The price for fairness in a regression framework

Thibaut Le Gouic *and Jean-Michel Loubes †

May 26, 2020

Abstract

We consider the problem of achieving fairness in a regression framework. Fairness is here expressed as demographic parity. We provide a control over the loss of the generalization error when fairness constraint is imposed, hence computing the cost for fairness for a regressor. Then, using optimal transport theory, we provide a way to construct a fair regressor which is optimal since it achieves the optimal generalization bound. This regressor is obtained by a post-processing methodology.

Contents

*thibaut.le_gouic@math.cnrs.fr, Massachusetts Institute of Technology, Department of Mathematics and Centrale Marseille, I2M, UMR 7373, CNRS, Aix-Marseille univ., Marseille, 13453, France,

†loubes@math.univ-toulouse.fr, Institut de Mathématiques de Toulouse, University Toulouse 3, Toulouse

1 Introduction

In the framework of Machine Learning, the concept of *fairness* has recently arisen due to the rapid growth of the uses of *automatic process* or algorithms in a growing number of parts of our society. In several areas such as justice, recruitment, this development comes with moral and legal issues. We refer for instance to [18], [12], [6] or [26] and references therein for examples of unfair treatment in automatic decisions. In particular, using automated procedures directly impacting humans requires being able to guarantee a *fair* treatment for all. Yet the presence of bias in the datasets used to train algorithm or in the algorithms themselves induces decision rules that favor majority groups as pointed out in [21], [19] or [7].

More formally, an algorithm is said to suffer from **unfairness** with respect to a variable S, if its outcome (or decision) is (partially) based on S. The variable S typically models a characteristic of a population that *should* not play a decisive role in the decision making process for legal, ethical or practical reasons. Making an "automatic process" *fair* aims to remove partially or totally the influence of S in the process. Note that it is possible for an automatic process to be unfair with respect to a variable S that it does not observe as an input, since this input might be correlated to S. For instance, an algorithm can easily make a good prediction of the gender of a person from data such its income, its level of study and the number of hours spent weekly at work as shown in [7] for instance with emphasis in the hiring policy of several companies as detailed in a recent newspapers article [16]

In this article, we consider the multidimensional regression framework. For a triple (X, Y, S) of random variables, the goal regression is to predict a variable of interest Y given (X, S), where X are characteristics while S denotes the sensitive variable. In this setting, the performance of a regressor $(X, S) \mapsto g(X, S)$ of Y can be quantified by its quadratic risk

$$\mathcal{R}(g) := \mathbf{E}\|Y - g(X, S)\|^2,$$

for $\|.\|$ a given norm. For a class of function $\mathcal{G}$, the best possible regressor $g \in \mathcal{G}$ as a function of (X, S), if it exists, has a risk

$$\mathcal{R}(\mathcal{G}) := \inf_{g \in \mathcal{G}} \mathcal{R}(g). \tag{1}$$

When $\mathcal{G}$ is the set of all measurable functions, the minimum is achieved for the conditional expectation of Y given (X, S), denoted $\eta_S(X) := \mathbf{E}(Y|(X, S))$. In this context, it is also called the *Bayes regressor*.

Many definitions of fairness have been considered in the literature, originally for classification problems. Some recent work deals with the regression case as in [13], [20] or [3] for instance. Fairness conditions in this setting aims at reducing or removing the possible relationships between $g(X, S)$ and S and the methodological choices consist in quantifying this notion of statistical relationship or correlation in order to remove it to obtain fair regressors.

Among all criterion, in this work, we consider the *demographic parity* framework where fairness requires that the regressor g is independent of the variable S in order to guarantee that the decision will not be impacted by the sensitive variable. In the classification case, when Y takes only the two possible values 0 and 1 demography parity of g implies that

$$P(g(X, S) = 1|S = 1) = P(g(X, S) = 1|S = 0).$$

When a certain amount of correlation between $g(X,S)$ and S is admitted, this constraint can be relaxed using the notion of Disparate Impact. The *disparate impact* of a classifier is a measures of the risk of discrimination when using the decision rules encoded in g by computing the ratio

$$DI(g) := \frac{P(g(X,S)=1|S=1)}{P(g(X,S)=1|S=0)}. \tag{2}$$

We refer to [19] and references therein for a description of this measure which has become popular due to its interpretation in terms of legislation in the USA for instance since 1971 to detect disparate treatment [1].

Yet other criterion have been used to model dependence relationships between S and $g(X,S)$ using quantitative indicators that can be used to calibrate the algorithms.
Several methods can be used, which can be divided into three main categories : pre-processing the observations to remove the influence of the sensitive variable, constraining the optimization process to obtain fair classifiers or post-processing the outcome of the algorithm. We refer to [31], [2] or [30] and references therein for a review on these methods.

In many paths to fairness, a key issue is to compare the distributions of an object (observations, regressor or scores of a regressor) for the different values of the sensitive variable S. Hence it amounts to comparing different conditional distributions and to study whether they are close with respect to a well chosen distance, which would convey the information that the sensitive variable plays little role and thus enhancing fairness with respect to this variable. Within this framework, the choice of a proper distance between the distributions is of high importance and Monge Kantorovich type distances have been studied in this context. In fact, many works dealing with optimal transport and fairness have been conducted in this direction.When dealing with the task of removing the influence of S in the training dataset (X,Y) (referred to as *repairing* the data) a solution with optimal transport was proposed in [19], or [24] with some theoretical results on the price for fairness provided in [22] or testing methods in [17]. In the framework of classification, in [25] a special attention is paid on the relation between W_1 distance and the error induced by a fair classifier. In their work, the authors provide a control of an integrated loss for the classification error which involves the 1-Wasserstein and for which the minimum is achieved by considering the 1-Wasserstein barycenter of what is called the model belief and which corresponds to the Bayes score $\eta_S(X) = P(Y=1|X,S)$ in the classification model with two classes. This result enables to post-process the scores of a classifier to gain fairness. Adding Wasserstein type constraint has been done in [33] while counterfactual examples are built using optimal transport methods in [8].

