A Sequentially Fair Mechanism for Multiple Sensitive Attributes

François Hu[1], **Philipp Ratz**[2], **Arthur Charpentier**[2]

[1]Université de Montréal, Montréal, Québec, Canada
[2]Université du Québec à Montréal, Montréal, Québec, Canada
francois.hu@umontreal.ca, ratz.philipp@courrier.uqam.ca, charpentier.arthur@uqam.ca

Abstract

In the standard use case of Algorithmic Fairness, the goal is to eliminate the relationship between a sensitive variable and a corresponding score. Throughout recent years, the scientific community has developed a host of definitions and tools to solve this task, which work well in many practical applications. However, the applicability and effectivity of these tools and definitions becomes less straightfoward in the case of multiple sensitive attributes. To tackle this issue, we propose a sequential framework, which allows to progressively achieve fairness across a set of sensitive features. We accomplish this by leveraging multi-marginal Wasserstein barycenters, which extends the standard notion of Strong Demographic Parity to the case with multiple sensitive characteristics. This method also provides a closed-form solution for the optimal, sequentially fair predictor, permitting a clear interpretation of intersensitive feature correlations. Our approach seamlessly extends to approximate fairness, enveloping a framework accommodating the trade-off between risk and unfairness. This extension permits a targeted prioritization of fairness improvements for a specific attribute within a set of sensitive attributes, allowing for a case specific adaptation. A data-driven estimation procedure for the derived solution is developed, and comprehensive numerical experiments are conducted on both synthetic and real datasets. Our empirical findings decisively underscore the practical efficacy of our post-processing approach in fostering fair decision-making.

Introduction

Recent media coverage has put the spotlight anew on a question that preoccupies the field of (algorithmic) fairness, namely what constitutes fairness and how to achieve it. We center our focus on group fairness, particularly on the concept of Demographic Parity fairness (Calders, Kamiran, and Pechenizkiy 2009), with the objective of achieving independence between attributes and predictions while bypassing the use of labels. A point raised frequently is that only considering a single (binary or discrete) attribute is insufficient to determine whether a system is truly fair (Kong 2022). Indeed studies emphasize that when a model considers only a single attribute, it overlooks subgroups defined by intersecting attributes. This notion is commonly referred as *fairness gerrymandering* (Kearns et al. 2018) and it occurs "when we only look for unfairness over a small number of pre-defined groups" that are arbitrarily selected. Scholars in the field of algorithmic fairness have devised a range of methods to counter unfairness in predictions, for both regression (Chzhen et al. 2020a,b; Gouic, Loubes, and Rigollet 2020) and classification (Hardt, Price, and Srebro 2016; Agarwal, Dudik, and Wu 2019; Chiappa et al. 2020; Denis et al. 2021). However, these approaches consider fairness with respect to a single feature, making them susceptible to the criticism from above. As an example, (Buolamwini and Gebru 2018) underscored this intersectional bias in Machine Learning (ML). They discovered that major face recognition algorithms exhibited preferences for recognizing men and lighter skin tones, resulting in reduced accuracy for women with darker skin tones, thereby exposing both gender and racial discrimination.

A naive solution to this problem would be to create a single discrete feature that groups a set of sensitive variables. However, this approach has several drawbacks: *i)* this methodology assigns similar weights to all attributes, hindering approximate fairness which allows the user to adjust the fairness constraint; *ii)* further, this method complicates tracking procedure effects, as disentangling them from a combined variable is complex and also lacks prioritization across attributes; *(iii)* in applications, some sensitive feature might need more attention, like bias due to gender over age (Macnicol 2006; Charpentier, Hu, and Ratz 2023). This prioritization bridges the gap between a status-quo of (unfair) predictions and a goal of fair predictions *w.r.t* a set of sensitive features. Fairness considerations often leads to reduced predictive performance (Menon and Williamson 2018; Chen, Johansson, and Sontag 2018), making adoption challenging. Presetting a level of unfairness could potentially facilitate its acceptance and adoption.

Main contributions

This highlights the need for a more holistic approach to consider fairness. In this article, we propose a methodology that is adaptable for optimal fair predictions involving Multiple Sensitive Attributes (MSA). More specifically:

- We address the learning problem under the Demographic Parity constraint, involving MSA, by constructing multi-marginal 2-Wasserstein barycenters. Our method offers a closed-form solution, allowing us to develop an empirical data-driven approach that enhances fairness for any off-the-shelf estimators.
- We rewrite the optimal fair solution into a sequential form by using the associativity of Wasserstein barycenters in

univariate settings. This formulation seamlessly extends to approximate fairness achieving fairness-risk trade-off.

- Our approach is demonstrated through numerical experiments on diverse datasets (both synthetic and real). It demonstrates high effectiveness in reducing unfairness while enabling clear interpretation of inter-sensitive feature correlations.

We begin by introducing some necessary notation before formally presenting the problem. After deriving the main results we conduct extensive numerical experiments and illustrate the use of our methodology on a both synthetic and real-world datasets. Note that all proofs are relegated to the supplementary materials to ease the lecture of the article.

Related work

Much of our work extends earlier findings from the literature of algorithmic fairness using optimal transport, a mathematical framework for measuring distributional differences. The aim is to transform biased scores into equitable ones while minimizing their impact to uphold predictive performance. There are several methods that can broadly be classified into pre-, in- and post-processing methods. Our approach developed here falls into the latter category, as post-processing is computationally advantageous and allows a clear interpretation of the outputs. In regression, methods like (Chzhen et al. 2020b) and (Gouic, Loubes, and Rigollet 2020) minimize Wasserstein distance to mitigate bias. Similarly, in classification, (Chiappa et al. 2020) and (Gaucher, Schreuder, and Chzhen 2023) use optimal transport for fairness. Instead of considering multiple fair attributes, (Hu, Ratz, and Charpentier 2023) consider multiple fair prediction tasks through joint optimization. However, despite optimal transport's widespread use in algorithmic fairness, there is limited research on MSA and intersectional fairness. This article aims to fill this research gap.

Notation

Consider a function f and a random tuple $(\boldsymbol{X}, \boldsymbol{A}) \in \mathcal{X} \times \mathcal{A} \subset \mathbb{R}^d \times \mathbb{N}^r$, with positive integers d and r. We denote $\mathcal{V}$ the space of probability measures on $\mathcal{Y} \subset \mathbb{R}$. Let $\nu_f \in \mathcal{V}$ and $\nu_{f|\boldsymbol{a}} \in \mathcal{V}$ be respectively the probability measure of $f(\boldsymbol{X}, \boldsymbol{A})$ and $f(\boldsymbol{X}, \boldsymbol{A})|\boldsymbol{A} = \boldsymbol{a}$. $F_{f|\boldsymbol{a}}(u) := \mathbb{P}\left(f(\boldsymbol{X}, \boldsymbol{A}) \leq u | \boldsymbol{A} = \boldsymbol{a}\right)$ corresponds to the cumulative distribution function (CDF) of $\nu_{f|\boldsymbol{a}}$ and $Q_{f|\boldsymbol{a}}(v) := \inf\{u \in \mathbb{R} : F_{f|\boldsymbol{a}}(u) \geq v\}$ its associated quantile function.

Background on Wasserstein barycenters

This section introduces the concepts of Wasserstein barycenter from one-dimensional optimal transport theory. Further details can be found in (Santambrogio 2015; Villani 2021).

Wasserstein distance

We consider two probability measures, ν_1 and ν_2. The *Wasserstein distance* quantifies the minimum "cost" of transforming one distribution into the other. Specifically, the squared Wasserstein distance between ν_1 and ν_2 is defined as

$$\mathcal{W}_2^2(\nu_1, \nu_2) = \inf_{\pi \in \Pi(\nu_1, \nu_2)} \mathbb{E}_{(Z_1, Z_2) \sim \pi} \left(Z_2 - Z_1\right)^2 \quad ,$$

where $\Pi(\nu_1, \nu_2)$ is the set of distributions on $\mathcal{Y} \times \mathcal{Y}$ having ν_1 and ν_2 as marginals. A coupling that achieves this infimum is called optimal coupling between ν_1 and ν_2.

