A Sequentially Fair Mechanism for Multiple Sensitive Attributes

François Hu1, Philipp Ratz2, Arthur Charpentier2

1Université de Montréal, Montréal, Québec, Canada
2Université 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 straightforward 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 inter-sensitive 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 across attributes; a combined variable is complex and also lacks prioritization challenging. Presetting a level of unfairness could prove detrimental. As an example, (Buolamwini and Gebru 2018) underscored this intersectional bias in Machine Learning (ML). They discovered that major face recognition algorithms exhibited reduced accuracy for women with darker skin tones, thereby exposing both gender and racial discrimination.

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, Dudík, and Wu 2019; Chiappa et al. 2020; Denis et al. 2021). However, these approaches consider fairness w.r.t. 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 reduced accuracy for women with darker skin tones, thereby exposing both gender and racial discrimination.

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 approach for approximating fairness across a set of sensitive features. Fairness considerations often lead 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.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
a 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.
- Our approach is demonstrated through numerical experiments on diverse datasets. 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 2023a) 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.

Whereas the majority of recent works in the intersectional fairness domain primarily focus on classification tasks (Gohar and Cheng 2023), this article explores both classification and often-overlooked regression tasks within the intersectional fairness framework. Our sequential approach, characterized by the barycenter associativity property, distinguishes itself from conventional methods by enhancing interpretability and naturally addressing inter-correlations among sensitive attributes. Our approach further mitigates data sparsity and uncertainty issues, providing a valuable complementary perspective to existing work in intersectional fairness (Gohar and Cheng 2023).

Within the MSA framework, multiple studies explore diverse perspectives distinct from ours. (Liu, Zhang, and Jiang 2023) and (Jalal et al. 2021) use sampling frameworks—one in an active learning setting and the other in image generation—to ensure equal representation of sensitive subgroups. Concurrently, (Kilbertus et al. 2018) addresses disparate impact within privacy constraints without explicit knowledge of sensitive attributes, operating within a fairness-unaware framework. In a similar vein, (Hu, Ratz, and Charpentier 2023b) introduces domain knowledge considerations to alleviate the negative effects of latent sensitive attributes when tackling biases associated with an observed sensitive attribute. In contrast, our method takes a direct approach by adjusting ML scores for fairness without the need to select fair samples, distinguishing it from the methodologies employed by (Liu, Zhang, and Jiang 2023) and (Jalal et al. 2021). Moreover, our approach operates within a fairness-aware framework, leveraging available sensitive attributes. This sets our method apart from (Kilbertus et al. 2018) and (Hu, Ratz, and Charpentier 2023b), which operate in fairness-unaware (or semi-unaware) frameworks.

### Notation

Consider a function $f$ and a random tuple $(X, A) \in \mathcal{X} \times \mathcal{A} \subset \mathbb{R}^d \times \mathbb{N}^r$, 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 $\nu_f|_a \in \mathcal{V}$ be respectively the probability measure of $f(X, A)$ and $f(X, A)|_A = a$. $F_{f|a}(u) := \mathbb{P}(f(X, A) \leq u|A = a)$ corresponds to the cumulative distribution function (CDF) of $\nu_f|a$ and $Q_{f|a}(v) := \inf\{u \in \mathbb{R} : F_{f|a}(u) \geq v\}$ 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

Let $\nu_1$ and $\nu_2$ be two probability measures. The Wasserstein distance quantifies the minimum ”cost” of transforming one distribution into the other. Specifically, the squared Wasserstein distance between $\nu_1$ and $\nu_2$ is defined as

$$W_2^2(\nu_1, \nu_2) = \inf_{\pi \in \Pi(\nu_1, \nu_2)} \mathbb{E}(Z_1, Z_2 \sim \pi) (Z_2 - Z_1)^2,$$

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

#### 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 $W_2$-barycenter), denoted as $\text{Bar}(w_k, \nu_k)_{k=1}^K$, is the minimizer

$$\text{Bar}(w_k, \nu_k)_{k=1}^K = \arg \min_{\nu} \sum_{k=1}^K w_k \cdot W_2^2(\nu_k, \nu).$$

