Fairness Through Matching

Anonymous authors
Paper under double-blind review

Abstract

Group fairness requires that different protected groups, characterized by a given sensitive attribute, receive equal outcomes overall. Typically, the level of group fairness is measured by the statistical gap between predictions from different protected groups. In this study, we reveal an implicit property of existing group fairness measures, which provides an insight into how the group-fair models behave. Then, we develop a new group-fair constraint based on this implicit property to learn group-fair models. To do so, we first introduce a notable theoretical observation: every group-fair model has an implicitly corresponding transport map between the input spaces of each protected group. Based on this observation, we introduce a new group fairness measure termed Matched Demographic Parity (MDP), which quantifies the averaged gap between predictions of two individuals (from different protected groups) matched by a given transport map. Then, we prove that any transport map can be used in MDP to learn group-fair models, and develop a novel algorithm called Fairness Through Matching (FTM), which learns a group-fair model using MDP constraint with an user-specified transport map. We specifically propose two favorable types of transport maps for MDP, based on the optimal transport theory, and discuss their advantages. Experiments reveal that FTM successfully trains group-fair models with certain desirable properties by choosing the transport map accordingly.

1 Introduction

Artificial Intelligence (AI) technologies based on machine learning algorithms have become increasingly prevalent as crucial decision-making tools across diverse areas, including credit scoring, criminal risk assessment, and college admissions. However, when observed data contains unfair biases, the resulting trained models may produce discriminatory decisions (Calders et al., 2009; Feldman et al., 2015; Angwin et al., 2016; Barocas & Selbst, 2016; Chouldechova, 2016; Kleinberg et al., 2018; Mehrabi et al., 2019; Zhou et al., 2021). For instance, several cases of unfair preferences favoring specific groups, such as white individuals or males, have been reported (Angwin et al., 2016; Ingold & Soper, 2016; Dua & Graff, 2017). To address these issues, there is a growing trend in non-discrimination laws that calls for the consideration of fair models (Hellman, 2019).

Under this social circumstance, ensuring algorithmic fairness in AI-based decision-making has become a crucial mission. Among several notions of algorithmic fairness, the notion of *group fairness* is the most explored one, which requires that certain statistics of each protected group should be similar. For example, the ratio of positive predictions should be similar across each protected group (Calders et al., 2009; Barocas & Selbst, 2016; Zafar et al., 2017; Donini et al., 2018; Agarwal et al., 2018).

Various algorithms have been proposed to learn models achieving group fairness. Existing methods for group fairness are roughly categorized into: pre-processing, in-processing and post-processing. Pre-processing approaches (Zemel et al., 2013; Feldman et al., 2015; Webster et al., 2018; Xu et al., 2018; Madras et al., 2018; Creager et al., 2019; Kim et al., 2022) aim to debias a given dataset, typically by learning fair representations whose distribution is independent of a given sensitive attribute. The debiased data (or fair representation) is then used to learn models. In-processing approaches (Kamishima et al., 2012; Goh et al., 2016; Zafar et al., 2017; Agarwal et al., 2018; Wu et al., 2019; Cotter et al., 2019; Celis et al., 2019; Zafar et al., 2019; Cho et al., 2020; Jiang et al., 2020a) train models by minimizing a given objective function under a specified group fairness constraint. Post-processing approaches (Kamiran et al., 2012; Hardt et al., 2016b; Fish et al.,

2016; Corbett-Davies et al., 2017; Pleiss et al., 2017; Chzhen et al., 2019; Wei et al., 2020; Jiang et al., 2020a) transform given prediction scores, typically provided by an unfair model, to satisfy a given group fairness level.

Most group-fair algorithms correspond to specific group fairness measures, typically defined by explicit quantities such as prediction scores and sensitive attributes. For example, demographic parity (Calders et al., 2009; Feldman et al., 2015; Angwin et al., 2016) considers the gap between two protected groups in terms of the positive prediction ratio. A shortcoming of such measures is that they only concern statistical disparities without accounting for implicit mechanisms about how a given model achieves group fairness. Thus, models can achieve high levels of group fairness in very undesirable ways (see Section C of Appendix for example). This observation serves as the motivation of this study.

In this paper, we first propose a new group fairness measure that reveals implicit behaviors of group-fair models. Based on the proposed measure, we develop an in-processing algorithm for learning group-fair models with controllable properties that cannot be controlled in existing fairness measures.

To do so, we begin by introducing a notable finding: every group-fair model implicitly corresponds to a transport map, which moves the measure of one protected group to another. Building on this observation, we propose a new measure for group fairness called Matched Demographic Parity (MDP), which quantifies the averaged gap between predictions of two individuals from different protected groups matched by a given transport map. We further prove that the reverse of this finding also holds, meaning that any transport map can be used in MDP to learn group-fair models. Finally, we develop an algorithm called Fairness Through Matching (FTM), designed to learn group-fair models under a fairness constraint based on MDP with a given transport map.

We propose two specific choices for the transport map: one is the optimal transport (OT) map in the input space, and the other is the OT map in the product space of input and output. Each is designed to achieve specific goals. For example, the former achieves higher fairness levels on subsets, while the latter yields better prediction performance and achieves higher levels of equalized odds compared to the former. Experiments on real benchmark datasets support our theoretical findings, showing that FTM successfully learns group-fair models, while the two proposed transport maps effectively achieve the intended purposes.

Main contributions

1. We introduce a notable observation that every group-fair model has an implicit transport map. Based on this finding, we present a novel measure of group fairness called *Matched Demographic Parity* (MDP).
2. We prove that any transport map can be used in MDP to learn group-fair models. Subsequently, we devise a novel algorithm called *Fairness Through Matching* (FTM), designed to find a group-fair model using a constraint based on MDP with a given transport map. We propose two favorable transport maps designed for specific purposes.
3. Experiments on benchmark datasets illustrate that FTM successfully learns group-fair models, and examine the effectiveness of the two transport maps in achieving their intended purposes.

2 Preliminaries

2.1 Notations & Problem setting

In this section, we outline the mathematical notations used throughout this paper. We focus on binary classification in this study. We denote $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ and $Y \in \mathcal{Y} = \{0,1\}$ as the d-dimensional input vector and the binary label, respectively. Assuming the pre-defined sensitive attribute is binary, we denote $S \in \{0,1\}$ as the binary sensitive attribute. For a given $s \in \{0,1\}$, the realization of S, we write $s' = 1-s$. For the probability distributions of these variables, let $\mathcal{P}$ and $\mathcal{P}_{\mathbf{X}}$ represent the joint distribution of $(\mathbf{X}, Y, S)$ and the marginal distribution of $\mathbf{X}$, respectively. Furthermore, let $\mathcal{P}_s = \mathcal{P}_{\mathbf{X}|S=s}, s \in \{0,1\}$ be the conditional distributions of $\mathbf{X}$ given $S = s$. We write $\mathbb{E}$ and $\mathbb{E}_s$ as the corresponding expectations of $\mathcal{P}$ and $\mathcal{P}_s$, respectively, and write $\mathcal{X}_s$

as the support of $\mathcal{P}_s$. For observed data, denote $\mathcal{D} = \{(\mathbf{x}_i, y_i, s_i)\}_{i=1}^n$ as the training dataset, consisting of n independent copies of the random tuple $(\mathbf{X}, Y, S) \sim \mathcal{P}$.

We denote the classification model as $f = f(\cdot, s), s \in \{0,1\}$, which is an estimator of $\mathcal{P}_s(Y = 1|\mathbf{X} = \cdot)$, belonging to a given hypothesis class $\mathcal{F} \subset \{f : \mathcal{X} \times \{0,1\} \to [0,1]\}$. For simplicity, we sometimes write $f(\cdot, s) = f_s(\cdot)$ wherever necessary. For given f and $s \in \{0,1\}$, we denote $\mathcal{P}_{f_s}$ the conditional distribution of $f(\boldsymbol{X}, s)$ given $S = s$. Furthermore, let C_{f_s} be the classification rule based on $f(\cdot, s)$, i.e., $C_{f_s}(\cdot) = \mathbb{I}(f(\cdot, s) \geq \tau)$ where τ is a specific threshold (typically, $\tau = 0.5$).

2.2 Measures for group fairness

In the context of group fairness, various measures have been introduced to quantify the gap between predictions of each protected group specified by a given sensitive attribute. The original measure for DP, $\Delta\mathrm{DP}(f) := |\mathcal{P}(C_{f_0}(\mathbf{X}) = 1|S = 0) - \mathcal{P}(C_{f_1}(\mathbf{X}) = 1|S = 1)|$, has been initially considered in various studies (Calders et al., 2009; Feldman et al., 2015; Donini et al., 2018; Agarwal et al., 2019; Zafar et al., 2019). Its relaxed version, $\Delta\overline{\mathrm{DP}}(f) := |\mathbb{E}(f_0(\mathbf{X}))|S = 0) - \mathbb{E}(f_1(\mathbf{X}))|S = 1)|$, has also been explored popularly (Madras et al., 2018; Chuang & Mroueh, 2021; Kim et al., 2022).

However, $\Delta\mathrm{DP}(f)$ has a limitation as it relies on a specific threshold τ (Silvia et al., 2020). To overcome this issue, the concept of strong DP (which requires similarity in the distributions of preditive values of each protected group) with several measures for quantifying the discrepancy between $\mathcal{P}_{f_0}$ and $\mathcal{P}_{f_1}$ have been considered (Jiang et al., 2020b; Chzhen et al., 2020; Silvia et al., 2020; Barata et al., 2021). $\Delta\mathrm{WDP}(f) := \mathcal{W}(\mathcal{P}_{f_0}, \mathcal{P}_{f_1})$, $\Delta\mathrm{TVDP}(f) := \mathrm{TV}(\mathcal{P}_{f_0}, \mathcal{P}_{f_1})$, and $\Delta\mathrm{KSDP}(f) := \mathrm{KS}(\mathcal{P}_{f_0}, \mathcal{P}_{f_1})$ are the examples of the measure for strong DP, where $\mathcal{W}, \mathrm{TV}$, and KS represent the Wasserstein distance, the Total Variation, and the Kolmogorov-Smirnov distance, respectively.

We denote $\Delta = \Delta(f)$ for all fairness measure Δ, provided the notation is clear for a given f.

2.3 Optimal transport

The concept of the Optimal Transport (OT) provides an approach for geometric comparison between two probability measures. For given two probability measures $\mathcal{Q}_1$ and $\mathcal{Q}_2$, a map $\mathbf{T}$ from $\mathrm{Supp}(\mathcal{Q}_1)$ to $\mathrm{Supp}(\mathcal{Q}_2)$ is called a *transport map* from $\mathcal{Q}_1$ to $\mathcal{Q}_2$ if the push-forward measure $\mathbf{T}_{\#}\mathcal{Q}_1$ is equal to $\mathcal{Q}_2$ (Villani, 2008). Here, the push-forward measure is defined by $\mathbf{T}_{\#}\mathcal{Q}_1(A) = \mathcal{Q}_1(\mathbf{T}^{-1}(A))$ for any measurable set A. The OT map is the optimal choice among all transport maps from a source distribution $\mathcal{Q}_1$ to a target distribution $\mathcal{Q}_2$. In this context, 'optimal' means minimizing transport cost, such as L_p distance in Euclidean space.

Monge (1781) originally formulates the OT problem: for given source and target distributions $\mathcal{Q}_1, \mathcal{Q}_2$ in $\mathbb{R}^d$ and a cost function c (e.g., L_2 distance), the OT map from $\mathcal{Q}_1$ to $\mathcal{Q}_2$ solves $\min_{\mathbf{T}:\mathbf{T}_{\#}\mathcal{Q}_1=\mathcal{Q}_2} \mathbb{E}_{\mathbf{X}\sim\mathcal{Q}_1}(c(\mathbf{X}, \mathbf{T}(\mathbf{X})))$, where $\mathbf{T}_{\#}\mathcal{Q}_1$ is the push-forward measure of $\mathcal{Q}_1$ induced by the map $\mathbf{T} : \mathrm{Supp}(\mathcal{Q}_1) \to \mathbb{R}^d$. The push-forward measure is defined by $\mathbf{T}_{\#}\mathcal{Q}_1(A) = \mathcal{Q}_1(\mathbf{T}^{-1}(A))$ for any measurable set A. If both $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are discrete with an identical number of support, a one-to-one mapping exists. For the case when $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are discrete but have different numbers of supports, Kantorovich relaxed the Monge problem by seeking the optimal coupling between two distributions (Kantorovich, 2006). The Kantorovich problem is formulated as $\inf_{\pi\in\Pi(\mathcal{Q}_1,\mathcal{Q}_2)} \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\pi}(c(\mathbf{X}, \mathbf{Y}))$ where $\Pi(\mathcal{Q}_1, \mathcal{Q}_2)$ is the set of all joint measures of $\mathcal{Q}_1$ and $\mathcal{Q}_2$. Note that the problem can be applied to the case of $\mathcal{Q}_1$ and $\mathcal{Q}_2$ with an identical number of support.

