Fairness Risks for Group-Conditionally Missing Demographics

Kaiqi Jiang[1] **Wenzhe Fan**[1] **Mao Li**[2] **Xinhua Zhang**[1]
[1] Department of Computer Science, University of Illinois at Chicago, Chicago IL 60607
[2] Amazon

Abstract

Fairness-aware classification models have gained increasing attention in recent years as concerns grow on discrimination against some demographic groups. Most existing models require full knowledge of the sensitive features, which can be impractical due to privacy, legal issues, and an individual's fear of discrimination. The key challenge we will address is the group dependency of the unavailability, e.g., people of some race may be more reluctant to reveal their race. Our solution augments general fairness risks with probabilistic imputations of the sensitive features, while jointly learning the *group-conditionally* missing probabilities in a variational auto-encoder. Our model is demonstrated effective on both image and tabular datasets, achieving an improved balance between accuracy and fairness.

1 Introduction

As machine learning systems rapidly acquire new capabilities and get widely deployed to make human-impacting decisions, ethical concerns such as fairness have recently attracted significant effort in the community. To combat societal bias and discrimination in the model that largely inherit from the training data, a bulk of research efforts have been devoted to addressing group unfairness, where the model performs more favorably (*e.g.*, accurately) to one demographic group than another (Hardt et al., 2016; Dwork et al., 2012; Ye and Xie, 2020; Buolamwini and Gebru, 2018; Giannfrancesco et al., 2018; Mehrabi et al., 2021; Yapo and Weiss, 2018). Group fairness typically concerns both the conventional classification labels (*e.g.*, recidivism) and socially sensitive group features (*e.g.*, gender, or race). As applications generally differ in their context of fairness, a number of quantitative metrics have been developed such as demographic parity, equal opportunity, and equalized odds. A necessarily outdated overview is available at Barocas et al. (2019) and Wikipedia contributors (2024).

Learning algorithms for group fairness can be broadly categorized into pre-, post-, and in-processing methods. Pre-processing methods transform the input data to remove dependence between the class and demographics according to a predefined fairness constraint (Kamiran and Calders, 2012). Post-processing methods warp the class labels (or their distributions) from any classifier to fulfill the desired fairness criteria (Hardt et al., 2016). In-processing approaches, which are most commonly employed including this work, learn to minimize the prediction loss while upholding fairness regularizations simultaneously. More discussions are given in Section 2.

Most of these methods require accessing the sensitive demographic feature, which is often unavailable due to fear of discrimination and social desirability (Krumpal, 2013), privacy concerns, and legal regulations (Coston et al., 2019; Lahoti et al., 2020). For example, people of some gender may be more inclined to withhold their gender information when applying for jobs dominated by other genders. Recently several methods have been developed for this data regime. Lahoti et al. (2020) achieved Rawlsian max-min fairness by leveraging computationally-identifiable errors in adversarial reweighted learning. Hashimoto et al. (2018) proposed a distributionally robust optimization to minimize the risk over the worst-case group distribution. However, they are not effective in group fairness and cannot be customized for different group fairness metrics. Yan et al. (2020) infers groups by clustering the data, but there is no guarantee that the uncovered groups are consistent with the real sensitive features of interest. For example, the former may identify race while we seek to be fair in gender.

We aim to tackle these issues in a complementary setting, where demographics are *partially* missing. Al-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

though their availability is limited, it is still often feasible to obtain a *small* amount of labeled demographics.[1] For example, some people may not mind disclosing their race or gender. Therefore, different from the aforementioned works that assume complete unavailability of demographics, we study in this paper the *semi-supervised* setting where they can be partially available at random in both training and test data. Similarly, the class labels can be missing in training.

Semi-supervised learning (SSL) has been well studied, and can be applied to impute missing demographics. Zhang and Long (2021) simply used the labeled subset to estimate the fairness metrics, but their focus is on analyzing its estimation bias instead of using it to train a classifier. Jung et al. (2022) assigned pseudo group labels by training an auxiliary group classifier, and assigned low-confidence samples to random groups. However, the confidence threshold needs to be tuned based on the conditional probability of groups given the class label, which becomes challenging as the latter is also only partially available in our setting.

Dai and Wang (2021) used SSL on graph neural networks to infer the missing demographics, but rounded the predictions to binary group memberships. This is suboptimal because, due to the scarcity of labeled data, there is marked *uncertainty* in demographic imputations. As a result, the common practice of "rounding" sensitive estimates—so that fairness metrics defined on categorical memberships can be enforced instead of mere independence—may over-commit to a group just by chance, dropping important uncertainty information when the results are fed to downstream learners. Moreover, most applications do not naturally employ an underlying graph, and the method only enforces min-max (i.e., adversarial) fairness instead of group fairness metrics.

Our **first** contribution, therefore, is to leverage the *probabilistic* imputation by designing a differentiable fairness risk in Section 4, such that it is customized for a *user-specified* fairness metric with discrete demographics as opposed to simple independence relationships, and can be directly integrated into general SSL algorithms to regularize the learned posterior towards low risks in both classification and fairness.

Our method is **distinguished** from generic SSL recipes which predict unobserved protected attributes and then optimize the fairness risk. Performing the two steps separately prevents the common backbone features to be synthesized that concurrently facilitate identifying demographics and class labels. This could be ameliorated via a bi-level optimization, but its computational cost can be high. We hence resort to flattening the two levels into one joint optimization, and then address the pathological phenomenon where the group memberships are also "learned" to promote fairness.

Interestingly, we found an efficient solution by stopping the gradient of fairness risk with respect to the group estimates. Furthermore, to effectively implement this principle, we employed a Monte-Carlo evaluation of the risk, which significantly accelerates inference and differentiation via a provable $O(1/\epsilon^2)$ sample complexity. So our **second** contribution hits two birds with one stone, as detailed in Section 4.3.

As our **third** contribution, we instantiated the semi-supervised classifier with a new encoder and decoder in a variational autoencoder (VAE, Narayanaswamy et al., 2017; Kingma et al., 2014), allowing **group-conditional missing demographics**, which has so far only been considered for noisy but not missing demographics. In general, the chance of unavailability does depend on specific groups – if people of a certain race are aware of the unfairness against them, they will be more reluctant to disclose their race. The comprehensive model will be introduced in Section 5.

Our **fourth** contribution is to demonstrate, in Section 6, that our method empirically outperforms state of the art for fair classification where both the demographics and class labels are only partially available. The paper is primed with preliminaries in Section 3.

2 Related Work

Noisy demographics have been tackled by, *e.g.*, Lamy et al. (2019); Celis et al. (2021a). Adversarially perturbed and privatized demographics were studied by Celis et al. (2021b) and Mozannar et al. (2020), which also account for group-conditional noise. However, although they conceptually subsume unavailability as a type of noise, their method and analysis do not carry through. Some of these methods, along with Wang et al. (2020), also require an auxiliary dataset to infer the noise model, which is much harder for missing demographics. Shah et al. (2023) dispenses with such a dataset, but its favorable theoretical properties apply only to Gaussian data, while the bootstrapping-based extension to non-Gaussian data does not model the conditional distribution of demographics.

Another strategy to tackle missing demographics resorts to proxy features (Gupta et al., 2018; Chen et al., 2019; Kallus et al., 2022). Based on domain knowledge, they are assumed to correlate with and allude to the sensitive feature in question. Zhao et al. (2022) optimizes

[1] According to Consumer Financial Protection Bureau (2023), "A creditor shall not inquire about the race, color, religion, national origin, or sex of an applicant." Exceptions include inquiry "for the purpose of conducting a self-test that meets the requirements of §1002.15", which essentially tests if the fair lending rules are conformed with.

the correlations between these features and the prediction, but it is not tailored to the specific group fairness metric in question. This is difficult because it relies on the binary group memberships which is not modeled by the method. Zhu et al. (2023) considered the accuracy of estimating the fairness metrics using proxies, but did not demonstrate its effectiveness when applied to train a fair classifier. In addition, proxy features can often be difficult to identify. In face recognition where the sensitive feature is age, race, or wearing glasses, raw pixels are no good proxies and identifying semantic features as proxies can be challenging.

3 Preliminary

We first consider supervised learning with group fairness, where samples are drawn from a distribution D over $\mathcal{X} \times \{0,1\} \times \{0,1\}$, with $(X, A, Y) \sim D$. Here X is the non-protected feature, Y is the binary label, and A is the binary sensitive feature which is also referred to as the group demographics. Our method can be extended to **multi-class** labels and sensitive features, under any fairness risk defined for fully observed A and Y. The details are deferred to Appendix A.1. Since our focus is on addressing missing demographics, we will stick with the binary formulation for ease of exposition.

We follow the setup in Menon and Williamson (2018) and aim to find a measurable *randomized classifier* parameterized by a function $f : \mathcal{X} \to [0,1]$ (*e.g.*, a neural network), such that it predicts an $X \in \mathcal{X}$ to be positive with probability $f(X)$. That is, the predicted label $\hat{Y} \in \{0,1\}$ follows the Bernoulli distribution $\hat{Y}|X \sim \text{Bern}(f(X))$. We denote the conditional distribution as $P_f(\hat{Y}|X)$, and omit the subscript f for brevity when the context is clear. This setup also subsumes a classifier based on a *class-probability estimator*, where $f(X) \in [0,1]$ is translated to a hard deterministic label via a learnable threshold τ: $P_f(\hat{Y} = 1|X) = [\![f(X) > \tau]\!]$. Here, the Iverson bracket $[\![\cdot]\!] = 1$ if $\cdot$ is true, and 0 otherwise.

Suppose we have a training set $\mathcal{D}_{tr} := \{\mathbf{x}_i, y_i, a_i\}_{i=1}^n$ with size n, where $\mathbf{x}_i \in \mathcal{X}$ is the nonsensitive feature of the i-th example, y_i is its label, and a_i is its sensitive feature. In a completely observed scenario, both y_i and a_i are 0 or 1, and we will later extend it to missing values. As a shorthand, let $\mathbf{a} = (a_1, \ldots, a_n)^\top$ and $\mathbf{y} = (y_1, \ldots, y_n)^\top$. We first review some standard group fairness metrics:

Demographic parity (Dwork et al., 2012):

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1), \tag{1}$$

Equalized odds (Hardt et al., 2016): for $y \in \{0,1\}$

$$P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y), \tag{2}$$

Equal opportunity (Hardt et al., 2016):

$$P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1). \tag{3}$$

In general, enforcing perfect fairness can be too restrictive, and approximate fairness can be considered via fairness measures. For example, the mean difference compares their difference (Calders and Verwer, 2010)

$$\text{MD}(f) = P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1). \tag{4}$$

and the disparate impact (DI) factor computes their ratio (Feldman et al., 2015). Similar measures can be defined for equal opportunity and equalized odds. For continuous variables, χ^2-divergence measures the independence as stipulated by separation and independence (Mary et al., 2019).