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 $\nu_i$ 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 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 Wasserstein 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:

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

where \( \{w_{i,k}\}_{i,k} \) and \( \{\nu_{i,k}\}_{i,k} \) represent univariate measures. In this context, the overall Wasserstein barycenter \( \text{Bar} \{B_1 \cup \cdots \cup B_r\} \) can be expressed as the barycenter \( \text{Bar} \{\bar{w}_i, \text{Bar}(\bar{B}_i)\}_{i=1, \ldots, r} \), where

\[
\bar{w}_i := \sum_{k=1}^{K_i} w_{i,k} \bar{w}_i \quad \text{and} \quad \bar{B}_i := (w_{i,k}/\bar{w}_i, \nu_{i,k})_{k=1, \ldots, K_i}.
\]

This formulation captures the relationship between the overall Wasserstein barycenter and the individual barycenters of its constituent updates, updating the relevant weights. Given measures \( \nu_1, \nu_2, \nu_3 \) with positive weights \( w_1, w_2, w_3 \) summing to 1, mirroring the aforementioned barycentric concept in Euclidean space, we derive the relations:

\[
\text{Bar} \{(w_1, \nu_1), (w_2, \nu_2), (w_3, \nu_3)\} = \text{Bar} \{(w_1, \nu_1), (w_2 + w_3, \nu_2 + 3\nu_3)\} = \text{Bar} \{(w_1 + w_2, \nu_1 + 2\nu_2), (w_3, \nu_3)\}.
\]

For any \( i \neq j \), the measure \( \nu_{i,j} = \text{Bar} \{\{\bar{w}_i, \nu_i\}, (\bar{w}_j, \nu_j)\} \) is defined with \( \bar{w}_k = w_k/(w_i + w_j) \).

**Problem Formulation**

Let \((X, A, Y)\) be a random tuple with distribution \( \mathbb{P} \). Here, \( X \in \mathcal{X} \subset \mathbb{R}^d \) denotes the \( d \)-non-sensitive features, \( Y \in \mathcal{Y} \subset \mathbb{R} \) represents the target task, and \( A = (A_1, \ldots, A_r) \in \mathcal{A} := \{A_1, \ldots, A_r\} \) is the \( r \)-discrete sensitive attributes, where \( A_i = \{1, \ldots, K_i\} \) with \( K_i \in \mathbb{N} \). For example, in a binary case, we could have \( A_1 = \text{gender} \) and \( A_2 = \text{age} \). For convenience, we denote \( A_{i+1:k} := (A_i, A_{i+1}, \ldots, A_{i+k}) \) as the sequence of \( k + 1 \) sensitive features ranging from \( i \) to \( i+k \) (so \( A = A_{1:r} \)). Further, we denote \( \mathcal{F} \) the set of predictors of the form \( f : \mathcal{X} \times A \to \mathcal{Y} \) where we assume that each measure \( \nu_{f|a} \) 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 \( \{\nu_{f|a}\}_{a \in \mathcal{A}} \) 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,

\[
\text{(Risk Measure)} \quad R(f) := \mathbb{E}(Y - f(X, A))^2,
\]

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

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

\[
U_i(f) = \max_{a_i \in A_i} \int_{u \in [0,1]} |Q_f(u) - Q_{f|a_i}(u)| \, du,
\]

while for the multiple sensitive features \( A_1, \ldots, A_{i+k} \), their collective unfairness is simply assessed through:

\[
U_{i+\ldots,i+k}(f) = U_{i+k}(f) = U_{i}(f) + \cdots + U_{i+k}(f).
\]

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 \( A = A_{1:r} \) can be quantified by the unfairness measure,

\[
\text{(Unfairness measure)} \quad U(f) = U_{1:r}(f).
\]

Then \( f \) is called exactly fair if and only if

\[
U(f) = 0.
\]

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

\[
U_i(f) \leq \varepsilon_i \cdot U(i^*)(f)
\]

Therefore, in the context of approximate fairness, we are interested in the relative fairness improvement of a fair predictor w.r.t. Bayes’ rule \( f^* \), see (Chzhen and Schreuder 2022).

