Journal of Machine Learning Research 27 (2026) 1-42 Submitted 8/24; Published 1/26

Demographic Parity in Regression and Classification Within the Unawareness Framework

Vincent Divol VINCENT.DIVOL@ENSAE.FR
CREST, ENSAE, IP Paris
5 avenue Henry Le Chatelier
91120 Palaiseau, France

Solenne Gaucher SOLENNE.GAUCHER@POLYTECHNIQUE.EDU
CMAP, cole polytechnique, IP Paris
Route de Saclay
91120 Palaiseau, France

Editor: Maya Gupta

Abstract

This paper explores the theoretical foundations of fair regression under the constraint of demographic parity within the unawareness framework, where disparate treatment is prohibited, extending existing results where such treatment is permitted. Specifically, we aim to characterize the optimal fair regression function when minimizing the quadratic loss. Our results reveal that this function is given by the solution to a barycenter problem with optimal transport costs. Additionally, we study the connection between optimal fair cost-sensitive classification, and optimal fair regression. We demonstrate that nestedness of the decision sets of the classifiers is both necessary and sufficient to establish a form of equivalence between classification and regression. Under this nestedness assumption, the optimal classifiers can be derived by applying thresholds to the optimal fair regression function; conversely, the optimal fair regression function is characterized by the family of cost-sensitive classifiers.

Keywords: statistical fairness, demographic parity, optimal transport.

1. Introduction

Recent breakthroughs in artificial intelligence have led to the widespread adoption of machine learning algorithms, exerting an increasingly influential and insidious impact on our lives. Essentially, these algorithms learn to detect and reproduce patterns using massive datasets. It is now widely recognized that these predictions carry the risk of perpetuating, or even exacerbating, the social discriminations and biases often present in these datasets (Angwin et al., 2016; Barocas et al., 2023). Algorithmic fairness seeks to measure and mitigate the unfair impact of algorithms; we refer the reader to the reviews by Barocas et al. (2023); del Barrio et al. (2020); Oneto and Chiappa (2020) for an introduction. Beyond machine learning, ethical theory offers three classical lenses for fairness: *consequentialist* (outcome-based), *deontological* (rule-based), and *virtue-ethics* (character-based). Most machine learning work has operationalized consequentialist notions via statistical criteria on outcomes or error rates. Deontological approaches encode inviolable constraints (rules) directly into models (Wang and Gupta, 2020), while recent LLM settings highlight a virtue-ethical perspective that asks about the dispositions systems exhibit (e.g., honesty, harmlessness). We position our contribution squarely within the consequentialist (or statistical) tradition.

Different approaches have been developed to mitigate algorithmic unfairness. One approach focuses on *individual fairness*, ensuring that similar individuals are treated similarly, regardless of potentially discriminatory factors. Another approach targets *group fairness*, aiming to prevent algorithmic predictions from discriminating against groups of individuals. *Statistical fairness* falls under the latter approach and relies on the formalism of supervised learning to impose fairness criteria while minimizing a risk measure. In this work, we study risk minimization under the demographic parity criterion, which requires that predictions be statistically independent of sensitive attributes. Although this criterion, introduced by Calders et al. (2009); Agarwal et al. (2019), has some known limitations (Hardt et al., 2016; Zafar et al., 2019), it finds application in a wide range of scenarios (Makhlouf et al., 2021; Denis et al., 2024). Its simplicity arguably makes it the most extensively studied criterion.

The statistical fairness literature can be broadly divided into two lines of work: the *awareness framework*, which allows using protected attributes in predictions (also called disparate treatment), and the *unawareness framework*, which prohibits such use. We will follow the unawareness framework. Empirical evidence from simulations (Lipton et al., 2018) indicates that within this framework, predictions often result in suboptimal trade-offs between fairness and accuracy and may induce within-group discrimination. Gaucher et al. (2023) proved that while the unawareness framework aims to prevent discrimination based on sensitive attributes, predictions in this setting implicitly rely on estimates of these attributes, which had previously been conjectured by Lipton et al. (2018).

In this paper, we investigate the problem of fair regression under demographic parity constraints within the unawareness framework. A key difficulty in overcoming algorithmic unfairness is the limited understanding of how fair algorithms make predictions. Therefore, we focus on providing a simple mathematical characterization of the optimal regression function in the presence of fairness constraints.

1.1 Problem statement

Let (X, S, Y) be a tuple in $\mathcal{X} \times \mathcal{S} \times \mathbb{R}$ with distribution $\mathbb{P}$, where X corresponds to a non-sensitive feature in a feature space $\mathcal{X}$, S is a sensitive attribute in a finite set $\mathcal{S}$, and Y is a response variable that we want to predict, which has a finite second moment. To illustrate this problem with an example, assume, as in Chzhen and Schreuder (2020a), that X represents a candidate's skill, S is an attribute indicating population groups, and Y is the current market salary of the candidate. Due to historical biases, the distribution of the salary may be unbalanced between the groups. Our aim is to make predictions that are fair, and as close as possible to the current market value Y. In the unawareness framework, we cannot make explicit use of the sensitive attribute to make our predictions. Therefore, we consider regression functions of the form $f : \mathcal{X} \to \mathbb{R}$ in the set of score functions $\mathcal{F}$. We want to ensure that our regression function satisfies the following demographic parity criterion.

Definition 1 (Demographic parity) *The function $f : \mathcal{X} \to \mathbb{R}$ satisfies the Demographic Parity criterion if*

$$f(X) \perp S.$$

In essence, the demographic parity criterion requires that the distribution of predictions (in our example, the salary) be identical across all groups. We assess the quality of a regression function f through its quadratic risk

$$\mathcal{R}_{sq}(f) = \mathbb{E}\left[(Y - f(X))^2\right].$$

Definition 2 (Fair regression) *An optimal fair regression function f^* satisfies*

$$f^* \in \arg\min_{f \in \mathcal{F}} \{\mathcal{R}_{sq}(f) \ : \ f(X) \perp S\}, \tag{1}$$

where $\mathcal{F}$ is the set of regression functions from $\mathcal{X}$ to $\mathbb{R}$.

Classical results show that when no fairness constraints are imposed, the Bayes regression function f^{Bayes} minimizing the squared risk $\mathcal{R}_{sq}$ is a.s. equal to the conditional expectation η, where

$$\eta(x) = \mathbb{E}[Y|X = x].$$

In this paper, we also investigate the relationship between classification and regression problems. When $Y \in \{0, 1\}$ a.s., the quality of a classification function $g : \mathcal{X} \to \{0, 1\}$ can be assessed through its expected weighted $0 - 1$ loss $\mathcal{R}_y(g)$, where for $y \in [0, 1]$, $\mathcal{R}_y(g)$ is defined as

$$\mathcal{R}_y(g) = y \cdot \mathbb{P}[Y = 0, g(X) = 1] + (1 - y) \cdot \mathbb{P}[Y = 1, g(X) = 0].$$

For the choice $y = 1/2$, minimizing this risk measure corresponds to maximizing the classical accuracy measure.

Definition 3 (Fair classification) *For a given value $y \in [0, 1]$, an optimal fair classification function g^*_y satisfies*

$$g^*_y \in \arg\min_{g \in \mathcal{G}} \{\mathcal{R}_y(g) \ : \ g(X) \perp S\}, \tag{2}$$

where $\mathcal{G}$ is the set of classification functions from $\mathcal{X}$ to $\{0, 1\}$.

Let us again illustrate this problem with an example from recruitment. Assume that X represents a candidate's skill, S is an attribute indicating different population groups, and Y denotes whether a human recruiter would consider the candidate qualified for a given position. Due to historical biases, the distribution of the binary response Y may be unbalanced across the groups. We aim to make a prediction, or equivalently, to take the decision to accept or reject a candidate. Our goal is to make predictions for the value of Y, or equivalently, to decide whether to accept or reject a candidate, in a way that is both accurate and fair. Specifically, under demographic parity, we aim to ensure that the probability of acceptance is the same across all groups.

Classical results show that when no fairness constraints are imposed, the classifier $g_y^{\text{Bayes}}(x) = \mathbb{1}\left\{f^{\text{Bayes}}(x) \geq y\right\}$ is a Bayes classifier that minimizes $\mathcal{R}_y(g)$. This relationship is at the heart of the design and study of plug-in classifiers (Yang, 1999; Massart and Ndlec, 2006; Audibert and Tsybakov, 2007; Biau et al., 2008). Interestingly, it was recently shown that a similar relationship holds under demographic parity constraints in the awareness framework Gaucher et al. (2023). Extending this result to the unawareness framework has remained an open problem, which we address in this paper.

Notation We first set some notation. Recall that we are given a tuple (X, S, Y) in $\mathcal{X} \times \mathcal{S} \times \mathbb{R}$ with distribution $\mathbb{P}$, where $\mathcal{X}$ is any measurable space (the space of features) and $\mathcal{S}$ is a finite set (the set of labels). For $s \in \mathcal{S}$, we denote by p_s the probability $\mathbb{P}(S = s)$ and by μ_s the conditional law of $X|S = s$. We let $\mu = \sum_{s\in\mathcal{S}} p_s\mu_s$ be the marginal distribution of X. We let $\mathcal{P}(\mathcal{X})$ be the set of probability measures on the measurable space $\mathcal{X}$. Moreover, we let $L^1(\nu)$ be the space of functions integrable with respect to the probability measure ν. Finally, $\mathring{C}$ denotes the interior of the set C.

1.2 Related work

Fair classification The majority of work on optimal prediction under demographic parity constraints has focused on classification tasks, where the goal is to predict a binary response in $\{0, 1\}$. Indeed, this problem is intrinsically linked to the issue of fair candidate selection, central in algorithmic fairness. This problem is well understood in the awareness setting from an algorithmic point of view (Feldman et al., 2015; Menon and Williamson, 2018; Yang et al., 2020; Schreuder and Chzhen, 2021; Zeng et al., 2022a; Denis et al., 2024). On the theoretical side, Gaucher et al. (2023) recently proved that the optimal classifier for the risk $\mathcal{R}_y$ can be obtained as the indicator that the optimal fair prediction function for the squared loss f^* is above the threshold y, a result that was later extended in Xian et al. (2023) to multi-class classification.

Less is known about fair classification in the unawareness framework. On the algorithmic side, several works have proposed various relaxations of the demographic parity constraint, leading to tractable algorithms for computing classifiers (Goh et al., 2016; Zafar et al., 2019; Oneto et al., 2020). On the theoretical side, Lipton et al. (2018) provided empirical evidence suggesting that fair classifiers may base their decisions on irrelevant features correlated with the sensitive attribute, potentially disrupting within-group ordering. This hypothesis was further confirmed by Gaucher et al. (2023), who characterized the optimal fair classifier in the unawareness framework. They showed that it is given by the indicator that the conditional expectation $\eta(X)$ is above a threshold, which depends on the probabilities that the individual described by X belongs to the different groups. Notably, the question of whether this classifier can be obtained by thresholding the optimal fair prediction function for the squared loss remains an open problem.

Fair regression In the awareness framework, fair regression is well understood both algorithmically and theoretically, see e.g., (Chzhen and Schreuder, 2020a; Gouic et al., 2020; Chzhen et al., 2020a) and also (Narasimhan et al., 2020), where a weaker notion of fairness related to fair ranking is studied. On the theoretical front, it has been shown that the problem of fair regression under demographic parity can be rephrased as the problem of finding the weighted barycenter of the distributions of $\eta(X,S) = \mathbb{E}[Y|X,S]$ across different groups, with costs given by optimal transport problems. Intuitively, the distribution of the fair predictions should remain close to the distributions of the optimal unfair predictions, $\eta(X,S) \mid S = s$, within each group. This ensures that the excess risk relative to these predictions remains as small as possible. Consequently, the problem is closely related to that of finding a barycenter of the distributions $\eta(X,S) \mid S = s$. More concretely, let ν_s be the distribution of the Bayes predictor $\eta(X,S)|S = s$ in the group s. Then, the optimal fair predictor f^* for an individual x in the group s is obtained by computing the quantile q of the Bayes predictor $\eta(x,s)$ in ν_s, and then predicting the weighted average of the qth quantiles of the distributions $\nu_{s'}$ over all groups s'. This result is summarized in the next theorem.

Theorem 4 (Chzhen and Schreuder (2020a); Gouic et al. (2020)) *Assume that for all $s \in \mathcal{S}$, the distribution ν_s of $\eta(X,S)$ for $S = s$ has no atoms, and let $p_s = \mathbb{P}(S = s)$. Then,*

$$\min_{f \text{ is fair}} \mathcal{R}_{sq}(f) = \min_{\nu \in \mathcal{P}(\mathbb{R})} \sum_{s \in \mathcal{S}} p_s \mathcal{W}_2^2(\nu_s, \nu)$$

where $\mathcal{W}_2^2(\nu_s, \nu)$ is the squared Wasserstein distance between ν_s and ν. Moreover, if f^* and ν solve the left-hand side and the right-hand side problems respectively, then ν is equal to the distribution of $f^*(X,S)$, and

$$f^*(x,s) = \left(\sum_{s' \in \mathcal{S}} p_{s'} \mathcal{Q}_{s'} \right) \circ F_s(\eta(x,s)),$$

where $\mathcal{Q}_s$ and F_s are respectively the quantile function and the c.d.f. of ν_s.

This result relates the problem of fair regression in the awareness framework to a more general optimal transport problem. Interestingly, this problem has an explicit solution, given by the quantiles and c.d.fs of the conditional expectation $\eta(X,S)$ across the different groups. This explicit formulation yields, as an immediate consequence, that the optimal fair regression function preserves order, a property introduced in Chzhen et al. (2020a); Chzhen and Schreuder (2020a) within the awareness framework. Recall that the Bayes regression function in the awareness framework is η. A prediction function f is said to *preserve order* if for any two candidates $(x,x') \in \mathcal{X}^2$ in the same group $s \in \mathcal{S}$, $\eta(x,s) \leq \eta(x',s)$ implies $f(x,s) \leq f(x',s)$. Thus, this property implies that the fairness correction does not alter the ordering of the predictions within a group.

In contrast, the problem of fair regression within the unawareness framework has seldom been studied, particularly from a theoretical perspective. One reason for this is that the demographic parity constraint is more challenging to implement without disparate treatment. Agarwal et al. (2019) propose an algorithm based on a discretization of the problem, followed by a reduction to cost-sensitive classification. However, their algorithm requires calling an oracle cost-sensitive classifier, which may not be available in practice. Additionally, their results are limited to a class of regression functions with bounded Rademacher complexity. Other algorithms complying with the demographic parity constraint have been proposed by Chzhen and Schreuder (2020b) and Zhou and Marecek (2023). However, the authors do not claim that the estimators obtained are optimal in terms of risk. Finally, in parallel with our work, Taturyan et al. (2024) proposed an algorithm for estimating a fair prediction function in the regression setting. In contrast to our approach, their method provides a computationally efficient algorithm that produces a piecewise constant *randomized* prediction function which is close to being optimal. By comparison, our work focuses on characterizing the theoretically optimal regression function and shows that, under mild assumptions, the optimal solution is deterministic. In this sense, the two approaches are complementary.

1.3 Outline and contribution

In this paper, we focus on the theoretical aspects of the problem of fair regression in the unawareness framework, specifically on characterizing and studying the optimal regression function. We extend results presented earlier in the awareness framework to this setting, albeit under the assumption that the sensitive attribute takes only two values; henceforth, we assume that $\mathcal{S} = \{1, 2\}$. Although restrictive, this assumption is not uncommon in the literature (Lipton et al., 2018) and covers the important case where one of the two groups includes protected individuals. Our results shed light on important phenomena, and we leave the extension to scenarios with more than two groups to future work. Below is a list of our main contributions:

- We derive the first theoretical characterization of the optimal fair regression function under demographic parity in the unawareness framework, as the solution of a barycenter problem with optimal transport costs.
- We show that, under mild assumptions, this optimal fair prediction is deterministic.
- We prove that, under mild assumptions, this predictor does not satisfy the desirable property of preserving order within groups.
- We prove that, under mild assumptions, the optimal fair classifier under demographic parity in the unawareness framework is not envy-free.
- We derive necessary and sufficient conditions for this classifier to be equivalent to the classifier obtained by thresholding the fair score function.

The paper is organized as follows. We begin in Section 2 with a brief introduction to optimal transport theory and to the main tools used in the proofs of our results. In Section 3, we characterize the optimal fair regression function. First, we prove in Proposition 8 that *in general, the optimal fair regression function f^* does not preserve order.* Next, we demonstrate the following result, which relates fair regression in the unawareness framework to an optimal transport problem.

Theorem 5 (Informal, see Theorem 12) *Under mild assumptions, the optimal fair regression function f^* is given by the solution to a barycenter problem with optimal transport costs. In particular, there exists a function $\boldsymbol{f}^*$ such that*

$$f^*(x) = \boldsymbol{f}^*(\eta(x), \Delta(x)),$$

where $\Delta(x) \propto \frac{\mathbb{P}(S=1|X=x)}{p_1} - \frac{\mathbb{P}(S=2|X=x)}{p_2}$.

This result shows that, similarly to the awareness case characterized in Theorem 4, the solution to the fair regression problem in the unawareness framework is given by the solution to a barycenter problem with optimal transport costs. In particular, the unawareness setting admits an *optimal deterministic* fair regressor. Deterministic predictors are often preferable in practice (Cotter et al., 2019); yet, to our knowledge, the only method with general theoretical guarantees currently yields *randomized* predictors (Taturyan et al., 2024).

Comparing Theorem 5 to the one provided in Theorem 4 within the awareness framework, we note that there is no explicit formula for the optimal fair regression function within the unawareness framework (as it is the case for solutions of optimal transport cost problems in more than one dimension). Moreover, Theorem 5 underscores that the optimal fair regression function effectively relies on an estimate $\Delta(X)$ of the unobserved sensitive attribute S to make predictions, thereby indirectly implementing disparate treatment. This result provides a theoretical explanation for the empirical phenomenon observed by Lipton et al. (2018). In this work, the authors argue that this behavior is problematic as it can lead to basing predictions on factors that do not have a direct causal effect on the outcome Y, but rather that have a causal effect on the sensitive attribute S.

In Section 4, we investigate the relationship between fair regression and fair classification when $Y \in \{0, 1\}$. We demonstrate the existence of a dichotomy based on a *nestedness* criterion. Recall that as the threshold y increases, the Bayes classifier g_y^{Bayes} predicts 1 for a decreasing proportion of candidates; we show that this also holds for the optimal fair classifier g_y^*. We say that the fair classification problem is *nested* if, almost surely with respect to the measure μ of X, the prediction $g_y^*(X)$ for the candidate X decreases as y increases. In other words, candidates rejected (i.e., with prediction 0) at low values of y cannot be accepted at higher values of y, when the proportion of accepted candidates is lower. For example, the Bayes classifier defined by $g_y^{\text{Bayes}}(x) = \mathbb{1}\left\{f^{\text{Bayes}}(x) \geq y\right\}$ satisfies this condition. When the nestedness criterion holds, the decision boundaries for the optimal fair classifier for different risk $\mathcal{R}_y$ form nested sets. The following informal result summarizes our findings.