Wasserstein barycenter

Throughout this article, we will frequently make use of *Wasserstein Barycenters* (Agueh and Carlier 2011). The Wasserstein barycenter finds a representative distribution that lies between multiple given distributions in the Wasserstein space. It is defined for a family of K measures $(\nu_1, \ldots, \nu_K)$ in $\mathcal{V}$ and some positive weights $(w_1, \ldots, w_K) \in \mathbb{R}_+^K$. The Wasserstein barycenter (or in short $\mathcal{W}_2$-barycenter), denoted as $\mathrm{Bar}\left\{(w_k, \nu_k)_{k=1}^K\right\}$ is the minimizer

$$\mathrm{Bar}(w_k, \nu_k)_{k=1}^K = \operatorname*{argmin}_{\nu} \sum_{k=1}^K w_k \cdot \mathcal{W}_2^2\left(\nu_k, \nu\right) \quad . \qquad (1)$$

The work in (Agueh and Carlier 2011) shows that in our configuration the barycenter exists and a sufficient condition of uniqueness is that one of the measures ν_i admits a density w.r.t. the Lebesgue measure.

Our study focuses on barycentric associativity within the unidimensional space $\mathcal{Y} \subset \mathbb{R}$. This principle asserts that the global barycenter coincides with the barycenter of barycenters. This associativity is clear in Euclidean spaces (Ungar 2010): the barycenter of points x_1, x_2, and x_3 in $\mathbb{R}^d$ with weights w_1, w_2, and w_3 aligns with the barycenter of $x_{1,2}$ and x_3 with weights $w_1 + w_2$ and w_3, where $x_{1,2}$ is the barycenter of x_1 and x_2. In the following proposition, we explore the relevance of this barycentric associativity, particularly for the 2-Wasserstein barycenter within the unidimensional space. It is important to note that the Wasserstein barycenter loses its associativity in a multi-dimensional framework.

Proposition 1 (Associativity of the $\mathcal{W}_2$-barycenter) *Consider a collection of positive integers $K_1, K_2, \ldots, K_r$, where their sum is denoted as $K = K_1 + K_2 + \cdots + K_r$. Let the sets be defined as follows:*

$$B_1 = (w_{1,k}, \nu_{1,k})_{k=1}^{K_1}, \ldots, B_r = (w_{r,k}, \nu_{r,k})_{k=1}^{K_r} \quad ,$$

where $\{w_{i,k}\}_{i,k}$ are positive and non-zero weights summing to 1 and $\{\nu_{i,k}\}_{i,k}$ represent univariate measures. In this context, the overall Wasserstein barycenter $\mathrm{Bar}\left\{B_1 \cup \cdots \cup B_r\right\}$ can be expressed as the barycenter $\mathrm{Bar}(\tilde{w}_i, \mathrm{Bar}(\tilde{B}_i))_{i=1,\ldots,r}$, where

$$\tilde{w}_i := \sum_{k'=1}^{K_i} w_{i,k'} \text{ and } \tilde{B}_i := \left(\frac{w_{i,k}}{\tilde{w}_i}\ ,\ \nu_{i,k}\right)_{k=1,\ldots,K_i} \quad .$$

In a nutshell, this formulation captures the relationship between the overall Wasserstein barycenter and the individual barycenters of its constituent sets, incorporating the relevant weights seamlessly. Given measures (ν_1, ν_2, ν_3) with positive weights (w_1, w_2, w_3) summing to 1, mirroring the aforementioned barycentric concept in Euclidean space, we derive the following relations:

$$\begin{aligned} &\mathrm{Bar}\left\{(w_1, \nu_1), (w_2, \nu_2), (w_3, \nu_3)\right\} \\ &\quad = \mathrm{Bar}\left\{(w_1, \nu_1), (w_2 + w_3, \nu_{2,3})\right\} \\ &\quad\quad = \mathrm{Bar}\left\{(w_1 + w_2, \nu_{1,2}), (w_3, \nu_3)\right\} \quad . \end{aligned}$$

Here, for $i, j \in \{1, 2, 3\}$ (where $i \neq j$), the measure $\nu_{i,j} = \text{Bar}\{(\tilde{w}_i, \nu_i), (\tilde{w}_j, \nu_j)\}$ is defined, with $\tilde{w}_k = w_k/(w_i + w_j)$.

Problem formulation

Let $(\boldsymbol{X}, \boldsymbol{A}, Y)$ be a random tuple with distribution $\mathbb{P}$. Here, $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^d$ denotes the d non-sensitive features, $Y \in \mathcal{Y} \subset \mathbb{R}$ represents the target task, and $\boldsymbol{A} = (A_1, \ldots, A_r) \in \mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_r$ the r discrete sensitive attributes, where $\mathcal{A}_i = \{1, \ldots, K_i\}$ with $K_i \in \mathbb{N}$. For example, in a binary case with $r = 2$, we could have $A_1 =$ gender and $A_2 =$ age. For convenience, we use the notation $A_{i:i+k} := (A_i, A_{i+1}, \cdots, A_{i+k})$ to denote the sequence of $k+1$ sensitive features ranging from i to $i + k$ (so $\boldsymbol{A} = A_{1:r}$). Further, we denote $\mathcal{F}$ the set of predictors of the form $f : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ where we assume that each measure $\nu_{f|\boldsymbol{a}}$ admits a density w.r.t. Lebesgue measure. More precisely, we require the following assumption to hold:

Assumption 2 *Given $f \in \mathcal{F}$, measures $\{\nu_{f|\boldsymbol{a}}\}_{\boldsymbol{a}\in\mathcal{A}}$ are non-atomic with finite second moments.*

Risk measure

Within the statistical learning community, a central pursuit revolves around the minimization of a designated risk measure across the set $\mathcal{F}$ encompassing all predictors. In particular, a Bayes regressor minimizing the squared risk,

$$\textbf{(Risk Measure)} \quad \mathcal{R}(f) := \mathbb{E}\left(Y - f(\boldsymbol{X}, \boldsymbol{A})\right)^2 \quad ,$$

over the set $\mathcal{F}$ is represented by $f^*(\boldsymbol{X}, \boldsymbol{A}) := \mathbb{E}[Y|\boldsymbol{X}, \boldsymbol{A}]$.

In our case, our objective is to characterize the optimal fair predictor, which minimizes the squared risk under a given fairness constraint. To do so, we introduce formally the Demographic Parity notion of fairness.

Unfairness measure

Demographic Parity (DP) will be used to determine the fairness of a predictor. Fairness considerations under DP offers the advantage of being applicable to both classification and regression tasks. In our study, the unfairness measure of the predictor on the feature A_i is given by

$$\mathcal{U}_i(f) = \max_{a_i \in \mathcal{A}} \int_{u\in[0,1]} \left| \, Q_f(u) - Q_{f|a_i}(u) \, \right| du \quad , \qquad (2)$$

while for the multiple sensitive features $A_i, \ldots, A_{i+k}$, their collective unfairness is simply assessed through:

$$\mathcal{U}_{\{i,\ldots,i+k\}}(f) = \mathcal{U}_{i:i+k}(f) = \mathcal{U}_i(f) + \cdots + \mathcal{U}_{i+k}(f) \quad . \qquad (3)$$

Hence, we naturally broaden the DP-fairness definition to encompass both exact and approximate fairness within the context of MSA framework.

Definition 3 (Fairness under Demographic Parity) *The overall unfairness of a predictor $f \in \mathcal{F}$ w.r.t. the feature $\boldsymbol{A} = A_{1:r}$, can be quantified by the unfairness measure,*

$$\textbf{\textit{(Unfairness measure)}} \quad \mathcal{U}(f) = \mathcal{U}_{1:r}(f) \quad .$$

Then f is called exactly fair if and only if

$$\mathcal{U}(f) = 0 \quad .$$

Given $\boldsymbol{\varepsilon} = \varepsilon_{1:r} := (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_r)$ where each $\varepsilon_i \in [0, 1]$, f is called approximately fair under DP with $\boldsymbol{\varepsilon}$ relative improvement ($\boldsymbol{\varepsilon}$-RI) if and only if each individual unfairness satisfies

$$\mathcal{U}_i(f) \leq \varepsilon_i \times \mathcal{U}_i(f^*) \quad .$$

In other words, in the context of approximate fairness, we are interested in the relative (fairness) improvement of a fair predictor with respect to Bayes' rule f^* (see (Chzhen and Schreuder 2022) for further details).