In this work, we tackle the problem of studying the price to pay to achieve fairness in a regression setting. Actually, promoting fairness in machine learning algorithms, as seen previously, consists in reducing the effect of some statistical correlations present in the dataset. In most of the cases it results in a loss of accuracy to predict data that share the same biases in their distribution since it removes some available information. So fairness constraint modifies and often reduces the estimation performance of the estimator when it is evaluated on a dataset sharing the same distribution. So there is a natural trade-off between the loss of theoretical performance and the level of fairness to be reached. An important task is thus to be able to quantify the exact loss of performance when looking for fair regressor.

We study the difference between the optimal risk of a regressor and a fair regressor, for the case of demographic parity. In this setting, we obtain a lower bound on the price to pay in performance for fairness in the regression problem. We also provide a method based on the Wasserstein barycenter on the conditional distribution of the Bayes regressor and prove it achieve our lower bound. Hence, we provide a fair regressor which achieves the best generalization error using optimal transport theory. Finally using the multimarginal formulation of the barycenter problem, we provide a feasible method to compute this fair regressor. This estimator is part of the family of post-processing estimator since it relies on an optimal combination of the approximated Bayes regressor.

The rest of the paper falls into the following parts. In Section 2, we provide a lower bound on the excess risk of the quadratic regression due to fairness, in a broad sense. In Section 3, we build an optimal fair regressor for demographic parity constraint. Section 4 is then devoted to the computation of an estimator of our optimal regressor, in the case the dimension of the outcome is one. Proofs are gathered in Section 5.

2 A general lower bound for fair regression

2.1 Notation and Definitions

Consider the following notation. For some $d \in \mathbf{N}$, let (X, Y, S) be a Borel random variable taking its values in $\mathcal{X} \times \mathbf{R}^d \times \mathcal{S}$, where $\mathcal{X}$ and $\mathcal{S}$ are topological spaces. We focus on the the usual setting where S is a random variable taking a finite number k of values. The distribution of (X, Y, S) is denoted by $\mathbf{P}$.

Recall that we consider the regression problem of the target variable Y on g, functions of the variables of X and S that belong to a given class $\mathcal{G}$ of measurable functions . Fairness constraints are modeled through constraints imposed on the space $\mathcal{G}$ of measurable functions that will be chosen later on (depending on the fairness framework we adopt).

When the regression is performed on $\mathcal{G} = \mathcal{F}$ the set of all measurable functions from $\mathcal{X} \times \mathcal{S}$ to $\mathbf{R}^d$, the optimal risk (a.k.a. Bayesian risk), is defined as

$$\mathcal{R}^\star := \mathcal{R}(\mathcal{F}) = \min_{g \in \mathcal{F}} \mathbf{E}\|Y - g(X, S)\|^2.$$

The minimum is achieved for $g(X, S) = \eta_S(X) = E(Y|X, S)$, known as the Bayes estimator. We are interested in the excess risk of a class of esitmators $\mathcal{G}$ defined as

$$\mathcal{E}(\mathcal{G}) := \mathcal{R}(\mathcal{G}) - \mathcal{R}^\star. \tag{3}$$

We introduce a class of functions $\mathcal{G}$ for which we will be able to determine the excess risk of an optimal *fair* regressor. Such classes of functions $g(X, S)$ should be well defined by their conditional distribution as explained in the following definition.

Definition 1 (CCD class)**.** *A class $\mathcal{G}$ of measurable functions from $\mathcal{X} \times \mathcal{S}$ to $\mathbf{R}^d$ is* characterized by conditional distributions *for the model (X, Y, S) (abbreviated* CCD(X, Y, S) *or simply CCD) if for every two measurable functions $g, g' : \mathcal{X} \times \mathcal{S} \to \mathbf{R}^d$, such that $g(X, S)$ and $g'(X, S)$ have same conditional distribution given S, $g \in \mathcal{G}$ implies $g' \in \mathcal{G}$.*

Remark 2 (Noisy CCD class)**.** *In some situation, it useful to be able to minimize over a class $\mathcal{G}$ of* random *measurable functions from $\mathcal{X} \times \mathcal{S}$ to $\mathbf{R}^d$. This can be formalized by allowing $\mathcal{G}$ to be a class of functions from $\mathcal{E} \times \mathcal{X} \times \mathcal{S}$ to $\mathbf{R}^d$ where $\mathcal{E}$ is equipped with a random variable ε independent to (X, S). The definition of CCD class in that context — we will referred to it as* noisy CCD class *— is essentially the same: $\mathcal{G}$ is a CCD class if $g \in \mathcal{G}$ and g and g' has same conditional distribution given S, then $g' \in \mathcal{G}$.*

Example 3 (CCD classes)**.** *The definition of CCD classes encompasses several notions that are used in the fairness literature.*

1. *Demographic parity*

   A canonical example of such CCD classes is the set $\mathcal{G}_{\perp\!\!\!\perp}$ of all measurable functions g such that $g(X, S)$ is independent of S. This is a class CCD as a consequence of the fact that $g(X, S)$ is independent of S if and only if the conditional distribution of $g(X, S)$ given S is equal to the distribution of $g(X, S)$. This class of function $\mathcal{G}$ corresponds to the case of demographic parity regression.

