FARE: Provably Fair Representation Learning with Practical Certificates

Nikola Jovanović [1] **Mislav Balunović** [1] **Dimitar I. Dimitrov** [1] **Martin Vechev** [1]

Abstract

Fair representation learning (FRL) is a popular class of methods aiming to produce fair classifiers via data preprocessing. Recent regulatory directives stress the need for FRL methods that provide *practical certificates*, i.e., provable upper bounds on the unfairness of any downstream classifier trained on preprocessed data, which directly provides assurance in a practical scenario. Creating such FRL methods is an important challenge that remains unsolved. In this work, we address that challenge and introduce FARE (*Fairness with Restricted Encoders*), the first FRL method with practical fairness certificates. FARE is based on our key insight that restricting the representation space of the encoder enables the derivation of practical guarantees, while still permitting favorable accuracy-fairness tradeoffs for suitable instantiations, such as one we propose based on fair trees. To produce a practical certificate, we develop and apply a statistical procedure that computes a finite sample high-confidence upper bound on the unfairness of any downstream classifier trained on FARE embeddings. In our comprehensive experimental evaluation, we demonstrate that FARE produces practical certificates that are tight and often even comparable with purely empirical results obtained by prior methods, which establishes the practical value of our approach.

1. Introduction

It has been repeatedly shown that machine learning systems deployed in real-world applications propagate training data biases (Corbett-Davies et al., 2017; Kleinberg et al., 2017a). This is especially concerning in decision-making applications which use data representing humans (e.g., financial or medical), where such biases can lead to unfavorable treatment of certain population subgroups (Brennan et al., 2009; Barocas & Selbst, 2016). For instance, a loan prediction system might recommend rejection based on a *sensitive attribute*, such as race or gender. The relevance of fairness (as perceived by companies) has increased the most over the past year compared to any other potential risk of AI (Chui et al., 2021; Benaich & Hogarth, 2021).

Fair representation learning A promising approach to addressing this issue is *fair representation learning* (FRL, Zemel et al. (2013)), which relies on separation of responsibility between two parties—a *data producer* and a *data consumer*. The key promise of FRL is that the data producer, who uses an encoder f to transform each point $\boldsymbol{x} \in \mathcal{X}$ into a debiased representation $\boldsymbol{z}$, can ensure that these representations can be directly given to any data consumer aiming to solve a prediction task, such that *any* classifier $g \in \mathcal{G}$ they train will have favorable fairness, even if they are unaware of fairness or actively aim to discriminate.

The need for certificates Recent regulatory directives (FTC, 2021; EU, 2021) *demand* that parties aiming to deploy ML models ensure the fairness of predictions (Dwork et al., 2012). In high-stakes applications (e.g., judicial), failing to guarantee fairness may thus lead to fines or reputational harm (Simonite, 2018), in addition to causing societal harm. This makes it essential that data producers provide *certificates* guaranteeing that any classifier trained on their representations must be fair, which follows the trends in other areas in which certificates are of interest, such as privacy (Abadi et al., 2016) or robustness (Gehr et al., 2018).

Practical certificates To achieve this goal, we are interested in *practical* certificates, which upper bound a fairness metric (such as demographic parity distance) *with high probability*, and are computed using a *finite dataset* that the data producer can access. The certificate should not make *any additional assumptions* about the data distribution, and it should hold for *any model* trained on the representations. Finally, a certificate should provide tight, non-vacuous bounds that can be *computed on real datasets*. While there has been a lot of work in this direction, none of the prior approaches satisfy all of these requirements (see Section 4).

This work We propose FARE (Fairness with Restricted Encoders, Figure 1)—the first FRL method that provides

[1]Department of Computer Science, ETH Zurich. Correspondence to: Nikola Jovanović <nikola.jovanovic@inf.ethz.ch>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

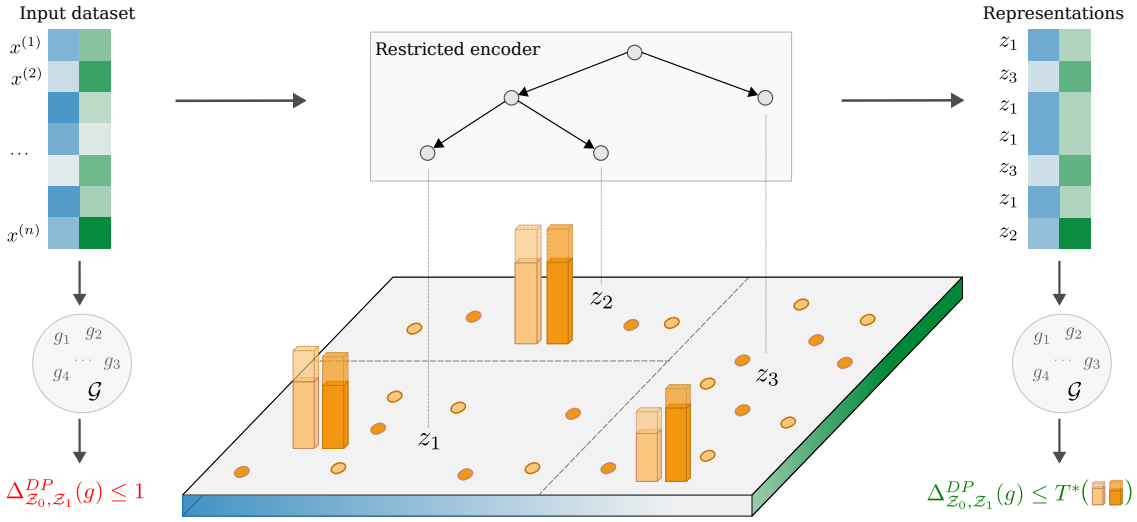


Figure 1: Overview of FARE. The input dataset is transformed into fair representations using a restricted encoder. FARE can compute a certificate $T^\star$ which provably upper-bounds the unfairness of any classifier $g \in \mathcal{G}$ trained on the representations.

practical certificates. Our key insight is that using a *restricted encoder*, i.e., limiting possible representations to a finite set $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\}$, allows us to derive a practical statistical procedure that computes a high-confidence upper bound on the unfairness of any g, i.e., a fairness certificate. We instantiate this with an encoder based on decision trees, leading to an efficient end-to-end FRL method, producing fair representations with practical fairness certificates.

Concretely, FARE starts from a given set of data samples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\}$ (left in Figure 1), partitions the input space into k *cells* (middle plane, here $k = 3$) using the decision tree encoder, and maps all samples from the same cell i into the same representation $\boldsymbol{z}_i$ (right). As usual in FRL, training classifiers on $\boldsymbol{z}_i$ improves fairness at the slight cost of accuracy. However, the key advantage of FARE is that using a restricted encoder allows us to estimate the distribution of sensitive groups in each cell—namely, we empirically estimate $P(s = 0|\boldsymbol{z}_i)$ and $P(s = 1|\boldsymbol{z}_i)$ (solid color orange bars) for all $\boldsymbol{z}_i$, and further upper-bound those values with high probability (transparent bars). This in turn leads to the key feature of FARE: a certificate $T^\star$, i.e., a tight upper bound on the unfairness of any $g \in \mathcal{G}$, where $\mathcal{G}$ is the set of all downstream classifiers that can be trained on $\boldsymbol{z}_i$.

In experiments on several real datasets, we demonstrate that FARE certificates are relatively tight and thus practical (as defined above and elaborated on in later sections). Further, we show that downstream classifiers trained on FARE representations can achieve empirical accuracy-fairness tradeoffs that are comparable to those of methods from prior work.

We believe our work represents a major step towards solving the important problem of preventing the deployment of discriminatory machine learning models in a provable way.

Main contributions Our key contributions are:

- A practical statistical procedure that, for a restricted encoder, produces a *practical certificate* (see Section 4), upper-bounding the unfairness of any downstream classifier trained on its representations (Section 5).
- An end-to-end FRL method, FARE, that instantiates this approach with a fair decision tree encoder (Section 6), applying our statistical procedure to augment the representations with a practical certificate. The implementation of FARE is publicly available at `https://github.com/eth-sri/fare`.
- An extensive experimental evaluation in several settings, demonstrating favorable empirical fairness results, as well as practical certificates, which were out of reach for prior work. Interestingly, FARE certificates are often comparable to purely empirical results of existing FRL methods (Section 7).

2. Related Work

We discuss related work on FRL for group fairness, provable fairness in orthogonal settings, and fair decision trees.

FRL for group fairness Following Zemel et al. (2013) which originally introduced FRL, a plethora of different methods have been proposed based on optimization (Calmon et al., 2017; Shui et al., 2022), adversarial training (Edwards & Storkey, 2016; Xie et al., 2017; Madras et al., 2018; Song et al., 2019; Feng et al., 2019; Roy & Boddeti, 2019; Kairouz et al., 2022; Kim et al., 2022), variational approaches (Louizos et al., 2016; Moyer et al., 2018; Oh

et al., 2022; Liu et al., 2022), disentanglement (Sarhan et al., 2020), mutual information (Gupta et al., 2021; Gitiaux & Rangwala, 2022), and normalizing flows (Balunović et al., 2022; Cerrato et al., 2022). The key issue is that most of these methods produce representations with no fairness certificate, meaning that a model trained on their representations could have much worse fairness. In fact, prior work (Elazar & Goldberg, 2018; Xu et al., 2020; Gupta et al., 2021; Gitiaux & Rangwala, 2021) has shown that adversarial training methods often significantly overestimate the fairness of their representations. We analyze FRL works attempting to provide guarantees in detail in Section 4.

Provable fairness in other settings Numerous related works on provable fairness operate in a different setting than ours—we investigate group fairness via FRL and scenarios where separation of responsibility is crucial, and following prior work, focus on other FRL methods in our evaluation. Several FRL methods have proposed approaches for learning individually fair representations (Lahoti et al., 2019; Ruoss et al., 2020; Peychev et al., 2021), a different notion of fairness than group fairness. Prior work has also examined fairness guarantees in problem settings such as ranking (Konstantinov & Lampert, 2021), distribution shifting (Kang et al., 2022; Jin et al., 2022), in-processing (Feldman et al., 2015; Donini et al., 2018; Celis et al., 2019; Shamsabadi et al., 2023), post-processing (Petersen et al., 2021; Li et al., 2022), and meta-learning (Oneto et al., 2020).

Decision trees for fairness The line of work focusing on adapting decision trees to fairness concerns includes a wide range of methods which differ mainly in the branching criterion. Common choices include variations of Gini impurity (Kamiran et al., 2010; Raff et al., 2018; Zhang & Ntoutsi, 2019), mixed-integer programming (Aghaei et al., 2019; Wang et al., 2022) or AUC (Barata & Veenman, 2021), while some apply adversarial training (Grari et al., 2020; Ranzato et al., 2021). Further, some works operate in a different setting such as online learning (Zhang & Ntoutsi, 2019) or post-processing (Abebe et al., 2022). The only works in this area that offer provable fairness guarantees are Ranzato et al. (2021), which certifies individual fairness for post-processing, Meyer et al. (2021), which certifies that predictions will not be affected by data changes, and Shamsabadi et al. (2023), who propose an in-processing method for training fair trees with zero-knowledge proofs. These fundamentally differ from our FRL setting where the goal is to certify fairness of any downstream classifier.

3. Background

We set up the notation and background on FRL, and recall some key results from prior work which we build upon.

Fair classification Assume tuples $(\boldsymbol{x}, s) \in \mathbb{R}^d \times \{0,1\}$ from distribution $\mathcal{X}$, where each datapoint belongs to a group with respect to a sensitive attribute s. We focus on binary classification, i.e., given $y \in \{0,1\}$ for each datapoint, the goal is to build a model $g\colon \mathbb{R}^d \to \{0,1\}$ that predicts y from $\boldsymbol{x}$. Besides maximizing accuracy of g, we aim to maximize its fairness with respect to s according to some definition, often for a slight accuracy cost. While we consider binary s and y, our results can be easily extended to other settings (see Appendix D.1).

Fair representation learning In FRL, a *data producer* applies an encoder $f\colon \mathbb{R}^d \to \mathbb{R}^{d'}$ to obtain a *representation* $\boldsymbol{z} = f(\boldsymbol{x})$ of each datapoint, inducing a joint distribution $\mathcal{Z}$ of $(\boldsymbol{z}, s)$. The representations are then published and used by various *data consumers*, who train downstream classifiers g to predict some y from $\boldsymbol{z}$, i.e., now we have $g\colon \mathbb{R}^{d'} \to \{0,1\}$. The key advantage of FRL is that by ensuring fairness properties of $\boldsymbol{z}$, we can limit the unfairness of *any* g trained on data from $\mathcal{Z}$, for any number of consumers. Namely, there is a separation of responsibility: data producers need to ensure fairness, such that even fairness-agnostic or adversarial consumers are unable to discriminate.