Preliminary results

Recall the Wasserstein barycenter defined in Eq. (1). We consider measures $(\nu_{f|\boldsymbol{a}})_{\boldsymbol{a}\in\mathcal{A}}$ with corresponding weights $(p_{\boldsymbol{a}})_{\boldsymbol{a}\in\mathcal{A}}$, where $p_{\boldsymbol{a}} := \mathbb{P}(\boldsymbol{A} = \boldsymbol{a})$. It is assumed that $\min_{\boldsymbol{a}}\{p_{\boldsymbol{a}}\} \geq 0$. Encompassing these measures is the Wasserstein barycenter, denoted $\mu_{\mathcal{A}} : \mathcal{V} \to \mathcal{V}$ and it is defined as

$$\mu_{\mathcal{A}}(\nu_f) := \text{Bar}(p_{\boldsymbol{a}}, \nu_{f|\boldsymbol{a}})_{\boldsymbol{a}\in\mathcal{A}} = \operatorname*{argmin}_{\nu} \sum_{\boldsymbol{a}\in\mathcal{A}} p_{\boldsymbol{a}} \cdot \mathcal{W}_2^2\left(\nu_{f|\boldsymbol{a}}, \nu\right) \quad .$$

Single Sensitive Attribute (SSA) case We consider a single sensitive attribute A, belonging to the set $\mathcal{A} = \{1, \ldots, K\}$, with $p_a := \mathbb{P}(A = a)$. Let $f_B \in \mathcal{F}$, and let its measure be the Wasserstein barycenter $\nu_{f_B} = \mu_{\mathcal{A}}(\nu_{f^*})$, where we recall that $f^*(\boldsymbol{X}, A) = \mathbb{E}[Y|\boldsymbol{X}, A]$ is the Bayes rule which minimizes the squared risk. Prior research conducted by (Chzhen et al. 2020b; Gouic, Loubes, and Rigollet 2020) demonstrates that,

$$f_B = \operatorname*{argmin}_{f\in\mathcal{F}} \{\mathcal{R}(f) : \mathcal{U}(f) = 0\} \quad .$$

Therefore, f_B represents the optimal fair predictor in terms of minimizing unfairness-risk. Additionally, previous studies have offered a closed-form solution: for all $(\boldsymbol{x}, a) \in \mathcal{X} \times \mathcal{A}$,

$$f_B(\boldsymbol{x}, a) = \left(\sum_{a'\in\mathcal{A}} p_{a'} Q_{f^*|a'}\right) \circ F_{f^*|a}\left(f^*(\boldsymbol{x}, a)\right) \quad .$$

Note that this solution can be easily adapted not only to classification tasks (Gaucher, Schreuder, and Chzhen 2023), but also to multi-task learning involving classification and regression (Hu, Ratz, and Charpentier 2023). In this article, we extend this formulation to the MSA case $\boldsymbol{A} = (A_1, \ldots, A_r) \in \mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_r$. For any $(\boldsymbol{x}, \boldsymbol{a}) \in \mathcal{X} \times \mathcal{A}$ we denote,

$$f_{B_i}(\boldsymbol{x}, \boldsymbol{a}) = \left(\sum_{a_i'\in\mathcal{A}_i} p_{a_i'} Q_{f^*|a_i'}\right) \circ F_{f^*|a_i}\left(f^*(\boldsymbol{x}, \boldsymbol{a})\right) \quad , \quad (4)$$

as the optimal fair predictor, ensuring fairness only across $\mathcal{A}_i$ ($\mathcal{A}_i$-fair for short). By abuse of notation, we denote $p_{a_i'} := \mathbb{P}(A_i = a_i')$, $F_{f^*|a_i'}(u) := \mathbb{P}\left(f(\boldsymbol{X}, \boldsymbol{A}) \leq u | A_i = a_i'\right)$ and $Q_{f^*|a_i'}$ its associated quantile function.

Optimal fair prediction with MSA

We extend the fair characterization into a sequential framework to accommodate MSA. Building upon previous research in the SSA case (Chzhen et al. 2020b; Gouic, Loubes, and Rigollet 2020), we demonstrate that fairness in the MSA problem can also be framed as the optimal transport problem involving the 2-Wasserstein distance. The relationship between these concepts is established in the following proposition.

Proposition 4 (Fair characterization: global approach) *We assume that Assumption 2 holds, and we let*

$$f_B = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \{\mathcal{R}(f) : \mathcal{U}(f) = 0\} \ .$$

Subsequently, its measure satisfies $\nu_{f_B} = \mu_{\mathcal{A}}(\nu_{f^*})$. Furthermore, this equation yields a closed-form solution for the optimal fair predictor

$$f_B(\boldsymbol{x}, \boldsymbol{a}) = \left(\sum_{\boldsymbol{a}' \in \mathcal{A}} p_{\boldsymbol{a}'} Q_{f^*|\boldsymbol{a}'} \right) \circ F_{f^*|\boldsymbol{a}} \left(f^*(\boldsymbol{x}, \boldsymbol{a}) \right) \ . \quad (5)$$

Considering the aforementioned proposition and Prop. 1, which confirms that the barycenter of barycenters aligns with the overall barycenter under suitable updated weights, a straightforward corollary arises. This corollary asserts that regardless of the selected "debiasing path" in the sequential fairness mechanism, the end result consistently leads to the same optimal fair solution.

Proposition 5 (Sequentially fair mechanism) *Under the assumption that Assumption 2 holds, the term $\mu_{\mathcal{A}}(\nu_{f^*})$ defined in Prop. 4 can be equivalently expressed as follows:*

$$\mu_{\mathcal{A}}(\nu_{f^*}) = \mu_{\mathcal{A}_1} \circ \mu_{\mathcal{A}_2} \circ \cdots \circ \mu_{\mathcal{A}_r}(\nu_{f^*}) \ .$$

More generally, under any permutation, i.e. bijection function of the form $\sigma : \mathcal{S} \to \mathcal{S}$ where $\mathcal{S} := \{1, 2, \ldots, r\}$, the above expression can be rewritten as

$$\mu_{\mathcal{A}}(\nu_{f^*}) = \mu_{\mathcal{A}_{\sigma(1)}} \circ \mu_{\mathcal{A}_{\sigma(2)}} \circ \cdots \circ \mu_{\mathcal{A}_{\sigma(r)}}(\nu_{f^*}) \ .$$

Notably the expressions proposed in Prop. 5 allows us to establish a link between Eq. (4) and Eq. (5) of the form:

$$\begin{aligned} f_B(\boldsymbol{X}, \boldsymbol{A}) &= \left(f_{B_1} \circ f_{B_2} \circ \ldots \circ f_{B_r}\right)(\boldsymbol{X}, \boldsymbol{A}) \\ &= \left(f_{B_{\sigma(1)}} \circ f_{B_{\sigma(2)}} \circ \ldots \circ f_{B_{\sigma(r)}}\right)(\boldsymbol{X}, \boldsymbol{A}) \ . \end{aligned}$$

Note that the $\circ$ notation is used in a relaxed manner with regard to predictors, aimed at streamlining the presentation and alleviating the complexities of notation. In particular, by abuse of notation we establish the definition of $f_{B_i} \circ f_{B_j}$ as follows (with f^* serving as the default function):

$$\begin{aligned} &\left(f_{B_i} \circ f_{B_j}\right)(\boldsymbol{x}, \boldsymbol{a}) \\ &\quad = \left(\sum_{a_i' \in \mathcal{A}_i} p_{a_i'} Q_{f_{B_j}|a_i'} \right) \circ F_{f_{B_j}|a_i} \left(f_{B_j}(\boldsymbol{x}, \boldsymbol{a}) \right) \ . \end{aligned}$$

Introducing a sequential approach is pivotal for enhancing clarity. Indeed, this methodology helps in comprehending intricate concepts like approximate fairness with ε-RI defined in Definition 3, which involves improving fairness relatively and approximately with $\boldsymbol{\varepsilon}$ a preset level of relative fairness improvement. This perspective enables us to break down how various components of the sequential fairness mechanism interact to achieve fairness goals and allows for the interpretation of the intrinsic effects of adjusting fairness. Therefore, adopting a sequential perspective is a key step in attaining a deeper understanding of fairness and bias.

