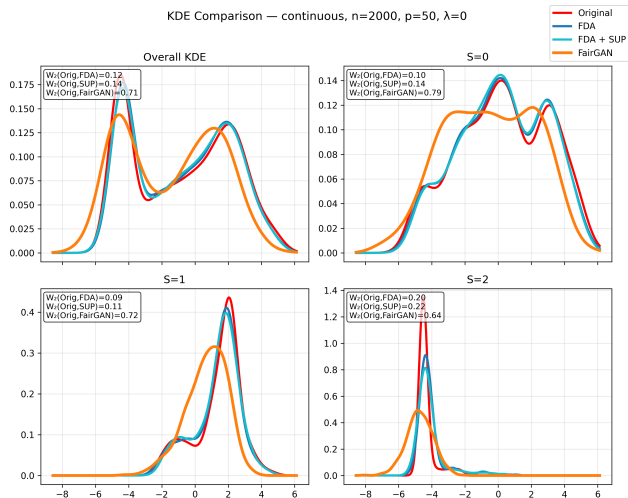
Our **fDA** method, compared with GAN-based approach **FairGAN**: $\lambda = 0$ no fairness constraint.



FairGAN does not provide a stable monotone relationship between λ , unfairness and fidelity.

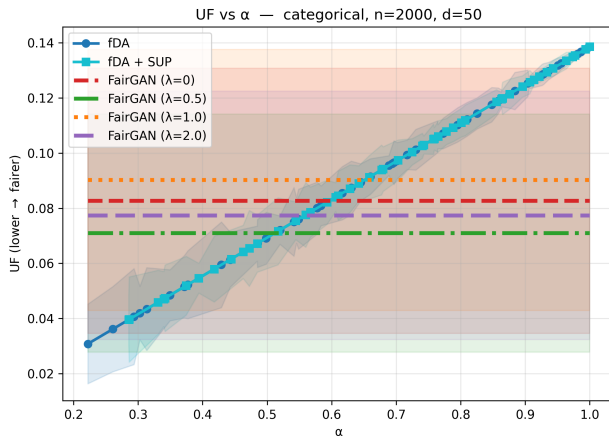
Simulation (continuous Y , $n = 2000$ and $d = 50$)

FairGAN produce synthetic labels $\hat{Y}$ whose distribution deviates substantially from that of the true labels Y even when $\lambda = 0$ (not imposing fairness constraint).



Simulation (categorical Y , $n = 2000$ and $d = 50$)

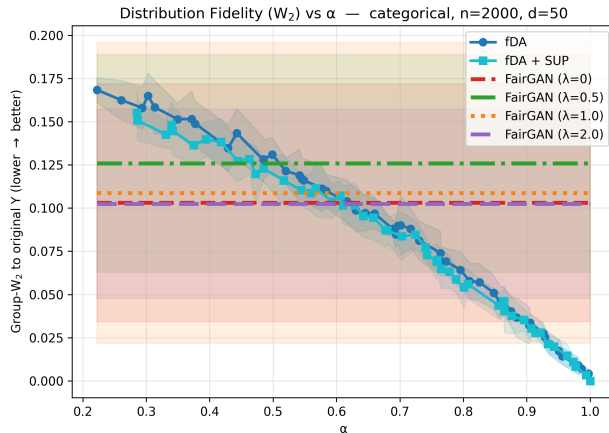
Our **fDA** method, compared with GAN-based approach **FairGAN**: $\lambda = 0$ no fairness constraint.



FairGAN does not provide a stable monotone relationship between λ , unfairness and fidelity.

Simulation (categorical Y , $n = 2000$ and $d = 50$)

Our **fDA** method, compared with GAN-based approach **FairGAN**: $\lambda = 0$ no fairness constraint.



FairGAN does not provide a stable monotone relationship between λ , unfairness and fidelity.

Overview:

- ▶ A widely used benchmark dataset from the U.S. Census Bureau, commonly used for **fairness and bias analysis**.
- ▶ Task: predict whether an individual's income exceeds \$50K/year.

Dataset summary:

- ▶ **Samples:** $\sim 48,842$ individuals.
- ▶ **Features:** 14 attributes (both categorical and continuous), e.g.:
 - ▶ Age, Education, Occupation, Workclass, Capital Gain/Loss, Hours per Week.
- ▶ **Sensitive attributes: Gender (Male/Female).**
- ▶ **Target:** binary income label (Y): Income $> \$50K$ vs. $\leq \$50K$.

