Fairness-Aware Post-Processing in Supervised Classification: L1/L2 Norm and Optimal Swapping Methods

Flore Vancompernolle Vromman*, **Sylvain Courtain, Pierre Leleux, Marco Saerens**

LouRIM & ICTEAM, Université catholique de Louvain, Belgium

Abstract

The growing use of AI in decision-making raises ethical concerns, including fairness in classification tasks. In this paper, we propose two fairness-aware post-processing methods – (1) the least L_1/L_2 norm with covariance constraints and (2) the optimal swapping – which address group fairness in the outputs of any probabilistic classifier. The first method incorporates fairness constraints and minimizes deviations from the classifier's initial probabilistic predictions through an interpolation between the L_1 and L_2 norms. The second method is a heuristic approach that directly modifies binary classification decisions through a simple, yet efficient swapping approach. We evaluate our methods on five benchmark datasets and compare them with some well-established baselines. Our results show that optimal swapping achieves the best trade-off between fairness and accuracy on the investigated datasets. While we focus on demographic parity and disparate impact, our swapping post-processing method is adaptable to multi-class settings, some other fairness definitions, and allows for additional linear constraints.

Introduction

General Introduction

Supervised classification algorithms are widely used in diverse and impactful domains, such as hiring, lending, criminal justice, and healthcare (Barocas, Hardt, and Narayanan 2023; Pessach and Shmueli 2022; Romei and Ruggieri 2014). Despite their utility, these models can inadvertently propagate or amplify biases present in the data, leading to unfair outcomes for certain groups of individuals. For example, a loan approval system might systematically deny loans to applicants from disadvantaged socioeconomic or racial groups. Such biases often stem from factors like non-representative training samples, proxy attributes (e.g., zip codes as a proxy for ethnicity), or historical discrimination encoded in the data (Mehrabi et al. 2021; Mitchell et al. 2021; Pessach and Shmueli 2022). These concerns have driven the field to focus on developing fairness-aware machine learning methods (see (Barocas, Hardt, and Narayanan 2023) and references therein). Among those fairness-aware methods, the post-processing ones operate directly on the model's predictions, ensuring compatibility with any probabilistic classifier and preserving the integrity of training workflows.

In this paper, we focus on developing and evaluating fairness-aware post-processing methods to address group fairness concerns in supervised multi-class classification problems with a preferred class. Specifically, we propose two methods designed to balance fairness and accuracy: (1) a *least L_1/L_2 norm with covariance constraints* (COV), which minimizes the divergence from the model's original probabilistic predictions while enforcing fairness constraints, and (2) an *optimal swapping* (OS) method, which adjusts classification decisions with two types of swaps to ensure fairness at the decision level.

In the experimental section, results obtained across five datasets and five classifiers show that the optimal swapping method achieves the best trade-off between fairness and accuracy, outperforming some established methods from the literature. Furthermore, our methods could be applied to any supervised classification model with probabilistic outputs, ensuring their applicability in diverse real-world scenarios. It will also be shown that they could incorporate other constraints, allowing different measures of fairness, or introducing other external constraints such as, for instance, bounding the total number of instances classified in the preferred class. In particular, while we focus on demographic parity and disparate impact, our methods are not limited to these fairness definitions: as shown later, any fairness constraint that can be expressed as a linear function (for the COV method) or as a known function (for the OS method) could be integrated in the same framework.

Notation and Background

Before presenting our work and methods, we define some notations used in the paper. Let us assume that a multi-class supervised classification model has been fitted on a training set and applied to a test set to be classified (Duda, Hart, and Stork 2001; Bishop 2006; Hastie, Tibshirani, and Friedman 2009; Sen, Hajra, and Ghosh 2020; Murphy 2022). This classifier is assumed to provide the a posteriori probability of belonging to one of m mutually exclusive classes (Santafe, Inza, and Lozano 2015). The classification model is denoted by $\hat{y}_{ik} = g(\mathbf{x}_i)$ where $\hat{y}_{ik} \in [0, 1]$ is the numerical prediction of class k membership for instance i, based on the fea-

*Contact: flore.vancompernolle@uclouvain.be

ture vector $\mathbf{x}_i$. The binary decision predicted by the classifier to assign the feature vector $\mathbf{x}_i$ to class k is $\hat{y}^{\mathrm{d}}_{ik} \in \{0, 1\}$ (d for decision – one-hot encoding technique). Moreover, the observed real categories for each instance of the training set are y^{d}_{ik}. The observed classes and the predicted probabilities are therefore encoded in the $n \times m$ (n instances and m classes) matrices $\mathbf{Y}^{\mathrm{d}}$ and $\hat{\mathbf{Y}}$. In the same way, $\mathbf{y}^{\mathrm{d}}_i$ and $\hat{\mathbf{y}}_i$ are respectively $m \times 1$ column vectors containing the observed classes and the predictions for instance i (the i-row of $\mathbf{Y}^{\mathrm{d}}$ and $\hat{\mathbf{Y}}$ viewed as column vectors).