Fairness metrics Let $\mathcal{Z}_0$ and $\mathcal{Z}_1$ denote conditional distributions of $\boldsymbol{z}$ where $s = 0$ and $s = 1$, respectively. We aim to minimize the *demographic parity (DP) distance* (i.e., *statistical parity*) of g:

$$\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}(g) := \left|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z}_0}[g(\boldsymbol{z})] - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z}_1}[g(\boldsymbol{z})]\right|.$$

This metric choice is motivated by prior work and enables comparison with a wide range of baselines, as many consider only DP distance (Moyer et al., 2018; Gupta et al., 2021; Kim et al., 2022)—other definitions may be more suitable for particular use-cases (Hardt et al., 2016), and our method can be adapted to support them. For instance, equalized odds and equal opportunity correspond to DP on $\mathcal{Z}_0$ and $\mathcal{Z}_1$ conditioned on the target label y, enabling application of our statistical procedure given in Section 5—see Appendix D.2 for more details and experimental results.

Towards provable FRL The goal of *provable FRL* is for the data producer to provide a *fairness certificate* $T^\star \in \mathbb{R}$, guaranteeing that the DP distance of *any* classifier trained by *any* data consumer on representations from $\mathcal{Z}$ is not higher than $T^\star$. We now recall results from prior work, which we also utilize as a first step towards provable FRL.

Consider $h\colon \mathbb{R}^{d'} \to \{0,1\}$, the adversary predicting group membership s, aiming to maximize its balanced accuracy:

$$BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h) := \frac{1}{2}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z}_0}[1 - h(\boldsymbol{z})] + \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z}_1}[h(\boldsymbol{z})]\right). \quad (1)$$

Let $h^\star$, such that for all h, $BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h^\star) \geq BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h)$, denote the *optimal adversary*. Intuitively, $h^\star$ predicts s for which the likelihood of $\boldsymbol{z}$ under the corresponding distribution ($\mathcal{Z}_0$ or $\mathcal{Z}_1$) is larger. More formally, $h^\star(\boldsymbol{z}) = \mathbb{1}\{P(\boldsymbol{z}|s=1) \geq P(\boldsymbol{z}|s=0)\}$ (see e.g., Balunović et al. (2022) for a proof). As shown in Feldman et al. (2015); McNamara et al. (2017); Madras et al. (2018),

$$\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}(g) \leq 2 \cdot BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h^\star) - 1 \tag{2}$$

holds for any g, i.e., $BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h^\star)$ represents a fairness certificate. Another similar upper bound noted by Madras et al. (2018); Zhao et al. (2020); Shen et al. (2021) uses the total variation distance $d_{TV}(\mathcal{Z}_0,\mathcal{Z}_1)$ (Levin et al., 2006) of distributions, as it closely relates to $h^\star$, namely $BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h^\star) = 1/2 + 1/2 \cdot d_{TV}(\mathcal{Z}_0,\mathcal{Z}_1)$. Utilizing these results to obtain certificates is not simple, as accurately approximating $h^\star$ or $d_{TV}(\mathcal{Z}_0,\mathcal{Z}_1)$ (or related distance measures from the family of IPM or ϕ-divergences) is hard (Devroye et al., 1996; Sriperumbudur et al., 2012; Bhattacharyya et al., 2022).

4. FRL with Practical Certificates

We now list requirements that a practical certificate should satisfy, and reflect on prior attempts to achieve this goal.

Practical certificates We start by stating an informal definition of a *practical DP distance certificate*:

Definition 4.1 (Informal)**.** Given small ϵ, finite dataset $D = \{(\boldsymbol{x}^{(j)}, s^{(j)})\}_{j=1}^n$ sampled from an unknown distribution $\mathcal{X}$, and an encoder f mapping each $\boldsymbol{x}^{(j)}$ into a corresponding representation $\boldsymbol{z}^{(j)}$, a *practical DP distance certificate* is a value $T^\star(n, D) \in \mathbb{R}$ such that

$$\sup_{g \in \mathcal{G}} \Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}(g) \leq T^\star(n, D)$$

holds with probability $1-\epsilon$ (over the data sampling), where $\mathcal{G}$ is the set of all downstream classifiers. In realistic use-cases, computing $T^\star(n, D)$ must lead to non-vacuous bounds, i.e., $T^\star(n, D) < 1$.

This definition naturally arises from the core principles of FRL, as discussed in Section 1, and is necessary to provide suitable assurances to involved parties. We next formalize the requirements implicitly stated in Definition 4.1 (denoted R1-R5), and discuss prior work, which has so far been unable to satisfy all requirements (see Appendix A for a more detailed exposition). Further, we make the case that R1-R5 should be used as a first step (a *sanity check*) in evaluating future attempts to obtain practical provable FRL.

We remark that prior work, while not solving the problem of practical certificates for FRL which we focus on, establishes useful theoretical results, proposes methods with good accuracy-fairness tradeoffs, or provides weaker assurances which may be sufficient for different use-cases than ours.

Nature of bounds A practical certificate should define a *high-probability bound (R1)* that holds for a *concrete sample of datapoints* D collected by the data producer. This is in contrast to expectation bounds (such as e.g., Gitiaux & Rangwala (2021)), which imply bounded unfairness of downstream classifiers only in expectation (with respect to data sampling), and do not provide *any* assurance for a particular sample D in the practical case we observe.

The certificate should also be a *finite-sample bound (R2)*, as opposed to an asymptotic bound, which would hold only for $n \to \infty$, providing no strict assurance in a concrete case of fixed and limited n. An example of this is a large set of works (e.g., Shen et al. (2021); Gupta et al. (2021); McNamara et al. (2017); Madras et al. (2018)), where DP distance is soundly upper-bounded by a quantity not directly computable in practice (e.g., $BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h^\star)$), for which a more feasible approximation (e.g., a 2-layer neural network modeling h) is then optimized in training (approximately, using SGD which generally has no guarantees) (McNamara et al., 2017; Madras et al., 2018). As the quality of approximations is either not discussed or is of the asymptotic nature, these works can not offer practical certificates.

Limiting assumptions A practical certificate should be *distribution-free (R3)*, not requiring any assumptions on the distribution $\mathcal{X}$, as otherwise the certificate may not hold for $\mathcal{X}$ which the data producer encounters in practice. For instance, the certificates of Kairouz et al. (2022) are valid only if $\mathcal{X}$ is a mixture of Gaussians. Similarly, Balunović et al. (2022) provide finite sample bounds that hold given high-confidence density estimation of $\mathcal{X}$, which is possible only under restrictive assumptions such as Lipschitz (Wasserman, 2019) or α-Hölder continuity (Jiang, 2017).

A practical certificate should also be *model-free (R4)*, with no assumptions on the hypothesis class of the downstream classifiers $\mathcal{G}$. Having such assumptions overestimates fairness (see Section 2) and offers *no protection* against data consumers using a model outside this family, an aspect which the data producer is by design unable to control. For instance, Grünewälder & Khaleghi (2021); Tan et al. (2020) focus on providing certificates only for the family of Kernel models, which is an unrealistic restriction in our setting.

Empirical validation Finally, Definition 4.1 notes that a practical certificate should be *empirically tested (R5)*, i.e., it is important to demonstrate that for a realistic dataset and encoder, the certification procedure can return a real value $T^\star$, which is non-vacuous, i.e., at most 1, the maximum possible value of $\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}(g)$. Unfortunately, most prior work (McNamara et al. (2017); Madras et al. (2018), etc.) violates R5, often producing certificates whose actual value can not be exactly computed in practice. We strongly argue that a realistic empirical study is crucial to demonstrate the

practical value of a proposed certification method.

The statistical procedure we will propose in Section 5 satisfies R1-R5: it produces $T^\star$, a high-probability finite-sample bound on $\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}(g)$, with no restrictive assumptions. We apply the proposed procedure in Section 7 to several real datasets, obtaining decidedly non-vacuous values of $T^\star$, which are in some cases lower than the purely empirical unfairness values of some prior methods.

5. Deriving Practical Certificates for Restricted Encoders

We describe our key contribution, the derivation of practical fairness certificates for restricted encoders (formally defined shortly). In Section 6 we instantiate this approach with a particular encoder based on fair decision trees.

Let p_0 and p_1 denote the PDFs of $\mathcal{Z}_0$ and $\mathcal{Z}_1$ respectively, i.e., $p_0(\boldsymbol{z}_i) = P(\boldsymbol{z}_i|s=0)$ and $p_1(\boldsymbol{z}_i) = P(\boldsymbol{z}_i|s=1)$ and p denote the PDF of the marginal distribution of $\boldsymbol{z}$. Similarly, we use q for the marginal distribution of s, and q_i for the conditional distribution of s for $\boldsymbol{z} = z_i$, i.e., $q_i(0) = P(s=0|\boldsymbol{z}=z_i)$ and $q_i(1) = P(s=1|\boldsymbol{z}=z_i)$.

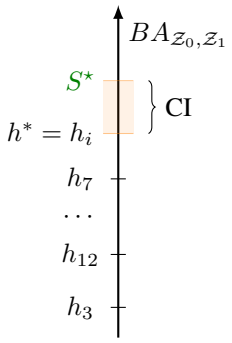


Figure 2: Restricted encoders enable practical certificates.

Restricted encoders As noted before, prior work is unable to directly utilize Equation (2) to obtain practical certificates, as for neural network encoders it is intractable to compute the densities $p_0(\boldsymbol{z})$ and $p_1(\boldsymbol{z})$ that define the optimal adversary h^*. To remedy this, we propose using *restricted encoders* $f\colon \mathbb{R}^d \to \{\boldsymbol{z}_1,\dots,\boldsymbol{z}_k\}$, i.e., encoders that map each $\boldsymbol{x}$ to one of k possible values (*cells*) $\boldsymbol{z}_i \in \mathbb{R}^{d'}$, i.e., $\mathcal{Z}$ is a discrete distribution with a finite support. We note that no prior work uses this concept—Zemel et al. (2013) map data to *prototypes*, but this mapping is probabilistic, thus incompatible with our definition.

As now there is a finite number of possible values for a representation, we can use a set D of n samples from $\mathcal{Z}$ (obtained by applying f to samples from $\mathcal{X}$) to analyze the optimal adversary h^* on each possible $\boldsymbol{z}$. Moreover, we can upper-bound its balanced accuracy (RHS of Equation (2)) on the whole distribution $\mathcal{Z}$ with some value $S^\star$ with high probability, using confidence intervals (CI, Figure 2). Finally, we can apply Equation (2) to obtain the certificate $\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}(g) \le 2S^\star - 1 = T^\star$. As in Definition 4.1, $T^\star$ has a dependency on D and n, which we omit for brevity. A detailed presentation of our certification procedure follows.

Upper-bounding the balanced accuracy Starting from Equation (1), we reformulate the balanced accuracy of $h^\star$:

$$
\begin{aligned}
&BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h^*)\\
&= \frac{1}{2}\left(\sum_{i=1}^{k} p_0(\boldsymbol{z}_i)\cdot[1-h^*(\boldsymbol{z}_i)] + \sum_{i=1}^{k} p_1(\boldsymbol{z}_i)\cdot[h^*(\boldsymbol{z}_i)]\right)\\
&= \frac{1}{2}\left(\sum_{i=1}^{k} \max\left(p_0(\boldsymbol{z}_i), p_1(\boldsymbol{z}_i)\right)\right)\\
&= \sum_{i=1}^{k} p(\boldsymbol{z}_i)\cdot\max\left(\underbrace{\frac{1}{2q(0)}}_{\alpha_0}\cdot q_i(0), \underbrace{\frac{1}{2q(1)}}_{\alpha_1}\cdot q_i(1)\right).
\end{aligned}
$$

The first line uses the definition of expectation of a discrete RV, the second the definition of $h^\star$, and the final line the two applications of Bayes' rule, namely $p_0(\boldsymbol{z}_i) = q_i(0)p(\boldsymbol{z}_i)/q(0)$ and $p_1(\boldsymbol{z}_i) = q_i(1)p(\boldsymbol{z}_i)/q(1)$. To upper bound $BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h^*)$ with high probability using given n samples, we focus on the final expression above, the prior-weighted (i.e., weighted by $p(\boldsymbol{z}_i)$) per-cell balanced accuracy (i.e., $\max(\alpha_0 q_i(0), \alpha_1 q_i(1))$ for each cell i), where we set $\alpha_0 = \frac{1}{2q(0)}$ and $\alpha_1 = \frac{1}{2q(1)}$.