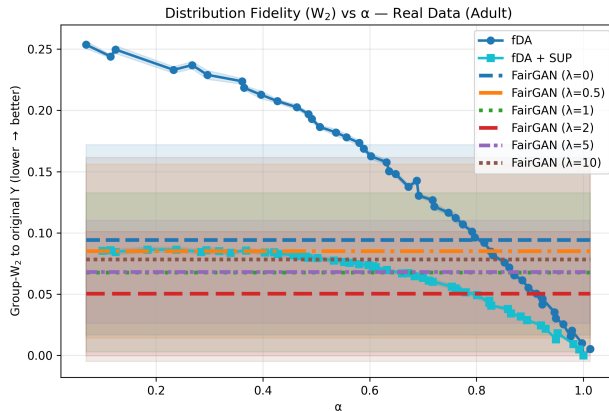
Overview:

- ▶ A widely used benchmark dataset from direct marketing campaigns of a Portuguese banking institution.
- ▶ Task: predict whether a client subscribes to a term deposit after a marketing contact.

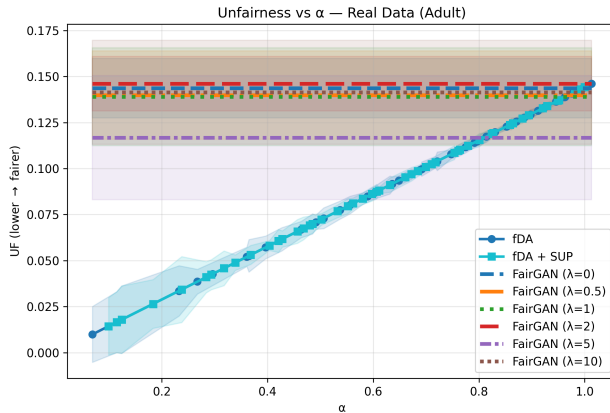
Dataset summary:

- ▶ **Samples:** $\sim 45,211$ clients.
- ▶ **Features:** 16 attributes, including both categorical and continuous variables, e.g.:
 - ▶ Age, Job, Marital Status, Education, Balance, Housing Loan, Personal Loan, Contact Type, Campaign History.
- ▶ **Sensitive attribute:** Age group, e.g. young vs. older clients.
- ▶ **Target:** binary subscription label Y : subscribed to term deposit (yes) vs. did not subscribe (no).

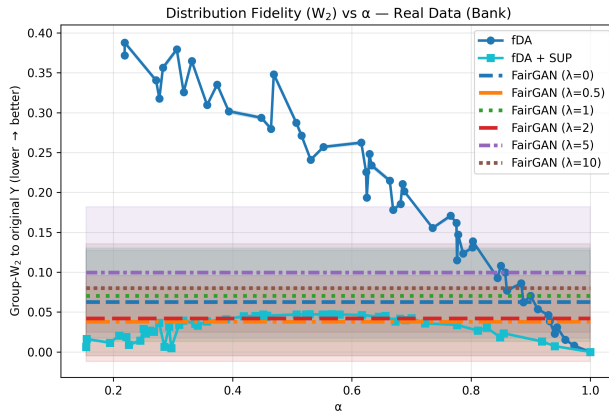
FairGAN can produce synthetic labels $\hat{Y}$ whose distribution deviates substantially from that of the true labels Y even when $\lambda = 0$ (not imposing fairness constraint).



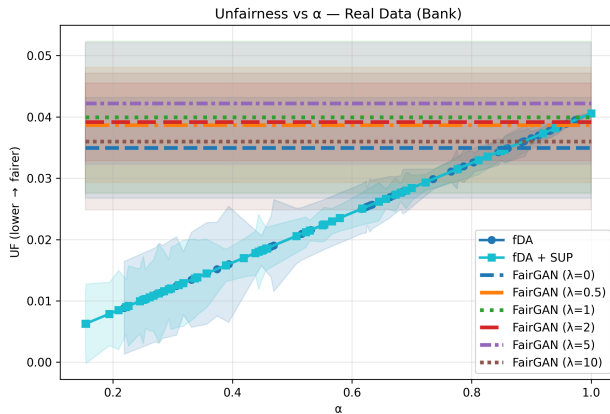
FairGAN does not provide a stable monotone relationship between λ , unfairness and fidelity.



FairGAN can produce synthetic labels $\hat{Y}$ whose distribution deviates substantially from that of the true labels Y even when $\lambda = 0$ (not imposing fairness constraint).



FairGAN does not provide a stable monotone relationship between λ , unfairness and fidelity.



Goal: Generate debiased synthetic data with an **tunable trade-off** between fairness and faithfulness.

- ▶ **Unfairness reduction factor** $\alpha \in [0, 1]$:
 - ▶ $\alpha = 0$: absolute fairness (maximal bias removal).
 - ▶ $\alpha = 1$: perfect faithfulness (synthetic data matches the original).
- ▶ **Compared to GAN-based methods:**
 - ▶ **Simple:** sampling only from basic distributions (Gaussian for continuous labels; Multinomial for categorical).
 - ▶ **No black-box training:** avoids unstable, costly optimization; in our experiments, **fDA** required orders-of-magnitude less computation (e.g., $\sim 10^3 \times$ fewer times).
 - ▶ **Theoretically grounded** and **user-friendly**.
- ▶ **Impact:** facilitates responsible data management and promotes equitable algorithmic outcomes.