Based on the prediction matrix $\hat{\mathbf{Y}}$ obtained by any classifier that provides membership probabilities, fairness-aware post-processing methods compute a new prediction matrix $\tilde{\mathbf{Y}}$, where the goal is usually to satisfy a specific definition of fairness while remaining as close as possible to the original predictions $\hat{\mathbf{Y}}$, to only allow minimal perturbations to the original model. These new predictions provide new degrees of membership to classes summing to 1. Again, $\tilde{\mathbf{y}}_i$ is the vector of revised predictions associated with instance i.

Although the methods are developed for multi-class problems, in the experimental section of this work, we mainly focus on binary classification problems, a prevalent type of supervised classification task, by evaluating our methods on five datasets designed for such problems. In that context, we consider one positive/preferred class in which it is preferable to be classified, $Y^{\mathrm{d}}_1 = 1$. As an example, if an applicant loan i is 'approved', $y^{\mathrm{d}}_{i1} = 1$; if it is 'denied', $y^{\mathrm{d}}_{i1} = 0$. In this example, with $\hat{y}_{i1} = 0.65$, the classifier estimates that the probability of instance i belonging to class 1 is 0.65 and the predicted class (decision) for this instance i is 1, $\hat{y}^{\mathrm{d}}_{i1} = 1$ (assuming that the decision threshold is 0.5, which minimizes the risk of error). After applying the post-processing method, the predicted class decision for instance i could change and become $\tilde{y}^{\mathrm{d}}_{i1} = 0$, meaning that the classifier predicts applicant loan i to be approved while the post-processing fairness adjustments recommend the applicant loan i to be declined.

Fairness may be defined from different perspectives, as detailed further. In our work, we focus on group fairness, which ensures that groups of individuals identified by a sensitive (random) variable should be treated similarly (Dwork et al. 2012). The sensitive variable is noted Z, with the protected group (i.e., the group disadvantaged/discriminated against) associated with $Z = 1$. With the example of loan approval, the sensitive variable could be the age, and the protected (i.e., discriminated) group could correspond to people under 25. In that example, a person i under 25 has $z_i = 1$, while someone over 25 has $z_i = 0$.

In our first post-processing method (COV), we use the centered sensitive variable to incorporate it in the covariance that is centered. We therefore define the centered sensitive variable as $\dot{\mathbf{z}} = \mathbf{H}\mathbf{z}$, where $\mathbf{H} = \left(\mathbf{I} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathsf{T}}\right)$ is the centering matrix (Mardia, Kent, and Bibby 1979), and $\mathbf{e}$ is a $n \times 1$ column vector containing 1's. Note that $\mathbf{H}$ centers any vector ($\mathbf{e}^{\mathsf{T}}\mathbf{H}\mathbf{z} = 0$) and that $\mathbf{H}\mathbf{H} = \mathbf{H}$.

To summarize those concepts, the main notations regarding fairness-aware supervised classification algorithms used in this work are presented in Table 1.

Notation	Meaning
y^{d}_{ik}	Real, observed, class k assigned to instance i.
$\hat{y}_{ik}$	Probabilistic output predicted (by the classifier) to assign instance i to class k.
$\hat{y}^{\mathrm{d}}_{ik}$	Binary decision predicted (by the classifier) to assign instance i to class k (one-hot encoding).
$\tilde{y}_{ik}$	New predicted probability (by the post-processing method) to assign instance i to class k.
$\tilde{y}^{\mathrm{d}}_{ik}$	New binary predicted decision (by the post-processing method) to assign instance i to class k.
z_i	Real, observed, binary sensitive variable for instance i.
$\bar{z}$	Mean of $\mathbf{z}$, containing the z_i.
$\dot{z}_i = z_i - \bar{z}$	Centered sensitive variable for instance i.