Theorem 6 (Informal, see Proposition 21 and Corollary 23) *Under mild assumptions, if the fair classification problem is nested, then the regression function*

$$f^*(x) = \sup\left\{y \in \mathbb{R} \ : \ g_y^*(x) = 1\right\}$$

is optimal for the fair regression problem (1)*; equivalently, the classifier*

$$g_y(x) = \mathbb{1}\left\{f^*(x) \geq y\right\}$$

is optimal for the fair classification problem (2) *for the risk* $\mathcal{R}_y$*. Conversely, if the classification problem is not nested and if* f^* *is the optimal fair regression function, then there exists* $y \in (0,1)$ *such that*

$$g_y(x) = \mathbb{1}\{f^*(x) \geq y\}$$

is sub-optimal for the fair classification problem with risk $\mathcal{R}_y$*.*

While nestedness may initially appear to be a natural assumption, it does not always hold. In Section 5, we show how to design examples of problems where this assumption is either met or violated.

2. A short introduction to optimal transport

In this section, we provide a brief introduction to optimal transport. We present the main tools that will be used in the proofs of Sections 3 and 4. We begin by providing an overview of optimal transport in Section 2.1, before discussing the multi-to-one dimensional transport problem in Section 2.2.

2.1 The optimal transport problem

Optimal transport provides a mathematical framework to compare probability distributions. Consider a Borel probability measure μ on a Polish space $\mathcal{X}$ and a Borel probability measure ν on some other Polish space $\mathcal{Y}$. We are given a continuous cost function $c : \mathcal{X} \times \mathcal{Y} \to [0,+\infty]$, where $c(x,y)$ represents the cost of moving a unit of mass from $x \in \mathcal{X}$ to $y \in \mathcal{Y}$. The optimal transport problem consists of finding the optimal way to move the distribution of mass μ to ν by minimizing the total displacement cost. Formally, a transport map is a measurable map $T : \mathcal{X} \to \mathcal{Y}$ such that the pushforward measure $T\sharp\mu$ of μ by T is equal to ν, where the pushforward measure is defined for all measurable sets $B \subset \mathcal{Y}$ by

$$T\sharp\mu(B) = \mu(T^{-1}(B)).$$

The optimal transport problem is then the following

$$\begin{aligned} &\text{minimize} \quad \int c(x, T(x))\mathrm{d}\mu(x) \\ &\text{under the constraint } T\sharp\mu = \nu. \end{aligned} \tag{3}$$

The existence of minimizers of the optimization problem (3) is a delicate problem that depends on both the regularity of the cost function c and the properties of μ and ν. For instance, when $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $c(x,y) = \|x-y\|^2$, a solution exists whenever μ gives zero mass to sets of dimension smaller than $d-1$; otherwise, a solution may not exist, see (Villani, 2009, Chapter 10). When $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $c(x,y) = \|x-y\|^2$, the corresponding minimum is known as the (squared) Wasserstein distance between μ and ν, denoted by $\mathcal{W}_2^2(\mu,\nu)$. More generally, an optimal transport map exists whenever μ gives zero mass to sets of dimensions smaller than $d-1$ and the cost function $c(x,y) = \|x-y\|^2$ is replaced by any smooth cost function c satisfying the so-called *twist condition*, which states that the determinant $\det(\frac{\partial^2 c}{\partial y_j \partial x_i})$ never vanishes.

The optimal transport problem also admits a relaxed version in terms of transport plans, which is often more convenient. A transport plan is a probability measure π on the product space $\mathcal{X} \times \mathcal{Y}$ which has first marginal equal to μ and second marginal equal to ν: for all measurable sets $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$,

$$\pi(A \times \mathcal{Y}) = \mu(A), \quad \pi(\mathcal{X} \times B) = \nu(B),$$

or in probabilistic terms, if $(X, Y) \sim \pi$, then $X \sim \mu$ and $Y \sim \nu$. Informally, for $x \in \mathcal{X}$, the conditional law of $Y|X = x$ describes the different locations where the mass initially at x will be sent. The cost of a transport plan π is given by

$$\iint c(x, y) \mathrm{d}\pi(x, y).$$

Observe that a transport map T induces a transport plan by considering the law π of $(X, T(X))$ (formally, $\pi = (\mathrm{id}, T)\sharp\mu$). The optimal transport cost is defined by the following minimization problem:

$$\mathrm{OT}_c(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \int c(x, y) \mathrm{d}\pi(x, y), \tag{4}$$

where $\Pi(\mu, \nu)$ is the set of transport plans between μ and ν. Optimal transport plans always exist, whereas optimal transport maps may fail to do so. When optimal transport maps exist and the source measure μ has no atoms, the minimization problem (3) gives the same value as the optimal transport cost defined in (4), see Pratelli (2007).

Our proofs will rely heavily on the dual formulation of the optimal transport problem, which we now introduce. The c-transform of a function $\varphi : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ is defined as

$$\forall x \in \mathcal{X},\ \varphi^c(x) = \sup_{y \in \mathcal{Y}} (\varphi(y) - c(x, y)).$$

The subdifferential of φ is defined as

$$\partial_c \varphi = \{(x, y) \in \mathcal{X} \times \mathcal{Y} :\ \varphi(y) - \varphi^c(x) = c(x, y)\}.$$

For the quadratic cost, the notion of c-transform is related to the notion of convex conjugate.

Kantorovich duality (Villani, 2009, Theorem 5.10) states that

$$\mathrm{OT}_c(\mu, \nu) = \sup_{\varphi \in L^1(\nu)} \left(\int \varphi(y) \mathrm{d}\nu(y) - \int \varphi^c(x) \mathrm{d}\mu(x) \right).$$

Moreover, under the mild assumption that there exist two functions $a \in L^1(\mu)$ and $b \in L^1(\nu)$ such that $c(x, y) \leq a(x) + b(y)$ for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, the previous supremum is attained by a function φ, which we call a Kantorovich potential. In that case, any optimal transport π is supported on the subdifferential of the c-convex function φ, meaning that

$$\pi(\partial_c \varphi) = 1.$$

This last condition imposes significant constraints on the structure of optimal transport plans. For the quadratic cost, this fact is the key ingredient in proving that optimal transport plans are induced by optimal transport maps.

2.2 Multi-to-one dimensional optimal transport

In the next section, we demonstrate that the fair regression problem within the unawareness framework can be reduced to a barycenter problem of the form:

$$\min_{\nu \in \mathcal{P}(\mathbb{R})} p_1 \mathrm{OT}_c(\mu_1, \nu) + p_2 \mathrm{OT}_c(\mu_2, \nu), \tag{5}$$

where μ_1, μ_2 are *two-dimensional* probability measures and $c : \mathbb{R}^2 \times \mathbb{R} \to [0, +\infty)$ is a cost function. This reduction raises the question of whether the solutions to the barycenter problem (5) can be characterized by transport maps.

Proving that the optimal transport problem $\mathrm{OT}_c(\mu_s, \nu)$ is solved by a transport map is nontrivial. Complications arise because the measures μ_s and ν are defined on spaces of different dimensions. Optimal transport problems involving spaces of different dimensions have not been as extensively studied and exhibit distinct properties compared to the standard case where both measures are defined on the same space, see (Pass, 2012; Chiappori et al., 2016, 2017; McCann and Pass, 2020). For instance, the classical twist condition $\det(\frac{\partial^2 c}{\partial y_j \partial x_i}) \neq 0$ does not make sense in this setting: the Hessian matrix of c is not squared, so that the determinant is not even well-defined.

Chiappori et al. (2017) focus on the optimal transport problem between a measure μ supported on a domain $\mathcal{X} \subset \mathbb{R}^m$ (with $m > 1$) and a measure ν on an interval $\mathcal{Y} \subset \mathbb{R}$ for some cost function $c : \mathcal{X} \times \mathcal{Y} \to [0, +\infty)$. They demonstrate that an optimal transport map $T : \mathcal{X} \to \mathcal{Y}$ between μ and ν exists under a natural condition on (c, μ, ν) known as *nestedness*. For $y \in \mathcal{Y}$, $k \in \mathbb{R}$, let

$$\mathcal{X}_{\leq}(y, k) = \{x \in \mathcal{X} : \ \partial_y c(x, y) \leq k\}.$$

Kantorovich duality implies that an optimal transport plan between μ and ν will match an interval $(-\infty, y]$ to a set $\mathcal{X}_{\leq}(y, k)$, where $k = k(y)$ is a solution of the equation $\nu((-\infty, y]) = \mu(\mathcal{X}_{\leq}(y, k))$. The triplet (c, μ, ν) is called nested if the collection of sets $(\mathcal{X}_{\leq}(y, k(y)))_y$ increases with y. Chiappori et al. (2017) prove that an optimal transport map T between μ and ν exists when the problem is nested: informally, the monotonicity of $(\mathcal{X}_{\leq}(y, k(y)))_y$ ensures that a given $x_0 \in \mathcal{X}$ belongs to the boundary of a single set $\mathcal{X}_{\leq}(y_0, k(y_0))$, with y_0 being equal to $T(x_0)$.

This nestedness condition will be crucial in Section 4, where it will be used to establish the equivalence between regression and classification problems. However, in Section 3, we will be able to show the existence of optimal transport maps for the barycenter problem (5) (and consequently of optimal fair regression functions) without any nestedness condition.

3. Fair regression and the barycenter problem

In this section, we characterize the solution to the fair regression problem. We begin by showing in Section 3.1 that, under mild assumptions, the fair regression function does not preserve order. Then, in Section 3.2, we show that the fair regression problem can be reduced to a barycenter problem with optimal transport costs. Using the tools introduced in Section 2, we prove the existence of a fair optimal prediction function and study some of its properties.

3.1 Fair regression functions do not preserve order

Before analyzing in detail the fair regression problem in the unawareness framework, we establish a simple yet important property of fair regression functions. We begin by extending the definition of order preservation (Chzhen et al., 2020a; Chzhen and Schreuder, 2020a) to the unawareness framework. Recall that in this case, the Bayes prediction for a candidate x is given by $\eta(x) = \mathbb{E}[Y|X = x]$. A prediction function f is said to preserve order if, for any two candidates $(x, x') \in \mathcal{X}^2$ in the same group $s \in \mathcal{S}$, $\eta(x) \leq \eta(x')$ implies $f(x) \leq f(x')$. This definition is formalized below.

Definition 7 (Order preservation in regression - unawareness framework) *A prediction function f preserves order if $\mathbb{P} \otimes \mathbb{P}$-almost surely,*

$$\big\{\eta(X) < \eta(X') \text{ and } S = S'\big\} \implies f(X) < f(X').$$

This property implies that the fairness correction does not alter the ordering of the predictions of the Bayes prediction function within a group. It is related to the concept of "rational ordering" introduced by Lipton et al. (2018) in the context of classification, where the authors require that within a group, the most able candidates are the ones accepted.

Proposition 8 *Let $f : \mathcal{X} \to \mathbb{R}$ be a regression function with $\mathbb{E}[f(X)^2] < \infty$ satisfying the demographic parity constraint. Assume that the Bayes regression function η does not satisfy the demographic parity constraint and that $\mathbb{P}(S = s|X = x) \in (0, 1)$ for all $s \in \mathcal{S}$, $x \in \mathcal{X}$. Then, f does not preserve order.*

Proof We prove the contrapositive: if f is a regression function satisfying the demographic parity constraint and preserving order, then the Bayes regression function also satisfies the demographic parity constraint. For a fixed group $s \in \mathcal{S}$, consider the joint law π_s of $(\eta(X), f(X))$, where $X \sim \mu_s$. As f is order-preserving, the support of the measure π_s is monotone, in the sense that

$$\forall (y_1, z_1), (y_2, z_2) \in \text{supp}(\pi_s),\ y_1 < y_2 \implies z_1 < z_2. \tag{6}$$

According to (Santambrogio, 2015, Lemma 2.8), this implies that π_s is actually the optimal transport plan for the quadratic cost between the first marginal of π_s, equal to $\eta\sharp\mu_s =: \nu_s$ and the second marginal of π_s, equal to $f\sharp\mu_s =: \nu$ (the second measure does not depend on s because of demographic parity). We claim that strict monotonicity implies that the transport plan π_s takes the form of a transport map T_s transporting ν towards ν_s, that is $\pi_s = (T_s, \text{id})\sharp\nu$ (see a proof below). So, if $X \sim \mu_s$, we have $(\eta(X), f(X)) = (T_s(f(X)), f(X))$ almost surely. To put it another way, we have for every s,

$$\eta(x) = T_s \circ f(x) \quad \mu_s\text{-almost everywhere.}$$

As $\mathbb{P}(S = s|X = x) > 0$ for all $x \in \mathcal{X}$, this equality is also satisfied μ-almost everywhere. Hence, for μ-almost all x, the quantity $T_s \circ f(x)$ does not depend on s (it is equal to $\eta(x)$). This defines a function T with $\eta\sharp\mu_s = T\sharp f\sharp\mu_s = T\sharp\nu$. As this measure does not depend on s, this proves that η satisfies the demographic parity constraint.

To conclude our proof, it remains to prove our claim. Decompose ν as $\nu_1+\nu_2$ where ν_2 is atomless and $\nu_1 = \sum_j p_j \delta_{z_j}$. If $f(X) = z_j$, then we have $\eta(X) = y_j$ for some value y_j: this value y_j has to be unique, for otherwise it would contradict the monotonicity assumption. Therefore, ν_s can be written as $\nu_s = \nu_{1s} + \nu_{2s}$, where $\nu_{1s} = \sum_j p_j \delta_{y_j}$. Consider the plan $\pi_1 = \sum_j p_j \delta_{(y_j,z_j)}$. Then $\pi - \pi_1$ is a plan between ν_{2s} and ν_2. By (Santambrogio, 2015, Lemma 2.8), as ν_2 is atomless, the monotonicity condition implies that it is induced by a transport map $\tilde{T}_s$. In total, we can define T_s by $T_s(z_j) = y_j$ and by $T_s = \tilde{T}_s$ on the complementary set of the atoms. ■

Proposition 8 implies, in particular, that in many instances, the optimal fair regression function does not preserve order. Consequently, highly qualified individuals who belong to protected groups could potentially suffer from fairness corrections due to the demographic parity constraint.

The condition that $\mathbb{P}(S = s|X = x) \in (0, 1)$ for all $s \in \mathcal{S}$, $x \in \mathcal{X}$ ensures that the sensitive attribute S cannot be determined from the observation of X. When this condition is not satisfied, the distinction between the unawareness and awareness frameworks becomes blurred: if S can be inferred from X alone, it becomes meaningless to differentiate between a regression function that depends on both X and S, and one that depends solely on X. Furthermore, it is important to note that in the awareness framework, there do exist regression functions that satisfy the demographic parity constraint and preserve order, with the optimal fair regression function described in Theorem 4 being one such example.

3.2 Reduction to an optimal transport problem

In the following, we let $\mathcal{S} = \{1, 2\}$. We assume that $\mu_1 \neq \mu_2$ (otherwise the Bayes regression function η already solves the fair regression problem). We now show how to transform the fair regression problem into a barycenter problem using optimal transport costs. To do so, we first leverage a reformulation of the demographic parity constraint due to Chzhen and Schreuder (2020b), which is based on the Jordan decomposition of the signed measure $\mu_1 - \mu_2$. Then, we show how to rephrase the regression problem as a barycenter problem, using this new constraint. Finally, we show that, under mild assumptions, the barycenter problem admits a unique solution, which is given by a transport map.

3.2.1 Reformulation of the demographic parity constraint

In the unaware setting, since we do not have access to the sensitive group S, it will be convenient to partition the space $\mathcal{X}$ into two subsets: the subset $\mathcal{X}_+$ where it is more likely that an individual $x \in \mathcal{X}_+$ belongs to the group 1, and the subset $\mathcal{X}_-$ where it is more likely that an individual $x \in \mathcal{X}_-$ belongs to the group 2. Formally, this partition is obtained by considering the Jordan decomposition of the measure $\mu_1 - \mu_2$. We will then define separate prediction functions on $\mathcal{X}_+$ and $\mathcal{X}_-$. Let $|\mu_1 - \mu_2|$ be the variation of $\mu_1 - \mu_2$ and define

$$\begin{cases} (\mu_1 - \mu_2)_+ = \frac{1}{2}(|\mu_1 - \mu_2| + \mu_1 - \mu_2), \\ (\mu_1 - \mu_2)_- = \frac{1}{2}(|\mu_1 - \mu_2| - \mu_1 + \mu_2) \end{cases}$$

the Jordan decomposition of $\mu_1 - \mu_2$. The two measures $(\mu_1 - \mu_2)_+$ and $(\mu_1 - \mu_2)_-$ have the same mass, which we denote by m. We define the scaled Jordan decomposition of $\mu_1 - \mu_2$

as the pair of probability measures

$$\mu_+ = (\mu_1 - \mu_2)_+/m \quad \text{and} \quad \mu_- = (\mu_1 - \mu_2)_-/m.$$

Let $\frac{\mathrm{d}\mu_+}{\mathrm{d}\mu}$ (resp. $\frac{\mathrm{d}\mu_-}{\mathrm{d}\mu}$) be the density of μ_+ (resp. μ_-) with respect to μ (that are defined uniquely μ-almost everywhere). As μ_+ and μ_- are mutually singular measures, we can always find versions of $\frac{\mathrm{d}\mu_+}{\mathrm{d}\mu}$ and $\frac{\mathrm{d}\mu_-}{\mathrm{d}\mu}$ such that the sets

$$\begin{cases} \mathcal{X}_+ = \{x \in \mathcal{X} : \frac{\mathrm{d}\mu_+}{\mathrm{d}\mu}(x) > 0\}, \\ \mathcal{X}_- = \{x \in \mathcal{X} : \frac{\mathrm{d}\mu_-}{\mathrm{d}\mu}(x) > 0\}, \\ \mathcal{X}_= = \mathcal{X} \backslash (\mathcal{X}_+ \sqcup \mathcal{X}_-). \end{cases}$$

form a partition of $\mathcal{X}$, with μ_+ giving mass 1 to $\mathcal{X}_+$ and μ_- giving mass 1 to $\mathcal{X}_-$. Then, for for any three functions f_+, f_-, and $f_=$ from $\mathcal{X}$ to $\mathbb{R}$, we can define the associated function $\mathcal{F}(f_+, f_-, f_=)$ equal to f_+ on $\mathcal{X}_+$, f_- on $\mathcal{X}_-$, and $f_=$ on $\mathcal{X}_=$:

$$\mathcal{F}(f_+, f_-, f_=)(x) = \begin{cases} f_+(x) & \text{if } x \in \mathcal{X}_+ \\ f_-(x) & \text{if } x \in \mathcal{X}_- \\ f_=(x) & \text{if } x \in \mathcal{X}_=. \end{cases}$$

Conversely, for any function $f : \mathcal{X} \to \mathbb{R}$, there exist functions f_+, f_-, and $f_=$ corresponding respectively to the restriction of f on $\mathcal{X}_+$, $\mathcal{X}_-$, and $\mathcal{X}_=$, i.e., such that $f = \mathcal{F}(f_+, f_-, f_=)$. The following lemma, due to Chzhen and Schreuder (2020b), rephrases the demographic parity constraint in terms of μ_+ and μ_-.