Extension to approximate fairness

Recall the previously introduced concept of $\boldsymbol{\varepsilon}$-RI fairness (or $\boldsymbol{\varepsilon}$-fairness for brevity), where $\boldsymbol{\varepsilon} = \varepsilon_{1:r} = (\varepsilon_1, \cdots, \varepsilon_r)$ and each $\varepsilon \in [0, 1]$. In the context of the SSA framework, a methodology introduced by (Chzhen and Schreuder 2022) employs geodesic parameterization. Specifically, considering $A_1 \in \mathcal{A}_1$ *w.l.o.g.*, the predictor of the form:

$$f_{B_1}^{\varepsilon_1}(\boldsymbol{X}, \boldsymbol{A}) = (1 - \varepsilon_1) \cdot f_{B_1}(\boldsymbol{X}, \boldsymbol{A}) + \varepsilon_1 \cdot f^*(\boldsymbol{X}, \boldsymbol{A}) \ ,$$

achieves the optimal risk-fairness trade-off. Notably, we have

$$f_{B_1}^{\varepsilon_1} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \{\mathcal{R}(f) : \mathcal{U}_1(f) \leq \varepsilon_1 \cdot \mathcal{U}_1(f^*)\} \ .$$

We denote the corresponding measure as $\mu_{\mathcal{A}_1}^{\varepsilon_1}(\nu_{f^*}) := \nu_{f_{B_1}^{\varepsilon_1}}$. In the following proposition, we extend sequentially this formulation to the context of MSA.

Proposition 6 (Characterization of approximate fairness) *Let Assumption 2 holds and let*

$$f_B^{\boldsymbol{\varepsilon}} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \mathcal{R}(f) : \mathcal{U}(f) \leq \sum_{i=1,\ldots,r} \varepsilon_i \cdot \mathcal{U}_i(f^*) \right\} \ ,$$

then

$$\nu_{f_B^{\boldsymbol{\varepsilon}}} = \mu_{\mathcal{A}}^{\boldsymbol{\varepsilon}}(\nu_{f^*}) := \mu_{\mathcal{A}_1}^{\varepsilon_1} \circ \cdots \circ \mu_{\mathcal{A}_r}^{\varepsilon_r}(\nu_{f^*}) \ .$$

Similarly to Prop. 5, this expression can also be reformulated by permuting indices.

The expression mentioned in Prop. 6 enables us to explicitly formulate an optimal closed-form predictor within the approximate fairness framework: for any permutation $\sigma \in \mathcal{P}(\mathcal{S})$,

$$f_B^{\boldsymbol{\varepsilon}}(\boldsymbol{X}, \boldsymbol{A}) = \left(f_{B_{\sigma(1)}}^{\varepsilon_{\sigma(1)}} \circ f_{B_{\sigma(2)}}^{\varepsilon_{\sigma(2)}} \circ \ldots \circ f_{B_{\sigma(r)}}^{\varepsilon_{\sigma(r)}} \right)(\boldsymbol{X}, \boldsymbol{A}) \ .$$

Data-driven procedure

For practical application on real data, the plug-in estimator of the Bayes rule f^* is denoted as $\hat{f}$—any DP-unconstrained ML model trained on a set *i.i.d.* instances of $(\boldsymbol{X}, \boldsymbol{A}, Y)$. Given $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{a} = (a_1, \ldots, a_r) \in \mathcal{A}$, the empirical counterpart of an optimal $\mathcal{A}_i$-fair predictor f_{B_i} is then defined as:

$$\widehat{f_{B_i}}(\boldsymbol{x}, \boldsymbol{a}) = \left(\sum_{a_i' \in \mathcal{A}_i} \hat{p}_{a_i'} \hat{Q}_{\hat{f}|a_i'} \right) \circ \hat{F}_{\hat{f}|a_i} \left(\hat{f}(\boldsymbol{x}, \boldsymbol{a}) \right) \ , \quad (6)$$

Here, $\hat{p}_{a_i}$, $\hat{F}_{\hat{f}|a_i}$, and $\hat{Q}_{\hat{f}|a_i}$ are empirical counterparts of p_{a_i}, $F_{f^*|a_i}$, and $Q_{f^*|a_i}$. Interestingly, aside from $\hat{f}$, the other quantities can be constructed using an unlabeled dataset. Notably, (Chzhen et al. 2020b) provides some statistical guarantees: if the estimator $\hat{f}$ approximates f^* well, then given mild assumptions on distribution $\mathbb{P}$, the post-processing method $\widehat{f_{B_i}}$ is a good estimator of f_{B_i}. By composition, $\widehat{f_B} = \widehat{f_{B_1}} \circ ... \circ \widehat{f_{B_r}}$ and $\widehat{f_B^{\boldsymbol{\varepsilon}}} = \widehat{f_{B_1}^{\varepsilon_1}} \circ ... \circ \widehat{f_{B_r}^{\varepsilon_r}}$ emerge as good estimators for f_B and $f_B^{\boldsymbol{\varepsilon}}$ respectively, enabling accurate and fair estimation of the instances. Note that the unfairness of f is assessed on a hold-out set (or test set) using $\widehat{\mathcal{U}}(f)$, the empirical counterpart of Eq. (3). Predictive performance uses mean squared error (MSE) for regression and Accuracy/F_1-score for classification on the same test set.

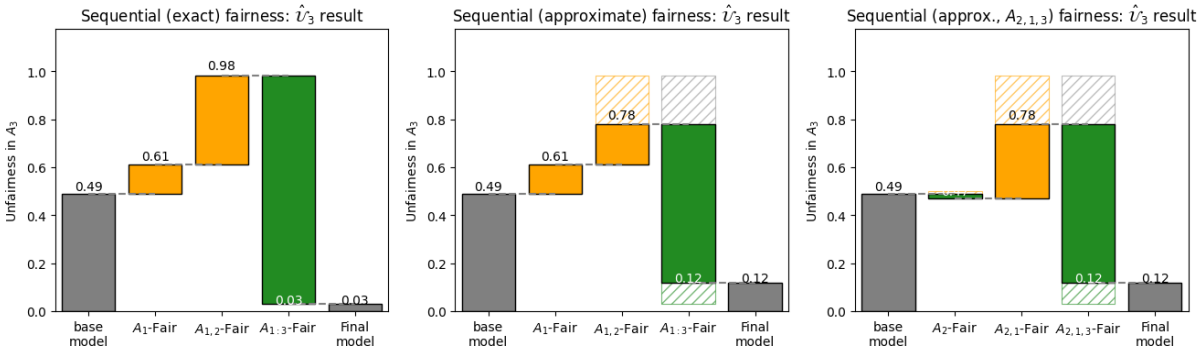

Figure 1: Synthetic data with parameter $\boldsymbol{\tau} = (0, 0.05, 0.1)$. A sequential unfairness evaluation, $\mathcal{U}_3$, of (*left pane*) exact fairness, (*middle pane*) approximate $A_{1,2,3}$-fairness with $\boldsymbol{\varepsilon}$-RI where $\boldsymbol{\varepsilon} = \varepsilon_{1,2,3} = (0.2, 0.5, 0.75)$ and (*right pane*) approximate $A_{2,1,3}$-fairness with $\varepsilon_{2,1,3}$-RI. Hashed color corresponds to exact fairness.