**Preliminary Results**

Recall the Wasserstein barycenter defined in Eq. (1). We consider measures \( \nu_{f|a} \) with corresponding weights \( p_a \in \mathcal{A} \), where \( p_a := \mathbb{P}(A = a) \). It is assumed that \( \min_a \{p_a\} \geq 0 \). Encompassing these measures is the Wasserstein barycenter, denoted \( \mu_A : \mathcal{V} \to \mathcal{V} \) with

\[
\mu_A(\nu_f) := \text{Bar}(p_a, \nu_{f|a})_{a \in \mathcal{A}} = \arg\min_{\nu} \sum_{a \in \mathcal{A}} p_a \cdot W_2^2(\nu_{f|a}, \nu).
\]
**Single Sensitive Attribute (SSA) Case** We consider a single sensitive attribute \( A \), belonging to the set \( A = \{1, \ldots, K\} \), with \( p_a := \mathbb{P}(A = a) \). Let \( f_B \in \mathcal{F} \), and let its measure be the Wasserstein barycenter \( \nu_{f_B} = \mu_A(v_{f^*}) \), where we recall that \( f^*(X, A) = \mathbb{E}[Y|X, A] \) is the Bayes rule which minimizes the squared risk. Prior research conducted by (Chzhen et al. 2020b) demonstrates that,

\[
f_B = \arg\min_{f \in \mathcal{F}} \{ \mathcal{R}(f) : \mathcal{U}(f) = 0 \}. 
\]

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

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

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

\[
f_{B_i}(x, a) = \left( \sum_{a' \in A_i} p_{a'} Q_{f^*|a'} \right) \circ F_{f^*|a} (f^*(x, a)),
\]

as the optimal fair predictor, ensuring fairness only across \( A_i \)-fair (short). By abuse of notation, we let \( p_{a'} := \mathbb{P}(A_i = a_i') \), \( F_{f^*|a} (u) := \mathbb{P}(f(X, A) \leq u|A_i = a_i') \) and \( Q_{f^*|a} \) 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; Gouder, Loubes, and Rigollet 2020), we demonstrate that fairness in the MSA problem can also be framed as the optimal transport problem involving the \( W_2 \)-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

\[
f_B = \arg\min_{f \in \mathcal{F}} \{ \mathcal{R}(f) : \mathcal{U}(f) = 0 \}. 
\]

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

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

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 \( \mu_A(\nu_{f^*}) \) defined in Prop. 4 can be equivalently expressed as follows:

\[
\mu_A(\nu_{f^*}) = \mu_{A_1} \circ \mu_{A_2} \circ \cdots \circ \mu_{A_r}(\nu_{f^*}).
\]

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

\[
\mu_A(\nu_{f^*}) = \mu_{A_{\sigma(1)}} \circ \mu_{A_{\sigma(2)}} \circ \cdots \circ \mu_{A_{\sigma(r)}}(\nu_{f^*}).
\]

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

\[
f_B(X, A) = (f_{B_1} \circ f_{B_2} \circ \cdots \circ f_{B_r})(X, A).
\]

We denote the corresponding measure as \( f_{B_i} \circ f_{B_{i+1}} \) as follows (with \( f^* \) serving as the default function):

\[
(f_{B_1} \circ f_{B_2})(x, a) = \left( \sum_{a_i' \in A_i} p_{a_i'} Q_{f_{B_i}|a_i'} \right) \circ F_{f_{B_i}|a_i} (f_{B_i}(x, a)).
\]

Introducing a sequential approach is pivotal for enhancing clarity. Indeed, this methodology helps in comprehending intricate concepts like approximate fairness with \( \varepsilon \)-RI defined in Definition 3, which involves improving fairness relatively and approximately with \( \varepsilon \) 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 \( \varepsilon \)-RI fairness (or \( \varepsilon \)-fairness for short), where \( \varepsilon = \varepsilon_{f^*} = (\varepsilon_1, \ldots, \varepsilon_r). \) In the context of the SSA framework (Chzhen and Schreuder 2022) employs geodesic parameterization. Specifically, considering \( A_1 \in \mathcal{A}_1 \) w.l.o.g., the predictor of the form:

\[
f_{B_1}^\varepsilon (X, A) = (1 - \varepsilon_1) \cdot f_{B_1}(X, A) + \varepsilon_1 \cdot f^*(X, A),
\]

achieves the optimal risk-fairness trade-off. Therefore,

\[
f_{B_1}^\varepsilon \in \arg\min_{f \in \mathcal{F}} \{ \mathcal{R}(f) : \mathcal{U}_1(f) \leq \varepsilon_1 \cdot \mathcal{U}_1(f^*) \}. 
\]

