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 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. 2020). 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 for the purposes of model comparison and selection.

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.

*These authors contributed equally.

1.1 Contributions

We first introduce two new flexible measures of conditional demographic disparity (CDD): the *CDD in the Wasserstein sense* and the *CDD in the ℓ^p sense*. Both measures are well defined regardless of whether the legitimate feature(s) and/or outcome are discrete or continuous. To our knowledge, no such general definitions of CDD have been proposed before. We then propose and evaluate regularization-based methods 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 *bi-causal transport distance*, a distributional distance recently studied in the optimal transport literature. We call the resulting 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*, in which 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). Our methods work regardless of whether the model 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. 2020), 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. (2020), 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. 2022; 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) 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

[A table of notation used throughout the paper is given in Appendix A[1].]

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 $Y \in \mathcal{Y} \subseteq \mathbb{R}$ 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 or to some other type of quantifiable output, e.g. an automated

[1]Please see the arXiv version of the paper for all appendices.

decision such as a loan approval decision. Everything that follows applies in either setting.

Definition 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 (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 $l \in \operatorname{supp}\mathbb{P}(L|A = 0) \cap \operatorname{supp}\mathbb{P}(L|A = 1)$.*

Here, supp refers to the support of a distribution. 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 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.

A crucial step towards enforcing conditional demographic 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 produce 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 C for more details.

Definition 3 (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

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

Intuitively, the "∞" value for $D(l, l')$ when $l \neq l'$ ensures that we only compare model outputs *within* rather than *across* levels of the legitimate feature (e.g. within income levels in the example in Table 1). 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 4 (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 $\mathsf{CDD}^{\ell_p}(f) := \left(\sum_{l \in \mathcal{L}} \mathbb{Q}(l)(D(l))^p\right)^{1/p}$, where $\mathbb{Q}(L)$ is a probability measure defined over $\mathcal{L}$.

The disparity in Definition 4 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 and 4 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}. (See Appendix D for pseudocode for our actual algorithms.)

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* (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 $(\tilde{U}, \tilde{V}) \sim \tilde{\mathbb{P}}$ and $(U, V) \sim \mathbb{P}$ 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 5 (Bi-causal transport distance (BCD))**.** *For any fixed $C > 0$, 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*

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

See Appendix C 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 of the CDD. See Appendix F for the proof.

Theorem 4.1. *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 $\mathsf{CDD}^{\mathsf{wass}}(f)$ with $\mathsf{d}(\cdot, \cdot) = \mathcal{W}_p^p(\cdot, \cdot)$.*

In other words, if we set the constant C in Definition 5 large enough, then the bi-causal distance is exactly equal to the disparity given in Definition 3. This enables us to use methods previously developed to approximate the BCD to approximate and minimize $\mathsf{CDD}^{\mathsf{wass}}(f)$. 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 Theorem 4.1. The bi-causal distance can be viewed as a nested Wasserstein distance (Proposition 2 in Appendix C), one that takes a different form from the nested Wasserstein distance expressed in Definition 3 but which is equivalent under the conditions given in Theorem 4.1. 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. We 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 D 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 $\mathsf{CDD}^{\ell_p}(f_\theta) = \left(\sum_{l \in \mathcal{L}} \mathbb{Q}(l) (D_l)^p \right)^{1/p}$, as described in Definition 4. 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 highlight in particular 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, while avoiding 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). Similarly 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 D 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. (2020). 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 1 shows that R_{DCFR} aims to minimize a quantity that does not uniformly bound the CDD (in the sense of either Definition 3 or 4). This is contrast with our proposed methods, which aim to directly minimize the CDD.

Proposition 1. *The DCFR regularizer 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 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, $L = \emptyset$, meaning there is no legitimate feature, so $\mathsf{CDD}^{\mathsf{wass}} = \mathsf{CDD}^{\ell_p}$, 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 E.

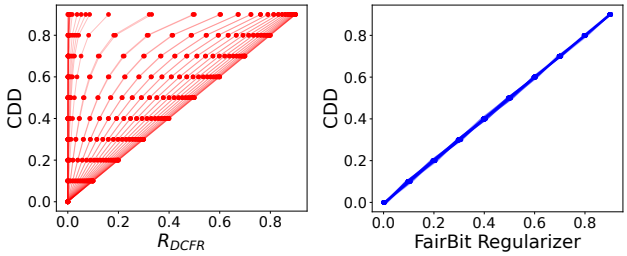

Figure 1: Conditional demographic disparity (CDD) versus R_{DCFR} (left) and the value of FairBiT regularizer (right) in a synthetic loan setting, varying the proportion of males vs. females and the loan acceptance rates. The proportion of males vs. females controls the slope of the CDD-R_{DCFR} curve, 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 two other fairness criteria: equalized odds and equality of opportunity). In the case of DP, the model $f(X)$ is trained to predict the target Y from the features X, while the adversary is trained to predict the sensitive feature A from the outcome of the model $f(X)$. In addition to using their original method, 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 1. *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 1, $f(X)$ 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.

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 modified not to require the sensitive feature at inference time, though this will in general result in a decrease in model performance.