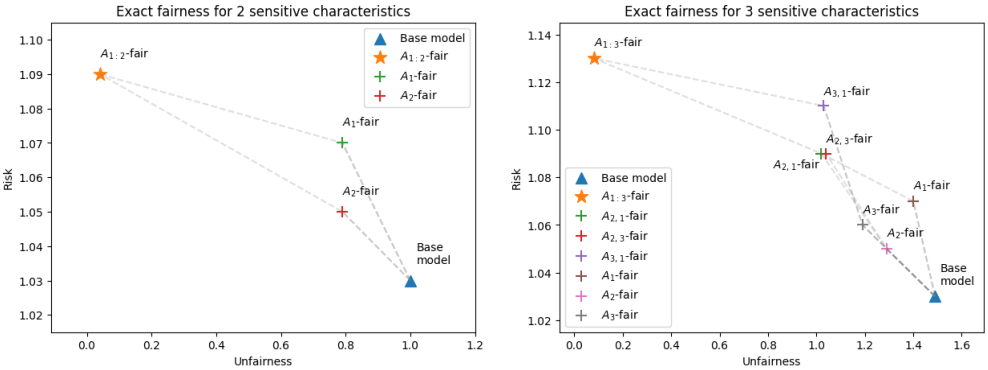

Figure 2: (Risk, Unfairness) phase diagrams that shows the sequential fairness approach for (*left pane*) two sensitive features; (*right pane*) three sensitive features. In this study, Unfairness represents the overall unfairness $\hat{\mathcal{U}} = \hat{\mathcal{U}}_{1:3}$. Bottom-left corner gives the best trade-off.

Numerical experiments

In this section, we conduct a comparative analysis of our methodology against DP-unconstrained methods and state-of-the-art approach given in (Agarwal, Dudik, and Wu 2019). Our findings demonstrate that our exact and approximate fairness approach stands out in terms of interpretability, adaptability and competitiveness.

Case study on synthetic data

Prior to showcasing our method on a real dataset, we opted to assess its performance using synthetic data. This step aims to provide a clearer insight into the effectiveness of the sequential fairness mechanism. Specifically, we consider synthetic data $(\boldsymbol{X}, \boldsymbol{A}, Y)$ with the following characteristics:

- $\boldsymbol{X} \in \mathbb{R}^d$: Comprises d non-sensitive features generated from a centered Gaussian distribution $\boldsymbol{X} \sim \mathcal{N}_d(0, \sigma_X I_d)$, where $\sigma_X > 0$ parameterizes its variance.
- $\boldsymbol{A} = A_{1:r} \in \{-1, 1\}^r$: Represents r binary sensitive features, with $A_i \sim 2 \cdot \mathcal{B}(q_i) - 1$ following a Bernoulli law with parameter $q_i = \mathbb{P}(\tilde{X} > \tau_i)$, where $\tilde{X} \sim \mathcal{N}(0, \sigma_X)$. Here, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_r)$ is a user-set parameter.
- $Y \sim \mathcal{N}(\mathbf{1}^T \boldsymbol{X} + \mathbf{1}^T \boldsymbol{A}, 1)$: Represents the regression task.

Simulation scheme Default parameters are set as follows: $d = 10$, $r = 3$, $\sigma_X = 0.15$, and $\boldsymbol{\tau} = (0, 0.05, 0.1)$. We generated 10,000 synthetic examples and divided the data into three sets (50% training, 25% testing, and 25% unlabeled). As a base model, we opt for a simple linear regression using default parameters from `scikit-learn` in Python. Comparatively, we assess our sequential approach against the uncalibrated base model.

Interpreting intersectional fairness In the context of $r = 3$ sensitive features, Figure 1 showcases $\hat{\mathcal{U}}_3$, an unfairness measure focusing on A_3. The sequential fair approach, detailed in Prop. 5, enhances interpretability by addressing intercorrelations among sensitive features while striving for fair predictions. Specifically, this highlights the A_1 and A_2 correlation with A_3. Pursuing exact fairness, the left side of Figure 1 demonstrates that making A_1 and A_2 fair can inadvertently introduce unfairness in A_3, revealing the fairness gerrymandering issue mentioned in the introduction. Aligning with Prop. 5 and Prop. 6, both Figure 1 and Figure 2 exhibit varied numerical debiasing paths to achieve fairness across the three sensitive features. Importantly, this implies that achieving fairness for A_1 before A_2 is equivalent to the reverse, evident in

Figure 2's (Risk, Unfairness) phase diagrams. Each step illustrated incurs a performance loss but garners fairness gains. The number of such points is influenced by the cardinality of the power set encompassing all sensitive features. Supplementary details on additional fairness experiments with this synthetic data are available in the supplementary materials.

Case study on real data with two sensitive attributes

To illustrate a possible application of our methodology and showcase its use in a real world use-case, we consider data collected in the *folktables* package of (Ding et al. 2021). This package compiles datasets sourced from the US Census, offering a basis for benchmarking ML models. Notably, with named features, we can add interpretation to our methodology. Our study centers on predicting an individual's total income (Income) within sunbelt states (AL, AZ, CA, FL, GA, LA, MS, NM, SC, TX) using standard filters (age $\geq$ 18, at least one hour of weekly work, total income $>$ 100\$). As a secondary task, we consider the problem of predicting whether an individual is covered by public health insurance (Coverage), again with the standard filters (age $<$ 65, income $\leq$ 30,000\$). For both prediction problems, we rely on provided standard features and treat *gender* and *race* as sensitive attributes[1].

Methodology As our approach is applicable to both regression and classification tasks, we consider both problems. For the regression task, we aim to predict the log-income and for the classification task, we aim to predict whether an individual's income exceeds 50,000\$. The sensitive features studied are *gender* and *race*. In total, we have 600,041 observations, with 52.3% Male participants and 10.7% participants from the minority racial class. The data is split into 64% train, 20% test and 16% unlabeled data. As a base model, we opt for a lightGBM (Ke et al. 2017) model with early stopping, where the early stopping iterations are optimized using 5-fold cross validation on the training data.

In an exact fairness framework ($\varepsilon = 0$), we evaluate numerical performance over ten runs with distinct seeds. Since, to our knowledge, there are no open-source fairness methods for MSA predictions, direct benchmarking is not feasible. Instead, for the classification task, we compare our approach with the state-of-the-art *ExponentiatedGradient* method from the *Fairlearn* package (Weerts et al. 2023). First, we ensure fairness on a single variable, enabling a comparison of baseline accuracy, F_1-Score, and unfairness. Next, our methodology is employed to ensure fairness across both variables, facilitating another comparison.

Results From Table 1, in regression, the outcomes align with expectations. Post-processed forecasts exhibit slightly reduced predictive performance, yet offer nearly complete fairness. This achievement comes with minimal added computational cost, quantified by seconds of compute time. In classification, our method proves competitive against the benchmark model. Indeed, both methods provide efficient and fair outcomes for a single feature, though the post-processing of our method is significantly faster. With two sensitive features, the benchmark model underperforms in fairness (as anticipated), while our method excels across both tasks. Though performance slightly dips compared to a single sensitive feature, competitiveness persists—even against the benchmark model, which fails to achieve fairness across both sensitive features.

Table 1: Results for the correction of the biases for the gender and race features. For the classification tasks we first performed the fairness calibration on gender (one sensitive) and then on gender and race (two sensitive).

	Uncorrected	***Our Method***	***Fairlearn***
	Regression		
MSE	0.547 ± 0.003	0.596 ± 0.003	N/A
Unfairness	0.378 ± 0.007	$\mathbf{0.019 \pm 0.005}$	N/A
Time (s)	N/A	8.981 ± 1.319	N/A
	Classification, Income - one sensitive		
Accuracy	0.820 ± 0.001	0.809 ± 0.001	0.808 ± 0.001
F1	0.753 ± 0.001	0.737 ± 0.001	0.73 ± 0.002
Unfairness	0.170 ± 0.001	$\mathbf{0.003 \pm 0.002}$	0.021 ± 0.002
Time (s)	N/A	6.319 ± 0.422	100.8 ± 10.467
	Classification, Income - two sensitive		
Accuracy	0.820 ± 0.001	0.804 ± 0.001	0.808 ± 0.001
F1	0.753 ± 0.001	0.73 ± 0.001	0.73 ± 0.002
Unfairness	0.354 ± 0.006	$\mathbf{0.009 \pm 0.005}$	0.207 ± 0.005
	Classification, Coverage		
Accuracy	0.805 ± 0.0	0.802 ± 0.0	0.805 ± 0.0
F1	0.587 ± 0.0	0.584 ± 0.0	0.574 ± 0.0
Unfairness	0.127 ± 0.0	$\mathbf{0.011 \pm 0.0}$	0.119 ± 0.0
Time (s)	N/A	10.661 ± 0.147	355.2 ± 6.689

Continuing with our application example, our data (see Figure 4) reveals that the median income for Male participants is 17,000\$ higher than for Female participants. Similarly, the sensitive race's members have incomes 10,000\$ lower than the rest of the population. While not the sole contributor, it's plausible that these sub-groups could gain from fair predictions. Figure 4 illustrates the log income distribution, highlighting significant subgroup disparities. For instance, the difference between genders within the sensitive racial group is smaller compared to other participants.

