Auditing and Enforcing Conditional Fairness via Optimal Transport

Mohsen Ghassemi*[1], Alan Mishler*[1], Niccolò Dalmasso*[1], Luhao Zhang[2†]
Vamsi K. Potluru[1], Tucker Balch[1], Manuela Veloso[1]
[1]J.P.Morgan AI Research
[2]Department of Applied Mathematics and Statistics, Johns Hopkins University
{mohsen.ghassemi, alan.mishler, niccolo.dalmasso}@jpmchase.com, luhao.zhang@jhu.edu

Abstract

Conditional demographic parity (CDP) is a measure of the demographic parity of a predictive model or decision process when conditioning on an additional feature or set of features. Many algorithmic fairness techniques exist to target demographic parity, but CDP is much harder to achieve, particularly when the conditioning variable has many levels and/or when the model outputs are continuous. The problem of auditing and enforcing CDP is understudied in the literature. In light of this, we propose novel measures of conditional demographic disparity (CDD) which rely on statistical distances borrowed from the optimal transport literature. We further design and evaluate regularization-based approaches based on these CDD measures. Our methods, *FairBiT* and *FairLeap*, allow us to target CDP even when the conditioning variable has many levels. When model outputs are continuous, our methods target full equality of the conditional distributions, unlike other methods that only consider first moments or related proxy quantities. We validate the efficacy of our approaches on real-world datasets.

1 Introduction

Algorithmic decision-making has an increasing impact on individuals' lives, in areas such as finance, healthcare, and hiring. The growing field of algorithmic fairness aims to define, measure, and prevent discrimination in such systems.

Much of the early work in algorithmic fairness adopted as a fairness criterion *demographic parity* (DP) (aka statistical parity), which requires the outputs of a model or decision process to be statistically independent of a sensitive feature such as race, gender, or disability status (Calders, Kamiran, and Pechenizkiy 2009; Kamiran and Calders 2010; Kamiran, Karim, and Zhang 2012). Demographic parity remains arguably the most widely studied fairness criterion to date (Hort et al. 2024), but it can permit an algorithm to behave in intuitively unfair ways (Dwork et al. 2012; Hardt, Price, and Srebro, Nathan 2016). Consider the hypothetical loan approval process summarized in Table 1. The overall approval rate is 50% for both male and female applicants, but within each income level, males are more likely to get approved. If applicants within income levels are equally qualified, then this process appears to discriminate against females.

Conditional demographic parity (CDP) instead requires independence of the output and the sensitive feature conditional on a *legitimate* or *explanatory* feature or set of features, such as income in the loan example (Zliobaite, Kamiran, and Calders 2011; Kamiran, Žliobaitė, and Calders 2013). One way to achieve CDP is to use only the legitimate features as model inputs, but this may result in unacceptably poor predictive performance. When the legitimate features have a small number of levels, CDP can also be satisfied by maintaining a separate model for each subgroup defined by values of the legitimate feature and applying a method that targets DP to each subgroup. However, this approach may be infeasible when the legitimate feature has many levels.

To our knowledge, there is only one existing method designed to target conditional demographic parity even in the presence of many-valued legitimate features and/or continuous outputs (Xu et al. 2020a). This method uses a regularizer that is derived from a particular characterization of conditional independence. We show, however, that except when the regularizer is exactly zero, minimizing this quantity does not provide guarantees with respect to the actual disparities.

Additionally, quantifying (un)fairness in the sense of conditional demographic parity remains underexplored in the literature. As it is generally impossible to exactly satisfy a fairness criterion without sacrificing model performance (Corbett-Davies et al. 2017; Zhao and Gordon 2019), it is crucial to define appropriate measures of *conditional demographic disparity (CDD)*, i.e. to quantify violations of CDP. Quantifying CDD entails measuring disparities at every level of the legitimate features. While any existing measure of demographic disparity may be used at each level, aggregating the level-wise disparities to obtain a single CDD value requires careful attention. We are not aware of any previous work investigating this issue.

1.1 Contributions

We first introduce two new generic measures of CDD: the *CDD in the Wasserstein sense* and the *CDD in the ℓ^p sense*. Both measures take the entire (conditional) distributions into account and work for both classification and regression. We propose and evaluate regularizers designed to minimize different versions of these disparities.

For CDD in the Wasserstein sense, minimizing the disparity is nontrivial. We propose a regularizer based on the

*These authors contributed equally.
†Work done during an internship at J.P.Morgan AI Research.

Income Level	Approval Rate		Applicants	
	Male	Female	Male	Female
High Income	80%	60%	10%	60%
Medium Income	60%	40%	30%	30%
Low Income	40%	20%	60%	10%
Overall	50%	50%	100%	100%

Table 1: Toy example illustrating a process which satisfies demographic parity but not conditional demographic parity. Males and females have the same marginal ("Overall") loan approval rate (50%), but within each income level, males have higher approval rates than females.

bi-causal transport distance, a distributional distance recently studied in the optimal transport literature. We call this method *FairBiT: conditional Fairness through Bi-causal Transport*. For CDD in the ℓ_p sense, we propose *FairLeap: conditional Fairness in the ℓ_p sense*, where the regularizer is essentially a weighted ℓ_p norm of the vector of level-wise disparities.

Our regularization-based approaches provide practitioners with a tunable knob for navigating fairness-performance trade-offs (Zhao and Gordon 2019; Kim, Chen, and Talwalkar 2020; Liu and Vicente 2022). All our regularizers work regardless of whether the algorithm output is continuous or discrete, and since they utilize the entire dataset, they can be meaningfully applied even when the legitimate feature has many levels. Unlike the existing state of the art (Xu et al. 2020a), our proposed methods do not require access to the sensitive feature at inference time. We show that our methods generally provide better fairness-performance trade-offs than other methods on a range of datasets.

1.2 Related Work

Demographic parity (DP) and conditional demographic parity (CDP) There are a vast number of methods designed to target DP (Hort et al. 2024), but there has been very little investigation of CDP (Zliobaite, Kamiran, and Calders 2011; Kamiran, Žliobaitė, and Calders 2013). A natural way to target CDP is to stratify on the legitimate feature and apply methods designed to achieve DP within each stratum. However, this may result in poor overall model performance, especially if there is a small amount of data in each level. Our methods utilize the entire dataset during model training. Additionally, the majority of methods for DP are designed for classification or only designed to equalize the first moments of the two distributions, whereas we target full conditional independence in both classification and regression settings.

The closest work to ours is Xu et al. (2020a), who propose DCFR, the Derivable Conditional Fairness Regularizer. DCFR also accommodates rich legitimate features and continuous outcomes and also involves a regularizer applied to the entire dataset. However, we show in Section 5.1 that the regularizer is equivalent to a proxy quantity that is distinct from the disparity of interest. Small values of the proxy quantity do not necessarily imply small values of the disparity. By contrast, we utilize regularizers that directly target the distances between the relevant conditional distributions. Additionally, DCFR requires access to the sensitive feature at inference time, which our methods do not.

Fairness for binary vs. continuous model outputs The majority of the algorithmic fairness literature considers classification settings in which the final model outputs are binary, though there is a growing set of methods that can handle regression settings (Chzhen et al. 2020; Chzhen and Schreuder 2022; Romano, Bates, and Candes 2020; Fukuchi and Sakuma 2024; Xian et al. 2024; Coston, Rambachan, and Chouldechova 2021; Mishler and Kennedy 2021; Franklin et al. 2023; Jin and Lai 2023). Many methods which can handle continuous outputs only aim to equalize the first moments of the output distributions across levels of the sensitive feature (e.g. the average loan amount for males vs. females); a smaller but growing number target full equality of the output distributions (Chzhen et al. 2020; Chzhen and Schreuder 2022; Romano, Bates, and Candes 2020; Fukuchi and Sakuma 2024). Our methods falls in the latter category, aiming to ensure equality of conditional output distributions across levels of the legitimate feature, regardless of whether the output is discrete or continuous.

Optimal transport for fairness Optimal transport has been increasingly used for fairness-related applications, including fair edge prediction, training fair predictors, and uncovering discrimination in predictors (Black, Yeom, and Fredrikson 2020; Silvia et al. 2020; Laclau et al. 2021; Si et al. 2021; Zehlike, Hacker, and Wiedemann 2020; Chiappa and Pacchiano 2021; Buyl and De Bie 2022; Yang et al. 2022b; Rychener, Taskesen, and Kuhn 2022; Jiang et al. 2020; Johndrow and Lum 2019; Gordaliza et al. 2019; Jiang et al. 2020; Rychener, Taskesen, and Kuhn 2022; Miroshnikov et al. 2022; Jourdan et al. 2023; Langbridge, Quinn, and Shorten 2024; Wang, Nguyen, and Hanasusanto 2024). Optimal transport techniques generally require the specification of a distributional distance function. Most papers that apply optimal transport to fairness rely on the *Wasserstein distance*. Our proposed methods utilize the Wasserstein distance at each level of the legitimate feature, and aggregate these distances using an outer measure. One of our disparity measures, the CDD in the Wasserstein sense (Definition 3.1) gives rise to a nested Wasserstein distance; to target this disparity, we utilize a distance known as the *bi-causal transport distance* (Backhoff et al. 2017). This distance does not appear to have been used previously in the context of algorithmic fairness.

2 Notation and Problem Setup

Throughout, we consider data drawn from a distribution $(X, A, Y) \sim \mathbb{P}$, where $X \in \mathcal{X}$ is a set of features, $A \in \{0, 1\}$ is a binary sensitive feature, and $Y \in \mathbb{R}$ is a prediction target. We use "$\perp\!\!\!\perp$" to denote statistical independence. In a slight abuse of notation, we use $\mathbb{P}$ to refer both to the probability measure and to its density, assuming the density is defined. We let $f : \mathcal{X} \to \mathbb{R}$ be a model whose (un)fairness we wish to measure. In practice, $f(X)$ might map to a prediction that is used downstream in some decision process, or it might map to an automated decision such as a loan approval decision. Everything that follows applies in either setting.

Definition 2.1 (Demographic parity (DP))**.** The model $f(X)$ satisfies *demographic parity (DP)* if $f(X) \perp\!\!\!\perp A$, or in other words if $\mathbb{P}(f(X) \mid A=0) \equiv \mathbb{P}(f(X) \mid A=1)$. (Agarwal, Dudík, and Wu 2019).

DP takes into consideration only the sensitive feature and model outputs; it is indifferent to the features X. As discussed in the introduction and illustrated in Table 1, this means that a model which satisfies DP may treat different groups differently within levels of X, which may result in intuitively unfair behavior. This motivates *conditional demographic parity* (Kamiran, Žliobaitė, and Calders 2013; Corbett-Davies et al. 2017; Ritov, Sun, and Zhao 2017) as an alternative.

Definition 2.2 (Conditional demographic parity (CDP))**.** The model $f(X)$ satisfies conditional demographic parity with respect to a feature or set of features $L \subset X$ if $f(X) \perp\!\!\!\perp A|L=l$, for all $l \in \operatorname{supp}\mathbb{P}(L|A=0) \cap \operatorname{supp}\mathbb{P}(L|A=1)$.

We follow Corbett-Davies et al. (2017) in referring to L as the *legitimate feature(s)*. In the loan approval example (Table 1), L would be the income level. We emphasize that the choice of L, and the choice of the sensitive feature A, are up to the user.

Throughout the remainder of the paper, we assume that $\operatorname{supp}\mathbb{P}(L|A=0) = \operatorname{supp}\mathbb{P}(L|A=1)$, i.e. the legitimate features have the same support for both groups represented by the sensitive feature. Since any method targeting CDP requires multiple samples at each level of L, we further assume that L is either naturally discrete or appropriately discretized. See Appendix B.2 for a discussion of this assumption.

In this work, we investigate how to *effectively promote conditional demographic parity in supervised learning problems*. A crucial step towards enforcing parity is to choose an appropriate disparity measure. We discuss this next.

3 Measuring disparities

Quantifying violations of parity informs the development of fairness methods, and it allows the fairness of different models and methods to be compared on a continuous scale. Any definition of the *conditional demographic disparity* (CDD), the violation of CDP, must take into account the conditional distributions for each possible level l of the legitimate features, as well as how to aggregate these level-wise features to output a single value.

In principle, we could utilize any measure of *demographic disparity* (the violation of DP) from the literature to measure violations at each level l. Most previous work defines the demographic disparity by $|\mathbb{E}[f(X) \mid A=1] - \mathbb{E}[f(X) \mid A=0]|$ or by $\mathbb{E}[f(X) \mid A=1]/\mathbb{E}[f(X) \mid A=0]$, considering only the first moments of the distributions. (See Appendix B.1.) We instead consider distances between the full conditional distributions. We introduce two notions of the conditional demographic disparity, *CDD in the Wasserstein sense* and *CDD in the ℓ^p sense*. Here, "ℓ^p" and "Wasserstein" refer to the aggregation method once the level-wise distances between the conditional distributions are obtained.

For any $p \in [1,\infty)$, let $\mathcal{W}_p(\cdot,\cdot;D)$ denote the p-Wasserstein distance with cost function D, so that $\mathcal{W}_p^p$ denotes $\mathcal{W}_p$ taken to the power p. The Wasserstein distance represents the smallest possible cost to transport all the probability mass from one distribution to another given cost function D. See Appendix A for more details.

Definition 3.1 (CDD in the Wasserstein sense)**.** Let $\mathsf{d}(\cdot,\cdot)$ denote a distance between distributions, let $p \in [1,\infty)$, and let $\mathcal{L}$ be the support of L. The conditional demographic disparity in the Wasserstein sense ($\mathsf{CDD}^{\mathsf{wass}}$) for model $f(X)$ with legitimate features L and sensitive feature A is

$$\mathsf{CDD}^{\mathsf{wass}}(f) := \mathcal{W}_p^p(\mathbb{P}(L|A=0), \mathbb{P}(L|A=1); D),$$

where the cost function $D : \mathcal{L} \times \mathcal{L} \mapsto \mathbb{R}$ is defined as

$$D(l,l') = \begin{cases} \mathsf{d}\big(\mathbb{P}(f(X) \mid L=l, A=0), \\ \qquad \mathbb{P}(f(X) \mid L=l', A=1)\big), & l=l', \\ \infty, & \text{elsewhere.} \end{cases}$$

Different distance functions d induce different notions of disparity. In particular, when d itself is the Wasserstein distance $\mathcal{W}_p^p$, the CDD in the Wasserstein sense becomes a nested Wasserstein distance. This nested Wasserstein distance is tightly related to the *bi-causal transport distance*, recently studied in the optimal transport literature. While this method of aggregating the level-wise disparities may seem unintuitive at first, we will see in Section 4 that the bicausal distance naturally captures the conditional independence relationships that define conditional demographic parity.

Definition 3.2 (CDD in the ℓ_p sense)**.** Let $\mathsf{d}(\cdot,\cdot)$ denote a distance between distributions, let $p \in [1,\infty)$, and let $\mathcal{L}$ be the support of L. Define $D : \mathcal{L} \mapsto \mathbb{R}$ as $D(l) =$

$$\mathsf{d}\left(\mathbb{P}(f(X) \mid L=l, A=0), \mathbb{P}(f(X) \mid L=l, A=1)\right).$$

The conditional demographic disparity in the ℓ_p sense (CDD^{ℓ_p}) of the model $f(X)$ is $\mathsf{CDD}^{\ell_p}(f) := \|D\|_{\ell^p(\mathcal{L};\mathbb{Q}(L))}$ where $\mathbb{Q}(L)$ is a probability measure defined over $\mathcal{L}$.

The disparity in Definition 3.2 is the weighted ℓ_p norm of the distances between the conditional distributions defined by levels l of the legitimate features, where the associated measure $\mathbb{Q}$ determines the weight assigned to each level.

In the definitions above, different values of d, p, and $\mathbb{Q}$ induce different notions of disparity. There is a growing literature that investigates how different quantitative notions of (un)fairness capture or fail to capture various legal and philosophical notions of fairness (Friedler, Scheidegger, and Venkatasubramanian 2021; Bothmann, Peters, and Bischl 2024). In Appendix B.3, we suggest some desiderata that are relevant to choosing these values, but we leave a fuller investigation for future work. In our regularizers and in our experiments below, we set d to be the Wasserstein distance for both definitions, and we fix $p=1$ for CDD^{ℓ_p} and $p=2$ for the inner and outer Wasserstein distances in $\mathsf{CDD}^{\mathsf{wass}}$. We investigate several versions of $\mathbb{Q}(L)$ for CDD^{ℓ_p}.

4 Regularized approaches to enforce CDP

In light of the discussion in Section 3, a natural approach to enforce CDP in risk minimization problems is to employ the CDD measures in Definitions 3.1 and 3.2 as regularizers. In particular, given an i.i.d. training sample $\{(X_i, Y_i, A_i)\}_{i=1}^N$, we consider the following target problem:

$$\min_{f_\theta \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} g(f_\theta(X_i), Y_i) + \lambda \,\mathsf{CDD}(f_\theta), \tag{1}$$

where g is a differentiable loss function, $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ is the class of models $f_\theta(X)$ under consideration, indexed by θ, $\lambda > 0$ is a penalty parameter to encourage fairness, and CDD represents either $\mathsf{CDD}^{\mathsf{wass}}$ or CDD^{ℓ_p}.

4.1 Enforcing CDD in the Wasserstein sense

The CDD measure $\mathsf{CDD}^{\mathsf{wass}}$ contains a discontinuous inner cost function and is non-differentiable; additionally, computing the Wasserstein distance exactly is known to be computationally intractable (Arjovsky, Chintala, and Bottou 2017; Salimans et al. 2018). To tackle these challenges, we leverage an interesting connection between $\mathsf{CDD}^{\mathsf{wass}}$ and the *bi-causal transport distance*, recently studied in the optimal transport literature (Backhoff et al. 2017). This allows us to find tractable approximations to $\mathsf{CDD}^{\mathsf{wass}}$ such that the resulting minimization problem can be solved using gradient-based methods.

Connecting $\mathsf{CDD}^{\mathsf{wass}}$ to bi-causal transport distance In order to define the bi-causal transport distance, we first need to define transport plans in general and bi-causal transport plans in particular. Consider two distributions $(U, V) \sim \mathbb{P}$ and $(\tilde{U}, \tilde{V}) \sim \tilde{\mathbb{P}}$ on a common measure space $\mathcal{U} \times \mathcal{V}$. The set $\Gamma(\tilde{\mathbb{P}}, \mathbb{P})$ of *transport plans* denotes the collection of all probability measures on the space $(\mathcal{U} \times \mathcal{V}) \times (\mathcal{U} \times \mathcal{V})$ with marginals $\tilde{\mathbb{P}}$ and $\mathbb{P}$. The set $\Gamma_{bc}(\tilde{\mathbb{P}}, \mathbb{P}) \subset \Gamma(\tilde{\mathbb{P}}, \mathbb{P})$ of *bicausal transport plans* is given by

$$\Gamma_{bc}(\tilde{\mathbb{P}}, \mathbb{P}) = \{\gamma \in \Gamma(\tilde{\mathbb{P}}, \mathbb{P}) \text{ s.t. for } ((\tilde{U}, \tilde{V}), (U, V)) \sim \gamma, \\ U \perp\!\!\!\perp \tilde{V} \mid \tilde{U} \text{ and } \tilde{U} \perp\!\!\!\perp V \mid U\}.$$

We are now ready to define the bi-causal transport distance.

Definition 4.1 (Bi-causal transport distance (BCD))**.** The *bi-causal transport distance* (BCD, referred to hereafter simply as the *bi-causal distance*) between $\tilde{\mathbb{P}}$ and $\mathbb{P}$, denoted by $\mathcal{C}_b^p(\tilde{\mathbb{P}}, \mathbb{P})$, is defined as

$$\inf_{\gamma \in \Gamma_{bc}(\tilde{\mathbb{P}}, \mathbb{P})} \mathbb{E}_{((\tilde{U}, \tilde{V}), (U, V)) \sim \gamma} \left[C \|\tilde{U} - U\|^p + \|\tilde{V} - V\|^p \right].$$

See Appendix A for more background on bi-causal transport. In our setting, U will correspond to the legitimate feature L, and V will correspond to the model output $f(X)$. The presence or absence of the tilde corresponds to the two levels of the sensitive feature A. The following result connects BCD to Definition 3.1 of CDD.

Proposition 4.2. *Consider a model $f : \mathcal{X} \to \mathcal{Y}$. Let $\mathcal{L}$ be the support of the legitimate feature L. Let $\underline{d} = \min_{l, l' \in \mathcal{L}} \|l - l'\|_p^p$ denote the minimum distance between two levels of the legitimate features. Moreover, let $\bar{d} = \max_{x, x' \in \mathcal{X}} \|f(x) - f(x')\|_p^p$ denote the diameter of $\mathcal{Y}$. Then, the bi-causal distance $\mathcal{C}_b^p(\mathbb{P}(f(X), L|A = 0), \mathbb{P}(f(X), L|A = 1))$ with $C > \frac{\bar{d}}{\underline{d}}$ is equivalent to $\mathsf{CDD}_f^{\mathsf{wass}}$ with $\mathsf{d}(\cdot, \cdot) = \mathcal{W}_p^p(\cdot, \cdot)$.*

See Appendix E for the proof. This result implies that for $C > \frac{\bar{d}}{\underline{d}}$, the bi-causal distance is equivalent to $\mathsf{CDD}^{\mathsf{wass}}$ with the inner metric $\mathsf{d}(\cdot, \cdot)$ equal to $\mathcal{W}_p^p(\cdot, \cdot)$. We describe our BCD-regularized approach next.