Table 1: Notation used in the paper. Corresponding column vectors will be in bold and matrices in uppercase bold.

Related Work

In this subsection, we first discuss the general concept of fairness in supervised classification and the specific interpretation of fairness used in our paper. Then, we detail the existing methods to incorporate fairness in supervised classification, specifically the post-processing methods.

Fairness in Supervised Classification A growing concern in supervised classification is the potential unfairness of algorithmic predictions. Fairness in classification can be defined as individual (i.e., individuals with similar characteristics should be treated similarly) or group (i.e., groups of individuals defined by sensitive variables should be treated similarly) (Dwork et al. 2012). In our work, we focus on group fairness, using a binary sensitive variable Z (e.g., gender, ethnicity, age, religion, etc.) , where $Z = 1$ indicates the protected group (i.e., the discriminated group).

Several metrics exist for group fairness. In this work, we focus on fairness metrics that do not require the true target Y, such as demographic parity (DP) or disparate impact (DI) (Barocas, Hardt, and Narayanan 2023), allowing broader applicability. Metrics requiring Y, such as equal opportunity (Mehrabi et al. 2021), are left for future work. Without knowledge of the true target value, group fairness could be ideally defined as the independence between predictions and the sensitive variable, i.e., $\mathbb{P}[\hat{Y}^{\mathrm{d}}_1 = 1|Z = 0] = \mathbb{P}[\hat{Y}^{\mathrm{d}}_1 = 1|Z = 1]$ (Barocas, Hardt, and Narayanan 2023). From this arise DP (absolute difference) and DI (ratio) (Mehrabi et al. 2021; Feldman et al. 2015):

$$\mathrm{DP} = \left|\mathbb{P}[\hat{Y}^{\mathrm{d}}_1 = 1|Z = 0] - \mathbb{P}[\hat{Y}^{\mathrm{d}}_1 = 1|Z = 1]\right| \quad (1)$$

$$\mathrm{DI} = \frac{\mathbb{P}[\hat{Y}^{\mathrm{d}}_1 = 1|Z = 1]}{\mathbb{P}[\hat{Y}^{\mathrm{d}}_1 = 1|Z = 0]} \quad (2)$$

A tolerance level is often allowed, expressed as $\mathrm{DP} \leq \varepsilon$ and $\mathrm{DI} \geq 1 - \varepsilon$ where $\varepsilon > 0$.

Interestingly, as shown in (Vancompernolle Vromman et al. 2024), the sample covariance between $\hat{Y}_1$ and Z can be used as a soft[1] proxy for DP, and has the advantage to also cover the case where the protected variable is continuous, in which case it can be used for de-correlating class predictions and protected variable. This covariance between $\hat{\mathbf{y}}_1$ and $\mathbf{z}$ (in absolute value) can be computed from

$$\text{cov}(\hat{\mathbf{y}}_1, \mathbf{z}) = \left| \frac{1}{n-1} \sum_{i=1}^{n} \hat{y}_{i1} \dot{z}_i \right| \tag{3}$$

Enforcing fairness often leads to a trade-off with accuracy (Berk et al. 2021; Kamiran and Calders 2012). To measure it, we use the F1 demographic parity (F1DP) (Vancompernolle Vromman et al. 2024), an adaptation of the F1-score:

$$\text{F1DP} = 2 \times \frac{\text{ACC} \times (1 - \text{DP})}{\text{ACC} + (1 - \text{DP})} \tag{4}$$

It favors models achieving high accuracy and fairness.

Fairness-Aware Post-Processing Methods for Supervised Classification Fairness in supervised classification can be improved via pre-processing (data transformation), in-processing (model training modification), or post-processing (adjusting predictions) (Alves et al. 2023; Chen, Wu, and Wang 2023; Ferrara 2023; Hort et al. 2024; Jui and Rivas 2024). The present work focuses on fairness-aware post-processing techniques. These methods offer several advantages: they suit any probabilistic classifier, require no re-training, enhance interpretability, and enable end-users to adjust fairness levels directly (Caton and Haas 2024; Kamiran, Karim, and Zhang 2012; Kim and Cho 2022).