Given a dataset $\mathcal{D}_{tr}$ with n examples, we can empirically estimate both sides of (1) by, *e.g.*, for $a \in \{0,1\}$

$$\begin{aligned} P(\hat{Y} = 1|A = a) &\approx \text{mean}(\{P_f(\hat{Y} = 1|x_i) : a_i = a\}) \\ &= \sum\nolimits_{i:a_i=a} P_f(\hat{Y} = 1|x_i) \Big/ \sum\nolimits_{i:a_i=a} 1. \end{aligned} \tag{5}$$

Here, for a finite set S, we denote the mean of its elements as mean (S). The above estimator is straightforward from Eq 1 and 7 of Menon and Williamson (2018) in the setting of randomized classifier. It is asymptotically unbiased, consistent, and **differentiable** in f. see the proof in Appendix E. Similarly, we can estimate $P(\hat{Y} = 1|A = a, Y = y) \approx \text{mean}\,(S)$, where $S := \{P_f(\hat{Y} = 1|x_i) : a_i = a, y_i = y\}$. Other estimators can also be adopted, such as kernel density estimation (Cho et al., 2020). Applying these estimates to MD, DI, χ^2-divergence, or any other metric, we obtain a fairness risk $\mathcal{F}(P_f, \mathbf{y}, \mathbf{a})$. We keep its expression general, and some concrete examples are given in (18) to (20) in Appendix A.1. An **example** based on (5) is

$$\begin{aligned} \mathcal{F}(P_f, \mathbf{y}, \mathbf{a}) := |&\text{mean}(\{P_f(\hat{Y} = 1|x_i) : a_i = 1\}) \\ &- \text{mean}(\{P_f(\hat{Y} = 1|x_i) : a_i = 0\})|. \end{aligned}$$

Classification risk. We denote the conventional classification risk as $\mathcal{R}(P_f, \mathbf{y})$. For example, the standard cross-entropy loss yields $-\frac{1}{n}\sum_i \log P_f(\hat{Y} = y_i|X = x_i)$. Combined with fairness risk, we can next find the optimal f by minimizing their sum weighted by $\lambda > 0$:

$$\mathcal{F}(P_f, \mathbf{y}, \mathbf{a}) + \lambda \cdot \mathcal{R}(P_f, \mathbf{y}). \tag{6}$$

In Williamson and Menon (2019), a more general framework of fairness risk is constructed, incorporating the loss ℓ into the definition of $\mathcal{F}$ itself (Lahoti et al., 2020). Our method developed below can be applied directly.

4 Fairness with Group-Conditionally Unavailable Demographics

The regularized objective (6) requires the demographic features which can often become unavailable due to a respondent's preference, privacy, and legal reasons. As a key contribution of this work, we further address **group-conditionally** unavailable demographics, where demographics can get unavailable in a *non-uniform* fashion, depending on the specific group. This matches reality. For example, people above some age may be more reluctant to reveal it when applying entry-level jobs.

We show that this new setting can be approached by extending the objective (6). To this end, we introduce a new random variable $\tilde{A}$, which is the observation of the true latent demographic A. For a principled treatment, we treat A as **never observed**, while $\tilde{A}$ can be equal to A, or take a distorted value, or be unavailable (denoted as $\emptyset$).[2] In other words, $\tilde{A}$ is **always observed**, including taking the value of $\emptyset$. As Section 5 shows, this novel model offers substantial flexibility, even if the demographic contains multiple features and a different subset of them is unavailable for different individuals.

We model the group-conditional unavailability with a learnable noising probability $P(\tilde{A}|A)$ (Yao et al., 2023). As a special case, one may assert that demographics cannot be misrepresented, i.e., $P(\tilde{A}|A) = 0$ for all $\tilde{A} \notin \{A, \emptyset\}$. In order to address isometry (e.g., totally swapping the concept of male and female), we impose some prior on $P(\tilde{A}|A)$. For example, $P(\tilde{A}|A)$ must be low for $\tilde{A} \notin \{A, \emptyset\}$. This can be enforced by a Dirichlet prior, e.g., $(P(\text{male}|\text{male}), P(\text{female}|\text{male}), P(\emptyset|\text{male})) \sim \text{Dir}(0.5, 0.1, 0.4)$. Similarly, we can define $P(\tilde{Y}|Y)$ for class-conditionally unavailable label, allowing $\tilde{Y} = \emptyset$.

In the sequel, we will assume there is an SSL algorithm that predicts Y and A with probabilistic models $P_f(Y|X, \tilde{Y})$ and $q_\phi(A|X, \tilde{A})$, respectively. For example, the encoder of a VAE, which will be detailed in Section 5 as (11). Both A and Y can be **multi-class**. It is now natural to extend the fairness risk by taking the *expectation* of these missing values:

$$\mathcal{E}_{\text{vanilla}}(P_f, q_\phi) \qquad (7)$$
$$:= \underset{Y_i \sim P_f(Y|x_i, \tilde{y}_i)}{\mathbb{E}} \; \underset{A_i \sim q_\phi(A|x_i, \tilde{a}_i)}{\mathbb{E}} \mathcal{F}(P_f, Y_{1:n}, A_{1:n}).$$

In the case of no misrepresentation, the expectation of A_i can be replaced by $A_i = \tilde{a}_i$ if $\tilde{a}_i \neq \emptyset$. Likewise for Y_i. Adding this vanilla semi-supervised regularizer to an existing learning objective of P_f and $q_\phi(A|X, \tilde{A})$ such as VAE gives a straightforward promotion of fairness, and we will also discuss its optimization in Section 4.4. However, despite its clear motivation, we now demonstrate that it is indeed plagued with conceptual and practical issues, which we will address next.

4.1 Rationalizing semi-supervised fairness risk

It is important to note that the onus of fairness is only supposed to be on the classifier P_f, while $q_\phi(A|X, \tilde{A})$—which infers the demographics to define the fairness metric itself—is supposed to be *recused* from fairness risk minimization. Reducing the risk by manipulating an individual's gender is not reasonable. This issue does not exist in a fully supervised setting, but complicates the regularizer when $q_\phi(A|X, \tilde{A})$ is *jointly* optimized with the classifier P_f. For example, although the risk can be reduced by making A independent of X, this should not be intentionally pursued unless the underlying data distribution supports it. The posterior $q_\phi(A|X, \tilde{A})$ is only supposed to accurately estimate the demographics, while the task of fairness should be left to the classifier P_f.

A natural workaround is a two-step approach: first train a semi-supervised learner for the distribution $A|X, \tilde{A}$, and denote it as $r(A|X, \tilde{A})$. Then in (7), we replace $A_i \sim q_\phi(A|x_i, \tilde{a}_i)$ with $A_i \sim r(A|x_i, \tilde{a}_i)$. Although this resolves the above issue, it decouples the learning of $A|X, \tilde{A}$ (first step) and $Y|X, \tilde{Y}$ (second step). This is not ideal because, as commonly recognized, the inference of A and Y can benefit from *shared* backbone feature extractors on X (*e.g.*, ResNet for images), followed by finetuning target-specific heads.

However, sharing backbones will allow $q_\phi(A|X, \tilde{A})$ to be influenced by the learning of $P_f(Y|X, \tilde{Y})$, conflicting with the aforementioned recusal principle. In light of this difficulty, we resort to the commonly used stop-gradient technique that is available in PyTorch (`detach`) and TensorFlow (`tf.stop_gradient`). While they still share backbones, the derivative of $\mathcal{E}_{\text{vanilla}}$ in (7) with respect to $q_\phi(A|X, \tilde{A})$ is stopped (set to 0) from backpropagation, *i.e.*, treating $q_\phi(A|X, \tilde{A})$ as a constant. It also much simplifies the training process. The stop-gradient method is a technique that blocks the flow of gradient through a specific part of the computational graph during backpropagation, which is commonly used to decouple certain computations from parameter updates. We present the corresponding results in our ablation study, shown by the blue curve among all experiment figures.

[2] We intentionally term it "unavailable" instead of "missing", because *missing* means unobserved in statistics, while we take *unavailable* (i.e., $\emptyset$) as an *observed* outcome.

4.2 Imputation of unavailable training labels

A similar issue is also present in the expectation $Y_i \sim P_f(Y|x_i, \tilde{y}_i)$ in (7). The ground truth label $Y_{1:n}$ is given in the supervised setting, but its missing values should *not* be inferred to account for fairness, because labels are part of the fairness definition itself, except for demographic parity. We could resort to the stop-gradient technique again to withhold the backpropagation through $Y_i \sim P_f(Y|x_i, \tilde{y}_i)$. However, different from unavailable demographics, the situation here is more intricate because P_f also serves as the first argument of $\mathcal{F}(P_f, Y_{1:n}, A_{1:n})$, where it indeed enforces fair classification. This conflicts with the fairness-oblivious requirement for imputing Y_i, which is also based on P_f. In other words, we cannot enforce P_f to both respect and disregard fairness at the same time.

To address this issue, we will train a *separate* semi-supervised classifier to impute Y_i without fairness concerns, and it does leverage the backbone features mentioned above. Suppose $\mathbf{h}_i$ is the current feature representation for x_i, e.g., the third last layer of f. Assuming there is no misrepresentation in $\tilde{y}_i$, we can clamp $Y|x_i$ to P_f for those examples with observed label ($\tilde{y}_i \neq \emptyset$). Then we infer a class probability $P_g(Y|x_i)$ for the rest examples by using $\mathbf{h}_i$ and any SSL algorithm such as graph-based Gaussian field (Zhu et al., 2003). This amounts to a fairness risk as

$$\boxed{\mathcal{E}(P_f, q_\phi)} \tag{8}$$
$$:= \mathop{\mathbb{E}}_{\{Y_i \sim P_g(Y|x_i): \tilde{y}_i = \emptyset\}} \ \mathop{\mathbb{E}}_{A_i \sim q_\phi(A|x_i, \tilde{a}_i)} \mathcal{F}(P_f, Y_{1:n}, A_{1:n}).$$

Note we stop the gradient on P_g and q_ϕ here, and P_g is used instead of P_f. Appendix A.7 extends this risk from randomized classifier to class-probability estimator.