FairBiT: Conditional Fairness through Bi-causal Transport For features X, legitimate features $L \subset X$, and sensitive feature A, we define the regularizer

$$\mathcal{B}(f) := \mathcal{C}_b^p \left(\mathbb{P}(L, f(X)|A = 0), \mathbb{P}(L, f(X)|A = 1) \right),$$

the bi-causal transport distance between $\mathbb{P}(L, f(X)|A = 0)$ and $\mathbb{P}(L, f(X)|A = 1)$, with $C > \frac{\bar{d}}{\underline{d}}$ as described in proposition 4.2. The bi-causal distance can be viewed as a nested Wasserstein distance (Proposition A.3 in the Appendix), one that takes a different form from the nested Wasserstein distance expressed in Definition 3.1 but which is equivalent under the conditions given in Proposition 4.2. The reformulated nested Wasserstein expression contains a smooth inner cost function; this enables us to estimate the BCD using a nested version of the Sinkhorn divergence (Sinkhorn 1964; Cuturi 2013; Pichler and Weinhardt 2021), which is an entropy-regularized version of the Wasserstein distance. Let us denote this estimate of the BCD by $\hat{\mathcal{B}}(f)$. Our proposed approach, *Fairness through Bi-causal Transport (FairBiT)*, aims to solve the version of Problem (1) with $\mathsf{CDD}(f_\theta)$ set to $\widehat{\mathcal{B}}(f_\theta)$.

Since $\widehat{\mathcal{B}}(f_\theta)$ is differentiable in θ, this problem is amenable to gradient-based solvers. The computational complexity of the *FairBiT* regularizer is $\mathcal{O}(n^2 + |L|^2) = \mathcal{O}(n^2)$, where $|L| \leq n$ is the number of levels of L observed in the sample. See Appendix C for further details as well as the nested Sinkhorn divergence algorithm.

4.2 Enforcing CDD in the ℓ_p sense

The regularization term $\mathsf{CDD}(f_\theta)$ in this case takes the form $\mathsf{CDD}^{\ell_p}(f_\theta) = \|D\|_{\ell^p(\mathcal{L}; \mathbb{Q}(L))}$ as described in Definition 3.2. Our proposed method based on CDD^{ℓ_p} is called *FairLeap: Conditional Fairness in the ℓ_p sense*. The parameters p (the order of the ℓ_p norm) and $\mathbb{Q}(L)$ (the probability measure over $\mathcal{L}$) determine how the level-wise disparities are aggregated. We particularly highlight three aggregation strategies, which result in three variants of *FairLeap*:

1. *Fairleap (uniform)*: A simple average with $\mathbb{Q} = \mathbb{U}(L)$, which we use from here forward to denote the uniform distribution over L: This puts equal emphasis on every observed level of L.
2. *Fairleap ($\mathbb{P}(L)$)*: A weighted average with $\mathbb{Q} = \mathbb{P}(L)$ and $p = 1$: This prioritizes levels of L with more mass.
3. *Fairleap (Ave. $\mathbb{P}(L|A)$)*: A weighted average with $\mathbb{Q} = \frac{\mathbb{P}(L|A=0) + \mathbb{P}(L|A=1)}{2}$ and $p = 1$: This prioritizes levels of L with more mass, within either class of A, and avoids favoring the majority class.

In practice, the unknown distribution $\mathbb{P}$ is replaced with the empirical distribution. A main advantage of these choices of p and $\mathbb{Q}$ is that they yield interpretable regularizers. In terms of the inner distance function d, the definition of CDD^{ℓ_p} is generic and admits any distributional distance. In our implementations, we choose Wasserstein distance due to the known connection between closeness of distribution in Wasserstein sense and parity of performance in downstream tasks (Villani et al. 2009; Santambrogio 2015; Xiong et al. 2023). Similar to *FairBiT*, here we estimate the Wasserstein distance by employing the Sinkhorn divergence, which is differentiable. The computational complexity of the *FairLeap* regularizer is $\mathcal{O}(\frac{n^2}{|L|})$. See Appendix C for further details.

5 Discussion of Existing Methods

In this section, we discuss the methods that we experimentally compare to FairBiT and FairLeap in the next section. We analyze the state of the art method for CDP with rich legitimate features and illustrate why it may not in fact minimize the disparity of interest. In order to widen the scope of our comparisons, we modify an existing approach for DP to apply to CDP, and we consider under what circumstances another existing approach for DP may result in (approximate) CDP without modification. Finally, we discuss an existing method for DP that is computationally closest to ours. Table 2 compares our methods to all these methods.

5.1 State of the art: DCFR

The current state of the art for CDP with rich legitimate features is the approach proposed in Xu et al. (2020a). This method, named Derivable Conditional Fairness Regularizer (DCFR), aims to learn a representation Z of the input features such that $Z \perp\!\!\!\perp A \mid L$, from which it follows that a model $f(Z)$ will satisfy CDP. The method utilizes an adversarial learning approach with a regularizer defined as $R_{\text{DCFR}}(Z, L, A) := \sup_{h \in H_{ZF}} Q(h)$, where H_{ZF} is the set of all bounded square-integrable functions with values in $[0, 1]$, and

$$Q(h) := \mathbb{E}[\mathbf{1}_{\{A=1\}}\, \mathbb{P}(A = 0|L) h(Z, L)] - \mathbb{E}[\mathbf{1}_{\{A=0\}}\, \mathbb{P}(A = 1|L) h(Z, L)]. \quad (2)$$

$Q(h)$ is motivated by a characterization of conditional independence given by Daudin (1980). However, Proposition 5.1 shows that R_{DCFR} aims to minimize a quantity that does not uniformly bound the CDD (in the sense of either Definition 3.1 or 3.2). This is contrast with our proposed methods, which aim to directly minimize the CDD.

Proposition 5.1. *The regularizer proposed by (Xu et al. 2020a) can be equivalently expressed as*

$$R_{DCFR} = \mathbb{E}[(\mathbb{P}(A = 1|Z, L) - \mathbb{P}(A = 1|L))_+]. \quad (3)$$

See Appendix E for the proof of Proposition 5.1. Note that the value of this regularizer depends on which level of the sensitive feature is labeled as 1. Furthermore, while CDP holds if $R_{\text{DCFR}} = 0$, small values of R_{DCFR} do not necessarily imply small values of CDD. Figure 1 illustrates this vis-a-vis FairBiT via a simple synthetic loan example in which we vary both the proportions of males vs. females applying for loans and the respective acceptance rates. (In this simple example, $L = \emptyset$, so $\text{CDD}^{\text{wass}} = \text{CDD}^{\ell_p}$, and the y-axis is therefore simply labeled "CDD.") Each line in Figure 1 is obtained by changing acceptance rates for a fixed proportion of male vs. female applicants; the more unbalanced this proportion, the steeper the CDD-R_{DCFR} curve gets, while *FairBiT* enjoys a consistent relationship with CDD regardless of this proportion. The same is true for *FairLeap*; details and further analysis are given in Appendix D.

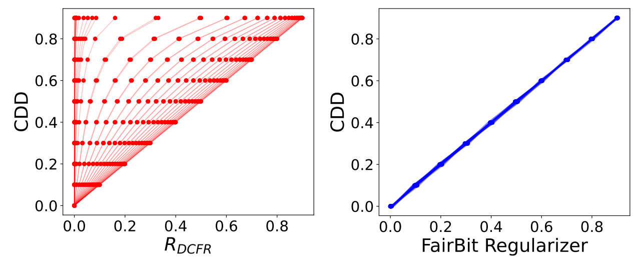

Figure 1: Conditional demographic disparity (CDD) versus R_{DCFR} (left) and the value of FairBiT regularizer (right) in a synthetic loan setting, varying the proportion of males vs. females and the loan acceptance rates. The proportion of males vs. females controls the slope of the CDD-R_{DCFR} curve, with higher ratios yielding steeper curves. A 45 degree line only occurs if the ratio of males to females is 1:1.

5.2 Modifying Adversarial Debiasing for CDP

Adversarial debiasing (Zhang, Lemoine, and Mitchell 2018) employs an adversarial training framework to target DP (as well as equality of odds and equality of opportunity). In the case of DP, the model $f(X)$ is trained to predict the target Y from the input features X, while the adversary is trained to predict the sensitive feature A from the outcome of the model $f(X)$. We modify their method to provide the adversary not only with $f(X)$ but also the legitimate features L. This is motivated by the following remark.

Remark 5.2. *It follows from the weak union property of conditional independence that if $(L, f(X)) \perp\!\!\!\perp A$, then $f(X) \perp\!\!\!\perp A \mid L$, i.e. CDP is satisfied.*

If the adversary is unable to predict A from $(f(X), L)$, then, by Remark 5.2, $f(X)$ will satisfy CDP. If L is correlated with A, then this will not happen in general. When L is correlated with A, the learning procedure may encourage $f(X) \perp\!\!\!\perp A \mid L$ (i.e. CDP) in order to minimize the amount of information the adversary is able to obtain from $(f(X), L)$. Hence, this approach may effectively target CDP. We note however that this method does not include a parameter to control the fairness-performance trade-off, and the choice of the adversary architecture is an additional hyper-parameter that impacts the success of the approach.

5.3 Employing DP methods for CDP without modification

For a more extensive empirical comparison, we consider how two existing methods designed for demographic parity may be used/modified to approximately target conditional demographic parity.

Methods	Perf.-fairness tradeoff param	Targets CDP	No sensitive feat. (training)†	No sensitive feat. (inference)
DCFR	✓	✗*	✗	✗‡
Pre-processing repair	✓	✗	✗	✗
Adversarial debiasing	✗	✓**	✓	✓
Wasserstein regularization	✓	✗	✓	✓
FairBiT & *FairLeap* (ours)	✓	✓	✓	✓

Table 2: Comparison of *FairBiT* and *FairLeap* to the methods described in Section 5. *DCFR is intended to target CDP but targets a proxy quantity that does not necessarily yield small CDD values (Section 5.1). **Adversarial debiasing here refers to our modified version in which we pass $(f(X), L)$ to the adversary instead of just $f(X)$. †For adv. debiasing, only the adversary requires access to the sensitive feature A. For Wasserstein reg. and FairBiT, the gradients for the regularization term may be computed separately on each iteration and passed to the analysts who are training the model $f(X)$. This is important for example in a financial setting where teams with access to sensitive features are distinct from model developers. ‡DCFR can be easily modified not to require the sensitive feature at inference time, though this will in general result in a decrease in model performance.

Pre-processing repair This method proposed by Feldman et al. (Feldman et al. 2015) is a pre-processing method which utilizes a transformation or "repair" $X \mapsto \tilde{X}$ such that $\tilde{X}$ is approximately independent of A. Since $L \subset X$, it follows from Remark 5.2 that $f(\tilde{X})$ is approximately independent of A conditional on $\tilde{L}$. Note that $f(\tilde{X}) \perp\!\!\!\perp A \mid \tilde{L} \not\Rightarrow f(\tilde{X}) \perp\!\!\!\perp A \mid L$, but if the transformation $L \mapsto \tilde{L}$ happens to be minimal, then this method may result in approximate CDP. This method includes a *repair level* parameter in $[0, 1]$, where 0 indicates no transformation, 1 indicates full orthogonalization, and values in between represent degrees of repair.

Wasserstein regularization We also investigate a method which penalizes the loss function with a Wasserstein distance penalty $\mathcal{W}_p^p(\mathbb{P}(f(X) \mid A = 0), \mathbb{P}(f(X) \mid A = 1))$ computed via the Sinkhorn algorithm (Cuturi 2013), which we refer to simply as *Wasserstein regularization*. This method, which targets DP, is the closest approach in the literature when it comes to regularization-based methods using optimal transport to enforce fairness (Rychener, Taskesen, and Kuhn 2022), which is our motivation for including it in our comparisons. However, as illustrated in Table 1 in the Introduction, DP does not imply CDP, and small values of this disparity do not imply small values of CDD.

6 Experiments

In this section, we compare the effectiveness of the methods described in Sections 4 and 5 on four real datasets commonly used in the fairness literature (Fabris et al. 2022).

6.1 Setup

We include two classification tasks on the `Drug` (Fehrman, Egan, and Mirkes 2016) and `Adult` (Becker and Kohavi 1996) datasets; and two regression tasks on the `Law School` (Ramsey, Wightman, and Council 1998) and `Communities and Crime` (Redmond 2009) datasets. In all experiments, we use a multi-layer perceptron (MLP) with two hidden layers containing 50 and 20 nodes and a rectified linear unit (ReLu) activation function. The loss function is set to be cross-entropy for classification (after a softmax activation) and mean squared error (MSE) for regression. We measure the *predictive power (PP)* of each method by the area under the ROC curve (AUC) for the classification tasks and MSE for the regression task. For CDD^{ℓ_p}, we report results for $\mathbb{Q}(L) = \mathbb{U}(L)$, the uniform distribution over L. (We include the results for the variants with $\mathbb{Q}(L) = \mathbb{P}(L)$ and $\mathbb{Q}(L) = \frac{\mathbb{P}(L|A=0)+\mathbb{P}(L|A=0)}{2}$ in Appendix F.3). We report the mean over 10 runs for every method-hyperparameter combination; we omit the hyperparameter labels in the figures for clarity. See Appendix F.2 for more details on hyperparameter values and implementation details.

We compare the following methods, along with two baselines for reference:

- *FairBiT* (Section 4.1),
- *FairLeap (uniform)*, *FairLeap ($\mathbb{P}(L)$)*, and *FairLeap (Ave. $\mathbb{P}(L|A)$)* (Section 4.2),
- *DCFR* (Section 5.1),
- *Adversarial debiasing* (Section 5.2),
- *Pre-processing repair* (Section 5.3),
- *Wasserstein regularization* (Section 5.3),
- *No regularization*, i.e., model training without enforcing any fairness constraint,
- *Legitimate only*, i.e., model training using only the legitimate feature L.

As DCFR and pre-processing repair require access to the sensitive feature, we make the sensitive feature available during training to all methods for the purpose of comparing downstream performance. However, we note that our proposed methods *FairBiT* and *FairLeap*, as well as the adversarial debiasing method, do not necessarily require direct access to the sensitive feature during training (see Table 2). We refer the reader to Appendix F for more comprehensive results, including means, standard deviations, details on hyper-parameter values, and evaluations of demographic disparity.

6.2 Results

Figure 2 reports the fairness-predictive power (referred to hereafter as fairness-PP) trade-offs for the four real datasets.

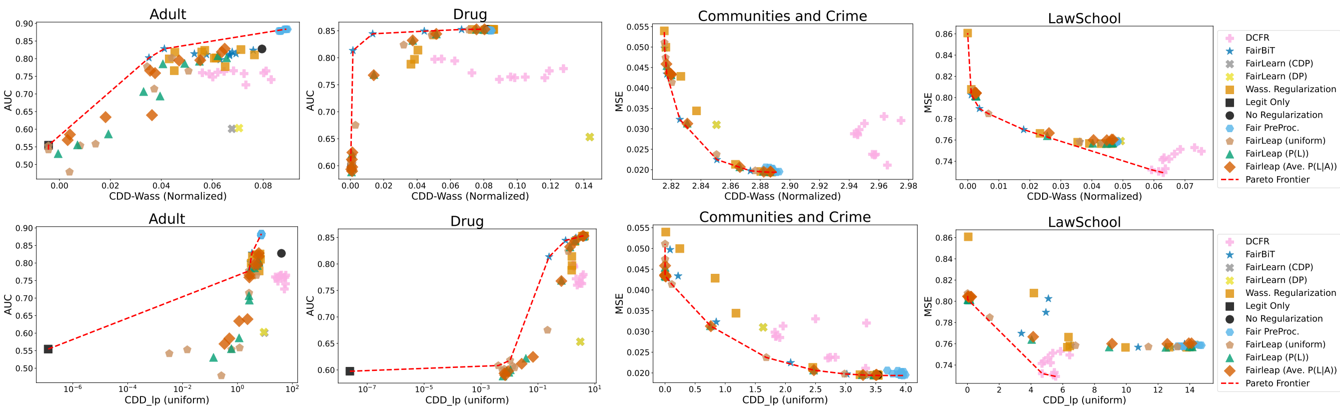

Figure 2: Fairness-predictive power trade-offs and Pareto frontiers for the four real datasets. *Top row:* Trade-offs when measuring fairness using a CDD$^{\text{wass}}$ (we use a 'normalized" version for presentation purposes. See Appendix F for details). *Bottom row:* Trade-offs when measuring fairness using CDD$^{\ell_p}$ (with $\mathbb{Q}(L) = \mathbb{U}(L)$ and $p = 1$). Predictive power (PP) is measured by AUC for classification tasks (first and second columns from left) and MSE for regression tasks (third and fourth columns from left). Results are averaged over 10 runs, with multiple points per model due indicating different hyper-parameter values. Overall, *FairBiT* and the variants of *FairLeap* are consistently part of the Pareto frontier, hence providing better fairness-PP trade-offs than many of the other proposed methods. See text and Appendix F for more details.

We report conditional demographic disparity in the Wasserstein sense with $p = 2$ (Figure 2, top) as well as in the ℓ_p sense with $\mathbb{Q}(L) = \mathbb{U}(L)$ and $p = 1$ (Figure 2, bottom). Each point represents a method-hyperparameter combination; we refer the reader to Appendix F for more details.

Overall, *FairBiT* and the three variants of *FairLeap* are consistently among the highest performing methods, generally providing better fairness-PP trade-offs than the other methods, as indicated by the Pareto frontiers. When evaluated by CDD$^{\text{wass}}$ (Figure 2, top row), *FairBiT* has points on the Pareto frontier for every experiment and generally outperforms all other methods. Similarly, when CDD$^{\ell_p}$ with $\mathbb{Q}(L) = \mathbb{U}(L)$ and $p = 1$ is the measure (Figure 2, bottom row), the method corresponding to this CDD measure, namely *FairLeap (uniform)*, has points on the Pareto frontier for all 4 datasets. Overall, we observe that whether CDD is measured by CDD$^{\text{wass}}$ or CDD$^{\ell_p}$, all our proposed methods are also among the best performing. Moreover, these results show that our regularized methods provide a wide range of points in fairness-PP space, allowing practitioners to choose their desired tradeoff point.

We summarize the performance of the other methods in our experiments as follows.

- Both versions of *Adversarial Debiasing* show poor predictive performance and conditional fairness on the `Adult`, `Drug`, and `LawSchool` datasets. On the `Communities and Crime` dataset, they slightly improve CDD levels but still underperform *FairBiT* and *FairLeap* at similar levels of MSE.
- *Preprocessing repair* achieves the best predictive performance levels (on par with the model with no regularization), but fails to enforce conditional fairness.
- *DCFR* performs well on `LawSchool` when CDD is measured by CDD$^{\ell_p}$, but is not part of the Pareto frontier in the other datasets.
- *Wasserstein regularization* performs well in terms of enforcing conditional fairness in regression tasks, but this comes at the cost of higher MSE than *FairBiT* and *FairLeap* variants for the same CDD values. In classification tasks, it provides reasonable fairness-PP trade-offs, often close to or on the Pareto frontier, albeit providing worse trade-offs than *FairBiT* and variants of *FairLeap*.
- *Legitimate only*, as expected, achieves full CDP but suffers significantly in terms of predictive performance. For presentation purposes, we have omitted *legit-only* from Figure 2, and present it in Figure 3 in Appendix F.

7 Conclusion

In this work, we propose novel measures to quantify the violation of conditional demographic parity (CDP), or conditional demographic disparity (CDD), in the Wasserstein sense and in the ℓ_p sense, based on distributional distances borrowed from the optimal transport literature. We design regularization-based approaches to enforce CDP based on these two measures, *FairBiT* and *FairLeap*. Our methods provide tunable knobs for navigating fairness-performance trade-offs, and they can be applied even when the conditioning variable has many levels. When model outputs are continuous, our methods targets CDP not just in the approximate sense of closeness of the first moments but in terms of full equality of the conditional distributions. We empirically compare our methods to the state of the art for CDP (*DCFR*) as well as methods designed to target (unconditional) demographic parity, including a method that we modify to approximately target CDP. Our evaluations over four real datasets show that our methods generally provide better fairness-performance trade-offs than existing methods.

Potential directions of future work include (i) extending the adoption of bi-causal distance to target other notions of conditional fairness such as equalized odds, (ii) combining methods proposed in this work (e.g. *pre-processing repair* and *FairBiT*) to further improve downstream fairness-performance trade-offs, and (iii) exploring in-training regularization approaches for fair synthetic data generation in healthcare and financial applications (Giuffrè and Shung 2023; Potluru et al. 2023).