2. *Fixed Disparate impact.*

   In the context of binary classification, (i.e. Y takes in values in $\{0, 1\}$), for any $\alpha > 0$, the class $\mathcal{G}_\alpha$ of all functions g such that $\mathrm{DI}(g) \leq \alpha$ — as defined in (2) — is also a CCD class. Indeed, the condition $\mathrm{DI}(g) = \alpha$ depends only on the conditional distribution of $g(X, S)$ given S, and thus if $g \in \mathcal{G}_\alpha$ and the conditional distributions of $g'(X, S)$ and $g(X, S)$ given S are identical, then $\mathrm{DI}(g') = \mathrm{DI}(g) = \alpha$.

3. *Bounded conditional variance*

   In the framework of regression, demographic parity requires that the predictions of the algorithm are the same for each class. This implies that $E((g(X, S)|S)$ should not depend on S, so $E((g(X, S)|S) = E(g(X, S))$. A weaker condition is given by the fact that the random variable $E(g(X, S)|S)$ has small variability, which warrants a similar behavior for all values of S. So define for any $\alpha > 0$, $\mathcal{G}_\alpha$ the class of functions g such that $\mathrm{Var}(E(g(X, S)|S)) \leq \alpha$. This class is a CCD class. Looking at small variability of the regressor is inspired by the idea pointed out in [38] in which authors consider the variability of the conditional loss of the regressor as a measure of fairness.

From now on, a regressor $g(X, S)$, such that $g \in \mathcal{G}$ will be referred to as *fair regressor*.

Our results use the notion of optimal transport between probabilities. We recall here briefly some main notions. We will use the 2-Wasserstein distance (a.k.a quadratic Monge-Kantorovitch distance) between two probability measures defined as follows. For P and Q two probability measures on $\mathbb{R}^d$ with norm $\|.\|$, the squared Wasserstein distance between P and Q is defined as

$$W_2^2(P, Q) := \min_{\pi \in \Pi(P,Q)} \int \|x - y\|^2 d\pi(x, y)$$

where $\Pi(P, Q)$ the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals P and Q. The function W_2 defines a metric on the set $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures on $\mathbb{R}^d$ with finite moment of order 2. The metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is called the Wasserstein space. More

specifically, our method relies on the notion of Wasserstein barycenter, that was first introduced in [4]. Several algorithm to compute its solution has been developed [14, 15, 35] and statistical guaranties of its solutions have been studied in [5, 11, 29, 10, 28, 39, 27]. We refer to the comprehensive books [36, 34] for more details on optinal transport.

2.2 A lower bound to quantifying the price for fairness

First recall that the optimal regressor when no constraint is specified is given by $\eta_S(X) = E(Y|X,S)$. For different fixed values of S, $\eta_S(X)$ is a random variable with respect to X with value in $\mathbb{R}^d$. Hence let denote by μ_S its conditional distribution with respect to X which defines a random measure on $\mathbb{R}^d$. Fairness conditions tend to impose that the distribution of the regressors conditionally to S are close. Hence it seems natural that the variability of the μ_S plays en essential role to quantify the difficulty to impose fairness.

More precisely, the following result relates the excess risk with a minimization problem in the Wasserstein space.

Theorem 4. *Consider a regressor $g(X,S)$ in a given functional class $\mathcal{G}$ which models a fairness condition. Seen as random variable, let $\nu_S(g)$ be its conditional distribution given S. Then, recall that*

$$\mathcal{E}(\mathcal{G}) = \min_{g\in\mathcal{G}} \mathbf{E}\|Y - g(X,S)\|^2 - \mathbf{E}\|Y - \eta_S(X)\|^2.$$

Then the following bound holds

$$\mathcal{E}(\mathcal{G}) \geq \inf_{g\in\mathcal{G}} \mathbf{E}W_2^2(\mu_S, \nu_S(g)). \tag{4}$$

Moreover, if $\mathcal{G}$ is CCD and μ_s has density w.r.t. Lebesgue measure for almost every s, or if $\mathcal{G}$ is a noisy CCD, then (4) becomes an equality

$$\mathcal{E}(\mathcal{G}) = \inf_{g\in\mathcal{G}} \mathbf{E}W_2^2(\mu_S, \nu_S(g)). \tag{5}$$

Theorem 4 controls the price for imposing the condition that $g \in \mathcal{G}$ on the performance of the regressor. Actually, the lower bound measures the loss in the quality of forecasting induced by the restriction to the fair $\mathcal{G}$ case. Hence for CCD classes, we quantify the price to pay for fairness as the quantity defined by $\inf_{g\in\mathcal{G}} \mathbf{E}W_2^2(\mu_S, \nu_S(g))$. This quantity is the mean distance between the actual distribution $\nu_S(g)$ of the chosen fair regressor g and the conditional distribution of the optimal unfair regressor μ_S, given the chosen subgroup defined by the sensitive variable S.
In the following section we will make this bound more explicit in the case of demographic parity fairness.

We point out that optimal fair regressors can thus be found by minimizing (5), i.e the quadratic Wasserstein distance between the distribution of the optimal regressors for each sub-group depending on the variable S — namely $\nu_S(X) \sim \mu_S$ at given S — and the distribution of the candidate function g.
This leads to the conclusion that optimal fair methods in regression can be achieved by post-processing methods : first compute the optimal algorithms for each sub-group indexed by the

values of S, and then averaging the different scores to obtain a mean estimator combining the sub-groups information. Theorem 4 can thus be given the following interpretation. The loss of any method to achieve fairness depends on how far the distributions of the optimal algorithms for each sub-groups are — *far* being defined with respect to Wasserstein distance. Moreover, this is done optimally via post-processing.