5.3 Employing DP Methods for CDP Without Modification

For a more extensive empirical comparison, we consider two additional existing methods designed for demographic parity, one of which may in some circumstances promote CDP and one of which is computationally related to our methods.

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 1 that $f(\tilde{X})$ is approximately independent of A conditional on $\tilde{L}$. Note that $f(\tilde{X}) \perp\!\!\!\perp A \mid \tilde{L} \implies f(\tilde{X}) \perp\!\!\!\perp A \mid L$, 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 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 tasks. For CDD^{ℓ_p}, we report results for $\mathbb{Q}(L) = \mathbb{U}(L)$, the uniform distribution over L. (We include results for 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 G.3). We report the mean over 10 runs for every method-hyperparameter combination, omitting the hyperparameter labels in the figures for clarity.

In sum, we compare the following methods, including the methods described above as well as two baseline methods:

- *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) for DP and for CDP,
- *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). Appendix G contains more comprehensive results, including means, standard deviations, and discussion and evaluations of demographic disparity (the violation of demographic parity, i.e. Definition 1).

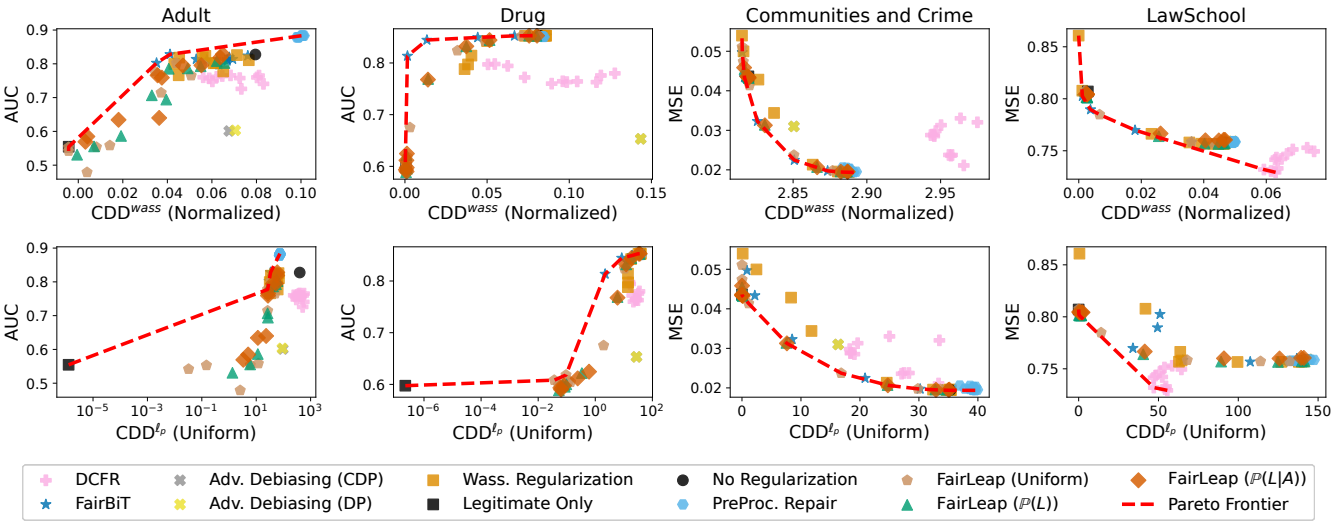

Figure 2: Fairness-predictive power (PP) trade-offs and Pareto frontiers for four datasets. *Top row:* Trade-offs when measuring fairness using CDD^{wass}. (We use a "normalized" version for presentation purposes. See Appendix G for details). *Bottom row:* Trade-offs when measuring fairness using CDD^{ℓ_p}, with $\mathbb{Q}(L) = \mathbb{U}(L)$ and $p = 1$. Predictive power (PP) is measured by AUC for classification tasks and MSE for regression tasks . 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 comparison methods.

6.2 Results

Figure 2 reports the fairness-predictive power (PP) trade-offs for the four datasets. We report conditional demographic disparity in the Wasserstein sense (CDD^{wass}) with $p = 2$ (Figure 2, top) as well as in the ℓ_p sense (CDD^{ℓ_p}) with $\mathbb{Q}(L) = \mathbb{U}(L)$ and $p = 1$, which we refer to as CDD^{ℓ_p} *(Uniform)* (Figure 2, bottom). Each point represents a method-hyperparameter combination; see Appendix G for 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^{wass} (Figure 2, top row), *FairBiT* has points on the Pareto frontier for every experiment and generally outperforms all other methods. Similarly, when CDD^{ℓ_p} (Uniform) 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^{wass} or CDD^{ℓ_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^{ℓ_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.

7 Conclusion

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. Experiments show that our methods generally provide better fairness-performance trade-offs than the existing state of the art method, DCFR, as well as methods designed to target (unconditional) demographic parity.

Acknowledgments

Luhao Zhang's work was done during an internship at JP Morgan AI research. The authors would like to thank Zikai Xiong for his invaluable feedback. 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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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*.

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

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. 2020. Algorithmic decision making with conditional fairness. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2125–2135.

Yang, M.; Sheng, J.; Liu, W.; Jin, B.; Wang, X.; and Wang, X. 2022. 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.