Lemma 9 *A regression function $f : \mathcal{X} \to \mathbb{R}$ satisfies the demographic parity constraint if and only if*

$$f\sharp\mu_+ = f\sharp\mu_-.$$

Lemma 9 reveals that for any functions f, and f_+, f_-, $f_=$ such that $f = \mathcal{F}(f_+, f_-, f_=)$, the regression function f satisfies the demography parity constraint if and only if $f_+\sharp\mu_+ = f_-\sharp\mu_-$. The two functions f_+ and f_- can be chosen with disjoint support (in $\mathcal{X}_+$ and $\mathcal{X}_-$, respectively). Thus, the demographic parity constraint essentially reduces to the equality of the pushforward measures of two distinct probabilities (μ_+ and μ_-) by two distinct functions (f_+ and f_-).

3.2.2 A barycenter problem

In order to rephrase the regression problem as a barycenter problem, we introduce further notation. We define

$$\begin{cases} \Delta(x) = \frac{\mathrm{d}\mu_+}{\mathrm{d}\mu}(x) & \text{if } x \in \mathcal{X}_+, \\ \Delta(x) = -\frac{\mathrm{d}\mu_-}{\mathrm{d}\mu}(x) & \text{if } x \in \mathcal{X}_-, \\ \Delta(x) = 0 & \text{if } x \in \mathcal{X}_=. \end{cases} \tag{7}$$

Equivalently, $\Delta(x)$ is proportional to $\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu}(x) - \frac{\mathrm{d}\mu_2}{\mathrm{d}\mu}(x)$. Observe that in the particular case where the sensitive attribute S can be perfectly determined from X, $\Delta(x)$ takes two different values: it is equal to p_1 for $x \in \mathcal{X}_+$ and equal to p_2 for $x \in \mathcal{X}_-$. More generally, the function

Δ partitions the space $\mathcal{X}$ into group of individuals that are equally likely to belong to a given group. These individuals should be treated in a similar manner by a fair unaware algorithm.

We define the cost function $c : \mathcal{X} \times \mathbb{R} \to [0, +\infty]$ given by $c(x, y) = \frac{(\eta(x)-y)^2}{|\Delta(x)|}$ for all $x \in \mathcal{X}$ and all $y \in \mathbb{R}$. When restricting the cost to individuals that are equally likely to belong to a given group, this cost function is simply a rescaling of the usual quadratic cost. When $X \sim \mu_\pm$, the variables $(\eta(X), \Delta(X))$ belong to the domain $\Omega := \{(h, d) \in \mathbb{R}^2 : \ d \neq 0\}$. In the following, we use bold notation to denote functions related to these two-dimensional variables. For example, we denote by $\boldsymbol{\mu}_+$ (resp. $\boldsymbol{\mu}_-$) the distributions of $(\eta(X), \Delta(X))$ when X follows the distribution μ_+ (resp. μ_-). Observe that the support of $\boldsymbol{\mu}_+$ is included in the upper half-plane $\{d > 0\}$ while $\boldsymbol{\mu}_-$ is included in the lower half-plane $\{d < 0\}$. We define the two-to-one dimensional cost $\mathbf{c}$, given by $\mathbf{c}(\mathbf{z}, y) = \frac{(h-y)^2}{|d|}$ for all $\mathbf{z} = (h, d) \in \Omega$ and $y \in \mathbb{R}$. Introducing this modified cost function will prove to be convenient when we will use theoretical results from optimal transport theory, as, unlike the initial cost function c, the cost function $\mathbf{c}$ is a smooth function on a Euclidean domain.

Consider the barycenter problem

$$\text{minimize} \quad \mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_+, \nu) + \mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_-, \nu) \text{ over } \nu \in \mathcal{P}(\mathbb{R}) \tag{8}$$

where we recall that $\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_\pm, \nu)$ is the optimal transport cost for sending $\boldsymbol{\mu}_\pm$ to ν with cost function $\mathbf{c}$, defined in Equation (4). In the situation where S can be perfectly determined from X, the objective in the barycenter problem (8) is equal to $p_1 W_2^2(\mu_1, \nu) + p_2 W_2^2(\mu_2, \nu)$, that is we find a standard Wasserstein barycenter problem. In general, we can think about (8) as a continuous Wasserstein barycenter problem where we consider a continuous family of groups indexed by $\Delta(x) = d$, that is, we group together individuals which are equally likely to belong to a given group. In the aware setting the fair prediction depends on both the Bayes predictor and on the group s. In the unaware setting, the fair prediction will depend on both the Bayes predictor and on the "group" represented by the value of $\Delta(x)$: in short, we will show that the optimal fair predictor at x is a function of $(\eta(x), \Delta(x))$.

We say that a solution ν^{bar} of the barycenter problem is solved by optimal transport maps if

$$\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_\pm, \nu^{\mathrm{bar}}) = \int \mathbf{c}(\mathbf{z}, \boldsymbol{f}_\pm(\mathbf{z})) \mathrm{d}\boldsymbol{\mu}_\pm(\mathbf{z})$$

for some transport maps $\boldsymbol{f}_\pm : \Omega \to \mathbb{R}$ from $\boldsymbol{\mu}_\pm$ to ν^{bar}.

Lemma 10 *There is a one-to-one correspondence between the set of solutions to the barycenter problem* (8) *solved by optimal transport maps and the set of optimal fair regression functions solving* (1)*. This correspondence associates a barycenter ν^{bar} with optimal transport maps $\boldsymbol{f}_\pm$ to the optimal fair regression function $f = \mathcal{F}(f_+, f_-, \eta)$, where $f_\pm(x) = \boldsymbol{f}_\pm(\eta(x), \Delta(x))$ for $x \in \mathcal{X}$.*

Proof Classical computations show that $\mathcal{R}_{sq}(f) = \mathbb{E}\left[(\eta(X) - f(X))^2\right] + \mathbb{E}\left[(\eta(X) - Y)^2\right]$. Thus, minimizing the risk is equivalent to minimizing $\mathbb{E}\left[(\eta(X) - f(X))^2\right]$. Now, as the space $\mathcal{X}$ is partitioned into $\mathcal{X}_+$, $\mathcal{X}_-$ and $\mathcal{X}_=$,

$$\begin{aligned}\mathbb{E}\left[(\eta(X)-f(X))^2\right] &= \int(\eta(x)-f(x))^2\mathrm{d}\mu(x)\\ &= \int_{\mathcal{X}_+}(\eta(x)-f(x))^2\frac{\mathrm{d}\mu}{\mathrm{d}\mu_+}(x)\mathrm{d}\mu_+(x)+\int_{\mathcal{X}_-}(\eta(x)-f(x))^2\frac{\mathrm{d}\mu}{\mathrm{d}\mu_-}(x)\mathrm{d}\mu_-(x)\\ &\quad+\int_{\mathcal{X}_=}(\eta(x)-f(x))^2\mathrm{d}\mu(x).\end{aligned} \tag{9}$$

Using the definition of Δ along with Lemma 9, we see that any solution to the fair regression problem can be written as $f=\mathcal{F}(f_+,f_-,\eta)$, where (f_+,f_-) is solution to the problem

$$\begin{aligned}&\text{minimize} \quad \int_{\mathcal{X}_+}\frac{(\eta(x)-f_+(x))^2}{|\Delta(x)|}\mathrm{d}\mu_+(x)+\int_{\mathcal{X}_-}\frac{(\eta(x)-f_-(x))^2}{|\Delta(x)|}\mathrm{d}\mu_-(x)\\ &\text{such that} \quad f_+\sharp\mu_+=f_-\sharp\mu_-.\end{aligned} \tag{10}$$

The triplet $(\eta(X),\Delta(X),f_+(X))$ for $X\sim\mu_+$ defines a coupling π_{f+} between $\boldsymbol{\mu}_+$ and $\nu_{f+}=f_+\sharp\mu_+$. Likewise, we define a coupling π_{f-} between $\boldsymbol{\mu}_-$ and ν_{f-}. Using that the space $\mathcal{X}$ is partitioned into $\mathcal{X}_+$, $\mathcal{X}_-$ and $\mathcal{X}_=$, we can rewrite (9) as

$$\begin{aligned}\mathbb{E}\left[(\eta(X)-\mathcal{F}(f_+,f_-,\eta)(X))^2\right] &= \int\mathbf{c}(\mathbf{z},y)\mathrm{d}\pi_{+f}(\mathbf{z},y)+\int c(\mathbf{z},y)\mathrm{d}\pi_{-f}(\mathbf{z},y)\\ &\geq \mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_+,\nu_{f+})+\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_-,\nu_{f-}).\end{aligned}$$

The constraint $f_+\sharp\mu_+=f_-\sharp\mu_-$ implies that $\nu_{f+}=\nu_{f-}$. Hence,

$$\inf_{f\text{ fair}}\mathbb{E}\left[(\eta(X)-f(X))^2\right]\geq\inf_{\nu\in\mathcal{P}(\mathbb{R})}\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_+,\nu)+\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_-,\nu). \tag{11}$$

Reciprocally, assume that there exists ν^{bar} solving the above barycenter problem, and that an optimal transport map between $\boldsymbol{\mu}_+$ and ν^{bar} is given by an application $\boldsymbol{f}_+:\Omega\to\mathbb{R}$, with $(\boldsymbol{f}_+)\sharp\boldsymbol{\mu}_+=\nu^{\mathrm{bar}}$. Likewise, we assume that there exists an optimal transport map $\boldsymbol{f}_-$ between $\boldsymbol{\mu}_-$ and ν^{bar}. Then, $\boldsymbol{f}_-\sharp\boldsymbol{\mu}_-=\boldsymbol{f}_+\sharp\boldsymbol{\mu}_+=\nu^{\mathrm{bar}}$. Defining $f_\pm(x)=\boldsymbol{f}_\pm(\eta(x),\Delta(x))$, we have $f_-\sharp\mu_-=f_+\sharp\mu_+=\nu^{\mathrm{bar}}$, and so $\mathcal{F}(f_+,f_-,\eta)$ is a fair regression function. Also, we have by optimality that

$$\begin{aligned}\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_+,\nu^{\mathrm{bar}})+\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_-,\nu^{\mathrm{bar}}) &= \int c(x,f_+(x))\mathrm{d}\mu_+(x)+\int c(x,f_-(x))\mathrm{d}\mu_-(x)\\ &= \mathbb{E}[(\eta(X)-\mathcal{F}(f_+,f_-,\eta)(X))^2].\end{aligned}$$

Hence, by (11), the regression function $\mathcal{F}(f_+,f_-,\eta)$ is optimal. This shows that the the inequality in (11) is actually an equality, and that any such barycenter ν yields an optimal fair regression function. Reciprocally, let (f_+^*,f_-^*) be a solution of the problem (10) and define $\pi_{\pm f^*}$ are transport plans between $\mu_\pm$ and $\nu^*=f_+^*\sharp\mu_+=f_-^*\sharp\mu_-$. The objective of the problem at (f_+^*,f_-^*) is equal to

$$\int\mathbf{c}(\mathbf{z},y)\mathrm{d}\pi_{+f^*}(\mathbf{z},y)+\int c(\mathbf{z},y)\mathrm{d}\pi_{-f^*}(\mathbf{z},y)\geq\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_+,\nu^*)+\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_-,\nu^*).$$

The fact that (11) is actually an equality then shows that ν^* is a barycenter and that the plans $\pi_{\pm f^*}$ are optimal transport plans. This concludes the proof of Lemma 10. ■

3.2.3 Transport maps for the barycenter problem

The rest of this section is devoted to proving that the barycenter problem indeed admits a solution given by transport maps, which will imply that there exists a solution to the fair regression problem. We show that this holds under the following mild regularity assumption.

$$\text{The measures } \boldsymbol{\mu}_+ \text{ and } \boldsymbol{\mu}_- \text{ give zero mass to graphs of functions in the sense that for any measurable function } F:\mathbb{R}\backslash\{0\}\to\mathbb{R},\ \boldsymbol{\mu}_\pm(\{(F(d),d):\ d\neq 0\})=0. \tag{A}$$

By Fubini's theorem, this assumption is trivially satisfied if $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ have a density with respect to the Lebesgue measure. Another interesting example is given by the awareness framework, seen as a particular instance of the unawareness framework.

Remark 11 (Awareness as a special case of unawareness) *Consider a triplet of random variable $(X,S,Y)\sim\mathbb{P}$, where $X\in\mathcal{X}$ is a feature, $S\in\{1,2\}$ is a sensitive attribute and $Y\in\mathbb{R}$ is a response variable of interest. Let $Z=(X,S)$ and let $\mathbb{Q}$ be the law of the triplet (Z,S,Y). Then, there is an equivalence between considering an aware regression function $f(X,S)$ under law $\mathbb{P}$ and an unaware regression function $f(Z)$ under law $\mathbb{Q}$. Observe that Z is a random variable on $\tilde{\mathcal{X}}=\mathcal{X}\times\{1,2\}$. The laws μ_1 of $Z|S=1$ and μ_2 of $Z|S=2$ have disjoint support. It follows that $\mathcal{X}_+=\mathcal{X}\times\{1\}$ with $\mu_+=\mu_1$ and $\mathcal{X}_-=\mathcal{X}\times\{2\}$ with $\mu_-=\mu_2$. Then, $\Delta(x)=1/p_1$ if $x\in\mathcal{X}_+$ and $\Delta(x)=-1/p_2$ if $x\in\mathcal{X}_-$. In particular, both measures $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ are supported on horizontal lines in Ω.*

In that case, Assumption A is equivalent to the fact that $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ have no atoms, which is exactly equivalent to the fact that the law of $\eta(X)$ (for $X\sim\mu$) has no atoms. This assumption is often considered to be a minimal assumption to ensure the existence of optimal fair regression functions in the awareness framework. Hence, Assumption A constitutes a generalization of this assumption to the unawareness framework.

Theorem 12 *Assume that $(X,Y,S)\sim\mathbb{P}$ is such that $\mathbb{E}[Y^2]<\infty$. Under Assumption A, there is a unique minimizer ν^{bar} of the barycenter problem*

$$\inf_{\nu\in\mathcal{P}(\mathbb{R})}\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_+,\nu)+\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_-,\nu). \tag{12}$$

Moreover, this problem is solved by optimal transport maps $\boldsymbol{f}_\pm$. In particular, there exists a unique solution f^* of the regression problem under the demographic parity constraint (1), *which is given by*

$$\forall x\in\mathcal{X},\ \ f^*(x)=\mathcal{F}\left(\boldsymbol{f}_+\big(\eta(x),\Delta(x)\big),\boldsymbol{f}_-\big(\eta(x),\Delta(x)\big),\eta(x)\right).$$

Proof Using Lemma 10, it is enough to show that the barycenter problem admits a unique solution ν^{bar} such that the corresponding transport problems $\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_+,\nu^{\mathrm{bar}})$ and $\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_-,\nu^{\mathrm{bar}})$ are solved by transport maps.

Step 1: reduction to a standard transport problem. We begin by reducing the barycenter problem (12) to a single two-to-two dimensional optimal transport problem $\mathrm{OT}_{\mathbf{C}}(\boldsymbol{\mu}_+,\boldsymbol{\mu}_-)$. The multimarginal version of the barycenter problem reads

$$\inf_{\nu\in\mathcal{P}(\mathbb{R})}\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_+,\nu)+\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_-,\nu)=\inf_{\rho\in\Pi(\cdot,\boldsymbol{\mu}_+,\boldsymbol{\mu}_-)}\int(\mathbf{c}(\mathbf{z}_1,y)+\mathbf{c}(\mathbf{z}_2,y))\mathrm{d}\rho(y,\mathbf{z}_1,\mathbf{z}_2), \tag{13}$$

where $\Pi(\cdot, \boldsymbol{\mu}_+, \boldsymbol{\mu}_-)$ stands for the set of measures on $\mathbb{R} \times \Omega \times \Omega$ with second marginal $\boldsymbol{\mu}_+$ and third marginal $\boldsymbol{\mu}_-$. Indeed, if $\rho \in \Pi(\cdot, \boldsymbol{\mu}_+, \boldsymbol{\mu}_-)$, then its two first marginals provide a transport plan between its first marginal ν and $\boldsymbol{\mu}_-$, while the first and last marginals provide a transport plan between ν and $\boldsymbol{\mu}_+$. This proves that the left-hand side of Equation (13) is smaller than the right-hand side. For the other inequality, consider $\nu \in \mathcal{P}(\mathbb{R})$, with associated optimal transport plans $\pi_+ \in \Pi(\boldsymbol{\mu}_+, \nu)$ and $\pi_- \in \Pi(\boldsymbol{\mu}_-, \nu)$. By the gluing lemma (see, e.g., Lemma 5.5 in Santambrogio (2015)), there exists $\rho \in \Pi(\cdot, \boldsymbol{\mu}_+, \boldsymbol{\mu}_-)$ such that the joint law of the first two marginals is equal to π_+, and the joint law of the first and last marginal is equal to π_-. Then, $\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_+, \nu) + \mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_-, \nu) = \int (\mathbf{c}(\mathbf{z}_1, y) + \mathbf{c}(\mathbf{z}_2, y)) \mathrm{d}\rho(y, \mathbf{z}_1, \mathbf{z}_2)$, proving that the right-hand side is smaller than the left-hand side in (13). This shows the validity of (13).

Furthermore, if ρ solves the right-hand side of (13), then its first marginal ν is a barycenter. Actually, by optimality, for any $(y, \mathbf{z}_1, \mathbf{z}_2)$ in the support of the optimal ρ, the point y necessarily minimizes the function $z \mapsto \mathbf{c}(\mathbf{z}_1, z) + \mathbf{c}(\mathbf{z}_2, z)$. Let us compute this minimizer. For $\mathbf{z}_1 = (h_1, d_1)$ and $\mathbf{z}_2 = (h_2, d_2)$, we have

$$\mathbf{c}(\mathbf{z}_1, y) + \mathbf{c}(\mathbf{z}_2, y) = (h_1 - y)^2/|d_1| + (h_2 - y)^2/|d_2|.$$

This function is convex in y. The first order condition for optimality reads

$$(y - h_1)/|d_1| + (y - h_2)/|d_2| = 0 \iff y = m(\mathbf{z}_1, \mathbf{z}_2) := \frac{h_1/|d_1| + h_2/|d_2|}{1/|d_1| + 1/|d_2|}. \tag{14}$$

Moreover, the cost $\mathbf{C}(\mathbf{z}_1, \mathbf{z}_2) := \inf_y \mathbf{c}(\mathbf{z}_1, y) + \mathbf{c}(\mathbf{z}_2, y)$ corresponding to this minimum is equal to

$$\mathbf{C}(\mathbf{z}_1, \mathbf{z}_2) = \frac{(h_2 - h_1)^2}{|d_1| + |d_2|}. \tag{15}$$

These considerations show that

$$\inf_{\nu \in \mathcal{P}(\mathbb{R})} \mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_+, \nu) + \mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_-, \nu) = \mathrm{OT}_{\mathbf{C}}(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-), \tag{16}$$

and that optimal transport plans $\pi^* \in \Pi(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-)$ are in correspondence with barycenters ν through the formula $\nu = m \sharp \pi^*$. In particular, as there exists at least one optimal transport plan, the infimum in the barycenter problem is actually a minimum.