This remark enables to build a new fair classification for regression type problems in Section 3.

Remark 5. *For the quadratic regression (i.e. quadratic loss for the risk), the excess risk and the average squared distance between a fair regressor $g(X,S)$ and the Bayes regressor $\eta_S(X)$ are equal (see equation (11)). This is the only part where we use the particular choice of quadratic regression. If we were interested in the quantity $\mathbf{E}\|g(X,S)-\eta_S(X)\|^2$ to quantify the performance of a fair regressor, then similar results hold for any kind of regression.*

Remark 6. *Note that Theorem 4 provides also a bound for the excess risk in the classification case. Let $Y \in \{0,1\}$ and $\mathcal{G}$ be a subset of functions with also binary values $\{0,1\}$. Note that*

$$\mathbf{E}\|Y-g(X,S)\|^2 = \mathbb{P}(Y \neq g(X,S))$$

is the classification risk, while $\eta_S(X) = E[Y|S,X] = P(Y=1|X,S)$ and. Hence (5) provides a control over

$$\mathbb{P}(Y \neq g(X,S)) - \mathbf{E}\|Y-\eta_S(X)\|^2$$

which is not the excess risk

$$\mathbb{P}(Y \neq g(X,S)) - \inf_g \mathbb{P}(Y \neq g(X,S))$$

since $\eta_S(X) \notin \{0,1\}$.

Yet this bound can still be used when trying to understand the prediction of scores used in classification before a threshold is applied.

3 Demographic parity regression

3.1 Fairness and Wasserstein barycenter

In this part, we develop the case of demographic parity regression, corresponding to the class $\mathcal{G}_{\perp\!\!\!\perp}$ introduced in Example 3. This fairness condition corresponds to removing all influence of S on the regressor $g(X,S)$. In other words, we aim at finding the best regressor $g(X,S)$ of Y such that $g(X,S)$ is independent of S.

Assume without loss of generality that S takes its values in $\mathcal{S} := \{1,\dots,k\}$ and set $\pi_s = \mathbf{P}(S=s)$ for $s=1,\dots,k$. In this case, Theorem 4 can be rewritten

$$\mathcal{E}(\mathcal{G}) = \inf_{g\in\mathcal{G}} \sum_{s=1}^{k} \pi_s W_2^2(\mu_s, \nu(g)), \tag{6}$$

where $\nu(g) := \nu_S(g)$ does not depend on S.
Finding the minimum in (4) amounts to solve the minimization problem

$$\nu \mapsto \mathbf{E}_S W_2^2(\nu, \mu_S) \tag{7}$$

where μ_S is a random variable following the empirical distribution $P_{\mu_S} := \sum_{j=1}^{k} \pi_j \delta_{\mu_j}$ on the set of distributions.

This problem corresponds to finding the so-called Wasserstein barycenter of the distribution P_{μ_S} generating the mixture of conditional distributions μ_s with weights π_s. A minimizer of (7) is called a *barycenter* of the empirical distribution P_{μ_S} and will be denoted by ν_B when exists. Finding a barycenter as pointed out in Section 4 of [4], is equivalent to the following multi-marginal problem. Note $\Gamma(\mu_1, \dots, \mu_k)$ the set of probability measures on $(\mathbb{R}^d)^k$ having marginals $\mu_1, \dots, \mu_k$. The multi-marginal problem consists in the following minimization problem

$$\min \left\{ \int \|y_s - b(\mathbf{y})\|^2 \mathrm{d}\gamma(y_1, \dots, y_k); \gamma \in \Gamma(\mu_1, \dots, \mu_k) \right\}, \tag{8}$$

where $\mathbf{y} = (y_1, \dots, y_k) \in (\mathbb{R}^d)^k$ and $b(\mathbf{R}^d)^k \to \mathbf{R}^d$ is the barycenter map defined by

$$b(\mathbf{y}) := \sum_{s=1}^{k} \pi_s y_s.$$

Proposition 4.2 in [4] shows that there always exists a solution $\gamma_\star$ of (8). Moreover, the barycenter ν_B of P_{μ_S} is the pushforward measure defined by $b_\# \gamma_\star := \gamma_\star \circ b^{-1}$. Hence the following holds

$$\inf_{\nu} \mathbf{E}_S W_2^2(\nu, \mu_S) = \sum_{s=1}^{k} \pi_s \int \|b(y) - y_s\|^2 \mathrm{d}\gamma_\star(y_1, \dots, y_k). \tag{9}$$

Thus, letting $\bar{\mathbf{Y}} = (\bar{Y}_1, \dots, \bar{Y}_k)$ be a random variable with distribution $\gamma_\star$, the distribution of $b(\bar{\mathbf{Y}})$ is ν_B, which is the distribution of the optimal fair regressor as shown in (6) according to Theorem 4. We use this multi-marginal formulation to exhibit an optimal fair regressor.

3.2 Optimal fair regressor

We now turn to the formulation of an optimal fair regressor $g_\star(X, S)$ that we have proven that is must have distribution ν_B.