Various feasible estimators have been developed (Cuturi, 2013; Genevay et al., 2016), and applied to diverse tasks such as domain adaptation (Damodaran et al., 2018; Forrow et al., 2019) and computer vision (Su et al., 2015; Li et al., 2015; Salimans et al., 2018), to name a few.

2.4 Related works

Algorithmic fairness *Group fairness* is a fairness notion aimed at preventing discriminatory predictions for protected (demographic) groups divided by pre-defined sensitive attributes. Among various notions of group fairness, Demographic Parity (DP) (Calders et al., 2009; Feldman et al., 2015; Agarwal et al., 2019; Jiang et al., 2020b; Chzhen et al., 2020) quantifies the statistical gap in predictions between two different

protected groups. Other measures, including Equal opportunity (Eqopp) and Equalized Odds (EO), consider protected groups conditioned on both the label and the sensitive attribute (Hardt et al., 2016a). Various algorithms have been developed to learn group-fair model with respect to these group fairness notions (Zafar et al., 2017; Donini et al., 2018; Agarwal et al., 2018; Madras et al., 2018; Zafar et al., 2019; Chuang & Mroueh, 2021; Kim et al., 2022). See Section B of Appendix for an introduction to fairness notions other than group fairness.

Applications of the OT map to algorithmic fairness Several studies, including Gordaliza et al. (2019); Jiang et al. (2020b); Chzhen et al. (2020); Silvia et al. (2020); Buyl & Bie (2022), have employed the OT map for algorithmic fairness. Gordaliza et al. (2019) introduced a fair representation learning method that aligns inputs from different protected groups using the OT map. Jiang et al. (2020b); Chzhen et al. (2020); Silvia et al. (2020) proposed aligning prediction scores from different protected groups using the OT map or OT-based barycenter. Buyl & Bie (2022) developed a method that projects prediction scores onto a fair space by optimizing the projection through minimizing the transport cost calculated on all pairs of inputs.

Our proposed algorithm also utilizes the OT map, being the first to define a group fairness measure based on the transport map and to propose reasonable choices for the transport map.

3 Learning group-fair model through matching

The goal of this section is to explore and specify the correspondence between group-fair models and transport maps. In Section 3.1, we show that ***every group-fair model has a corresponding implicit transport map*** in the input space that matches two individuals from different protected groups. We then introduce a new fairness measure based on the correspondence. In Section 3.2, we show the reverse, i.e., ***any given transport map can be used to learn a group-fair model***, then present our proposed algorithm for learning group-fair models under a fairness constraint based on a given transport map.

3.1 Matched Demographic Parity (MDP)

Proposition 3.1 below shows that for a given perfectly group-fair model f (i.e., $\mathcal{P}_{f_0} = \mathcal{P}_{f_1}$ or equivalently $\Delta = 0$), there exists an implicit transport map in the input space that matches two individuals from different protected groups. Its proof is in Section A of Appendix.

Proposition 3.1 (Fair model ⇒ Transport map: perfect fairness case)**.** *For any perfectly group-fair model f, i.e., $\mathcal{P}_{f_0} = \mathcal{P}_{f_1}$, there exists a transport map $\mathbf{T}_s = \mathbf{T}_s(f)$ satisfying $f(\mathbf{X}, s) = f(\mathbf{T}_s(\mathbf{X}), s'), a.e.$*

The key implication of this proposition is that all perfectly group-fair models are not the same and the differences can be identified by the corresponding transport maps. We can also define a transport map for a not-perfectly group-fair model (i.e., $\Delta > 0$) similarly, by using a novel fairness measure termed **Matched Demographic Parity (MDP)**. Let $\mathcal{T}_s^{\text{trans}}$ be the set of all transport maps from $\mathcal{P}_s$ to $\mathcal{P}_{s'}$.

Definition 3.2 (Matched Demographic Parity)**.** *For a given model $f \in \mathcal{F}$ and a transport map $\mathbf{T}_s \in \mathcal{T}_s^{\text{trans}}$, the measure for MDP is defined as*

$$\Delta\text{MDP}(f, \mathbf{T}_s) := \mathbb{E}_s \left| f(\mathbf{X}, s) - f(\mathbf{T}_s(\mathbf{X}), s') \right|. \tag{1}$$

The idea behind MDP is that two individuals from different protected groups are matched by $\mathbf{T}_s$, and $\Delta\text{MDP}(f, \mathbf{T}_s)$ quantifies the similarity between the predictions of these two matched individuals. Subsequently, Theorem 3.3 below presents a relaxed version of Proposition 3.1, showing that any approximately group-fair model f has a transport map $\mathbf{T}_s$ such that $\Delta\text{MDP}(f, \mathbf{T}_s)$ is small. Suppose that (C1) $\mathcal{F}$ is the collection of bounded functions and (C2) $\mathcal{P}_s, s = 0, 1$, are absolutely continuous with respect to the Lebesgue measure. Refer to Section A of Appendix for the proof of Theorem 3.3.

Theorem 3.3 (Fair model ⇒ Transport map: relaxed fairness case)**.** *Fix a fairness level $\delta \geq 0$. Under* (C1) *and* (C2)*, for any given group-fair model f such that $\Delta\text{TVDP}(f) \leq \delta$, there exists a transport map $\mathbf{T}_s \in \mathcal{T}_s^{\text{trans}}$ satisfying $\Delta\text{MDP}(f, \mathbf{T}_s) \leq C\delta$ for some constant $C > 0$.*

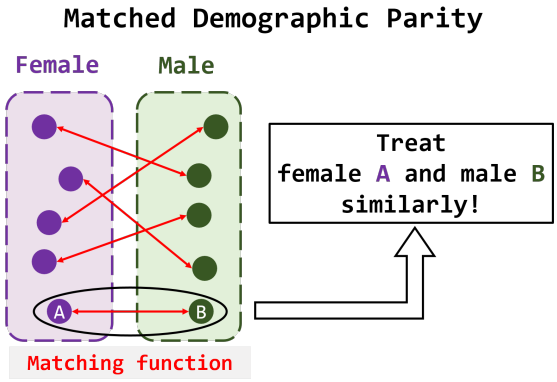

Figure 1: Simplified illustration of MDP. Once two individuals A and B are matched (↔), a model treats the pair of matched individuals A and B similarly (as well as other pairs). This implicit mechanism contributes to making the model fair.

We define the matching function of a given f as the transport map that attains the infimum of equation (1), as formulated in Definition 3.4 below. Through investigating the matching function, we can understand the mechanism behind how a group-fair model behaves. That is, the matching function specifically reveals which pairs of individuals from two protected groups are treated similarly by the group-fair model. See Figure 1 for the illustration of the matching function.

Definition 3.4 (Matching function of f)**.** *For a given f, denote $\mathbf{T}_s^f := \arg\min_{\mathbf{T}_s \in \mathcal{T}_s^{\text{trans}}} \Delta\text{MDP}(f, \mathbf{T}_s), s \in \{0,1\}$. For $\hat{s} := \arg\min_{s\in\{0,1\}} \Delta\text{MDP}(f, \mathbf{T}_s^f)$, the matching function of f is defined as $\mathbf{T}^f := \mathbf{T}_{\hat{s}}^f$.*

3.2 FTM: learning a group-fair model for a given transport map

The goal of this section is to formulate our proposed algorithm. Before introducing our proposed algorithm, we provide a theoretical support, which shows that a group-fair model can be constructed by MDP using any transport map.

Theorem 3.5 below, which is the reverse of Theorem 3.3, shows that any transport map in the input space can construct a group-fair model. That is, for a given transport map, if a model provides similar predictions for two individuals who are matched by the transport map, then it is group-fair. The proof is given in Section A of Appendix.

Theorem 3.5 (Transport map ⇒ Group-fair model)**.** *For a given $\mathbf{T}_s \in \mathcal{T}_s^{\text{trans}}$, if $\Delta\text{MDP}(f, \mathbf{T}_s) \le \delta$, then we have $\Delta\text{WDP}(f) \le \delta$ and $\Delta\overline{\text{DP}}(f) \le \delta$.*

Again, it is remarkable that a group-fair model and its corresponding transport map are closely related, i.e., ***every group-fair model has its corresponding implicit transport map, and vice versa***. Now, we are ready to devise an algorithm for learning group-fair models using MDP constraint. Based on Theorem 3.5, we develop a learning algorithm named **Fairness Through Matching (FTM)**, which learns a group-fair model subject to MDP being small with a given transport map.

Once a transport map $\mathbf{T}_s$ is given, FTM solves the following objective for a given loss function l (e.g., cross-entropy) and a pre-defined fairness level $\delta \ge 0$:

$$f^{\text{FTM}}(\mathbf{T}_s) := \arg\min_{f\in\mathcal{F}} \mathbb{E}l(Y, f(\mathbf{X}, S)) \text{ s.t. } \min_{s\in\{0,1\}} \Delta\text{MDP}(f, \mathbf{T}_s) \le \delta. \tag{2}$$

Unless there is any confusion, we write f^{FTM} instead of $f^{\text{FTM}}(\mathbf{T}_s)$ for simplicity. By Theorem 3.5, it is clear that f^{FTM} is fair (i.e., $\Delta\text{WDP}(f^{\text{FTM}}), \Delta\overline{\text{DP}}(f^{\text{FTM}}) \le \delta$) for any transport map $\mathbf{T}_s$. In practice, we estimate f^{FTM} with observed data $\mathcal{D}$ using mini-batch technique along with a stochastic gradient descent based algorithm (see Section 4 for details).

As any transport map can be used to build a group-fair model, it is possible to learn a group-fair model with a specific property by choosing an appropriate transport map. In the following section, we introduce and discuss two desirable options for the transport map with their advantages.

4 Choice of the transport map in FTM

We propose two favorable choices of the transport map $\mathbf{T}_s$ used in FTM. In Section 4.1, we suggest using the OT map in the input space $\mathcal{X}$, which minimizes the transport cost, resulting in a group-fair model with a higher fairness level on subsets. In Section 4.2, we propose using the OT map in the product space $\mathcal{X} \times \mathcal{Y}$, to improve prediction performance and the level of equalized odds, when compared to the OT map in the input space.

4.1 OT map on $\mathcal{X}$

First, we propose using the OT map for the transport map between the two input spaces, which is the minimizer of the transport cost on $\mathcal{X}$ among all transport maps. For a given $\mathbf{T}_s \in \mathcal{T}_s^{\text{trans}}$, the *transport cost* of $\mathbf{T}_s$ is defined by $\mathbb{E}_s\|\mathbf{X} - \mathbf{T}_s(\mathbf{X})\|^2$. From now on, we call this OT map on $\mathcal{X}$ as the ***marginal OT map***.

We explore a benefit of using the marginal OT map (i.e., low transport cost) by showing a theoretical relationship between the transport cost and fairness on subsets. Many undesirable behaviors of group-fair models have been recognized and discussed (Dwork et al., 2012; Kearns et al., 2018a; Wachter et al., 2020; Mougan et al., 2024). *Subset fairness*, which is a similar concept to subset targeting in Dwork et al. (2012), is one of such undesirable behaviors. We say that a group-fair model is subset-unfair if it is not group-fair against a certain subset (e.g., aged over 60s) even if it is group-fair overall. A mathematical definition of subset fairness can be done as follows.

Definition 4.1 (Subset fairness)**.** *Let A be a subset of $\mathcal{X}$. The level of subset fairness over A is defined as*

$$\Delta\overline{\text{DP}}_A(f) := |\mathbb{E}(f(\mathbf{X},0)|S=0,\mathbf{X}\in A) - \mathbb{E}(f(\mathbf{X},1)|S=1,\mathbf{X}\in A)|.$$

Intuitively, we expect that a group-fair model with a low transport cost would exhibit a high level of subset fairness. This is because the chance of two matched individuals (from different protected groups) belonging to the same subset A tends to be higher when the transport cost is smaller. Theorem 4.2 theoretically supports this conjecture, whose proof is provided in Section A of Appendix.

Theorem 4.2 (Low transport cost benefits subset fairness)**.** *Suppose $\mathcal{F}$ is the collection of L-Lipschitz functions. Let A be a given subset in $\mathcal{X}$. Then, for all f satisfying $\Delta\text{MDP}(f, \mathbf{T}_s^f) \le \delta$, we have*

$$\Delta\overline{\text{DP}}_A(f) \le L\left(\mathbb{E}_s\|\mathbf{X} - \mathbf{T}_s^f(\mathbf{X})\|^2\right)^{\frac{1}{2}} + \text{TV}(\mathcal{P}_{0,A}, \mathcal{P}_{1,A}) + U\delta, \tag{3}$$

where $\mathcal{P}_{s,A}$ is the distribution of $\mathbf{X}|S=s, \mathbf{X}\in A$, and $U>0$ is a constant only depending on A and $\mathcal{P}_s, s=0,1$.