Step 2: existence of a transport map. Observe that

$$\mathbf{C}(\mathbf{z}_1, \mathbf{z}_2) = \frac{(h_1 - h_2)^2}{|d_1| + |d_2|} \leq 2\frac{h_1^2}{|d_1|} + 2\frac{h_2^2}{|d_2|}.$$

This quantity is integrable against $\boldsymbol{\mu}_+ \otimes \boldsymbol{\mu}_-$. Indeed,

$$\int \frac{h_1^2}{|d_1|} \mathrm{d}\boldsymbol{\mu}_+(h_1, d_1) = \int \frac{\eta(x)^2}{\Delta(x)} \mathrm{d}\mu_+(x) = \int_{\mathcal{X}_+} \eta(x)^2 \mathrm{d}\mu(x) \leq \mathbb{E}[\mathbb{E}[Y|X]^2] \leq \mathbb{E}[Y^2] < \infty.$$

In particular, the optimal cost $\mathrm{OT}_{\mathbf{C}}(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-)$ is finite. Hence, by Kantorovich duality (see Section 2), there is a $\mathbf{C}$-convex function (called a Kantorovich potential) $\varphi : \Omega \to \mathbb{R} \cup \{+\infty\}$ such that if we let

$$\Gamma = \{(\mathbf{z}_1, \mathbf{z}_2) : \ \varphi(\mathbf{z}_1) - \varphi^{\mathbf{C}}(\mathbf{z}_2) = \mathbf{C}(\mathbf{z}_1, \mathbf{z}_2)\} \tag{17}$$

be the subdifferential of φ, then *any* optimal transport plan π satisfies $\pi(\Gamma) = 1$, see Section 2. We show in Appendix A.1 the following lemma.

Lemma 13 *Let $\varphi : \Omega \to \mathbb{R} \cup \{+\infty\}$ be a $\mathbf{C}$-convex function with $\mathrm{dom}(\varphi) := \{\mathbf{z} :\ \varphi(\mathbf{z}) < +\infty\}$. Then, the set of points $\mathbf{z} \in \mathrm{dom}(\varphi)$ such that the partial derivative $\partial_h \varphi(\mathbf{z})$ does not exist is included in a countable union of graphs of measurable functions $F : d \in \mathbb{R}\backslash\{0\} \mapsto F(d) \in \mathbb{R}$.*

Let Σ be the countable union of graphs given by Lemma 13 for the Kantorovich potential φ. According to Assumption A, if we let $\Omega_0 = \Omega \backslash \Sigma$, then $\boldsymbol{\mu}_+(\Omega_0) = 1$.

Let $\mathbf{z}_1 \in \Omega_0$ and let $(\mathbf{z}_1, \mathbf{z}_2) \in \Gamma$. Consider the function $g_{\mathbf{z}_2} : \mathbf{z} \in \Omega \mapsto \varphi(\mathbf{z}) - \mathbf{C}(\mathbf{z}, \mathbf{z}_2)$. As $\varphi^{\mathbf{C}}(\mathbf{z}_2) = \varphi(\mathbf{z}_1) - \mathbf{C}(\mathbf{z}_1, \mathbf{z}_2)$, by definition of the $\mathbf{C}$-transform, the function $g_{\mathbf{z}_2}$ attains its maximum at $\mathbf{z}_1$. In particular, as $\partial_{h_1}\varphi(\mathbf{z}_1)$ exists by assumption, we have

$$\partial_h \varphi(\mathbf{z}_1) = \partial_{h_1} \mathbf{C}(\mathbf{z}_1, \mathbf{z}_2) = \frac{2(h_1 - h_2)}{|d_1| + |d_2|}.$$

This implies that

$$h_2 = h_1 - \frac{|d_1| + |d_2|}{2} \partial_{h_1} \varphi(\mathbf{z}_1).$$

Using this expression, we find that

$$m(\mathbf{z}_1, \mathbf{z}_2) = h_1 - \frac{|d_1| \partial_{h_1} \varphi(\mathbf{z}_1)}{2}.$$

In particular, $m(\mathbf{z}_1, \mathbf{z}_2)$ is uniquely determined by $\mathbf{z}_1$. This defines a measurable map $\mathbf{z}_1 \in \Omega_0 \mapsto \boldsymbol{f}_+(\mathbf{z}_1)$. We extend $\boldsymbol{f}_+$ on Ω by setting $\boldsymbol{f}_+(\mathbf{z}_1) = 0$ if $\mathbf{z}_1 \in \Omega \backslash \Omega_0$. As explained in **Step 1**, for $(\mathbf{z}_1, \mathbf{z}_2) \sim \pi^*$, the law ν of $m(\mathbf{z}_1, \mathbf{z}_2)$ solves the barycenter problem $\mathrm{OT}_{\mathbf{C}}(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-)$. Hence, $\nu = m \sharp \pi^* = (\mathrm{id}, \boldsymbol{f}_+) \sharp \boldsymbol{\mu}_+$ is a barycenter.

Step 3: uniqueness of a transport map. Likewise, we show the existence of a function $\boldsymbol{f}_-$ such that $\nu' = (\mathrm{id}, \boldsymbol{f}_-) \sharp \boldsymbol{\mu}_-$ is a barycenter. If we show that there is a unique barycenter, then $\nu = \nu' = \nu^{\mathrm{bar}}$, and the theorem is proven.

We now show uniqueness of the barycenter. Let ν be any measure that solves the barycenter problem. Let π_+ (resp. π_-) be an optimal transport plan for $\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_+, \nu)$ (resp. $\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_-, \nu)$). By the gluing lemma, there exists $\rho \in \Pi(\cdot, \boldsymbol{\mu}_+, \boldsymbol{\mu}_-)$ whose joint law of the two first marginals is equal to π_+, and whose joint law of the first and last marginal is equal to π_-. The joint distribution π between the second and last marginal is a transport plan between $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$. Furthermore, as ν is a barycenter and by definition of $\mathbf{C}$, we have

$$\begin{aligned}\mathrm{OT}_{\mathbf{C}}(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-) = \mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_+, \nu) + \mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_-, \nu) &= \int (\mathbf{c}(\mathbf{z}_1, y) + c(\mathbf{z}_2, y)) \mathrm{d}\rho(y, \mathbf{z}_1, \mathbf{z}_2) \\ &\geq \int \mathbf{C}(\mathbf{z}_1, \mathbf{z}_2) \mathrm{d}\pi(\mathbf{z}_1, \mathbf{z}_2),\end{aligned}$$

so that π is an optimal transport plan between $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$, with $\nu = m \sharp \pi$. But then, recall that (17) holds for *any* optimal transport plan π (for the same potential φ). Hence, by the same arguments as before, we have $\nu = (\boldsymbol{f}_+) \sharp \boldsymbol{\mu}_+$ for the map $\boldsymbol{f}_+ : \mathbf{z}_1 \mapsto h_1 - \frac{d_1 \partial_{h_1} \varphi(\mathbf{z}_1)}{2}$

(defined $\boldsymbol{\mu}_+$-almost everywhere). In particular, ν is uniquely determined by $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ through the potential φ. ■

Theorem 12 is the counterpart of Theorem 4, established by Chzhen et al. (2020b) and Gouic et al. (2020) within the awareness framework. Both theorems demonstrate that the optimal fair regression function solves a barycenter problem with optimal transport costs. Remark 11 further indicates that Theorem 4 generalizes Theorem 2.3 in Chzhen et al. (2020b), as the awareness framework can be considered as a special case of the unawareness framework. However, unlike in the awareness framework, there is no explicit formulation of the optimal fair prediction function in the unawareness framework, as the corresponding barycenter problem involves multi-to-one dimensional transport costs with no explicit solutions.

Theorem 12 reveals that the fair prediction $f^*(x)$ only depends on the bi-dimensional feature $(\eta(x), \Delta(x))$ of the candidate x. By definition, $\Delta(x) \propto \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu}(x) - \frac{\mathrm{d}\mu_2}{\mathrm{d}\mu}(x)$. Moreover, we have $\mathbb{P}(S = 1|X = x) = p_1 \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu}(x)$ and $\mathbb{P}(S = 2|X = x) = p_2 \frac{\mathrm{d}\mu_2}{\mathrm{d}\mu}(x)$. Thus, $\Delta(x) \propto \frac{\mathbb{P}(S=1|X=x)}{p_1} - \frac{\mathbb{P}(S=2|X=x)}{p_2}$. In other words, $\Delta(x)$ reflects the probability that x belongs to the different groups. Hence, in the unawareness framework, the optimal fair regression function effectively relies on estimates of S to make its prediction. This result provides a theoretical justification for the empirical observations of Lipton et al. (2018). As noted by these authors, this phenomenon may be undesirable, as it means that the predictions can rely on features not relevant to predict the response Y, simply because they are predictive of the group S.

4. Links between classification and regression problems

We now turn to the study of the relationship between fair regression and fair classification problems within the unawareness framework. When $Y \in \{0, 1\}$, classical results show that the Bayes classifier g_y^{Bayes} minimizing the risk

$$\mathcal{R}_y(g) = y \cdot \mathbb{P}\left[Y = 0, g(X) = 1\right] + (1 - y) \cdot \mathbb{P}\left[Y = 1, g(X) = 0\right].$$

is given by $g_y^{\text{Bayes}}(x) = \mathbb{1}\left\{f^{\text{Bayes}}(x) \geq y\right\}$, where f^{Bayes} is the Bayes regression function minimizing $\mathcal{R}_{sq}$. Similarly, recent results by Gaucher et al. (2023) demonstrate that in the awareness framework, the optimal fair classifier g_y^* minimizing the risk $\mathcal{R}_y$ is given by $g_y^*(x, s) = \mathbb{1}\left\{f^*(x, s) \geq y\right\}$, where f^* is the optimal fair regression function minimizing $\mathcal{R}_{sq}$. These results can be leveraged to obtain plug-in classifiers $\hat{g}$ using estimates $\hat{f}$ of the regression function.

Somewhat less explored is the converse relationship: given a family of optimal classifiers $(g_y)_{y\in[0,1]}$ for the risks $(\mathcal{R}_y)_{y\in[0,1]}$, one could define a regression function f of the form $f(x) = \sup\{y : g_y(x) = 1\}$. For example, this formulation yields the Bayes regression function when using Bayes classifiers and the optimal fair regression function when using optimal fair classifiers in the awareness framework. In both examples, this relationship may not be particularly useful since there already exists an explicit characterization of the optimal regression function. However, if this relationship were to hold in the unawareness framework, it would be significantly more valuable. Indeed, Theorem 12 rephrases the fair regression problem as a barycenter problem with optimal transport costs but does not provide an explicit solution.

Agarwal et al. (2019) proposed leveraging this relationship to address the problem of fair regression using cost-sensitive classifiers. The authors demonstrate an equivalence between minimizing a discretized version of the risk $\mathcal{R}_{sq}$ and minimizing the average of the cost-sensitive risks $(\mathcal{R}_y(g_{f,y}))_{y\in\mathcal{Z}}$ for a finite set $\mathcal{Z}$, where $g_{f,y}$ is defined as $g_{f,y}(x) = \mathbb{1}\{f(x) \geq y\}$. To obtain the optimal fair regression function for this discretized risk, the authors assume access to an oracle that returns the regression function f such that $g_{f,y}$ minimizes the average of the risks $(\mathcal{R}_y(g_{f,y}))_{y\in\mathcal{Z}}$. We emphasize that minimizing the average of the risks $(\mathcal{R}_y(g_{f,y}))_{y\in\mathcal{Z}}$ remains an open and challenging problem.

In contrast, to define a regression function of the form $f(x) = \sup\{y : g_y(x) = 1\}$, one only needs to solve independent cost-sensitive classification problems. Recent results by Gaucher et al. (2023) offer an explicit characterization of these classifiers. This raises the intriguing possibility of constructing the optimal fair regression function in the unawareness framework using these fair classifiers. In this section, we demonstrate that such a construction is not always possible. To do so, we begin by providing some reminders on fair classification in the unawareness framework.

4.1 Fair classification

In this section, we assume that $Y \in \{0, 1\}$ almost surely. We consider the problem of minimizing a family of risk measures $\mathcal{R}_y$ under the demographic parity constraint. We show that the optimal fair classifier for the risk $\mathcal{R}_y$ is of the form g_y^κ for some $\kappa \in \mathbb{R}$, where g_y^κ is given by

$$\forall x \in \mathcal{X}, \quad g_y^\kappa(x) = \mathbf{1}\{\eta(x) \geq y + \kappa\Delta(x)\}. \tag{18}$$

The following proposition extends Proposition 5.3 in Gaucher et al. (2023), and characterizes the optimal fair classifier.

Proposition 14 *Let $y \in \mathbb{R}$, and let $\kappa^* \in \mathbb{R}$ satisfy*

$$\mu_+\left(\eta(X) \geq y + \kappa^*\Delta(X)\right) = \mu_-\left(\eta(X) \geq y + \kappa^*\Delta(X)\right).$$

Under Assumption A, $g_y^{\kappa^*}$ solves the fair classification problem

$$\begin{cases} \textit{minimize} & \mathcal{R}_y(g) \\ \textit{such that} & \mathbb{E}\left[g(X)|S=1\right] = \mathbb{E}\left[g(X)|S=2\right]. \end{cases} \tag{C_y}$$

Moreover, all solutions to (C_y) are a.s. equal to $g_y^{\kappa^*}$ on $\mathcal{X} \setminus \{x \in \mathcal{X} : \eta(x) = y$ and $\Delta(x) = 0\}$.

The optimal classifier is uniquely defined outside of $\{\eta(x) = y\}$. While the set $\{\eta(x) = y$ and $\Delta(x) \neq 0\}$ has null measure under Assumption A, the set $\{\eta(x) = y$ and $\Delta(x) = 0\}$ may have positive measure. On this set, the classifiers $\mathbf{1}\{\eta(x) \geq y\}$ and $\mathbf{1}\{\eta(x) > y\}$ differ, yet they are both optimal for the risk $\mathcal{R}_y(g)$.

The proof of this proposition is postponed to Appendix A.2. As discussed in the previous section, Assumption A encompasses as a special case the awareness framework. In this case, the optimal classifier presented in Proposition 14 reduces to the optimal fair classifier in the awareness framework given by

$$\forall (x,s) \in \mathcal{X} \times \{1,2\}, \quad g_y^{\text{aware}}(x,s) = \begin{cases} \mathbf{1}\{\eta(x) \geq y + \frac{\kappa^*}{p_1}\} & \text{if } s = 1 \\ \mathbf{1}\{\eta(x) \geq y - \frac{\kappa^*}{p_2}\} & \text{if } s = 2, \end{cases} \tag{19}$$

as described in Schreuder and Chzhen (2021); Zeng et al. (2022b).

In the unawareness framework, the optimal classifier relies on the probability that the observation X belongs to the different groups $\Delta(X)$. This behavior is similar to that of the optimal fair regression function, as established in Theorem 12. Next, we investigate whether the optimal classifier is envy-free.

Definition 15 (Envy-free classifiers) *We say that a classifier $g : \mathcal{X} \to \{0,1\}$ is envy-free within group if $\mathbb{P} \otimes \mathbb{P}$-a.s., for (X,S) and (X',S') such that $S = S'$ and $g^{Bayes}(X) > g^{Bayes}(X')$, we have $g(X) \geq g(X')$.*

In essence, this property ensures that no candidate who would have been accepted by the Bayes classifier but is rejected after fairness correction envies another candidate *from the same group* who would have been rejected by the Bayes classifier but is accepted after fairness correction. Observe that this property is weaker than order preservation, as a classifier that preserves order is envy-free within groups.

Proposition 16 reveals that in the unawareness framework, the optimal fair classifier is generally not envy-free. This behavior contrasts with that of optimal fair classifiers in the awareness framework: indeed, since these classifiers preserve order, they are also envy-free.

Proposition 16 *Let $y \in \mathbb{R}$. Under Assumption A, if $\mathbb{P}(S = s|X = x) \in (0,1)$ for all $s \in \mathcal{S}$, $x \in \mathcal{X}$, then one of the following cases hold:*

1. $g_y^{Bayes}(x) = 1 \implies g_y^{\kappa^*}(x) = 1$ *μ-a.s.*
2. $g_y^{\kappa^*}(x) = 1 \implies g_y^{Bayes}(x) = 1$ *μ-a.s.*
3. *the classifier $g_y^{\kappa^*}$ is not envy-free within group.*

Proof Assume that 1. and 2. do not hold. Then, we have $\kappa^* \neq 0$, and we can assume without loss of generality that $\kappa^* > 0$. Since 1. does not hold, we have that $\mu(g_y^{\text{Bayes}}(X) = 1 \text{ and } g_y^{\kappa^*}(X) = 0) > 0$. This implies in turn that $\mu_+(g^{\text{Bayes}}(X) = 1 \text{ and } g_y^{\kappa^*}(X) = 0) > 0$, since g_y^{Bayes} and $g_y^{\kappa^*}$ coincide on $\mathcal{X}_=$, and since by definition when $\kappa^* > 0$, we have $\mu_-(g_y^{\text{Bayes}}(X) = 1 \text{ and } g_y^{\kappa^*}(X) = 0) = 0$. Similarly, we can show that since 2. does not hold, $\mu_-(g_y^{\text{Bayes}}(X) = 0 \text{ and } g_y^{\kappa^*}(X) = 1) > 0$. Now, $\mathbb{P}(S = s|X = x) \in (0,1)$ for all $s \in \mathcal{S}$, so $\mu_1 \gg \mu_+$ and $\mu_1 \gg \mu_-$. Therefore, $\mu_1(g_y^{\text{Bayes}}(X) = 1 \text{ and } g_y^{\kappa^*}(X) = 0) > 0$, and $\mu_1(g_y^{\text{Bayes}}(X) = 0 \text{ and } g_y^{\kappa^*}(X) = 1) > 0$. This implies

$$\mathbb{P}\left(g_y^{\text{Bayes}}(X) > g_y^{\text{Bayes}}(X') \text{ and } g_y^{\kappa^*}(X) < g_y^{\kappa^*}(X') \middle| S = S' = 1\right) > 0$$

which concludes the proof. ■

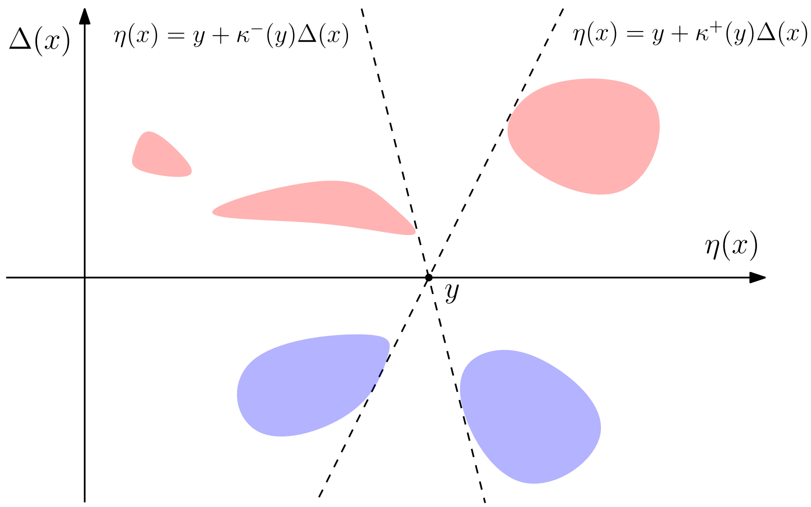

Figure 1: The measure $\boldsymbol{\mu}_+$ is displayed in red and the measure $\boldsymbol{\mu}_-$ is displayed in blue. By definition of $\kappa^+(y)$ and $\kappa^-(y)$, the red region and the blue region to the right of the two dotted lines have equal masses. The region in between the two lines contains no mass.