4.3 Efficient evaluation of fairness risk $\mathcal{E}$

A major challenge in risk-based methods is the cost of computing the expectations in $\mathcal{E}$, along with its derivatives. In the sequel, we will resort to a Monte Carlo based method, such that for n training examples, an ϵ accurate approximation with confidence $1 - \delta$ can be found with $O(\frac{c_n}{\epsilon^2} \log \frac{1}{\delta})$ computation, where c_n is the cost of evaluating $\mathcal{F}$ and it is $O(n)$ in our considered cases. Although some customized algorithms can be developed by exploiting the specific structures in $\mathcal{F}$ such as demographic parity, we prefer a more general approach. Despite the inevitable inexactness in the result, we prove tight concentration bounds that turn out sufficiently accurate in practice.

Our method simply draws N number of iid samples from $A_i \sim q_\phi(A|x_i, \tilde{a}_i)$ and $Y_i \sim P_g(Y|x_i)$. For $s = 1, \ldots, N$, let $Z_s := \{a_i^{(s)}, y_i^{(s)}\}_{i=1}^n$. We estimate $\mathcal{E}$ by

$$\hat{\mathcal{E}}_n(Z_1, \ldots, Z_N) := \tfrac{1}{N} \sum\nolimits_{s=1}^{N} \mathcal{F}(P_f, Z_s). \tag{9}$$

We prove the following sample complexity for this estimator by leveraging McDiarmid's inequality.

Theorem 1. *Suppose $\mathcal{F} \in [0, C]$ where $C > 0$ is a constant. Then for all $\epsilon > 0$,*

$$P(|\hat{\mathcal{E}}_n(Z_1, \ldots, Z_N) - \mathcal{E}| \geq \epsilon) \leq \ 2\exp(-2N\epsilon^2/C^2).$$

As a result, to guarantee an estimation error of ϵ with confidence $1 - \delta$, it suffices to draw $N = \frac{C^2}{2\epsilon^2} \log \frac{1}{\delta}$ samples. The proof is available in Appendix A.2.

This theorem is non-trivial because the fairness risk can depend on the predicted probabilities in a complex manner, creating nonlinear dependencies among all the training examples; see Appendix A.1. The analysis also reveals the limitations of the sampler, as certain risks, such as disparate impact, are unbounded.

As $\hat{\mathcal{E}}_n(Z_1, \ldots, Z_N)$ costs $O(nN)$ to compute, the total cost is $\frac{nC^2}{2\epsilon^2} \log \frac{2}{\delta}$. When $C = 1$, $\epsilon = 0.01$ and $n = 10^4$, this is order of 10^8, which is quite affordable. In experiments, a sample size **as low as** $\underline{N = 100}$ was sufficient for our method to outperform the state of the art; see Section 6. Most risks we consider satisfy the boundedness assumption, although exceptions exist such as disparate impact.

4.4 Differentiation of vanilla fairness risk

Since we stop the gradient with respect to $q_\phi(A|x_i, \tilde{a}_i)$ and $P_g(Y|x_i)$, the whole regularizer $\mathcal{E}(P_f, q_\phi)$ in (8) can be easily differentiated with respect to f.

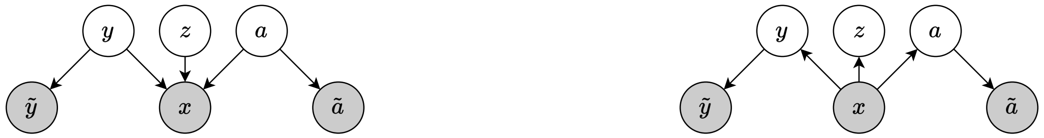
In practice, we also would like to conduct an ablation study by comparing with $\mathcal{E}_{\text{vanilla}}$, where challenges arise from differentiation with respect to $q_\phi(A|x_i, \tilde{a}_i)$ and $y_i \sim P_f(Y|x_i, \tilde{y}_i)$. We resolve this issue by using the straight-through Gumbel-Softmax method (Jang et al., 2017; Maddison et al., 2017). A self-contained description is given in Appendix A.3.

5 Integrating Fairness Risk with SSL

We next illustrate how the fairness risk can be integrated with SSL. As an example, we recap the semi-supervised VAE (SS-VAE, Kingma et al., 2014) with missing a and y values.[3]

Why VAE? In terms of SSL, while some discriminative methods exist such as Zhu et al. (2003), the most

[3] It is customary in statistics to denote random variables by capital letters. However, the VAE literature generally uses lowercase letters. We thus switch to their custom.



(a) Decoder (b) Encoder

Figure 1: SS-VAE decoder and encoder with unavailable demographic/label conditioned on group/class

prevalent and effective approaches are based on generative models, such as SS-VAE. Indeed, most conventional generative models for SSL are essentially VAEs, especially when they are trained with variational inference.

Secondly, our approach to fairness risk relies on an encoder capable of estimating the posterior probability of missing labels or demographics. This rules out normalizing flows or generative adversarial networks unsuitable, as the latter's discriminator does not serve as an encoder. Moreover, transformers are not considered since our formulation does not involve sequential data modeling.

In contrast, VAE employs an encoder and provides the conditional independencies that allow us to more finely specify the constituent distributions as in (10) and (11). Diffusion models, as a more refined variant of VAE, can also be adopted. However, we intentionally keep the generative model simple, prioritizing fairness modeling, while future work can incorporate additional refinements to the VAE.

5.1 Encoders and decoders

SS-VAE employs a decoder/generation process and an encoder/inference process parameterized by θ and ϕ respectively. In the decoder whose graphical model is shown in Figure 1a, the observation $\mathbf{x}$ is generated conditioned on the latent variable $\mathbf{z}$, latent class y and latent demographics a:

$$\begin{aligned} p_\theta(\mathbf{x}, \tilde{a}, \tilde{y} | a, y, \mathbf{z}) &= \mathcal{N}(\mathbf{x} | g_\theta(a, y, \mathbf{z})) \cdot P_\theta(\tilde{a}|a) \cdot P_\theta(\tilde{y}|y), \\ p(y) &= \mathrm{Cat}(y|\pi_y), \quad p(\pi_y) = \mathrm{SymDir}(\gamma_y), \\ p(a) &= \mathrm{Cat}(a|\pi_a), \quad p(\pi_a) = \mathrm{SymDir}(\gamma_a), \\ p(\mathbf{z}) &= \mathcal{N}(\mathbf{z}|0, I). \end{aligned} \tag{10}$$

Here $\mathrm{Cat}(y|\pi_y)$ is the multinoulli prior of the class variable. SymDir is the symmetric Dirichlet with hyperparameter γ_y. The prior of a is analogously defined.

The inference process finds conditionally independent representations $\mathbf{z}$, a, and y under a given $\mathbf{x}$, $\tilde{a}$, and $\tilde{y}$. Accordingly, the approximate posterior/encoder $q_\phi(a, y, \mathbf{z}|\mathbf{x}, \tilde{a}, \tilde{y})$ factors as in Figure 1b:

$$\begin{aligned} q_\phi(a, y, \mathbf{z}|\mathbf{x}, \tilde{a}, \tilde{y}) &= q_\phi(\mathbf{z}|\mathbf{x}) \cdot q_\phi(y|\mathbf{x}, \tilde{y}) \cdot q_\phi(a|\mathbf{x}, \tilde{a}) \\ \text{where} \quad q_\phi(\mathbf{z}|\mathbf{x}) &= \mathcal{N}(\mathbf{z}|\mu_\phi(\mathbf{x}), \sigma^2_\phi(\mathbf{x})), \\ q_\phi(y|\mathbf{x}, \tilde{y}) &\propto q_\phi(y|\mathbf{x}) \cdot P_\theta(\tilde{y}|y) \\ &= \mathrm{Cat}(y|g_\phi(\mathbf{x})) \cdot P_\theta(\tilde{y}|y), \\ q_\phi(a|\mathbf{x}, \tilde{a}) &\propto q_\phi(a|\mathbf{x}) \cdot P_\theta(\tilde{a}|a) \\ &= \mathrm{Cat}(a|h_\phi(\mathbf{x})) \cdot P_\theta(\tilde{a}|a). \end{aligned} \tag{11}$$

Here $\mu_\phi, \sigma_\phi, g_\phi, h_\phi$ are all defined as neural networks. Compared with Figure 1a, here we only reversed the arrows connected with $\mathbf{x}$, following a standard assumption in VAE that, in posterior, y, z, a are independent given $\mathbf{x}$. We maximally preserved the other arrows, namely $y \to \tilde{y}$ and $a \to \tilde{a}$. The chain of $\mathbf{x} \to a \to \tilde{a}$ can be interpreted as first probabilistically determining a based on $\mathbf{x}$, and then adding noise to it producing $\tilde{a}$ (including $\emptyset$). This chain structure allows us to derive $q_\phi(a|\mathbf{x}, \tilde{a})$ in (11) (and analogously $q_\phi(y|\mathbf{x}, \tilde{y})$) as:

$$\begin{aligned} p(a|\mathbf{x}, \tilde{a}) &\propto p(a)p(\mathbf{x}, \tilde{a}|a) = p(a)p(\mathbf{x}|a)p(\tilde{a}|a) \\ &= p(\mathbf{x})p(a|\mathbf{x})p(\tilde{a}|a) \propto p(a|\mathbf{x})p(\tilde{a}|a). \end{aligned}$$

In training, the fairness risks in (7) and (8) require $P_f(y|\mathbf{x}, \tilde{y})$, which can be served by $q_\phi(y|\mathbf{x}, \tilde{y})$.

5.2 Instilling Fairness to SS-VAE

The evidence lower bound (ELBO) of $\log p_\theta(\mathbf{x}, \tilde{a}, \tilde{y})$ can be derived in a standard fashion, and we relegate the details to Appendix A.4. Denoting it as $\mathrm{ELBO}(\mathbf{x}, \tilde{a}, \tilde{y})$, we extend the SS-VAE objective as follows, to be minimized over θ and ϕ:

$$\begin{aligned} \mathcal{L}(\theta, \phi) = \Omega(P_\theta(\tilde{a}|a), P_\theta(\tilde{y}|y)) \\ - \mathop{\mathbb{E}}_{(\mathbf{x}, \tilde{a}, \tilde{y}) \sim \mathcal{D}_{tr}} \Big[\mathrm{ELBO}(\mathbf{x}, \tilde{a}, \tilde{y}) \\ + [\![\tilde{a} \neq \emptyset]\!] \cdot \log \sum\nolimits_a \mathrm{Cat}(a|h_\phi(\mathbf{x})) P_\theta(\tilde{a}|a) \\ + [\![\tilde{y} \neq \emptyset]\!] \cdot \log \sum\nolimits_y \mathrm{Cat}(y|g_\phi(\mathbf{x})) P_\theta(\tilde{y}|y) \Big]. \end{aligned} \tag{12}$$

Here, the regularizer Ω enforces the Dirichlet prior. For example, let $a = P(\text{male}|\text{male})$, $b = P(\text{female}|\text{male})$, $c = P(\emptyset|\text{male})$. Then a prior of $\mathrm{Dir}(0.5, 0.2, 0.4)$ leads to a regularizer of $\log(a^{0.5-1} b^{0.2-1} c^{0.4-1})$. The regularization is beneficial when prior knowledge about the distributions is available, such as the likelihood that a male is more likely to be labeled as male rather than female. However our experiments did not make any such assumptions and thus did not incorporate this regularizer. Suppose $\tilde{y} \neq \emptyset$. If the provided $\tilde{y}$ always truthfully represents y, then $P(\tilde{y}|y) = [\![\tilde{y} = y]\!]$. As a result, the last line in (12) equals $\mathrm{Cat}(\tilde{y}|g_\phi(\mathbf{x}))$, the standard supervised loss in SS-VAE.