The first term of RHS, $L\left(\mathbb{E}_s\|\mathbf{X} - \mathbf{T}_s^f(\mathbf{X})\|^2\right)^{1/2}$, implies that using a transport map with a small transport cost helps improve the level of subset fairness. The uncontrollable term $\text{TV}(\mathcal{P}_{0,A}, \mathcal{P}_{1,A})$ can be small for certain subsets. For example, for disjoint sets $A_1, \cdots, A_K$ of A, suppose that $\mathcal{P}_s$ is a mixture of uniform distribution given as $\mathcal{P}_s(\cdot) = \sum_{k=1}^K p_{sk}\mathbb{I}(\cdot \in A_k)$ with $p_{sk} \ge 0$ and $\sum_{k=1}^K p_{sk} = 1$ (e.g., the histogram). Then, $\text{TV}(\mathcal{P}_{0,A}, \mathcal{P}_{1,A})$ becomes zero for all $A_k, k \in [K]$. The last term $U\delta$ becomes small when δ is small.

We further present an example in Section C showing that a transport map with a high transport cost can lead to a problematic group-fair model. Moreover, Section 5.2.2 empirically shows that group-fair models learned by FTM with the marginal OT map attain higher fairness levels on various subsets that are not explicitly considered in the training phase, when compared to group-fair models learned by existing algorithms.

Practical algorithm for FTM with the marginal OT map To estimate the marginal OT map in practice, we sample two random mini-batches $\tilde{\mathcal{D}}_0 = \{\mathbf{x}_i^{(0)}\}_{i=1}^m \subset \mathcal{D}_0$ and $\tilde{\mathcal{D}}_1 = \{\mathbf{x}_j^{(1)}\}_{j=1}^m \subset \mathcal{D}_1$ with an identical size $m \le n$. For given two empirical distributions on $\tilde{\mathcal{D}}_0$ and $\tilde{\mathcal{D}}_1$, the cost matrix between the two is

defined by $\mathbf{C} \coloneqq [c_{i,j}] \in \mathbb{R}_+^{m \times m}$ where $c_{i,j} = \|\mathbf{x}_i^{(0)} - \mathbf{x}_j^{(1)}\|^2$. The optimal coupling is then defined by the matrix $\Gamma = [\gamma_{i,j}] \in \mathbb{R}_+^{m \times m}$, which solves the following objective:

$$\min_{\Gamma} \|\mathbf{C} \odot \Gamma\|_1 = \min_{\gamma_{i,j}} c_{i,j}\gamma_{i,j} \text{ s.t. } \sum_{i=1}^{m} \gamma_{i,j} = \sum_{j=1}^{m} \gamma_{i,j} = \frac{1}{m}, \gamma_{i,j} \geq 0. \tag{4}$$

Due to the equal sample sizes of $\tilde{\mathcal{D}}_s, s \in \{0,1\}$, the optimal coupling has only one non-zero (positive) entry for each row and column. Then, the marginal OT map for each $\mathbf{x}_i^{(0)} \in \tilde{\mathcal{D}}_0$ is defined by $\mathbf{T}_{0,\tilde{\mathcal{D}}}(\mathbf{x}_i^{(0)}) = \mathbf{x}_j^{(1)} \mathbb{1}(\gamma_{i,j} > 0)$ and $\mathbf{T}_{1,\tilde{\mathcal{D}}}$ is defined similarly.

Finally, we learn f with a stochastic gradient descent algorithm. For each update, to calculate the expected loss, we sample a random mini-batch $\mathcal{D}' \subset \mathcal{D}$ of size $n' \leq n$. Then, we update the solution using the gradient of the following objective function

$$\mathcal{L}(f) \coloneqq \frac{1}{n'} \sum_{(\mathbf{x}_i, y_i, s_i) \in \mathcal{D}'} l(y_i, f(\mathbf{x}_i, s_i)) + \lambda \frac{1}{m} \sum_{\mathbf{x}_i^{(s)} \in \tilde{\mathcal{D}}_s} \left| f(\mathbf{x}_i^{(s)}, s) - f(\mathbf{T}_{s,\tilde{\mathcal{D}}}(\mathbf{x}_i^{(s)}), s') \right| \tag{5}$$

for any $s \in \{0,1\}$, where $\lambda > 0$ is the Lagrange multiplier and $\mathbf{T}_{s,\tilde{\mathcal{D}}}$ is the marginal OT map from $\tilde{\mathcal{D}}_s$ to $\tilde{\mathcal{D}}_{s'}$.

4.2 OT map on $\mathcal{X} \times \mathcal{Y}$

One might argue that using the marginal OT map as the matching function could degrade the prediction performance much, since the matchings done by the marginal OT map do not consider the similarity in Y. As a remedy for this issue, we consider incorporating the label Y into the cost matrix calculation to avoid substantial degradation in prediction performance.

For this purpose, we define a new cost function on $\mathcal{X} \times \mathcal{Y}$. Let α be a given positive constant. The new cost function $c^\alpha : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to \mathbb{R}_+$ is defined by: $c^\alpha((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)) \coloneqq \|\mathbf{x}_1 - \mathbf{x}_2\|^2 + \alpha|y_1 - y_2|$, for given $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ and $y_1, y_2 \in \mathbb{R}$. Among all transport maps from the distribution of $\mathbf{X}, Y|S = s$ to the distribution of $\mathbf{X}, Y|S = s'$, we find the OT map that minimizes the transport cost (i.e., the expected value of c^α, where the expectation is taken over the distribution of $\mathbf{X}, Y|S = s$). Once this OT map on $\mathcal{X} \times \mathcal{Y}$ transports a given pair of input $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$, we focus only on the components corresponding to the input. For example, we select the first d values from the $d+1$ dimensional vector transported by the OT map. This map, which outputs only the components corresponding to the input, is called the ***joint OT map***. Note that the joint OT map can be used as a transport map for $\mathbf{X}$, where the transport cost is calculated based on both $\mathbf{X}$ and Y.

Clearly, using the new transport cost with a positive α can contribute to the improvement in prediction accuracy compared to the marginal OT map. This is because, while the marginal OT map does not care labels when matching individuals, the joint OT map tends to match individuals with the same label as much as possible when α is sufficiently large.

Not only the prediction accuracy, but also the level of equalized odds (i.e., demographic parities on the subsets consisting of those with $Y = 0$ and $Y = 1$, respectively) can be improved. This is because, FTM with the joint OT map tends to predict similarly for individuals with the same labels but from different protected groups, which directly aligns with the concept of equalized odds. For example, when $\Delta\text{MDP}(f, \mathbf{T}_s) = 0$ for the joint OT map $\mathbf{T}_s$ with a large α, equalized odds is perfectly achieved.

We empirically confirm this conjecture in Section 5.3, by showing that group-fair models learned by FTM with the joint OT map offer improved prediction accuracies as well as improved levels of equalized odds, when compared to FTM with the marginal OT map.

Practical algorithm for finding the joint OT map We apply a similar technique to the marginal OT map case, starting by sampling two random mini-batches $\tilde{\mathcal{D}}_0$ and $\tilde{\mathcal{D}}_1$ with an identical size $m \leq n$. Let $y_i^{(0)}$ and $y_j^{(1)}$ be the corresponding labels of $\mathbf{x}_i^{(0)} \in \tilde{\mathcal{D}}_0$ and $\mathbf{x}_j^{(1)} \in \tilde{\mathcal{D}}_1$, respectively. For a given $\alpha \geq 0$, we modify the cost matrix as follows: $\mathbf{C}^\alpha \coloneqq [c_{i,j}^\alpha] \in \mathbb{R}_+^{m \times m}$ where $c_{i,j}^\alpha = \|\mathbf{x}_i^{(0)} - \mathbf{x}_j^{(1)}\|^2 + \alpha|y_i^{(0)} - y_j^{(1)}|$. Note that when $\alpha = 0$, this problem becomes equivalent to the case of the marginal OT map in equation (4). We similarly calculate

the optimal coupling the matrix by solving the following objective:

$$\min_{\Gamma} \|\mathbf{C}^{\alpha} \odot \Gamma\|_1 = \min_{\gamma_{i,j}} c_{i,j}^{\alpha} \gamma_{i,j} \text{ s.t. } \sum_{i=1}^{m} \gamma_{i,j} = \sum_{j=1}^{m} \gamma_{i,j} = \frac{1}{m}, \gamma_{i,j} \geq 0. \tag{6}$$

Then, the joint OT map for each $\mathbf{x}_i^{(0)} \in \tilde{\mathcal{D}}_0$ is defined by $\mathbf{T}_{0,\tilde{\mathcal{D}}}(\mathbf{x}_i^{(0)}) = \mathbf{x}_j^{(1)} \mathbb{1}(\gamma_{i,j} > 0)$ and $\mathbf{T}_{1,\tilde{\mathcal{D}}}$ is defined similarly.

Finally, we learn f with a stochastic gradient descent algorithm. For each update, to calculate the expected loss, we sample a random mini-batch $\mathcal{D}' \subset \mathcal{D}$ of size $n' \leq n$. Then, we update the solution using the gradient of the following objective function

$$\mathcal{L}(f) \coloneqq \frac{1}{n'} \sum_{(\mathbf{x}_i, y_i, s_i) \in \mathcal{D}'} l(y_i, f(\mathbf{x}_i, s_i)) + \lambda \frac{1}{m} \sum_{\mathbf{x}_i^{(s)} \in \tilde{\mathcal{D}}_s} \left| f(\mathbf{x}_i^{(s)}, s) - f(\mathbf{T}_{s,\tilde{\mathcal{D}}}(\mathbf{x}_i^{(s)}), s') \right| \tag{7}$$

for any $s \in \{0, 1\}$, where $\lambda > 0$ is the Lagrange multiplier and $\mathbf{T}_{s,\tilde{\mathcal{D}}}$ is the joint OT map from $\tilde{\mathcal{D}}_s$ to $\tilde{\mathcal{D}}_{s'}$.

5 Experiments

This section presents our experimental results, showing that FTM with the proposed transport maps in Section 4 empirically work well to learn group-fair models. The key findings throughout this section are summarized as follows.

- FTM with the marginal OT map successfully learns group-fair models that exhibit (i) competitive prediction performance (Section 5.2.1) and (ii) higher levels of subset fairness (Section 5.2.2), when compared to other group-fair models learned by existing baseline algorithms. Beyond subset fairness, we further evaluate the self-fulfilling prophecy (Dwork et al., 2012) as an additional benefit of low transport cost (see Table 7 and 8 in Section E of Appendix).
- FTM with the joint OT map has the ability to learn group-fair models with improved prediction performance as well as improved levels of equalized odds, when compared to FTM with the marginal OT map (Section 5.3).

5.1 Settings

Datasets We use four real benchmark tabular datasets in our experiments: Adult (Dua & Graff, 2017), German (Dua & Graff, 2017), Dutch (Van der Laan, 2001), and Bank (Moro et al., 2014). The basic information about these datasets is provided in Table 1. We randomly partition each dataset into training and test datasets with the 8:2 ratio. This procedure is repeated 5 times, and we take the average of results on the test dataset.

Baseline algorithms and implementation details For the baseline algorithms, we consider three most popular state-of-the-art methods: Reduction (Agarwal et al., 2018), Reg (minimizing cross-entropy + $\lambda \Delta\overline{\text{DP}}^2$) (Donini et al., 2018; Chuang & Mroueh, 2021), and Adv (learning a model which cannot predict the sensitive attribute) (Zhang et al., 2018). Additionally, we consider the unfair baseline (abbr. Unfair), the ERM model trained without fairness regularization. For the measure of prediction performance, we use the classification accuracy (abbr. `Acc`). For fairness measures, we consider ΔDP, $\Delta\overline{\text{DP}}$ and ΔWDP, which are defined in Section 2.2.

For all algorithms, we employ MLP networks with ReLU activation and two hidden layers, where the hidden size is equal to the input dimension. We run all algorithms for 200 epochs and report their final performances on the test dataset. The Adam optimizer (Kingma & Ba, 2014) with the initial learning rate of 0.001 is used. To obtain the OT map, i.e, to solve the linear program, we utilize the POT library (Flamary et al., 2021). We utilize several Intel Xeon Silver 4410Y CPU cores and RTX 3090 GPU processors. More implementation details with `Pytorch`-style psuedo-code are provided in Section D.2 and D.3 of Appendix.

Table 1: The description of the tabular datasets: ADULT, GERMAN, BANK, and DUTCH. $\mathbf{X}$ denotes the input vector, S denotes the sensitive attribute, Y denotes the target label information, and d denotes the dimension of $\mathbf{X}$. Train/Test data sizes are the number of samples.