Disclaimer

This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JP Morgan") and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

Acciaio, B.; Backhoff-Veraguas, J.; and Carmona, R. 2019. Extended mean field control problems: stochastic maximum principle and transport perspective. *SIAM journal on Control and Optimization*, 57(6): 3666–3693.

Acciaio, B.; Backhoff-Veraguas, J.; and Zalashko, A. 2020. Causal optimal transport and its links to enlargement of filtrations and continuous-time stochastic optimization. *Stochastic Processes and their Applications*, 130(5): 2918–2953.

Agarwal, A.; Dudík, M.; and Wu, Z. S. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, 120–129. PMLR.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, 214–223. PMLR.

Backhoff, J.; Beiglbock, M.; Lin, Y.; and Zalashko, A. 2017. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4): 2528–2562.

Backhoff-Veraguas, J.; Bartl, D.; Beiglböck, M.; and Eder, M. 2020. Adapted Wasserstein distances and stability in mathematical finance. *Finance and Stochastics*, 24(3): 601–632.

Becker, B.; and Kohavi, R. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Black, E.; Yeom, S.; and Fredrikson, M. 2020. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 111–121.

Bothmann, L.; Peters, K.; and Bischl, B. 2024. What Is Fairness? On the Role of Protected Attributes and Fictitious Worlds. arXiv:2205.09622.

Buyl, M.; and De Bie, T. 2022. Optimal Transport of Binary Classifiers to Fairness. *arXiv preprint arXiv:2202.03814*.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building Classifiers with Independency Constraints. In *2009 IEEE International Conference on Data Mining Workshops*, 13–18. Miami, FL, USA: IEEE. ISBN 978-1-4244-5384-9.

Carlier, G.; Duval, V.; Peyré, G.; and Schmitzer, B. 2017. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2): 1385–1418.

Chiappa, S.; and Pacchiano, A. 2021. Fairness with Continuous Optimal Transport. *arXiv preprint arXiv:2101.02084*.

Chzhen, E.; Denis, C.; Hebiri, M.; Oneto, L.; and Pontil, M. 2020. Fair regression with wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33: 7321–7331.

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

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Coston, A.; Rambachan, A.; and Chouldechova, A. 2021. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*, 2144–2155. PMLR.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Daudin, J. 1980. Partial association measures and an application to qualitative regression. *Biometrika*, 67(3): 581–590.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*, 214–226. ACM Press. ISBN 978-1-4503-1115-1.

Fabris, A.; Messina, S.; Silvello, G.; and Susto, G. A. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6): 2074–2152.

Fehrman, E.; Egan, V.; and Mirkes, E. 2016. Drug Consumption (Quantified). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5TC7S.

Fehrman, E.; Muhammad, A. K.; Mirkes, E. M.; Egan, V.; and Gorban, A. N. 2017. The five factor model of personality and evaluation of drug consumption risk. In *Data science*, 231–242. Springer.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 259–268. New York, NY, USA: Association for Computing Machinery. ISBN 9781450336642.

Flamary, R.; Courty, N.; Gramfort, A.; Alaya, M. Z.; Boisbunon, A.; Chambon, S.; Chapel, L.; Corenflos, A.; Fatras, K.; Fournier, N.; Gautheron, L.; Gayraud, N. T.; Janati, H.; Rakotomamonjy, A.; Redko, I.; Rolet, A.; Schutz, A.; Seguy, V.; Sutherland, D. J.; Tavenard, R.; Tong, A.; and Vayer, T. 2021. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78): 1–8.

Franklin, J. S.; Powers, H.; Erickson, J. S.; McCusker, J.; McGuinness, D. L.; and Bennett, K. P. 2023. An Ontology for Reasoning About Fairness in Regression and Machine Learning. In *Iberoamerican Knowledge Graphs and Semantic Web Conference*, 243–261. Springer.

Friedler, S. A.; Scheidegger, C.; and Venkatasubramanian, S. 2021. The (Im)Possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making. *Communications of the ACM*, 64(4): 136–143.

Fukuchi, K.; and Sakuma, J. 2024. Demographic parity constrained minimax optimal regression under linear model. *Advances in Neural Information Processing Systems*, 36.

Ghosal, P.; and Nutz, M. 2022. On the Convergence Rate of Sinkhorn's Algorithm. *arXiv preprint arXiv:2212.06000*.

Giuffrè, M.; and Shung, D. L. 2023. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ digital medicine*, 6(1): 186.

Gordaliza, P.; Del Barrio, E.; Fabrice, G.; and Loubes, J.-M. 2019. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, 2357–2365. PMLR.

Hardt, M.; Price, E.; and Srebro, Nathan. 2016. Equality of Opportunity in Supervised Learning. In *30th Conference on Neural Information Processing Systems*, 9. Barcelona, Spain.

Heitsch, H.; and Römisch, W. 2009. Scenario tree modeling for multistage stochastic programs. *Mathematical Programming*, 118(2): 371–406.

Hort, M.; Chen, Z.; Zhang, J. M.; Harman, M.; and Sarro, F. 2024. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 1(2): 1–52.

Jean, J. 1980. Weak and strong solutions of stochastic differential equations. *Stochastics*, 3(1-4): 171–191.

Jiang, R.; Pacchiano, A.; Stepleton, T.; Jiang, H.; and Chiappa, S. 2020. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, 862–872. PMLR.

Jin, Y.; and Lai, L. 2023. Fairness-aware regression robust to adversarial attacks. *IEEE Transactions on Signal Processing*.

Johndrow, J. E.; and Lum, K. 2019. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1): 189–220.

Jourdan, F.; Kaninku, T. T.; Asher, N.; Loubes, J.-M.; and Risser, L. 2023. How optimal transport can tackle gender biases in multi-class neural network classifiers for job recommendations. *Algorithms*, 16(3): 174.

Kamiran, F.; and Calders, T. 2010. Classification with No Discrimination by Preferential Sampling. In *Proceedings of the 19th Machine Learning Conference*.

Kamiran, F.; Karim, A.; and Zhang, X. 2012. Decision Theory for Discrimination-Aware Classification. In *2012 IEEE 12th International Conference on Data Mining*, 924–929. Brussels, Belgium: IEEE. ISBN 978-1-4673-4649-8 978-0-7695-4905-7.

Kamiran, F.; Žliobaitė, I.; and Calders, T. 2013. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35(3): 613–644.

Kantorovitch, L. 1958. On the translocation of masses. *Management science*, 5(1): 1–4.

Kim, J. S.; Chen, J.; and Talwalkar, A. 2020. FACT: A diagnostic for group fairness trade-offs. In *International Conference on Machine Learning*, 5264–5274. PMLR.

Kurtz, T. 2014. Weak and strong solutions of general stochastic models. *Electronic Communications in Probability*, 19: 1–16.

Laclau, C.; Redko, I.; Choudhary, M.; and Largeron, C. 2021. All of the fairness for edge prediction with optimal transport. In *International Conference on Artificial Intelligence and Statistics*, 1774–1782. PMLR.

Lakshmanan, R.; and Pichler, A. 2023. Fast Approximation of Unbalanced Optimal Transport and Maximum Mean Discrepancies. *arXiv preprint arXiv:2306.13618*.

Langbridge, A.; Quinn, A.; and Shorten, R. 2024. Optimal Transport for Fairness: Archival Data Repair using Small Research Data Sets. In *2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW)*, 237–245. IEEE.

Lassalle, R. 2013. Causal transference plans and their Monge-Kantorovich problems. *arXiv preprint arXiv:1303.6925*.

Liu, S.; and Vicente, L. N. 2022. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Computational Management Science*, 19(3): 513–537.

Luise, G.; Rudi, A.; Pontil, M.; and Ciliberto, C. 2018. Differential properties of sinkhorn approximation for learning with wasserstein distance. *Advances in Neural Information Processing Systems*, 31.

Miroshnikov, A.; Kotsiopoulos, K.; Franks, R.; and Ravi Kannan, A. 2022. Wasserstein-based fairness interpretability framework for machine learning models. *Machine Learning*, 111(9): 3307–3357.

Mishler, A.; and Kennedy, E. 2021. FADE: FAir Double Ensemble Learning for Observable and Counterfactual Outcomes. *ArXiv preprint arXiv:2109.00173*.

Monge, G. 1781. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, 666–704.

Oneto, L.; Donini, M.; Luise, G.; Ciliberto, C.; Maurer, A.; and Pontil, M. 2020. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. *Advances in Neural Information Processing Systems*, 33: 15360–15370.

Pessach, D.; and Shmueli, E. 2022. A Review on Fairness in Machine Learning. *ACM Computing Surveys*, 55(3): 1–44.

Pflug, G. C. 2010. Version-independence and nested distributions in multistage stochastic optimization. *SIAM Journal on Optimization*, 20(3): 1406–1420.

Pflug, G. C.; and Pichler, A. 2012. A distance for multistage stochastic optimization models. *SIAM Journal on Optimization*, 22(1): 1–23.

Pflug, G. C.; and Pichler, A. 2014. *Multistage stochastic optimization*. Springer.

Pflug, G. C.; and Pichler, A. 2015. Dynamic generation of scenario trees. *Computational Optimization and Applications*, 62(3): 641–668.

Pflug, G. C.; and Pichler, A. 2016. From empirical observations to tree models for stochastic optimization: convergence properties. *SIAM Journal on Optimization*, 26(3): 1715–1740.

Pichler, A.; and Weinhardt, M. 2021. Nested Sinkhorn Divergence To Compute The Nested Distance. *arXiv preprint arXiv:2102.05413*.

Potluru, V. K.; Borrajo, D.; Coletta, A.; Dalmasso, N.; El-Laham, Y.; Fons, E.; Ghassemi, M.; Gopalakrishnan, S.; Gosai, V.; Kreačić, E.; et al. 2023. Synthetic data applications in finance. *arXiv preprint arXiv:2401.00081*.

Ramsey, H.; Wightman, L.; and Council, L. S. A. 1998. *LSAC National Longitudinal Bar Passage Study*. LSAC research report series. Law School Admission Council.

Redmond, M. 2009. Communities and Crime. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C53W3X.

Ritov, Y.; Sun, Y.; and Zhao, R. 2017. On conditional parity as a notion of non-discrimination in machine learning. *arXiv preprint arXiv:1706.08519*.

Romano, Y.; Bates, S.; and Candes, E. 2020. Achieving Equalized Odds by Resampling Sensitive Attributes. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 361–371. Curran Associates, Inc.

Rychener, Y.; Taskesen, B.; and Kuhn, D. 2022. Metrizing Fairness. *arXiv preprint arXiv:2205.15049*.

Salimans, T.; Zhang, H.; Radford, A.; and Metaxas, D. 2018. Improving GANs using optimal transport. *arXiv preprint arXiv:1803.05573*.

Santambrogio, F. 2015. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63): 94.

Si, N.; Murthy, K.; Blanchet, J.; and Nguyen, V. A. 2021. Testing group fairness via optimal transport projections. In *International Conference on Machine Learning*, 9649–9659. PMLR.

Silvia, C.; Ray, J.; Tom, S.; Aldo, P.; Heinrich, J.; and John, A. 2020. A general approach to fairness with optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3633–3640.

Sinkhorn, R. 1964. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2): 876–879.

Villani, C.; et al. 2009. *Optimal transport: Old and new*, volume 338. Springer.

Wang, Y.; Nguyen, V. A.; and Hanasusanto, G. A. 2024. Wasserstein robust classification with fairness constraints. *Manufacturing & Service Operations Management*.

Xian, R.; Li, Q.; Kamath, G.; and Zhao, H. 2024. Differentially Private Post-Processing for Fair Regression. *arXiv preprint arXiv:2405.04034*.

Xiong, Z.; Dalmasso, N.; Potluru, V. K.; Balch, T.; and Veloso, M. 2023. Fair Wasserstein Coresets. *arXiv preprint arXiv:2311.05436*.

Xu, R.; Cui, P.; Kuang, K.; Li, B.; Zhou, L.; Shen, Z.; and Cui, W. 2020a. Algorithmic decision making with conditional fairness. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2125–2135.

Xu, T.; Wenliang, L. K.; Munn, M.; and Acciaio, B. 2020b. Cot-gan: Generating sequential data via causal optimal transport. *Advances in neural information processing systems*, 33: 8798–8809.

Yamada, T.; and Watanabe, S. 1971. On the uniqueness of solutions of stochastic differential equations. *Journal of Mathematics of Kyoto University*, 11(1): 155–167.

Yang, J.; Zhang, L.; Chen, N.; Gao, R.; and Hu, M. 2022a. Decision-making with Side Information: A Causal Transport Robust Approach. *Available at Optimization Online*.

Yang, M.; Sheng, J.; Liu, W.; Jin, B.; Wang, X.; and Wang, X. 2022b. Obtaining Dyadic Fairness by Optimal Transport. In *2022 IEEE International Conference on Big Data (Big Data)*, 4726–4732. IEEE.

Zehlike, M.; Hacker, P.; and Wiedemann, E. 2020. Matching code and law: achieving algorithmic fairness with optimal transport. *Data Mining and Knowledge Discovery*, 34(1): 163–200.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhao, H.; and Gordon, G. 2019. Inherent Tradeoffs in Learning Fair Representations. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zliobaite, I.; Kamiran, F.; and Calders, T. 2011. Handling Conditional Discrimination. In *2011 IEEE 11th International Conference on Data Mining*, 992–1001. Vancouver, BC, Canada: IEEE. ISBN 978-1-4577-2075-8 978-0-7695-4408-3.

Appendix

A Details on bi-causal transport distance

The bi-causal transport distance, aka the nested transport distance, is defined through a constrained optimal transport problem. The optimal transport problem was originally formulated as the problem of transporting one distribution to another while incurring minimal cost (Monge 1781; Kantorovitch 1958). In our setup, for probability measures $\tilde{\mathbb{P}}, \mathbb{P}$ on measure space $\mathcal{U} \times \mathcal{V}$, the set $\Gamma(\tilde{\mathbb{P}}, \mathbb{P})$ of *transport plans* denotes the collection of all probability measures on the space $(\mathcal{U} \times \mathcal{V}) \times (\mathcal{U} \times \mathcal{V})$ with marginals $\tilde{\mathbb{P}}$ and $\mathbb{P}$. Then for a cost function $\mathsf{d} : ((\mathcal{U} \times \mathcal{V}) \times (\mathcal{U} \times \mathcal{V})) \to [0, \infty]$, the optimal transport problem is that of finding the transport plan γ^* that attains the infimum

$$\inf_{\gamma \in \Gamma(\tilde{\mathbb{P}}, \mathbb{P})} \int_{(\mathcal{U} \times \mathcal{V}) \times (\mathcal{U} \times \mathcal{V})} \mathsf{d}((\tilde{u}, \tilde{v}), (u, v)) d\gamma((\tilde{u}, \tilde{v}), (u, v)).$$

A common distance defined through optimal transport is the *Wasserstein distance*. For $\tilde{\mathbb{P}}, \mathbb{P}$ with finite p-th moment, the Wasserstein distance of order p with metric d is defined by $\mathcal{W}_p^p(\tilde{\mathbb{P}}, \mathbb{P}; \mathsf{d}) := \inf_{\gamma \in \Gamma(\tilde{\mathbb{P}}, \mathbb{P})} \mathbb{E}_{(\tilde{U}, \tilde{V}), (U, V) \sim \gamma} \left[\mathsf{d}((\tilde{u}, \tilde{v}), (u, v))^p\right]$. (Note that $\mathcal{W}_p$ refers to the Wasserstein distance of order p, while $\mathcal{W}_p^p$ refers to the Wasserstein distance taken to the power p, for simplicity of expression.)

With the common choice of the ℓ_p norm for d, $\mathcal{W}_p^p$ becomes

$$\inf_{\gamma \in \Gamma(\tilde{\mathbb{P}}, \mathbb{P})} \mathbb{E}_{(\tilde{U}, \tilde{V}), (U, V) \sim \gamma} \left[\|\tilde{U} - U\|_p^p + \|\tilde{V} - V\|_p^p\right].$$

The bi-causal transport distance can be defined in the framework above with a constrained set of transport plans Γ_{bc} (Backhoff et al. 2017). The bi-causal transport distance is defined in Definition 4.1 in the main body of the paper; for clarity of exposition, we reiterate the definition here with some additional detail.

Definition A.1 (Causal and bi-causal transport plans)**.** A joint distribution $\gamma \in \Gamma(\tilde{\mathbb{P}}, \mathbb{P})$ is called a *causal transport plan* if for $((\tilde{U}, \tilde{V}), (U, V)) \sim \gamma$, U and $\tilde{V}$ are conditionally independent given $\tilde{U}$:

$$U \perp\!\!\!\perp \tilde{V} \mid \tilde{U}.$$

Analogously, bi-causal transport plans are transport plans that are "causal in both directions." The set of all such plans is given by

$$\Gamma_{bc}(\tilde{\mathbb{P}}, \mathbb{P}) = \{\gamma \in \Gamma(\tilde{\mathbb{P}}, \mathbb{P}) \text{ s.t. for } ((\tilde{U}, \tilde{V}), (U, V)) \sim \gamma, \\ U \perp\!\!\!\perp \tilde{V} \mid \tilde{U} \text{ and } \tilde{U} \perp\!\!\!\perp V \mid U\}.$$

The origin of causal transport can be traced back to the Yamada-Watanabe criterion for stochastic differential equations (Yamada and Watanabe 1971; Jean 1980; Kurtz 2014). In the theory of optimal transport, Lassalle (2013) studied the transport problem in continuous time under the so-called causality constraint, and Backhoff et al. (2017) considers a discrete-time analogue. Causal transport has been applied to stochastic optimization (Acciaio, Backhoff-Veraguas, and Zalashko 2020), as well as other areas such as stochastic control (Acciaio, Backhoff-Veraguas, and Carmona 2019) and machine learning (Xu et al. 2020b). The bicausal transport distance is a symmetrized analog of the casual transport distance, which has been exploited to study the stability and sensitivity of multistage stochastic programming (Pflug 2010; Pflug and Pichler 2012, 2014, 2015, 2016). Most previous works consider a multi-period definition of bi-causal transport distance, whose main motivation is to investigate optimal transportation problems with filtrations and their applications to stochastic calculus. In our problem, our definition can be viewed as a bi-causal transport plan (Backhoff et al. 2017) with two periods, which is also considered in (Yang et al. 2022a), to capture the conditional independence relations that we care about in the context of conditional demographic parity.

Definition A.2 (Bi-causal transport distance (BCD))**.** The *bi-causal transport distance* (BCD, referred to hereafter simply as the *bi-causal distance*) between $\tilde{\mathbb{P}}$ and $\mathbb{P}$, denoted by $\mathcal{C}_b^p(\tilde{\mathbb{P}}, \mathbb{P})$, is defined as

$$\inf_{\gamma \in \Gamma_{bc}(\tilde{\mathbb{P}}, \mathbb{P})} \mathbb{E}_{((\tilde{U}, \tilde{V}), (U, V)) \sim \gamma} \left[C\|\tilde{U} - U\|_p^p + \|\tilde{V} - V\|_p^p\right].$$

The following proposition shows that the bi-causal distance can be viewed as a nested Wasserstein distance (Backhoff-Veraguas et al. 2020).

Proposition A.3. *The bi-causal distance can equivalently be written as* $\mathcal{C}_b^p(\tilde{\mathbb{P}}, \mathbb{P}) = \mathcal{W}_p^p(\tilde{\mathbb{P}}_{\tilde{U}}, \mathbb{P}_U; D)$ *where*

$$D^p(\tilde{U}, U) = \mathcal{W}_p^p(\tilde{\mathbb{P}}_{\tilde{V}|\tilde{U}}, \mathbb{P}_{V|U}) + C\|\tilde{U} - U\|_p^p.$$

As mentioned in Section 4.1, U in our setting corresponds to the legitimate feature L, while V corresponds to the model output $f(X)$. The presence or absence of the tilde corresponds to the two levels of the sensitive feature A, so that $\mathcal{W}_p^p(\tilde{\mathbb{P}}_{\tilde{V}|\tilde{U}}, \mathbb{P}_{V|U})$ represents the Wasserstein distance between $\mathbb{P}(f(X) \mid L = l, A = 0)$ and $\mathbb{P}(f(X) \mid L = l, A = 1)$.

B Additional discussion on disparity definitions

B.1 Demographic disparity

In Appendix F.3 below, we provide additional results from our experiments illustrating that methods which target demographic parity are not necessarily successful at targeting conditional demographic parity. In this section, therefore, we define the *demographic disparity*, the violation of demographic parity.

Most previous work defines the demographic disparity (the violation of demographic parity) by $|\mathbb{E}[f(X) \mid A = 1] - \mathbb{E}[f(X) \mid A = 0]|$ or by $\mathbb{E}[f(X) \mid A = 1]/\mathbb{E}[f(X) \mid A = 0]$ and aims to make this quantity as close to 0 (for the difference) or 1 (for the ratio) as possible (Pessach and Shmueli 2022). When $f(X)$ is binary, equalizing the first moments in this way is equivalent to enforcing that $f(X) \perp\!\!\!\perp A$, but if $f(X)$ is continuous, then $f(X)$ may behave very differently for the two groups even if the means of the two conditional distributions are equal. As with the conditional demographic disparity (Definitions 3.1 and 3.2, we define the demographic disparity via a distributional distances, as we are interested in minimizing the discrepancy between the entire distributions, not just the first moments.