Various post-processing techniques exist (see, e.g., (Chen, Wu, and Wang 2023; Pessach and Shmueli 2022)). One widely recognized post-processing method is proposed in (Hardt, Price, and Srebro 2016). This method enforces equalized odds as a fairness measure, which imposes that the classifier's true positive rate and false positive rate are identical across all groups defined by the sensitive variable. This method requires access to the true target values Y for adjustment, which can be a limitation in scenarios where these values are unavailable. Similarly, (Mishler, Kennedy, and Chouldechova 2021) extend this method by further exploring its implications in counterfactual settings. Calibration is another post-processing technique, adjusting a model output to match the true probability of an event (Jui and Rivas 2024). (Pleiss et al. 2017) develop a post-processing method that includes calibration for equalized odds. Some post-processing methods address both individual and group fairness (Lohia et al. 2019; Noriega-Campero et al. 2019; Small et al. 2024). Other adjust classification thresholds across groups to enhance fairness. For example, (Jang, Shi, and Wang 2022) adapt classification thresholds for each demographic group based on the confusion matrix. Another well-known method is Reject Option Classification (ROC) (Kamiran, Karim, and Zhang 2012), which modifies predictions near the decision boundary. (Kleinberg et al. 2018) propose a simple ranking-based approach that selects top-scoring individuals from each group for the positive class. However, this method requires prior knowledge of how many instances should be selected from each group. This ranking-based approach has inspired our optimal swapping method (OS). Similarly, the massaging method developed by (Kamiran and Calders 2009), initially introduced as a pre-processing method, has directly influenced the design of our swapping mechanism. Recent works have explored the use of optimal transport formulations to address fairness in classification (Gordaliza et al. 2019; Lazar Reich and Vijaykumar 2021; Xian, Yin, and Zhao 2023) as a post-processing step. In these approaches, fairness constraints are incorporated in a transportation optimization problem, often with convex cost functions between $\tilde{\mathbf{y}}_i$ and $\hat{\mathbf{y}}_i$ (Lazar Reich and Vijaykumar 2021; Xian, Yin, and Zhao 2023). These methods offer a flexible mathematical framework for reasoning about fairness through cost minimization, and they allow for modeling the trade-off between accuracy and fairness. The main relationships between our proposed techniques and the optimal transport ones will be discussed later in the paper.

[1]Using probabilistic outputs, $\hat{Y}_1$, instead of discrete class assignments, $\hat{Y}_1^{\text{d}}$. In that case, for a given test set, the covariance is proportional to the soft DP without taking the absolute value, $\mathbb{P}[\hat{Y}_1 = 1|Z = 0] - \mathbb{P}[\hat{Y}_1 = 1|Z = 1]$.

Contributions and Content

This work contributes to (1) the development of two novel fairness-aware post-processing methods, least L_1/L_2 norm with covariance constraints and optimal swapping, which balance fairness and accuracy through optimization formulations, including fairness constraints and the use of a parameter (γ) for interpolation between L_1 and L_2 norms; (2) the use of the massaging method of Kamiran and Calders (Kamiran and Calders 2009) by adapting label swapping from a pre-processing to a post-processing setting; (3) the validation and experimental comparisons of the proposed methods with established methods on five datasets, showing that optimal swapping achieves the best DP and trade-off between fairness and accuracy, evaluated using F1DP on the investigated datasets; (4) the ability to extend the proposed methods by incorporating additional linear constraints, broadening the scope of their applicability to different definitions of fairness.

The paper is organized as follows. We first present the two proposed post-processing methods: *least L_1/L_2 norm with covariance constraints* (COV), and *optimal swapping* (OS). Then, we describe the experimental setup used to compare our methods with those of (Kamiran and Calders 2009) and (Kamiran, Karim, and Zhang 2012) on five datasets. Finally, we conclude and outline future research directions.

Development of Fairness-Aware Post-Processing Methods

In this section, we develop two different fairness-aware post-processing methods that determine the new predicted decisions $\tilde{\mathbf{Y}}^{\text{d}}$ in case of DP/DI and multi-class, as well as simple binary, classification problems. To the best of our knowledge, such methods have not yet been proposed in the literature.

Least L1/L2 Norm with Covariance Constraints

This section introduces the first fairness-aware post-processing method called *least L_1/L_2 norm with covariance constraints* (shortened later as COV) which is particularly simple, but also quite general. This method can be seen as a simple instance of an optimal mapping problem, where predicted probability distributions $\hat{\mathbf{Y}}$ are adjusted to continuous fairer alternatives $\tilde{\mathbf{Y}}$ at minimal cost. It is therefore related to optimal transport techniques which have recently been applied to fairness-aware classification (Gordaliza et al. 2019; Lazar Reich and Vijaykumar 2021; Xian, Yin, and Zhao 2023). Such mapping to fairer continuous predictions (instead of binary decision for supervised classification) especially makes sense in applications where these predictions are used for computing derived quantities of interest, depending on the probability of membership to the different output classes like, for instance, in cost-sensitive learning (Elkan 2001).