Dataset	Variable	Description	Dataset	Variable	Description
ADULT	$\mathbf{X}$	Personal attributes	GERMAN	$\mathbf{X}$	Personal attributes
	S	Gender		S	Gender
	Y	Outcome over \$50k		Y	High credit score
	d	101		d	60
	Train data size	30,136		Train data size	800
	Test data size	15,086		Test data size	200
BANK	$\mathbf{X}$	Personal attributes	DUTCH	$\mathbf{X}$	Personal attributes
	S	Binarized age		S	Gender
	Y	Subscribing a term deposit		Y	High-level occupation
	d	57		d	58
	Train data size	24,390		Train data size	48,336
	Test data size	6,098		Test data size	12,084

5.2 FTM with the marginal OT map

This section shows the success of FTM with the marginal OT map, in terms of (i) fairness-prediction trade-off and (ii) improvement in subset fairness.

5.2.1 Fairness-prediction trade-off

In this section, we empirically verify that learned models by FTM successfully achieves (strong) demographic parity. For the transport map used in MDP constraint, we choose the marginal OT map. Figure 2 below clearly shows that FTM successfully learns group-fair models for various fairness levels.

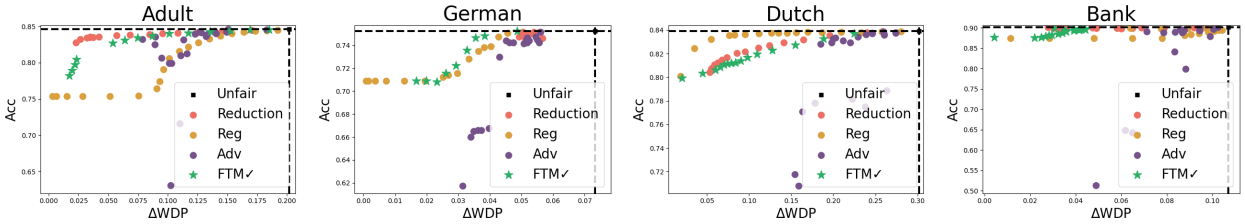

Figure 2: **Fairness-prediction trade-offs**: Plots of ΔWDP vs. Acc. (Left to right) ADULT, GERMAN, DUTCH, BANK. Refer to Figure 6 in Section E of Appendix for other (but not strong) fairness measures.

Another main implication is that using the marginal OT map does not hamper prediction performance much. Figure 2 supports this assertion in terms of fairness-prediction trade-off, that is, FTM is competitive with the three baselines. In most datasets, the performance of FTM is not significantly worse than that of the top-performing algorithm (i.e., Reduction). Notably, on GERMAN dataset, FTM performs the best, whereas Reduction fails to learn group-fair models with fairness level under 0.06. Additionally, FTM mostly outperforms the other two baseline algorithms, Reg and Adv. Hence, we can conclude that FTM is also a promising algorithm for strong group fairness.

5.2.2 Improvement in subset fairness

This section highlights the key advantages of using the marginal OT map in terms of subset fairness, which is theoretically supported by Theorem 4.2. We examine two scenarios for the subset A in Definition 4.1: (1) random subsets and (2) subsets defined by specific input variables.

Random subsets First, we generate a random subset $\mathcal{D}_{\text{sub}}$ of the test data defined as $\mathcal{D}_{\text{sub}} = \{i : \mathbf{v}^\top \mathbf{x}_i \geq 0\}$, using a random vector $\mathbf{v}$ drawn from the uniform distribution on $[-1, 1]^d$. Then, we calculate $\Delta\overline{\text{DP}}$ on $\mathcal{D}_{\text{sub}}$.

Figure 3 presents boxplots of the $\Delta\overline{\text{DP}}$ values calculated on 1,000 randomly generated $\mathcal{D}_{\text{sub}}$. Outliers in the boxplots (points in red boxes) represent the example instances of subset unfairness. For a fair comparison, we evaluate under a given $\Delta\overline{\text{DP}}$ for each dataset: 0.06 for Adult, 0.01 for German, 0.07 for Dutch, and 0.04 for Bank. Notably, FTM consistently has the fewest outliers than all the baselines, indicating that FTM achieves higher fairness on random subsets.

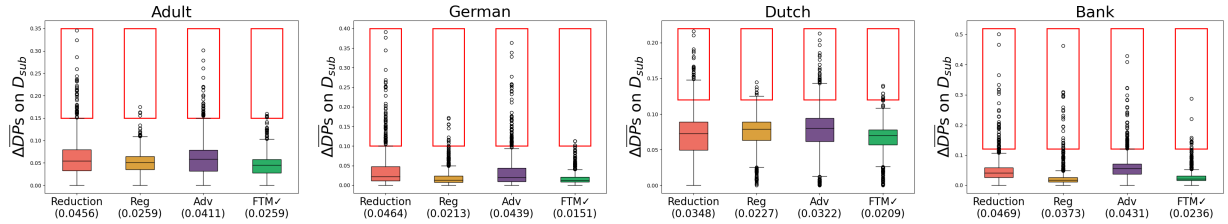

Figure 3: **Fairness on random subsets**: Boxplots of the levels of $\Delta\overline{\text{DP}}$ on 1,000 randomly generated subsets $\mathcal{D}_{\text{sub}}$ of test datasets. (Left to right) Adult, German, Dutch and Bank. The values presented under the algorithm name (e.g., 0.0151 for FTM in German) are the standard deviations.

Subsets defined by specific input variables Second, we focus on subsets defined by a specific input variable. For this scenario, we construct two subsets by binarizing a specific input variable using its median value. Note that we learn models considering only gender as the sensitive attribute.

Table 2 presents the fairness levels (with respect to gender) in the two subsets of the learned models. We consider German and Dutch datasets for this analysis. For German dataset, the two subsets are defined by {**x** of high age} and {**x** of low age}. For Dutch dataset, the two subsets are defined by {**x** who is married} and {**x** who is not married}. The results highlight the superiority of FTM in achieving higher fairness in these specific subsets.

Table 2: Fairness levels on the subsets defined by specific input variables: **Bold** faced ones highlight the best results, and underlined ones are the second best ones. Standard errors are reported in Tables 5 and 6 in Section E of Appendix. (Top) The subsets are defined by the input variable 'age' on German dataset under a given $\Delta\overline{\text{DP}} = 0.045$. (Bottom) The subsets are defined by the input variable 'marital status' on Dutch dataset under a given $\Delta\overline{\text{DP}} = 0.12$.

German					
Subset	Fairness measure	Reduction	Reg	Adv	FTM ✓
High age	ΔDP	0.073	0.077	0.048	**0.045**
	$\Delta\overline{\text{DP}}$	0.049	0.029	0.028	**0.026**
	ΔWDP	0.053	0.039	0.042	**0.038**
Low age	ΔDP	0.118	0.116	0.122	**0.077**
	$\Delta\overline{\text{DP}}$	**0.047**	0.050	0.053	**0.047**
	ΔWDP	0.058	0.059	0.061	**0.054**

Dutch					
Subset	Fairness measure	Reduction	Reg	Adv	FTM ✓
Married	ΔDP	0.258	0.372	0.237	**0.204**
	$\Delta\overline{\text{DP}}$	0.182	0.164	0.187	**0.152**
	ΔWDP	0.183	0.172	0.193	**0.152**
Not married	ΔDP	**0.061**	0.131	0.095	0.068
	$\Delta\overline{\text{DP}}$	0.045	0.062	0.098	**0.036**
	ΔWDP	**0.045**	0.072	0.098	**0.045**

5.3 FTM with the joint OT map

This section shows the effect of using the joint OT map, in terms of (i) prediction accuracy as well as level of equalized odds, and (ii) the transport cost. To do so, we compare the group-fair models learned by FTM with the two maps, the marginal OT map and the joint OT map, by using ADULT dataset for this analysis.

5.3.1 Improvement in prediction accuracy and equalized odds

Table 3 shows that FTM with the joint OT map can provide higher prediction performance in certain scenarios where more accurate group-fair models than FTM with the marginal OT map exist (e.g., $\Delta\text{DP} \leq 0.06$ in Figure 2). Furthermore, we observe that the level of equalized odds can be improved in this scenarios. To assess the level of equalized odds, we use a measure defined as $\Delta\text{EO} := \frac{1}{2}\sum_{y=0,1} |\mathbb{E}(f_0(\mathbf{X}))|S=0, Y=y) - \mathbb{E}(f_1(\mathbf{X}))|S=1, Y=y)|$.

Note that we fix $\alpha = 100$ for FTM with the joint OT map, because the results with $\alpha > 100$ are almost identical to those with $\alpha = 100$. In other words, 100 is the minimum value for α where $\alpha|y_i^{(0)} - y_j^{(1)}|$ fully dominates the transport cost $\|\mathbf{x}_i^{(0)} - \mathbf{x}_j^{(1)}\|^2$.

Table 3: Comparison between (i) the marginal OT map and (ii) the joint OT map in terms of prediction accuracy and level of equalized odds, with the two fixed fairness level ΔDPs at 0.033 and 0.054.

$\mathbf{T}_s$ in FTM	$\Delta\text{DP} = 0.033$		$\Delta\text{DP} = 0.054$	
	Acc (↑)	ΔEO (↓)	Acc (↑)	ΔEO (↓)
Marginal OT map	0.806	0.030	0.826	0.025
Joint OT map	0.810	0.026	0.830	0.017

5.3.2 Increase in transport cost

However, using the joint OT map can result in a higher transport cost compared to the marginal OT map. The left panel of Figure 4 illustrates that increasing α can improve prediction accuracy, though this improvement comes with a higher transport cost, especially when group-fair models more accurate than FTM with the marginal OT map exist. That is, at $\Delta\text{DP} = 0.025$, a point where a group-fair model more accurate than FTM with the marginal OT map exists (e.g., Reduction in Figure 2), both accuracy and transport cost increase, as α increases.

In contrast, the right panel of Figure 4 shows that increasing α does not significantly improve prediction accuracy while still incurring a higher transport cost, when FTM with the marginal OT map is competitive with other group-fair models in terms of accuracy. That is, at $\Delta\text{DP} = 0.073$, a point where the accuracy of FTM with the marginal OT map is similar to that of other group-fair models, increasing α does not yield notably beneficial results.

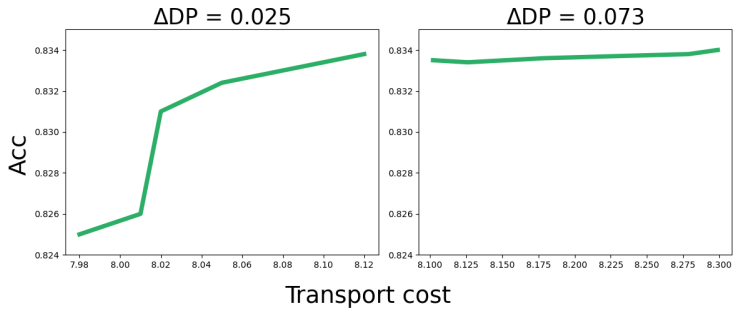

Figure 4: **Transport cost-prediction trade-offs**: Transport cost of the matching function vs. Acc of FTM with the joint OT map of $\alpha \in \{0, 1, 5, 10, 50, 100\}$ on ADULT dataset.

Overall, FTM with the joint OT map using an appropriately tuned α could be a desirable solution, especially when seeking for a group-fair model that is more accurate than a group-fair model learned by FTM with the marginal OT map. However, tuning α can be challenging, and compromising subset fairness would be generally not advisable. Additionally, as discussed in Section 5.2.1, using the marginal OT map is also competitive with other baselines in terms of prediction accuracy. Therefore, we basically recommend using the marginal OT map for FTM, while considering the joint OT map is particularly useful when the prediction accuracy of a group-fair model learned by FTM with the marginal OT map is suboptimal.

6 Conclusion and discussion

In this paper, we have discussed the existence of implicit transport maps for all group-fair models. Specifically, we have introduced a novel group fairness measure named MDP. Building upon MDP, we propose a novel algorithm, FTM, designed for learning group-fair models with high levels of subset fairness. Experimental results demonstrate that FTM with the marginal OT map effectively produces group-fair models with improved levels of subset fairness on various subsets compared to baseline models, while maintaining reasonable prediction performance. Moreover, we have proposed to use the joint OT map to improve the prediction accuracy and equalized odds of FTM.

We have only considered the two transport maps for FTM. There could be other transport maps which provide other types of group-fair models. It would be interesting to search for other useful transport maps, which we leave as a future work.

A key social benefit of the proposed methods is that we are able to train group-fair models with higher levels of subset fairness without the need to collect and process additional sensitive data. By doing so, the proposed algorithm is expected to transcend the fairness-privacy trade-off, making it practical for use without conflicting with data protection laws.

Broader Impact Statement

A broad goal of this study is to caution users of fair AI models, such as social planners and courts, against solely pursuing group fairness without accounting for the risks of potential discrimination, e.g., subset fairness. Additionally, it aims to equip them with a tool to enhance these aspects. Even though it is rather technical, we believe that our work provides a new perspective on algorithmic fairness and could possibly impact policy-making and regulation in related fields.