Definition B.1 (Demographic disparity (DD))**.** Let $d(\cdot,\cdot)$ denote a distance between distributions. The demographic disparity for model $f(X)$ with respect to $\mathsf{d}(\cdot,\cdot)$ is defined by

$$\mathsf{d}(\mathbb{P}_{f(X)|A=0}, \mathbb{P}_{f(X)|A=1}). \tag{4}$$

In other words, we define the demographic disparity as the distance between the conditional distributions of $f(X)$ for the two groups represented by the sensitive feature.

The choice of distance function d may depend on the problem setting. For example, minimizing the Kolmogorov distance in some sense requires fairness across the entire range of values of $f(X)$. In some settings, this may be ethically appropriate, while in others it may be unnecessary and might incur an unacceptable trade-off in performance. The Wasserstein distance has previously been used to target demographic parity, in part because the closeness of two distributions in the Wasserstein sense is related to closeness in the downstream performance of models trained on those distributions (Villani et al. 2009; Santambrogio 2015; Xiong et al. 2023). We utilize the Wasserstein distance to measures DD in our additional results below.

B.2 CDP with continuous legitimate features

The conditional demographic disparity (CDP) is defined over the set $\mathcal{S} = \operatorname{supp}\mathbb{P}(L|A=0) \cap \operatorname{supp}\mathbb{P}(L|A=1)$, i.e. on the common support of the legitimate feature across the two groups defined by the sensitive feature. In practice, if the legitimate feature is continuous, then the empirical common support in any finite sample will be empty, which means the CDD cannot be estimated. Moreover, any method targeting CDP will require multiple samples at each level l of L. This is why we need to assume that L is either naturally discrete or appropriately discretized. This is not overly limiting, since continuous features may be discretized using appropriate binning strategies informed by the fairness requirements and the problem setting.

As an alternative, we envision a "smooth CDP" definition, where parity is required not only at each level but to some degree across different levels. The extent to which model output distributions are allowed to differ across levels will be reversely related to the distance between levels. Such a definition would be applicable to both discrete and continuous legitimate features, without any binning required. The implications and appropriateness of such a definition need careful attention and are beyond the scope of this work.

B.3 Aggregation strategies for CDD in the ℓ_p sense

The conditional demographic disparity (CDD) in the ℓ_p sense (Definition 3.2) aggregates the level-wise disparities at each value of the legitimate features according to a measure $\mathbb{Q}$. An obvious candidate for $\mathbb{Q}$ is $\mathbb{P}(L)$, where the weights are proportional to the amount of data at each level. However, this measure may overly "favor" the majority class in cases where the data is unbalanced with respect to the sensitive feature. For example, using the same scenario as in Table 1 in the Introduction, suppose there were 10 times as many female as male loan applicants, and suppose that most females had high incomes and most males had low incomes. Suppose that the approval probability was the same for high income female vs. male applicants, but that low income females had a higher approval probability than low income males. Suppose that p were set to 1, so that the disparity were simply $\mathbb{E}[D(L)]$, the average distributional distance across levels of L. In this example, $\mathbb{E}[D(L)]$ would be relatively small, since much of the mass of L would be concentrated in the high income level, but the majority of males would be subject to the disparate outcomes represented by the low income tier.

Alternatively, one could use $\mathbb{Q} = \frac{\mathbb{P}(L|A=0)+\mathbb{P}(L|A=1)}{2}$ to avoid favoring the majority class. This probability measure ensures that disparities which primarily affect only one level of the sensitive feature (e.g. males in the example in the preceeding paragraph) are more influential in the overall disparity measure.

A third possible choice is $\mathbb{Q} = \mathbb{U}(L)$, where $\mathbb{U}(L)$ is the uniform distribution over values of L. This puts equal emphasis on every level of L, which may or may not be desirable. Any distribution $\mathbb{Q}$ results in a valid CDD measure; the choice depends on the user's own values and problem-specific characteristics like the distributions of A and L.

In all the disparity definitions given above, different distance functions d and different values of p induce different notions of disparity. For example, in the CDD in the ℓ_p sense (Definition 3.2), if $\mathsf{d}(\cdot,\cdot)$ is the Kolmogorov distance and $p=\infty$, then the conditional demographic disparity of $f(X)$ is ε if the Kolmogorov distance between $\mathbb{P}(f(X) \mid L=l, A=0)$ and $\mathbb{P}(f(X) \mid L=l, A=1)$ is no more than ε for every possible l. Setting $p=\infty$ and enforcing a small corresponding disparity requires the model $f(X)$ to behave similarly for the two groups within every level of the legitimate feature $L=l$. On the other hand, setting p to, say, 2, allows $f(X)$ to potentially behave quite differently for the two groups over areas of $\mathcal{L}$ with small measure.

C Computing the bi-causal distance

As shown in Proposition A.3, the bi-causal distance can be viewed as a nested Wasserstein distance with particular inner and outer distance functions. Computing the Wasserstein distance exactly is known to be computationally intractable (Arjovsky, Chintala, and Bottou 2017; Salimans et al. 2018). We approximate the Wasserstein distance by the Sinkhorn divergence (Sinkhorn 1964; Cuturi 2013), which is computationally efficient and corresponds to regularizing the Wasserstein distance with an entropy term. It also improves the scalability of traditional methods. The convergence of this entropic regularized optimal transport problem is theoretically studied in Carlier et al. (2017), while Ghosal and Nutz (2022) obtain the rate $\mathcal{O}(t^{-1})$ for a large class of cost functions and marginals. The consistency and differential properties are studied in Luise et al. (2018).

To summarize the idea behind computing the bi-causal transport distance, the nested Sinkhorn divergence is a multistage stochastic program. Assuming the observed samples (X_i, Y_i, A_i) are i.i.d. samples from a stochastic process, the algorithm proposed by Pichler and Weinhardt (2021) discretizes the whole space and filtration of the stochastic process (i.e., the whole feature space) using scenario trees and uses those compute the optimal transport plan (Heitsch and Römisch 2009). We compute the gradient through the Sinkhorn divergence algorithm and propagate them via stochastic gradient descent to minimize the empirical risk as in Oneto et al. (2020).

Algorithm 1 presents the details of the nested algorithm used to calculate the bi-causal distance; it can be viewed as a special case of Algorithm 1 in (Pichler and Weinhardt 2021) with two time periods. For each period, we use the Sinkhorn algorithm ((Cuturi 2013; Flamary et al. 2021)). To simplify the notation, denote $\mathbb{P}_0 := (f \otimes \mathrm{Id})_{\#}\mathbb{P}_{X|A=0}$ and $\mathbb{P}_1 := (f \otimes \mathrm{Id})_{\#}\mathbb{P}_{X|A=1}$. For the discrete nested distance, we use scenario trees to model the whole space and filtration (i.e., the whole feature space). Denote $\mathcal{N}_0^X, \mathcal{N}_0^Y$ ($\mathcal{N}_1^X, \mathcal{N}_1^Y$, resp.) the set of all nodes of $\mathbb{P}_0$ ($\mathbb{P}_1$, resp.). We use $m \prec i$ to denote the predecessor m of the node i, and similarly $i \succ m$ to denote the successors i of the node m. In this notation, we can define the two probability distributions as:

$$\mathbb{P}_0 = \sum_{m \in \mathcal{N}_0^X} p_0^m \sum_{i \succ m} q_0^i \delta_{(x_0^m, y_0^i)},$$
$$\mathbb{P}_1 = \sum_{n \in \mathcal{N}_1^X} p_1^n \sum_{j \succ n} q_1^j, \delta_{(x_1^n, y_1^j)},$$

where $\delta_{a,b}$ is the Kronecker delta function, i.e. $\delta_{a,b}(x, y) = 1$ only for $x = a$ and $y = b$, and otherwise 0. Moreover, p_k^m denotes the marginal distribution of the predecessor m, where q_k^i is the conditional distribution of the node i for $k = 0, 1$. The Sinkhorn algorithm (Cuturi 2013) cam be viewed as an entropic regularized optimal transport problem, by adding a logarithmic penalty term $H(\gamma) = \sum_{i,j} \gamma_{i,j} \log \gamma_{i,j}$ to the objective.

The inputs to Algorithm 1 are the distributions $\mathbb{P}_0$ and $\mathbb{P}_1$, which in practice correspond to the features $(x_0, y_0) \sim \mathbb{P}_0$ and the marginal probabilities p_0 — equivalently, (x_1, y_1) and p_1 for $\mathbb{P}_1$. As we discretize the feature space using scenario trees, we denote x_0^m the features belonging to the node $m \in \mathcal{N}_0^X$, and y_0^i the labels belonging to the node $m \in \mathcal{N}_0^Y$ (and equivalently for $\mathbb{P}_1$).

Remark C.1. *In our algorithms, we use the Sinkhorn solver (Flamary et al. 2021) whose time complexity has proven to be nearly $\mathcal{O}(n^2)$. This means that the time complexity of computing the regularizer in FairLeap (Section 4.2) is $\mathcal{O}(|L|) + \sum_{j=1}^{|L|} \mathcal{O}(n_j^2) = \mathcal{O}(\frac{n^2}{|L|})$ where n_j is the number of samples with legitimate feature value l_j and all n_j's are comparable. Moreover, it follows that our nested Sinkhorn algorithm in FairBiT has time complexity of $|L|^2 + \sum_j \sum_k n_j n_k = |L|^2 + n^2 = \mathcal{O}(n^2)$ where n_j is the number of samples with legitimate feature value l_j and all n_j's are comparable, $n = \sum_{k=1}^K n_k$ is the sample size, and $|L|$ is the number of the levels the legitimate features take in the training sample (so that necessarily $|L| \leq n$).*

Algorithm 1: Nested Sinkhorn iteration of the bi-causal distance $\mathcal{C}_b^p(\mathbb{P}_0, \mathbb{P}_1)$

Input : The distribution $\mathbb{P}_0, \mathbb{P}_1$ by specifying x_0^m, y_0^i, p_0^i, x_1^n, y_1^j, p_1^j for $m \in \mathcal{N}_0^X$, $i \in \mathcal{N}_0^Y$, $n \in \mathcal{N}_1^X$, $j \in \mathcal{N}_1^Y$, and regularization parameter $\lambda > 0$.

Output : Bi-causal distance between the two distributions $\mathbb{P}_0$ and $\mathbb{P}_1$.

for *every combination of nodes* $m \in \mathcal{N}_0^X$,$n \in \mathcal{N}_1^X$ **do**

For all combination of leaf nodes $i \in \mathcal{N}_0^Y$ and $j \in \mathcal{N}_1^Y$ with predecessor $m \prec i$ and $n \prec j$:

$$d_Y(i,j)^p = |x_0^m - x_1^n|^p + |y_0^i - y_1^j|^p$$

solve the linear program

$$\gamma_Y^\star(m,n)^p = \min_{\gamma_Y} \sum_{\substack{i \succ m \\ j \succ n}} \gamma_Y(i,j|m,n) d_Y(i,j)^p - \frac{1}{\lambda} H(\gamma_Y)$$
$$\text{subject to} \sum_{j \succ n} \gamma_Y(i,j|m,n) = q_0^i, \qquad i \succ m,$$
$$\sum_{i \succ m} \gamma_Y(i,j|m,n) = q_1^j, \qquad j \succ n,$$
$$\gamma_Y(i,j|m,n) \geq 0.$$

end

solve the linear program

$$\min_{\gamma_X} \sum_{\substack{i \succ m \\ j \succ n}} \gamma_X(m,n) \gamma_Y^\star(m,n)^p - \frac{1}{\lambda} H(\gamma_X)$$
$$\text{subject to} \sum_{j \succ n} \gamma_X(m,n) = p_0^m, \qquad i \succ m,$$
$$\sum_{i \succ m} \gamma_X(m,n) = p_1^n, \qquad j \succ n,$$
$$\gamma_X(m,n) \geq 0. \tag{5}$$

return *The bi-causal distance between* $\mathbb{P}_0$ *and* $\mathbb{P}_1$ *is the* p*-th root of the optimal value of* (5).

Remark C.2. *In the above algorithm, the complexity mostly depends on the complexity of the Sinkhorn solver we used. A recent work by Lakshmanan and Pichler (2023) has significantly reduce the arithmetic operations to compute the distances from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log n)$, which enables access to large and high-dimensional data sets. By employing nonequispaced fast Fourier transform as what they do, we may be able to reduce our time complexity of $K^2 + \mathcal{O}(n \log n)$. We leave this to future work.*

D Relationship between DCFR and CDD

Here we consider in further detail the relationship between conditional demographic disparity (CDD) and the Derivable Conditional Fairness Regularizer (DCFR) method proposed by Xu et al. (2020a).

The CDD considers the distances between the conditional distributions of the model $f(X)$ given the sensitive feature A and the legitimate feature L, i.e. between $\mathbb{P}(f(X) \mid L, A = 1)$ and $\mathbb{P}(f(X) \mid L, A = 0)$. The DCFR regularizer R_{DCFR}, by contrast, aims to minimize $(\mathbb{P}(A = 1 \mid Z, L) - \mathbb{P}(A = 1 \mid L))_+$, the (positive part function of the) differences between the conditional distributions of the sensitive feature given the legitimate feature and a variable $Z = g(X, A)$ that is a transformation of the input features. DCFR promotes the learning of the transformation $g(X, A)$ simultaneously with the training of a model $f(Z)$. (See Figure 2 in Xu et al. 2020a.)

In the special case that $R_{\text{DCFR}} = 0$, the CDD is also 0, meaning that Conditional Demographic Parity holds. This is illustrated via the following chain of reasoning:

$$
\begin{aligned}
&R_{\text{DCFR}} := \mathbb{E}\left[(\mathbb{P}(A = 1 \mid Z, L) - \mathbb{P}(A = 1 \mid L))_+\right] = 0\\
&\implies \mathbb{P}(A = 1 \mid Z = z, L = \ell) = \mathbb{P}(A = 1 \mid L = \ell)\ \forall \ell, z\\
&\implies Z \perp\!\!\!\perp A \mid L\\
&\implies f(Z) \perp\!\!\!\perp A \mid L.
\end{aligned}
$$

However, it does not follow that if R_{DCFR} is small, then $f(Z)$ is approximately independent of A given L. More precisely, the magnitude of R_{DCFR} does not necessarily guarantee anything about the magnitude of the CDD. One reason for this is that without further constraints, closeness in the conditional distributions $\mathbb{P}(A \mid Z, L)$ does not imply closeness in the distributions $\mathbb{P}(Z \mid A, L)$. The following proposition illustrates this point. The proof is given below in Section D.2.

Proposition D.1. *Assuming each of the conditional probabilities and densities are well-defined and the denominators are non-zero in the following expression, we have*

$$
\frac{|\mathbb{P}(Z = z \mid A = 1, L = \ell) - \mathbb{P}(Z = z \mid A = 0, L = \ell)|}{|\mathbb{P}(A = 1 \mid Z = z, L = \ell) - \mathbb{P}(A = 1 \mid L = \ell)|} = \frac{\mathbb{P}(Z = z \mid L = \ell)}{\mathbb{P}(A = 0 \mid L = \ell)\mathbb{P}(A = 1|L = \ell)}.
$$

This follows by a simple application of Bayes' rule. The numerator on the left represents the difference in the densities of Z at z conditional on $L = \ell$ for the two groups, while the denominator represents the differences at $Z = z$ in the probability mass functions that appear in R_{DCFR}. The expression on the right shows that their ratio can be arbitrarily large, since $\mathbb{P}(A = 0 \mid L = \ell)\mathbb{P}(A = 1 \mid L = \ell)$ can be arbitrarily small and the density $\mathbb{P}(Z = z \mid L = \ell)$ could be arbitrarily large. The relationship between the conditional distributions on the left therefore varies across levels of z, ℓ in a problem-dependent way. By contrast, in *FairLeap* we directly use (an empirical estimate of) CDD in the ℓ_p sense as the regularizer. Moreover, Proposition 4.2 shows that the bi-causal distance (with large enough C) employed in *FairBiT* is equivalent to the CDD in the Wasserstein sense, irrespective of the data generating distribution.

Note additionally that when $f(Z)$ is continuous, closeness in the distributions $\mathbb{P}(Z \mid A)$ does not necessarily imply closeness in the distributions $\mathbb{P}(f(Z) \mid A)$. For example, consider a simple setting in which $Z \mid A = 0 \sim N(0, 1)$ and $Z \mid A = 1 \sim N(\delta, 1)$ with $\delta > 0$. It is easy to see that for common distance measures such as $1-$ and $2-$Wasserstein distance, total variation distance, etc., the distance between these two distributions can be made arbitrarily small by making δ small. However, for any fixed δ, the distance between $\mathbb{P}(f(Z) \mid A = 1)$ and $\mathbb{P}(f(Z) \mid A = 0)$ can be made arbitrarily large by making $f(Z)$ an appropriate increasing function of Z. This further complicates the use of DCFR outside a classification setting.

D.1 Data generating process for Figure 1

To illustrate the relationship between DCFR and CC, we consider a simple loan setting, in which the sensitive feature A represents sex (male vs. female), $X = L = \emptyset$, and $f(X, A) \in \{0, 1\}$ represents whether an applicant is approved for a loan or not, with a, b, c, d representing the probability mass in each cell:

	Male	Female
Approved	a	b
Not approved	c	d

In this simple setting, since $L = \emptyset$, there are no levels to aggregate across when computing the conditional demographic disparity, so $\mathsf{CDD}_f^{\mathsf{wass}} = \mathsf{CDD}_f^{\ell_p}$, and we denote them both by CDD. We have:

$$
\begin{aligned}
\mathsf{CDD} &= \frac{a}{a + c} - \frac{b}{b + d}\\
R_{\text{DCFR}} &= \frac{a}{a + b} - \frac{c}{c + d}.
\end{aligned}
$$

In this simple example, reducing CDD implies minimizing the difference in acceptance rate by sex, while reducing R_{DCFR} implies minimizing the difference between the ratio of admitted males and the proportion of males in the applicant group.

To produce Figure 1 in Section 5.1, we create a simulation in which we repeatedly sample $10,000$ applicants, varying both the proportion of males in the applicant pool r_m and the difference in acceptance rate by a hyperparameter δ as follows:

$$
\begin{aligned}
&\mathbb{P}(\text{Approve}|\text{Male}) = 0.5 + \delta,\\
&\mathbb{P}(\text{Approve}|\text{Female}) = 0.5 - \delta, \qquad \text{for } \delta \in [0, 0.5).
\end{aligned}
$$

For each sample, we compute R_{DCFR} and CDD as above, and we compute the *FairBiT* regularizer (which here is equivalent to the *FairLeap* regularizer due to the fact that $L = \emptyset$) as described in Section 4.1. Figure 1 illustrates how the relationship between R_{DCFR} and CDD depends on the proportion of males. Each line represents different values of δ for a fixed proportion of males r_m. The closer r_m is to 1, the steeper the slope in the CDD - R_{DCFR} curve. A given small value of R_{DCFR} can correspond to an arbitrarily large value of CDD. The only case where CDD and R_{DCFR} are equal in this example is when $r_m = 0.5$, i.e., the proportions in the candidate pool are balanced. In contrast, *FairBiT* (and equivalently, *FairLeap*) enjoy a consistent relationship with CDD, with some variance along the 45 degree line due to the finite sample size in the example.

Note that since $L = \emptyset$ in this example, this example is equivalent to fixing the *FairBiT* regularizer and R_{DCFR} for a specific level $L = l$. This example shows that R_{DCFR} does not promote small demographic disparity in each level, and therefore it does not promote small CDD, regardless of how level-wise disparities are aggregated.