More precisely, the objective function minimizes a (usually convex) loss function between the $\tilde{\mathbf{y}}_i$ and $\hat{\mathbf{y}}_i$ under a fairness constraint. Several options exist for this objective function. A very common choice is to use the least squares, or L_2 norm, which is the sum of squared differences (Golub and Van Loan 1996; Hansen, Pereyra, and Scherer 2013; Hastie, Tibshirani, and Friedman 2009),

$$L_2(\tilde{\mathbf{Y}}, \hat{\mathbf{Y}}) = \sum_{i=1}^{n} \|\tilde{\mathbf{y}}_i - \hat{\mathbf{y}}_i\|_2^2 = \sum_{i=1}^{n}\sum_{k=1}^{m} (\tilde{y}_{ik} - \hat{y}_{ik})^2 \quad (5)$$

where n is the number of instances in the test set and m is the number of classes. The L_2 norm is often used because it penalizes large deviations more heavily than small ones due to squaring the differences. This property can be beneficial when stability and global convergence are desired, but it is also quite sensitive to outliers (Golub and Van Loan 1996; Hastie, Tibshirani, and Friedman 2009).

We could also consider the L_1 norm, which is the absolute value of differences (Hastie, Tibshirani, and Friedman 2009), leading to the least absolute values loss,

$$L_1(\tilde{\mathbf{Y}}, \hat{\mathbf{Y}}) = \sum_{i=1}^{n} \|\tilde{\mathbf{y}}_i - \hat{\mathbf{y}}_i\|_1 = \sum_{i=1}^{n}\sum_{k=1}^{m} |\tilde{y}_{ik} - \hat{y}_{ik}| \quad (6)$$

The L_1 norm is advantageous in some situations because it is more robust to outliers than the L_2 norm. By taking the absolute values of the differences, extreme values less influence the L_1 norm, making it suitable for scenarios where the dataset may contain noisy or erroneous outliers (Hastie, Tibshirani, and Friedman 2009).

Since L_1 and L_2 norms have their specificities in measuring the difference between the previous and the new predictions, we decided to test combinations of those measures by incorporating a parameter $\gamma \in [0, 1]$ interpolating between the two norms, providing the following loss

$$L(\tilde{\mathbf{Y}}, \hat{\mathbf{Y}}) = \sum_{i=1}^{n} \ell_i(\tilde{\mathbf{y}}_i, \hat{\mathbf{y}}_i)$$

$$= \sum_{i=1}^{n} \underbrace{\sum_{k=1}^{m} \left(\gamma|\tilde{y}_{ik} - \hat{y}_{ik}| + (1-\gamma)(\tilde{y}_{ik} - \hat{y}_{ik})^2\right)}_{= \ell_i(\tilde{\mathbf{y}}_i, \hat{\mathbf{y}}_i)} \quad (7)$$

Equation (7) is used as the loss function for our fitting problem, leading to the following convex optimization problem to determine the new matrix of predictions $\tilde{\mathbf{Y}}$,

$$\begin{array}{lll} \underset{\{\tilde{y}_{ik}\}}{\text{minimize}} & \displaystyle\sum_{i=1}^{n}\sum_{k=1}^{m} \left(\gamma|\tilde{y}_{ik} - \hat{y}_{ik}| + (1-\gamma)(\tilde{y}_{ik} - \hat{y}_{ik})^2\right) & \\ \text{subject to} & \tilde{y}_{ik} \geq 0 & \text{for all } i, k \\ & \sum_{k=1}^{m} \tilde{y}_{ik} = 1 & \text{for all } i \\ & \left|\frac{1}{n-1}\sum_{i=1}^{n} \tilde{y}_{ik}\dot{z}_i\right| \leq \varepsilon & \text{for some } k \end{array} \quad (8)$$

The first two constraints impose $\tilde{y}_{ik}$ to belong to the interval $[0, 1]$ and sum to 1. The last constraint of the optimization problem is the fairness constraint. As described in the Related Work, the sample covariance between the categorical variable Z, representing the protected group, and the new predicted target variable $\tilde{Y}$ can be used as a soft proxy of DP. Therefore, this constraint imposes that the sample covariance on the test set should be no larger than a given threshold $\varepsilon \geq 0$ (a tolerance level; see (Vancompernolle Vromman et al. 2024) for details).