Another social impact of our study is that the relationship between the transport maps and group-fair models may help us form a new concept of fairness that can be easily accepted by society. Our approach explores the micro-level behavior of a given group-fair model (i.e., how the model matches individuals rather than simply looking at the statistics), which could enable finding reasonable compromises for seemingly paradoxical existing concepts of fairness.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 60–69. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/agarwal18a.html`.

Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, 2019. URL `http://proceedings.mlr.press/v97/agarwal19d.html`.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.

António Pereira Barata, Frank W. Takes, H. Jaap van den Herik, and Cor J. Veenman. Fair tree classifier using strong demographic parity, 2021.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018. URL https://arxiv.org/abs/1810.01943.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. Technical Report MSR-TR-2020-32, Microsoft, May 2020.

Maarten Buyl and Tijl De Bie. Optimal transport of classifiers to fairness. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-welFirjMss.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18. IEEE, 2009.

Jean-Paul Carvalho, Bary Pradelski, and Cole Williams. Affirmative action with multidimensional identities. *Available at SSRN 4070930*, 2022.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 319–328, 2019.

Jaewoong Cho, Changho Suh, and Gyeongjo Hwang. A fair classifier using kernel density estimation. In *34th Conference on Neural Information Processing Systems, NeurIPS 2020*. Conference on Neural Information Processing Systems, 2020.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2016. URL https://arxiv.org/abs/1610.07524.

Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=DNl5s5BXeBn.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. In *Advances in Neural Information Processing Systems*, pp. 12760–12770, 2019.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with wasserstein barycenters. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7321–7331. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/51cdbd2611e844ece5d80878eb770436-Paper.pdf.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797–806, 2017.

Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. Two-player games for efficient non-convex constrained optimization. In *ALT*, 2019.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, pp. 1436–1445. PMLR, 2019.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, pp. 467–483, Berlin, Heidelberg, 2018. Springer-Verlag. ISBN 978-3-030-01224-3. doi: 10.1007/978-3-030-01225-0_28. URL https://doi.org/10.1007/978-3-030-01225-0_28.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pp. 2791–2801, 2018.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pp. 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 144–152. SIAM, 2016.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. Statistical optimal transport via factored couplings. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2454–2465. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/forrow19a.html.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3440–3448, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, pp. 2415–2423, 2016.

Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2357–2365. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/gordaliza19a.html.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016a. URL https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016b.

Deborah Hellman. Measuring algorithmic fairness. *Criminal Procedure eJournal*, 2019. URL https://api.semanticscholar.org/CorpusID:199002104.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166 – 1194, 2021. doi: 10.1214/20-AOS1997. URL https://doi.org/10.1214/20-AOS1997.

David Ingold and Spencer Soper. Amazon doesn't consider the race of its customers. should it. *Bloomberg, April*, 1, 2016.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pp. 862–872. PMLR, 2020a.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In Ryan P. Adams and Vibhav Gogate (eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 862–872. PMLR, 22–25 Jul 2020b. URL https://proceedings.mlr.press/v115/jiang20a.html.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pp. 924–929. IEEE, 2012.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer, 2012.

Lev Kantorovich. On the translocation of masses. *Journal of Mathematical Sciences*, 133:1381–1382, 2006.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning, 2018a.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2564–2572. PMLR, 10–15 Jul 2018b. URL https://proceedings.mlr.press/v80/kearns18a.html.

Dongha Kim, Kunwoong Kim, Insung Kong, Ilsang Ohn, and Yongdai Kim. Learning fair representation with a parametric integral probability metric. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11074–11101. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/kim22b.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pp. 22–27, 2018.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1718–1727, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/li15.html.

David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In *ICML*, 2018.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

Anay Mehrotra, Bary S. R. Pradelski, and Nisheeth K. Vishnoi. Selection in the presence of implicit bias: The advantage of intersectional constraints. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 599–609, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533124. URL https://doi.org/10.1145/3531146.3533124.

Mathieu Molina and Patrick Loiseau. Bounding and approximating intersectional fairness through marginal fairness. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=LffWuGtC9BE.

Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2014.03.001. URL https://www.sciencedirect.com/science/article/pii/S016792361400061X.

Carlos Mougan, Antonio Ferrara, Laura State, Salvatore Ruggieri, and Steffen Staab. Beyond demographic parity: Redefining equal treatment, 2024. URL https://openreview.net/forum?id=cVea4KQ4xm.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689, 2017.

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*, 12(3), mar 2022. doi: 10.1002/widm.1452. URL https://doi.org/10.1002%2Fwidm.1452.

Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport, 2018.

Changjian Shui, Gezheng Xu, Qi CHEN, Jiaqi Li, Charles Ling, Tal Arbel, Boyu Wang, and Christian Gagné. On learning fairness and accuracy on multiple subgroups. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=YsRH6uVcx2l.

Chiappa Silvia, Jiang Ray, Stepleton Tom, Pacchiano Aldo, Jiang Heinrich, and Aslanides John. A general approach to fairness with optimal transport. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3633–3640, Apr. 2020. doi: 10.1609/aaai.v34i04.5771. URL https://ojs.aaai.org/index.php/AAAI/article/view/5771.

Joshua Simons, Sophia Adams Bhatti, and Adrian Weller. Machine learning and the meaning of equal treatment. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pp. 956–966, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462556. URL https://doi.org/10.1145/3461702.3462556.

Zhengyu Su, Yalin Wang, Rui Shi, Wei Zeng, Jian Sun, Feng Luo, and Xianfeng Gu. Optimal mass transport for shape matching and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37 (11):2246–2259, 2015. doi: 10.1109/TPAMI.2015.2408346.

Paul Van der Laan. *The 2001 Census in the Netherlands: Integration of Registers and Surveys*, pp. 39–52. 12 2001.

C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509. URL https://books.google.co.kr/books?id=hV8o5R7_5tkC.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law. *W. Va. L. Rev.*, 123:735, 2020.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617, 2018.

Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio Calmon. Optimized score transformation for fair classification. volume 108 of *Proceedings of Machine Learning Research*, pp. 1673–1683, Online, 26–28 Aug 2020. PMLR. URL http://proceedings.mlr.press/v108/wei20a.html.

Yongkai Wu, Lu Zhang, and Xintao Wu. On convexity and bounds of fairness-aware classification. In *The World Wide Web Conference*, WWW '19, pp. 3356–3362, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313723. URL https://doi.org/10.1145/3308558.3313723.

Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 570–575. IEEE, 2018.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness Constraints: A Flexible Approach for Fair Classification. *J. Mach. Learn. Res.*, 20(75):1–42, 2019.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/zemel13.html.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 335–340, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278779. URL https://doi.org/10.1145/3278721.3278779.

Yuyin Zhou, Shih-Cheng Huang, Jason Alan Fries, Alaa Youssef, Timothy J. Amrhein, Marcello Chang, Imon Banerjee, Daniel Rubin, Lei Xing, Nigam Shah, and Matthew P. Lungren. Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr, 2021. URL https://arxiv.org/abs/2111.11665.

A Proofs of theorems

Let $B > 0$ be the bound of sup-norm of f, i.e., $\sup_{\mathbf{x}} |f(\mathbf{x}, s)| \in [-B, B]$.

Definition A.1. *For $s = 0, 1$ and any measurable set $A \subseteq [-B, B]$, we denote*

$$f_s^{-1}(A) := \{\mathbf{x} \in \mathcal{X}_s : f(\mathbf{x}, s) \in A\}$$

where $\mathcal{X}_s$ as the domain of $\mathbf{X}|S = s$.

Proposition 3.1 For any perfectly group-fair model f, i.e., $\mathcal{P}_{f_0} = \mathcal{P}_{f_1}$, there exists a transport map $\mathbf{T}_s = \mathbf{T}_s(f)$ satisfying $f(\mathbf{X}, s) = f(\mathbf{T}_s(\mathbf{X}), s'), a.e.$

Proof of Proposition 3.1. By letting $\delta = 0$, Theorem 3.3 below implies $\mathbb{E}_s |f(\mathbf{X}, s) - f(\mathbf{T}_s(\mathbf{X}), s')| = 0$. This implies that $|f(\mathbf{X}, s) - f(\mathbf{T}_s(\mathbf{X}), s')| = 0$ almost everywhere, which concludes the proof. □

Theorem 3.3 Fix a fairness level $\delta \geq 0$. Under (C1) and (C2), for any given group-fair model f such that $\Delta\text{TVDP}(f) \leq \delta$, there exists a transport map $\mathbf{T}_s \in \mathcal{T}_s^{\text{trans}}$ satisfying $\Delta\text{MDP}(f, \mathbf{T}_s) \leq C\delta$ for some constant $C > 0$.

Proof of Theorem 3.3. Without loss of generality, let $s = 0$ and $s' = 1$.

Denote $F_0 : [-B, B] \to [0, 1]$ and $F_1 : [-B, B] \to [0, 1]$ the CDFs (Cumulative Distribution Function) of $f(\mathbf{X}, 0)|S = 0$ and $f(\mathbf{X}, 1)|S = 1$, respectively. Note that F_0 and F_1 have at most countably many discontinuous points. We define the set of all discontinuous points of F_s as D_s, which is countable.

1. We define the sub-CDF $F_s^{cont}(v) := F_s(v) - \sum_{t \in D_s, t \leq v} \Delta F_s(t)$ for $v \in [-B, B]$.

   We prove the lemma for the case of $F_1^{cont}(B) \leq F_0^{cont}(B)$. The case of $F_1^{cont}(B) > F_0^{cont}(B)$ can be treated similarly.

   There exists $z \leq B$ such that $F_1^{cont}(B) = F_0^{cont}(z)$. Define $v_k = -B + \lfloor \delta \rfloor k$ for $k \in \{0, \ldots, m-1\}$ where $m \in \mathbb{N}$ satisfies $-B + (m-1)\lfloor \delta \rfloor \leq z \leq -B + m\lfloor \delta \rfloor$. We also let $v_m = z$.

   Fix $k \in \{1, \ldots, m\}$. Suppose that $F_1^{cont}((v_{k-1}, v_k]) \leq F_0^{cont}((v_{k-1}, v_k])$. Then, there exists $z_k \leq v_k$ such that $F_1^{cont}((v_{k-1}, v_k]) = F_0^{cont}((v_{k-1}, z_k])$. Define $\mathcal{X}_{0,k} := f_0^{-1}((v_{k-1}, z_k] \setminus D_0)$ and $\mathcal{X}_{1,k} := f_1^{-1}((v_{k-1}, v_k] \setminus D_1)$. We can define $\mathcal{X}_{0,k}$ and $\mathcal{X}_{1,k}$ similarly when $F_1^{cont}((v_{k-1}, v_k]) \geq F_0^{cont}((v_{k-1}, v_k])$.

   For each $k \in \{1, \ldots, m\}$, we define probability measures $\mathcal{P}_{s,k}, s \in \{0, 1\}$ such that $\mathcal{P}_{s,k}(A) := \mathcal{P}_s(A \cap \mathcal{X}_{s,k})/\mathcal{P}_s(\mathcal{X}_{s,k})$ for measurable subsets $A \subseteq \mathcal{X}$.

   Then, there exists a transport map $\mathbf{T}_{0,k}^{(1)}$ from $\mathcal{P}_{0,k}(\cdot)$ to $\mathcal{P}_{1,k}(\cdot)$, by Breiner's Theorem (Villani, 2008; Hütter & Rigollet, 2021) under (C2). Since $v_k - v_{k-1} \leq \delta, \forall k$, we have that $|f(\mathbf{x}, 0) - f(\mathbf{T}_{0,k}^{(1)}(\mathbf{x}), 1)| \leq \delta$ for $\mathbf{x} \in \mathcal{X}_{0,k}$.

2. Second, we consider the intersection of D_0 and D_1. Let $D_{0,1} := D_0 \cap D_1$.

   Fix $d \in D_{0,1}$. Suppose that $\mathcal{P}_1(f_1^{-1}(\{d\})) \leq \mathcal{P}_0(f_0^{-1}(\{d\}))$. Then, there exists $f_0^{-1}(\{d\})' \subset f_0^{-1}(\{d\})$ such that $\mathcal{P}_0(f_0^{-1}(\{d\}))' = \mathcal{P}_1(f_1^{-1}(\{d\}))$. Define $\tilde{\mathcal{X}}_{0,d} := f_0^{-1}(\{d\})'$ and $\tilde{\mathcal{X}}_{1,d} := f_1^{-1}(\{d\})$. We can define $\tilde{\mathcal{X}}_{0,d}$ and $\tilde{\mathcal{X}}_{1,d}$ similarly when $\mathcal{P}_1(f_1^{-1}(\{d\})) > \mathcal{P}_0(f_0^{-1}(\{d\}))$.