D.2 Proof of Proposition D.1

Proof. Let $N_{\text{left}}, D_{\text{left}}$ denote the numerator and denominator on the left hand side of the proposition:

$$N_{\text{left}} = |\mathbb{P}(Z = z \mid A = 1, L = \ell) - \mathbb{P}(Z = z \mid A = 0, L = \ell)|$$

$$D_{\text{left}} = |\mathbb{P}(A = 1 \mid Z = z, L = \ell) - \mathbb{P}(A = 1 \mid L = \ell)|$$

First we note that:

$$\mathbb{P}(Z = z|L = \ell) = \mathbb{P}(Z = z \mid A = 1, L = \ell)\mathbb{P}(A = 1 \mid L = \ell) + \mathbb{P}(Z = z \mid A = 0, L = \ell)\mathbb{P}(A = 0 \mid L = \ell)$$

Therefore,

$$\mathbb{P}(Z = z \mid A = 0, L = \ell) = \frac{\mathbb{P}(Z = z \mid L = \ell)}{\mathbb{P}(A = 0 \mid L = \ell)} - \frac{\mathbb{P}(Z = z \mid A = 1, L = \ell)\mathbb{P}(A = 1 \mid L = \ell)}{\mathbb{P}(A = 0 \mid L = \ell)},$$

which we can use to rewrite N_{left} as:

$$\begin{aligned}
&|\mathbb{P}(Z = z \mid A = 1, L = \ell) - \frac{\mathbb{P}(Z = z \mid L = \ell)}{\mathbb{P}(A = 0 \mid L = \ell)} + \frac{\mathbb{P}(Z = z \mid A = 1, L = \ell)\mathbb{P}(A = 1 \mid L = \ell)}{\mathbb{P}(A = 0 \mid L = \ell)}| \\
&= |\frac{-\mathbb{P}(Z = z \mid L = \ell)}{\mathbb{P}(A = 0 \mid L = \ell)} + \frac{\mathbb{P}(Z = z \mid A = 1, L = \ell)[\mathbb{P}(A = 0) + \mathbb{P}(A = 1)]}{\mathbb{P}(A = 0 \mid L = \ell)}| \\
&= |\frac{\mathbb{P}(Z = z \mid A = 1, L = \ell) - \mathbb{P}(Z = z \mid L = \ell)}{\mathbb{P}(A = 0 \mid L = \ell)}|,
\end{aligned}$$

since $\mathbb{P}(A = 0) + \mathbb{P}(A = 1) = 1$ as the sensitive feature A is binary. We then note that:

$$\mathbb{P}(A = 1 \mid Z = z, L = \ell) = \frac{\mathbb{P}(Z = z \mid A = 1, L = \ell)\mathbb{P}(A = 1 \mid L = \ell)}{\mathbb{P}(Z = z \mid L = \ell)}$$

Therefore,

$$\frac{\mathbb{P}(A = 1 \mid Z = z, L = \ell)}{\mathbb{P}(A = 1 \mid L = \ell)} = \frac{\mathbb{P}(Z = z \mid A = 1, L = \ell)}{\mathbb{P}(Z = z \mid L = \ell)} := \gamma.$$

Given the definitions above, we have that:

$$\begin{aligned}
&\frac{N_{\text{left}}\mathbb{P}(A = 0 \mid L = \ell)}{\mathbb{P}(Z = z \mid L = \ell)} = |\gamma - 1|, \\
&\frac{D_{\text{left}}}{\mathbb{P}(A = 1 \mid L = \ell)} = |\gamma - 1| \\
\implies &\frac{N_{\text{left}}}{D_{\text{left}}} = \frac{\mathbb{P}(Z = z \mid L = \ell)}{\mathbb{P}(A = 0 \mid L = \ell)\mathbb{P}(A = 1 \mid L = \ell)}
\end{aligned}$$

□

E Proofs

In this section, we provide the proofs of Propositions A.3, 4.2, and 5.1.

Proof of Proposition 4.2. Consider CDD in Wasserstein sense $\mathsf{CDD}_f^{\mathsf{wass}} = \mathcal{W}_p^p(\mathbb{P}(L|A = 0), \mathbb{P}(L|A = 1); D)$, with

$$D_{cdd}(l, l') = \begin{cases} \mathsf{d}_{cdd}\big(\mathbb{P}(f(X) \mid L = l, A = 0), \\ \qquad \mathbb{P}(f(X) \mid L = l', A = 1)\big), \ l = l', \\ \infty, \qquad \text{elsewhere.} \end{cases}$$

Define the set $\mathbb{B} = \{(l, l')| \, l, l' \in \mathcal{S}, \ l \neq l'\}$. Consider two transport plans $\gamma_1, \gamma_2 \in \Gamma(\mathbb{P}(L|A = 0), \mathbb{P}(L|A = 1)$ such that $\forall(l, l') \in \mathbb{B}: \ \gamma_1(l, l') = 0$ but $\exists(l, l') \in \mathbb{B}: \ \gamma_2(l, l') \neq 0$. It follows from the definition of $\mathsf{CDD}_f^{\mathsf{wass}}$ that for bounded d, we have $D_{cdd}(l, l) < D_{cdd}(l, l')$ for all $l \neq l'$. Therefore, transport plan γ_2 cannot be an optimal transport plan of the optimal transport problem with cost function $D_{cdd}(l, l')$, since any transport plan γ_1 such that $\forall(l, l') \in \mathbb{B}: \ \gamma_1(l, l') = 0$ will have a lower cost.

On the other hand, Proposition A.3 states that $\mathcal{C}_b^p(\tilde{\mathbb{P}}, \mathbb{P})$ can be written as $\mathcal{W}_p^p(\tilde{\mathbb{P}}_{\tilde{U}}, \mathbb{P}_U; D_{bcd})$ where

$$D_{bcd}^p(\tilde{U}, U) = \mathcal{W}_p^p(\tilde{\mathbb{P}}_{\tilde{V}|\tilde{U}}, \mathbb{P}_{V|U}) + C\|\tilde{U} - U\|_p^p.$$

Therefore, when $C > \frac{\bar{d}}{\underline{d}}$, we have

$$\begin{aligned}
\forall(l, l') \in \mathbb{B}: \quad &C\|l' - l\|_p^p \\
&> \bar{d} \\
&\geq \mathcal{W}_p^p\big(\mathbb{P}(f(X) \mid L = l, A = 0), \\
&\qquad \mathbb{P}(f(X) \mid L = l, A = 1)\big).
\end{aligned}$$

It follows that any transport at the same level of L has a lower cost than any transport between different levels. Therefore, the optimal transport plan γ with cost function D_{bcd} must satisfy $\forall(l, l') \in \mathbb{B}: \ \gamma(l, l') = 0$.

Therefore, when $\mathsf{d}_{cdd}(\cdot, \cdot) = \mathcal{W}_p^p(\cdot, \cdot)$ and $C > \frac{\bar{d}}{\underline{d}}$, both $\mathsf{CDD}_f^{\mathsf{wass}}$ and $\mathcal{C}_b^p$ correspond to optimal transport plans with zero mass on $\mathbb{B}$ and have the same cost functions when $l = l'$, and therefore they are equivalent. □

Proof of Proposition A.3. It follows from Definition 4.1 that bi-causal distance can be written as

$$\mathcal{C}_b^p(\tilde{\mathbb{P}}, \mathbb{P}) :=$$

$$\inf_{\gamma \in \Gamma_{bc}(\tilde{\mathbb{P}}, \mathbb{P})} \mathbb{E}_{((\tilde{U}, \tilde{V}),(U,V)) \sim \gamma} \left[C \|\tilde{U} - U\|_p^p + \|\tilde{V} - V\|_p^p \right] =$$

$$\inf_{\gamma_1 \in \Gamma(\tilde{\mathbb{P}}_{\tilde{U}}, \mathbb{P}_U)} \mathbb{E}_{(\tilde{U}, U) \sim \gamma_1} \Bigg[\inf_{\gamma_2 \in \Gamma(\tilde{\mathbb{P}}_{\tilde{V}|\tilde{U}}, \mathbb{P}_{V|U})} \mathbb{E}_{(\tilde{V}, V) \sim \gamma_2} \Big[C \|\tilde{U} - U\|_p^p + \|\tilde{V} - V\|_p^p \mid (\tilde{U}, U) \Big] \Bigg]$$

□

Proof of Proposition 5.1. By definition of $Q(h)$ in Xu et al. (2020a):

$$\begin{aligned} Q(h) &= \mathbb{E}[\mathbf{1}_{\{A=1\}} \, \mathbb{P}(A=0|L) h(Z,L)] \\ &\quad - \mathbb{E}[\mathbf{1}_{\{A=0\}} \, \mathbb{P}(A=1|L) h(Z,L)] \\ &= \mathbb{E}\left[\left(\mathbf{1}_{\{A=1\}} \, \mathbb{P}(A=0|L) \right.\right. \\ &\quad \left.\left. - \mathbf{1}_{\{A=0\}} \, \mathbb{P}(A=1|L)\right) h(Z,L)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\{A=1\}} \, \mathbb{P}(A=0|L) \right.\right. \\ &\quad \left.\left. - \mathbf{1}_{\{A=0\}} \, \mathbb{P}(A=1|L) | Z, L\right] h(Z,L)\right] \\ &= \mathbb{E}\Bigg[\Big(\mathbb{P}(A=1|Z,L) \, \mathbb{P}(A=0|L) \\ &\quad - \mathbb{P}(A=0|Z,L) \, \mathbb{P}(A=1|L)\Big) h(Z,L)\Bigg] \\ &= \mathbb{E}[g(Z,L) h(Z,L)]. \end{aligned}$$

Since $h \in L^2$ and $0 \leq h(Z,L) \leq 1$, then it follows that:

$$\begin{aligned} h^\star &= \arg\sup_h Q(h) \\ &= \mathbf{1}_{\{g(Z,L)>0\}} \\ &= \mathbf{1}_{\{\mathbb{P}(A=1|Z,L) \, \mathbb{P}(A=0|L) - \mathbb{P}(A=0|Z,L) \, \mathbb{P}(A=1|L) > 0\}} \\ &= \mathbf{1}_{\left\{\frac{\mathbb{P}(A=1|Z,L)}{\mathbb{P}(A=1|L)} > \frac{\mathbb{P}(A=0|Z,L)}{\mathbb{P}(A=0|L)}\right\}} \\ &= \mathbf{1}_{\{\mathbb{P}(A=1|Z,L) > \mathbb{P}(A=1|L)\}}. \end{aligned}$$

Hence we have that:

$$\begin{aligned} R_{\text{DCFR}} &= Q(h^\star) \\ &= \mathbb{E}[\big(\mathbb{P}(A=1|Z,L) \, \mathbb{P}(A=0|L) \\ &\quad - \mathbb{P}(A=0|Z,L) \, \mathbb{P}(A=1|L)\big)_+] \\ &= \mathbb{E}[(\mathbb{P}(A=1|Z,L) - \mathbb{P}(A=1|L))_+] \end{aligned}$$

□

F Experiments Details

This Section includes a more detailed breakdown of the numerical experiments. Section F.1 provides more information on the datasets used, Section F.2 lists all the implementation details including hyper-parameter ranges for each method, Section F.3 provides further discussion on the fairness-performance tradeoffs of all methods, Section F.4 provides an exhaustive inventory of all results, including additional figures in this appendix. Finally, Section F.5 includes any further considerations which was not included in the main paper due to space concerns.

F.1 Datasets Details

1. `Adult` dataset (Becker and Kohavi 1996). The Adult dataset is based on the 1994 U.S. census data. It contains demographic information for $48,842$ respondents. It contains 14 features and a target variable (`income`: whether income is over \$$50,000$ or not). We consider `sex` as the sensitive feature and `occupation` as the legitimate feature. In this dataset, the occupation `Armed Forces` and `Priv-House-serv` are very rare, so we combine them with `Protective Services`. We encode the values of `occupation` as integer values, i.e. $L \in \{0,..,11\}$, resulting in 10 levels. We use a subsample (selected uniformly at random without replacement) of size $5,000$ for computational reasons (see details below).
2. `Drug` dataset (Fehrman et al. 2017). This dataset contains self-declared drug usage history for 1885 respondents, for 19 different drugs. For each drug, the respondents declare the time of the last consumption (possible responses are: never, over a decade ago, in the last decade/year/month/week, or on the last day). It also has demographic variables and scores on several psychological tests. In our classification problem, the task is to predict the response "never used" versus "others" (i.e., used) for `cannabis`. We treat `education` as the legitimate feature ($L \in \{0,..,8\}$, resulting in 9 levels) and `gender` as the sensitive feature.
3. `Law School` dataset (Ramsey, Wightman, and Council 1998). This dataset comes from a survey conducted by the Law School Admission Council (LSAC) across 163 law schools in the United States in 1991. It contains law school admission records and law school performance of $21,027$ students. The prediction task is to predict the students' first-year average grade (`zfygpa`). We use the candidate LSAT score as legitimate feature (discretized in deciles, $L \in \{0,..,9\}$, resulting in 10 levels) and the candidate gender as the sensitive feature.
4. `Communities and Crime` dataset (Redmond 2009). This dataset combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. It includes data for $1,994$ communities across the United States, represented by 126 socio-economic, demographic, and law enforcement-related features. The target variable is `ViolentCrimesPerPop` which represents the rate of violent crimes per $100,000$ population in each community. We cosider `racepctblack` (percentage of the population in a community that is Black) as the sensitive feature. We cosider `medIncome` (the median income of the community) as the legitimate feature (discretized in deciles, $L \in \{0,..,9\}$, resulting in 10 levels).

See Table 5 for the list of the features used in each dataset.

Unfairness metric	*Figures*
$\mathsf{CDD}_f^{\mathsf{wass}}$	2(top row)
$\mathsf{CDD}_f^{\ell_p}$(uniform)	2(bottom row)
$\mathsf{CDD}_f^{\ell_p}(P(L))$	4(top row)
$\mathsf{CDD}_f^{\ell_p}$(Ave. $P(L\|A)$)	4(middle row)
DP	4(bottom row)

Table 3: List of all figures for all experiments across different fairness metrics.

Dataset	*Tables*
`Adult`	6, 7
`Drug`	8, 9
`Law School`	10, 11
`Communities and Crime`	12, 13

Table 4: List of all tables for all experiments across the four datasets.

Dataset	Input features	Legitimate feature	Sensitive feature	Target
`Adult`	Age, Workclass, Fnlwgt, Education, Education-num, Capital-loss, Race, Capital-gain, Relationship, Hours-per-week, Native-country, Marital-status	Occupation	Sex	Income
`Drug`	Education, Country, Nscore, Escore, Oscore, Ascore, Cscore, Impulsive, SS	Education	gender	Cannabis
`Law School`	ugpa, fam_inc, fulltime, tier, lsat_decile	lsat_decile	gender	zfygpa
`Communities and Crime`	124 features (see Section F.1)	medIncomeDecile	racepctblack	ViolentCrimesPerPop

Table 5: List of variables for `Adult`, `Drug`, `Law School`, and `Communities and crime` datasets.

F.2 Implementation Details

In all experiments we use a two-layer deep multi-layer perceptron (MLP), with hidden layers including 50 and 20 nodes, and a rectified linear unit (ReLu) activation function. We train the MLP using stochastic Adam optimizer, with 500 training epochs and a learning rate of $l_r = 0.001$, for the all methods and datasets. All methods are equipped with early stopping, with a patience set to 50 epochs. We chose these values since they resulted in the convergence of the training procedure for all methods under consideration to non-trivial predictors, with reasonable model variance across 10 runs. The constant C in the computation of $\mathsf{CDD}_f^{\mathsf{wass}}$ (see Proposition 4.2) is set to $C = 1$ for all four datasets, which is larger than the lower bound $\frac{\bar{d}}{\underline{d}}$ in all cases and hence satifies Proposition 4.2. Finally, during training we apply stratified sampling for generating each training batch to ensure at least one sample from each level of the legimitate feature L is included on a given batch. To do so, we set the batch size to 256 for `Drug`, 1024 for `Adult` and 128 for `Law School` and `Crime`.

For *Wasserstein regularization* and *FairLeap* we compute the 1-Wasserstein distance, i.e., setting $p = 1$. For *FairBiT*, we compute the Sinkhorn approximation using the `sinkhorn` and `sinkhorn2` functions from the Python optimal transport (POT) package (Flamary et al. 2021), which optimizes for the 2-Wasserstein distance, i.e., $p = 2$ for both the inner and outer metrics in $\mathsf{CDD}^{\mathsf{wass}}$. The details for each method are as follows:

- For *FairBiT* and *Wasserstein regularization*, values of λ_b and λ_w are 10 logarithmically-spaced numbers in the range $[0.0001, 10]$:

$$\lambda_b, \lambda_w \in \{0.0001, 0.0005, 0.0022, 0.0100, 0.0464, \\ 0.2154, 1.0000, 4.6416, 21.5443, 100\}$$

- For all 3 versions of *FairLeap*, values of λ_{lp} are 10 logarithmically-spaced numbers in the range $[0.0001, 100]$:

$$\lambda_b, \lambda_w \in \{0.0001, 0.0004, 0.0013, 0.0046, 0.0167, \\ 0.0599, 0.2154, 0.7743, 2.7826, 10\}$$

- For Pre-processing repair, we select $\lambda_r \in [0, 1]$, with the following values:

$$\lambda_r \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$$

- For DCFR, we consider the same range $(0, 20]$ as in (Xu et al. 2020a), with the following values:

$$\lambda_{DCFR} \in \{0.1, 0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 5.0, 10.0, 15.0, 20.0\}$$

Finally, our *FairBiT* method requires substantial computation due to our non-optimized Python implementation, unlike all other approaches which utilize direct calls from the Python Optimal transport package (POT, Flamary et al. 2021), which are natively written in C++ with a Python binding. For instance, it requires around 2 hours to train one model using the full `Law School` dataset on a machine equipped with 32GB of RAM, a 16GB Tesla T4 and a 2.50GHz Intel(R) Xeon(R) Platinum 8259CL CPU, while for *FairLeap*, as well as other methods, the same training is executed in less than 10 minutes.

F.3 Additional Results

Figure 3 includes the performance of all methods (including *legit-only*) with fairness metrics $\mathsf{CDD}^{\mathsf{wass}}$ (normalized version) on the top row and CDD^{ℓ_p} (uniform) on the bottom row. We normalize the $\mathsf{CDD}^{\mathsf{wass}}$ distance by substracting the Wasserstein distance between the distribution of the legitimate feature at the two levels of the sensitive feature, i.e., subtracting $\mathcal{W}(\mathbb{P}(L|A=0), \mathbb{P}(L|A=1))$. This corresponds to the term $C\|\tilde{U} - U\|^p$ in Definition 4.1, which is not dependent on the model $f(X)$ and as such equal across all methods. Figure 3 is a repetition of the results presented in Figure 2 in the main body, with the addition of the results for *legit-only*. *Legit-only* completely removes conditional disparities but significantly suffers in terms of predictive performance. This is expected behavior since it relies only on a few (in our experiments, only 1) feature(s).

Figure 4 presents the performance of all methods (including *legit-only*) with fairness metrics CDD^{ℓ_p} $(P(L))$, CDD^{ℓ_p} (Ave. $P(L|A)$), and demographic parity on the top, middle, and bottom rows, respectively. Overall, when fairness is measured by CDD^{ℓ_p} metrics, *FairBiT* and the variants of *FairLeap* are consistently among the highest performing, often providing better fairness-predictive power trade-offs than the other proposed methods. When fairness is measured by DD, unsurprisingly *Wasserstein Reg.* performs the best. In this case, DCFR has good performance especially on regression datasets with some points on the Pareto frontier. *FairBiT* and the variants of *FairLeap* generally do not perform as well on demographic disparity across all datasets, although they are still part of or close to the Pareto frontier in classification settings.

F.4 Results Details

We provide the mean and standard deviation results for all our methods (for all regularization parameter values where applicable) in Tables 6 through 13. Tables 3 and 4 below provide exhaustive pointers to the respective result figures and tables for each experiment across different datasets and fairness metrics. See Sections 6, F.3 and Figures captions for comments and takeaways on models performance.

F.5 Discussion on our choices of evaluation metrics for predictive performance

While some previous papers use the AUC of the performance-fairness curves (see e.g., Rychener, Taskesen, and Kuhn 2022) to describe the overall fairness-performance behavior of each predictor, in our experiments we find this measure to be misleading since the ranges of the fairness and performance results are different across different methods. As we are not aware of any single measure that can appropriately summarize the overall fairness-performance behavior of a predictor, we report results as the means and standard deviations over 10 runs across separate random seeds for each method.

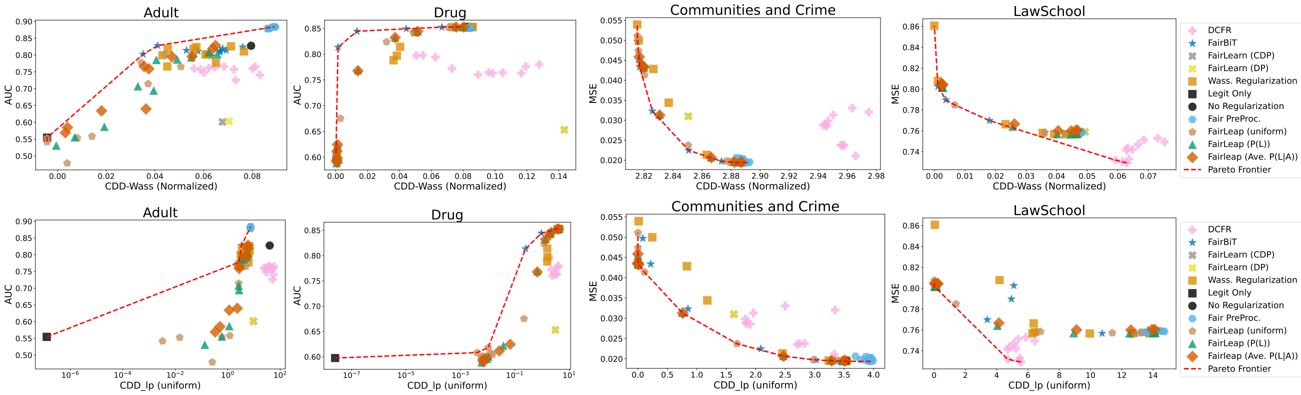

Figure 3: Fairness-predictive power trade-offs for `Adult`, `Drug`, `Communities and Crime`, and `LawSchool` datasets. The figures in the top row present the results when fairness is measured by $\text{CDD}_f^{\text{wass}}$. The results when fairness metric is $\text{CDD}_f^{\ell_p}$ (with $\mathbb{Q}(L) = \mathbb{U}(L)$ and $p = 1$) are presented in the bottom row. Predictive power (PP) is measured by AUC for Classification and MSE for regression. Results are averaged over 10 runs, with different values for the same methods due to different hyper-parameter settings; see Appendix F for details. These figures include *legit-only* as a reference. *Legit-only*, unsurprisingly, achieves full conditional parity but significantly suffers in terms of predictive performance.