Next, we introduce 3 lemmas, and later combine them to obtain the desired certificate. We use $B(p; v, w)$ to denote the p-th quantile of a beta distribution with parameters v and w. Note that for Lemma 5.1 we do not use the values $\boldsymbol{z}^{(j)}$ in the proof, but still introduce them for consistency.

Lemma 5.1 (Bounding base rates)**.** *Given n independent samples $\{(\boldsymbol{z}^{(1)}, s^{(1)}), \dots, (\boldsymbol{z}^{(n)}, s^{(n)})\} \sim \mathcal{Z}$ and a parameter ϵ_b, for α_0 and α_1 as defined above, it holds that*

$$
\alpha_0 < \frac{1}{2B(\frac{\epsilon_b}{2}; m, n-m+1)}, \quad \textit{and}
$$
$$
\alpha_1 < \frac{1}{2(1-B(1-\frac{\epsilon_b}{2}; m+1, n-m))},
$$

with confidence $1-\epsilon_b$, where $m = \sum_{j=1}^{n} \mathbb{1}\{s^{(j)}=0\}$.

Proof. We define n independent Bernoulli random variables $X^{(j)} := \mathbb{1}\{s^{(j)}=0\}$ with same unknown success probability $q(0)$. Using the Clopper-Pearson binomial CI (Clopper & Pearson (1934), see Appendix B) to estimate the probability $q(0)$ we get $P(q(0) \le B(\frac{\epsilon_b}{2}; m, n-m+1)) \le \epsilon_b/2$ and $P(q(0) \ge B(1-\frac{\epsilon_b}{2}; m+1, n-m)) \le \epsilon_b/2$. Substituting $q(0) = 1-q(1)$ in the latter, as well as the definitions of α_0 and α_1 in both inequalities recovers the inequalities from the lemma statement, which simultaneously hold with confidence $1-\epsilon_b$. □

Lemma 5.2 (Bounding balanced accuracy for a cell)**.** *Given n independent samples $\{(\boldsymbol{z}^{(1)}, s^{(1)}), \ldots, (\boldsymbol{z}^{(n)}, s^{(n)})\} \sim \mathcal{Z}$, constants $\bar{\alpha_0}$ and $\bar{\alpha_1}$ such that $\alpha_0 < \bar{\alpha_0}$ and $\alpha_1 < \bar{\alpha_1}$, and a parameter ϵ_c, it holds for each cell $i \in \{1, \ldots, k\}$, with total confidence $1 - \epsilon_c$, that*

$$\max(\alpha_0 \cdot q_i(0), \alpha_1 \cdot q_i(1)) \leq t_i, \tag{3}$$

where t_i is the maximum of $\bar{\alpha_0} B(1 - \frac{\epsilon_c}{2k}; m_i + 1, n_i - m_i)$ and $\bar{\alpha_1}(1 - B(\frac{\epsilon_c}{2k}; m_i, n_i - m_i + 1))$. Here, $n_i = |Z_i|$, and $m_i = \sum_{j \in Z_i} \mathbb{1}\{s^{(j)} = 0\}$, where $Z_i = \{j | \boldsymbol{z}^{(j)} = \boldsymbol{z}_i\}$.

Proof. As in Lemma 5.1, for each cell we use the Clopper-Pearson CI to estimate $q_i(0)$ with samples indexed by Z_i and confidence $1 - \epsilon_c/k$. As before, we apply $q_i(0) = 1 - q_i(1)$ to arrive at k inequalities of the form Equation (3), which per union bound jointly hold with confidence $1 - \epsilon_c$. □

Lemma 5.3 (Bounding the sum)**.** *Given n independent samples $\{(\boldsymbol{z}^{(1)}, s^{(1)}), \ldots, (\boldsymbol{z}^{(n)}, s^{(n)})\} \sim \mathcal{Z}$, where for each $j \in \{1, \ldots, n\}$ we define a function $idx(\boldsymbol{z}^{(j)}) = i$ such that $\boldsymbol{z}^{(j)} = \boldsymbol{z}_i$ (cell index), parameter ϵ_s, and a set of real-valued constants $\{t_1, \ldots, t_k\}$, it holds that*

$$P\left(\sum_{i=1}^{k} p(\boldsymbol{z}_i) t_i \leq S^\star\right) \geq 1 - \epsilon_s, \quad \textit{where}$$

$$S^\star = \frac{1}{n} \sum_{j=1}^{n} t_{idx(\boldsymbol{z}^{(j)})} + (b - a)\sqrt{\frac{-\log \epsilon_s}{2n}},$$

and we set $a = \min\{t_1, \ldots, t_k\}$ and $b = \max\{t_1, \ldots, t_k\}$.

Proof. For each j let $X^{(j)} := t_{idx(\boldsymbol{z}^{(j)})}$ denote a random variable. As for all j, $X^{(j)} \in [a, b]$ with probability 1 and $X^{(j)}$ are independent, we can apply Hoeffding's inequality (Hoeffding (1963), restated in Appendix B) to upper-bound the difference between the population mean $\sum_{i=1}^{k} p(\boldsymbol{z}_i) t_i = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{Z}} t_{idx(\boldsymbol{z})}$ and its empirical estimate $\frac{1}{n} \sum_{j=1}^{n} X^{(j)}$. Setting the upper bound such that the error is below ϵ_s directly recovers $S^\star$ and the lemma statement. □

Applying the lemmas We now describe how we apply the lemmas in practice to upper-bound $BA_{\mathcal{Z}_0, \mathcal{Z}_1}(h^\star)$, and in turn upper-bound $\Delta^{DP}_{\mathcal{Z}_0, \mathcal{Z}_1}(g)$ for any downstream classifier g. We assume a standard setting, where a set D of datapoints $\{(\boldsymbol{x}^{(j)}, s^{(j)})\}$ from $\mathcal{X}$ is split into a training set D_{train}, used to train f, validation set D_{val}, held-out for the upper-bounding procedure (and not used in training of f in any capacity), and a test set D_{test}, used to evaluate the empirical accuracy and fairness of downstream classifiers.

After training the encoder and applying it to produce representations $(\boldsymbol{z}^{(j)}, s^{(j)}) \sim \mathcal{Z}$ for all three data subsets, we aim to derive an upper bound on $\Delta^{DP}_{\mathcal{Z}_0, \mathcal{Z}_1}(g)$ for any g, that holds with confidence at least $1 - \epsilon$, where ϵ is the hyperparameter of the procedure (we use $\epsilon = 0.05$). To this end, we heuristically choose some decomposition $\epsilon = \epsilon_b + \epsilon_c + \epsilon_s$, and apply Lemma 5.1 on D_{train} to obtain upper bounds $\alpha_0 < \bar{\alpha_0}$ and $\alpha_1 < \bar{\alpha_1}$ with error probability ϵ_b. As mentioned above, using D_{train} in this step is sound as estimated probabilities $q(0)$ and $q(1)$ are independent of the encoder f. Next, we use $\bar{\alpha_0}$, $\bar{\alpha_1}$ and D_{val} in Lemma 5.2, to obtain upper bounds $t_1, \ldots, t_k$ on per-cell accuracy that jointly hold with error probability ϵ_c. Finally, we upper-bound the sum $\sum_{i=1}^{k} p(\boldsymbol{z}_i) t_i \leq S^\star$ with error probability ϵ_s using Lemma 5.3 on D_{test} with previously computed $t_1, \ldots, t_k$. Combining this with Equation (2) finally gives the certificate

$$\Delta^{DP}_{\mathcal{Z}_0, \mathcal{Z}_1}(g) \leq 2 \cdot BA_{\mathcal{Z}_0, \mathcal{Z}_1}(h^\star) - 1 \leq 2S^\star - 1 = T^\star, \tag{4}$$

which per union bound holds with desired error probability ϵ, with respect to the sampling process.

Practicality of the certificate According to requirements in Section 4, our certificate is practical, matching Definition 4.1. We provided a high-probability (R1) finite-sample (R2) bound, with no restrictive assumptions on $\mathcal{X}$ (R3) or the family of downstream classifiers (R4). Our procedure can be applied to *any* restricted f—as the certificate is derived for a fixed f, this is not an obstacle to certificate practicality, like R3 and R4. We further hypothesize that for a suitable instantiation of restricted f (i.e., the one we introduce in Section 6, based on decision trees) the accuracy-fairness tradeoffs obtained are favorable, and that our procedure leads to non-vacuous certificates on real datasets. We confirm this in our extensive experiments in Section 7 (satisfying R5).

6. Fair Decision Trees as Restricted Encoders

In this section, we instantiate a restricted encoder which allows for favorable accuracy-fairness tradeoffs along with practical certificates. We use decision trees, motivated by strong results of tree-based models on tabular data, their efficiency and interpretability (Borisov et al., 2021; Gardner et al., 2022), real-life uses (e.g., in finance (Ghatasheh, 2014; Guntay et al., 2022)), and their feature splitting procedure, which allows control over the tightness of our certificate.

In Appendix C we present an experiment with another class of restricted encoders based on k-means clustering. This substantiates our claim that our procedure can directly produce practical certificates for *any* restricted f, and illustrates that choosing a suitable restricted f is hard, as fairness-unaware encoders (e.g., k-means) likely lead to unfavorable results.

Classification trees Starting from the training set D_{root} of examples $(\boldsymbol{x}, y) \in \mathbb{R}^d \times \{0, 1\}$, a binary *classification tree* f repeatedly *splits* some leaf node P with assigned D_P, i.e., picks a split feature $j \in \{1, \ldots, d\}$ and

a split threshold v, and adds two nodes L and R as children of P, such that $D_L = \{(\boldsymbol{x}, y) \in D_P \mid x_j \leq v\}$ and $D_R = D_P \setminus D_L$. j and v are picked to minimize a chosen criterion, weighted by $|D_L|$ and $|D_R|$, aiming to produce y-homogeneous leaves. We focus on Gini impurity, computed as $Gini_y(D) = 2p_y(1-p_y) \in [0, 0.5]$ where $p_y = \sum_{(\boldsymbol{x},y)\in D} \mathbb{1}\{y=1\}/|D|$. At test time, an example $\boldsymbol{x}$ is propagated to a leaf l, predicting the majority class of D_l.

Trees as encoders Our key idea is to train a classification tree f with k leaves, encoding all examples in leaf i to the same $\boldsymbol{z}_i$. We construct $\boldsymbol{z}_i$ based on all training examples in leaf i, taking the median for continuous, and the most common value for categorical features (so we have $d' = d$).

Fairness-aware criterion Criteria such as $Gini_y(D)$ aim to maximize the accuracy by making the distribution of the label y in each leaf highly unbalanced. Using such a tree as an encoder leads to high unfairness, making it necessary to introduce a way to prioritize more fair tree structures.

We use, similar to Kamiran et al. (2010) and others (see Section 2), the criterion $FairGini(D) = (1-\gamma)Gini_y(D) + \gamma(0.5 - Gini_s(D)) \in [0, 0.5]$. The second term aims to *maximize* $Gini_s(D)$, i.e., make the distribution of s in each leaf i as close to uniform (making it hard to infer s from $\boldsymbol{z}_i$), while γ controls the accuracy-fairness tradeoff.

Fairness-aware categorical splits While usual splits of the form $x_j \leq v$ are suitable for continuous, they are inefficient for categorical (usually one-hot) variables, as only 1 category can be isolated. This increases the number of cells, making our certificates loose. Instead, we represent n_j categories for feature j as integers $c \in \{1, 2, ..., n_j\}$. To avoid evaluating all $2^{n_j} - 1$ possible partitions, we sort the values by $p_y(c) = \sum_{(\boldsymbol{x},y)\in D_c} \mathbb{1}\{y=1\}/|D_c|$ where we set $D_c = \{\boldsymbol{x} \in D \mid \boldsymbol{x}_j = c\}$, and consider all prefix-suffix partitions (*Breiman shortcut*).

This ordering focuses on accuracy and is provably optimal for $FairGini(D)$ with $\gamma = 0$ (Breiman et al., 1984). However, as it ignores fairness, it is inefficient for $\gamma > 0$. To alleviate this, we generalize the Breiman shortcut, and explore all prefix-suffix partitions under several orderings. Namely, for several values of the parameter q, we split the set of categories $\{1, 2, \ldots, n_j\}$ in q-quantiles with respect to $p_s(c)$ (defined analogous to $p_y(c)$), and sort each quantile by $p_y(c)$ as before, interspersing q obtained arrays to obtain the final ordering. While this offers no optimality guarantees, it is an efficient way to consider both objectives.