   For each $d \in D_{0,1}$, we define probability measures $\tilde{\mathcal{P}}_{s,d}, s \in \{0, 1\}$ such that $\tilde{\mathcal{P}}_{s,d}(A) := \mathcal{P}_s(A \cap \tilde{\mathcal{X}}_{s,d})/\mathcal{P}_s(\tilde{\mathcal{X}}_{s,d})$ for measurable subsets $A \subseteq \mathcal{X}$.

   Then, there exists a transport map $\mathbf{T}_{0,d}^{(2)}$ from $\tilde{\mathcal{P}}_{0,d}(\cdot)$ to $\tilde{\mathcal{P}}_{1,d}(\cdot)$. By definition of $D_{0,1}$, we note that $f(\mathbf{x}, 0) = f(\mathbf{T}_{0,d}^{(2)}(\mathbf{x}), 1)$ for $\mathbf{x} \in \tilde{\mathcal{X}}_{0,d}$.

3. Third, we collect the complement sets as

$$\mathcal{X}_0' \coloneqq \mathcal{X}_0 \setminus \left(\bigcup_{k\in\{1,\dots,m\}} \mathcal{X}_{0,k} \cup \bigcup_{d\in D_{0,1}} \tilde{\mathcal{X}}_{0,d} \right)$$

and

$$\mathcal{X}_1' \coloneqq \mathcal{X}_1 \setminus \left(\bigcup_{k\in\{1,\dots,m\}} \mathcal{X}_{1,k} \cup \bigcup_{d\in D_{0,1}} \tilde{\mathcal{X}}_{1,d} \right).$$

Because $\mathcal{P}_0(\bigcup_{k\in\{1,\dots,m\}} \mathcal{X}_{0,k}) = \mathcal{P}_1(\bigcup_{k\in\{1,\dots,m\}} \mathcal{X}_{1,k})$ and $\mathcal{P}_0(\bigcup_{d\in D_{0,1}} \tilde{\mathcal{X}}_{0,d}) = \mathcal{P}_1(\bigcup_{d\in D_{0,1}} \tilde{\mathcal{X}}_{1,d})$, we have $\mathcal{P}_0(\mathcal{X}_0') = 1 - \mathcal{P}_0(\bigcup_{k\in\{1,\dots,m\}} \mathcal{X}_{0,k}) - \mathcal{P}_0(\bigcup_{d\in D_{0,1}} \tilde{\mathcal{X}}_{0,d}) = \mathcal{P}_1(\mathcal{X}_1')$.

We define probability measures $\mathcal{P}_s', s \in \{0,1\}$ such that $\mathcal{P}_s'(A) \coloneqq \mathcal{P}_s(A \cap \mathcal{X}_s')/\mathcal{P}_s(\mathcal{X}_s')$ for measurable subsets $A \subseteq \mathcal{X}$.

Then, there exists a transport map $\mathbf{T}_0^{(3)}$ from $\mathcal{P}_0'(\cdot)$ to $\mathcal{P}_1'(\cdot)$.

Furthermore, because $\mathrm{TV}\left(\mathcal{P}_{f(\mathbf{X},0)|S=0}, \mathcal{P}_{f(\mathbf{X},1)|S=1}\right) \le \delta$, we have $\mathcal{P}_0(\mathcal{X}_0') \le \delta$, and by $f(\cdot) \in [-B, B]$, it holds that

$$\begin{aligned} &\mathbb{E}_0\left(|f(\mathbf{X},0) - f(\mathbf{T}_0^{(3)}(\mathbf{X}),1)| \cdot \mathbb{I}(\mathbf{X} \in \mathcal{X}_0')\right) \\ &= \int |f(\mathbf{X},0) - f(\mathbf{T}_0^{(3)}(\mathbf{X}),1)| \cdot \mathbb{I}(\mathbf{X} \in \mathcal{X}_0') d\mathcal{P}_0(\mathbf{X}) \\ &\le 2B \int \mathbb{I}(\mathbf{X} \in \mathcal{X}_0') d\mathcal{P}_0(\mathbf{X}) = 2B\mathcal{P}_0(\mathcal{X}_0') \le 2B\delta. \end{aligned} \tag{8}$$

4. Finally, combining 1 to 3, we define

$$\mathbf{T}_0(\cdot) \coloneqq \sum_{k=1}^{m} \mathbf{T}_{0,k}^{(1)}(\cdot)\mathbb{I}(\cdot \in \mathcal{X}_{0,k}) + \sum_{d\in D_{0,1}} \mathbf{T}_{0,d}^{(2)}(\cdot)\mathbb{I}(\cdot \in \tilde{\mathcal{X}}_{0,d}) + \mathbf{T}_0^{(3)}(\cdot)\mathbb{I}(\cdot \in \mathcal{X}_0'). \tag{9}$$

We note that $\left\{\{\mathcal{X}_{0,k}\}_{k=1}^m, \{\tilde{\mathcal{X}}_{0,d}\}_{d\in D_{0,1}}, \mathcal{X}_0'\right\}$ and $\left\{\{\mathcal{X}_{1,k}\}_{k=1}^m, \{\tilde{\mathcal{X}}_{1,d}\}_{d\in D_{0,1}}, \mathcal{X}_1'\right\}$ are partitions of $\mathcal{X}_0$ and $\mathcal{X}_1$, respectively. Moreover, $\mathcal{P}_0(\mathcal{X}_{0,k}) = \mathcal{P}_1(\mathcal{X}_{1,k}), \forall k$, $\mathcal{P}_0(\tilde{\mathcal{X}}_{0,d}) = \mathcal{P}_1(\tilde{\mathcal{X}}_{1,d}), \forall d$, and $\mathcal{P}_0(\mathcal{X}_0') = \mathcal{P}_1(\mathcal{X}_1')$. Hence, $\mathbf{T}_0$ is a transport map from $\mathcal{P}_0$ to $\mathcal{P}_1$.

Furthermore, we have that

$$\begin{aligned} &\mathbb{E}_0|f(\mathbf{X},0) - f(\mathbf{T}_0(\mathbf{X}),1)| = \int |f(\mathbf{X},0) - f(\mathbf{T}_0(\mathbf{X}),1)| d\mathcal{P}_0(\mathbf{X}) \\ &= \sum_{k\in\{1,\dots,m\}} \int |f(\mathbf{X},0) - f(\mathbf{T}_{0,k}^{(1)}(\mathbf{X}),1)| \cdot \mathbb{I}(\mathbf{X} \in \mathcal{X}_{0,k}) d\mathcal{P}_0(\mathbf{X}) \\ &+ \sum_{d\in D_{0,1}} \int |f(\mathbf{X},0) - f(\mathbf{T}_{0,d}^{(2)}(\mathbf{X}),1)| \cdot \mathbb{I}(\mathbf{X} \in \tilde{\mathcal{X}}_{0,d}) d\mathcal{P}_0(\mathbf{X}) \\ &+ \int |f(\mathbf{X},0) - f(\mathbf{T}_0^{(3)}(\mathbf{X}),1)| \cdot \mathbb{I}(\cdot \in \mathcal{X}_0') d\mathcal{P}_0(\mathbf{X}) \\ &\le \delta \sum_{k\in\{1,\dots,m\}} \mathcal{P}_0(\mathcal{X}_{0,k}) + \int |f(\mathbf{X},0) - f(\mathbf{T}_0^{(3)}(\mathbf{X}),1)| \cdot \mathbb{I}(\cdot \in \mathcal{X}_0') d\mathcal{P}_0(\mathbf{X}) \\ &\le \delta + 2B\delta = (1+2B)\delta. \end{aligned} \tag{10}$$

Letting $C = 1 + 2B$ completes the proof.

□

Theorem 3.5 For a given $\mathbf{T}_s \in \mathcal{T}_s^{\text{trans}}$, if $\Delta\text{MDP}(f, \mathbf{T}_s) \leq \delta$, then we have $\Delta\text{WDP}(f) \leq \delta$ and $\Delta\overline{\text{DP}}(f) \leq \delta$.

Proof of Theorem 3.5. Fix $f \in \{f \in \mathcal{F} : \mathbb{E}_s|f(\mathbf{X}, s) - f(\mathbf{T}_s(\mathbf{X}), s')| \leq \delta\}$. Let $\mathcal{L}_1$ the set of all 1-Lipschitz functions. Using the fact that Wasserstein-1 distance is equivalent to IPM induced by set of 1-Lipschitz function (Villani, 2008), we have that

$$
\begin{aligned}
\Delta\text{WDP}(f) &= \mathcal{W}\left(\mathcal{P}_{f(\mathbf{X},0)|S=0}, \mathcal{P}_{f(\mathbf{X},1)|S=1}\right) \\
&= \sup_{u\in\mathcal{L}_1} |\mathbb{E}_s(u \circ f(\mathbf{X}, s)) - \mathbb{E}_{s'}(u \circ f(\mathbf{X}, s'))| \\
&\leq \sup_{u\in\mathcal{L}_1} |\mathbb{E}_s(u \circ f(\mathbf{X}, s)) - \mathbb{E}_s(u \circ f(\mathbf{T}_s(\mathbf{X}), s'))| \\
&+ \sup_{u\in\mathcal{L}_1} |\mathbb{E}_s(u \circ f(\mathbf{T}_s(\mathbf{X}), s')) - \mathbb{E}_{s'}(u \circ f(\mathbf{X}, s'))| \\
&\leq \sup_{u\in\mathcal{L}_1} \mathbb{E}_s |u \circ f(\mathbf{X}, s) - u \circ f(\mathbf{T}_s(\mathbf{X}), s')| \\
&+ \sup_{u\in\mathcal{L}_1} |\mathbb{E}_s(u \circ f(\mathbf{T}_s(\mathbf{X}), s')) - \mathbb{E}_{s'}(u \circ f(\mathbf{X}, s'))| \\
&\overset{u\in\mathcal{L}_1}{\leq} \mathbb{E}_s |f(\mathbf{X}, s) - f(\mathbf{T}_s(\mathbf{X}), s')| \\
&+ \sup_{u\in\mathcal{L}_1} |\mathbb{E}_s(u \circ f(\mathbf{T}_s(\mathbf{X}), s')) - \mathbb{E}_{s'}(u \circ f(\mathbf{X}, s'))| \\
&\leq \delta + \sup_{u\in\mathcal{L}_1} |\mathbb{E}_s(u \circ f(\mathbf{T}_s(\mathbf{X}), s')) - \mathbb{E}_{s'}(u \circ f(\mathbf{X}, s'))| \\
&\leq \delta + \sup_{f\in\mathcal{F}} \sup_{u\in\mathcal{L}_1} |\mathbb{E}_s(u \circ f(\mathbf{T}_s(\mathbf{X}), s')) - \mathbb{E}_{s'}(u \circ f(\mathbf{X}, s'))| \\
&\leq \delta + \text{TV}(\mathbf{T}_{s\#}\mathcal{P}_s, \mathcal{P}_{s'}) = \delta.
\end{aligned}
\tag{11}
$$

The last equality holds since $\text{TV}(\mathbf{T}_{s\#}\mathcal{P}_s, \mathcal{P}_{s'}) = 0$ for any transport map $\mathbf{T}_s$.
For $\Delta\overline{\text{DP}}(f)$, because the identity map is 1-Lipschitz, we have that $\Delta\overline{\text{DP}}(f) \leq \Delta\text{WDP}(f)$, which completes the proof. □

Theorem 4.2 Suppose $\mathcal{F}$ is the collection of L-Lipschitz functions. Let A be a given subset in $\mathcal{X}$. Then, for all f satisfying $\Delta\mathrm{MDP}(f, \mathbf{T}_s^f) \le \delta$, we have

$$\Delta\overline{\mathrm{DP}}_A(f) \le L\left(\mathbb{E}_s\|\mathbf{X} - \mathbf{T}_s^f(\mathbf{X})\|^2\right)^{\frac{1}{2}} + \mathrm{TV}(\mathcal{P}_{0,A}, \mathcal{P}_{1,A}) + U\delta, \tag{12}$$

where $\mathcal{P}_{s,A}$ is the distribution of $\mathbf{X}|S = s, \mathbf{X} \in A$, and $U > 0$ is a constant only depending on A and $\mathcal{P}_s, s = 0, 1$.