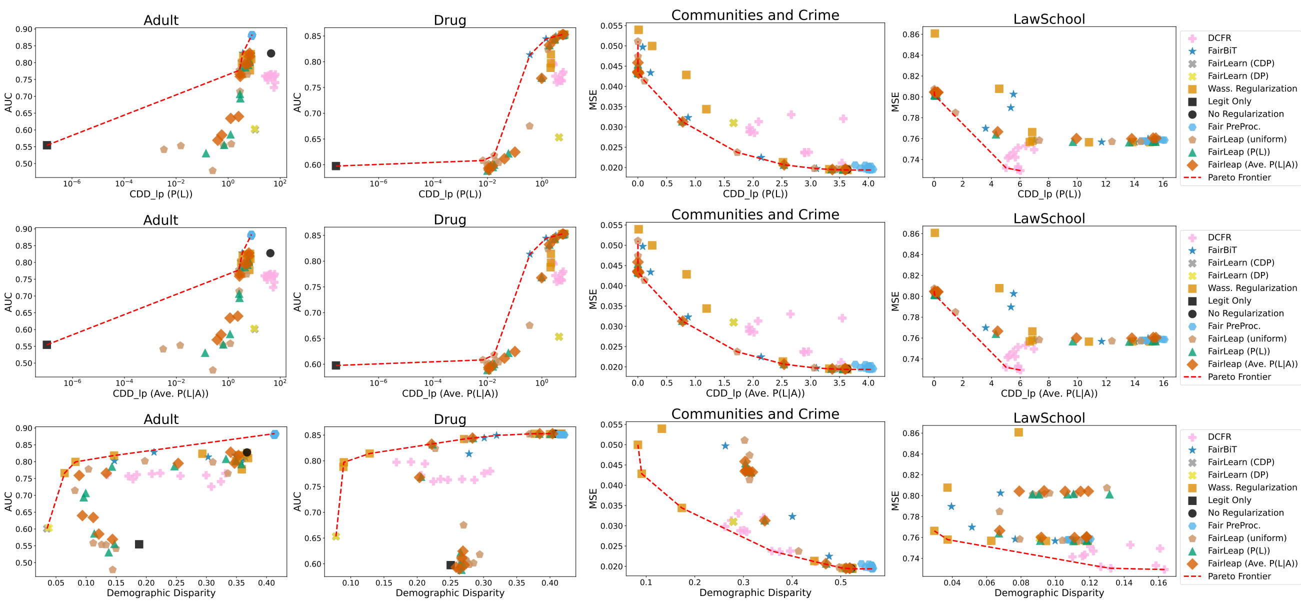

Figure 4: Fairness-performance trade-offs for `Adult`, `Drug`, `Communities and Crime`, and `LawSchool` datasets. In the top row, we show the results when fairness is measured by $\text{CDD}_f^{\ell_p}$ with $\mathbb{Q}(L) = \mathbb{P}(L)$ and $p = 1$. The figures in the middle row show the results when fairness is measured by $\text{CDD}_f^{\ell_p}$ with $\mathbb{Q}(L) = \frac{\mathbb{P}(L|A=0)+\mathbb{P}(L|A=1)}{2}$ and $p = 1$. The figures in the middle row show the results when fairness is measured by demographic disparity (DD), specifically using the 1-Wasserstein distance. Predictive power is measured by AUC for Classification and MSE for regression. Results are averaged over 10 runs, with different values for the same methods due to different hyper-parameter settings; see Appendix F for details. Overall, when fairness is measured by CDD metrics, *FairBiT* and the variants of *FairLeap* are consistently among the highest performing, often providing better fairness-predictive power trade-offs than the other proposed methods. When fairness is measured by DD, unsurprisingly *Wasserstein Reg.* performs the best. In this case, DCFR has good performance especially on regression datasets with some points on the frontier. *FairBiT* and the variants of *FairLeap* generally do not improve demographic disparity by much, although they still have many points on the Pareto frontier in classification.

Method	AUC	$\mathsf{CDD}^{\mathsf{wass}}$ (normalized)	CDD^{ℓ_p} (uniform)	CDD^{ℓ_p} (Ave. $\mathbb{P}(L\|A)$)	CDD^{ℓ_p} ($\mathbb{P}(L)$)
DCFR, λ_d :0.10	0.74 ± 0.12	0.083 ± 0.029	561.824 ± 97.038	63.648 ± 11.645	59.971 ± 10.741
DCFR, λ_d :0.25	0.76 ± 0.13	0.080 ± 0.029	533.695 ± 99.986	60.697 ± 12.268	57.169 ± 11.290
DCFR, λ_d :0.50	0.76 ± 0.13	0.073 ± 0.030	452.759 ± 92.157	51.592 ± 11.407	48.545 ± 10.507
DCFR, λ_d :0.75	0.77 ± 0.12	0.068 ± 0.029	398.082 ± 92.588	45.366 ± 11.173	42.859 ± 10.354
DCFR, λ_d :1.00	0.76 ± 0.12	0.065 ± 0.027	363.625 ± 88.308	41.365 ± 10.721	39.031 ± 9.847
DCFR, λ_d :1.50	0.75 ± 0.13	0.060 ± 0.027	297.195 ± 91.580	33.682 ± 10.967	31.769 ± 10.187
DCFR, λ_d :2.00	0.76 ± 0.10	0.059 ± 0.030	278.981 ± 73.521	31.542 ± 8.868	29.892 ± 8.116
DCFR, λ_d :5.00	0.76 ± 0.11	0.056 ± 0.035	239.655 ± 72.522	26.636 ± 8.521	25.056 ± 7.963
DCFR, λ_d :10.00	0.76 ± 0.12	0.062 ± 0.035	318.831 ± 223.980	35.677 ± 25.750	33.521 ± 24.104
DCFR, λ_d :15.00	0.77 ± 0.12	0.081 ± 0.022	574.569 ± 134.654	65.306 ± 16.085	61.141 ± 14.891
DCFR, λ_d :20.00	0.73 ± 0.09	0.073 ± 0.031	511.350 ± 103.351	57.783 ± 11.787	54.074 ± 11.099
FairBiT, λ_b :0.00	0.81 ± 0.01	0.058 ± 0.016	55.469 ± 7.486	6.168 ± 0.990	5.726 ± 0.870
FairBiT, λ_b :0.00	0.81 ± 0.03	0.067 ± 0.022	61.142 ± 13.468	6.867 ± 1.717	6.369 ± 1.532
FairBiT, λ_b :0.00	0.81 ± 0.05	0.067 ± 0.017	62.883 ± 9.146	7.109 ± 1.211	6.589 ± 1.096
FairBiT, λ_b :0.00	0.82 ± 0.02	0.076 ± 0.021	67.350 ± 9.651	7.573 ± 1.386	7.037 ± 1.225
FairBiT, λ_b :0.02	0.80 ± 0.05	0.063 ± 0.021	60.261 ± 10.298	6.749 ± 1.354	6.286 ± 1.198
FairBiT, λ_b :0.06	0.81 ± 0.06	0.069 ± 0.018	63.161 ± 7.356	7.079 ± 1.109	6.576 ± 0.910
FairBiT, λ_b :0.22	0.82 ± 0.04	0.068 ± 0.018	61.223 ± 9.513	6.841 ± 1.230	6.391 ± 1.125
FairBiT, λ_b :0.77	0.81 ± 0.06	0.053 ± 0.019	46.851 ± 8.784	5.195 ± 1.199	4.878 ± 1.011
FairBiT, λ_b :2.78	0.83 ± 0.02	0.041 ± 0.012	33.622 ± 4.609	3.655 ± 0.628	3.457 ± 0.542
FairBiT, λ_b :10.00	0.80 ± 0.02	0.035 ± 0.015	29.978 ± 3.358	3.239 ± 0.398	3.011 ± 0.355
FairLearn (CDP)	0.60 ± 0.04	0.068 ± 0.038	95.417 ± 85.756	10.887 ± 9.986	10.340 ± 9.487
FairLearn (DP)	0.60 ± 0.04	0.070 ± 0.040	91.254 ± 89.047	10.468 ± 10.417	9.919 ± 9.871
Wass. Regularization, λ_w :0.00	0.82 ± 0.01	0.056 ± 0.015	55.296 ± 10.003	6.249 ± 1.330	5.786 ± 1.141
Wass. Regularization, λ_w :0.00	0.80 ± 0.05	0.062 ± 0.022	58.566 ± 12.635	6.611 ± 1.669	6.135 ± 1.453
Wass. Regularization, λ_w :0.00	0.83 ± 0.02	0.071 ± 0.023	66.117 ± 12.194	7.477 ± 1.608	6.944 ± 1.432
Wass. Regularization, λ_w :0.00	0.81 ± 0.05	0.077 ± 0.028	68.589 ± 17.184	7.735 ± 2.068	7.182 ± 1.882
Wass. Regularization, λ_w :0.02	0.78 ± 0.08	0.065 ± 0.016	61.582 ± 11.624	6.929 ± 1.359	6.459 ± 1.254
Wass. Regularization, λ_w :0.06	0.80 ± 0.06	0.061 ± 0.022	57.394 ± 9.826	6.481 ± 1.250	6.046 ± 1.073
Wass. Regularization, λ_w :0.22	0.82 ± 0.04	0.057 ± 0.022	49.628 ± 11.510	5.516 ± 1.428	5.211 ± 1.281
Wass. Regularization, λ_w :0.77	0.82 ± 0.01	0.045 ± 0.016	34.830 ± 4.871	3.834 ± 0.676	3.605 ± 0.609
Wass. Regularization, λ_w :2.78	0.80 ± 0.02	0.043 ± 0.012	31.387 ± 3.768	3.391 ± 0.489	3.171 ± 0.434
Wass. Regularization, λ_w :10.00	0.77 ± 0.04	0.045 ± 0.012	32.547 ± 3.389	3.493 ± 0.458	3.249 ± 0.395
Legit Only	0.55 ± 0.02	-0.004 ± 0.014	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
No Regularization	0.83 ± 0.03	0.080 ± 0.034	39.842 ± 35.694	44.793 ± 40.119	41.785 ± 37.418
Fair PreProc., λ_r :0.00	0.88 ± 0.01	0.090 ± 0.014	7.445 ± 0.560	8.352 ± 0.734	7.850 ± 0.641
Fair PreProc., λ_r :0.10	0.88 ± 0.01	0.089 ± 0.014	7.433 ± 0.570	8.339 ± 0.754	7.836 ± 0.658
Fair PreProc., λ_r :0.20	0.88 ± 0.01	0.089 ± 0.014	7.431 ± 0.535	8.337 ± 0.740	7.834 ± 0.636
Fair PreProc., λ_r :0.30	0.88 ± 0.01	0.089 ± 0.014	7.422 ± 0.478	8.327 ± 0.687	7.827 ± 0.587
Fair PreProc., λ_r :0.40	0.88 ± 0.01	0.089 ± 0.014	7.439 ± 0.493	8.346 ± 0.711	7.842 ± 0.603
Fair PreProc., λ_r :0.50	0.88 ± 0.01	0.088 ± 0.014	7.415 ± 0.506	8.319 ± 0.744	7.816 ± 0.620
Fair PreProc., λ_r :0.60	0.88 ± 0.01	0.087 ± 0.012	7.301 ± 0.681	8.201 ± 0.928	7.703 ± 0.809
Fair PreProc., λ_r :0.70	0.88 ± 0.01	0.087 ± 0.013	7.276 ± 0.696	8.174 ± 0.956	7.678 ± 0.834
Fair PreProc., λ_r :0.80	0.88 ± 0.01	0.087 ± 0.013	7.282 ± 0.717	8.184 ± 0.984	7.687 ± 0.854
Fair PreProc., λ_r :0.90	0.88 ± 0.01	0.087 ± 0.014	7.274 ± 0.834	8.173 ± 1.116	7.674 ± 0.974
Fair PreProc., λ_r :1.00	0.88 ± 0.01	0.086 ± 0.014	7.257 ± 0.840	8.154 ± 1.122	7.655 ± 0.985

Table 6: Mean and one standard deviation results for all methods-hyperparameter combinations for all methods except *FairLeap*, for the `Adult` dataset across 10 separate runs.

Method	AUC	$\mathsf{CDD}^{\mathsf{wass}}$ (normalized)	CDD^{ℓ_p} (uniform)	CDD^{ℓ_p} (Ave. $\mathbb{P}(L\|A)$)	CDD^{ℓ_p} ($\mathbb{P}(L)$)
FairLeap (uniform), λ_w :0.00	0.77 ± 0.08	0.051 ± 0.018	49.145 ± 13.342	5.508 ± 1.748	5.127 ± 1.584
FairLeap (uniform), λ_w :0.00	0.80 ± 0.05	0.056 ± 0.024	51.151 ± 10.382	5.711 ± 1.349	5.318 ± 1.229
FairLeap (uniform), λ_w :0.00	0.80 ± 0.03	0.044 ± 0.017	35.180 ± 7.448	3.764 ± 0.963	3.546 ± 0.833
FairLeap (uniform), λ_w :0.01	0.78 ± 0.06	0.034 ± 0.015	26.346 ± 5.901	2.739 ± 0.606	2.579 ± 0.607
FairLeap (uniform), λ_w :0.05	0.71 ± 0.04	0.037 ± 0.012	26.331 ± 6.101	2.865 ± 0.756	2.640 ± 0.676
FairLeap (uniform), λ_w :0.22	0.56 ± 0.12	0.014 ± 0.027	11.964 ± 10.538	1.328 ± 1.195	1.202 ± 1.080
FairLeap (uniform), λ_w :1.00	0.55 ± 0.13	0.008 ± 0.027	6.054 ± 9.312	0.672 ± 1.054	0.621 ± 0.980
FairLeap (uniform), λ_w :4.64	0.55 ± 0.20	-0.004 ± 0.014	0.148 ± 0.112	0.015 ± 0.011	0.014 ± 0.010
FairLeap (uniform), λ_w :21.54	0.54 ± 0.21	-0.004 ± 0.014	0.033 ± 0.024	0.003 ± 0.002	0.003 ± 0.002
FairLeap (uniform), λ_w :100.00	0.48 ± 0.18	0.004 ± 0.012	2.580 ± 5.533	0.254 ± 0.556	0.255 ± 0.560
FairLeap (P(L)), λ_w :0.00	0.79 ± 0.04	0.055 ± 0.017	53.119 ± 9.445	5.962 ± 1.330	5.537 ± 1.181
FairLeap (P(L)), λ_w :0.00	0.80 ± 0.05	0.066 ± 0.018	60.711 ± 11.942	6.823 ± 1.558	6.334 ± 1.416
FairLeap (P(L)), λ_w :0.00	0.81 ± 0.05	0.062 ± 0.018	56.668 ± 9.152	6.349 ± 1.309	5.919 ± 1.150
FairLeap (P(L)), λ_w :0.01	0.79 ± 0.07	0.049 ± 0.025	41.535 ± 11.905	4.595 ± 1.403	4.307 ± 1.302
FairLeap (P(L)), λ_w :0.05	0.79 ± 0.06	0.041 ± 0.016	30.153 ± 4.635	3.219 ± 0.611	3.012 ± 0.524
FairLeap (P(L)), λ_w :0.22	0.71 ± 0.09	0.033 ± 0.015	26.212 ± 7.478	2.851 ± 0.897	2.627 ± 0.809
FairLeap (P(L)), λ_w :1.00	0.69 ± 0.05	0.040 ± 0.017	27.009 ± 7.529	2.950 ± 0.852	2.717 ± 0.770
FairLeap (P(L)), λ_w :4.64	0.59 ± 0.12	0.019 ± 0.021	11.460 ± 8.641	1.246 ± 0.963	1.170 ± 0.910
FairLeap (P(L)), λ_w :21.54	0.56 ± 0.16	0.007 ± 0.029	5.971 ± 12.050	0.705 ± 1.434	0.641 ± 1.297
FairLeap (P(L)), λ_w :100.00	0.53 ± 0.15	-0.001 ± 0.019	1.337 ± 4.006	0.141 ± 0.425	0.128 ± 0.387
Fairleap (Ave. P(L|A)), λ_w :0.00	0.80 ± 0.05	0.055 ± 0.019	53.008 ± 13.259	5.913 ± 1.707	5.490 ± 1.540
Fairleap (Ave. P(L|A)), λ_w :0.00	0.82 ± 0.01	0.064 ± 0.020	59.219 ± 9.992	6.634 ± 1.402	6.167 ± 1.258
Fairleap (Ave. P(L|A)), λ_w :0.00	0.83 ± 0.02	0.065 ± 0.019	59.554 ± 8.023	6.659 ± 1.163	6.205 ± 1.017
Fairleap (Ave. P(L|A)), λ_w :0.01	0.79 ± 0.05	0.047 ± 0.023	40.801 ± 9.111	4.484 ± 1.107	4.221 ± 1.032
Fairleap (Ave. P(L|A)), λ_w :0.05	0.77 ± 0.07	0.036 ± 0.014	27.195 ± 6.888	2.883 ± 0.761	2.720 ± 0.719
Fairleap (Ave. P(L|A)), λ_w :0.22	0.76 ± 0.04	0.038 ± 0.015	27.067 ± 5.693	2.929 ± 0.767	2.715 ± 0.677
Fairleap (Ave. P(L|A)), λ_w :1.00	0.64 ± 0.11	0.036 ± 0.017	23.157 ± 5.814	2.522 ± 0.742	2.322 ± 0.640
Fairleap (Ave. P(L|A)), λ_w :4.64	0.63 ± 0.10	0.018 ± 0.022	11.498 ± 10.472	1.261 ± 1.228	1.165 ± 1.105
Fairleap (Ave. P(L|A)), λ_w :21.54	0.57 ± 0.16	0.003 ± 0.029	3.356 ± 9.768	0.402 ± 1.184	0.374 ± 1.103
Fairleap (Ave. P(L|A)), λ_w :100	0.58 ± 0.19	0.004 ± 0.025	5.057 ± 11.993	0.569 ± 1.372	0.529 ± 1.283

Table 7: Mean and one standard deviation results for all methods-hyperparameter combinations for the 3 variants of *FairLeap*, for the `Adult` dataset across 10 separate runs.