7. Experimental Evaluation

We evaluate FARE on several datasets, showing that it produces representations with fairness-accuracy tradeoffs comparable to prior work, while for the first time offering practical certificates. We then present several additional studies.

Experimental setup We consider common fairness datasets: Health (Kaggle, 2012), ACSIncome-CA (only California), and ACSIncome-US (US-wide) (Ding et al., 2021). The sensitive attributes are age and sex, respectively. We include the following FRL baselines: LAFTR (Madras et al., 2018), CVIB (Moyer et al., 2018), FCRL (Gupta et al., 2021), FNF (Balunović et al., 2022), sIPM (Kim et al., 2022), and FairPath (Shui et al., 2022). All omitted details regarding our experiments are given in Appendix E.

Main experiments We explore the fairness-accuracy tradeoff of each method by running it with various hyperparameters to obtain different representations, further used to train a 1-hidden-layer neural network (1-NN, other classifiers explored in Appendix H.1) for a certain prediction task, whose DP distance and prediction accuracy are then plotted. All hyperparameters of FARE are listed in Appendix E and we present an ablation study of the key parameter $\bar{k}$ in Appendix H.2, which can be used alongside our main results to guide parameter choices in practical applications of FARE.

We show a test set Pareto front for each method. For FARE, we also show a Pareto front of the certificate, i.e., a 95% confidence provable upper bound on DP distance (results with other metrics are deferred to Appendix D.2 and lead to similar conclusions). As noted before, no other method provides practical certificates that could be compared to FARE's. Finally, we include the Unfair Baseline, i.e., an identity encoder. The results are shown in Figure 3 and Figure 4. We omit FairPath and LAFTR from the main plots (see extended results in Appendix F), as LAFTR has stability issues (Gupta et al., 2021; Kim et al., 2022), and FairPath uses a different metric to us (Shui et al., 2022).

FARE achieves a better or comparable accuracy-fairness tradeoff compared to baselines. Crucially, the baselines cannot guarantee that there is no classifier with a worse DP distance when trained on their representations. This cannot happen for FARE—we compute a *provable* upper bound on DP distance of *any* such classifier. Our certificate is often comparable to *empirical* values of baselines. We note a small gap ($\leq 1.5\%$) between the best accuracy of FARE and the unfair baseline, indicating a tradeoff of restricted encoders—FARE computes a practical certificate, but loses some information, limiting the predictive power of classifiers. This is rarely a practical issue, as achieving meaningful fairness generally requires a non-trivial accuracy loss, especially when s and y are correlated. In Appendix H.3 we show that this gap is unaffected by dataset size. Finally, another important advantage of FARE is efficiency, with runtime of only several seconds, as opposed to minutes or hours for all other baselines (we measure this in Appendix G).

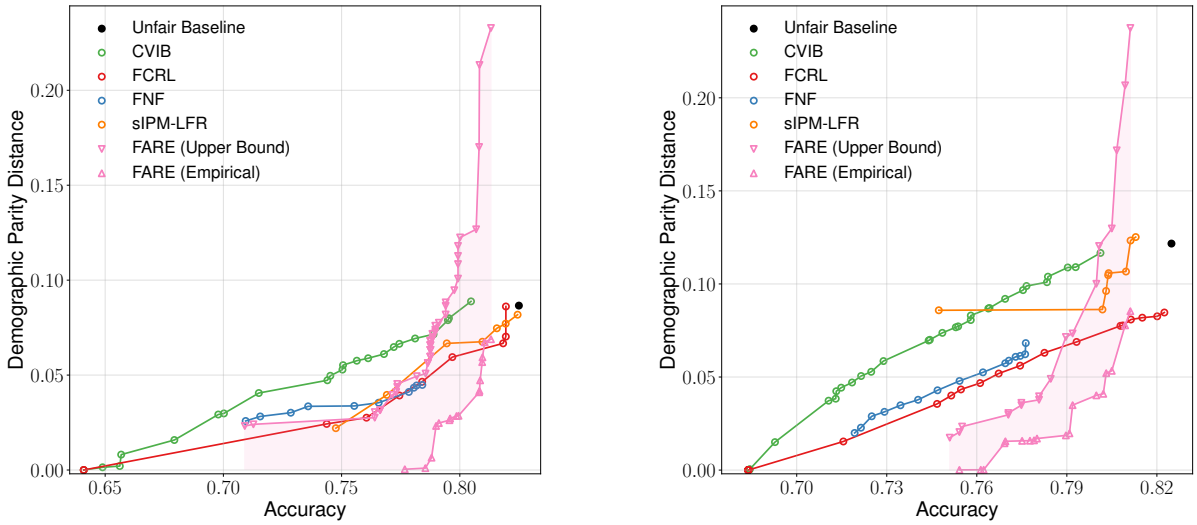


Figure 3: Evaluation of fair representation learning methods on ACSIncome-CA (left) and ACSIncome-US (right) datasets.

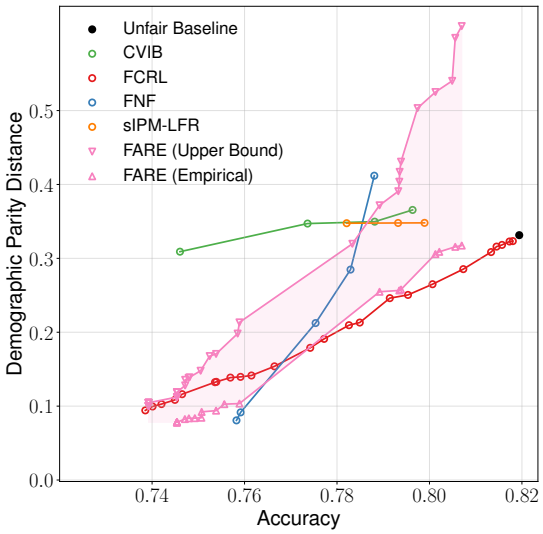


Figure 4: Evaluation of FRL methods on the Health dataset.

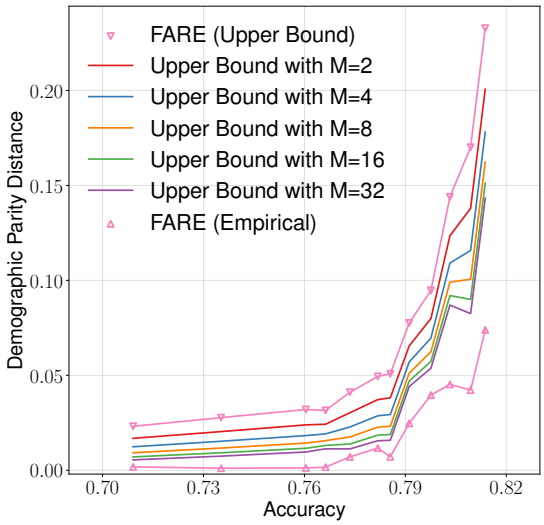


Figure 5: The impact of increasing the dataset size M times on the tightness of the FARE fairness certificate.

Data improves bounds In Figure 3 we see that FARE certificates are tighter for ACSIncome-US, suggesting that using more samples improves the bounds. To investigate this further in a controlled manner, we choose a representative set of FARE points from Figure 3 (left), and repeat the certification procedure with the dataset repeated M times, showing the resulting upper bounds for $M \in \{2, 4, 8, 16, 32\}$ in Figure 5. We observe a significant improvement in the certificate for larger M, further reinforcing our intuition that FARE is well-suited for large datasets and will benefit from ever-increasing amounts of data used in ML (Villalobos & Ho, 2022). Note that the bounds obtained for some $M > 1$ are only valid for the corresponding duplicated dataset (where we assume it was directly sampled), and do not hold on the original one, i.e., we cannot simply boost bound tightness on the original problem via data duplication. In Appendix H.4 we study the robustness of FARE to distribution shifts and missing data, and in Appendix H.5 show that it is robust to sensitive attribute imbalance.

Certificate validity In the next experiment we investigate a representative point obtained by FARE from Figure 4 with accuracy 79.3%, empirical DP of 26.2% and DP upper bound of 39.1%. In Figure 6, we show this point together with 24 other downstream classifiers (see Appendix E) obtained by training diverse model classes on the same representations (simulating a real use case where each data

consumer might prefer to use a different model). Half of the models are trained to maximize unfairness, and half to maximize accuracy. We make two observations. The left cluster shows that the data consumer can train models with high unfairness, especially in the worst case when they intentionally discriminate. The right cluster reaffirms a known limitation of prior work (Xu et al., 2020; Gupta et al., 2021): evaluating representations with some model class (triangle shows the 1-NN model from Figure 4) can underestimate unfairness, as data consumers using a different model class (e.g., SVM) can be more unfair. Both of these highlight the value of FARE which provides a practical certificate—all unfairness values still remain below a known upper bound.

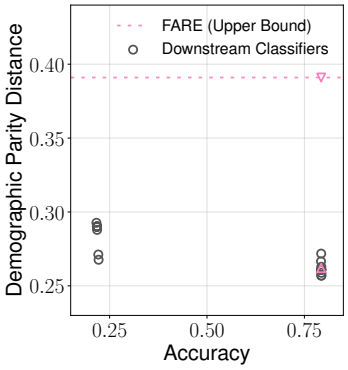


Figure 6: Downstream classifiers are below the FARE certificate.

Interpretability Another advantage of FARE is that its tree-based encoder enables direct interpretation of representations. To illustrate this, we explore a point from Figure 3 (right), with accuracy 75.1% and DP distance of 0.005, where we have $k = 6$. We can easily find that, for example, the representation z_6 is assigned to each person older that 24, with at least a Bachelor's degree, and an occupation in management, business or science. We note that this illustrative point has a particularly low value of k, greatly sacrificing accuracy for fairness—other Pareto front points often use larger values of k (see Appendix E).

Transfer learning Finally, we analyze the transferability of learned representations across tasks. We produce a diverse set of representations on the Health dataset with each method, and following the procedure from prior work (Madras et al., 2018; Balunović et al., 2022; Kim et al., 2022) evaluate them on five unseen tasks y (see Appendix E for details), where for each the goal is to predict a certain primary condition group. For each task and each method, we identify the highest accuracy obtained while keeping $\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}$ not above 0.20 (or 0.05). Moreover, we show $T^\star$, the provable DP distance upper bound of FARE.

The results are shown in Table 1. We observe that some methods are unable to reduce $\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}$ below the given threshold. FARE can always reduce $\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}$ sufficiently, but due to our restriction which enables the practical certificate, we often lose more accuracy for high $\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}$ thresholds.

y	$\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}$	$T^\star$	FARE	FCRL	FNF	sIPM
MIS	≤ 0.20	0.73	79.3	78.6	78.9	79.8
	≤ 0.05	0.57	78.7	78.6	78.7	78.6
NEU	≤ 0.20	0.72	73.2	72.4	71.9	76.6
	≤ 0.05	0.43	72.1	71.4	71.7	/
ART	≤ 0.20	0.55	74.4	70.7	68.9	78.3
	≤ 0.05	0.12	69.5	69.5	68.5	/
MET	≤ 0.20	0.48	69.8	69.2	75.0	/
	≤ 0.05	0.12	66.1	65.3	/	/
MSC	≤ 0.20	0.48	67.4	70.5	73.0	/
	≤ 0.05	0.12	63.0	/	/	/

Table 1: The results of transfer learning on Health.

8. Limitations and Future Work

Here we reflect on the limitations of FARE and highlight interesting avenues for future work. While FARE is the first FRL method to provide practical certificates, and its empirical tradeoffs are generally favorable, our results in Section 7 illustrate two main directions in which its results could be improved: (i) investigating and reducing the gap between the best achievable accuracy and the unfair baseline, (ii) tightening the certificate in high-accuracy regions. To this end, future work may attempt to extend the tree-based instantiation, or look for other more suitable instantiations of restricted encoders, analyzing their fundamental tradeoffs. Further, an important step for future work is the improvement of transfer learning results which are especially relevant in the FRL setting. Finally, important settings that we briefly study such as multivalued y and s (Appendix D.1) or robustness to distribution shifts (Appendix H.4) could be studied in more detail in the context of restricted encoders.