Casting SS-VAE into the fairness risk framework, the role of f is played by $q_\phi(y|\mathbf{x})$ and we only need to augment the objective (12) into

$$\boxed{\mathcal{L}(\theta, \phi) + \lambda \cdot \mathcal{E}(q_\phi(y|\mathbf{x}), q_\phi(a|\mathbf{x})),} \tag{13}$$

where $\lambda > 0$ is a tradeoff hyperparameter and $\mathcal{E}$ takes the same form as in (8). We will henceforth refer to this model as **Fair-SS-VAE**.

Classifying test data. We simply predict by

$$P_f(y|\mathbf{x}) := q_\phi(y|\mathbf{x}) = \mathrm{Cat}(y|g_\phi(\mathbf{x})), \tag{14}$$

with no access to $\tilde{a}$ (Lipton et al., 2018). A more principled approach is to minimize the risk of fairness and classification, which is relegated to Appendix A.6.

Note our new VAE can readily extend y and a from binary values to **multi-class** and **continuous**, with little impact on Monte Carlo sampling in (9). So our Fair-SS-VAE is applicable wherever the underlying fairness risk is available; see Appendix A.1.

6 Experimental Results

We now show empirically that with group-conditionally missing demographics and generally missing labels, our Fair-SS-VAE significantly outperforms various state-of-the-art semi-supervised fair classification methods in terms of various fairness metrics, while keeping a similar accuracy.

Datasets We used two datasets to create *three* sets of experiments:

CelebA (Liu et al., 2015). CelebA contains face images annotated with 40 binary attributes. We sampled 45k images as our training and validation dataset, and 5k as test set. Following (Jung et al., 2022), we set "Attractive" as the target label, and "Gender" as the sensitive group attribute.

UCI Adult (Becker and Kohavi, 1996). The Adult dataset is a tabular dataset where the target label is whether the income exceeds \$50K per year given a person's attributes. We conducted two experiments with **gender** and **race** serving the sensitive attribute. The same pre-processing routine as in Bellamy et al. (2018) was adopted.

It is worth mentioning that conducting experiments where there is a significant correlation between target labels and groups is necessary. If the labels and groups are already uncorrelated, fairness modeling is likely less critical. For example, in CelebA, when considering a label/group pair with very low correlation, such as 'Eyeglasses' and 'Male', the model already achieves high accuracy with a DEO close to zero (ranging from 0.000 to 0.003) even without any fairness interventions. This suggests that fairness-aware models, including ours, have limited potential for improvement, making them less ideal for experiments.

To simulate the missing data, we randomly masked 25% labels, and masked the demographics as:

$$P(\tilde{A}{=}\emptyset|A{=}1) = \alpha, \quad P(\tilde{A}{=}1|A{=}1) = 1{-}\alpha,$$
$$P(\tilde{A}{=}\emptyset|A{=}0) = \beta, \quad P(\tilde{A}{=}0|A{=}0) = 1{-}\beta.$$

We considered three levels of missing demographics: **sparse** ($\alpha = 0.4, \beta = 0.8$), **medium** ($\alpha = 0.2, \beta = 0.4$), and **dense** ($\alpha = 0.1, \beta = 0.2$). Fair-SS-VAE can infer α and β from the data. Each dataset was first randomly partitioned into training, validation, and testing. Masking was applied to the first two, and no group information is used in testing.

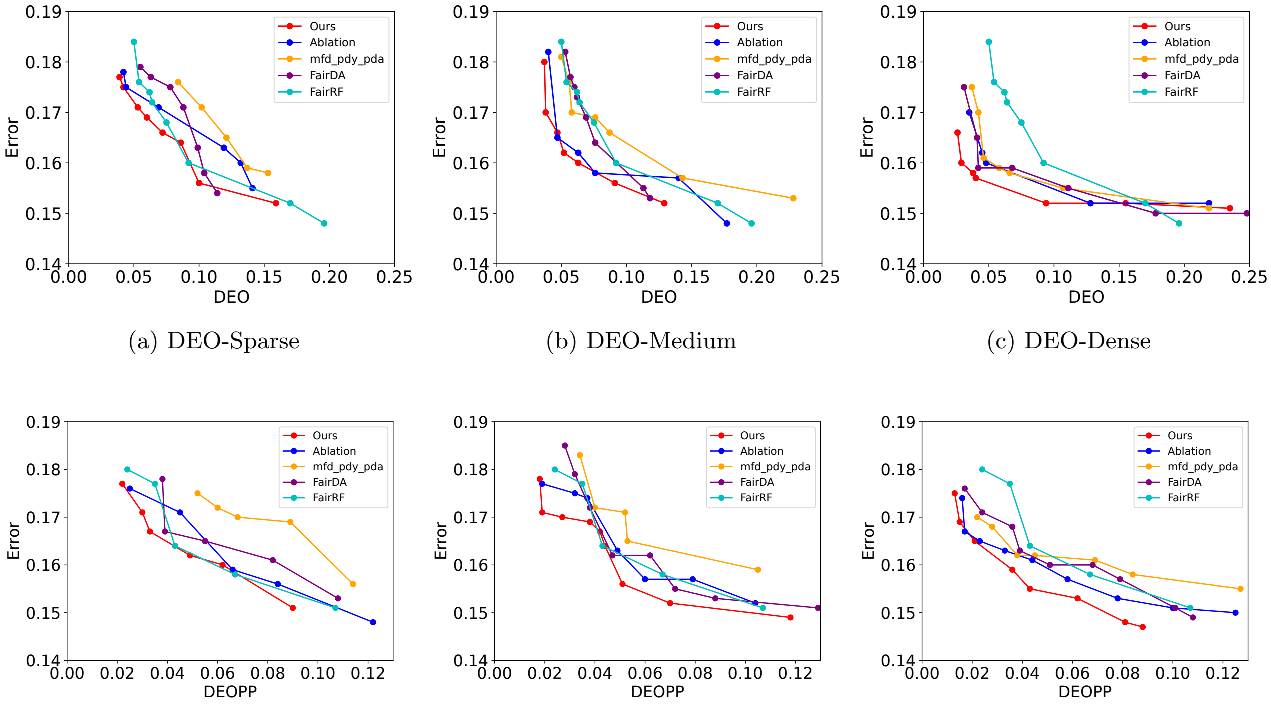
Baseline fair classifiers We compared with three types of methods. MMD-based Fair Distillation (**MFD**, Jung et al., 2021) encourages fairness through feature distillation, but it requires observed labels and demographics. So we imputed y and a with three heuristics:

- (**gt y, gt a**): feed the ground-truth y and a. It forms an unfair comparison with our method but sheds light on the best possible performance of the method;
- (**gt y, rand a**): feed the ground-truth y and impute a uniformly at random to eliminate the correlation between labels and groups;
- (**pred y, pred a**): train a classifier to predict the missing y and a;

As plotting three curves makes the figures overly crowded, we defer the full plots to Appendix C, and only show in the main paper the results without involving ground-truth imputation (the last variant).

The second baseline is **FairRF** (Zhao et al., 2022), which uses proxy features. For Adult, they used *age*, *relation* and *marital status* as the proxy features. We found empirically that using *age* alone gave even better results. So we just used *age* as the proxy feature. We did not apply FairRF to CelebA because proxy features are difficult to identify for images.

The third baseline is **FairDA** (Liang et al., 2023). It divides all examples into source and target domains based



(a) DEO-Sparse

(b) DEO-Medium

(c) DEO-Dense

(d) DEOPP-Sparse

(e) DEOPP-Medium

(f) DEOPP-Dense

Figure 2: Pareto frontier of error versus DEO/DEOPP for **Adult-Gender**

on the availability of sensitive feature, and estimates that for the target domain by domain adaptation.

Implementation details of Fair-SS-VAE We employed M2-VAE (Kingma et al., 2014) for Fair-SS-VAE, with WideResNet-28-2 used as the feature extractor for CelebA, and three fully connected layers for Adult. We kept the sample size to 100 in the Monte-Carlo evaluation of $\mathcal{E}$ in (9).