Method	AUC	CDD$^{\text{wass}}$ (normalized)	CDD$^{\ell_p}$ (uniform)	CDD$^{\ell_p}$ (Ave. $\mathbb{P}(L\|A)$)	CDD$^{\ell_p}$ ($\mathbb{P}(L)$)
DCFR, λ_d :0.10	0.78 ± 0.02	0.128 ± 0.019	35.404 ± 5.013	6.070 ± 0.892	6.061 ± 0.892
DCFR, λ_d :0.25	0.78 ± 0.01	0.119 ± 0.022	32.381 ± 6.650	5.504 ± 1.211	5.498 ± 1.216
DCFR, λ_d :0.50	0.76 ± 0.02	0.117 ± 0.017	31.468 ± 5.937	5.411 ± 1.159	5.404 ± 1.146
DCFR, λ_d :0.75	0.76 ± 0.02	0.105 ± 0.016	27.762 ± 4.501	4.827 ± 0.745	4.822 ± 0.751
DCFR, λ_d :1.00	0.76 ± 0.02	0.099 ± 0.011	26.282 ± 2.846	4.403 ± 0.524	4.399 ± 0.526
DCFR, λ_d :1.50	0.76 ± 0.02	0.096 ± 0.014	25.750 ± 5.251	4.189 ± 0.960	4.184 ± 0.964
DCFR, λ_d :2.00	0.76 ± 0.02	0.089 ± 0.018	21.700 ± 6.444	3.530 ± 1.041	3.527 ± 1.043
DCFR, λ_d :5.00	0.77 ± 0.02	0.073 ± 0.011	19.651 ± 3.494	3.117 ± 0.667	3.113 ± 0.664
DCFR, λ_d :10.00	0.79 ± 0.02	0.063 ± 0.009	16.741 ± 4.430	2.664 ± 0.824	2.663 ± 0.825
DCFR, λ_d :15.00	0.80 ± 0.02	0.055 ± 0.008	15.813 ± 2.478	2.444 ± 0.415	2.442 ± 0.416
DCFR, λ_d :20.00	0.80 ± 0.02	0.051 ± 0.007	15.213 ± 4.191	2.436 ± 0.709	2.434 ± 0.708
FairBiT, λ_b :0.00	0.85 ± 0.02	0.081 ± 0.012	36.468 ± 5.642	6.232 ± 1.079	6.238 ± 1.080
FairBiT, λ_b :0.00	0.85 ± 0.02	0.086 ± 0.011	38.809 ± 3.924	6.630 ± 0.697	6.635 ± 0.699
FairBiT, λ_b :0.00	0.85 ± 0.02	0.086 ± 0.011	38.823 ± 3.954	6.632 ± 0.701	6.638 ± 0.703
FairBiT, λ_b :0.00	0.85 ± 0.02	0.085 ± 0.011	38.641 ± 3.955	6.600 ± 0.701	6.606 ± 0.704
FairBiT, λ_b :0.02	0.85 ± 0.02	0.084 ± 0.011	38.058 ± 3.941	6.498 ± 0.698	6.503 ± 0.700
FairBiT, λ_b :0.06	0.85 ± 0.02	0.079 ± 0.010	36.089 ± 3.867	6.152 ± 0.682	6.158 ± 0.685
FairBiT, λ_b :0.22	0.85 ± 0.02	0.067 ± 0.008	30.326 ± 3.605	5.141 ± 0.638	5.146 ± 0.640
FairBiT, λ_b :0.77	0.85 ± 0.02	0.044 ± 0.006	19.536 ± 3.026	3.262 ± 0.550	3.267 ± 0.551
FairBiT, λ_b :2.78	0.84 ± 0.02	0.013 ± 0.002	8.458 ± 1.443	1.421 ± 0.286	1.423 ± 0.286
FairBiT, λ_b :10.00	0.81 ± 0.02	0.002 ± 0.000	2.220 ± 0.420	0.363 ± 0.086	0.364 ± 0.086
FairLearn (CDP)	0.65 ± 0.02	0.144 ± 0.030	27.817 ± 9.400	4.239 ± 1.449	4.246 ± 1.458
FairLearn (DP)	0.65 ± 0.02	0.144 ± 0.030	27.817 ± 9.400	4.239 ± 1.449	4.246 ± 1.458
Wass. Regularization, λ_w :0.00	0.85 ± 0.02	0.081 ± 0.012	36.200 ± 6.182	6.184 ± 1.170	6.191 ± 1.172
Wass. Regularization, λ_w :0.00	0.85 ± 0.02	0.086 ± 0.011	38.842 ± 3.981	6.635 ± 0.702	6.641 ± 0.704
Wass. Regularization, λ_w :0.00	0.85 ± 0.02	0.086 ± 0.011	38.772 ± 3.958	6.623 ± 0.701	6.628 ± 0.704
Wass. Regularization, λ_w :0.00	0.85 ± 0.02	0.085 ± 0.011	38.394 ± 3.970	6.556 ± 0.705	6.562 ± 0.707
Wass. Regularization, λ_w :0.02	0.85 ± 0.02	0.082 ± 0.011	37.124 ± 4.007	6.329 ± 0.707	6.335 ± 0.709
Wass. Regularization, λ_w :0.06	0.85 ± 0.02	0.073 ± 0.010	32.556 ± 4.114	5.514 ± 0.712	5.520 ± 0.713
Wass. Regularization, λ_w :0.22	0.84 ± 0.02	0.050 ± 0.007	17.982 ± 3.748	2.882 ± 0.673	2.887 ± 0.674
Wass. Regularization, λ_w :0.77	0.81 ± 0.02	0.040 ± 0.005	14.006 ± 2.199	2.139 ± 0.447	2.142 ± 0.449
Wass. Regularization, λ_w :2.78	0.80 ± 0.02	0.038 ± 0.005	14.166 ± 1.957	2.147 ± 0.337	2.149 ± 0.339
Wass. Regularization, λ_w :10.00	0.79 ± 0.02	0.036 ± 0.006	13.862 ± 2.145	2.074 ± 0.349	2.076 ± 0.350
Legit Only	0.60 ± 0.02	0.001 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
No Regularization	0.85 ± 0.02	0.082 ± 0.009	4.128 ± 0.391	6.343 ± 0.664	6.349 ± 0.666
Fair PreProc., λ_r :0.00	0.85 ± 0.02	0.084 ± 0.010	4.292 ± 0.375	6.597 ± 0.639	6.603 ± 0.641
Fair PreProc., λ_r :0.10	0.85 ± 0.02	0.084 ± 0.010	4.291 ± 0.375	6.586 ± 0.646	6.592 ± 0.649
Fair PreProc., λ_r :0.20	0.85 ± 0.02	0.084 ± 0.010	4.293 ± 0.395	6.599 ± 0.702	6.605 ± 0.704
Fair PreProc., λ_r :0.30	0.85 ± 0.02	0.083 ± 0.009	4.223 ± 0.326	6.483 ± 0.626	6.489 ± 0.628
Fair PreProc., λ_r :0.40	0.85 ± 0.02	0.083 ± 0.009	4.225 ± 0.337	6.488 ± 0.638	6.494 ± 0.640
Fair PreProc., λ_r :0.50	0.85 ± 0.02	0.082 ± 0.009	4.206 ± 0.341	6.459 ± 0.658	6.465 ± 0.660
Fair PreProc., λ_r :0.60	0.85 ± 0.02	0.084 ± 0.009	4.311 ± 0.344	6.623 ± 0.624	6.629 ± 0.624
Fair PreProc., λ_r :0.70	0.85 ± 0.02	0.085 ± 0.009	4.351 ± 0.366	6.701 ± 0.715	6.707 ± 0.714
Fair PreProc., λ_r :0.80	0.85 ± 0.02	0.084 ± 0.009	4.301 ± 0.327	6.624 ± 0.696	6.630 ± 0.695
Fair PreProc., λ_r :0.90	0.85 ± 0.02	0.084 ± 0.009	4.278 ± 0.353	6.586 ± 0.750	6.592 ± 0.748
Fair PreProc., λ_r :1.00	0.85 ± 0.02	0.084 ± 0.009	4.285 ± 0.342	6.597 ± 0.732	6.603 ± 0.731

Table 8: Mean and one standard deviation results for all methods-hyperparameter combinations for all methods except *FairLeap*, for the `Drug` dataset across 10 separate runs.

Method	AUC	$\mathsf{CDD}^{\mathsf{wass}}$ (normalized)	CDD^{ℓ_p} (uniform)	CDD^{ℓ_p} (Ave. $\mathbb{P}(L\|A)$)	CDD^{ℓ_p} ($\mathbb{P}(L)$)
FairLeap (uniform), λ_w :0.00	0.85 $\pm$ 0.02	0.071 $\pm$ 0.009	31.246 $\pm$ 3.632	5.289 $\pm$ 0.633	5.295 $\pm$ 0.635
FairLeap (uniform), λ_w :0.00	0.84 $\pm$ 0.02	0.049 $\pm$ 0.006	16.752 $\pm$ 3.309	2.693 $\pm$ 0.615	2.698 $\pm$ 0.616
FairLeap (uniform), λ_w :0.00	0.82 $\pm$ 0.02	0.032 $\pm$ 0.005	10.529 $\pm$ 2.412	1.667 $\pm$ 0.448	1.669 $\pm$ 0.449
FairLeap (uniform), λ_w :0.01	0.68 $\pm$ 0.04	0.003 $\pm$ 0.003	1.956 $\pm$ 1.463	0.353 $\pm$ 0.278	0.353 $\pm$ 0.277
FairLeap (uniform), λ_w :0.05	0.61 $\pm$ 0.03	0.001 $\pm$ 0.000	0.117 $\pm$ 0.050	0.021 $\pm$ 0.010	0.021 $\pm$ 0.010
FairLeap (uniform), λ_w :0.22	0.60 $\pm$ 0.04	0.001 $\pm$ 0.000	0.136 $\pm$ 0.088	0.026 $\pm$ 0.017	0.026 $\pm$ 0.017
FairLeap (uniform), λ_w :1.00	0.62 $\pm$ 0.03	0.001 $\pm$ 0.000	0.095 $\pm$ 0.045	0.018 $\pm$ 0.009	0.018 $\pm$ 0.009
FairLeap (uniform), λ_w :4.64	0.60 $\pm$ 0.02	0.001 $\pm$ 0.000	0.051 $\pm$ 0.017	0.010 $\pm$ 0.004	0.010 $\pm$ 0.004
FairLeap (uniform), λ_w :21.54	0.60 $\pm$ 0.05	0.001 $\pm$ 0.000	0.059 $\pm$ 0.023	0.011 $\pm$ 0.004	0.011 $\pm$ 0.004
FairLeap (uniform), λ_w :100.00	0.61 $\pm$ 0.04	0.000 $\pm$ 0.000	0.037 $\pm$ 0.013	0.007 $\pm$ 0.002	0.007 $\pm$ 0.002
FairLeap (P(L)), λ_w :0.00	0.85 $\pm$ 0.02	0.080 $\pm$ 0.009	36.136 $\pm$ 3.628	6.157 $\pm$ 0.651	6.163 $\pm$ 0.655
FairLeap (P(L)), λ_w :0.00	0.85 $\pm$ 0.02	0.076 $\pm$ 0.010	33.516 $\pm$ 4.020	5.689 $\pm$ 0.704	5.695 $\pm$ 0.706
FairLeap (P(L)), λ_w :0.00	0.84 $\pm$ 0.02	0.052 $\pm$ 0.007	18.332 $\pm$ 3.602	2.962 $\pm$ 0.651	2.967 $\pm$ 0.652
FairLeap (P(L)), λ_w :0.01	0.83 $\pm$ 0.02	0.037 $\pm$ 0.005	12.080 $\pm$ 2.786	1.887 $\pm$ 0.506	1.890 $\pm$ 0.508
FairLeap (P(L)), λ_w :0.05	0.77 $\pm$ 0.03	0.014 $\pm$ 0.005	6.028 $\pm$ 1.694	0.977 $\pm$ 0.265	0.979 $\pm$ 0.266
FairLeap (P(L)), λ_w :0.22	0.62 $\pm$ 0.03	0.001 $\pm$ 0.000	0.340 $\pm$ 0.396	0.059 $\pm$ 0.072	0.059 $\pm$ 0.072
FairLeap (P(L)), λ_w :1.00	0.60 $\pm$ 0.03	0.001 $\pm$ 0.000	0.097 $\pm$ 0.042	0.018 $\pm$ 0.008	0.018 $\pm$ 0.008
FairLeap (P(L)), λ_w :4.64	0.60 $\pm$ 0.04	0.001 $\pm$ 0.000	0.086 $\pm$ 0.043	0.016 $\pm$ 0.008	0.016 $\pm$ 0.008
FairLeap (P(L)), λ_w :21.54	0.60 $\pm$ 0.04	0.001 $\pm$ 0.000	0.072 $\pm$ 0.047	0.013 $\pm$ 0.009	0.013 $\pm$ 0.009
FairLeap (P(L)), λ_w :100.00	0.59 $\pm$ 0.04	0.001 $\pm$ 0.000	0.053 $\pm$ 0.019	0.010 $\pm$ 0.004	0.010 $\pm$ 0.004
Fairleap (Ave. P(L|A)), λ_w :0.00	0.85 $\pm$ 0.02	0.080 $\pm$ 0.009	36.337 $\pm$ 3.452	6.198 $\pm$ 0.652	6.204 $\pm$ 0.655
Fairleap (Ave. P(L|A)), λ_w :0.00	0.85 $\pm$ 0.02	0.076 $\pm$ 0.010	33.548 $\pm$ 4.008	5.695 $\pm$ 0.699	5.701 $\pm$ 0.701
Fairleap (Ave. P(L|A)), λ_w :0.00	0.84 $\pm$ 0.02	0.051 $\pm$ 0.007	18.292 $\pm$ 3.542	2.956 $\pm$ 0.639	2.961 $\pm$ 0.641
Fairleap (Ave. P(L|A)), λ_w :0.01	0.83 $\pm$ 0.02	0.037 $\pm$ 0.005	12.087 $\pm$ 2.736	1.886 $\pm$ 0.501	1.889 $\pm$ 0.503
Fairleap (Ave. P(L|A)), λ_w :0.05	0.77 $\pm$ 0.03	0.014 $\pm$ 0.005	6.013 $\pm$ 1.764	0.980 $\pm$ 0.279	0.982 $\pm$ 0.280
Fairleap (Ave. P(L|A)), λ_w :0.22	0.62 $\pm$ 0.05	0.001 $\pm$ 0.001	0.605 $\pm$ 1.024	0.102 $\pm$ 0.161	0.102 $\pm$ 0.162
Fairleap (Ave. P(L|A)), λ_w :1.00	0.61 $\pm$ 0.04	0.001 $\pm$ 0.000	0.242 $\pm$ 0.364	0.042 $\pm$ 0.061	0.042 $\pm$ 0.061
Fairleap (Ave. P(L|A)), λ_w :4.64	0.60 $\pm$ 0.04	0.001 $\pm$ 0.000	0.081 $\pm$ 0.044	0.015 $\pm$ 0.009	0.015 $\pm$ 0.009
Fairleap (Ave. P(L|A)), λ_w :21.54	0.59 $\pm$ 0.03	0.001 $\pm$ 0.000	0.063 $\pm$ 0.036	0.011 $\pm$ 0.007	0.011 $\pm$ 0.007
Fairleap (Ave. P(L|A)), λ_w :100	0.59 $\pm$ 0.03	0.001 $\pm$ 0.000	0.060 $\pm$ 0.028	0.011 $\pm$ 0.006	0.011 $\pm$ 0.006

Table 9: Mean and one standard deviation results for all methods-hyperparameter combinations for the 3 variants of *FairLeap*, for the `Drug` dataset across 10 separate runs.

Method	AUC	CDD$^{\text{wass}}$ (normalized)	CDD$^{\ell_p}$ (uniform)	CDD$^{\ell_p}$ (Ave. $\mathbb{P}(L\|A)$)	CDD$^{\ell_p}$ ($\mathbb{P}(L)$)
DCFR, λ_d :0.10	0.75 ± 0.03	0.075 ± 0.010	64.365 ± 6.454	6.989 ± 0.807	6.962 ± 0.793
DCFR, λ_d :0.25	0.75 ± 0.03	0.073 ± 0.010	58.616 ± 6.114	6.407 ± 0.734	6.382 ± 0.718
DCFR, λ_d :0.50	0.75 ± 0.03	0.069 ± 0.009	53.749 ± 9.718	5.871 ± 1.030	5.850 ± 1.021
DCFR, λ_d :0.75	0.75 ± 0.03	0.067 ± 0.008	50.864 ± 11.320	5.547 ± 1.268	5.522 ± 1.266
DCFR, λ_d :1.00	0.74 ± 0.03	0.063 ± 0.007	48.152 ± 6.224	5.273 ± 0.666	5.250 ± 0.660
DCFR, λ_d :1.50	0.74 ± 0.03	0.063 ± 0.007	46.919 ± 7.567	5.159 ± 0.802	5.141 ± 0.802
DCFR, λ_d :2.00	0.74 ± 0.03	0.065 ± 0.012	52.374 ± 14.938	5.676 ± 1.728	5.655 ± 1.720
DCFR, λ_d :5.00	0.73 ± 0.02	0.059 ± 0.009	47.078 ± 12.969	5.027 ± 1.437	5.018 ± 1.433
DCFR, λ_d :10.00	0.73 ± 0.02	0.064 ± 0.014	54.736 ± 19.358	5.863 ± 2.115	5.855 ± 2.112
DCFR, λ_d :15.00	0.73 ± 0.03	0.063 ± 0.007	55.557 ± 12.458	6.034 ± 1.416	6.025 ± 1.411
DCFR, λ_d :20.00	0.73 ± 0.03	0.062 ± 0.016	54.114 ± 18.042	5.838 ± 2.050	5.818 ± 2.036
FairBiT, λ_b :0.00	0.76 ± 0.02	0.046 ± 0.004	136.669 ± 24.976	14.949 ± 2.789	14.911 ± 2.781
FairBiT, λ_b :0.00	0.76 ± 0.01	0.046 ± 0.003	138.824 ± 18.970	15.171 ± 2.138	15.133 ± 2.129
FairBiT, λ_b :0.00	0.76 ± 0.02	0.046 ± 0.003	138.424 ± 18.793	15.126 ± 2.110	15.088 ± 2.102
FairBiT, λ_b :0.00	0.76 ± 0.02	0.046 ± 0.003	136.560 ± 19.170	14.919 ± 2.125	14.881 ± 2.115
FairBiT, λ_b :0.02	0.76 ± 0.02	0.045 ± 0.003	128.668 ± 18.087	14.048 ± 1.988	14.011 ± 1.977
FairBiT, λ_b :0.06	0.76 ± 0.02	0.043 ± 0.002	107.480 ± 14.507	11.709 ± 1.587	11.676 ± 1.576
FairBiT, λ_b :0.22	0.76 ± 0.02	0.036 ± 0.002	65.246 ± 9.924	7.015 ± 1.088	6.991 ± 1.074
FairBiT, λ_b :0.77	0.77 ± 0.02	0.018 ± 0.002	33.994 ± 6.489	3.602 ± 0.706	3.596 ± 0.705
FairBiT, λ_b :2.78	0.79 ± 0.01	0.004 ± 0.000	49.416 ± 11.565	5.359 ± 1.245	5.357 ± 1.244
FairBiT, λ_b :10.00	0.80 ± 0.01	0.001 ± 0.000	50.967 ± 8.455	5.547 ± 0.934	5.543 ± 0.932
FairLearn (CDP)	0.76 ± 0.03	0.049 ± 0.010	143.814 ± 75.740	15.553 ± 8.365	15.520 ± 8.354
FairLearn (DP)	0.76 ± 0.03	0.049 ± 0.010	143.814 ± 75.740	15.553 ± 8.365	15.520 ± 8.354
Wass. Regularization, λ_w :0.00	0.76 ± 0.02	0.047 ± 0.002	140.926 ± 13.829	15.402 ± 1.521	15.364 ± 1.510
Wass. Regularization, λ_w :0.00	0.76 ± 0.02	0.046 ± 0.002	139.740 ± 12.253	15.272 ± 1.337	15.234 ± 1.330
Wass. Regularization, λ_w :0.00	0.76 ± 0.02	0.046 ± 0.002	137.978 ± 12.025	15.082 ± 1.406	15.044 ± 1.407
Wass. Regularization, λ_w :0.00	0.76 ± 0.02	0.045 ± 0.002	126.986 ± 12.668	13.864 ± 1.494	13.828 ± 1.496
Wass. Regularization, λ_w :0.02	0.76 ± 0.02	0.043 ± 0.001	99.530 ± 8.605	10.823 ± 1.009	10.791 ± 1.005
Wass. Regularization, λ_w :0.06	0.76 ± 0.02	0.039 ± 0.001	62.866 ± 7.935	6.689 ± 0.890	6.668 ± 0.879
Wass. Regularization, λ_w :0.22	0.76 ± 0.01	0.035 ± 0.001	64.476 ± 12.326	6.867 ± 1.375	6.858 ± 1.372
Wass. Regularization, λ_w :0.77	0.77 ± 0.01	0.023 ± 0.002	63.646 ± 12.699	6.852 ± 1.337	6.850 ± 1.339
Wass. Regularization, λ_w :2.78	0.81 ± 0.01	0.001 ± 0.000	41.850 ± 5.725	4.549 ± 0.630	4.546 ± 0.628
Wass. Regularization, λ_w :10.00	0.86 ± 0.02	-0.000 ± 0.000	0.494 ± 0.479	0.054 ± 0.053	0.054 ± 0.053
Legit Only	0.81 ± 0.01	0.003 ± 0.001	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
No Regularization	0.76 ± 0.02	0.047 ± 0.002	14.133 ± 1.406	15.444 ± 1.500	15.406 ± 1.491
Fair PreProc., λ_r :0.00	0.76 ± 0.02	0.048 ± 0.003	14.364 ± 3.107	15.683 ± 3.610	15.644 ± 3.613
Fair PreProc., λ_r :0.10	0.76 ± 0.02	0.048 ± 0.003	14.243 ± 3.231	15.547 ± 3.732	15.509 ± 3.736
Fair PreProc., λ_r :0.20	0.76 ± 0.02	0.048 ± 0.003	14.264 ± 3.322	15.571 ± 3.842	15.532 ± 3.843
Fair PreProc., λ_r :0.30	0.76 ± 0.02	0.048 ± 0.004	14.734 ± 3.860	16.100 ± 4.436	16.059 ± 4.435
Fair PreProc., λ_r :0.40	0.76 ± 0.02	0.048 ± 0.004	14.175 ± 4.103	15.482 ± 4.707	15.441 ± 4.703
Fair PreProc., λ_r :0.50	0.76 ± 0.02	0.048 ± 0.005	14.632 ± 4.545	15.984 ± 5.201	15.943 ± 5.195
Fair PreProc., λ_r :0.60	0.76 ± 0.02	0.047 ± 0.005	13.237 ± 5.258	14.477 ± 5.981	14.435 ± 5.971
Fair PreProc., λ_r :0.70	0.76 ± 0.02	0.047 ± 0.006	13.397 ± 5.474	14.657 ± 6.223	14.614 ± 6.213
Fair PreProc., λ_r :0.80	0.76 ± 0.02	0.047 ± 0.007	13.573 ± 6.436	14.864 ± 7.262	14.821 ± 7.246
Fair PreProc., λ_r :0.90	0.76 ± 0.02	0.048 ± 0.008	13.771 ± 6.689	15.076 ± 7.556	15.031 ± 7.536
Fair PreProc., λ_r :1.00	0.76 ± 0.02	0.048 ± 0.007	13.636 ± 6.555	14.933 ± 7.408	14.888 ± 7.392

Table 10: Mean and one standard deviation results for all methods-hyperparameter combinations for all methods except *FairLeap*, for the `LawSchool` dataset across 10 separate runs.