9. Conclusion

We introduced FARE, a method for provably fair representation learning with practical certificates. The key idea was that using restricted encoders enables a practical statistical procedure for computing a high probability finite-sample upper bound on the unfairness of any downstream classifier. We instantiated this idea with a tree-based encoder, and experimentally demonstrated that FARE can, for the first time, obtain tight fairness bounds on several datasets, while simultaneously producing empirical fairness-accuracy tradeoffs similar to prior work which offers no practical guarantees.

Acknowledgements

We thank Angéline Pouget and Nikola Konstantinov for helpful feedback on previous versions of this paper. This work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI) (SERI-funded ERC Consolidator Grant).

References

Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *CCS*, pp. 308–318. ACM, 2016.

Abebe, S. A., Lucchese, C., and Orlando, S. Eifffel: enforcing fairness in forests by flipping leaves. In *SAC*, 2022.

Aghaei, S., Azizi, M. J., and Vayanos, P. Learning optimal and fair decision trees for non-discriminative decision-making. In *AAAI*, 2019.

Balunović, M., Ruoss, A., and Vechev, M. T. Fair normalizing flows. In *ICLR*, 2022.

Barata, A. P. and Veenman, C. J. Fair tree learning. *CoRR*, 2021.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *California Law Review*, 2016.

Benaich, N. and Hogarth, I. State of ai report 2021. 2021. https://www.stateof.ai/2021, accessed: 2023-01-16.

Bhattacharyya, A., Gayen, S., Meel, K. S., Myrisiotis, D., Pavan, A., and Vinodchandran, N. V. On approximating total variation distance. *CoRR*, 2022.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. Deep neural networks and tabular data: A survey. *arXiv*, 2021.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. *Classification and Regression Trees*. Wadsworth, 1984.

Brennan, T., Dieterich, W., and Ehret, B. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 2009.

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *NeurIPS*, 2017.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *FAT*, 2019.

Cerrato, M., Köppel, M., Segner, A., and Kramer, S. Fair group-shared representations with normalizing flows. *CoRR*, 2022.

Chui, M., Hall, B., Singla, A., and Sukharevsky, A. The state of ai in 2021. 2021. https://www.mckinsey.com/capabilities/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021, accessed: 2023-01-16.

Clopper, C. J. and Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 1934.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *ACM SIGKDD*, 2017.

Devroye, L., Györfi, L., and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*. 1996.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. Empirical risk minimization under fairness constraints. In *NeurIPS*, pp. 2796–2806, 2018.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. Fairness through awareness. In *ITCS*, 2012.

Edwards, H. and Storkey, A. J. Censoring representations with an adversary. In *ICLR*, 2016.

Elazar, Y. and Goldberg, Y. Adversarial removal of demographic attributes from text data. In *EMNLP*, 2018.

EU. Proposal for a regulation laying down harmonised rules on artificial intelligence, 2021.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *KDD*, 2015.

Feng, R., Yang, Y., Lyu, Y., Tan, C., Sun, Y., and Wang, C. Learning fair representations via an adversarial framework. *CoRR*, 2019.

FTC. Aiming for truth, fairness, and equity in your company's use of ai, 2021. https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-\companys-use-ai, accessed: 2023-01-16.

Gardner, J., Popovic, Z., and Schmidt, L. Subgroup robustness grows on trees: An empirical baseline investigation. *CoRR*, 2022.

Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., and Vechev, M. T. AI2: safety and robustness certification of neural networks with abstract interpretation. In *IEEE Symposium on Security and Privacy*, pp. 3–18. IEEE Computer Society, 2018.

Ghatasheh, N. Business analytics using random forest trees for credit risk prediction: A comparison study. *International Journal of Advanced Science and Technology*, 2014.

Gitiaux, X. and Rangwala, H. Learning smooth and fair representations. In *AISTATS*, 2021.

Gitiaux, X. and Rangwala, H. Sofair: Single shot fair representation learning. In *IJCAI*, 2022.

Grari, V., Ruf, B., Lamprier, S., and Detyniecki, M. Achieving fairness with decision trees: An adversarial approach. *Data Sci. Eng.*, 2020.

Grünewälder, S. and Khaleghi, A. Oblivious data for fairness with kernels. *J. Mach. Learn. Res.*, 22:208:1–208:36, 2021.

Guntay, L., Bozan, E., Tigrak, U., Durdu, T., and Ozkahya, G. E. An explainable credit scoring framework: A use case of addressing challenges in applied machine learning. In *TEMSCON EUROPE*, 2022.

Gupta, U., Ferber, A. M., Dilkina, B., and Steeg, G. V. Controllable guarantees for fair outcomes via contrastive information estimation. In *AAAI*, 2021.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *NeurIPS*, 2016.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *JSTOR*, (301), 1963.

Jiang, H. Uniform convergence rates for kernel density estimation. In *ICML*, 2017.

Jin, J., Zhang, Z., Zhou, Y., and Wu, L. Input-agnostic certified group fairness via gaussian parameter smoothing. In *ICML*, 2022.

Kaggle. Health heritage prize, 2012. URL https://www.kaggle.com/c/hhp.

Kairouz, P., Liao, J., Huang, C., Vyas, M., Welfert, M., and Sankar, L. Generating fair universal representations using adversarial models. *IEEE TIFS*, 2022.

Kamiran, F., Calders, T., and Pechenizkiy, M. Discrimination aware decision tree learning. In *ICDM*, 2010.

Kang, M., Li, L., Weber, M., Liu, Y., Zhang, C., and Li, B. Certifying some distributional fairness with subpopulation decomposition. *CoRR*, 2022.

Kim, D., Kim, K., Kong, I., Ohn, I., and Kim, Y. Learning fair representation with a parametric integral probability metric. In *ICML*, 2022.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, 2017a.

Kleinberg, J. M., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, 2017b.

Konstantinov, N. and Lampert, C. H. Fairness through regularization for learning to rank. *CoRR*, 2021.

Lahoti, P., Gummadi, K. P., and Weikum, G. ifair: Learning individually fair data representations for algorithmic decision making. In *ICDE*, 2019.

Lechner, T., Ben-David, S., Agarwal, S., and Ananthakrishnan, N. Impossibility results for fair representations. *CoRR*, 2021.

Levin, D. A., Peres, Y., and Wilmer, E. L. *Markov chains and mixing times*. American Mathematical Society, 2006.

Li, P., Zou, J., and Zhang, L. Fairee: Fair classification with finite-sample and distribution-free guarantee. *CoRR*, 2022.

Liu, J., Li, Z., Yao, Y., Xu, F., Ma, X., Xu, M., and Tong, H. Fair representation learning: An alternative to mutual information. In *KDD*, 2022.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. S. The variational fair autoencoder. In *ICLR*, 2016.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. S. Learning adversarially fair and transferable representations. In *ICML*, 2018.

McNamara, D., Ong, C. S., and Williamson, R. C. Provably fair representations. *CoRR*, abs/1710.04394, 2017.

McNamara, D., Ong, C. S., and Williamson, R. C. Costs and benefits of fair representation learning. In *AIES*, 2019.

Meyer, A. P., Albarghouthi, A., and D'Antoni, L. Certifying robustness to programmable data bias in decision trees. In *NeurIPS*, 2021.

Moyer, D., Gao, S., Brekelmans, R., Galstyan, A., and Steeg, G. V. Invariant representations without adversarial training. In *NeurIPS*, 2018.

Oh, C., Won, H., So, J., Kim, T., Kim, Y., Choi, H., and Song, K. Learning fair representation via distributional contrastive disentanglement. In *KDD*, 2022.

Oneto, L., Donini, M., Pontil, M., and Maurer, A. Learning fair and transferable representations with theoretical guarantees. In *DSAA*, 2020.

Petersen, F., Mukherjee, D., Sun, Y., and Yurochkin, M. Post-processing for individual fairness. In *NeurIPS*, 2021.

Peychev, M., Ruoss, A., Balunovic, M., Baader, M., and Vechev, M. T. Latent space smoothing for individually fair representations. *CoRR*, 2021.

Raff, E., Sylvester, J., and Mills, S. Fair forests: Regularized tree induction to minimize model bias. In *AIES*, 2018.

Rannen-Triki, A., Berman, M., Kolmogorov, V., and Blaschko, M. B. Function norms for neural networks. In *ICCV Workshops*, 2019.

Ranzato, F., Urban, C., and Zanella, M. Fair training of decision tree classifiers. *CoRR*, 2021.

Roy, P. C. and Boddeti, V. N. Mitigating information leakage in image representations: A maximum entropy approach. In *CVPR*, 2019.

Ruoss, A., Balunovic, M., Fischer, M., and Vechev, M. T. Learning certified individually fair representations. In *NeurIPS*, 2020.

Sarhan, M. H., Navab, N., Eslami, A., and Albarqouni, S. Fairness by learning orthogonal disentangled representations. In *ECCV*, 2020.

Shamsabadi, A. S., Wyllie, S. C., Franzese, N., Dullerud, N., Gambs, S., Papernot, N., Wang, X., and Weller, A. Confidential-PROFITT: Confidential PROof of fair training of trees. In *ICLR*, 2023.

Shen, X., Wong, Y., and Kankanhalli, M. S. Fair representation: Guaranteeing approximate multiple group fairness for unknown tasks. *CoRR*, abs/2109.00545, 2021.

Shui, C., Chen, Q., Li, J., Wang, B., and Gagné, C. Fair representation learning through implicit path alignment. In *ICML*, 2022.

Simonite, T. When it comes to gorillas, google photos remains blind, 2018. https://www.wired.com/story/when-it-comes-to-gorillas-google-\photos-remains-blind/, accessed: 2023-01-16.

Song, J., Kalluri, P., Grover, A., Zhao, S., and Ermon, S. Learning controllable fair representations. In *AISTATS*, 2019.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 2012.

Tan, Z., Yeom, S., Fredrikson, M., and Talwalkar, A. Learning fair representations for kernel models. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 155–166. PMLR, 2020.

Villalobos, P. and Ho, A. Trends in training dataset sizes. https://epochai.org/blog/trends-in-training-dataset-sizes, 2022. Accessed: 2023-01-18.

Wang, J., Li, Y., and Wang, C. Synthesizing fair decision trees via iterative constraint solving. In *CAV*, 2022.

Wasserman, L. Density estimation. 2019. URL https://www.stat.cmu.edu/~larry/=sml/densityestimation.pdf.

Xie, Q., Dai, Z., Du, Y., Hovy, E. H., and Neubig, G. Controllable invariance through adversarial feature learning. In *NIPS*, pp. 585–596, 2017.

Xu, Y., Zhao, S., Song, J., Stewart, R., and Ermon, S. A theory of usable information under computational constraints. In *ICLR*, 2020.

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *ICML*, 2013.

Zhang, W. and Ntoutsi, E. FAHT: an adaptive fairness-aware decision tree classifier. In *IJCAI*, 2019.

Zhao, H. and Gordon, G. J. Inherent tradeoffs in learning fair representations. *J. Mach. Learn. Res.*, 2022.

Zhao, H., Coston, A., Adel, T., and Gordon, G. J. Conditional learning of fair representations. In *ICLR*. OpenReview.net, 2020.

A. Prior Work on Provably Fair Representation Learning

Here we describe prior work on FRL aiming to produce provable guarantees. Works that solely propose FRL methods or consider other theoretical aspects were covered in Section 2. We remark that most works, in addition to the requirement from Section 4 we will note is violated, usually also violate R5, providing no empirical evaluation of their certificate, often due to it not being exactly computable in practice.

The work of Feldman et al. (2015) was the first to establish the link between the balanced accuracy of the optimal adversary and fairness (which we have discussed in Section 3) but for the case of disparate impact (the 80% rule). They use this to motivate their method, which is in the category of in-processing methods (SVM training) and thus not FRL. McNamara et al. (2017) further establish the same link for demographic parity, as well as a similar formulation based on entropy. Their adversarial training method is inspired by this but states no formal certificate. Madras et al. (2018) provides proof of similar results for more metrics, and discusses the relationship to total variation distance. As other works, they use this to motivate their adversarial training approach, where, as discussed in Section 4 they approximately (with stochastic gradient descent) optimize an approximation of an uncomputable certificate. Zhao et al. (2020) also show that TV distance can bound fairness metrics but focus on simultaneously achieving equalized odds and accuracy parity, showing that perfect equalized odds imply bounds on demographic parity; their method is, similar to previous works, a min-max approximate optimization with no strict certificate. None of these works thus provide a practical certificate, at best violating requirements R1, R2 and R5.