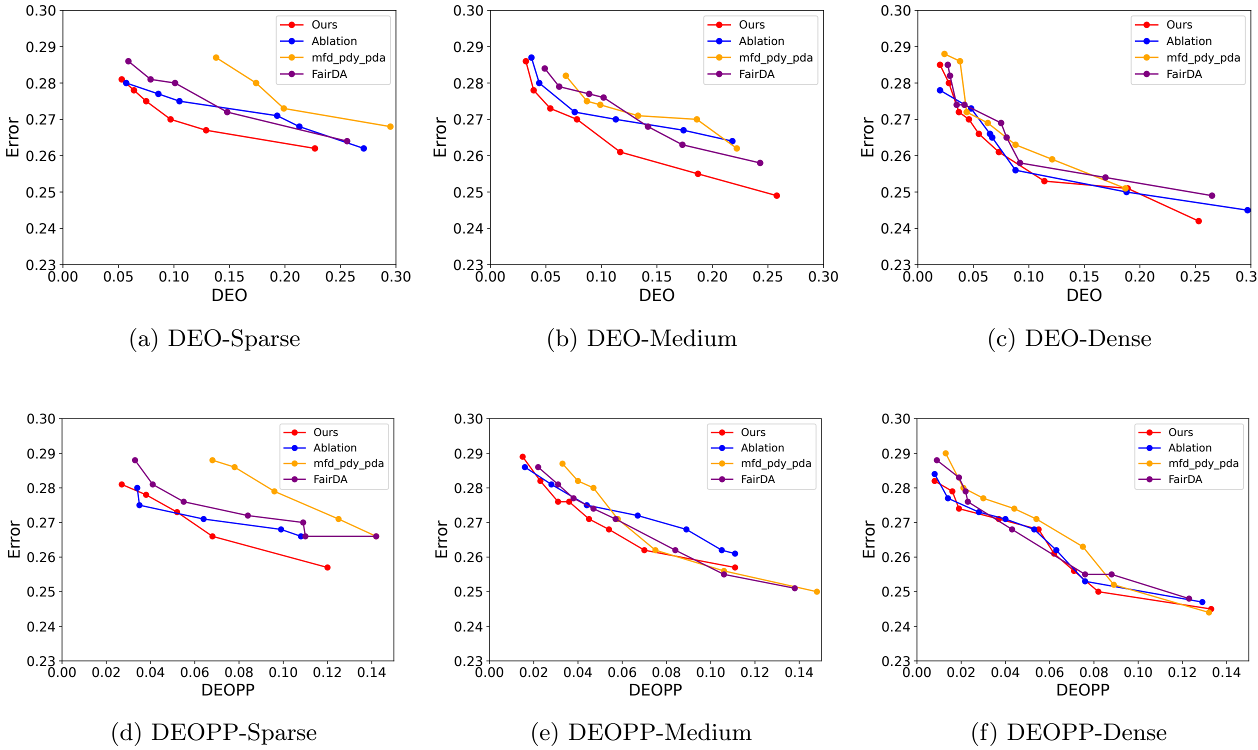
Fairness performance metrics We measured the fairness of the predictions by *difference of equalized odds* (**DEO**) and *difference of equal opportunity* (**DEOPP**). On a test set, the prediction $P(\hat{Y} = 1|x_i)$ is valued 0 or 1, simply rounded from $q_\phi(y|x)$. Surprisingly, it provides a very strong initialization for minimizing the expected fairness risk in (25). In fact, it directly hits a local optimal, leaving no improvement possible from coordinate descent. This suggests that the fairness regularization in Fair-SS-VAE allows VAE to learn a posterior model that already accounts for the desired fairness, dispensing with the need of risk minimization at test time.

Model selection Since we will plot the Pareto frontier of classification error and fairness metrics, ideally we only need a training and test set, and to enumerate all the hyperparameter values, based on which the Pareto frontier can be plotted. However, doing so would be too expensive in computation. Therefore, we used 10% of the data to tune all the other hyper-parameters (network architecture, optimization, etc), and varied the tradeoff weight λ in (13) to roll out the frontier.

Precisely, we focused on varying the tradeoff weight between accuracy and fairness while keeping other hyperparameters fixed, such as batch size, learning rate, and the weights for KL divergence and reconstruction error in the VAE. Since these hyperparameters are already part of the standard SS-VAE, we relied on SS-VAE to select their values. After determining these values, we put back the fairness risk and further adjusted the selected hyperparameters, generating a new set of candidate values for model selection based on the validation set.

6.1 Main results

Figures 2 and 3 show respectively the Pareto frontier of classification error and fairness metrics for Adult-Gender (with gender being the sensitive feature) and CelebA. Due to space limitation, we defer the result of Adult-Race to Figure 4 in Appendix C. As the missing demographics decrease from left to right (sparse to dense), the Pareto Frontier tends to shift from the upper right corner to the bottom left corner, suggesting that



(a) DEO-Sparse

(b) DEO-Medium

(c) DEO-Dense

(d) DEOPP-Sparse

(e) DEOPP-Medium

(f) DEOPP-Dense

Figure 3: Pareto frontier of error versus DEO/DEOPP for **CelebA**

all methods show improvement in the fairness metric (both DEO and DEOPP) while maintaining the same level of accuracy.

Fair-SS-VAE excels in producing a Pareto Frontier significantly closer to the origin across diverse datasets and varying levels of sparsity, indicating its superiority in balancing the tradeoff between accuracy and the fairness metric, unless the ground-truth y and a are used, which would make the comparison unfair to Fair-SS-VAE (see Appendix C). Overall, FairDA is the second most effective.

Recovery of group-conditional missing probability Table 1 presents the value of α and β that Fair-SS-VAE finds for the three datasets and three sparsity levels. Such a recovery is difficult because, once masked, the data does not carry the ground-truth of group feature, precluding counting for rate estimation. Although the exact values are hard to recover, in many cases, they turn out close, and the overall trend correctly decreases from sparse to dense.

Ablation study of stopping gradient To investigate the significance of stop gradient, we set up an ablation experiment that directly uses $\mathcal{E}_{\text{vanilla}}$. The corresponding results are labeled "Ablation" in Figure 2, 3 and 4. Under most circumstances, it performs worse than Fair-SS-VAE, with occasional advantages when applied to datasets with dense sparsity levels. Additionally, we carried out experiments titled CelebA-HighCheekbones, as illustrated in 8, where high cheekbones is the target label and gender serves as the group attribute.

Table 1: Recovery rate of sensitive attribute by Fair-SS-VAE. **Ad-G**: Adult-Gender, **Ad-R**: Adult-Race, **CelA**: CelebA

	Sparse		Medium		Dense	
	$\alpha = .4$	$\beta = .8$	$\alpha = .2$	$\beta = .4$	$\alpha = .1$	$\beta = .2$
Ad-G	.43±.16	.86±.17	.32±.06	.35±.06	.11±.07	.28±.04
Ad-R	.39±.23	.82±.15	.30±.04	.42±.13	.17±.01	.25±.12
CelA	.34±.16	.67±.02	.08±.02	.44±.13	.16±.04	.29±.04

Conclusion, limitation, broader impact, and future work We proposed a new fair classifier that addresses group-conditionally missing demographics. Promising empirical performance is shown. As a limitation, we did not incorporate proxy features when they are available. For future work, we will enable it by conditioning the VAEs on the proxy features. We will also model the bias between *different* sensitive groups, *e.g.*, people of some *race* are less reluctant to reveal their *gender*.

Acknowledgement We thank the reviewers and the meta-reviewer for assessing our paper and for their constructive feedback. This work is supported by NSF grant RI:1939743.

References

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018.

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, sep 2010.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning (ICML)*, 2021a.

L Elisa Celis, Anay Mehrotra, and Nisheeth Vishnoi. Fair classification with adversarial perturbations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using kernel density estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

Consumer Financial Protection Bureau. 12 CFR Part 1002 - Equal Credit Opportunity Act (Regulation B), 1002.5 Rules concerning requests for information. 2023.

Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

Enyan Dai and Suhang Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *International Conference on Web Search and Data Mining (WSDM)*, 2021.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2015.

Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine*, 178 (11):1544–1547, 2018.

Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *arXiv:1806.11212*, 2018.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3315–3323, 2016.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning (ICML)*, 2018.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumbel-softmax. In *International Conference on Learning Representations (ICLR)*, 2017.

Taeuk Jang, Pengyi Shi, and Xiaoqian Wang. Group-aware threshold adaptation for fair classification. In *National Conference of Artificial Intelligence (AAAI)*, 2022.

Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12115–12124, 2021.

Sangwon Jung, Sanghyuk Chun, and Taesup Moon. Learning fair classifiers with partially annotated group labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10348–10357, June 2022.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981, 2022.

F. Kamiran and T.G.K. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1): 1–33, 2012.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Ivar Krumpal. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity*, 47(4):2025–2047, 2013.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Alexandre Lamy, Ziyuan Zhong, Aditya Krishna Menon, and Nakul Verma. Noise-tolerant fair classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yueqing Liang, Canyu Chen, Tian Tian, and Kai Shu. Fair classification via domain adaptation: A dual adversarial learning approach. *Frontiers in Big Data*, 5:129, 2023.

Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ml's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations (ICLR)*, 2017.

Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning (ICML)*, pages 4382–4391. PMLR, 2019.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), 2021.

Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018.

Hussein Mozannar, Mesrob Ohannessian, and Nathan Srebro. Fair learning with private demographic data. In *International Conference on Machine Learning (ICML)*, 2020.

Siddharth Narayanaswamy, Timothy Paige, Jan-Willem Van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Julien Rouzot, Julien Ferry, and Marie-José Huguet. Learning optimal fair scoring systems for multi-class classification. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 197–204. IEEE, 2022.

Abhin Shah, Maohao Shen, Jongha Jon Ryu, Subhro Das, Prasanna Sattigeri, Yuheng Bu, and Gregory W Wornell. Group fairness with uncertainty in sensitive attributes. *arXiv preprint arXiv:2302.08077*, 2023.

Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. Robust optimization for fairness with noisy protected groups. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Wikipedia contributors. Fairness (machine learning) — Wikipedia, the free encyclopedia, 2024. URL https://en.wikipedia.org/w/index.php?title=Fairness_(machine_learning)&oldid=1201888401. [Online; accessed 1-February-2024].

Robert Williamson and Aditya Menon. Fairness risk measures. In *International Conference on Machine Learning (ICML)*, 2019.

Ruicheng Xian and Han Zhao. Efficient post-processing for equal opportunity in fair multi-class classification, 2023. URL https://openreview.net/forum?id=zKjSmbYFZe.

Shen Yan, Hsien-te Kao, and Emilio Ferrara. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *International Conference on Information and Knowledge Management (CIKM)*, 2020.

Jiangchao Yao, Bo Han, Zhihan Zhou, Ya Zhang, and Ivor W. Tsang. Latent class-conditional noise model.

IEEE Trans. Pattern Anal. Mach. Intell., 45(8):9964–9980, 2023.

Adrienne Yapo and Joseph W. Weiss. Ethical implications of bias in machine learning. In *Hawaii International Conference on System Sciences*, 2018.

Qing Ye and Weijun Xie. Unbiased subdata selection for fair classification: A unified framework and scalable algorithms. *arXiv preprint arXiv:2012.12356*, 2020.

Yiliang Zhang and Qi Long. Assessing fairness in the presence of missing data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022.

X. Zhu, J. Lafferty, and Z. Ghahramani. Semi-supervised learning using gaussian fields and harmonic functions. In *Proc. Intl. Conf. Machine Learning*, 2003.

Zhaowei Zhu, Yuanshun Yao, Jiankai Sun, Hang Li, and Yang Liu. Weak proxies are sufficient and preferable for fairness with missing sensitive attributes. *arXiv:2210.03175*, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Conceptual Details

In this appendix section, we fill in more technical details and proofs from the main paper.