Assume that for some $s \in \{1, \dots, k\}$, the distribution μ_s has density w.r.t. Lebesgue measure on $\mathbb{R}^d$. In this case, Theorem 4.1 in [4] ensures that there exist measurable maps $T_s : \mathbf{R}^d \to \mathbf{R}^d$ and $T^t : \mathbf{R}^d \to \mathbf{R}^d$ which are optimal transport maps, pushing μ_s towards ν_B and ν_B towards μ_t respectively. Hence if a random variable $\bar{Y}_s$ has distribution μ_s, then $T_s(\bar{Y}_s)$ has distribution ν_B and for $s \in \{1, \dots, k\}$ such that μ_s has density w.r.t. Lebesgue measure,

$$\gamma_\star = (T^1 \circ T_s, \dots, T^k \circ T_s)_\# \nu_s.$$

Since $b_\# \gamma_\star = \nu_B$, in particular, setting $T_s^t := T^t \circ T_s$, we have that, for any $s \in \mathcal{S}$ such that μ_s has density w.r.t. Lebesgue measure,

$$b((T_s^t(\eta_s(X)))_{t \in \mathcal{S}}) \sim \nu_B.$$

Hence we have proven the following theorem which provides an expression of an optimal fair regressor for the demographic parity criterion.

Theorem 7. *If for each $s \in \{1, \dots, k\}$, the conditional distributions μ_s has density w.r.t. the Lebesgue measure on $\mathbb{R}^d$, a minimal risk regressor respecting demographic parity (i.e. achieving minimum of (3) for $\mathcal{G} = \mathcal{G}_{\perp\!\!\!\perp}$), is given by*

$$g_\star(X, S) := b((T_S^t(\eta_S(X)))_{t \in \mathcal{S}}) = \sum_{t=1}^{k} \pi_t T_S^t(\eta_S(X)).$$

This theorem provides an insight on how to build a fair procedure which preserves as much as possible its prediction efficiency, in the sense that the corresponding prediction risk is optimal among all possible fair regressors. Recall that X represents the characteristics of the individuals that will be used to predict Y and $S \in \{1, \dots, k\}$ represents the community or category to which they belong. The conditional expectation $\eta_S(X)$ is the optimal regressor but the variable Y is unfairly predicted by $\eta_S(X)$ in the sense that the distribution $\eta_S(X)$ depends on the protected variable S — thus violating the property of demographic parity. Then the solution of the multi-marginal problem $\gamma_\star$ represents a pairing of all possible values of the regressor $\eta_S(X)$ among all categories indexed by the values of S. In other words, each individual $X = x$ in the category $S = s$ is paired up through the value of $\eta_s(x)$ with individuals $X = x'$ from every other category $t \neq s$ using the transportation maps T_s^t on $\eta_t(x')$. This allows to compare the optimal unfair regressor for *similar individuals* — in the sense that they should provide similar output in a fair context — but belonging to distinct categories.

Finally, the fair regressor $g_\star$ transforms a prediction $\eta_S(X)$ into the average of all predictions paired up with the others $\eta_t(X)$ belonging to the other categories for different values of t.

Note that this fair regressor is the theoretical fair regressor since it depends on unknown quantities η_S and π_S.

Remark 8. *A similar result holds when the distributions μ_s are not all absolutely continuous w.r.t. the Lebesgue measure. In the case, the optimal regressor $g_\star$ belongs to a noisy CCD class (see remark 2) and is constructed as follows. Denote by $\gamma_\star^{-s}$ the conditional distribution of $\gamma_\star$ given its s coordinate. That is, for a random variable $\bar{\mathbf{Y}} = (\bar{Y}_1, \dots, \bar{Y}_k) \sim \gamma_\star$, $\gamma_\star^{-s}(\bar{Y}_s)$ is the conditional distribution of $(\bar{Y}_1, \dots, \bar{Y}_k)$ given $\bar{Y}_s$. Then, when $S = s$, draw $\bar{\mathbf{Y}}^s$ from distribution $\gamma_\star^{-s}(\eta_s(X))$ and set*

$$g_\star(X, S) = b(\bar{\mathbf{Y}}^S).$$

4 Estimation of optimal fair regressor

In practice, the exact distribution of (X, Y, S) that is required to compute the fair regressor $g_\star$ is unknown. Building up on the previous section, we provide some guidelines to construct an estimator of $g_\star$ based on empirical observations.

4.1 General estimator of optimal fair regressor

Theorem 7 provides a way to construct an optimal fair regressor $g_\star$ for the regression case. Yet, it depends on unknown quantities, namely, the probabilities of occurrence of this variable,

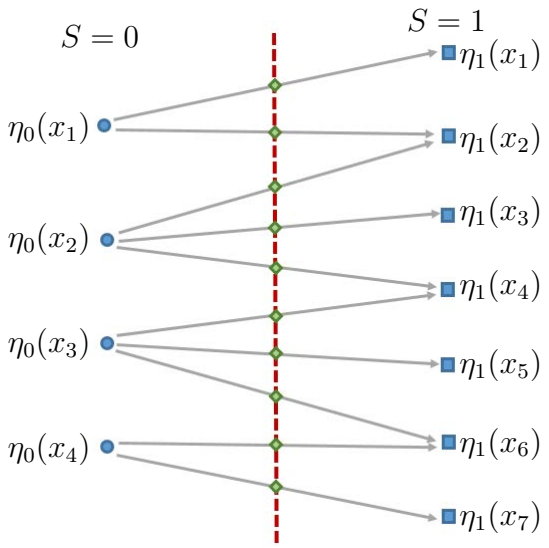

Figure 1: Example of construction of a fair regressor as in Remark 8. For $S = 0$ and $i = 1, \dots, 4$, the values $\eta_0(x_i)$ of the unfair regressor represented by blue dots on the left are matched with values $\eta_1(x_j)$ of the unfair regressor for $S = 1$. For each $i = 1, \dots, 4$, the optimal fair regressor for $X = x_i$ and $S = 0$ is drawn among the green dots on the arrows emanating from $\eta_0(x_i)$.