Extending cost-sensitive classification to $\mathcal{Y} = \mathbb{R}$ In the following, we consider the more general case where $\mathcal{Y} = \mathbb{R}$. Although the interpretation in terms of optimal classification is no longer applicable, we can still analyze the family of functions g_y^κ defined in Equation (18). The following proposition characterizes the values of the parameter κ^* such that $g_y^{\kappa^*}$ satisfies demographic parity. These values partition the feature space equally, see Figure 1.

Proposition 17 *Let $y \in \mathbb{R}$. Under Assumption A, the set of numbers $\kappa \in \mathbb{R}$ such that*

$$\mu_+ \left(\eta(X) \geq y + \kappa\Delta(X)\right) = \mu_- \left(\eta(X) \geq y + \kappa\Delta(X)\right) \tag{20}$$

is a nonempty closed interval $I(y) = [\kappa^-(y), \kappa^+(y)]$. The function $y \mapsto \kappa^+(y)$ is upper semicontinuous and the function $y \mapsto \kappa^-(y)$ is lower semicontinuous. Moreover, it holds that for μ-almost all x, for all $y \in \mathbb{R}$ and all $\kappa, \kappa' \in I(y)$, $g_y^\kappa(x) = g_y^{\kappa'}(x)$.

Proof Introduce the function

$$G : (\kappa, y) \mapsto \mu_+ \left(\eta(X) \geq y + \kappa\Delta(X)\right) - \mu_- \left(\eta(X) \geq y + \kappa\Delta(X)\right).$$

Under Assumption A, the measures $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ give zero mass to non-horizontal lines, implying that the function G is continuous. Furthermore, for $y \in \mathbb{R}$, the function $G(\cdot, y)$ is nonincreasing (recall that $\Delta(X) < 0$ for $X \sim \mu_-$). Hence, its zeroes form a closed interval $I(y)$. For $\kappa \in \mathbb{R}$, the set $\{y \in \mathbb{R} : \ \kappa^-(y) > \kappa\}$ is equal to the set $\{y \in \mathbb{R} : \ G(\kappa, y) > 0\}$, which is an open set by continuity of G. This proves that κ^- is lower semicontinuous. We prove similarly that κ^+ is upper semicontinuous.

It remains to prove the last statement. Fix $y \in \mathbb{R}$. First, we may assume without loss of generality that $\kappa = \kappa^-(y)$ and that $\kappa' = \kappa^+(y)$. We have $\mu_+ \left(\eta(X) \geq y + \kappa^+(y)\Delta(X)\right) =$

$\mu_+ (\eta(X) \geq y + \kappa^-(y)\Delta(X))$ (and likewise for μ_-). Thus, we have

$$\begin{aligned}\mu_+(g_y^\kappa(X) \neq g_y^{\kappa'}(X)) &= \mu_+(g_y^\kappa(X) = 1,\ g_y^{\kappa'}(X) = 0) \\ &= \mu_+\left(\frac{\eta(X) - y}{\Delta(X)} \in [\kappa, \kappa']\right) = 0.\end{aligned}$$

The same equality holds for μ_-. Also, the equality $g_y^{\kappa^+(y)}(x) = g_y^{\kappa^-(y)}(x)$ holds on $\mathcal{X}_=$ (as $\Delta(x) = 0$ on $\mathcal{X}_=$). Hence, for a fixed y, the equality $g_y^{\kappa^+(y)}(x) = g_y^{\kappa^-(y)}(x)$ holds for μ-almost all x.

However, the set of points x (of full measure) where this equality is satisfied depends on y, so that it is not trivial to show that this equality holds simultaneously for all $y \in \mathbb{R}$, almost surely.

To do so, we show that the set $\{(h, d) \in \Omega :\ \exists y \in \mathbb{R},\ y + \kappa^-(y)d \leq h \leq y + \kappa^+(y)d\}$ has mass 0 under $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$. For $y \in \mathbb{R}$, let $C_y = \{(h, d) \in \Omega :\ y + \kappa^-(y)d \leq h \leq y + \kappa^+(y)d\}$. We have previously shown that for any given $y \in \mathbb{R}$, $g_y^{\kappa^-(y)}(x) = g_y^{\kappa^+(y)}(x)$ for μ-almost all x, implying that $\boldsymbol{\mu}_\pm(C_y) = 0$. Let $C = \bigcup_{y\in\mathbb{R}} C_y$. Let us show that $\boldsymbol{\mu}_+(C) = 0$. Let $C_1 = \bigcup_{y\in\mathbb{R}} \mathring{C}_y$. First, it holds that $\boldsymbol{\mu}_+(C_1) = 0$. If it were not the case, as the measure $\boldsymbol{\mu}_+$ is inner regular, there would exist a compact set $K \subset C_1$ with $\boldsymbol{\mu}_+(K) > 0$. But then, the compact set K is covered by the family of open sets $(\mathring{C}_y)_{y\in\mathbb{R}}$. By compactness, there exists a finite cover $\mathring{C}_{y_1}, \ldots, \mathring{C}_{y_N}$ covering K. As each $\mathring{C}_{y_i}$ has zero mass, we obtain a contradiction with the positivity of $\boldsymbol{\mu}_+(K)$. Furthermore, if $(h, d) \in C \backslash C_1$, then there exists y_0 with either $h = y_0 + \kappa^-(y_0)d$ or $h = y_0 + \kappa^+(y_0)d$, with also $y + \kappa^-(y)d \leq h \leq y + \kappa^+(y)d$ for all $y \in \mathbb{R}$. This implies that $C\backslash C_1$ is included in the union of the graphs of the two functions $d \mapsto \sup_y(y + \kappa^-(y)d)$ and $d \mapsto \inf_y(y + \kappa^+(y)d)$. These two functions can easily be seen to be measurable because of the semicontinuity of κ^- and κ^+. Thus, by Assumption A, $\boldsymbol{\mu}_+(C\backslash C_1) = 0$. In conclusion, we have proven that $\boldsymbol{\mu}_+(C) = 0$. We show likewise that $\boldsymbol{\mu}_-(C) = 0$. ■

4.2 The nestedness assumption

Recall that our goal is to determine whether the optimal fair regression function can be expressed as $f^*(x) = \sup\{y : g_y^{\kappa(y)}(x) = 1\}$ for a certain choice $\kappa(y) \in I(y)$. In this section, we introduce an assumption regarding the family of classifiers $g_y^{\kappa(y)}$ and demonstrate that this assumption is necessary for the relationship to hold. Specifically, we wish to use the decision boundaries of optimal classifiers at different risk levels to define the regression function. For this to be possible, the function $y \mapsto g_y^{\kappa(y)}(x)$ must be nonincreasing for any choice of x: in other words, the rejection regions $\{x : g_y^{\kappa(y)}(x) < y\}$ must be nested. We formalize this assumption in the following definition.

Definition 18 (Nestedness) *We say that the problem corresponding to $(X, Y, S) \sim \mathbb{P}$ is nested if there exists a choice of $\kappa(y) \in I(y)$ for all $y \in \mathbb{R}$ such that*

$$\text{for } \mu\text{-almost all } x \in \mathcal{X},\ \text{ the map } y \in \mathbb{R} \mapsto g_y^{\kappa(y)}(x) \text{ is nonincreasing.} \qquad \textbf{(Nested)}$$

A straightforward (but key) property implied by nestedness is the fact that the sets

$$\forall y \in \mathbb{R},\ A(y) = \{x \in \mathcal{X} :\ \eta(x) < y + \kappa(y) \cdot \Delta(x)\} \tag{21}$$

are "almost" nested, in the sense that there exists a set $\tilde{\mathcal{X}}$ of full μ-measure such that for all $y' \leq y$, it holds that $A(y') \cap \tilde{\mathcal{X}} \subseteq A(y) \cap \tilde{\mathcal{X}}$.

Lemma 19 *Assume that Assumption A holds. Then, the problem is nested with choice $\kappa(y) \in I(y)$ for all $y \in \mathbb{R}$ if and only if for all $y < y'$,*

$$\mu(g_y^{\kappa(y)}(X) = 0 \text{ and } g_{y'}^{\kappa(y')}(X) = 1) = 0. \tag{22}$$

Furthermore, if the problem is nested, one can always choose $\kappa(y) = \kappa^+(y)$ for all $y \in \mathbb{R}$ in the definition of $g_y^{\kappa(y)}$.

Proof The direct implication is clear. For the converse one, assume that the nestedness assumption does not hold. By definition, there exists a measurable set $\mathcal{X}_0$ of positive μ-mass such that $y \mapsto g_y^{\kappa(y)}(x)$ is not nonincreasing for all $x \in \mathcal{X}_0$. It holds that either $\mu_+(\mathcal{X}_0) > 0$ or that $\mu_-(\mathcal{X}_0) > 0$. Assume without loss of generality that the first condition is satisfied, and let $\tilde{\mathcal{X}}$ be the set of points x in $\mathcal{X}_0 \cap \mathcal{X}_+$ that satisfy

$$\forall y \in \mathbb{R},\ \frac{\eta(x) - y}{\Delta(x)} \notin [\kappa^-(y), \kappa^+(y)] \tag{23}$$

According to Proposition 17, $\mu_+(\tilde{\mathcal{X}}) = \mu_+(\mathcal{X}_0 \cap \mathcal{X}_+) > 0$. For $x \in \tilde{\mathcal{X}}$, there exists $y < y'$ with $g_y^{\kappa(y)}(x) = 0$ and $g_{y'}^{\kappa(y')}(x) = 1$. As x satisfies (23), we have that

$$\begin{cases} \eta(x) < y + \kappa^-(y)\Delta(x) \\ \eta(x) > y' + \kappa^+(y')\Delta(x). \end{cases}$$

Because the function κ^- is lower semicontinuous, for $\tilde{y}$ close enough to y, we also have $\eta(x) < \tilde{y} + \kappa^-(\tilde{y})\Delta(x)$. Likewise, there exists $\tilde{y}'$ close enough to y' with $\eta(x) > \tilde{y}' + \kappa^+(\tilde{y}')\Delta(x)$. In conclusion, we have shown that

$$\tilde{\mathcal{X}} \subset \bigcup_{\substack{y,y' \in \mathbb{Q} \\ y < y'}} \{x :\ \eta(x) < y + \kappa^-(y)\Delta(x) \text{ and } \eta(x) > y' + \kappa^+(y')\Delta(x)\}$$

In particular, as $\mu_+(\tilde{\mathcal{X}}) > 0$, there exists $y < y' \in \mathbb{Q}$ with

$$\mu_+(\eta(X) < y + \kappa^-(y)\Delta(X) \text{ and } \eta(X) > y' + \kappa^+(y')\Delta(X)) > 0.$$

According to Proposition 17 and Assumption A, as the equality $\eta(X) = y' + \kappa^+(y')\Delta(X)$ happens with zero μ_+-probability, we have

$$\mu_+(g_y^{\kappa(y)}(X) = 0 \text{ and } g_{y'}^{\kappa(y')}(X) = 1) > 0,$$

proving the first claim.

The second claim follows from the characterization of nestedness that we have just established. Indeed, let $y < y'$. By Proposition 17, for μ-almost all x, $g_y^{\kappa(y)}(x) = g_y^{\kappa^+(y)}(x)$ and $g_{y'}^{\kappa(y')}(x) = g_{y'}^{\kappa^+(y')}(x)$. Thus, if (22) holds for $\kappa(y)$ and $\kappa(y')$, it also holds for $\kappa^+(y)$ and $\kappa^+(y')$. ■

As a warm-up, we begin by showing that the nestedness assumption is always verified in the awareness setting.

Proposition 20 *Assume that S is X-measurable. Then, under Assumption A, the classification problem is nested.*

Proof We prove Proposition 20 by contradiction. Assume that the problem is not nested. Using Lemma 19, there exist $y < y'$ and a set $\mathcal{X}_0$ of positive μ probability such that for all $x \in \mathcal{X}_0$, $g_y^{\kappa(y)}(x) = 0$ and $g_{y'}^{\kappa(y')}(x) = 1$. Using Proposition 17, we can also assume without loss of generality that $\eta(x) < y + \kappa^-(y)\Delta(x)$ and $\eta(x) > y' + \kappa^+(y')\Delta(x)$ for $x \in \mathcal{X}_0$. Letting $x \in \mathcal{X}_0$, that we assume without loss of generality is in $\mathcal{X}_+$, the previous inequalities become

$$y' + \frac{\kappa^+(y')}{p_1} < \eta(x) < y + \frac{\kappa^-(y)}{p_1}.$$

In words, the threshold for admission is lower at level y' that at level y. This implies in particular that $\mu_+(g_{y'}^{\kappa^+(y')}(X) = 1) \geq \mu_+(g_y^{\kappa^-(y)}(X) = 1)$. On the other hand, since $y < y'$, it also implies that $\kappa^+(y') < \kappa^-(y)$. Therefore, $y - \frac{\kappa^-(y)}{p_2} < y' - \frac{\kappa^+(y')}{p_2}$, so $\mu_-(g_{y'}^{\kappa^+(y')}(X) = 1) \leq \mu_-(g_y^{\kappa^-(y)}(X) = 1)$. Using $\mu_+(g_{y'}^{\kappa^+(y')}(X) = 1) = \mu_-(g_{y'}^{\kappa^+(y')}(X) = 1)$ and $\mu_+(g_y^{\kappa^-(y)}(X) = 1) = \mu_-(g_y^{\kappa^-(y)}(X) = 1)$, we find that $\mu_+(g_{y'}^{\kappa^+(y')}(X) = 1) = \mu_+(g_y^{\kappa^-(y)}(X) = 1)$. It implies that

$$\mu_+ \left(\eta(X) \in \left[y' + \frac{\kappa^+(y')}{p_1}, y + \frac{\kappa^-(y)}{p_1} \right] \right) = 0.$$

Likewise,

$$\mu_- \left(\eta(X) \in \left[y - \frac{\kappa^-(y)}{p_2}, y' - \frac{\kappa^+(y')}{p_2} \right] \right) = 0.$$

In particular, for $\kappa = \kappa^+(y') + p_1(y' - y)$, we see that $y + \frac{\kappa}{p_1} = y' + \frac{\kappa^+(y')}{p_1} < y + \frac{\kappa^-(y)}{p_1}$. Thus, $\kappa < \kappa^-(y)$, and $y - \frac{\kappa}{p_2} \geq y - \frac{\kappa^-(y)}{p_2}$. Moreover, $\kappa > \kappa^+(y')$, and $y < y'$, so $y - \frac{\kappa}{p_2} < y' - \frac{\kappa^+(y')}{p_2}$. This implies that

$$y' - \frac{\kappa^+(y')}{p_2} - (y - \frac{\kappa}{p_2}) = y' - y + \frac{1}{p_2} \left(\kappa^+(y') + p_1(y' - y) - \kappa^+(y') \right) > 0.$$

so $y - \frac{\kappa}{p_2} \in \left[y - \frac{\kappa^-(y)}{p_2}, y' - \frac{\kappa^+(y')}{p_2} \right]$. Thus,

$$\mu_+ \left(\eta(X) \geq y + \frac{\kappa}{p_1} \right) = \mu_- \left(\eta(X) \geq y - \frac{\kappa}{p_2} \right)$$

and $\kappa \in I(y)$. Since $\kappa < \kappa^-(y)$, this yields a contradiction. ■

Somewhat surprisingly, although the nestedness assumption may appear intuitive, it is not always verified. In Section 5, we present examples where this assumption holds and others where it does not.

Before proving in the next section that under the nestedness assumption, the optimal fair classification functions $g_y^{\kappa(y)}$ can be recovered by thresholding the optimal fair regression function f^*, we prove the converse: if the problem is not nested, there exists a value of $y \in \mathbb{R}$ where the classifier $\mathbf{1}\{f^*(x) \geq y\}$ is suboptimal for the fair classification problem (C_y).

Proposition 21 *Assume that $\mathcal{Y} = \{0, 1\}$, that Assumption A holds and that the classification problem is not nested. Let f^* be the optimal fair regression function in the unawareness framework. Then, there exists $y \in \mathbb{R}$ such that the classifier $x \mapsto \mathbf{1}\{f^*(x) \geq y\}$ is not the optimal fair classifier for the risk $\mathcal{R}_y$.*

Proof According to Lemma 19, there exists $y < y'$ with

$$\mu(g_y^{\kappa(y)}(X) = 0 \text{ and } g_{y'}^{\kappa(y')}(X) = 1) > 0.$$

Let $\mathcal{X}_0$ be the set corresponding to this event. Let us consider a classifier of the form $g_y(x) = \mathbf{1}\{f(x) \geq y\}$. On the one hand, if $\mu\,(X \in \mathcal{X}_0 \text{ and } f(X) < y) > 0$, then the probability $\mu\,(X \in \mathcal{X}_0 \text{ and } f(X) < y')$ is also positive, so $g_{y'}(x)$ and $g_{y'}^{\kappa(y')}(x)$ disagree on a set of positive probability. Now, Proposition 14 implies that all optimal classifiers are a.s. equal, so $g_{y'}$ is sub-optimal. On the other hand, if $\mu\,(X \in \mathcal{X}_0 \text{ and } f(X) < y) = 0$, then $g_y(x) = 1$ for almost all $x \in \mathcal{X}_0$. This implies that $g_y(x)$ and $g_y^{\kappa(y)}(x)$ disagree on a set of positive probability, so g_y is sub-optimal. ■

4.3 Constructing a regression function using nested classifiers

In the previous section, we proved that under mild assumptions, nestedness is a necessary condition for the relationship $g_y^*(x) = \mathbf{1}\{f^*(x) \geq y\}$ between the optimal fair classification and regression functions to hold. We now conclude by showing that nestedness is also a sufficient condition for this relationship to hold.

We begin by defining the function $f^* : \mathcal{X} \to \mathbb{R}$ as

$$\forall x \in \mathcal{X}, \;\; f^*(x) = \sup\{y : g_y^{\kappa(y)}(x) = 1\} \tag{24}$$

where $g_y^{\kappa(y)}$ is given by Equation (18). We assume without loss of generality (using Lemma 19) that $\kappa(y) = \kappa^+(y)$ for all $y \in \mathbb{R}$. Observe that f^* is then almost measurable because of the upper semicontinuity of κ^+, in the sense that its restriction to some set of full measure is measurable (here given by the set of full measure where $y \mapsto g_y^{\kappa(y)}(x)$ is nonincreasing).