Proof. We write $\mathbf{T}_s = \mathbf{T}_s^f$ for notational simplicity.

$$\begin{aligned}
&|\mathbb{E}(f(\mathbf{X},0)|S=0,\mathbf{X}\in A) - \mathbb{E}(f(\mathbf{X},1)|S=1,\mathbf{X}\in A)| \\
&\le |\mathbb{E}(f(\mathbf{X},0)|S=1,\mathbf{X}\in A) - \mathbb{E}(f(\mathbf{T}_1(\mathbf{X}),0)|S=1,\mathbf{X}\in A)| \\
&+ |\mathbb{E}(f(\mathbf{T}_1(\mathbf{X}),0)|S=1,\mathbf{X}\in A) - \mathbb{E}(f(\mathbf{X},1)|S=1,\mathbf{X}\in A)| \\
&+ |\mathbb{E}(f(\mathbf{X},0)|S=0,\mathbf{X}\in A) - \mathbb{E}(f(\mathbf{X},0)|S=1,\mathbf{X}\in A)|.
\end{aligned} \tag{13}$$

By (C1), the first term is bounded by $L\mathbb{E}_1\|\mathbf{X} - \mathbf{T}_1(\mathbf{X})\|$, which is also bounded by $L\left(\mathbb{E}_1\|\mathbf{X} - \mathbf{T}_1(\mathbf{X})\|^2\right)^{1/2}$. The second term is bounded by δ up to a constant for all f satisfying $\Delta\mathrm{MDP}(f, \mathbf{T}_s) \le \delta$. That is, we have

$$\begin{aligned}
&|\mathbb{E}(f(\mathbf{T}_1(\mathbf{X}),0)|S=1,\mathbf{X}\in A) - \mathbb{E}(f(\mathbf{X},1)|S=1,\mathbf{X}\in A)| \\
&= \left|\frac{\int f(\mathbf{T}_1(\mathbf{X}),0)\mathbb{I}(\mathbf{X}\in A)d\mathcal{P}_1(\mathbf{X})}{\int \mathbb{I}(\mathbf{X}\in A)d\mathcal{P}_1(\mathbf{X})} - \frac{\int f(\mathbf{X},1)\mathbb{I}(\mathbf{X}\in A)d\mathcal{P}_1(\mathbf{X})}{\int \mathbb{I}(\mathbf{X}\in A)d\mathcal{P}_1(\mathbf{X})}\right| \\
&\le \frac{1}{\int \mathbb{I}(\mathbf{X}\in A)d\mathcal{P}_1(\mathbf{X})}\int_{\mathbf{X}\in A}|f(\mathbf{T}_1(\mathbf{X}),0) - f(\mathbf{X},1)|d\mathcal{P}_1(\mathbf{X}) \\
&\le \frac{1}{\int \mathbb{I}(\mathbf{X}\in A)d\mathcal{P}_1(\mathbf{X})}\int_{\mathbf{X}\in\mathcal{X}}|f(\mathbf{T}_1(\mathbf{X}),0) - f(\mathbf{X},1)|d\mathcal{P}_1(\mathbf{X}) \\
&= U'(A,\mathcal{P}_1)\times\mathbb{E}_1|f(\mathbf{T}_1(\mathbf{X}),0) - f(\mathbf{X},1)| \\
&\le U'(A,\mathcal{P}_1)\times\delta
\end{aligned} \tag{14}$$

where $U'(A,\mathcal{P}_1) = 1/\int \mathbb{I}(\mathbf{X}\in A)d\mathcal{P}_1(\mathbf{X}) = 1/\mathcal{P}(\mathbf{X}\in A|S=1)$ is a constant only depending on $\mathcal{P}_1$ and A.

The third term $|\mathbb{E}(f(\mathbf{X},0)|S=0,\mathbf{X}\in A) - \mathbb{E}(f(\mathbf{X},0)|S=1,\mathbf{X}\in A)|$ is not controllable by either the transport map or δ but depends on the given distributions and A. That is,

$$\begin{aligned}
&|\mathbb{E}(f(\mathbf{X},0)|S=0,\mathbf{X}\in A) - \mathbb{E}(f(\mathbf{X},0)|S=1,\mathbf{X}\in A)| \\
&= \left|\int_A f(\mathbf{X},0)d\mathcal{P}_0(\mathbf{X}) - \int_A f(\mathbf{X},0)d\mathcal{P}_1(\mathbf{X})\right| \\
&\le \sup_{f\in\mathcal{F}}\left|\int_A f(\mathbf{X},0)d\mathcal{P}_0(\mathbf{X}) - \int_A f(\mathbf{X},0)d\mathcal{P}_1(\mathbf{X})\right| \\
&\le \mathrm{TV}(\mathcal{P}_{0,A},\mathcal{P}_{1,A}).
\end{aligned} \tag{15}$$

Hence, we have

$$\begin{aligned}
&|\mathbb{E}(f(\mathbf{X},0)|S=0,\mathbf{X}\in A) - \mathbb{E}(f(\mathbf{X},1)|S=1,\mathbf{X}\in A)| \\
&\le L(\mathbb{E}_1\|\mathbf{X} - \mathbf{T}_1(\mathbf{X})\|^2)^{1/2} + \mathrm{TV}(\mathcal{P}_{0,A},\mathcal{P}_{1,A}) + U'(A,\mathcal{P}_1)\delta.
\end{aligned} \tag{16}$$

We can similarly derive

$$\begin{aligned}
&|\mathbb{E}(f(\mathbf{X},0)|S=0,\mathbf{X}\in A) - \mathbb{E}(f(\mathbf{X},1)|S=1,\mathbf{X}\in A)| \\
&\le L(\mathbb{E}_0\|\mathbf{X} - \mathbf{T}_0(\mathbf{X})\|^2)^{1/2} + \mathrm{TV}(\mathcal{P}_{0,A},\mathcal{P}_{1,A}) + U'(A,\mathcal{P}_0)\delta.
\end{aligned} \tag{17}$$

Letting $U := \max\{U'(A,\mathcal{P}_0), U'(A,\mathcal{P}_1)\}$ completes the proof. □

B More related works on algorithmic fairness

Several recent studies have raised concerns that focusing solely on group fairness is not always a complete answer. For instance, Dwork et al. (2012) highlighted issues such as subset targeting and self-fulfilling prophecy, leading to the notion of individual fairness. On the other hand, individual fairness alone does not ensure group fairness when the distributions of protected groups differ significantly.

In presence of multiple sensitive attributes, the concept of subgroup fairness has emerged (Kearns et al., 2018a;b; Shui et al., 2022; Mehrotra et al., 2022; Molina & Loiseau, 2022; Carvalho et al., 2022), emphasizing the need for models that satisfy fairness for all subgroups defined over the multiple sensitive attributes. However, collecting all sensitive attributes a priori would be challenging and also can undermine privacy and security.

In addition, Wachter et al. (2020); Simons et al. (2021); Mougan et al. (2024) have suggested that not only equal outcome, but also the notion of equal treatment, i.e., treating individuals with equal reasons, should be considered.

C Disadvantage of high transport cost

This section presents an example of two completely different group-fair models where one is unreasonable and the other is reasonable, particularly in terms of subset fairness. This example suggests that *not all group-fair models are acceptable*, thereby emphasizing the necessity of finding group-fair models with favorable implicit transport maps.

Suppose that the distribution of the input variable is given as $\mathbf{X}|S = s \sim \text{Unif}(0,1)$, for $s \in \{0,1\}$. Consider the following two classification models: $\hat{f}(\mathbf{x}, s) = \frac{\text{sign}(2\mathbf{x}-1)(1-2s)+1}{2}$ and $\tilde{f}(\mathbf{x}, s) = \frac{\text{sign}(2\mathbf{x}-1)+1}{2}$.

It is clear that both $\hat{f}$ and $\tilde{f}$ are perfectly fair, i.e., $\mathcal{P}_{\hat{f}_0} = \mathcal{P}_{\hat{f}_1}$ and $\mathcal{P}_{\tilde{f}_0} = \mathcal{P}_{\tilde{f}_1}$. However, $\hat{f}$ has a notable unfairness issue in its treatments of individuals within the subset $\{\mathbf{x} \geq 1/2\}$ (as well as $\{\mathbf{x} < 1/2\}$); for when $\mathbf{x} > 1/2$, $\hat{f}$ assigns label 1 for all individuals of $s = 0$ while it assigns label 0 for all individuals of $s = 1$. This indicates that $\hat{f}$ discriminates against individuals in the subset $\{\mathbf{x} \geq 1/2\}$ (and also $\{\mathbf{x} < 1/2\}$). In contrast, $\tilde{f}$ does not exhibit such undesirable discrimination against the subsets. Hence, we can say that $\tilde{f}$ has less discrimination on the subsets than $\hat{f}$. Figure 5 provides a comparative illustration of $\hat{f}$ and $\tilde{f}$.

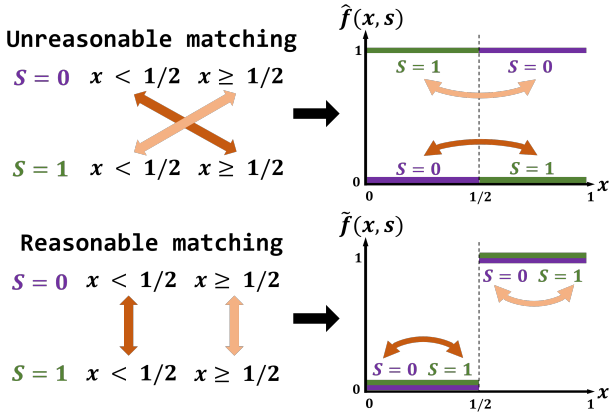

Figure 5: (Top) A group-fair model with the risk of discrimination on subsets. (Bottom) A group-fair model without the risk of discrimination on subsets.

The observed discrimination of $\hat{f}$ on subsets can be attributed to the unreasonable matching function of $\hat{f}$. It turns out that the matching function of $\hat{f}$ is $\mathbf{T}^{\hat{f}}(\mathbf{x}) = \mathbf{x} - \frac{\text{sign}((2\mathbf{x}-1)(1-2s))}{2}$. This function matches an individual in $\{\mathbf{x} < \frac{1}{2}, S = s\}$ with one in $\{\mathbf{x} \geq \frac{1}{2}, S = s'\}$, who are far apart from each other. In contrast, the matching function of $\tilde{f}$ is $\mathbf{T}^{\tilde{f}}(\mathbf{x}) = \mathbf{x}$.

This example emphasizes the need of group-fair models whose matching function have low transport costs.

D Implementation details

In this section, we provide detailed descriptions for the implementation of the experiments.

D.1 Datasets

First, the URLs of these datasets are provided.

ADULT: the Adult income dataset (Dua & Graff, 2017) can be downloaded from the UCI repository[1].

GERMAN: the German credit dataset (Dua & Graff, 2017) can be downloaded from the UCI repository[2].

DUTCH: the Dutch census dataset can be downloaded from the public Github of Quy et al. (2022) [3].

BANK: the Bank marketing dataset can be downloaded from the UCI repository[4].

Second, we describe pre-processing method of the datasets used. For ADULT, GERMAN, and BANK datasets, we follow the pre-processing of the implementation of IBM's AIF360 (Bellamy et al., 2018) [5]. For DUTCH dataset, we follow the pre-processing of Quy et al. (2022)'s Github[6]. Basically, continuous input variables are normalized by min-max scaling and categorical input variables are one-hot encoded. We set batch size as 1024, 200, 1024, and 512 for ADULT, GERMAN, DUTCH, and BANK datasets, respectively.

D.2 Algorithms

This section provides more detailed descriptions of the baseline algorithms used in our experiments.

- Reduction (Agarwal et al., 2018): This algorithm is an in-processing method that learns a fair classifier with the lowest empirical fairness level ΔDP. To implement this method for MLP model architecture, we employ FairTorch[7]. It minimizes cross-entropy + $\lambda \cdot$ Reduction regularizer for a given $\lambda > 0$.
- Reg (Donini et al., 2018; Chuang & Mroueh, 2021): This method is a regularizing approach that minimizes cross-entropy + $\lambda \cdot \Delta\overline{\text{DP}}^2$ for a given $\lambda > 0$. In Chuang & Mroueh (2021), they call this algorithm GapReg. This is also similar to the approach of Donini et al. (2018) in the sense that the model is learned with a constraint having a given level of $\Delta\overline{\text{DP}}$.
- Adv (Zhang et al., 2018): This algorithm is an in-processing method that regularizes the model outputs with an adversarial network so that the adversarial network is learned to predict the sensitive attribute using the model outputs as the inputs. It minimizes cross-entropy + $\lambda \cdot$ Adversarial loss for a given $\lambda > 0$.

Note that Reduction and Adv are ones of the most popular in-processing algorithms, as widely-used libraries AIF360 (Bellamy et al., 2018) and Fairlearn (Bird et al., 2020) provide the usage and implementation of the two algorithms. Reg is a vanilla approach of adding the regularization term in the loss function to learn the most accurate model among models satisfying a given level of group fairness.

We basically train models with various fairness levels by controlling the Lagrangian multiplier λ. The values are reported in the following table.

The Adam optimizer (Kingma & Ba, 2014) with an initial learning rate of 0.001 is used, and the learning rate is scheduled by multiplying 0.95 at each epoch.

[1] https://archive.ics.uci.edu/ml/datasets/adult
[2] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
[3] https://github.com/tailequy/fairness_dataset/tree/main/experiments/data/dutch.csv
[4] https://archive.ics.uci.edu/dataset/222/bank+marketing
[5] https://aif360.readthedocs.io/en/stable/
[6] https://github.com/tailequy/fairness_dataset/tree/main/experiments/data/
[7] `https://github.com/wbawakate/fairtorch`

Table 4: Hyper-parameters used for controlling fairness levels for each algorithm.