Method	AUC	$\mathsf{CDD}^{\mathsf{wass}}$ (normalized)	CDD^{ℓ_p} (uniform)	CDD^{ℓ_p} (Ave. $\mathbb{P}(L\|A)$)	CDD^{ℓ_p} ($\mathbb{P}(L)$)
FairLeap (uniform), λ_w :0.00	0.76 ± 0.02	0.046 ± 0.002	132.110 ± 14.357	14.432 ± 1.624	14.395 ± 1.619
FairLeap (uniform), λ_w :0.00	0.76 ± 0.02	0.043 ± 0.002	113.947 ± 14.296	12.429 ± 1.589	12.395 ± 1.581
FairLeap (uniform), λ_w :0.00	0.76 ± 0.02	0.036 ± 0.002	68.021 ± 9.092	7.347 ± 1.048	7.323 ± 1.038
FairLeap (uniform), λ_w :0.01	0.78 ± 0.01	0.007 ± 0.001	14.041 ± 3.161	1.491 ± 0.346	1.486 ± 0.345
FairLeap (uniform), λ_w :0.05	0.80 ± 0.01	0.003 ± 0.000	1.980 ± 1.219	0.217 ± 0.134	0.216 ± 0.134
FairLeap (uniform), λ_w :0.22	0.80 ± 0.01	0.003 ± 0.000	1.332 ± 0.542	0.146 ± 0.059	0.145 ± 0.059
FairLeap (uniform), λ_w :1.00	0.80 ± 0.01	0.003 ± 0.000	0.664 ± 0.415	0.072 ± 0.045	0.072 ± 0.045
FairLeap (uniform), λ_w :4.64	0.80 ± 0.01	0.003 ± 0.000	0.404 ± 0.336	0.044 ± 0.037	0.044 ± 0.037
FairLeap (uniform), λ_w :21.54	0.80 ± 0.01	0.002 ± 0.001	0.171 ± 0.143	0.019 ± 0.016	0.019 ± 0.016
FairLeap (uniform), λ_w :100.00	0.81 ± 0.01	0.002 ± 0.001	0.137 ± 0.128	0.015 ± 0.014	0.015 ± 0.014
FairLeap (P(L)), λ_w :0.00	0.76 ± 0.02	0.047 ± 0.002	141.579 ± 17.141	15.474 ± 1.950	15.435 ± 1.944
FairLeap (P(L)), λ_w :0.00	0.76 ± 0.02	0.046 ± 0.002	139.786 ± 15.280	15.275 ± 1.747	15.237 ± 1.741
FairLeap (P(L)), λ_w :0.00	0.76 ± 0.02	0.045 ± 0.002	124.893 ± 14.577	13.638 ± 1.697	13.603 ± 1.696
FairLeap (P(L)), λ_w :0.01	0.76 ± 0.02	0.040 ± 0.002	89.241 ± 11.527	9.698 ± 1.320	9.669 ± 1.312
FairLeap (P(L)), λ_w :0.05	0.76 ± 0.01	0.026 ± 0.001	40.443 ± 6.568	4.320 ± 0.751	4.305 ± 0.743
FairLeap (P(L)), λ_w :0.22	0.80 ± 0.01	0.003 ± 0.000	1.684 ± 0.915	0.183 ± 0.099	0.182 ± 0.099
FairLeap (P(L)), λ_w :1.00	0.80 ± 0.01	0.003 ± 0.000	1.545 ± 0.469	0.168 ± 0.051	0.168 ± 0.051
FairLeap (P(L)), λ_w :4.64	0.80 ± 0.01	0.003 ± 0.000	0.829 ± 0.552	0.090 ± 0.060	0.090 ± 0.060
FairLeap (P(L)), λ_w :21.54	0.80 ± 0.01	0.003 ± 0.000	0.437 ± 0.351	0.047 ± 0.038	0.047 ± 0.038
FairLeap (P(L)), λ_w :100.00	0.80 ± 0.01	0.002 ± 0.000	0.210 ± 0.124	0.023 ± 0.013	0.023 ± 0.013
Fairleap (Ave. P(L|A)), λ_w :0.00	0.76 ± 0.01	0.047 ± 0.002	141.081 ± 10.070	15.495 ± 1.004	15.462 ± 1.003
Fairleap (Ave. P(L|A)), λ_w :0.00	0.76 ± 0.02	0.046 ± 0.002	139.381 ± 8.513	15.306 ± 0.910	15.274 ± 0.915
Fairleap (Ave. P(L|A)), λ_w :0.00	0.76 ± 0.01	0.045 ± 0.002	125.836 ± 9.474	13.811 ± 1.067	13.781 ± 1.074
Fairleap (Ave. P(L|A)), λ_w :0.01	0.76 ± 0.01	0.041 ± 0.002	91.037 ± 9.269	9.947 ± 1.031	9.923 ± 1.021
Fairleap (Ave. P(L|A)), λ_w :0.05	0.77 ± 0.01	0.026 ± 0.001	41.407 ± 7.574	4.440 ± 0.853	4.428 ± 0.844
Fairleap (Ave. P(L|A)), λ_w :0.22	0.80 ± 0.01	0.003 ± 0.001	2.453 ± 1.993	0.267 ± 0.220	0.267 ± 0.220
Fairleap (Ave. P(L|A)), λ_w :1.00	0.80 ± 0.01	0.003 ± 0.001	2.062 ± 0.795	0.225 ± 0.088	0.225 ± 0.088
Fairleap (Ave. P(L|A)), λ_w :4.64	0.80 ± 0.01	0.003 ± 0.000	0.829 ± 0.869	0.091 ± 0.096	0.091 ± 0.096
Fairleap (Ave. P(L|A)), λ_w :21.54	0.80 ± 0.01	0.003 ± 0.000	0.737 ± 0.591	0.081 ± 0.065	0.081 ± 0.065
Fairleap (Ave. P(L|A)), λ_w :100	0.80 ± 0.01	0.002 ± 0.000	0.303 ± 0.217	0.033 ± 0.024	0.033 ± 0.024

Table 11: Mean and one standard deviation results for all methods-hyperparameter combinations for the 3 variants of *FairLeap*, for the `LawSchool` dataset across 10 separate runs.

Method	AUC	$\mathsf{CDD}^{\mathsf{wass}}$ (normalized)	CDD^{ℓ_p} (uniform)	CDD^{ℓ_p} (Ave. $\mathbb{P}(L\|A)$)	CDD^{ℓ_p} ($\mathbb{P}(L)$)
DCFR, λ_d :0.10	0.03 ± 0.00	2.975 ± 1.093	33.472 ± 5.698	3.566 ± 0.715	3.548 ± 0.722
DCFR, λ_d :0.25	0.03 ± 0.00	2.963 ± 1.092	25.007 ± 4.511	2.662 ± 0.583	2.649 ± 0.588
DCFR, λ_d :0.50	0.03 ± 0.00	2.950 ± 1.092	19.614 ± 5.013	2.095 ± 0.626	2.085 ± 0.627
DCFR, λ_d :0.75	0.03 ± 0.00	2.947 ± 1.093	18.263 ± 4.990	1.939 ± 0.624	1.931 ± 0.623
DCFR, λ_d :1.00	0.03 ± 0.00	2.945 ± 1.093	18.464 ± 3.793	1.957 ± 0.487	1.950 ± 0.490
DCFR, λ_d :1.50	0.03 ± 0.00	2.946 ± 1.097	19.063 ± 6.640	2.023 ± 0.730	2.017 ± 0.744
DCFR, λ_d :2.00	0.03 ± 0.00	2.943 ± 1.095	18.377 ± 4.436	1.941 ± 0.509	1.934 ± 0.512
DCFR, λ_d :5.00	0.02 ± 0.00	2.957 ± 1.101	27.067 ± 15.656	2.894 ± 1.747	2.887 ± 1.748
DCFR, λ_d :10.00	0.02 ± 0.00	2.958 ± 1.101	28.300 ± 17.502	2.964 ± 1.833	2.959 ± 1.842
DCFR, λ_d :15.00	0.02 ± 0.00	2.966 ± 1.096	33.330 ± 14.067	3.542 ± 1.528	3.529 ± 1.528
DCFR, λ_d :20.00	0.02 ± 0.00	2.956 ± 1.084	27.457 ± 14.423	2.891 ± 1.592	2.876 ± 1.583
FairBiT, λ_b :0.00	0.02 ± 0.00	2.887 ± 0.971	35.226 ± 3.210	3.615 ± 0.350	3.614 ± 0.353
FairBiT, λ_b :0.00	0.02 ± 0.00	2.888 ± 0.971	35.491 ± 3.707	3.642 ± 0.399	3.641 ± 0.403
FairBiT, λ_b :0.00	0.02 ± 0.00	2.888 ± 0.971	35.475 ± 3.764	3.640 ± 0.407	3.639 ± 0.410
FairBiT, λ_b :0.00	0.02 ± 0.00	2.887 ± 0.971	35.171 ± 3.787	3.608 ± 0.410	3.607 ± 0.414
FairBiT, λ_b :0.02	0.02 ± 0.00	2.884 ± 0.971	34.018 ± 3.718	3.489 ± 0.403	3.488 ± 0.406
FairBiT, λ_b :0.06	0.02 ± 0.00	2.873 ± 0.971	30.206 ± 3.313	3.098 ± 0.360	3.097 ± 0.363
FairBiT, λ_b :0.22	0.02 ± 0.00	2.851 ± 0.970	20.869 ± 2.518	2.140 ± 0.273	2.139 ± 0.275
FairBiT, λ_b :0.77	0.03 ± 0.00	2.826 ± 0.970	8.504 ± 1.165	0.871 ± 0.125	0.871 ± 0.126
FairBiT, λ_b :2.78	0.04 ± 0.00	2.817 ± 0.970	2.175 ± 0.278	0.220 ± 0.030	0.220 ± 0.030
FairBiT, λ_b :10.00	0.05 ± 0.00	2.816 ± 0.970	0.838 ± 0.143	0.084 ± 0.015	0.084 ± 0.015
FairLearn (CDP)	0.03 ± 0.01	2.851 ± 0.969	16.307 ± 4.753	1.660 ± 0.489	1.661 ± 0.490
FairLearn (DP)	0.03 ± 0.01	2.851 ± 0.969	16.307 ± 4.753	1.660 ± 0.489	1.661 ± 0.490
Wass. Regularization, λ_w :0.00	0.02 ± 0.00	2.887 ± 0.971	35.248 ± 3.370	3.616 ± 0.375	3.615 ± 0.378
Wass. Regularization, λ_w :0.00	0.02 ± 0.00	2.888 ± 0.971	35.439 ± 3.633	3.636 ± 0.397	3.635 ± 0.401
Wass. Regularization, λ_w :0.00	0.02 ± 0.00	2.887 ± 0.971	35.322 ± 3.720	3.623 ± 0.406	3.622 ± 0.409
Wass. Regularization, λ_w :0.00	0.02 ± 0.00	2.886 ± 0.971	34.715 ± 3.736	3.561 ± 0.407	3.560 ± 0.410
Wass. Regularization, λ_w :0.02	0.02 ± 0.00	2.880 ± 0.971	32.396 ± 3.645	3.323 ± 0.397	3.322 ± 0.400
Wass. Regularization, λ_w :0.06	0.02 ± 0.00	2.863 ± 0.970	24.545 ± 3.189	2.518 ± 0.348	2.517 ± 0.351
Wass. Regularization, λ_w :0.22	0.03 ± 0.00	2.837 ± 0.968	11.789 ± 2.384	1.189 ± 0.245	1.192 ± 0.246
Wass. Regularization, λ_w :0.77	0.04 ± 0.00	2.826 ± 0.969	8.325 ± 1.393	0.841 ± 0.146	0.845 ± 0.147
Wass. Regularization, λ_w :2.78	0.05 ± 0.00	2.816 ± 0.970	2.453 ± 0.492	0.247 ± 0.051	0.248 ± 0.051
Wass. Regularization, λ_w :10.00	0.05 ± 0.00	2.815 ± 0.969	0.121 ± 0.066	0.012 ± 0.007	0.012 ± 0.007
Legit Only	0.04 ± 0.00	2.820 ± 0.970	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
No Regularization	0.02 ± 0.00	2.887 ± 0.971	3.537 ± 0.317	3.629 ± 0.353	3.628 ± 0.356
Fair PreProc., λ_r :0.00	0.02 ± 0.00	2.893 ± 0.971	3.992 ± 0.420	4.081 ± 0.444	4.081 ± 0.447
Fair PreProc., λ_r :0.10	0.02 ± 0.00	2.891 ± 0.971	3.955 ± 0.321	4.044 ± 0.345	4.045 ± 0.348
Fair PreProc., λ_r :0.20	0.02 ± 0.00	2.890 ± 0.970	3.909 ± 0.388	3.997 ± 0.408	3.998 ± 0.411
Fair PreProc., λ_r :0.30	0.02 ± 0.00	2.889 ± 0.970	3.914 ± 0.326	3.998 ± 0.348	3.999 ± 0.351
Fair PreProc., λ_r :0.40	0.02 ± 0.00	2.886 ± 0.970	3.780 ± 0.376	3.863 ± 0.386	3.864 ± 0.389
Fair PreProc., λ_r :0.50	0.02 ± 0.00	2.886 ± 0.969	3.868 ± 0.362	3.948 ± 0.381	3.950 ± 0.385
Fair PreProc., λ_r :0.60	0.02 ± 0.00	2.888 ± 0.970	3.931 ± 0.485	4.011 ± 0.493	4.013 ± 0.497
Fair PreProc., λ_r :0.70	0.02 ± 0.00	2.888 ± 0.970	3.924 ± 0.487	4.001 ± 0.499	4.004 ± 0.503
Fair PreProc., λ_r :0.80	0.02 ± 0.00	2.888 ± 0.967	3.962 ± 0.605	4.042 ± 0.610	4.045 ± 0.615
Fair PreProc., λ_r :0.90	0.02 ± 0.00	2.886 ± 0.968	3.868 ± 0.606	3.944 ± 0.613	3.947 ± 0.618
Fair PreProc., λ_r :1.00	0.02 ± 0.00	2.884 ± 0.968	3.693 ± 0.576	3.770 ± 0.577	3.773 ± 0.582

Table 12: Mean and one standard deviation results for all methods-hyperparameter combinations for all methods except *FairLeap*, for the `Communities and Crime` dataset across 10 separate runs.

Method	AUC	$\mathsf{CDD}^{\mathsf{wass}}$ (normalized)	CDD^{ℓ_p} (uniform)	CDD^{ℓ_p} (Ave. $\mathbb{P}(L\|A)$)	CDD^{ℓ_p} ($\mathbb{P}(L)$)
FairLeap (uniform), λ_w :0.00	0.02 ± 0.00	2.884 ± 0.971	33.601 ± 2.883	3.449 ± 0.314	3.447 ± 0.318
FairLeap (uniform), λ_w :0.00	0.02 ± 0.00	2.877 ± 0.971	29.820 ± 3.179	3.062 ± 0.349	3.060 ± 0.352
FairLeap (uniform), λ_w :0.00	0.02 ± 0.00	2.851 ± 0.970	16.821 ± 2.517	1.725 ± 0.269	1.724 ± 0.271
FairLeap (uniform), λ_w :0.01	0.04 ± 0.00	2.820 ± 0.970	1.133 ± 0.599	0.116 ± 0.063	0.116 ± 0.063
FairLeap (uniform), λ_w :0.05	0.04 ± 0.00	2.820 ± 0.971	0.058 ± 0.021	0.006 ± 0.002	0.006 ± 0.002
FairLeap (uniform), λ_w :0.22	0.04 ± 0.00	2.820 ± 0.971	0.027 ± 0.007	0.003 ± 0.001	0.003 ± 0.001
FairLeap (uniform), λ_w :1.00	0.04 ± 0.00	2.819 ± 0.970	0.016 ± 0.005	0.002 ± 0.001	0.002 ± 0.001
FairLeap (uniform), λ_w :4.64	0.04 ± 0.00	2.818 ± 0.970	0.013 ± 0.005	0.001 ± 0.000	0.001 ± 0.000
FairLeap (uniform), λ_w :21.54	0.05 ± 0.00	2.816 ± 0.970	0.010 ± 0.005	0.001 ± 0.001	0.001 ± 0.001
FairLeap (uniform), λ_w :100.00	0.05 ± 0.00	2.815 ± 0.970	0.006 ± 0.002	0.001 ± 0.000	0.001 ± 0.000
FairLeap (P(L)), λ_w :0.00	0.02 ± 0.00	2.886 ± 0.972	34.671 ± 3.001	3.560 ± 0.331	3.559 ± 0.334
FairLeap (P(L)), λ_w :0.00	0.02 ± 0.00	2.887 ± 0.971	34.868 ± 3.827	3.577 ± 0.414	3.576 ± 0.417
FairLeap (P(L)), λ_w :0.00	0.02 ± 0.00	2.882 ± 0.971	32.780 ± 3.571	3.364 ± 0.390	3.363 ± 0.393
FairLeap (P(L)), λ_w :0.01	0.02 ± 0.00	2.866 ± 0.970	24.682 ± 3.075	2.534 ± 0.334	2.533 ± 0.336
FairLeap (P(L)), λ_w :0.05	0.03 ± 0.00	2.831 ± 0.969	7.530 ± 1.620	0.770 ± 0.169	0.771 ± 0.170
FairLeap (P(L)), λ_w :0.22	0.04 ± 0.00	2.820 ± 0.971	0.133 ± 0.081	0.014 ± 0.008	0.014 ± 0.008
FairLeap (P(L)), λ_w :1.00	0.04 ± 0.00	2.820 ± 0.971	0.033 ± 0.008	0.003 ± 0.001	0.003 ± 0.001
FairLeap (P(L)), λ_w :4.64	0.04 ± 0.00	2.820 ± 0.971	0.017 ± 0.004	0.002 ± 0.000	0.002 ± 0.000
FairLeap (P(L)), λ_w :21.54	0.04 ± 0.00	2.819 ± 0.970	0.013 ± 0.005	0.001 ± 0.001	0.001 ± 0.001
FairLeap (P(L)), λ_w :100.00	0.05 ± 0.00	2.817 ± 0.970	0.009 ± 0.003	0.001 ± 0.000	0.001 ± 0.000
Fairleap (Ave. P(L|A)), λ_w :0.00	0.02 ± 0.00	2.887 ± 0.971	35.155 ± 3.198	3.607 ± 0.356	3.605 ± 0.358
Fairleap (Ave. P(L|A)), λ_w :0.00	0.02 ± 0.00	2.887 ± 0.971	35.033 ± 3.772	3.593 ± 0.408	3.592 ± 0.411
Fairleap (Ave. P(L|A)), λ_w :0.00	0.02 ± 0.00	2.882 ± 0.971	32.893 ± 3.562	3.375 ± 0.387	3.374 ± 0.390
Fairleap (Ave. P(L|A)), λ_w :0.01	0.02 ± 0.00	2.866 ± 0.970	24.730 ± 3.019	2.539 ± 0.328	2.538 ± 0.331
Fairleap (Ave. P(L|A)), λ_w :0.05	0.03 ± 0.00	2.831 ± 0.970	7.615 ± 1.553	0.778 ± 0.161	0.778 ± 0.162
Fairleap (Ave. P(L|A)), λ_w :0.22	0.04 ± 0.00	2.820 ± 0.971	0.126 ± 0.129	0.013 ± 0.013	0.013 ± 0.013
Fairleap (Ave. P(L|A)), λ_w :1.00	0.04 ± 0.00	2.820 ± 0.971	0.027 ± 0.006	0.003 ± 0.001	0.003 ± 0.001
Fairleap (Ave. P(L|A)), λ_w :4.64	0.04 ± 0.00	2.819 ± 0.970	0.018 ± 0.005	0.002 ± 0.001	0.002 ± 0.001
Fairleap (Ave. P(L|A)), λ_w :21.54	0.04 ± 0.00	2.819 ± 0.970	0.013 ± 0.004	0.001 ± 0.000	0.001 ± 0.000
Fairleap (Ave. P(L|A)), λ_w :100	0.05 ± 0.00	2.817 ± 0.970	0.010 ± 0.002	0.001 ± 0.000	0.001 ± 0.000

Table 13: Mean and one standard deviation results for all methods-hyperparameter combinations for the 3 variants of *FairLeap*, for the `Commuinites and Crime` dataset across 10 separate runs.