Theorem 22 *Assume that the classification problem is nested and that Assumption A is satisfied. Then, the regression function f^* is optimal for the fair regression problem (1).*

Before proving Theorem 22, we state the following corollary.

Corollary 23 *Assume that the classification problem is nested, that Assumption A is satisfied, and that $\mathcal{Y} = \{0,1\}$. Then, the classification function $g_y : y \mapsto \mathbf{1}\{f^*(x) \geq y\}$ is optimal for the fair classification problem problem with cost $\mathcal{R}_y$, where f^* is the solution to the fair regression problem* (1).

The proof of Corollary 23 follows immediately by noticing that by Theorem 12, f^* is uniquely defined, and that the nestedness assumption and Theorem 22 imply that $g_y(x) = g_y^{\kappa(y)}(x)$ a.s.

The rest of the section is devoted to proving Theorem 22. To do so, we begin by proving that f^* is a fair regression function, and by defining F, the c.d.f. of the predictions under μ_+ and μ_-.

Lemma 24 *Assume that the problem is nested and that Assumption A is satisfied. Let $F : \mathbb{R} \to \mathbb{R}$ be defined by*

$$\forall y \in \mathbb{R},\ F(y) = \mu_+ \left(\eta(X) \leq y + \kappa(y)\Delta(X)\right) = \mu_- \left(\eta(X) \leq y + \kappa(y)\Delta(X)\right). \tag{25}$$

Then, there exists a probability measure ν^* with continuous c.d.f. F and finite second moment such that $f^*\sharp\mu_+ = f^*\sharp\mu_- = \nu^*$. In particular, f^* is a fair regression function.

Proof The "almost" nestedness of the sets $(A(y))_y$ implies that F is nondecreasing. Let us show that the function F is the c.d.f. of some continuous random variable, i.e., that it goes to 0 in $-\infty$, that it goes to 1 in $+\infty$, and that it is continuous.

First, recall that $\Delta(X) > 0$ for $X \sim \mu_+$ and that $\Delta(X) < 0$ for $X \sim \mu_-$. Thus, if $\kappa(y) \leq 0$, then $F(y) \leq \mu_+ \left(\eta(X) - y \leq 0\right)$ and if $\kappa(y) \geq 0$, then $F(y) \leq \mu_- \left(\eta(X) - y \leq 0\right)$. Hence,

$$F(y) \leq \max\left\{\mu_+ \left(\eta(X) - y \leq 0\right), \mu_- \left(\eta(X) - y \leq 0\right)\right\}, \tag{26}$$

and F converges to 0 in $-\infty$. Similarly, $F(y) \to 1$ when y converges to $+\infty$.

Next, let us show that F is continuous. Let $y_0, y_1 \in \mathbb{R}$ be such that $y_0 < y_1$. Now, if $\kappa(y_0) \geq \kappa(y_1)$, then

$$\begin{aligned} F(y_1) - F(y_0) &= \mu_+ \left(\eta(X) \leq y_1 + \kappa(y_1)\Delta(X)\right) - \mu_+ \left(\eta(X) \leq y_0 + \kappa(y_0)\Delta(X)\right) \\ &\leq \mu_+ \left(\eta(X) \leq y_1 + \kappa(y_0)\Delta(X)\right) - \mu_+ \left(\eta(X) \leq y_0 + \kappa(y_0)\Delta(X)\right) \end{aligned}$$

Similarly, if $\kappa(y_0) \leq \kappa(y_1)$, then, (recalling that $\Delta(X) < 0$ for $X \sim \mu_-$)

$$\begin{aligned} F(y_1) - F(y_0) &= \mu_- \left(\eta(X) \leq y_1 + \kappa(y_1)\Delta(X)\right) - \mu_- \left(\eta(X) \leq y_0 + \kappa(y_0)\Delta(X)\right) \\ &\leq \mu_- \left(\eta(X) \leq y_1 + \kappa(y_0)\Delta(X)\right) - \mu_- \left(\eta(X) \leq y_0 + \kappa(y_0)\Delta(X)\right). \end{aligned}$$

Thus,

$$F(y_1) - F(y_0) \leq \mu_+ \left(\eta(X) - \kappa(y_0)\Delta(X) \in [y_0, y_1]\right) + \mu_- \left(\eta(X) - \kappa(y_0)\Delta(X) \in [y_0, y_1]\right)$$

We have shown that F is non-decreasing, so $F(y_1) - F(y_0) \geq 0$. Under Assumption A, μ_+ and μ_- give zero mass to the sets $\{\eta(X) = y_0 + \kappa(y_0)\Delta(X)\}$, so $F(y_1) - F(y_0) \to 0$ as

$y_1 \to y_0^+$. This proves that F is right-continuous. To show that F is left-continuous, we Observe that if $\kappa(y_0) \geq \kappa(y_1)$, then

$$F(y_1) - F(y_0) \leq \mu_+ \left(\eta(X) \leq y_1 + \kappa(y_1)\Delta(X)\right) - \mu_+ \left(\eta(X) \leq y_0 + \kappa(y_1)\Delta(X)\right).$$

Similarly, if $\kappa(y_0) \leq \kappa(y_1)$, then,

$$F(y_1) - F(y_0) \leq \mu_- \left(\eta(X) \leq y_1 + \kappa(y_1)\Delta(X)\right) - \mu_- \left(\eta(X) \leq y_0 + \kappa(y_1)\Delta(X)\right).$$

Thus,

$$F(y_1) - F(y_0) \leq \mu_+ \left(\eta(X) - \kappa(y_1)\Delta(X) \in [y_0, y_1]\right) + \mu_- \left(\eta(X) - \kappa(y_1)\Delta(X) \in [y_0, y_1]\right)$$

and F is also left-continuous.

Then, let us show that ν^* has finite second moment. Let $Z \sim \nu^*$. We have

$$\mathbb{E}[Y^2] = \int_0^{+\infty} \mathbb{P}(Z^2 > t)\mathrm{d}t = \int_0^{+\infty} (F(\sqrt{t}) + (1 - F(-\sqrt{t})))\mathrm{d}t$$

We use (26) to obtain that for $y \in \mathbb{R}$, $F(y) \leq \max(\mu_+(\eta(X) \leq y), \mu_-(\eta(X) \leq y))$. But, as $\mathbb{E}[Y^2] < +\infty$, the random variable $\eta(X)$ has a finite second moment under the law of either μ_+ or μ_-. In particular, $\int_0^{+\infty} F(\sqrt{t})\mathrm{d}t$ is finite. Similarly, $\int_0^{+\infty}(1 - F(-\sqrt{t}))\mathrm{d}t$ is finite.

Finally, we prove the statement $f^*\sharp\mu_+ = \nu^*$. Indeed, for all $y_0 \in \mathbb{R}$, we have using that upper semicontinuity of $\kappa(y) = \kappa^+(y)$ that

$$\begin{aligned} f^*\sharp\mu_+((-\infty, y_0]) &= \mu_+ \left(\sup\{y : g_y^{\kappa(y)}(X) = 1\} \leq y_0\right) \\ &= \mu_+ \left(g_{y_0}^{\kappa(y_0)}(X) = 0\right) = \mu_+(\eta(X) < y_0 + \kappa(y_0)\Delta(X)) = F(y_0). \end{aligned}$$

where the second line follows from the nestedness assumption and the fact that the line $\{\eta(X) = y_0 + \kappa(y_0)\Delta(X)\}$ has zero mass. We show similarly that $f^*\sharp\mu_- = \nu^*$, thus concluding the proof of Lemma 24. ■

The function f^* depends only on x through the pair $(\eta(x), \Delta(x))$. Let $\boldsymbol{f}^* : \Omega \to \mathbb{R}$ be defined by the relation $f^*(x) = \boldsymbol{f}^*(\eta(x), \Delta(x))$ for $x \in \mathcal{X}_\pm$. We show that $\boldsymbol{f}^*$ defines an optimal transport map between $\boldsymbol{\mu}_+$ and ν^* with respect to the cost $\mathbf{c}$.

Lemma 25 *Assume that the problem is nested. Then, $\boldsymbol{f}^*$ is an optimal transport map between $\boldsymbol{\mu}_+$ and ν^* for the cost $\mathbf{c}$, with Kantorovich potential between ν^* and $\boldsymbol{\mu}_+$ given by $v : y \mapsto -2\int_0^y \kappa(t)\mathrm{d}t$. The same holds for $\boldsymbol{\mu}_-$, with Kantorovich potential given by $-v$.*

The proof of Lemma 25 relies on the following technical lemma, whose proof is postponned to Appendix A.4.

Lemma 26 *The function $y \mapsto \kappa(y)$ satisfies $|\kappa(y)| \leq C(1 + |y|)$ for some $C > 0$.*

Proof To prove Lemma 25, we begin by remarking that the potential v is in $L^1(\nu^*)$ because of Lemma 24 and Lemma 26. Let us now show that for almost all $x \in \mathcal{X}_+$,

$$v^c(x) := \sup_{y\in\mathbb{R}}(v(y) - c(x,y)) = v(f^*(x)) - c(x, f^*(x)). \tag{27}$$

Let $x \in \mathcal{X}_+$ be a point such that $y \mapsto g_y^{\kappa(y)}(x)$ is nonincreasing (almost all points satisfy this condition by nestedness). We note that $\partial_y c(x,y) = \frac{2(y-\eta(x))}{\Delta(x)}$. Hence,

$$\begin{aligned}
c(x,y) - c(x, f^*(x)) &= \int_{f^*(x)}^{y} \frac{2(t-\eta(x))}{\Delta(x)}\mathrm{d}t \\
&= -2\int_{f^*(x)}^{y} \left(\frac{\eta(x)-t}{\Delta(x)} - \kappa(t)\right)\mathrm{d}t - 2\int_{f^*(x)}^{y}\kappa(t)\mathrm{d}t \\
&= -2\int_{f^*(x)}^{y} \left(\frac{\eta(x)-t}{\Delta(x)} - \kappa(t)\right)\mathrm{d}t + v(y) - v(f^*(x)).
\end{aligned}$$

Assume that $y \geq f^*(x)$. For $t \in (f^*(x), y]$, by definition of f^* and by nestedness, $g_t^{\kappa(t)}(x) = 0$. Thus,

$$\frac{\eta(x)-t}{\Delta(x)} - \kappa(t) < 0.$$

This implies that

$$-2\int_{f(x)}^{y} \left(\frac{\eta(x)-t}{\Delta(x)} - \kappa(t)\right)\mathrm{d}t \geq 0.$$

We obtain that

$$c(x,y) - c(x, f^*(x)) \geq v(y) - v(f^*(x)).$$

The same result holds when $y < f^*(x)$. Indeed, in that case, for all $t \in [y, f^*(x))$,

$$\frac{\eta(x)-t}{\Delta(x)} - \kappa(t) \geq 0.$$

Hence,

$$-2\int_{f^*(x)}^{y} \left(\frac{\eta(x)-t}{\Delta(x)} - \kappa(t)\right)\mathrm{d}t = 2\int_{y}^{f^*(x)} \left(\frac{\eta(x)-t}{\Delta(x)} - \kappa(t)\right)\mathrm{d}t \geq 0.$$

This proves (27). This relation implies that the $\mathbf{c}$-transform of the function $v \in L^1(\nu^*)$ is a function $w : \Omega \to \mathbb{R}$ satisfying for μ_+-almost all $x \in \mathcal{X}_+$ (with $\mathbf{z} = (\eta(x), \Delta(x))$)

$$w(\mathbf{z}) = v(f^*(x)) - c(x, f^*(x)) = v(\boldsymbol{f}^*(\mathbf{z})) - \mathbf{c}(\mathbf{z}, \boldsymbol{f}^*(\mathbf{z})).$$

As $v \in L^1(\nu^*)$, Kantorovich duality implies

$$\begin{aligned}
\mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_+, \nu^*) &\geq \int v(y)\mathrm{d}\nu^*(y) - \int w(\mathbf{z})\mathrm{d}\boldsymbol{\mu}_+(\mathbf{z}) = \int v(\boldsymbol{f}^*(\mathbf{z}))\mathrm{d}\boldsymbol{\mu}_+(\mathbf{z}) - \int w(\mathbf{z})\mathrm{d}\boldsymbol{\mu}_+(\mathbf{z}) \\
&= \int \mathbf{c}(\mathbf{z}, \boldsymbol{f}^*(\mathbf{z}))\mathrm{d}\boldsymbol{\mu}_+(\mathbf{z}),
\end{aligned}$$

see Section 2. This shows that $\boldsymbol{f}^*$ is the optimal transport map between $\boldsymbol{\mu}_+$ and ν^*. The same holds for $\boldsymbol{\mu}_-$, where we use the potential $-v$ instead of v: precisely, we can show that we have for almost all $x \in \mathcal{X}_-$

$$(-v)^c(x) := \sup_{y\in\mathbb{R}}(-v(y) - c(x,y)) = -v(f^*(x)) - c(x, f^*(x)). \tag{28}$$

This concludes the proof of Lemma 25. ■

Lemma 25 shows that $\boldsymbol{f}^*$ defines an optimal transport map from $\boldsymbol{\mu}_+$ to ν^*, and from $\boldsymbol{\mu}_-$ to ν^*. To conclude the proof of Theorem 4, it remains to show that ν^* is solution to the barycenter problem described in Lemma 10.

Lemma 27 *The distribution ν^* is solution to the barycenter problem described in Lemma 10.*

Proof Let $\varphi : \mathbf{z}_1 \in \Omega \mapsto \mathbf{c}(\mathbf{z}_1, \boldsymbol{f}^*(\mathbf{z}_1)) - v(\boldsymbol{f}^*(\mathbf{z}_1))$ and let $\psi : \mathbf{z}_2 \in \Omega \mapsto \mathbf{c}(\mathbf{z}_2, \boldsymbol{f}^*(\mathbf{z}_2)) + v(\boldsymbol{f}^*(\mathbf{z}_2))$. Using (27) and (28), we see that for all $y \in \mathbb{R}$, for $\boldsymbol{\mu}_+$-almost all $\mathbf{z}_1$ and $\boldsymbol{\mu}_-$-almost all $\mathbf{z}_2$, it holds that

$$\begin{aligned}
\varphi(\mathbf{z}_1) + \psi(\mathbf{z}_2) &= \mathbf{c}(\mathbf{z}_1, \boldsymbol{f}^*(\mathbf{z}_1)) - v(\boldsymbol{f}^*(\mathbf{z}_1)) + \mathbf{c}(\mathbf{z}_2, \boldsymbol{f}^*(\mathbf{z}_2)) + v(\boldsymbol{f}^*(\mathbf{z}_2)) \\
&\leq \mathbf{c}(\mathbf{z}_1, y) - v(y) + \mathbf{c}(\mathbf{z}_2, y) + v(y) \\
&= \mathbf{c}(\mathbf{z}_1, y) + \mathbf{c}(\mathbf{z}_2, y).
\end{aligned}$$

By taking the value y that minimizes this last term, we obtain that

$$\varphi(\mathbf{z}_1) + \psi(\mathbf{z}_2) \leq \mathbf{C}(\mathbf{z}_1, \mathbf{z}_2),$$

where $\mathbf{C}$ is the cost function defined in (15). In particular, $-\varphi(\mathbf{z}_1) \geq \psi^{\mathbf{C}}(\mathbf{z}_1)$. Furthermore, observe that

$$-v(\boldsymbol{f}^*(\mathbf{z}_1)) \leq \varphi(\mathbf{z}_1) \leq c(\boldsymbol{x}_1, 0).$$

Thus, as $v \in L^1(\nu^*)$ and $\int \frac{h^2}{d}\mathrm{d}\boldsymbol{\mu}_+(h,d) < +\infty$, it holds that $\varphi \in L^1(\boldsymbol{\mu}_+)$. Likewise, $\psi \in L^1(\boldsymbol{\mu}_-)$. By Kantorovich duality, it holds that

$$\begin{aligned}
\mathrm{OT}_{\mathbf{C}}(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-) &\geq \int \psi(\mathbf{z}_2)\mathrm{d}\boldsymbol{\mu}_-(\mathbf{z}_2) - \int \psi^{\mathbf{C}}(\mathbf{z}_1)\mathrm{d}\boldsymbol{\mu}_+(\mathbf{z}_1) \\
&\geq \int \psi(\mathbf{z}_2)\mathrm{d}\boldsymbol{\mu}_-(\mathbf{z}_2) + \int \varphi(\mathbf{z}_1)\mathrm{d}\boldsymbol{\mu}_+(\mathbf{z}_1) \\
&= \int \mathbf{c}(\mathbf{z}_1, \boldsymbol{f}^*(\mathbf{z}_1))\mathrm{d}\boldsymbol{\mu}_+(\mathbf{z}_1) + \int \mathbf{c}(\mathbf{z}_2, \boldsymbol{f}^*(\mathbf{z}_2))\mathrm{d}\boldsymbol{\mu}_-(\mathbf{z}_2) \\
&\quad - \int v(\boldsymbol{f}^*(\mathbf{z}_1))\mathrm{d}\boldsymbol{\mu}_+(\mathbf{z}_1) + \int v(\boldsymbol{f}^*(\mathbf{z}_2))\mathrm{d}\boldsymbol{\mu}_-(\mathbf{z}_2) \\
&= \mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_+, \nu^*) + \mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_-, \nu^*) + \int v(y)\mathrm{d}\nu^*(y) - \int v(y)\mathrm{d}\nu^*(y) \\
&= \mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_+, \nu^*) + \mathrm{OT}_{\mathbf{c}}(\boldsymbol{\mu}_-, \nu^*) \geq \mathrm{OT}_{\mathbf{C}}(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-).
\end{aligned}$$

This proves that ν^* is the solution to the barycenter problem, and that f^* is an optimal regression function. ■

5. Building examples and counterexamples

In the previous section, we proved that under mild assumptions, the relationship $g_y^*(x) = \mathbf{1}\{f^*(x) \geq y\}$ only holds under the nestedness assumption. In this section, we now explain how to build large classes of triplets $(X, Y, S) \in \mathcal{X} \times \mathbb{R} \times \{1, 2\}$ whose distributions $\mathbb{P}$ either satisfy or do not satisfy this criterion. The starting point of our approach consisted in associating to each distribution $\mathbb{P}$ a pair of distributions $(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-) = (\boldsymbol{\mu}_+(\mathbb{P}), \boldsymbol{\mu}_-(\mathbb{P}))$ on Ω, where we recall that $\boldsymbol{\mu}_\pm(\mathbb{P})$ is the distribution of $(\eta(X), \Delta(X))$ when $X \sim \mu_\pm$. Then, both the optimal fair regression function and the nestedness criterion are best understood in terms of the pair $(\boldsymbol{\mu}_+(\mathbb{P}), \boldsymbol{\mu}_-(\mathbb{P}))$.

However, given a pair of measure $(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-)$ on Ω, it is not a priori clear whether there exists a triplet $(X, Y, S) \sim \mathbb{P}$ with $(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-) = (\boldsymbol{\mu}_+(\mathbb{P}), \boldsymbol{\mu}_-(\mathbb{P}))$. We give a definitive answer to this problem by providing a list of necessary and sufficient conditions for the existence of such a probability distribution $\mathbb{P}$. We then use this theoretical result to build probability distributions $\mathbb{P}$ for which the associated fair classification problem is either nested or not nested.