We denote the corresponding measure as \( \mu_{A_1}^\varepsilon (\nu_{f^*}) := \nu_{f_{B_1}^\varepsilon}(\nu_{f^*}). \)

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

\[
f_B^\varepsilon = \arg\min_{f \in \mathcal{F}} \{ \mathcal{R}(f) : \mathcal{U}(f) \leq \sum_{i=1,\ldots,r} \varepsilon_i \cdot \mathcal{U}_i(f^*) \},
\]

then

\[
\mu_A^\varepsilon (\nu_{f^*}) := \mu_{A_1}^\varepsilon \circ \cdots \circ \mu_{A_r}^\varepsilon (\nu_{f^*}).
\]

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 setting: for any permutation \( \sigma \in \mathcal{P}(S) \),

\[
    f_B^\sigma(X, A) = \left( f_{B_{\sigma(1)}}^\sigma \circ f_{B_{\sigma(2)}}^\sigma \circ \cdots \circ f_{B_{\sigma(r)}}^\sigma \right)(X, A) .
\]

**Data-Driven Procedure**

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

\[
    \hat{f}_{B_i}(x, a) = \left( \sum_{a_i' \in A_i} \hat{p}_{a_i'}\hat{Q}_{f_{A[a_i']}} \circ \hat{P}_{f_{A[a_i']}} \right)(\hat{f}(x, a)) , \tag{6}
\]

Here, \( \hat{p}_{a_i}, \hat{F}_{f_{A[a_i]}}, \) and \( \hat{Q}_{f_{A[a_i]}}, \) are empirical counterparts of \( p_{a_i}, F_{f_{A[a_i]}}, \) and \( Q_{f_{A[a_i]}}, \) respectively. 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 \( \mathcal{P} \), the post-processing method \( \hat{f}_{B_i} \) is a good estimator of \( f_{B_i} \). By composition, \( \hat{f}_{B} = \hat{f}_{B_{\sigma(1)}} \circ \cdots \circ \hat{f}_{B_{\sigma(r)}} \) and \( \hat{f}_{B}^\sigma = \hat{f}_{B_{\sigma(1)}}^\sigma \circ \cdots \circ \hat{f}_{B_{\sigma(r)}}^\sigma \) emerge as good estimators for \( f_{B} \) and \( f_{B}^\sigma \) respectively, enabling accurate and fair estimation of the instances. Note that the unfairness of \( \hat{f} \) is assessed on a test set using \( U(f) \), the empirical counterpart of Eq. (3). Predictive performance uses mean squared error (MSE) for regression and \( F_1 \)-score for classification.

**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 with Two Sensitive Attributes**

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 \((X, A, Y)\) with the following characteristics:

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

**Simulation Scheme**

Default parameters are set as follows: \( d = 10, r = 3, \sigma_X = 0.15, \) and \( \tau = (0, 0.05, 0.1) \). We generated 10,000 synthetic examples and divided the data into three sets (50% train, 25% test and 25% unlabeled). As a base model, we opt for a simple linear regression using default parameters from scikit-learn in Python. Then, we compare our sequential approach to this base model.

**Interpreting Intersectional Fairness**

In the context of \( r = 3 \) sensitive features, Figure 1 showcases \( \mathcal{U} \), an unfairness measure focusing on \( A_3 \). The sequential fair approach, detailed in Prop. 5, enhances interpretability by addressing inter-correlations 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 inures 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. Additional fairness experiments with this synthetic data are detailed in the supplementary materials.

**Case Study 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 approach. 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, income > 100K). 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 tasks, we rely on provided standard features and treat gender and race as sensitive attributes.