The work of Shen et al. (2021) analyzes transferability between 7 group fairness notions under perfect or approximate fairness and discriminativeness. However, their method aims to minimize a finite sample estimate of MMD, which is introduced as an asymptotic approximation to TV distance (violating R2). As noted in Section 4, while reporting the obtained certificate (which the authors do) may be useful in some cases, the certificate is not valid in a strict sense, as the approximations are not accounted for.

Kairouz et al. (2022) state that perfect total variation distance implies perfect demographic parity. Further, they formulate the optimal adversary for each type of the reconstruction loss, but then note that it is not possible to compute those in practice (except for the restricted case when input distribution is e.g., a mixture of Gaussians, violating R3), and they simply use it as a motivation for min-max optimization of an adversary in training.

Kim et al. (2022) consider the family of integral probability metrics (IPM), and restrict the witness function class (compared to TV distance) to obtain a SigmoidIPM metric. As noted above, this effectively restricts the domain of classifiers (and is thus not model-free, violating R4) that the certificate applies to. The authors note this, and provide some classes for which the certificate would hold, however under the assumption of SGD optimality.

The works of Moyer et al. (2018); Song et al. (2019); Gupta et al. (2021) all focus on the unfairness bounds via mutual information, stating bounds on this uncomputable quantity and approximately optimizing Monte Carlo estimates, thus as noted above, producing no strict certificate. The missing further analysis of the approximations would likely lead to violations of R1 or R2.

Several works study provable FRL restricted to kernel models, thus not model-free (violating R4). Donini et al. (2018) study in-processing with kernel-based methods, and their results apply to FRL (as noted) only in the special case of linear models. Tan et al. (2020) provide model-aware bounds for kernel models under mild further assumptions, similar to Grünewälder & Khaleghi (2021) who optimize a relaxation of the MMD metric. All these methods are restricted by their dependency on the model class, as kernel models, despite having specific use-cases, are generally unable to obtain state-of-the-art results for common problems of interest, and are often not scalable to complex data.

Feng et al. (2019) bound the accuracy of the optimal Lipschitz-continuous adversary (though the general optimal adversary does not have to be Lipschitz continuous, thus violating R4) using Wasserstein distance (an instance of the IPM) and the Lipschitz constant. They perform adversarial training with K-Lipschitz adversaries, however it is not possible to exactly compute the Wasserstein distance in practice, and produce the certificate.

Gitiaux & Rangwala (2021) propose to add Gaussian noise to representations to obtain a finite sample two-sided expectation bound (violating R1) on how well DP distance can be approximated. The bound depends on the l_∞ norm of the encoder (obtained by upper bounding and approximating the ξ^2 mutual information). As computing this norm is NP-hard (Rannen-Triki et al., 2019), this is also not computable, making it practically impossible to satisfy R5.

Finally, Balunović et al. (2022) propose an FRL method where the encoder is a normalizing flow, which allows computation of probability densities in the latent space from densities in the input space, and bounding of the total variation distance. Under the assumption that the input density is known, this provides a practical certificate. However, as noted in Section 4, for this certificate to be valid it is necessary to estimate densities with high confidence, which is generally only feasible under restrictive assumptions on the distribution, which in turn violates R3.

Other theoretical contributions Finally, it is worth mentioning that a line of theoretical work studies other aspects of FRL besides designing fairness certificates, such as trade-offs between different notions, fairness-utility tradeoffs, and impossibility results (Kleinberg et al., 2017b; McNamara et al., 2019; Lechner et al., 2021; Zhao & Gordon, 2022).

B. Mathematical Details

We first derive Equation (2). More details can be found in Madras et al. (2018), and here we provide an overview:

$$\begin{aligned}\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}(g) &= |\mathbb{E}_{z\sim\mathcal{Z}_0}[g(z)] - \mathbb{E}_{z\sim\mathcal{Z}_1}[g(z)]| \\ &= |\mathbb{E}_{z\sim\mathcal{Z}_0}[-g(z)] + \mathbb{E}_{z\sim\mathcal{Z}_1}[g(z)]| \\ &= |\mathbb{E}_{z\sim\mathcal{Z}_0}[1-g(z)] + \mathbb{E}_{z\sim\mathcal{Z}_1}[g(z)] - 1| \\ &= |2BA_{\mathcal{Z}_0,\mathcal{Z}_1}(g) - 1|\end{aligned}$$

From this, we can argue that we can drop the absolute value and bound the balanced accuracy of g with the balanced accuracy of h^*, finally arriving at Equation (2).

Then, we formally state the Hoeffding's inequality and the Clopper-Pearson binomial confidence intervals, used in our upper-bounding procedure in Section 5.

Hoeffding's inequality (Hoeffding, 1963): Let $X^{(1)}, \ldots, X^{(n)}$ be independent random variables such that $P(X^{(j)} \in [a^{(j)}, b^{(j)}]) = 1$. Let $\hat{\mu} = \frac{X^{(1)}+\ldots X^{(n)}}{n}$ and $\mu = \mathbb{E}[\hat{\mu}]$. It holds that:

$$P(\mu - \hat{\mu} \geq t) \leq \exp\left(\frac{-2n^2t^2}{\sum_{i=1}^{n}(b^{(i)} - a^{(i)})^2}\right).$$

Clopper-Pearson binomial proportion confidence intervals (Clopper & Pearson, 1934): Assume a binomial distribution with an unknown success probability θ. Given m successes out of n experiments, it holds that:

$$B(\frac{\alpha}{2}; m, n-m+1) < \theta < B(1-\frac{\alpha}{2}; m+1, n-m) \quad (5)$$

with confidence at least $1-\alpha$ over the sampling process, where $B(p; v, w)$ denotes the p-th quantile of a beta distribution with parameters v and w. The Clopper-Pearson confidence interval has symmetric coverage probability, i.e., each side of Equation (5) holds with confidence $1-\alpha/2$.

C. K-means Restricted Encoders

To substantiate our key claim from Section 5 that our statistical procedure can be applied to any restricted encoder, we consider encoders based on *k-means clustering*—i.e., for a given k, we cluster the input representations, and map all examples from the same cluster to the respective cluster center. This fits our definition of restricted encoders given in Section 5, as the distribution $\mathcal{Z}$ has finite support.

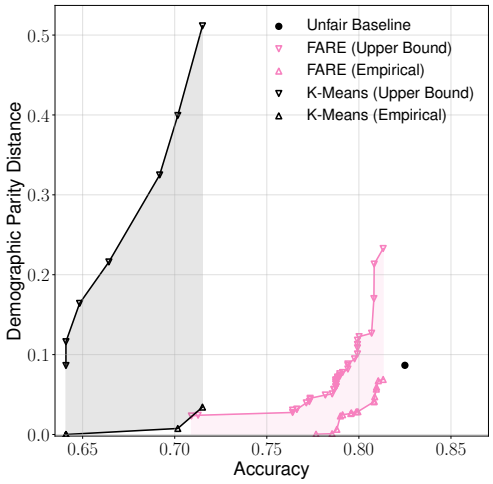


Figure 7: Comparing the k-means restricted encoders and FARE on ACSIncome-CA.

Results For $k \in [2, 500]$, in Figure 7 we report the DP distance (empirical and upper bound, i.e., the certificate) of the k-means restricted encoder on the ACSIncome-CA dataset, alongside the results of FARE from Figure 3 (left). While we were able to directly apply our statistical procedure to obtain upper bounds on unfairness, as k-means encoders are fairness-unaware, both empirical results and upper bounds are unfavorable and greatly outperformed by FARE. This illustrates that finding suitable classes of restricted encoders is not simple, and highlights the problem of finding other well-performing encoder classes.

D. Generalizations of FARE

In this section we discuss the generalizations of FARE beyond the settings considered in the main paper. In Appendix D.1 we focus on the relaxation of the requirements of binary classification and binary sensitive attribute, and in Appendix D.2 on additional unfairness metrics.

D.1. Beyond binary sensitive attributes and labels

We now describe how to extend FARE to the case of multivalued s and y, demonstrating that our method is not fundamentally limited to the binary setting. We use a common way to generalize DP distance, by considering the *maximal* absolute difference in prediction rates of *any* class y, w.r.t. *any* two sensitive groups i and j.

First, notice that our current certificate holds for any cardinality of y. Namely, we bound the unfairness of any binary classifier, and it can be easily seen that the most unfair clas-

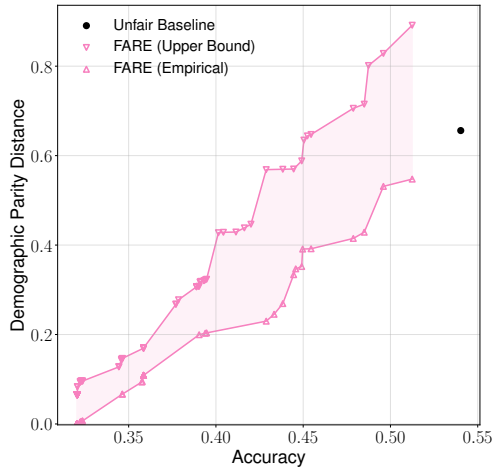


Figure 8: FARE can be generalized to multivalued y/s.

sifier will always be binary (i.e., will only predict 2 distinct classes, as this always maximizes the difference in prediction rates). Next, for non-binary s, we can simply invoke the same procedure for each pair of sensitive groups (i, j) by providing the procedure with only samples from those groups. Reducing ϵ in each of these invocations $\binom{|s|}{2}$ times leads to the same $1-\epsilon$ confidence for our certificate as before. Regarding our training procedure, while in the main paper we state the binary formulation of Gini impurity, the original definition directly supports multiple values of y/s, thus FARE training is already applicable to the general case.

Results To demonstrate that FARE can be applied to this setting on a real example, we provide preliminary results on a modified version of the ACSIncome-CA dataset, which represents 4-class classification (income classes thresholded at $[20k, 50k, 100k]$ dollars) and uses 3 sensitive groups (instead of sex we use a coarsening of race). We run FARE with above changes on this dataset and compute the certificate; we naively use the same hyperparameters as for ACSIncome-CA. The results are shown in Figure 8.

While we presented a straightforward extension, we believe future work on these settings may be able to further improve the tradeoffs and proof tightness by e.g., hyperparameter tuning or more elaborate extensions of the statistical procedure for the case of multiple sensitive groups.

D.2. Other fairness metrics

We demonstrate that FARE can easily be extended beyond demographic parity distance $\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}$, the unfairness metric used in our main results, by considering the commonly used *equal opportunity distance*

$$\Delta^{EOpp}_{\mathcal{Z}_0,\mathcal{Z}_1}(g) := \left|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z}_0^1}[g(\boldsymbol{z})] - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z}_1^1}[g(\boldsymbol{z})]\right|,$$

and the *equalized odds distance*

$$\Delta^{EO}_{\mathcal{Z}_0,\mathcal{Z}_1}(g) := \frac{1}{2}\sum_{y=0}^{1}\left|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z}_0^y}[g(\boldsymbol{z})] - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z}_1^y}[g(\boldsymbol{z})]\right|,$$

where we use $\mathcal{Z}_S^Y$ to denote the conditional distribution of $\boldsymbol{z}$ where $s = S$ and $y = Y$.

First, we need to extend our statistical procedure from Section 5 to compute certificates of other metrics. To this end, we observe that

$$\Delta^{EOpp}_{\mathcal{Z}_0,\mathcal{Z}_1} = \Delta^{DP}_{\mathcal{Z}_0^1,\mathcal{Z}_1^1},$$

and similarly

$$\Delta^{EO}_{\mathcal{Z}_0,\mathcal{Z}_1} = \frac{1}{2}\sum_{y=0}^{1}\Delta^{DP}_{\mathcal{Z}_0^y,\mathcal{Z}_1^y}.$$

This implies that the identical statistical procedure as used for the DP distance can be used to bound EOpp/EO distances, with the only change being the distribution it operates on, i.e., we restrict the data we provide to the statistical procedure to only samples where $y = Y$. Note that for EO we apply the procedure twice with $\epsilon/2$ to retain the total confidence of $1-\epsilon$, per union bound.

Next, we need to adapt the heuristics used in our tree-based instantiation, to optimize EOpp/EO instead of DP. To do this, we simply apply $Gini_s(D)$ in the definition of $FairGini(D)$ only to samples with $y = 1$ (for EOpp) and independently to samples with $y = 0$ and $y = 1$ (for EO), where we take a weighted sum based on the number of samples with each y.