For a decision maker, even when focusing solely on exact fairness, there are multiple routes to achieving fairness across both features. In our example, one could prioritize fairness in race scores first, followed by gender, or vice versa. Each step involves a performance loss but a fairness gain, depicted in the left panel of Figure 3. Although theory and the Figure demonstrate that the final outcome will be identical either way, practical implementation in steps presents a dilemma. The center panel of Figure 3 displays our method's average corrective effects on predictions for two subgroups. For women in the

[1]Code and synthetic data available at: https://github.com/phi-ra/SequentialFairness

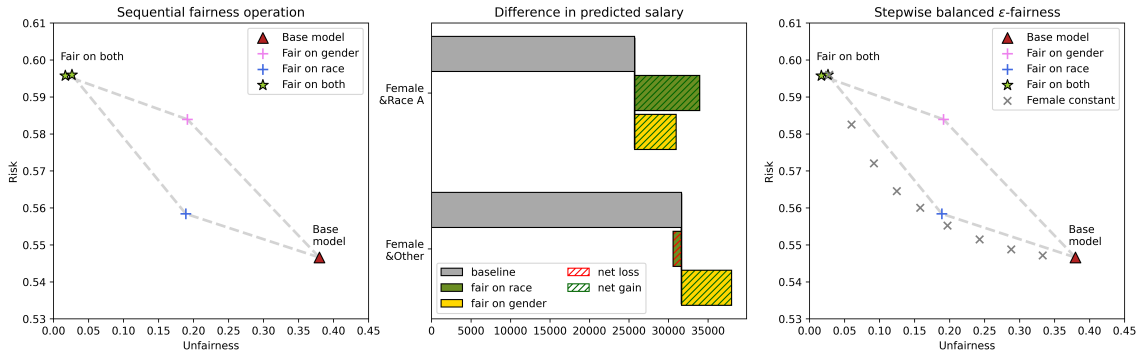

Figure 3: Applications on the *folktables* data set. Left pane, visualisation of the combined unfairness across two sensitive attributes and intermediate solutions rendering predictions fair on only one of them. Center pane, marginal changes to predicted income when rendering fair the predictions with respect to a single variable and the baseline predictions. Right pane, visualization of global metrics when correcting the score first for race, but keeping the average predicted salary of female individuals constant.

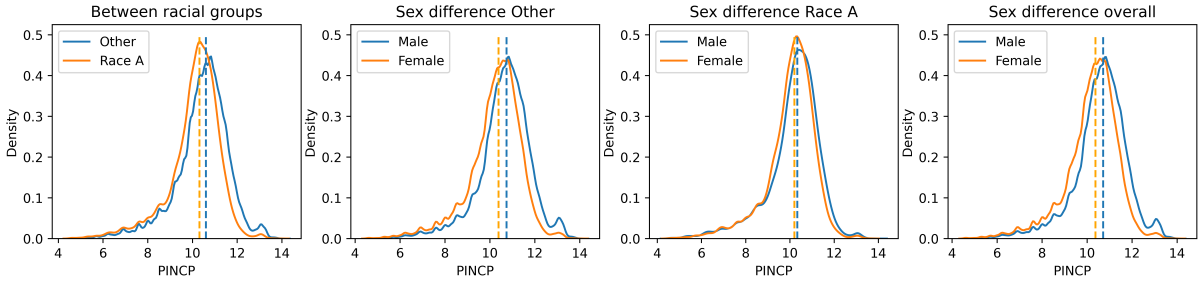

Figure 4: Visualisations of the log-income distribution, with different filters. The two center panes are calculated from the subset of observations defined by the racial groups.

sensitive racial group, correcting for race and gender boosts their predicted income. However, this poses a problem if both effects aren't corrected simultaneously. Women not in the sensitive racial group may oppose race-gender correction order due to a net deficit from race-only corrections. Conversely, sensitive subgroup women gain more from race corrections than gender, favoring the race-gender order. Our methodology offers flexibility for multiple fairness constraints. For instance, instead of consecutively rendering predictions fair on each feature, simultaneous steps are possible. In our example, if fair race predictions are sought without disadvantaging women, adjusting ε for gender to one-sixth of race's size maintains average women's predictions while correcting for race. This adaptable approach is illustrated in the right panel of Figure 3, highlighting our methodology's strength. It ensures exact fairness results remain consistent regardless of the order in which scores become fair, while enabling decision makers to analyze stepwise fairness implementation effects.

Conclusion

We proposed a framework that expands the standard concept of fair scores from SSA to MSA. This extension ensures exact sequential fairness yields the same predictions regardless of correction order for the sensitive features. However, intermediate solutions can yield significant subgroup differences influenced by sensitive features. Our approach quantifies these differences and offers a way to mitigate them using approximate solutions. Our analysis raises intriguing research questions, including how optimal solutions change with fairness metrics beyond the employed DP constraint. The flexibility and user-friendliness of our methodology, supporting a comprehensive fairness approach across multiple debiasing steps, promotes the adoption of fair decision-making practices.

References

Agarwal, A.; Dudik, M.; and Wu, S. 2019. Fair Regression: Quantitative Definitions and Reduction-Based Algorithms. In *International Conference on Machine Learning*.

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

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building classifiers with independency constraints. In *IEEE international conference on Data mining*.

Charpentier, A.; Hu, F.; and Ratz, P. 2023. Mitigating Discrimination in Insurance with Wasserstein Barycenters. arXiv:2306.12912.

Chen, I.; Johansson, F. D.; and Sontag, D. 2018. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31.

Chiappa, S.; Jiang, R.; Stepleton, T.; Pacchiano, A.; Jiang, H.; and Aslanides, J. 2020. A general approach to fairness with optimal transport. In *AAAI*.

Chzhen, E.; Denis, C.; Hebiri, M.; Oneto, L.; and Pontil, M. 2020a. Fair Regression via Plug-in Estimator and RecalibrationWith Statistical Guarantees. In *Advances in Neural Information Processing Systems*.

Chzhen, E.; Denis, C.; Hebiri, M.; Oneto, L.; and Pontil, M. 2020b. Fair Regression with Wasserstein Barycenters. In *Advances in Neural Information Processing Systems*.

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

Denis, C.; Elie, R.; Hebiri, M.; and Hu, F. 2021. Fairness guarantee in multi-class classification. *arXiv preprint arXiv:2109.13642*.

Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring Adult: New Datasets for Fair Machine Learning. *Advances in Neural Information Processing Systems*, 34.

Gaucher, S.; Schreuder, N.; and Chzhen, E. 2023. Fair learning with Wasserstein barycenters for non-decomposable performance measures. In *International Conference on Artificial Intelligence and Statistics*, 2436–2459. PMLR.

Gouic, T. L.; Loubes, J.-M.; and Rigollet, P. 2020. Projection to Fairness in Statistical Learning. *arXiv preprint arXiv:2005.11720*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Neural Information Processing Systems*.

Hu, F.; Ratz, P.; and Charpentier, A. 2023. Fairness in Multi-Task Learning via Wasserstein Barycenters. arXiv:2306.10155.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, 2564–2572. PMLR.

Kong, Y. 2022. Are "intersectionally fair" ai algorithms really fair to women of color? a philosophical analysis. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 485–494.

Macnicol, J. 2006. *Age discrimination: An historical and contemporary analysis*. Cambridge University Press.

Menon, A. K.; and Williamson, R. C. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*.

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

Ungar, A. A. 2010. *Barycentric calculus in Euclidean and hyperbolic geometry: A comparative introduction*. World Scientific.

Villani, C. 2021. *Topics in optimal transportation*, volume 58. American Mathematical Soc.