**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.

\(^1\)Code source at: https://p私立/phi-ra/SequentialFairness.
Figure 1: Synthetic data with $\tau = (0, 0.05, 0.1)$. A sequential unfairness evaluation, $U_3$, of (left pane) exact fairness, (middle) approximate $A_{1,3}$-fairness with $\varepsilon$-RI where $\varepsilon = \varepsilon_{1,2,3} = (0.2, 0.5, 0.75)$ and (right) approximate $A_{2,1,3}$-fairness with $\varepsilon_{2,1,3}$-RI.

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

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 easily achieved. For the main 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 $F_1$-Score, unfairness and compute-time. Next, our methodology is employed to ensure fairness across both variables, facilitating another comparison. To create further performance comparisons, we also run benchmarks on well-known data sets with differing number of sensitive variables ($s$ in Table 1) using Fairlearn and FairBalance from (Yu, Chakraborty, and Menzies 2021). Here, we create sensitive hyper-features to allow easier learning for the two benchmarking methods. This should help them learn a fair representation but will likely increase compute time.

**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. The last three sections in Table 1 contain the further performance benchmarks. Across all experiments, our method performs consistently fastest and best on the Fairness metric and competitive on the performance metric.

Continuing with our application example (see Figure 3), our data 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, these sub-groups could gain from fair predictions.

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
and benchmarking experiments.

Table 1: Results for the correction of the biases for specific attributes and intermediate solutions rendering predictions fair on only one of them. Center pane, marginal changes to predicted income when rendering fair the predictions w.r.t. 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.

<table>
<thead>
<tr>
<th></th>
<th>Uncorrected</th>
<th>Our Method</th>
<th>Fairlearn</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Regression</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>0.547 ± 0.003</td>
<td>0.596 ± 0.003</td>
<td>N/A</td>
</tr>
<tr>
<td>Unfairness</td>
<td>0.378 ± 0.007</td>
<td><strong>0.019 ± 0.005</strong></td>
<td>N/A</td>
</tr>
<tr>
<td>Time (s)</td>
<td>N/A</td>
<td>8.981 ± 1.319</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Classification, Income - one sensitive</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>0.753 ± 0.001</td>
<td>0.737 ± 0.001</td>
<td>0.73 ± 0.002</td>
</tr>
<tr>
<td>Unfairness</td>
<td>0.170 ± 0.001</td>
<td><strong>0.003 ± 0.002</strong></td>
<td>0.021 ± 0.002</td>
</tr>
<tr>
<td>Time (s)</td>
<td>N/A</td>
<td>6.319 ± 0.422</td>
<td>100.8 ± 10.467</td>
</tr>
<tr>
<td><strong>Classification, Income - two sensitive</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>0.753 ± 0.001</td>
<td>0.73 ± 0.001</td>
<td>0.73 ± 0.002</td>
</tr>
<tr>
<td>Unfairness</td>
<td>0.354 ± 0.006</td>
<td><strong>0.009 ± 0.005</strong></td>
<td>0.207 ± 0.005</td>
</tr>
<tr>
<td>Time (s)</td>
<td>N/A</td>
<td>10.661 ± 0.140</td>
<td>355.275 ± 6.346</td>
</tr>
</tbody>
</table>

Table 1: Results for the correction of the biases for specific and benchmarking experiments.

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 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 $\epsilon$ 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 that 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 DP constraint, particularly in the case of label-conditional unfairness metrics like equalized odds. The flexibility and user-friendliness of our method, supporting a comprehensive fairness approach across multiple debiasing steps, promotes the adoption of fair decision-making practices.
Acknowledgments

AC acknowledges Canada’s National Sciences and Engineering Research Council (NSERC) for funding (RGPIN-2019-07077) and AC and PR acknowledge funding from the SCOR Foundation for Science. This research was enabled in part by support provided by Calcul Québec (calculquebec.ca) and the Digital Research Alliance of Canada (alliancecan.ca). François Hu acknowledges funding from the NSERC Emerging Infectious Diseases Modelling Initiative (EIDM), awarded to the Mathematics for Public Health (MiPH) program.

References