Results We perform two experiments on the Health dataset (for EO and EOpp) with extensions listed above. For baselines, we run FNF with the EO/EOpp argument set, and use the same runs of other baselines as in Figure 4, as these methods do not provide code for metrics other than DP and generally leave such extensions to future work (which originally prompted our focus on DP as the most broadly considered metric).

The results are presented in Figure 9. In both cases, we can come to similar conclusions as in the corresponding DP experiment in Figure 4. FARE obtains favorable empirical tradeoffs, while being the only one with an unfairness certificate. The certificates are relatively tight, but progressively get looser for higher accuracies—this is much more pronounced for equalized odds.

E. Details of Experimental Evaluation

In this section we provide details of our experimental evaluation omitted from the main text.

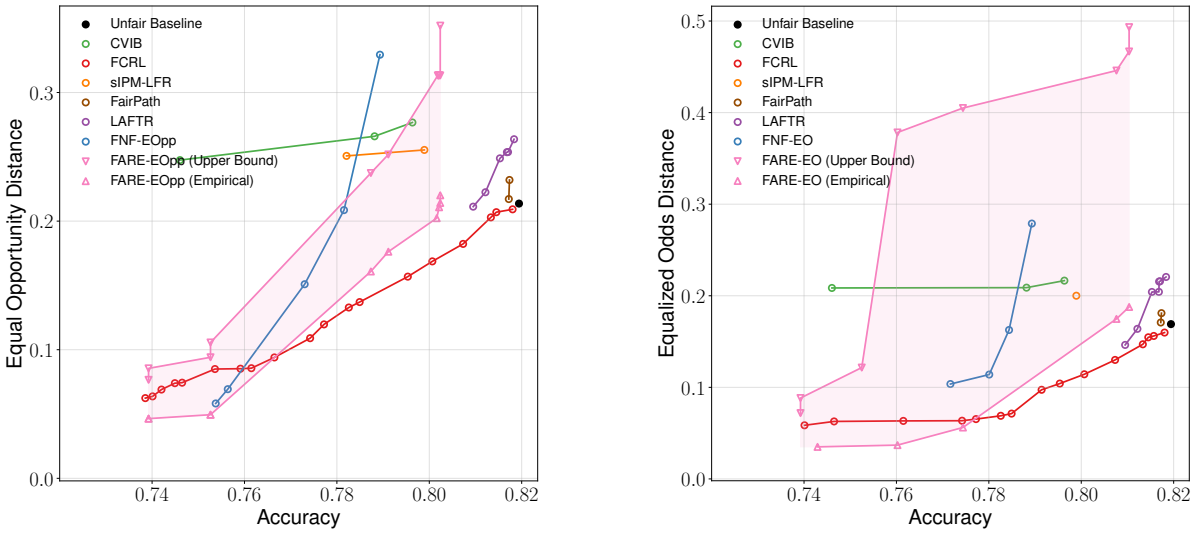


Figure 9: Generalizing FARE to other metrics: equal opportunity (left) and equalized odds (right), on the Health dataset.

Dataset	n_{train}	n_{test}	R_s	R_y
ACSIncome-CA	165 546	18 395	0.46	0.64
ACSIncome-US	1 429 070	158 786	0.48	0.68
Health	174 732	43 683	0.35	0.68

Table 2: Statistics of evaluated datasets.

Datasets As mentioned in Section 7, we perform our experiments on ACSIncome (Ding et al., 2021) and Health (Kaggle, 2012) datasets. In Table 2 we show some general statistics about the datasets: size of the training and test set (n_{train} and n_{test}), base rate for the sensitive attribute s (R_s, percentage of the majority group out of the total population), and base rate for the label y (R_y, accuracy of the majority class predictor).

ACSIncome is a dataset recently proposed by Ding et al. (2021) as an improved version of UCI Adult, with comprehensive data from US Census collected across all states and several years (we use 2014). The task is to predict whether an individual's income is above \$50,000, and we consider sex as a sensitive attribute. We evaluate our method on two variants of the dataset: ACSIncome-CA, which contains only data from California, and ACSIncome-US, which merges data from all states and is thus significantly larger but also more difficult, due to distribution shift. 10% of the total dataset is used as the test set. We also use the Health dataset (Kaggle, 2012), where the goal is to predict the Charlson Comorbidity Index, and we consider age as a sensitive attribute (binarized by thresholding at 60 years). For this dataset perform the same preprocessing as Balunović et al. (2022), and use 20% of the total dataset as the test set.

Evaluation procedure For our main experiments, as a downstream classifier we use a 1-hidden-layer neural network with hidden layer size 50, trained until convergence on representations normalized such that their mean is approximately 0 and standard deviation approximately 1. We train the classifier 5 times and in our main figures report the average test set accuracy, and the maximal DP distance obtained, following the procedure of Gupta et al. (2021).

Hyperparameters For baselines, we follow the instructions in respective writeups, as well as Gupta et al. (2021) to densely explore an appropriate parameter range for each value (linearly, or exponentially where appropriate), aiming to obtain different points on the accuracy-fairness curve. For CVIB, we explore $\lambda \in [0.01, 1]$ and $\beta \in [0.001, 0.1]$. For FCRL on ACSIncome we explore $\lambda = \beta \in [0.01, 2]$, and for Health $\lambda \in [0.01, 2]$ and $\beta = 0.5\lambda$. For FNF, we explore $\gamma \in [0, 1]$. For sIPM, we use $\lambda \in [0.0001, 1.0]$ and $\lambda_F \in [0.0001, 100.0]$, extending the suggested ranges. For FairPath we set the parameter $\kappa \in [0, 100]$. Finally, for LAFTR we use $g \in [0.1, 50]$, extending the range of $[0, 4]$ suggested by (Gupta et al., 2021). We adjust the parameters for transfer learning whenever supported by the method.

For FARE, there are four hyperparameters: γ (used for the criterion, where larger γ puts more focus on fairness), $\bar{k}$ (upper bound for the number of leaves), $\underline{n_i}$ (lower bound for the number of examples in a leaf), and v (the ratio of the training set to be used as a validation set). Note that all parameters affect accuracy, empirical fairness, and the tightness of the fairness bound. For example, larger $\underline{n_i}$ is likely to improve the bound by making Lemma 5.2 tighter, as more samples can be used for estimation. For the same reason, increasing v improves the tightness of the bound, but

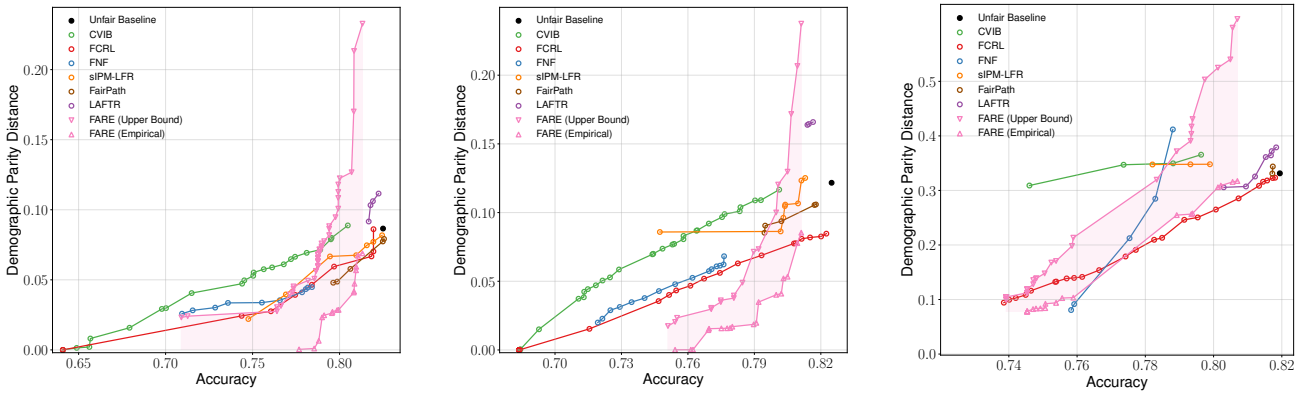


Figure 10: Extended evaluation of FRL methods on ACSIncome-CA (left), ACSIncome-US (middle) and Health (right).

y	$\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}$	$T^\star$	FARE	FCRL	FNF	sIPM
MIS	≤ 0.30	0.73	79.3	78.6	79.2	79.8
	≤ 0.20	0.73	79.3	78.6	78.9	79.8
	≤ 0.15	0.73	79.3	78.6	78.9	79.6
	≤ 0.10	0.59	79.0	78.6	78.9	79.0
	≤ 0.05	0.57	78.7	78.6	78.7	78.6
NEU	≤ 0.30	0.72	73.2	72.4	71.9	78.8
	≤ 0.20	0.72	73.2	72.4	71.9	76.6
	≤ 0.15	0.72	73.2	72.4	71.8	73.2
	≤ 0.10	0.42	73.1	72.2	71.8	/
	≤ 0.05	0.43	72.1	71.4	71.7	/
ART	≤ 0.30	0.55	74.4	70.7	68.9	78.3
	≤ 0.20	0.55	74.4	70.7	68.9	78.3
	≤ 0.15	0.48	74.2	70.1	68.9	/
	≤ 0.10	0.12	69.5	69.6	68.7	/
	≤ 0.05	0.12	69.5	69.5	68.5	/
MET	≤ 0.30	0.48	74.0	72.5	76.2	/
	≤ 0.20	0.48	69.8	69.2	75.0	/
	≤ 0.15	0.33	68.7	67.9	73.2	/
	≤ 0.10	0.12	66.1	66.7	73.2	/
	≤ 0.05	0.12	66.1	65.3	/	/
MSC	≤ 0.30	0.59	71.3	70.5	73.5	77.6
	≤ 0.20	0.48	67.4	70.5	73.0	/
	≤ 0.15	0.12	63.0	69.7	/	/
	≤ 0.10	0.12	63.0	69.0	/	/
	≤ 0.05	0.12	63.0	/	/	/

Table 3: Extended transfer learning results on Health.

may slightly reduce the accuracy as fewer samples remain in the training set used to train the tree. In our experiments we investigate $\gamma \in [0, 1]$, $\bar{k} \in [2, 200]$, $\underline{n_i} \in [50, 1000]$, $v \in \{0.1, 0.2, 0.3, 0.5\}$. For the upper-bounding procedure, we always set $\epsilon = 0.05$, $\epsilon_b = \epsilon_s = 0.005$, and thus $\epsilon_c = 0.04$. Finally, when sorting categorical features as described in Section 6, we use $q \in \{1, 2, 4\}$ in all cases.

Omitted details of additional experiments For the experiment in Figure 6 we explore the following classifiers: (i) 1-hidden-layer neural network (1-NN) with hidden layer sizes 50 and 200, (ii) 2-NN with hidden layers of size $(50, 50)$, as well as $(200, 100)$, (iii) logistic regression, (iv) random forest classifier with 100 and 1000 estimators, (v) decision tree with 100 and an unlimited number of leaf nodes. We train all these classifiers with a standardization preprocessing step as described above. We further train one variant of 1-NN, 2-NN, random forest, and logistic regression, on unnormalized data. All described models are trained both to predict the task label y, and to maximize unfairness, i.e., predict s, leading to 24 evaluated models.

For transfer learning (Table 1), the five transfer tasks represent prediction of the following attributes from the Health dataset: MISCHRT (MIS), NEUMENT (NEU), ARTHSPIN (ART), METAB3 (MET), MSC2A3 (MSC).

F. Extended Results

We provide the extended results of our main experiments, including two originally excluded methods, LAFTR and FairPath in Figure 10, corresponding to Figure 3 and Figure 4.

Additionally, in Table 3 we provide extended results of our transfer learning experiments, showing the accuracy values for thresholds $\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1} \in \{0.30, 0.20, 0.15, 0.10, 0.05\}$. We can observe similar trends as shown in Table 1.