A.1 Multi-class and Continuous Valued Fairness Risks

Fairness metrics can be extended to **multi-class** labels and sensitive features in a number of different ways (Rouzot et al., 2022; Xian and Zhao, 2023). Here we recap the disparity of true positive rates (TPRs). Let the label space be $[n_y] := \{1, \ldots, n_y\}$, and the sensitive feature space be $[n_a]$. Denote

$$\mathrm{TPR}(y, a) := P(\hat{Y} = y | Y = y, A = a), \forall\, y \in [n_y],\ a \in [n_a].$$

Let S_y be the set of labels of interest, *e.g.*, $S_y = [n_y]$. Then we can define the fairness metric as

$$\Delta_{\mathrm{TPR}} := \max_{a, a' \in [n_a]} \max_{y, y' \in S_y} \left| P(\hat{Y} = y | Y = y, A = a) - P(\hat{Y} = y' | Y = y', A = a') \right|. \tag{15}$$

When $\Delta_{\mathrm{TPR}} = 0$ and the classes are binary, this fairness notion recovers equalized odds with $S_y = [n_y]$, equal opportunity with $S_y = \{1\}$, and predictive equality with $S_y = \{0\}$ (Chouldechova, 2017).

Suppose there is a classifier f that assigns the probability $P_f(\hat{Y} = y|x)$. Then we can use the data to estimate

$$P(\hat{Y} = y | Y = y, A = a) \approx \operatorname{mean} S, \quad \text{where} \quad S := \{P_f(\hat{Y} = y | x_i) : a_i = a, y_i = y\}.$$

Plugging it into Δ_{TPR} would directly provide a differentiable fairness risk $\mathcal{F}$ for regularization.

Demographic parity: we need $\hat{Y}$ to be independent of A, and the degree of dependence can be measured by any existing applicable metric such as mutual information or

$$\max_{a, a' \in [n_a]} \max_{y \in S_y} \left| P(\hat{Y} = y | A = a) - P(\hat{Y} = y | A = a') \right| \tag{16}$$

$$\text{or} \quad \max_{a \in [n_a]} \max_{y \in S_y} \left| P(\hat{Y} = y | A = a) - P(\hat{Y} = y) \right| \tag{17}$$

Again, we can estimate the relevant probabilities from data by

$$P(\hat{Y} = y | A = a) \approx \operatorname{mean} S, \quad \text{where} \quad S := \{P_f(\hat{Y} = y | x_i) : a_i = a\}$$

$$P(\hat{Y} = y) \approx \operatorname{mean} S, \quad \text{where} \quad S := \{P_f(\hat{Y} = y | x_i)\}.$$

Concrete fairness risk expressions For ease of reference, we explicitly write out the expressions.

For equalized odds

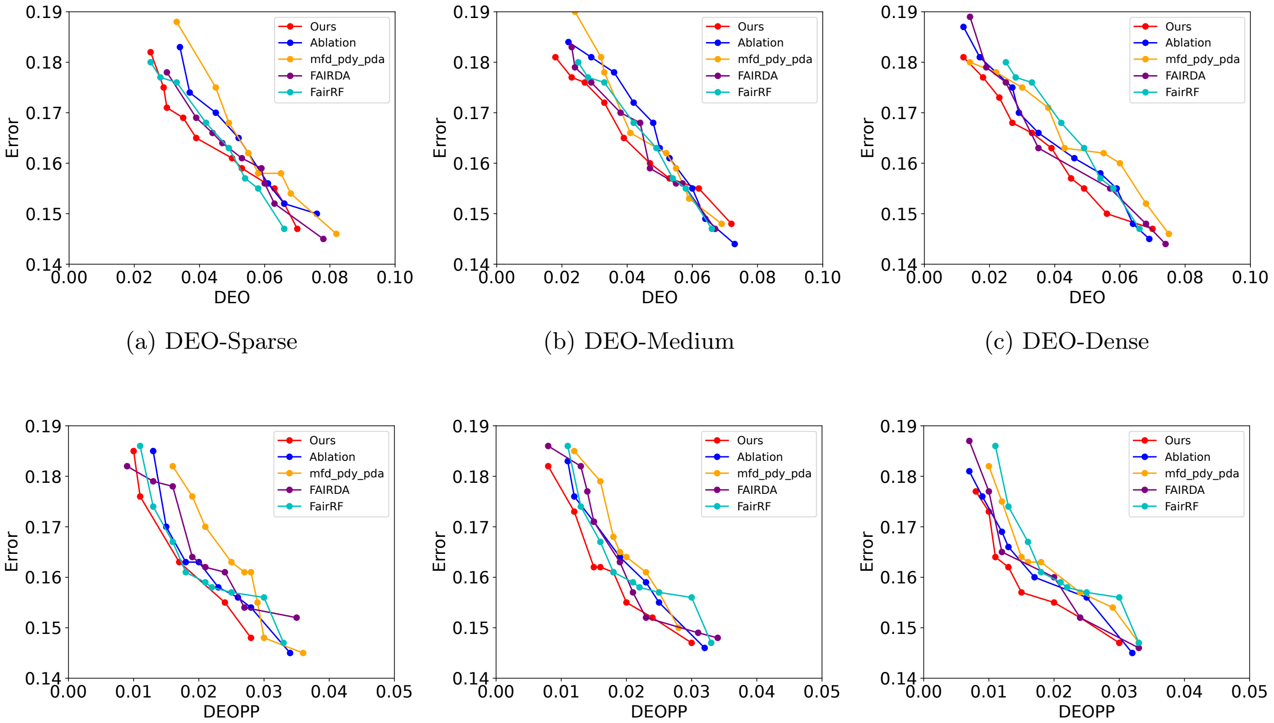
$$\begin{aligned}\mathcal{F}_{\mathrm{DEO}} &= \max_{y \in \{0,1\}} |\operatorname{mean}\{P_f(\hat{Y} = y | x_i) : a_i = 0, y_i = y\} - \operatorname{mean}\{P_f(\hat{Y} = y | x_i) : a_i = 1, y_i = y\}| \\ &= \max_{y \in \{0,1\}} |\operatorname{mean}\{P_f(\hat{Y} = 1 | x_i) : a_i = 0, y_i = y\} - \operatorname{mean}\{P_f(\hat{Y} = 1 | x_i) : a_i = 1, y_i = y\}|. \end{aligned} \tag{18}$$

For equal opportunity:

$$\mathcal{F}_{\mathrm{DEOPP}} = |\operatorname{mean}\{P_f(\hat{Y} = 1 | x_i) : a_i = 0, y_i = 1\} - \operatorname{mean}\{P_f(\hat{Y} = 1 | x_i) : a_i = 1, y_i = 1\}|. \tag{19}$$

For demographic parity:

$$\begin{aligned}\mathcal{F}_{\mathrm{DDP}} &= \max_{y \in \{0,1\}} |\operatorname{mean}\{P_f(\hat{Y} = y | x_i) : a_i = 0\} - \operatorname{mean}\{P_f(\hat{Y} = y | x_i) : a_i = 1\}| \\ &= |\operatorname{mean}\{P_f(\hat{Y} = 1 | x_i) : a_i = 0\} - \operatorname{mean}\{P_f(\hat{Y} = 1 | x_i) : a_i = 1\}|. \end{aligned} \tag{20}$$



(a) DEO-Sparse

(b) DEO-Medium

(c) DEO-Dense

(d) DEOPP-Sparse

(e) DEOPP-Medium

(f) DEOPP-Dense

Figure 4: Pareto frontier of error versus DEO/DEOPP for **Adult-Race**

Continuously valued label and sensitive features When the label and sensitive features are **continuous**, the above methods cease to be applicable. We will resort to variational estimators such as InfoNCE, and they require paired sampled of $\hat{Y}$ and A. However, we can do better than that because the classifier f provides a distribution of $\hat{Y}$ instead of a single sample. For example, in mutual information neural estimation (MINE), we can estimate the KL-divergence between $P_{A,\hat{Y}}$ and $P_A P_{\hat{Y}}$ by

$$\sup_{t:(A,\hat{Y})\to\mathbb{R}} \left\{ \underbrace{\mathbb{E}_{P_{A,\hat{Y}}}[t(A,\hat{Y})]}_{\textcircled{1}} - \underbrace{\mathbb{E}_{P_A}\mathbb{E}_{P_{\hat{Y}}}[\exp(t(A,\hat{Y})-1)]}_{\textcircled{2}} \right\}. \tag{21}$$

Then we can estimate both terms by

$$\textcircled{1} \approx \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\hat{Y}\sim P_f(\hat{Y}|x_i)}[t(a_i,\hat{Y})] \quad \textcircled{2} \approx \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}_{\hat{Y}\sim P_f(\hat{Y}|x_j)}[\exp(t(a_i,\hat{Y})-1)]. \tag{22}$$

If the function space of t is simple enough, we can directly evaluate the two expectations in a closed form. Otherwise, we can sample $\hat{Y}$ from $P_f(\hat{Y}|x_j)$ and apply Gumbel-softmax for differentiation.

A.2 Proof of Theorem 1

Proof. First we evaluate the bounded difference: replace Z_s with Z'_s while keeping all other $Z_{s'}$ intact. Here Z'_s is any admissible assignment of $a_i^{(s)}$ and $y_i^{(s)}$. Since $\mathcal{F}$ takes value in $[0,1]$, it is trivial that $|\hat{\mathcal{E}}_n(Z_1,\ldots,Z_N) - \hat{\mathcal{E}}_n(Z_1,\ldots,Z_{s-1},Z'_s,Z_{s+1},\ldots,Z_N)| = \frac{1}{N}|\mathcal{F}(P_f,Z_{s'}) - \mathcal{F}(P_f,Z_s)| \leq C/N$. Further noting that the expectation of $\hat{\mathcal{E}}_n(Z_1,\ldots,Z_N)$ over $(Z_1,\ldots,Z_N)$ is $\mathcal{E}$, the McDiarmid's inequality immediately implies the theorem. □

A.3 Straight-through Gumbel-Softmax for vanilla fairness risk

Suppose a Bernoulli variable y_i has $P(y_i = 1) = q_i$. Then the vanilla Gumbel-Softmax method first draws i.i.d. samples from a Gumbel$(0, 1)$ distribution:

$$\{\alpha_i^{(s)}, \beta_i^{(s)} : i = 1, \ldots, n, \text{ and } \ s = 1, \ldots, N\}. \tag{23}$$

Then a sample of $y_i^{(s)}$ is constructed by

$$y_i^{(s)} = \frac{\exp(\frac{\log q_i + \alpha_i^{(s)}}{T})}{\exp(\frac{\log q_i + \alpha_i^{(s)}}{T}) + \exp(\frac{\log(1-q_i) + \beta_i^{(s)}}{T})}\ , \tag{24}$$

where $T > 0$ is a temperature parameter. A larger value of T leads to a more smooth and uniform sample, i.e., $y_i^{(s)}$ will be closer to 0.5. Clearly $y_i^{(s)}$ is not discrete, but for reasonably small T, it will be close to 0 or 1. To patch up the non-integrality of $y_i^{(s)}$, the straight-through Gumbel-Softmax gradient estimator only uses the differentiable variable in the backward gradient propagation, while the forward pass still uses categorical variables (*i.e.*, turning $y_i^{(s)}$ into 1 if it is above 0.5, and 0 otherwise). The same approach can be applied to sample a_i.