π_s for all s, transport maps T_s and T_t and $\eta_s(X)$ for each $s \in \{1, \dots, k\}$. In the context of i.i.d. sampling, we are provided with a sample (Y_i, X_i, S_i) for $i = 1, \dots, n$, drawn from the same distribution, together with a sample $(X_i, S_i)_{i=n+1,\dots,n+m}$ of which we want to estimate our regressor $g_\star$. We can then use the following scheme to obtain an estimator of $g_\star$.

Algorithm to construct a fair regressor

1. Estimate $x \mapsto \eta_s(x)$ for each $s \in \mathcal{S}$ by an estimator of the conditional expectation $\hat{\eta}_s(x)$ using the observations (Y_i, X_i, S_i) for $i = 1, \dots, n$.

2. Approximate the distribution of each $\hat{\eta}_s(X)$ for $s = 1, \dots, k$ by the empirical measure using the whole dataset $(X_i, S_i)_{i=1,\dots,n+m}$:

$$\hat{\mu}_s := \frac{1}{N_s} \sum_{i=1}^{N_s} \delta_{\hat{\eta}_s(x_i^s)}$$

   where x_i^s are the values of the observations X_i such that $S_i = s$, while $N_s = \#\{i \in \{1, \dots, n+m\} | S_i = s\}$ denotes their total number.

3. Set $\hat{\pi}_s = N_s/(n+m)$ and

$$\hat{b} := (y_1, \dots, y_k) \mapsto \sum_{s=1}^{k} \hat{\pi}_s y_s$$

   and solve the multimarginal problem (8) for distributions $\hat{\mu}_s$, which can be written as

$$\min \left\{ \int \|y_s - \hat{b}(\mathbf{y})\|^2 \mathrm{d}\gamma(y_1, \dots, y_k); \gamma \in \Gamma(\hat{\mu}_1, \dots, \hat{\mu}_k) \right\}, \tag{10}$$

We denote by $\hat{\gamma}$ its solution.

4. As in Remark 8, denote by $\hat{\gamma}^{-s}$ the conditional distribution of $\hat{\gamma}$ given its s coordinate. For each $s \in \mathcal{S}$ and $i \in \{1, \dots, n+m\}$ such that $S_i = s$, draw $\hat{\mathbf{Y}}^s$ from $\hat{\gamma}^{-s}(\hat{\eta}_s(X_i))$. Then set
$$\hat{g}(X_i, S_i) := \hat{b}(\hat{\mathbf{Y}}^s).$$

This procedure provides an approximation of the fair regressor $g_\star$. Note that $(x, s) \mapsto \hat{g}(x, s)$ is well defined only when (x, s) lies in the predetermined set $(X_i, S_i)_{1 \le i \le n+m}$ for which we want a fair regressor. Hence the whole procedure consists in post-processing the output of the algorithm and aggregate them.

When the sample $\{(X_i, S_i)_{1 \le i \le n+m}\}$ is i.i.d., the distribution $\nu(\hat{g})$ of $\hat{g}(X, S)$ when (X, S) are distributed according to the empirical measure
$$\frac{1}{n+m} \sum_{i=1}^{n+m} \delta_{(X_i, S_i)}$$
is the Wassertein barycenter of the measure
$$\hat{P}_{\hat{\mu}_S} = \sum_{s=1}^{n} \hat{\pi}_s \delta_{\hat{\mu}_s}.$$
If the estimator $\hat{\eta}_s$ is consistent then $\hat{P}_{\hat{\mu}_S}$ converges to
$$P_{\mu_S} := \sum_{s=1}^{n} \pi_s \delta_{\mu_s},$$
as $n \to \infty$. Therefore, using consistency of the barycenter (see [28, Theorem 2]) we obtain the following theorem.

Theorem 9. *Suppose that $\hat{\eta}_s$ is a bounded L_2-consistent estimator of η_s for each $s \in \mathcal{S}$ and that $\hat{\eta}_s$ are uniformly Lipschitz for each $s \in \mathcal{S}$. Assume also that the distribution $\nu(g_\star)$ of $g_\star(X, S)$ is unique. Then, as $n \to \infty$, the empirical distribution of $g(X_i, S_i)$ converges to $\nu(g_\star)$ in W_2, a.s..*

This approximation requires estimators of the Bayes predictor that can be found for instance in [23].

There exist several algorithms to compute a multimarginal solution $\hat{\gamma}$ of (10). Such algorithms perform well only in low dimension d. We refer for instance to [14, 15, 35].

Hence, we have achieved the construction of a *fair regressor* by considering the empirical barycenter of the approximated empirical distributions of the Bayes regressor. So the procedure is thus a post-processing method that enables to reweigh the contribution of each individual by comparing its score to the corresponding scores of individual with different sensitive attribute.

In the following section, we consider the special case of uni-dimensional regression.

4.2 Special case of regression on the real line ($d = 1$)

In this section, we focus on the real valued regression case where computations are easier. Consider the model where we observe $Y_i \in \mathbb{R}$ responses of $i = 1, \dots, n$ individuals with characteristics $X_i \in \mathbb{R}^d$ and discrete sensible variables S_i with values in $\mathcal{S} = \{1, \dots, k\}$. Define the Wasserstein space $\mathcal{P}_2(\mathbf{R})$ as the space of probabilities over $\mathbb{R}$ with finite second moments. In that case, the Wasserstein distance between two measures μ and ν can be expressed with their quantile (or inverse cumulative distribution) functions F_μ^{-1} and F_ν^{-1}, by the formula [32, Remark 2.30]

$$W_2^2(\mu, \nu) = \int_0^1 \|F_\mu^{-1}(x) - F_\nu^{-1}(x)\|^2 \mathrm{d}x.$$

In particular, $\mu \mapsto F_\mu^{-1} \in L^2([0;1])$ is an isometry from $\mathcal{P}_2(\mathbf{R})$ to the convex set of non-decreasing l.s.c. function of $L^2([0;1])$.