G. Computational Efficiency Experiments

We investigate the computational efficiency and scalability of FARE in comparison to the baselines. We perform our measurements using the ACSIncome-CA dataset repeated M times (to observe the effects of data size, as in Figure 5). For each method, we use a single set of parameters (i.e., a single point from Figure 3, left). We use a *single core* of the i9-7900X CPU Intel CPU that has clock speed of 3.30GHz. All methods were given a single NVIDIA 1080 Ti GPU with 12 GB of VRAM, except FARE which does not require a GPU. We report CPU RAM, GPU RAM (VRAM) and

	Time			RAM			VRAM		
	M=1	M=4	M=16	M=1	M=4	M=16	M=1	M=4	M=16
FARE	**3 sec**	**11 sec**	**55 sec**	2.6 GB	3.0 GB	9.1 GB	**0.0 GB**	**0.0 GB**	**0.0 GB**
FNF	19 min	1 h 10 min	4 h 15 min	**2.3 GB**	4.3 GB	11.1 GB	1.6 GB	2.0 GB	3.2 GB
FCRL	57 min	3 h 51 min	15 h 38 min	2.4 GB	**2.4 GB**	**2.5 GB**	1.2 GB	2.7 GB	8.7 GB
CVIB	42 min	2 h 54 min	11 h 30 min	2.4 GB	**2.4 GB**	**2.5 GB**	1.2 GB	2.7 GB	8.7 GB
sIPM-LFR	23 min	1 h 35 min	OOM	3.1 GB	7.4 GB	>24.5 GB	2.0 GB	6.4 GB	>12 GB

Table 4: A study of computational efficiency of FRL methods.

runtimes for all methods for $M \in \{1, 4, 16\}$.

The results are shown in Table 4. As we noted in the main paper, FARE takes seconds to execute, which is orders of magnitude faster compared to all other methods, while also not requiring GPU support (i.e., we have VRAM requirements of 0.0 GB in Table 4). Further, FARE's runtime scales at most linearly with the dataset size, and its RAM usage scales well with M, and can easily be scaled to very large M on moderate hardware (as few 100s of GBs of CPU RAM are common on modern systems). In contrast, all other methods take hours for large M even with GPUs. FCRL, CVIB and sIPM-LFR use significant GPU RAM (scaling poorly with M). While FNF's GPU memory usage comparatively scales well with M, it still uses more CPU RAM than FARE.

H. Additional Experiments

In this section, we provide an ablation study of $\bar{k}$ and additional experiments investigating the robustness of FARE to different downstream classifiers, dataset size, distribution shifts, missing data, and sensitive attribute imbalance.

H.1. Experiments with different downstream classifiers

We compare different FRL methods on ACSIncome-CA, similarly to Figure 3 (left), in the case of different downstream classifiers. In Figure 12, we show the results on 4 additional downstream classifiers:

- Decision Tree with maximum 2500 leaves
- Random Forest using 100 trees
- Logistic Regression
- Two-layer neural network with 50 neurons per layer

We observe that the general trends observed in Figure 3 (left) for the different FRL methods hold regardless of the downstream classifier choice. We also see that the gap of our method to the maximum achievable accuracy is the smallest when the downstream classifier is a tree. This is unsurprising given that FARE's own representations are based on trees. Further, we see that the complex feature extraction of FCRL and sIPM enables higher accuracy than the unfair baseline in case of a simple classifier such as a decision tree.

H.2. Ablation study

We present an ablation study of $\bar{k}$ in Table 5. We explore 3 different settings of FARE: (i) *Fair*, with $\gamma = 0.999, \underline{n_i} = 1000, \upsilon = 0.5$, *Balanced*, with $\gamma = 0.85, \underline{n_i} = 100, \upsilon = 0.3$, and *Accurate*, with $\gamma = 0.3, \underline{n_i} = 10, \upsilon = 0.1$. In each setting we do a run with each $\bar{k} \in \{3, 5, 8, 20, 50\}$, and measure the accuracy, DP distance (unfairness), and $T^\star$, the DP distance certificate. We can observe that in all three settings, as expected, increasing k generally improves accuracy, but makes the certificate higher (and looser).

H.3. Performance gap on larger datasets

Next, we explore whether the small performance gap we observed in our main results, between FARE's most accurate model and the unfair baseline, widens for larger datasets. To this end, we merge two ACSIncome-US datasets from two consecutive years (2014 and 2015) and compare the results to the single year dataset from 2014, shown in Figure 3 (right). We note that the merged dataset has roughly 2x the number of data points. The comparison between the merged and single-year datasets is shown in Figure 13. We observe almost no difference between the results on the two datasets for the unfair baseline as well as the empirical and provable fairness of our method. This suggests that the complexity of the dataset is a more important factor than the data volume for the observed performance gap.

H.4. Distribution shifts and imputation

Next, we briefly consider two aspects that were not our main focus—the robustness to non-IID (distribution shift) and missing data. To measure distribution shift, we train FARE on ACSIncome-CA data from 2015 and compare its results when evaluated on test data from 2015 (the standard case) and 2016 (the *FARE Shifted* case). Regarding missing data,

	Fair			*Balanced*			*Accurate*		
$\bar{k}$	Acc.	DP Dist.	$T^\star$	Acc.	DP Dist.	$T^\star$	Acc.	DP Dist.	$T^\star$
3	0.735	**0.001**	**0.028**	0.786	**0.029**	**0.067**	0.786	0.081	**0.138**
5	0.760	**0.001**	0.035	0.800	0.057	0.105	0.799	0.074	0.150
8	0.773	0.007	0.041	0.800	0.057	0.113	0.802	0.080	0.163
20	0.774	0.008	0.072	0.803	0.045	0.144	0.810	0.062	0.229
50	**0.777**	0.004	0.082	**0.805**	0.049	0.191	**0.812**	**0.058**	0.303

Table 5: An ablation study of the FARE hyperparameter $\bar{k}$.

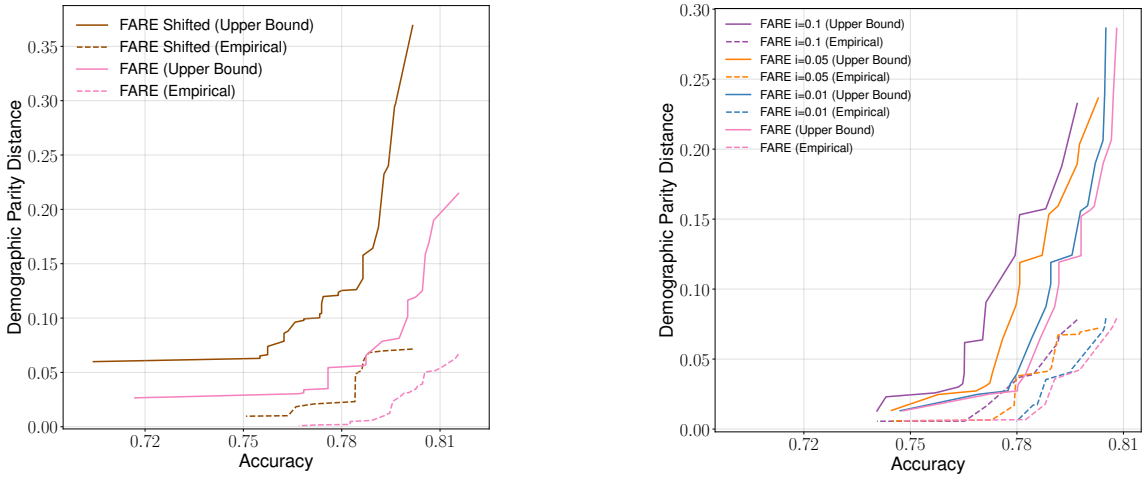


Figure 11: Exploring the robustness of FARE to distribution shifts and missing values.

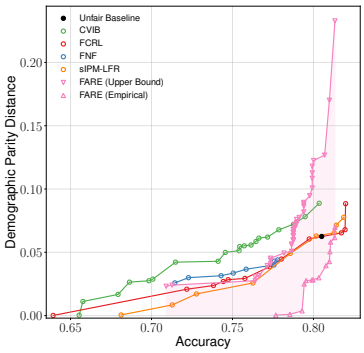
we use ACSIncome-US (as in Figure 3, right), and split its data in two parts, training FARE on the first part, and evaluating it (i.e., training downstream classifiers and computing the certificate) on the other part, where we randomly remove the fraction i of feature values, and impute them (using the most common value for categorical, and mean for continuous features). We use values $i \in \{0, 0.01, 0.05, 0.1\}$.

The results are shown in Figure 11. In both cases, as expected, we see some degradation of results—such concerns could be studied more in follow-up work.

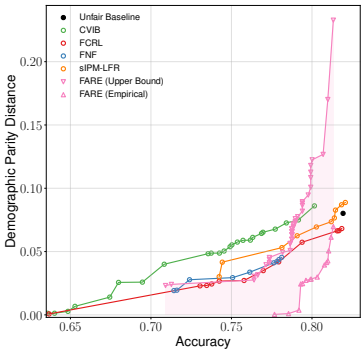
H.5. Effect of sensitive attribute imbalance

Finally, we study the effect of imbalance in the sensitive attribute on the fairness and accuracy of FARE. Let c denote the level of imbalance of each training set (i.e., the number of data points in the more common sensitive class divided by the total number of data points in the set). For each value of c we are interested in, we sample a random subset of size 49 053 from the original ACSIncome-CA training dataset (out of 165 546 data points in total), ensuring that the level of imbalance is exactly c. We use this number of samples, as this is the largest number for which we can have the same dataset size for each c, ensuring the fair comparison.

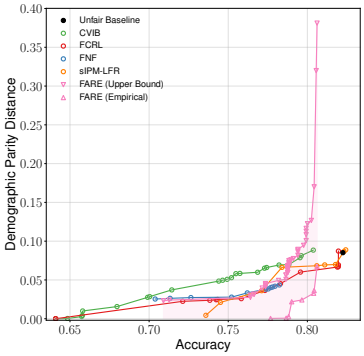
We train FARE on each subset separately and show Pareto plots, similar to those in Figure 3 and Figure 4, in Figure 14. We observe that FARE is robust to imbalance, as even for $c = 0.9$, we only see a small difference in our Pareto curves (and only in the low-accuracy regime).



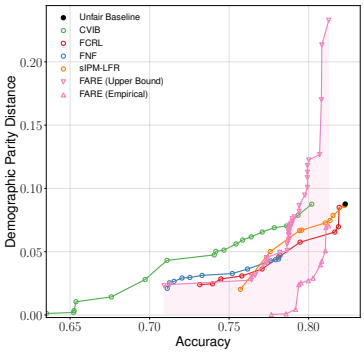
(a) Decision Tree



(b) Random Forest



(c) Logistic Regression



(d) Two-Layer NN

Figure 12: Comparison between different downstream classifiers on different FRL methods on ACSIncome-CA.

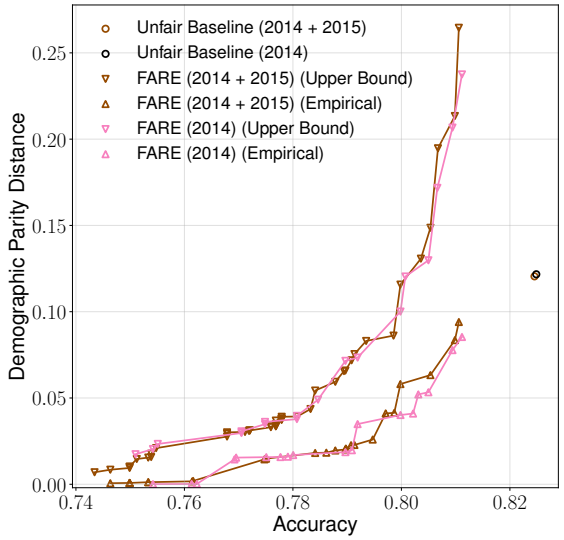


Figure 13: Comparison of the performance gap between FARE and the Unfair Baseline on ACSIncome-US for a single year and two years.

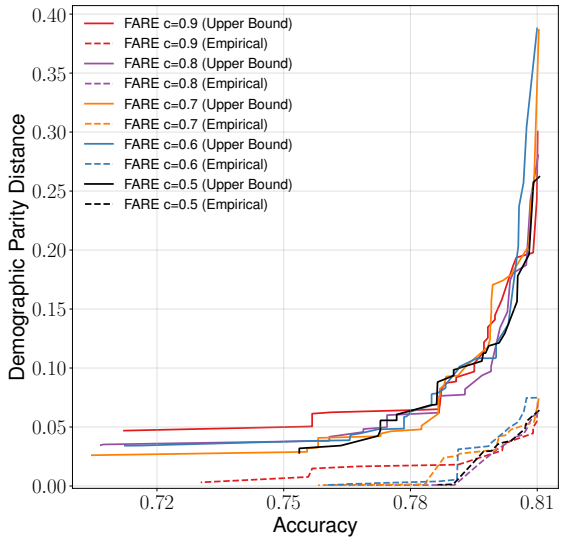


Figure 14: Evaluation of FARE at different levels of imbalance in the sensitive attribute (denoted by c) on randomly sampled subsets of ACSIncome-CA of the same size.