A.4 ELBO

We can extend the ELBO as

$$\begin{aligned}\log p_\theta(\mathbf{x}, \tilde{a}, \tilde{y}) \geq & \mathop{\mathbb{E}}_{(\mathbf{z},a,y)\sim q_\phi(\cdot|\mathbf{x},\tilde{a},\tilde{y})} [\log p_\theta(\mathbf{x}, \tilde{a}, \tilde{y}|a, y, \mathbf{z}) + \log(p(y)p(a)p(\mathbf{z})) - \log q_\phi(a, y, \mathbf{z}|\mathbf{x}, \tilde{a}, \tilde{y})] \\ = & \mathop{\mathbb{E}}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(x|a, y, \mathbf{z}) + \mathop{\mathbb{E}}_{y\sim q_\phi(y|\mathbf{x},\tilde{y})} \log P_\theta(\tilde{y}|y) + \mathop{\mathbb{E}}_{a\sim q_\phi(a|\mathbf{x},\tilde{a})} \log P_\theta(\tilde{a}|a) \\ & - \mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] - \mathrm{KL}[q_\phi(y|\mathbf{x}, \tilde{y})||p(y)] - \mathrm{KL}[q_\phi(a|\mathbf{x}, \tilde{a})||p(a)] \\ =: & \mathrm{ELBO}(\mathbf{x}, \tilde{a}, \tilde{y}).\end{aligned}$$

A.5 Soften the step function to enable differentiation

The step function $[\![\cdot > 0]\!]$ offers no useful derivative. Noting that taking a threshold of a probability $f(X)$ is equivalent to thresholding its logit $\sigma^{-1}(f(X))$, we improve numerical performance by setting $P_f(Y = 1|X) = \sigma(\frac{\sigma^{-1}(f(X))-\tau}{T})$, where the temperature T controls the steepness of the approximation. Finally, group-specific threshold can be enabled by replacing τ with $\tau_1 a_i + \tau_0(1 - a_i)$. Given the values of a_i—either from the training data or from the N samples—it is straightforward to optimize this objective.

A.6 Labeling test data by minimizing fairness and classification risks

With the learned ϕ, the test data can be labeled by simply seeking the $\mathbf{y}$ that maximizes the likelihood while also minimizing the expected fairness risk:

$$\arg \min_{y_i\in\{0,1\}} \Big\{ -\frac{1}{n_{\text{test}}} \sum_i \log q_\phi(y_i|\mathbf{x}_i) + \mathop{\mathbb{E}}_{A_i\sim q_\phi(A|\mathbf{x}_i,\tilde{a}_i)} \mathcal{F}(P_{q_\phi(y|\mathbf{x})}, \mathbf{y}, A_{1:n_{\text{test}}}) \Big\}.$$

When all demographics are observed, this can be solved by tuning group-specific thresholds on $q_\phi(y_i = 1|\mathbf{x}_i)$. When test demographics are missing (possibly in their entirety), we can first compute the binary demographics via $\arg\max_A q_\phi(A|\mathbf{x}_i, \tilde{a}_i)$, and then tune the group-specific threshold on $q_\phi(y_i = 1|\mathbf{x}_i)$. Afterwards, we do a coordinate descent to finetune $\mathbf{y}$ in the face of full expectation on A_i for $a_i = \emptyset$ using Monte Carlo.

A.7 Extension to class-probability estimation

The above discussion has been intentionally kept general by considering a randomized classifier for $Y|X$. In practice, however, this is not quite realistic because one does not get to predict multiple times for a single test set. Therefore, we would like to specialize the framework to classification based on class-probability, *i.e.*,

$P_f(Y=1|x) = [\![f(x) - \tau > 0]\!]$. It also provides the convenience of tuning the parameter τ. Once h is learned and fixed thereafter, different test scenarios may demand different trade-offs between risks in classification and fairness. For example, Jang et al. (2022) proposed adapting the thresholds for each demographic group. To enable differentiation, the $[\![\cdot]\!]$ operator can be softened as shown in Appendix A.

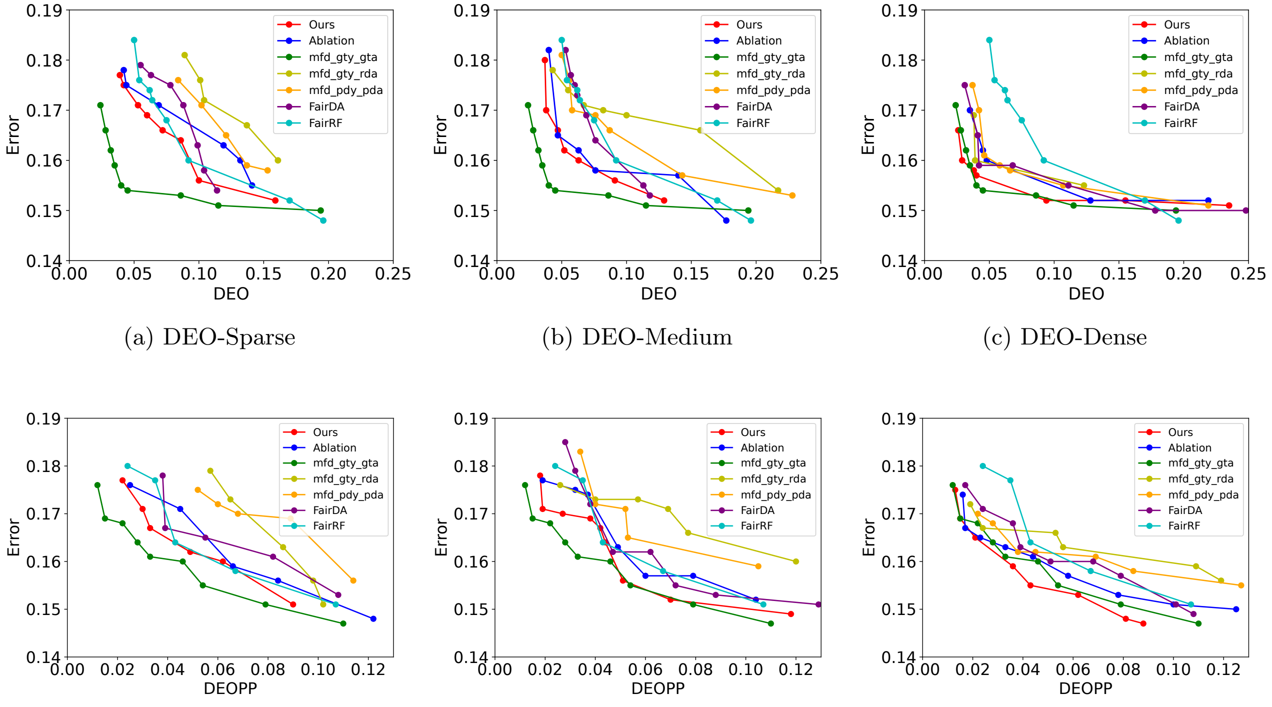
B Experiment Settings

We implemented our algorithm using Python 3.11.5 and PyTorch 2.1.1. The code runs on a server with Ubuntu 18.04, powered by an Intel(R) Core(TM) i9-9900X CPU @ 3.50GHz and 128 GB of RAM. For GPU resources, we have four GeForce RTX 2080 Ti cards, each with 11 GB of graphic memory, and the GPU driver version is 455.23.05.

C Additional Experimental Results

Figure 4 shows the Pareto frontier of error versus DEO/DEOPP for Adult-Race. Figure 8 shows the results for Celeba-HighCheekbones. Here, we use high cheekbones as the label and gender as the group attribute.

We additionally present the Pareto frontier attained by all the baseline methods. Figures 5 to 7 are crowded, hence moved to the appendix. The curve of mfd_gty_gta remains intact across all sparsity levels for a given dataset, because it imputes the sensitive feature with its ground truth. Note that it is not fair to compare Fair-SS-VAE against mfd_gty_gta, because the latter has full access to the ground truth of both labels and sensitive features. Nonetheless, we still visualize its Pareto frontier because it serves as a rough upper bound for all the methods.