The coupling $\gamma_\star$ minimizing the multimarginal problem (9) can thus be expressed as follows. Recall that for all given $s \in \{1, \dots, k\}$, $\eta_s(X) = E(Y|X, S = s)$ is a real random variable with distribution μ_s. Denote by $F_{\mu_s}^{-1}$ the quantile function of μ_s for all $s = 1, \dots, k$. Then, whenever $U \sim \mathcal{U}(0, 1)$ is the uniform distribution on $[0, 1]$,

$$(F_{\mu_1}^{-1}(U), \cdots, F_{\mu_k}^{-1}(U)) \sim \gamma_\star,$$

has distribution $\gamma_\star$. Moreover, if μ_s is absolutely continuous w.r.t. Lebesgue measure, then

$$F_{\mu_s}(F_{\mu_s}^{-1}(U)) = U \sim \mathcal{U}(0, 1).$$

And thus, in this case, according to Theorem 7, the optimal regressor that respects demographic parity is given by

$$g_\star(X, S) = \sum_{s=1}^{k} \pi_s F_{\mu_s}^{-1}(F_{\mu_S}(X)).$$

For a measure $\hat{P}_{\hat{\mu}_S}$ approximating P_{μ_S}, supported on absolutely continuous measures $\hat{\mu}_s$ (for instance, convolutions of empirical measures), then we can define $\hat{g}$ similarly to $g_\star$ as

$$\hat{g}(x, s) = \sum_{j=1}^{k} \hat{\pi}_j F_{\hat{\mu}_j}^{-1}(F_{\hat{\mu}_s}(x)).$$

5 Proofs

Proof of Theorem 4. Denote $\eta_S(X) := \mathbf{E}[Y|(X, S)]$. By definition of the conditional expectation, $Y - \eta_S(X)$ is orthogonal to the space of $\sigma(X, S)$-measurable functions in $L^2(\mathbf{P})$. Therefore,

$$\mathbf{E}\|Y - \eta_S(X)\|^2 + \mathbf{E}\|g(X, S) - \eta_S(X)\|^2 = \mathbf{E}\|Y - g(X, S)\|^2. \tag{11}$$

So,

$$\inf_{g \in \mathcal{G}} [\mathbf{E}\|Y - g(X, S)\|^2] - \mathbf{E}\|Y - \eta_S(X)]\|^2 = \inf_{g \in \mathcal{G}} \mathbf{E}[\mathbf{E}(\|g(X, S) - \eta_S(X)\|^2|S)]$$

For almost every value s of S, the conditional distribution of $(g(X,S),\eta_S(X))$ given S is a coupling between μ_s and ν_s, hence by definition of the Wasserstein distance

$$\mathbf{E}[\|g(X,S)-\eta_S(X)\|^2|S]\geq W_2^2(\nu_S(g),\mu_S).$$

Integrating with respect to S proves (4).

If μ_s has density w.r.t. the Lebesgue measure, then there exists a map $T_s:\mathbf{R}^d\to\mathbf{R}^d$ such that the distribution of $T_s(\eta_s(X))$ is $\nu_s(g)$ and

$$W_2^2(\nu_s(g),\mu_s)=\mathbf{E}\|T_s(\eta_s(X))-\eta_s(X)\|^2.$$

If $\mathcal{G}$ is CCD, then for each g,

$$g':=(x,s)\mapsto T_s(\eta_s(x))\in\mathcal{G}.$$

Since moreover, we have almost surely

$$\mathbf{E}[\|g'(X,S)-\eta_S(X)\|^2|S]=W_2^2(\nu_S(g),\mu_S),$$

the function g' is an admissible minimum and thus (4) is an equality.

The case of noisy CCD is handled similarly. □

Proof of Theorem 9. Using consistency of the Wasserstein barycenter [28, Theorem 2], we just need to prove that $\hat{P}_{\hat{\mu}_S}\to P_{\mu_S}$ in Wasserstein distance. This is a consequence of $\hat{\pi}_s\to\pi_s$ — which holds due to the law of large number, and $\hat{\mu}_s\to\mu_s$ in Wasserstein distance. Therefore, it just remains to prove that, as $n+m\to\infty$,

$$W_2^2(\hat{\mu}_s,\mu_s)\to 0,\quad\forall s\in\mathcal{S}.$$

Denote by $T^{n,m}:(x,s)\mapsto(T_x^{n,m}(x,s),T_s^{n,m}(x,s))\in\mathbf{R}\times\mathbf{R}^d$ the optimal transport from of the distribution $\mathbf{P}_{(X,S)}$ of (X,S) to the empirical measure

$$\mathbf{P}_{(X,S)}^{n+m}:=\frac{1}{n+m}\sum_{i=1}^{n+m}\delta_{(X_i,S_i)}.$$

Denote by L an upper bound of the Lipschitz constants of $\hat{\eta}_s$. We can compute