Weerts, H.; Dudík, M.; Edgar, R.; Jalali, A.; Lutz, R.; and Madaio, M. 2023. Fairlearn: Assessing and Improving Fairness of AI Systems. arXiv:2303.16626.

Supplementary Materials

The supplementary material consists of two parts. One part contains all the proofs of our results, while the other part deals with additional numerical considerations.

A. Proof of main results

Before providing the proof of the $\mathcal{W}_2$-barycentric associativity, we consider the following classical result in optimal transport theory in univariate measures (Santambrogio 2015; Agueh and Carlier 2011).

Lemma 7 *Let $\nu_1, \dots, \nu_K$ be K univariate probability measures admitting densities, for all $w_1, \dots, w_K \geq 0$ summing to 1, the CDF of the optimal measure $\nu_{1:K} := \mathrm{Bar}(w_i, \nu_i)_{i=1}^K$ is given by*

$$F_{\nu_{1:K}}(\cdot) = \left(\sum_{i=1}^K w_i Q_{\nu_i} \right)^{-1} (\cdot) \ .$$

Equivalently, the associated quantile function is given by

$$Q_{\nu_{1:K}}(\cdot) = \left(\sum_{i=1}^K w_i Q_{\nu_i} \right) (\cdot) \ .$$

Proof of Proposition 1 Given $K = K_1 + K_2 + \cdots + K_r$ and the sets:

$$B_1 = (w_{1,k}, \nu_{1,k})_{k=1}^{K_1}, \dots, B_r = (w_{r,k}, \nu_{r,k})_{k=1}^{K_r} \ ,$$

where $\{w_{i,k}\}$ are positive and non-zero weights summing to 1 and $\{\nu_{i,k}\}$ represent univariate measures. In this context, the quantile function of the overall Wasserstein barycenter $\mathrm{Bar}\{B\}$, where $B := B_1 \cup \cdots \cup B_r$ is given by (see above Lemma),

$$\begin{aligned} Q_{\nu_B}(u) &= \sum_{i=1}^r \sum_{k=1}^{K_i} w_{i,k} Q_{\nu_{i,k}}(u) \\ &= \sum_{i=1}^r \underbrace{\left(\sum_{k'=1}^{K_i} w_{i,k'} \right)}_{=:\tilde{w}_i} \underbrace{\left(\sum_{k=1}^{K_i} \frac{w_{i,k}}{\sum_{k'=1}^{K_i} w_{i,k'}} Q_{\nu_{i,k}}(u) \right)}_{=:Q_{\nu_{\tilde{B}_i}}(u)} \\ &= \sum_{i=1}^r \tilde{w}_i Q_{\nu_{\tilde{B}_i}} \end{aligned}$$

where $Q_{\nu_{\tilde{B}_i}}(u) := \sum_{k=1}^{K_i} \frac{w_{i,k}}{\tilde{w}_i} Q_{\nu_{i,k}}(u)$ corresponds to the quantile function of the measure

$$\nu_{\tilde{B}_i} := \mathrm{Bar}\left(\frac{w_{i,k}}{\tilde{w}_i}, \nu_{i,k} \right)_{k=1,\dots,K_i} \ .$$

Thus,

$$\mathrm{Bar}\left(B_i \cup \cdots \cup B_r\right\} = \mathrm{Bar}\left(\tilde{w}_i, \nu_{\tilde{B}_i}\right)_{i=1,\dots,r} \ ,$$

which concludes the proof.

Recall that $f^*(\boldsymbol{X}, \boldsymbol{A}) = \mathbb{E}[Y|\boldsymbol{X}, \boldsymbol{A}]$ and let us define the excess risk as $\mathcal{E}(\mathcal{H}) = \mathcal{R}(f^*) - \inf_{h \in \mathcal{H}} \mathcal{R}(h)$ with $\mathcal{H} \subset \mathcal{F}$ a subclass of regressors. The following lemma is adapted from (Gouic, Loubes, and Rigollet 2020) where we consider multiple sensitive attributes.

Lemma 8 *Let $\mathcal{H} \subset \mathcal{F}$ be a subclass of regressors. If for any $h \in \mathcal{H}$ and $f \in \mathcal{F}$, where,*

$$\nu_{f|\boldsymbol{a}} = \nu_{h|\boldsymbol{a}} \quad \textit{for all } \boldsymbol{a} \in \mathcal{A} \ ,$$

we have $f \in \mathcal{H}$, then the excess-risk of $\mathcal{H}$ is expressed as,

$$\mathcal{E}(\mathcal{H}) = \inf_{h \in \mathcal{H}} \sum_{s \in \mathcal{S}} p_s \mathcal{W}_2^2 \left(\nu_{f^*|\boldsymbol{a}}, \nu_{h|\boldsymbol{a}} \right) \ .$$

Proof of Lemma 8 *From Pythagoras' Theorem, we can derive directly:*

$$\begin{aligned} \mathbb{E}[(Y - h(\boldsymbol{X}, \boldsymbol{A}))^2] &= \mathbb{E}[(Y - f^*(\boldsymbol{X}, \boldsymbol{A}))^2] \\ &\quad + \mathbb{E}[(h(\boldsymbol{X}, \boldsymbol{A}) - f^*(\boldsymbol{X}, \boldsymbol{A}))^2] \ , \end{aligned}$$

and therefore

$$\begin{aligned} &\inf_{h \in \mathcal{H}} \mathbb{E}\big[\mathbb{E}[(h(\boldsymbol{X}, \boldsymbol{A}) - f^*(\boldsymbol{X}, \boldsymbol{A}))^2 | \boldsymbol{A}]\big] \\ &\quad = \inf_{h \in \mathcal{H}} \mathbb{E}[(Y - h(\boldsymbol{X}, \boldsymbol{A}))^2] - \mathbb{E}[(Y - f^*(\boldsymbol{X}, \boldsymbol{A}))^2] \ . \end{aligned}$$

From the definition of the Wasserstein distance, we have

$$\mathbb{E}[(h(\boldsymbol{X}, \boldsymbol{A}) - f^*(\boldsymbol{X}, \boldsymbol{A}))^2 | \boldsymbol{A}] \geq \mathbb{E}[\mathcal{W}_2^2(\nu_{f^*|\boldsymbol{A}}, \nu_{h|\boldsymbol{A}})] \ .$$

Finally, since Assumption 2 holds, there also exists an optimal transport map $T_{\boldsymbol{a}} : \mathbb{R} \to \mathbb{R}$ s.t. $T_{\boldsymbol{a}} \circ f^*(\boldsymbol{X}, \boldsymbol{a}) \sim \nu_{h|\boldsymbol{a}}$ and

$$\mathcal{W}_2^2(\nu_{f^*|\boldsymbol{a}}, \nu_{h|\boldsymbol{a}}) = \mathbb{E}[(T_{\boldsymbol{a}} \circ f^*(\boldsymbol{X}, \boldsymbol{a}) - f^*(\boldsymbol{X}, \boldsymbol{a}))^2] \ .$$

Since $f \in \mathcal{H}$ for any $(\nu_{f|\boldsymbol{a}})_{\boldsymbol{a}} = (\nu_{h|\boldsymbol{a}})_{\boldsymbol{a}}$, it implies that

$$\begin{aligned} \mathbb{E}[\mathcal{W}_2^2(\nu_{f^*|\boldsymbol{A}}, \nu_{h|\boldsymbol{A}})] &= \mathbb{E}[(T_{\boldsymbol{A}}(f^*(\boldsymbol{X}, \boldsymbol{A})) - f^*(\boldsymbol{X}, \boldsymbol{A}))^2] \\ &\geq \inf_{h \in \mathcal{H}} \mathbb{E}[(h(\boldsymbol{X}, \boldsymbol{A}) - f^*(\boldsymbol{X}, \boldsymbol{A}))^2] \ , \end{aligned}$$

which concludes the proof of the lemma.