(a) DEO-Sparse

(b) DEO-Medium

(c) DEO-Dense

(d) DEOPP-Sparse

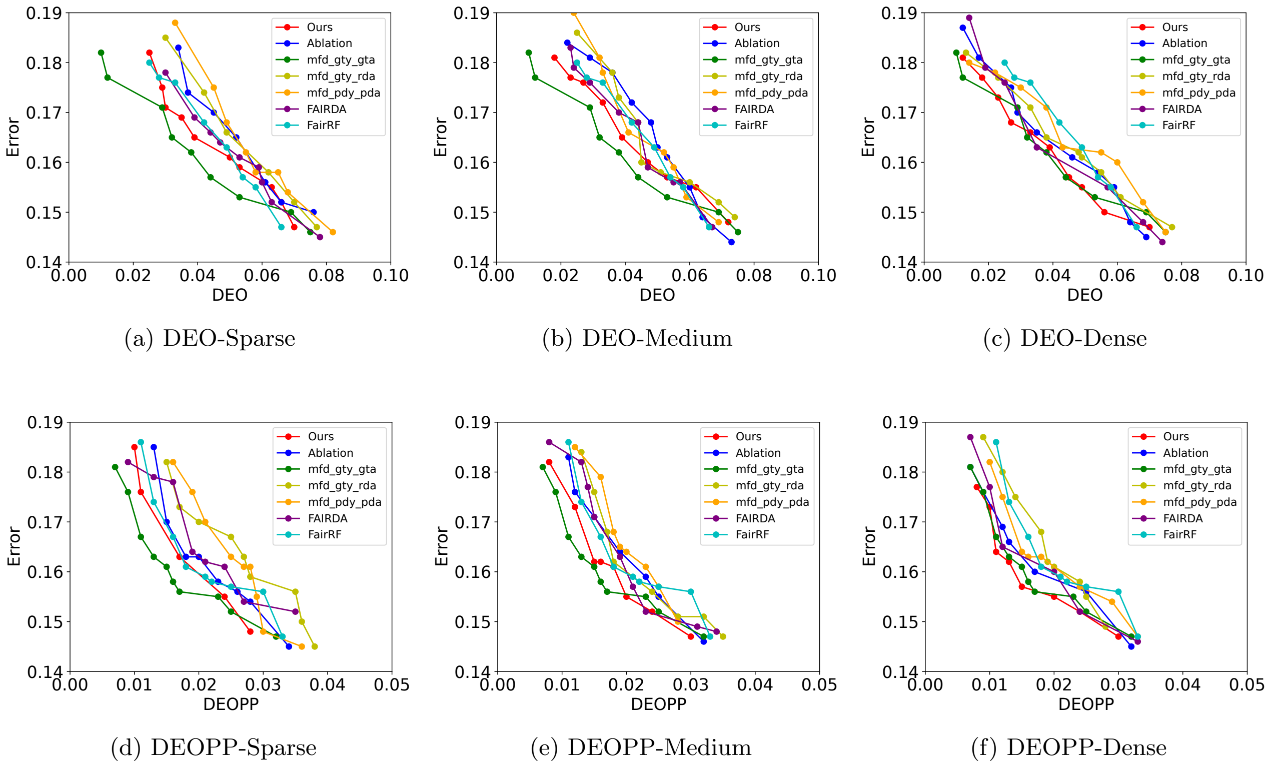
(e) DEOPP-Medium

(f) DEOPP-Dense

Figure 5: Pareto frontier of error versus DEO/DEOPP for Adult-Gender

D Source Code

The experiment code is available at `https://tinyurl.com/4wndtfbb`.



(a) DEO-Sparse

(b) DEO-Medium

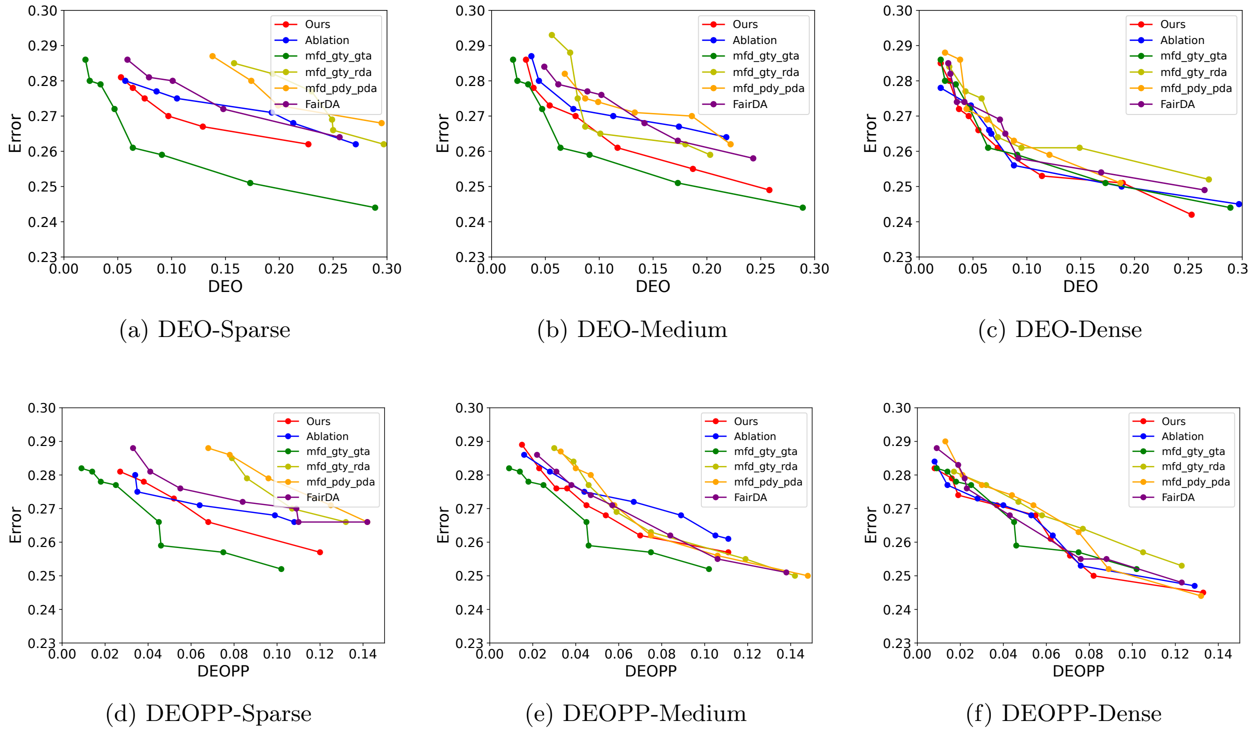
(c) DEO-Dense

(d) DEOPP-Sparse

(e) DEOPP-Medium

(f) DEOPP-Dense

Figure 6: Pareto frontier of error versus DEO/DEOPP for Adult-Race



(a) DEO-Sparse

(b) DEO-Medium

(c) DEO-Dense

(d) DEOPP-Sparse

(e) DEOPP-Medium

(f) DEOPP-Dense

Figure 7: Pareto frontier of error versus DEO/DEOPP for CelebA

E Statistical Property of Estimator (5)

Theorem 2. *The following estimator of $\theta := P(\hat{Y} = 1|A = a)$ is consistent and* ***asymptotically*** *unbiased:*

$$\theta_n := \sum\nolimits_{i:a_i=a} P(\hat{Y} = 1|x_i) \Big/ \sum\nolimits_{i:a_i=a} 1 \,. \tag{25}$$

However, it is in general not unbiased.

Proof. To show consistency, note

$$\frac{1}{n} \sum\nolimits_{i:a_i=a} 1 \;\xrightarrow{P}\; P(A = a) \tag{26}$$

$$\frac{1}{n} \sum\nolimits_{i:a_i=a} P(\hat{Y} = 1|x_i) \;\xrightarrow{P}\; P(\hat{Y} = 1, A = a). \tag{27}$$

Therefore,

$$\frac{\sum_{i:a_i=a} P(\hat{Y} = 1|x_i)}{\sum_{i:a_i=a} 1} = \frac{\frac{1}{n}\sum_{i:a_i=a} P(\hat{Y} = 1|x_i)}{\frac{1}{n}\sum_{i:a_i=a} 1} \;\xrightarrow{P}\; \frac{P(\hat{Y} = 1, A = a)}{P(A = a)} = P(\hat{Y} = 1|A = a). \tag{28}$$

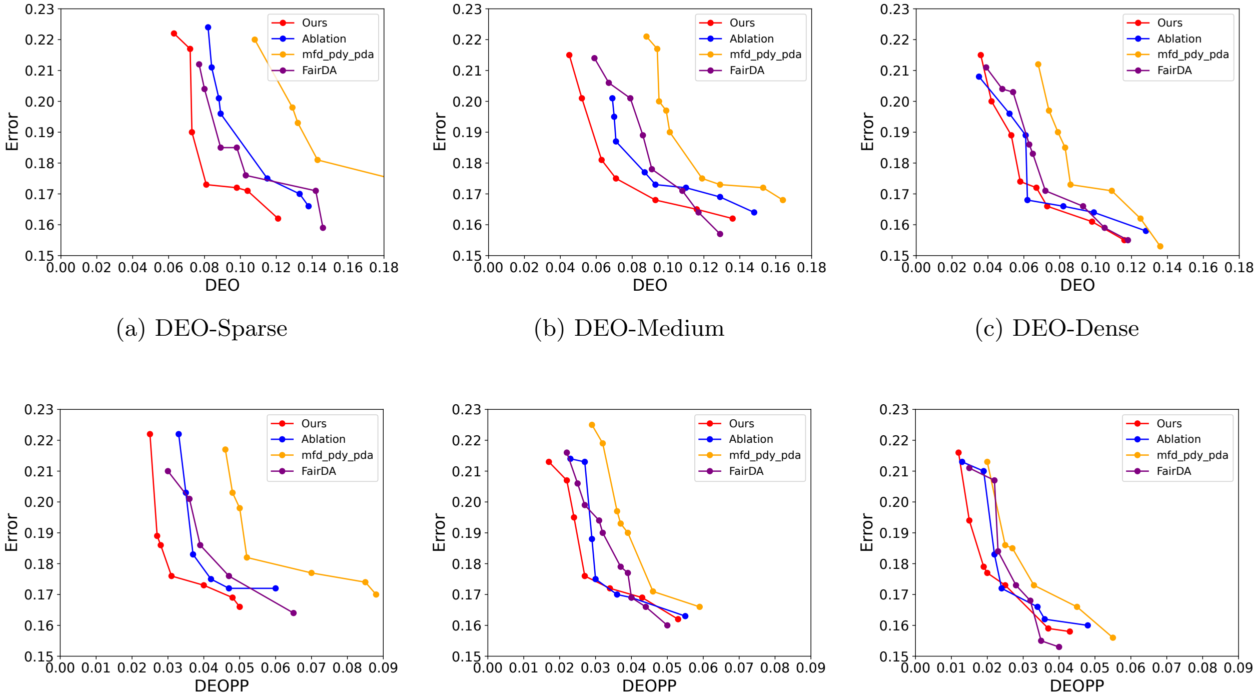
To show asymptotic unbiasedness, note

$$\mathbb{E}[|\theta_n - \theta|] = \mathbb{E}[|\theta_n - \theta| \cdot [\![|\theta_n - \theta| < \epsilon]\!]] + \mathbb{E}[|\theta_n - \theta| \cdot [\![|\theta_n - \theta| \geq \epsilon]\!]] \tag{29}$$

$$\leq \epsilon + \mathbb{E}[|\theta_n - \theta| \cdot [\![|\theta_n - \theta| \geq \epsilon]\!]] \tag{30}$$

$$\text{(by Cauchy-Schwarz)} \quad \leq \epsilon + \sqrt{\mathbb{E}[|\theta_n - \theta|^2] \, \mathrm{Pr}(|\theta_n - \theta| \geq \epsilon)}. \tag{31}$$

Consistency has already established that $\mathrm{Pr}(|\theta_n - \theta| \geq \epsilon) \to 0$ as $n \to \infty$. Since θ_n is clearly bounded by 1, the right-hand side falls below 2ϵ for sufficiently large n. □



(a) DEO-Sparse

(b) DEO-Medium

(c) DEO-Dense

(d) DEOPP-Sparse

(e) DEOPP-Medium

(f) DEOPP-Dense

Figure 8: Pareto frontier of error versus DEO/DEOPP for **CelebA-HighCheekbones**