$$\begin{aligned}
W_2^2(\hat{\mu}_s,\mu_s)&\leq\mathbf{E}\|\eta_S(X)-\hat{\eta}_{T_s(X,S)}(T_x(X,S))\|^2\\
&\leq 2\mathbf{E}\|\eta_S(X)-\hat{\eta}_S(X)\|^2+2\mathbf{E}\mathbf{1}_{T_s(X,S)=S}\|\hat{\eta}_S(X)-\hat{\eta}_S(T_x(X,S))\|^2\\
&\quad+2\mathbf{E}\mathbf{1}_{T_s(X,S)\neq S}\|\hat{\eta}_S(X)-\hat{\eta}_{T_s(X,S)}(T_x(X,S))\|^2\\
&\leq 2\mathbf{E}\|\eta_S(X)-\hat{\eta}_S(X)\|^2+2L^2\mathbf{E}\|X-T_x(X,S)\|^2+8M\mathbf{P}(T_s(X,S)\neq S),
\end{aligned}$$

where M is the bound of $\hat{\eta}_s$. Now, $\mathbf{E}\|\eta_S(X)-\hat{\eta}_S(X)\|^2\to 0$ as $\hat{\eta}_s$ is L_2-consistent. Then,

$$\mathbf{E}\|X-T_x(X,S)\|^2=W_2^2(\mathbf{P}_{(X,S)},\mathbf{P}_{(X,S)}^{n+m})\to 0$$

since the empirical measure is consistent in W_2 (this is a consequence of Varadarajan's Theorem on compact spaces, see for instance [9, 37] for rates of convergence). Finally, since $|s-s'|\geq 1$ for $s\neq s'\in\mathcal{S}$, then,

$$\mathbf{P}(T_s(X,S)\neq S)\leq W_2^2(\mathbf{P}_{(X,S)},\mathbf{P}_{(X,S)}^{n+m})\to 0$$

a.s., which concludes the proof. □

Acknowledgments.
Thibaut Le Gouic was supported by ONR grant N00014-17-1-2147 and NSF IIS-1838071.
Jean-Michel Loubes thank the AI interdisciplinary institute ANITI, grant agreement ANR-19-PI3A-0004 under the French investing for the future PIA3 program.

References

[1] Code of federal regulations: Title 29 - labor, 07 2017.

[2] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. *arXiv e-prints*, page arXiv:1507.05259, 7 2015.

[3] A. Agarwal, M. Dudík, and Z. S. Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *ICML*, 2019.

[4] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[5] A. Ahidar-Coutrix, T. Le Gouic, and Q. Paris. Convergence rates for empirical barycenters in metric spaces: Curvature, convexity and extendable geodesics. pages 1–46.

[6] S. Barocas and A. D. Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[7] P. Besse, E. del Barrio, P. Gordaliza, J.-M. Loubes, and L. Risser. A survey of bias in machine learning through the prism of statistical parity for the adult data set. *arXiv preprint arXiv:2003.14263*, 2020.

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

[9] E. Boissard and T. Le Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. In *Annales de l'IHP Probabilités et Statistiques*, volume 50, pages 539–563.

[10] E. Boissard, T. Le Gouic, and J.-M. Loubes. Distribution's template estimate with wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.

[11] S. Chewi, T. Maunu, P. Rigollet, and A. J. Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters.

[12] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[13] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair Regression via Plug-in Estimator and Recalibration With Statistical Guarantees. working paper or preprint, Mar. 2020.

[14] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. pages 2292–2300.

[15] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693.

[16] J. Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, 10 2018.

[17] E. Del Barrio, P. Gordaliza, and J.-M. Loubes. A central limit theorem for lp transportation cost on the real line with application to fairness assessment in machine learning. *Information and Inference: A Journal of the IMA*, 8(4):817–849, 2019.

[18] J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), 2018.

[19] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. *arXiv e-prints*, page arXiv:1412.3756, 12 2014.

[20] J. Fitzsimons, A. Al Ali, M. Osborne, and S. Roberts. A general framework for fair regression. *Entropy*, 21(8):741, 2019.

[21] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. *arXiv e-prints*, 02 2018.

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

[23] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media.

[24] P. Hacker and E. Wiedemann. A continuous framework for fairness. *ArXiv*, abs/1712.07924, 2017.

[25] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein fair classification. *stat*, 1050:28, 2019.

[26] P. T. Kim. Data-driven discrimination at work. *Wm. & Mary L. Rev.*, 58:857, 2016.

[27] A. Kroshnin, V. Spokoiny, and A. Suvorikova. Statistical inference for Bures-Wasserstein barycenters.

[28] T. Le Gouic and J.-M. Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168(3-4):901–917, 2017.

[29] T. Le Gouic, Q. Paris, P. Rigollet, and A. J. Stromme. Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space.

[30] A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118, New York, NY, USA, 2 2018. PMLR.

[31] L. Oneto and S. Chiappa. *Fairness in Machine Learning*, pages 155–196. Springer International Publishing, Cham, 2020.

[32] G. Peyré and M. Cuturi. *Computational Optimal Transport*. Mar. 2018. arXiv: 1803.00567.

[33] L. Risser, Q. Vincenot, N. Couellan, and J.-M. Loubes. Using wasserstein-2 regularization to ensure fair decisions with neural-network classifiers. *arXiv preprint arXiv:1908.05783*, 2019.

[34] F. Santambrogio. *Optimal transport for applied mathematicians*. Birkhauser, 2015.

[35] J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. 34(4):66.

[36] C. Villani. *Optimal transport: Old and new*. Springer, 2008.

[37] J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. 25:2620–2648.

[38] R. C. Williamson and A. Krishna Menon. Fairness risk measures. *arXiv e-prints*, page arXiv:1901.08665, 1 2019.

[39] P. C. Álvarez Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744 – 762, 2016.