Proof of Proposition 4 Considering Lemma 8, as we consider $\mathcal{F}_{\mathcal{A}-\text{fair}}$ of the form:

$$\mathcal{F}_{\mathcal{A}-\text{fair}} := \{f \in \mathcal{F} : \mathcal{U}(f) = 0\} \ ,$$

representing the subclass of DP-fair predictors, we can easily infer that if

$$f_B = \operatorname*{argmin}_{f \in \mathcal{F}_{\mathcal{A}-\text{fair}}} \mathcal{R}(f) \ ,$$

then it follows, directly from Lemma 8, that $\nu_{f_B} = \mu_{\mathcal{A}}(\nu_{f^*})$. This, combined with Lemma 7, furnishes us with the explicit closed-form solution:

$$f_B(\boldsymbol{x}, \boldsymbol{a}) = \left(\sum_{\boldsymbol{a}' \in \mathcal{A}} p_{\boldsymbol{a}'} Q_{f^*|\boldsymbol{a}'} \right) \circ F_{f^*|\boldsymbol{a}} \left(f^*(\boldsymbol{x}, \boldsymbol{a}) \right) \ ,$$

concluding the proof of Proposition 4.

Note that Proposition 5 directly follows from Proposition 6.

Proof of Propositions 5 and 6 *w.l.o.g.*, we focus on the scenario that involves two sensitive characteristics, $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$. Leveraging Proposition 1 and the sets

$$\begin{aligned} B_{1,2} &= \left(\mathbb{P}(A_1 = a_1, A_2 = a_2), \nu_{f^*|a_1,a_2} \right)_{(a_1,a_2) \in \mathcal{A}_1 \times \mathcal{A}_2} \\ B_{1|a_2} &= \left(\mathbb{P}(A_1 = a_1 | A_2 = a_2), \nu_{f^*|a_1,a_2} \right)_{a_1 \in \mathcal{A}_1} \text{ for } a_2 \in \mathcal{A}_2 \\ B_{2|a_1} &= \left(\mathbb{P}(A_2 = a_2 | A_1 = a_1), \nu_{f^*|a_2,a_1} \right)_{a_2 \in \mathcal{A}_2} \text{ for } a_1 \in \mathcal{A}_1 \end{aligned}$$

together with the Bayes theorem, we have,

$$\begin{aligned}\mathrm{Bar}(B_{1,2}) &= \mathrm{Bar}(\mathbb{P}(A_2 = a_2), \mathrm{Bar}(B_{1|a_2}))_{a_2 \in \mathcal{A}_2} \\ &= \mathrm{Bar}(\mathbb{P}(A_1 = a_1), \mathrm{Bar}(B_{2|a_1}))_{a_1 \in \mathcal{A}_1} \quad .\end{aligned}$$

Therefore,

$$\mu_{\mathcal{A}}(\nu_{f^*}) = \mu_{\mathcal{A}_2} \circ \mu_{\mathcal{A}_1}(\nu_{f^*}) = \mu_{\mathcal{A}_1} \circ \mu_{\mathcal{A}_2}(\nu_{f^*}) \quad ,$$

which concludes the proof of Proposition 5.

Given the definition of the Unfairness measure in the MSA scenario as a sum of unfairness measures pertaining to individual sensitive attributes (refer to Eq. (3)), Proposition 6 can be directly derived from Proposition 4.1 in (Chzhen and Schreuder 2022), in conjunction with Proposition 5.

B. Additional numerical experiments

This section extends the analysis using synthetic data from the main body. Figure 5 presents $\hat{\mathcal{U}}_{1:r}$, an overall unfairness measure for all sensitive attributes. The proposed sequential fair approach (Prop. 5) enhances interpretability by addressing inter-correlations among sensitive features while striving for fair predictions. Notably, this approach highlights the correlation of A_1 and A_2 with A_3. Pursuing (exact or approxiamte) fairness, as shown in Figure 5, indicates that making either A_1 or A_2 fair first impacts the overall unfairness differently in the initial step of the debiasing path. Specifically, making A_2 fair first reduces overall fairness, while making A_1 fair slightly enhances fairness. However, beyond this initial step, both paths eventually converge to the same performance.

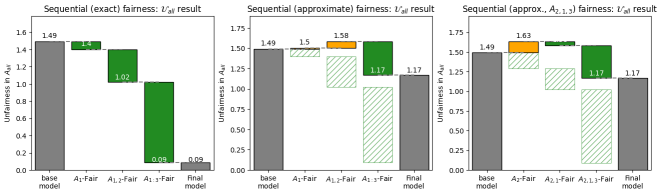

Figure 5: Synthetic data with parameter $\tau = (0, 0.05, 0.1)$. A sequential unfairness evaluation, $\mathcal{U} = \mathcal{U}_{1:3}$, of (*left pane*) exact fairness, (*middle pane*) approximate $A_{1,2,3}$-fairness with ε-RI where $\boldsymbol{\varepsilon} = \varepsilon_{1,2,3} = (0.2, 0.5, 0.75)$ and (*right pane*) approximate $A_{2,1,3}$-fairness with $\varepsilon_{2,1,3}$-RI. Hashed color corresponds to exact fairness.

Figures 6 and 7 display intermediate solutions of our sequential fair mechanism on the synthetic data. Figure 6 showcases exact fairness steps, while Figure 7 focuses on approximate fairness. Each row corresponds to enforcing fairness on an (additional) sensitive feature, while each column pertains to studying a specific sensitive feature. For example, the last (resp. first) row corresponds to enforcing fairness on A_1 (resp. A_3), and the last (resp. first) column corresponds to studying feature A_1 (resp. A_3). Both reveal intended score distribution differences, aligning perfectly (for exact fairness) or nearly so with some shifts (for approximate fairness), based on the chosen sensitive features (observed along the figures' diagonals).

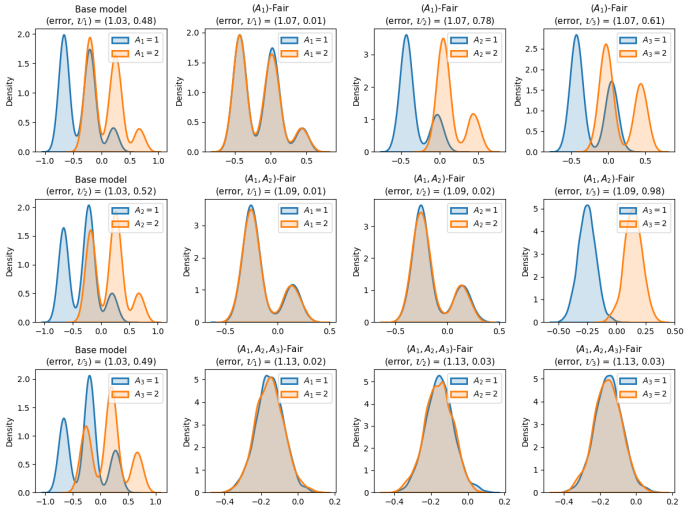

Figure 6: Synthetic data with parameters $\boldsymbol{\tau} = (0, 0.05, 0.1)$: sequential evaluation of exact fairness.

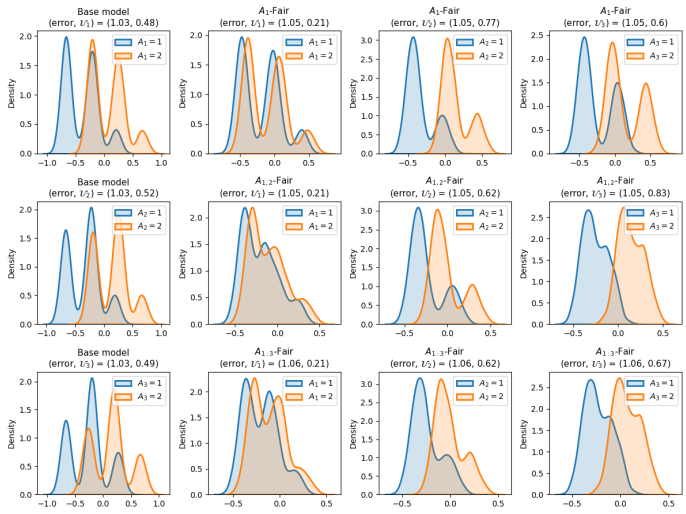

Figure 7: Synthetic data with parameters $\boldsymbol{\tau} = (0, 0.05, 0.1)$: evaluation of approximate fairness where $\boldsymbol{\varepsilon} = (0.2, 0.5, 0.75)$.