Let $\mathbb{P}$ be the distribution of a triplet $(X, Y, S) \in \mathcal{X} \times \mathbb{R} \times \{1, 2\}$, with $\mathbb{E}[Y^2] < +\infty$. Let $(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-) = (\boldsymbol{\mu}_+(\mathbb{P}), \boldsymbol{\mu}_-(\mathbb{P}))$ be the associated pair of measures on Ω. Then, it always holds that

$$\int_\Omega |d|^{-1} \mathrm{d}\boldsymbol{\mu}_+(h, d) = \int_\mathcal{X} \frac{1}{\Delta(x)} \mathrm{d}\mu_+(x) = \int_\mathcal{X} \frac{\mathrm{d}\mu}{\mathrm{d}\mu_+}(x) \mathrm{d}\mu_+(x) = \mu(\mathcal{X}_+),$$

while $\int_\Omega |d|^{-1} \mathrm{d}\boldsymbol{\mu}_-(h, d) = \mu(\mathcal{X}_-)$. In particular,

$$0 < \int_\Omega |d|^{-1} \mathrm{d}\boldsymbol{\mu}_+(h, d) + \int_\Omega |d|^{-1} \mathrm{d}\boldsymbol{\mu}_-(h, d) \leq 1. \tag{29}$$

Also, Observe that $\mu = p_1\mu_1 + p_2\mu_2$, so that $\Delta(x) = \frac{\mathrm{d}\mu_+}{\mathrm{d}\mu}(x) \leq \frac{1}{p_1 m}$ when $x \in \mathcal{X}_+$, whereas $\Delta(x) \geq -\frac{1}{p_2 m}$ when $x \in \mathcal{X}_-$ (recall that m is the mass of the measure $(\mu_1 - \mu_2)_+$). In particular, the supports of $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ are located in an horizontal strip of the form $\{(h, d) : -M \leq d \leq M\}$ for some $M > 0$. The next proposition states that these two conditions are actually sufficient for the existence of a probability measure $\mathbb{P}$ with $(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-) = (\boldsymbol{\mu}_+(\mathbb{P}), \boldsymbol{\mu}_-(\mathbb{P}))$.

Proposition 28 *Assume that $\mathcal{X}$ is an uncountable standard Borel space (e.g., $\mathcal{X} = [0, 1]$). Then, the set of pairs of measures $(\boldsymbol{\mu}_+(\mathbb{P}), \boldsymbol{\mu}_-(\mathbb{P}))$ that can be obtained from a distribution $\mathbb{P}$ of a triplet $(X, Y, S) \in \mathcal{X} \times \mathbb{R} \times \{1, 2\}$ with $\mathbb{E}[Y^2] < \infty$ is exactly equal to the set of pairs $(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-)$ supported on bounded horizontal strips, satisfying Equation (29), with $\boldsymbol{\mu}_+$ supported on $\{d > 0\}$ and $\boldsymbol{\mu}_-$ supported on $\{d < 0\}$.*

This proposition allows us to easily build examples where either nestedness or nonnestedness is satisfied: one does not need to build from scratch a joint distribution on $\mathcal{X} \times \mathbb{R} \times \{1, 2\}$, but can simply define a pair of measures $(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-)$ on Ω. As long as this pair satisfies the conditions given in Proposition 28, the existence of a probability distribution $\mathbb{P}$ such that $(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-) = (\boldsymbol{\mu}_+(\mathbb{P}), \boldsymbol{\mu}_-(\mathbb{P}))$ is ensured.

Proof We have already established that the pairs of measures $(\boldsymbol{\mu}_+(\mathbb{P}), \boldsymbol{\mu}_-(\mathbb{P}))$ satisfy the conditions stated in Proposition 28. Reciprocally, consider a pair $(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-)$ satisfying Equation (29), supported on bounded horizontal strips, with $\boldsymbol{\mu}_+$ supported on $\{d > 0\}$ and $\boldsymbol{\mu}_-$ supported on $\{d < 0\}$. Let $a_\pm = \int_\Omega |d|^{-1} \mathrm{d}\boldsymbol{\mu}_\pm(h, d)$.

Due to the Borel isomorphism theorem, $\mathcal{X}$ is Borel isomorphic to $\mathbb{R}^2$, so we may assume without loss of generality that $\mathcal{X} = \mathbb{R}^2$. Let $\mathcal{X}_+ = \{(h, d) \in \mathbb{R}^2 : \ d > 0\}$, $\mathcal{X}_- = \{(h, d) \in \mathbb{R}^2 : \ d < 0\}$ and $\mathcal{X}_= = \{(h, 0) : \ h \in \mathbb{R}\}$. Let $\mu_= = \delta_0$. We let $\mu_+ = \boldsymbol{\mu}_+$ and $\mu_- = \boldsymbol{\mu}_-$.

Let

$$\mathrm{d}\mu(h, d) = \frac{1}{|d|}\mathrm{d}\mu_+(h, d) + \frac{1}{|d|}\mathrm{d}\mu_-(h, d) + (1 - a_+ - a_-)\mathrm{d}\mu_=(h, d).$$

Note that μ is a probability measure:

$$\int \mathrm{d}\mu = \int \frac{1}{|d|}\mathrm{d}\mu_+(h, d) + \int \frac{1}{|d|}\mathrm{d}\mu_-(h, d) + (1 - a_+ - a_-)\int \mathrm{d}\mu_= = 1.$$

Consider m small enough so that the inequality $m|d|/2 \leq 1$ holds on the support of μ (this is possible because the d coordinate is bounded in the support of $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$). We define

$$\mathrm{d}\mu_1(h, d) = (1 + \frac{m}{2}d)\mathrm{d}\mu(h, d) \quad \text{and} \quad \mathrm{d}\mu_2(h, d) = (1 - \frac{m}{2}d)\mathrm{d}\mu(h, d)$$

Observe that $\mu = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2$. Also, μ_1 and μ_2 are probability measures, as

$$\int d\mathrm{d}\mu = \int \frac{d}{|d|}\mathrm{d}\mu_+(h, d) + \int \frac{d}{|d|}\mathrm{d}\mu_-(h, d) = 1 - 1 = 0.$$

Let $\eta(h, d) = h$. We define the triplet (X, Y, S) by letting S be uniform on $\{1, 2\}$. If $S = 1$, we draw $X \sim \mu_1$ and let $Y = \eta(X)$. If $S = 2$, we draw $X \sim \mu_2$ and let $Y = \eta(X)$. Let $\mathbb{P}$ be the distribution of (X, Y, S). One can easily check that $(\boldsymbol{\mu}_+(\mathbb{P}), \boldsymbol{\mu}_-(\mathbb{P})) = (\boldsymbol{\mu}_+, \boldsymbol{\mu}_-)$, as desired. ■

To build examples of nested and non-nested problems, we consider probability measures $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ supported on small horizontal segments:

$$\boldsymbol{\mu}_\pm = \frac{1}{K}\sum_{i=1}^{K} \nu_\pm^{(i)}$$

where $\nu_\pm^{(i)}$ is the uniform measure on $[a_\pm^{(i)}, a_\pm^{(i)} + 1] \times \{d_\pm^{(i)}\}$.

Example 1 (A nested classification problem) *Take $K = 1$, $d_+^{(1)} = d_-^{(1)} = 1$ and $a_+^{(1)} = 0$, $a_-^{(1)} = -1$. Let $\mathbb{P}$ be the probability associated with the pair $(\boldsymbol{\mu}_+, \boldsymbol{\mu}_-)$ defined for this choice of parameters. Then, it holds that $1/2 \in I(y)$ for all $y \in \mathbb{R}$. By choosing $\kappa(y) = 1/2$ for all $y \in \mathbb{R}$, we see that the classification problem associated with $\mathbb{P}$ is nested. See also Figure 2.*

Example 2 (A non-nested classification problem) *Take $K = 2$. Let $d_+^{(1)} = d_-^{(1)} = 1$ and $a_+^{(1)} = a_-^{(1)} = 0$. Let $d_+^{(2)} = d_-^{(2)} = 1/2$, and $a_+^{(2)} = -1$, $a_-^{(2)} = -6$. Then, for $y = 0$, $I(y) = \{0\}$, so the support of $\nu_+^{(2)}$ is to the left of the classification threshold for $y = 0$. But for $y = -3$, we have $I(y) = \{4\}$ and the support of $\nu_+^{(2)}$ is to the right of the classification threshold. Hence, the classification problem is non-nested. See also Figure 2.*

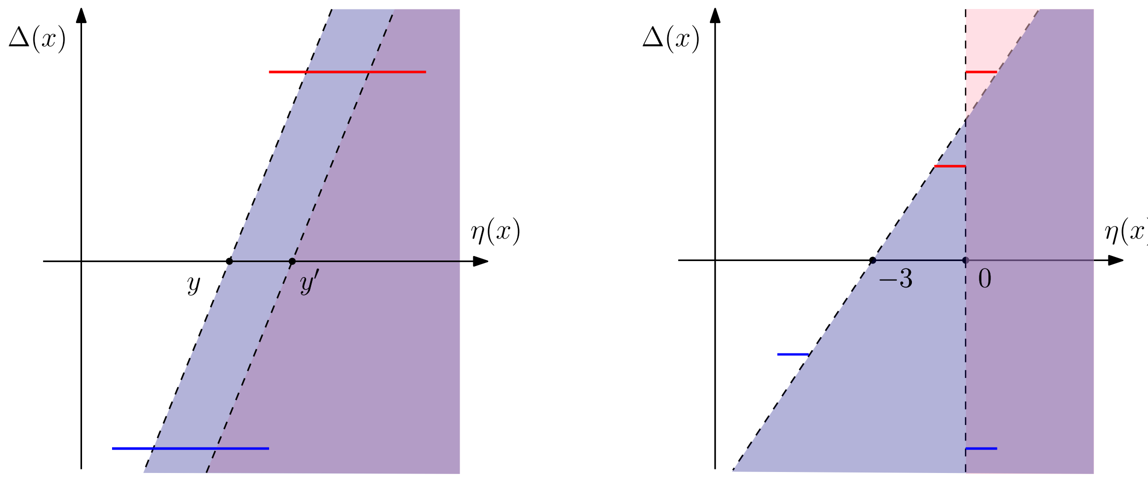

Figure 2: Left: example of a nested problem. The distributions of $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ are depicted in red and blue, corresponding to the distributions given in Example 1. The acceptance region for $g_y^{\kappa(y)}$ and $g_{y'}^{\kappa'(y)}$ are so that the masses of $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ to the right of the decision boundaries are equal. One can observe that these two regions are nested. Right: example of a non-nested problem. The distributions $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ are the ones described in Example 2. The region in pink is rejected for $y = -3$ but accepted for $y = 0$, contradicting the nestedness assumption.

6. Conclusion and future work

This work presents the first theoretical characterization of the optimal fair regression function in the unawareness framework as the solution to a barycenter problem with an optimal transport cost. Our results also demonstrate that, under the nestedness assumption, the optimal fair regression function can be represented by the family of classifiers $g_y^{\kappa(y)}$. Although both approaches—whether based on optimal transport or cost-sensitive classifiers—depend on the underlying distribution $\mathbb{P}$ which is generally unknown, they pave the way for developing new algorithms that estimate these unknown quantities from observed data. Designing such estimators, along with bounding their excess risk and potential unfairness, provides critical insight into fair algorithm.

While this work provides an initial characterization of the optimal fair regression function in the unawareness framework, it also has notable limitations. For instance, our results are currently limited to cases where the sensitive attribute is binary and apply only to univariate regression. Addressing these limitations and extending our findings to more general cases would be a valuable direction for future research.

Acknowledgments

The authors thank Evgenii Chzhen and Nicolas Schreuder for valuable and insightful discussions. S. G. gratefully acknowledges funding from the Fondation Mathmatique Jacques Hadamard.

Appendix A. Additional proofs

A.1 Proof of Lemma 13

Let $f:\mathbb{R}\to\mathbb{R}\cup\{+\infty\}$ be a lower semicontinuous convex function. The domain of such a function is an interval $\mathrm{dom}(f)$. Its right derivative f'_+ is defined and finite everywhere on $\mathrm{dom}(f)$, except on the right endpoint of the interval (should the right endpoint be included in $\mathrm{dom}(f)$) where it is equal to $+\infty$. Such a function is upper semicontinuous, with the representation:

$$\forall h\in\mathrm{dom}(f),\ f'_+(h)=\inf_{u>h}\frac{f(u)-f(h)}{u-h}, \tag{30}$$

where the infimum can be restricted to a countable dense collection of values u if needed. Likewise, the right derivative f'_- can be defined on $\mathrm{dom}(f)$, and is lower semicontinuous. Note also that the oscillation function $\mathrm{osc}(f)=f'_+-f'_-\in[0,+\infty]$ can be defined on $\mathrm{dom}(f)$, and is upper semicontinuous. Indeed, only one of f'_+ and f'_- can be infinite on $\mathrm{dom}(f)$ (and only at one of the endpoints of the domain), so that the difference is well defined.

Recall that a function φ is $\mathbf{C}$-convex if

$$\forall (h,d)\in\Omega,\ \varphi(h,d)=\sup_{(h',d')\in\Omega}\left(\varphi^{\mathbf{C}}(h',d')-\frac{(h-h')^2}{|d|+|d'|}\right). \tag{31}$$

The function φ is lower semicontinuous as a supremum of continuous functions. Furthermore, for any $d\neq 0$, the function

$$\varphi_d: h\mapsto \varphi(h,d)+\frac{h^2}{|d|}=\sup_{(h',d')\in\Omega}\left(\varphi^{\mathbf{C}}(h',d')+\frac{h^2}{|d|}-\frac{(h-h')^2}{|d|+|d'|}\right)$$

is convex as a supremum of lower semicontinuous convex functions. Let $G:\mathrm{dom}(\varphi)\to[0,+\infty]$ be defined as $G(h,d)=\mathrm{osc}(\varphi_d)(h)$ for $(h,d)\in\mathrm{dom}(\varphi)$. Let $L>0$, and consider the set $\Sigma_{d,L}$ defined as the set of points $h\in\mathrm{dom}(\varphi_d)$ such that $G(h,d)\geq L^{-1}$, $(\varphi_d)'_+(h)\geq -L$ and $(\varphi_d)'_-(h)\leq L$. As the left and right derivatives of φ_d are nondecreasing, the set $\Sigma_{d,L}$ is finite and its cardinality is bounded by a constant depending only on L. Let $\Sigma_L=\bigcup_{d\neq 0}\Sigma_{d,L}$ and $\Sigma=\bigcup_{L\in\mathbb{N}}\Sigma_L$. The set of points $\mathbf{z}\in\mathrm{dom}(\varphi)$ such that $\partial_h\varphi(\mathbf{z})$ does not exist is equal to Σ. Let us show that for any integer L, the set Σ_L is included in a countable union of graphs of measurable functions.

To do so, we use the following general result, see (Aliprantis and Border, 1999, Corollary 18.14). A correspondence Φ from a measurable set S to a topological space X assigns to each $s\in S$ a subset $\Phi(s)$ of X. We say that the correspondence is (weakly) measurable if for each open subset $U\subset X$, the set $\Phi^\ell(U)=\{s\in S:\ \Phi(s)\cap U\neq\emptyset\}$ is measurable.

Theorem 29 (Castaing's theorem) *If X is a Polish space and Φ is a measurable correspondence with non-empty closed values between S and X, then there exists a sequence $(f_n)_n$ of measurable functions $S\to X$ such that for every $s\in S$, $\Phi(s)=\overline{\{f_1(s),f_2(s),\cdots\}}$.*

Let Φ be the correspondence that assigns to each $d\neq 0$ the subset $\Sigma_{d,L}\cup\{0\}$ of $\mathbb{R}$. As each set $\Sigma_{d,L}$ is finite, this correspondence takes non-empty closed values. If we show that this correspondence is measurable, then Castaing theorem asserts the existence of a sequence of measurable functions $(f_n)_n$ such that for every $d \neq 0$, $\Sigma_{d,L} \cup \{0\} = \overline{\{f_1(d), f_2(d), \cdots\}}$. For each d, the set $\Sigma_{d,L}$ is finite, so that $\Sigma_{d,L} \cup \{0\} = \{f_1(d), f_2(d), \cdots\}$, implying that Σ_L is included in a countable union of graphs of measurable functions. It remains to show the measurability of Φ.

Lemma 30 *The function G is measurable.*

Proof The representation (30) implies that $(h,d) \mapsto (\varphi_d)'_+(h)$ is given by a countable infimum of measurable functions, and is therefore measurable. Likewise, $(h,d) \mapsto (\varphi_d)'_-(h)$ is measurable, so that G is also measurable. ■

Let $U \subset \mathbb{R}$ be an open set. If $0 \in U$, then $\Phi^\ell(U) = \mathbb{R}\backslash\{0\}$ is measurable. If $0 \notin U$, we have

$$\Phi^\ell(U) = \{d \neq 0 : \ \exists h \in [-L, L] \cap U,\ G(h,d) \geq L^{-1},\ (\varphi_d)'_+(h) \geq -L,\ (\varphi_d)'_-(h) \leq L\}.$$

This set is the projection on the d-axis of the measurable set

$$B = \{(h,d) \in \Omega : \ h \in [-L, L] \cap U,\ G(h,d) \geq L^{-1},\ (\varphi_d)'_+(h) \geq -L,\ (\varphi_d)'_-(h) \leq L\}$$

Furthermore, for each d, the section $\{h \in \mathbb{R} : \ (h,d) \in B\} = \Sigma_{d,L}$ is compact. By Theorem 4.7.11 in Srivastava (2008), this implies that $\Phi^\ell(U)$ is measurable, concluding the proof of Lemma 13.

A.2 Proof of Proposition 14

Classical manipulations show that the risk $\mathcal{R}_y(g)$ of a classifier g can be expressed as

$$\begin{aligned} \mathcal{R}_y(g) &= y\mathbb{E}\left[(1-Y)g(X)\right] + (1-y)\mathbb{E}\left[Y(1-g(X))\right] \\ &= (1-y)\mathbb{E}[Y] + \mathbb{E}\left[g(X)\left(y - \eta(X)\right)\right]. \end{aligned}$$

Using the definition of μ_+, μ_- and Δ given in Section 3, we find that

$$\begin{aligned} &\mathbb{E}\left[g(X)\left(y-\eta(X)\right)\right] \\ &= \int_{\mathcal{X}_+} g(x)(y-\eta(x))\frac{\mathrm{d}\mu}{\mathrm{d}\mu_+}(x)\mathrm{d}\mu_+(x) + \int_{\mathcal{X}_-} g(x)(y-\eta(x))\frac{\mathrm{d}\mu}{\mathrm{d}\mu_-}(x)\mathrm{d}\mu_-(x) \\ &\quad + \int_{\mathcal{X}_=} g(x)(y-\eta(x))\mathrm{d}\mu(x) \\ &= \int_{\mathcal{X}_+} g(x)\frac{y-\eta(x)}{|\Delta(x)|}\mathrm{d}\mu_+(x) + \int_{\mathcal{X}_-} g(x)\frac{y-\eta(x)}{|\Delta(x)|}\mathrm{d}\mu_-(x) + \int_{\mathcal{X}_=} g(x)(y-\eta(x))\mathrm{d}\mu(x). \end{aligned}$$

Moreover, Lemma 9 implies that the demographic parity constraint is equivalent to the constraint $\mathbb{E}_{X\sim\mu_+}[g(X)] = \mathbb{E}_{X\sim\mu_-}[g(X)]$. Using the decomposition $g = \mathcal{F}(g_+, g_-, g_=)$, we see that the fair classification problem can be rephrased as follows

$$\begin{cases} \text{minimize} \quad \mathbb{E}_{\mu_+}\left[g_+(X)\frac{y-\eta(X)}{\Delta(X)}\right] - \mathbb{E}_{\mu_-}\left[g_-(X)\frac{y-\eta(X)}{\Delta(X)}\right] \\ \qquad\qquad\quad +\mathbb{E}_{\mu}\left[\mathbb{1}_{\mathcal{X}_=}(X)g_+(X)(y-\eta(X))\right] \\ \text{such that} \quad \mathbb{E}_{\mu_+}[g_+(X)] = \mathbb{E}_{\mu_-}[g_-(X)]. \end{cases} \tag{C'_y}$$

The following lemma characterizes the solutions to the problem (C'_y).