Algorithm	λ
Reduction	$\{0.5, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 8.0, 10.0, 20.0, 30.0, 40.0, 50.0, 60.0, 80.0, 100.0, 150.0, 200.0, 300.0, 500.0\}$
Reg	$\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.2, 1.5, 1.8, 2.0, 3.0, 5.0, 10.0, 20.0, 50.0, 100.0\}$
Adv	$\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.5, 2.0, 3.0, 5.0, 10.0, 15.0, 20.0, 30.0, 50.0, 100.0\}$
FTM	$\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.5, 2.0, 3.0, 5.0, 10.0\}$

D.3 Pseudo-code

Here, we provide a `Pytorch`-style psuedo code of calculating the matching constraint in FTM.

Algorithm 1: `PyTorch`-style pseudo-code of calculating the matching constraint in FTM.

```
# xs, xt: input vectors from the source, target distribution, respectively.
# model: a classifier to be trained
import ot
# The matching constraint: matching with the OT map
weight_s = torch.ones(size=(xs.size(0), )) / xs.size(0)
weight_t = torch.ones(size=(xt.size(0), )) / xt.size(0) # identical to weight_s
M = ot.dist(xs, xt)
G = ot.emd(weight_s, weight_t, M)
matched_xs = xt[torch.argmax(G, dim=1)]
output, matched_output = model(xs), model(matched_xs)
FTM_REG = (output - matched_output).abs().mean()
```

E Auxillary experimental results

In this section, we provide auxillary experimental results that are not displayed in the main body.

E.1 Fairness-prediction trade-off (Section 5.2.1)

Figure 6 shows the trade-offs between the fairness levels with respect to $\Delta\mathrm{DP}$, $\Delta\overline{\mathrm{DP}}$ and classification accuracy.

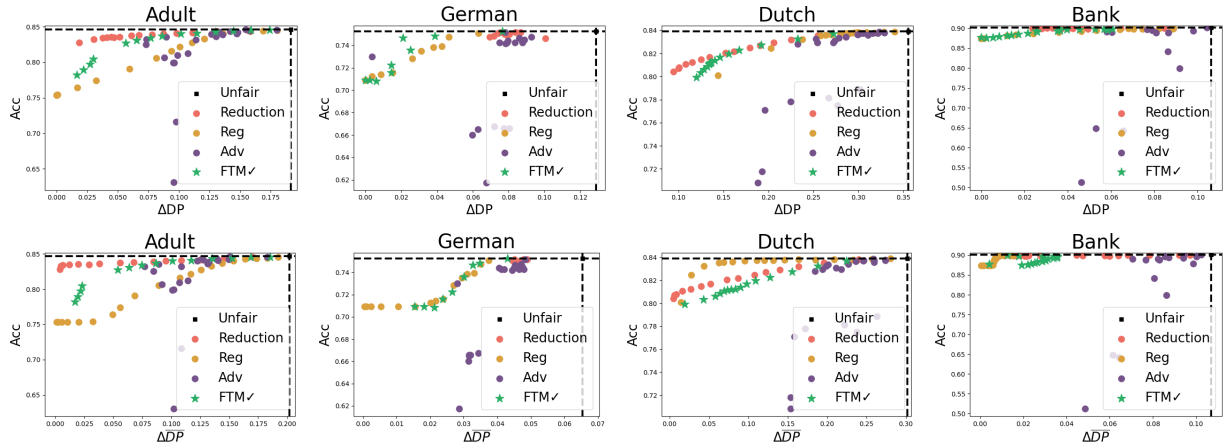

Figure 6: **Fairness-prediction trade-offs**: (Left to right) ADULT, GERMAN, DUTCH, BANK. (Top to bottom) $\Delta\mathrm{DP}$ vs. `Acc`, $\Delta\overline{\mathrm{DP}}$ vs. `Acc`.

E.2 Improvement in subset fairness (Section 5.2.2)

Here, we provide experimental results showing fairness levels on subsets defined by input variables. Tables 5 and 6 are copies of Table 2 with standard errors.

Table 5: **Fairness on subsets defined by the input variable age**: Fairness levels on subsets defined by the input variable age on GERMAN dataset under a given $\Delta\overline{\mathrm{DP}} = 0.045$ with standard errors (s.e.).

Algorithm		Reduction	Reg	Adv	FTM ✓
High age	$\Delta\mathrm{DP}$ (s.e.)	0.073 (0.015)	0.077 (0.013)	<u>0.048</u> (0.020)	**0.045** (0.021)
	$\Delta\overline{\mathrm{DP}}$ (s.e.)	0.049 (0.006)	0.029 (0.008)	<u>0.028</u> (0.012)	**0.026** (0.006)
	$\Delta\mathrm{WDP}$ (s.e.)	0.053 (0.005)	<u>0.039</u> (0.003)	0.042 (0.008)	**0.038** (0.003)
Low age	$\Delta\mathrm{DP}$ (s.e.)	0.118 (0.035)	<u>0.116</u> (0.037)	0.122 (0.047)	**0.077** (0.032)
	$\Delta\overline{\mathrm{DP}}$ (s.e.)	**0.047** (0.015)	<u>0.050</u> (0.009)	0.053 (0.017)	**0.047** (0.007)
	$\Delta\mathrm{WDP}$ (s.e.)	<u>0.058</u> (0.011)	0.059 (0.007)	0.061 (0.015)	**0.054** (0.006)

Table 6: **Fairness on subsets defined by the input variable marital status**: Fairness levels on subsets defined by the input variable marital status on DUTCH dataset under a given $\Delta\overline{\mathrm{DP}} = 0.12$ with standard errors (s.e.).

Algorithm		Reduction	Reg	Adv	FTM ✓
Married	$\Delta\mathrm{DP}$ (s.e.)	0.258 (0.005)	0.372 (0.003)	<u>0.237</u> (0.083)	**0.204** (0.003)
	$\Delta\overline{\mathrm{DP}}$ (s.e.)	0.182 (0.002)	<u>0.164</u> (0.001)	0.187 (0.073)	**0.152** (0.002)
	$\Delta\mathrm{WDP}$ (s.e.)	0.183 (0.002)	<u>0.172</u> (0.001)	0.193 (0.071)	**0.152** (0.002)
Not married	$\Delta\mathrm{DP}$ (s.e.)	**0.061** (0.007)	0.131 (0.006)	0.095 (0.038)	<u>0.068</u> (0.005)
	$\Delta\overline{\mathrm{DP}}$ (s.e.)	<u>0.045</u> (0.002)	0.062 (0.003)	0.098 (0.035)	**0.036** (0.003)
	$\Delta\mathrm{WDP}$ (s.e.)	**0.045** (0.003)	0.072 (0.002)	0.098 (0.034)	**0.045** (0.005)

E.3 An additional advantage of using the marginal OT map: reducing the risk of self-fulfilling prophecy

We compare the risks of discrimination in the context of self-fulfilling prophecy in Dwork et al. (2012), a critical limitation that can arise when focusing solely on group fairness: *unqualified individuals with relatively low scores can be chosen to be qualified, while other individuals with relatively high scores are chosen to be unqualified.* To quantify the risk of self-fulfilling prophecy, we assume that the unfair model is optimal for predicting the true score of each individual. We consider the following two evaluation approaches under this assumption. For the transport map used in MDP constraint, we choose the marginal OT map.

(Evaluation 1) The first measure for the risk of self-fulfilling prophecy is the *Spearman's rank correlation* between unfair and fair prediction scores at each protected group: a higher rank correlation implies a lower risk of self-fulfilling prophecy. Table 7 shows that FTM has lower risks of suffering from self-fulfilling prophecy, in most cases.

Table 7: Spearman's correlation coefficients between the scores of the unfair model and group-fair models under fixed levels of $\Delta\overline{\text{DP}}$ with standard errors (s.e.). **Bold** faces are the best ones, and underlined ones are the second bests.

Dataset	Adult		German	
$\Delta\overline{\text{DP}}$	0.10		0.05	
Sensitive attribute S	0	1	0	1
Reduction (s.e.)	0.935 (0.006)	0.987 (0.001)	0.996 (0.001)	0.997 (0.001)
Reg (s.e.)	0.762 (0.087)	0.806 (0.084)	**0.997** (0.000)	**0.998** (0.000)
Adv (s.e.)	0.876 (0.003)	0.979 (0.001)	0.986 (0.009)	0.986 (0.010)
FTM ✓ (s.e.)	**0.968** (0.003)	**0.989** (0.001)	0.993 (0.002)	0.995 (0.001)
Dataset	Dutch		Bank	
$\Delta\overline{\text{DP}}$	0.01		0.02	
Sensitive attribute S	0	1	0	1
Reduction (s.e.)	0.940 (0.001)	0.922 (0.001)	0.958 (0.010)	0.978 (0.005)
Reg (s.e.)	0.872 (0.003)	0.972 (0.003)	0.784 (0.031)	0.974 (0.003)
Adv (s.e.)	0.659 (0.171)	0.693 (0.185)	0.603 (0.207)	0.505 (0.238)
FTM ✓ (s.e.)	**0.973** (0.002)	**0.991** (0.000)	**0.964** (0.007)	**0.979** (0.004)

(Evaluation 2) For the second approach, we employ 2×2 confusion matrices to compare the predicted labels of the unfair and the group-fair models. In specific, in the privileged group $S = 1$, individuals predicted as $\hat{Y} = 0$ (i.e., unqualified) by the unfair model but $\hat{Y} = 1$ (i.e., chosen to be qualified) by the group-fair model are considered as undesirable instances in the context of self-fulfilling prophecy. Likewise, in the unprivileged group $S = 0$, individuals predicted as $\hat{Y} = 1$ by the unfair model but $\hat{Y} = 0$ by the group-fair model are similarly considered undesirable.

That is, for the risk of self-fulfilling prophecy, we count *the number of individuals whose prediction is undesirably flipped* (i.e., # of $\hat{Y} = 0$ (Unfair) $\rightarrow \hat{Y} = 1$ (Fair) for $S = 1$, and # of $\hat{Y} = 1$ (Unfair) $\rightarrow \hat{Y} = 0$ (Fair) for $S = 0$). Table 8 shows that the undesirable treatments of FTM are less observed than those of baseline methods, in most cases.

Table 8: 2×2 confusion matrices comparing the predicted labels of the unfair model and the group-fair models. The encircled numbers are the counts of undesirable instances. **Bold** faces are the best ones and underlined ones are the second bests.

Dataset ($\Delta\overline{\text{DP}}$)		Adult (0.05)		German (0.05)		Dutch (0.15)		Bank (0.04)	
		Unfair							
$S = 1$		$\hat{Y} = 0$	$\hat{Y} = 1$	$\hat{Y} = 0$	$\hat{Y} = 1$	$\hat{Y} = 0$	$\hat{Y} = 1$	$\hat{Y} = 0$	$\hat{Y} = 1$
Reduction	$\hat{Y} = 0$	6124	629	98	1	2170	662	5220	62
	$\hat{Y} = 1$	(22)	1701	(1)	32	(8)	3158	(62)	579
Reg	$\hat{Y} = 0$	6144	2198	99	8	2164	311	5265	229
	$\hat{Y} = 1$	(2)	132	(**0**)	25	(14)	3509	(17)	412
Adv	$\hat{Y} = 0$	6121	977	95	0	2152	1127	5255	516
	$\hat{Y} = 1$	(25)	1353	(4)	33	(26)	2693	(27)	125
FTM ✓	$\hat{Y} = 0$	6146	1364	99	1	2174	862	5279	397
	$\hat{Y} = 1$	(**0**)	966	(**0**)	32	(**4**)	2958	(**3**)	244
		Unfair							
$S = 0$		$\hat{Y} = 0$	$\hat{Y} = 1$	$\hat{Y} = 0$	$\hat{Y} = 1$	$\hat{Y} = 0$	$\hat{Y} = 1$	$\hat{Y} = 0$	$\hat{Y} = 1$
Reduction	$\hat{Y} = 0$	3486	(13)	54	(**3**)	4137	(**0**)	129	(14)
	$\hat{Y} = 1$	262	341	0	11	226	1723	1	31
Reg	$\hat{Y} = 0$	3748	(104)	54	(4)	4300	(13)	128	(45)
	$\hat{Y} = 1$	0	250	0	10	63	1710	2	0
Adv	$\hat{Y} = 0$	3655	(52)	53	(**3**)	3917	(85)	125	(34)
	$\hat{Y} = 1$	93	302	1	11	446	1638	5	11
FTM ✓	$\hat{Y} = 0$	3719	(**11**)	54	(**3**)	4217	(6)	120	(**10**)
	$\hat{Y} = 1$	29	343	0	11	146	1717	10	35