Lemma 31 *Under Assumption A, for any optimal classifier g, there exist κ^+, κ^- such that $g = \mathcal{F}(g^{\kappa^+}, g^{\kappa^-}, g_=)$, with*

$$\begin{aligned} g_=(x) = \mathbb{1}\{\eta(x) > y\} \quad &\textit{or} \quad g_=(x) = \mathbb{1}\{\eta(x) \geq y\}, \\ \textit{and} \quad g^\kappa(x) = \mathbb{1}&\left\{\eta(x) \geq y + \kappa\Delta(x)\right\}. \end{aligned}$$

To conclude the proof of Proposition 14, it remains to prove that all optimal classifier are a.s. equal when $\Delta(X) \neq 0$, and that the optimal classifier can be chosen as $g^* = \mathcal{F}(g^{\kappa^*}, g^{\kappa^*}, g_=)$ for some κ^*.

Denote by F_+ the c.d.f. of the random variable $Z_+ = \frac{\eta(X)-y}{\Delta(X)}$ when $X \sim \mu_+$ and by F_- the c.d.f. of $Z_- = \frac{y-\eta(X)}{\Delta(X)}$ when $X \sim \mu_-$. Let $\mathcal{Q}_+$ (resp. $\mathcal{Q}_-$) be the associated quantile function. To satisfy the demographic parity constraint, the classifier $\mathcal{F}(g^{\kappa^+}, g^{\kappa^-}, g_=)$ must be such that

$$F_+(\kappa^+) = F_-(-\kappa^-)$$

(recall that $\Delta(X) < 0$ when $X \sim \mu_-$, so that $g^{\kappa^-}(X) = 1$ if and only if $Z_- \geq -\kappa^-$). Denoting $\beta = F_+(\kappa^+) = F_-(-\kappa^-)$ and using the definition of the quantile function, we see that the law of $g^{\kappa^\pm}(X)$, where $X \sim \mu_\pm$ is equal to the law of

$$\mathbf{1}\{U \geq \beta\} = \mathbf{1}\{\mathcal{Q}_\pm(U) \geq \pm\kappa^\pm\},$$

where U is a uniform random variable on $[0,1]$.

Then, minimizing the risk of the classifier is equivalent to maximizing

$$\mathbb{E}\left[(\mathcal{Q}_+(U) + \mathcal{Q}_-(U))\,\mathbb{1}\left\{U \geq \beta\right\}\right] \tag{32}$$

Since F_+ and F_- are continuous, $u \to \mathcal{Q}_+(u) + \mathcal{Q}_-(u)$ is strictly increasing and left-continuous. Straightforward computations show that the expression in (32) has a unique maximum, which is attained for

$$\beta^* = \max\left\{\beta : \mathcal{Q}_+(\beta) + \mathcal{Q}_-(\beta) \leq 0\right\}.$$

Hence, it holds that $F_+(\kappa^+) = F_-(-\kappa^-) = \beta^*$. Let $g_1^* = \mathcal{F}(g^{\kappa_1^+}, g^{\kappa_1^-}, g_=)$ and $g_2^* = \mathcal{F}(g^{\kappa_2^+}, g^{\kappa_2^-}, g_=)$ be two optimal classifiers. For $\Delta(X) > 0$, they take different values only if $Z_+ \in [\kappa_1^+, \kappa_2^+]$. As $F_+(\kappa_1^+) = F_+(\kappa_2^+) = \beta^*$, this happens with zero probability. Likewise, the two classifiers are a.s. equal when $\Delta(X) < 0$.

It remains to show that we can pick $\kappa^+ = \kappa^-$. If $\mathcal{Q}_+ + \mathcal{Q}_-$ is continuous at β^*, the proof is complete: in this case, $\mathcal{Q}_+(\beta^*) + \mathcal{Q}_-(\beta^*) = 0$, and the choice $\kappa^* = \mathcal{Q}_+(\beta^*) = -\mathcal{Q}_-(\beta^*)$ satisfies $F_+(\kappa^*) = F_-(-\kappa^*) = \beta^*$.

Otherwise, $\mathcal{Q}_+(\beta^*) + \mathcal{Q}_-(\beta^*) < 0$. Defining

$$q_+ = \liminf_{\beta \to \beta^*_+} \mathcal{Q}_+(\beta) \quad \text{and} \quad q_- = \liminf_{\beta \to \beta^*_+} \mathcal{Q}_-(\beta),$$

we have $q_+ + q_- > 0$. Because $q_+ > -q_-$ and $\mathcal{Q}_+(\beta^*) < -\mathcal{Q}_-(\beta^*)$, there exists $\kappa^* \in [\mathcal{Q}_+(\beta^*), q_+] \cap [-q_-, -\mathcal{Q}_-(\beta^*)]$. By construction, $F_+(\kappa^*) = F_-(-\kappa^*) = \beta^*$, concluding the proof.

A.3 Proof of Lemma 31

Let g^* be a solution to the problem (C'_y), and let g^*_+, g^*_- and $g^*_=$ be the restrictions of g^* to $\mathcal{X}_+$, $\mathcal{X}_=$ and $\mathcal{X}_=$, so that $g^* = \mathcal{F}(g^*_+, g^*_-, g^*_=)$. Straightforward computations show that we necessarily have $g^*_=(X) = \mathbb{1}\{\eta(X) \geq y\}\mathbb{1}_{\mathcal{X}_=}(X)$ or $g^*_=(X) = \mathbb{1}\{\eta(X) > y\}\mathbb{1}_{\mathcal{X}_=}(X)$.

Now, let us assume (without loss of generality) that g^*_+ is not of the form g^{κ^+}. More precisely, assume that for κ^+ such that $\mathbb{E}_{\mu_+}[g^{\kappa^+}(X)] = \mathbb{E}_{\mu_+}[g^*_+(X)]$, we have $g^{\kappa^+}(X) \neq g^*_+(X)$ with positive μ_+-probability. Then, the classifier $\mathcal{F}(g^{\kappa^+}, g^*_-, g^*_=)$ satisfies the demographic parity constraint. Moreover, we have

$$\begin{aligned}
&\mathbb{E}_{\mu_+}\left[g^{\kappa^+}(X)\frac{y-\eta(X)}{\Delta(X)}\right] - \mathbb{E}_{\mu_+}\left[g^*_+(X)\frac{y-\eta(X)}{\Delta(X)}\right] \\
&= \mathbb{E}_{\mu_+}\left[\frac{y-\eta(X)}{\Delta(X)}g^{\kappa^+}(X)(1-g^*_+(X)) - \frac{y-\eta(X)}{\Delta(X)}(1-g^{\kappa^+}(X))g^*_+(X)\right] \\
&= \mathbb{E}_{\mu_+}\left[\left(\kappa^+ - \frac{\eta(X)-y}{\Delta(X)}\right)g^{\kappa^+}(X)(1-g^*_+(X))\right] - \kappa^+\mathbb{E}_{\mu_+}\left[g^{\kappa^+}(X)(1-g^*_+(X))\right] \\
&\quad - \mathbb{E}_{\mu_+}\left[\left(\kappa^+ - \frac{\eta(X)-y}{\Delta(X)}\right)(1-g^{\kappa^+}(X))g^*_+(X)\right] + \kappa^+\mathbb{E}_{\mu_+}\left[(1-g^{\kappa^+}(X))g^*_+(X)\right].
\end{aligned}$$

Since $\mathbb{E}_{\mu_+}[g^{\kappa^+}(X)] = \mathbb{E}_{\mu_+}[g^*_+(X)]$, we obtain that

$$\mathbb{E}_{\mu_+}\left[g^{\kappa^+}(X)(1-g^*_+(X))\right] = \mathbb{E}_{\mu_+}\left[(1-g^{\kappa^+}(X))g^*_+(X)\right].$$

Then, the definition of g^{κ^+} implies

$$\begin{aligned}
&\mathbb{E}_{\mu_+}\left[g^{\kappa^+}(X)\frac{y-\eta(X)}{\Delta(X)}\right] - \mathbb{E}_{\mu_+}\left[g^*_+(X)\frac{y-\eta(X)}{\Delta(X)}\right] \\
&= -\mathbb{E}_{\mu_+}\left[\left(\kappa^+ - \frac{\eta(X)-y}{\Delta(X)}\right)_-(1-g^*_+(X))\right] - \mathbb{E}_{\mu_+}\left[\left(\kappa^+ - \frac{\eta(X)-y}{\Delta(X)}\right)_+ g^*_+(X)\right] < 0.
\end{aligned}$$

This implies that $\mathcal{R}_y(\mathcal{F}(g^{\kappa^+}, g^*_-, g^*_=)) < \mathcal{R}_y(\mathcal{F}(g^*_+, g^*_-, g^*_=))$, which is absurd. Using a similar argument for g^*_-, we arrive to the conclusion that any optimal classifier is of the form $\mathcal{F}(g^{\kappa^+}, g^{\kappa^-}, g^*_=)$ with $g^*_=(X) = \mathbb{1}\{\eta(X) \geq y\}\mathbb{1}_{\mathcal{X}_=}(X)$ or $g^*_=(X) = \mathbb{1}\{\eta(X) > y\}\mathbb{1}_{\mathcal{X}_=}(X)$.

A.4 Proof of Lemma 26

Let us first show that κ is locally bounded. Let $[y_0, y_1]$ be a bounded interval. Then, if $\kappa(y) \geq M$ for some $y \in [y_0, y_1]$,

$$F(y) = \mu_+\left(\eta(X) \leq y + \kappa(y)\Delta(X)\right) \geq \mu_+\left(\eta(X) \leq y_0 + M\Delta(X)\right)$$

and

$$F(y) = \mu_-\left(\eta(X) \leq y + \kappa(y)\Delta(X)\right) \leq \mu_-\left(\eta(X) \leq y_1 + M\Delta(X)\right).$$

However, for M large enough,

$$\mu_-\left(\eta(X) \le y_1 + M\Delta(X)\right) < \mu_+\left(\eta(X) \le y_0 + M\Delta(X)\right),$$

and therefore it holds that $\kappa(y) \le M$ for all $y \in [y_0, y_1]$. Likewise, we show that for M large enough, $\kappa(y) \ge -M$ for all $y \in [y_0, y_1]$. Hence, κ is locally bounded.

We refine this argument to obtain a control on κ for large values of y. Let $L > 1$ and let $y > 1$ be such that $\kappa(y) \ge Ly$. Then, because of the definition of $\kappa(y)$ and as $\Delta(X) < 0$ for $X \sim \mu_-$,

$$\begin{aligned}\mu_+(\eta(X) \ge y) &\ge \mu_+(\eta(X) \ge y + \kappa(y)\Delta(X)) \\ &= \mu_-(\eta(X) \ge y + \kappa(y)\Delta(X)) \ge \mu_-(\eta(X) \ge y + Ly\Delta(X)).\end{aligned}$$

The complementary of the region $\{(h,d) \in \Omega : \ d < 0, \ h \ge y + Lyd\}$ in the lower half-plane $\{(h,d) \in \Omega : \ d < 0\}$ is contained in the set A_L given by the union of the horizontal strip $\{d \ge -1/L\}$ with the region $\{(h,d) \in \Omega : \ d < 0, \ h \le 1 + Ld\}$. For L large enough, $\boldsymbol{\mu}_-(A_L) < 1/2$. Then it holds that $\mu_+(\eta(X) \ge y) \ge 1/2$. For y large enough, this is not possible. Thus, we have shown that there exist $L > 0$ and $C > 0$ such that for $y > C$, we have $\kappa(y) \le Ly$. Likewise, we show that there exist constants $L, C > 0$ such that for $|y| > C$, $|\kappa(y)| \le L|y|$. As κ is also locally bounded, the conclusion follows.

References

Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 120–129. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/agarwal19d.html`.

Charalambos D. Aliprantis and Kim C. Border. *Measurable Correspondences*, pages 557–586. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999. ISBN 978-3-662-03961-8. doi: 10.1007/978-3-662-03961-8_17. URL `https://doi.org/10.1007/978-3-662-03961-8_17`.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias – Theres software used across the country to predict future criminals. And its biased against blacks. *ProPublica*, 2016.

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608 – 633, 2007. doi: 10.1214/009053606000001217. URL `https://doi.org/10.1214/009053606000001217`.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

Gerard Biau, Luc Devroye, and Gbor Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9:2015–2033, 09 2008. doi: 10.1145/1390681.1442799.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, 2009. doi: 10.1109/ICDMW.2009.83.

Pierre-André Chiappori, Robert J. McCann, and Brendan Pass. Multidimensional matching. *arXiv preprint arXiv:1604.05771*, 2016.

Pierre-Andr Chiappori, Robert J. McCann, and Brendan Pass. Multi-to one-dimensional optimal transport. *Communications on Pure and Applied Mathematics*, 70(12):2405–2444, 2017. doi: https://doi.org/10.1002/cpa.21707. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21707.

Evgenii Chzhen and Nicolas Schreuder. A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 2020a. URL https://api.semanticscholar.org/CorpusID:220830762.

Evgenii Chzhen and Nicolas Schreuder. An example of prediction which complies with demographic parity and equalizes group-wise risks in the context of regression. In *NeurIPS 2020 Workshop on Algorithmic Fairness through the Lens of Causality and Interpretability*, 2020b.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression via plug-in estimator and recalibration with statistical guarantees. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19137–19148. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/ddd808772c035aed516d42ad3559be5f-Paper.pdf.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with Wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33:7321–7331, 2020b.

Andrew Cotter, Maya Gupta, and Harikrishna Narasimhan. On making stochastic classifiers deterministic. *Advances in Neural Information Processing Systems*, 32, 2019.

Eustasio del Barrio, Paula Gordaliza, and Jean-Michel Loubes. Review of mathematical frameworks for fairness in machine learning. October 2020. URL https://hal.science/hal-02964972.

Christophe Denis, Romuald Elie, Mohamed Hebiri, and François Hu. Fairness guarantees in multi-class classification with demographic parity. *Journal of Machine Learning Research*, 25(130):1–46, 2024. URL http://jmlr.org/papers/v25/23-0322.html.

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*, KDD '15, page 259268, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2783311. URL https://doi.org/10.1145/2783258.2783311.

Solenne Gaucher, Nicolas Schreuder, and Evgenii Chzhen. Fair learning with Wasserstein barycenters for non-decomposable performance measures. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2436–2459. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/gaucher23a.html.

Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/dc4c44f624d600aa568390f1f1104aa0-Paper.pdf.

Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020.

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, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating MLs impact disparity require treatment disparity? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/8e0384779e58ce2af40eb365b318cc32-Paper.pdf.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. On the applicability of machine learning fairness notions. *SIGKDD Explor. Newsl.*, 23(1):1423, may 2021. ISSN 1931-0145. doi: 10.1145/3468507.3468511. URL https://doi.org/10.1145/3468507.3468511.

Pascal Massart and lodie Ndlec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5), October 2006. ISSN 0090-5364. doi: 10.1214/009053606000000786. URL http://dx.doi.org/10.1214/009053606000000786.

Robert J McCann and Brendan Pass. Optimal transportation between unequal dimensions. *Archive for Rational Mechanics and Analysis*, 238(3):1475–1520, 2020.

Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/menon18a.html.

Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Serena Wang. Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5248–5255, 2020.

Luca Oneto and Silvia Chiappa. *Fairness in Machine Learning*, page 155196. Springer International Publishing, 2020. ISBN 9783030438838. doi: 10.1007/978-3-030-43883-8_7. URL http://dx.doi.org/10.1007/978-3-030-43883-8_7.

Luca Oneto, Michele Donini, and Massimiliano Pontil. General fair empirical risk minimization. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020. doi: 10.1109/IJCNN48605.2020.9206819.

Brendan Pass. Regularity of optimal transportation between spaces with different dimensions. *Mathematical Research Letters*, 19(2):291–307, 2012.

Aldo Pratelli. On the equality between Monge's infimum and Kantorovich's minimum in optimal mass transportation. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 43, pages 1–13. Elsevier, 2007.

Filippo Santambrogio. Optimal Transport for Applied Mathematicians. *Birkäuser, NY*, 55 (58-63):94, 2015.

Nicolas Schreuder and Evgenii Chzhen. Classification with abstention but without disparities. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1227–1236. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/schreuder21a.html.

Sashi Mohan Srivastava. *A Course on Borel sets*, volume 180. Springer Science & Business Media, 2008.

Gayane Taturyan, Evgenii Chzhen, and Mohamed Hebiri. Regression under demographic parity constraints via unlabeled post-processing. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 117917–117953. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d5c3ecf397fff63419bb5f5f2d8afe33-Paper-Conference.pdf.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

Serena Wang and Maya Gupta. Deontological ethics by monotonicity shape constraints. In *International conference on artificial intelligence and statistics*, pages 2043–2054. PMLR, 2020.

Ruicheng Xian, Lang Yin, and Han Zhao. Fair and optimal classification via post-processing. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 37977–38012. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/xian23b.html.

Forest Yang, Mouhamadou Cisse, and Sanmi Koyejo. Fairness with overlapping groups; a probabilistic perspective. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4067–4078. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/29c0605a3bab4229e46723f89cf59d83-Paper.pdf.

Yuhong Yang. Minimax nonparametric classification. I. Rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, 1999. doi: 10.1109/18.796368.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019. URL http://jmlr.org/papers/v20/18-262.html.

Xianli Zeng, Edgar Dobriban, and Guang Cheng. Fair Bayes-optimal classifiers under predictive parity. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27692–27705. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1d9c7e7bd265d81aae8d74a7a6bd7f1-Paper-Conference.pdf.

Xianli Zeng, Edgar Dobriban, and Guang Cheng. Bayes-optimal classifiers under group fairness. *ArXiv*, abs/2202.09724, 2022b. URL https://api.semanticscholar.org/CorpusID:247011916.

Quan Zhou and Jakub Marecek. Group-blind optimal transport to group parity and its constrained variants, 2023.