Notice that fairness constraints can be introduced for several classes, depending on the application. Our framework also allows the integration of other fairness constraints beyond DP, provided they are expressed as linear constraints. However, incorporating multiple fairness constraints could be conflicting, leading to an infeasible problem (Barocas, Hardt, and Narayanan 2023; Castelnovo et al. 2022; Kleinberg, Mullainathan, and Raghavan 2016). If this occurs, we are in a problem setting for which there is no optimal solution. In that case, some constraints have to be relaxed, and the minimal ε for each constraint could be found by minimizing the covariance related to this constraint while maintaining the other constraints active (see (Vancompernolle Vromman et al. 2024) for an example in a related context). In our experiments, the convex optimization problem (8) will be solved using the CVXPY library (Diamond and Boyd 2016).

This method is particularly relevant when fairness must be enforced on probabilistic scores, as in risk assessment or applications relying on soft outputs. However, the main drawback of it is that it constrains the soft DP, although what often matters in supervised classification is the DP based on the decisions. We now introduce a swapping mechanism to tackle this issue.

Optimal Swapping

This section aims to develop a class-swapping mechanism that minimizes the same cumulated incurred loss as for the previous method, capturing the 'distance' to the predictions of the trained classifier until the fairness constraint is satisfied. As before, we assume that one class, say class 1, is the positive/preferred class, inducing one fairness constraint.

The situation with several positive classes is more complex and left for future work.

Swapping Mechanism Our second fairness-aware post-processing method, called *optimal swapping* (shortened later as OS), aims to enhance fairness based on a switching approach inspired by (Kamiran and Calders 2009; Kamiran, Karim, and Zhang 2012; Kleinberg et al. 2018; Hardt, Price, and Srebro 2016) that modifies the initial classifier decisions. Like our first method, optimal swapping uses the initial probabilistic predictions of the classifier $\hat{\mathbf{Y}}$, and the binary sensitive attribute vector $\mathbf{z}$. However, unlike the previous method, which adjusted probability predictions $\tilde{\mathbf{Y}}$, the OS method focuses on adjusting decision predictions $\tilde{\mathbf{Y}}^{\mathrm{d}}$. For example, if a classifier assigns a probability prediction $\hat{y}_{i1} = 0.67$ (inducing $\hat{y}^{\mathrm{d}}_{i1} = 1$), our previous COV post-processing method could determine an adjusted probability prediction $\tilde{y}_{i1} = 0.45$ while our OS post-processing method would directly provide an adjusted decision prediction $\tilde{y}^{\mathrm{d}}_{i1} = 0$.

This method is also related to a discrete version of an optimal transport problem, where instances are reassigned to different classes through an optimal mapping into the barycenter, and the cost corresponds to the change in distance (for instance L_1) from the original classifier outputs (Xian, Yin, and Zhao 2023). Optimal transport methods also find the optimal mapping and are more generic, but only compute one, optimal, solution, while our proposed swapping algorithm provides a whole set of candidate solutions together with their quality, is able to integrate additional constraints (see Subsection *Additional constraints* further for more details), is easy to implement and fast to compute.

Unfairness, as defined by DP and DI (see Equations (1) and (2)), is detected when the probability of being classified as positive is lower for individuals from the protected group ($Z = 1$) than for individuals from the unprotected group ($Z = 0$). The intuition behind our swapping method is to reduce this unfairness by swapping some of the original classification decisions. This mechanism solves the main restriction of the COV method, namely that DP was only guaranteed probabilistically based on the degrees of membership provided by the classifier, $\tilde{\mathbf{Y}}$ (see last equation of Problem 8). With the COV method, even if constraints are satisfied, once the decision $\tilde{\mathbf{Y}}^{\mathrm{d}}$ is taken, by assigning each instance to the class showing the highest prediction score, the fairness constraints in terms of crisp decisions could be violated. However, what often matters in many applications is the final decision, not the probabilistic output. The swapping method guarantees fairness on the test set, and no more probabilistically as in our previous method. In this setting, the natural extension of Problem (8) $\tilde{y}^{\mathrm{d}}_{ik}$, aims at replacing the domain of values $\tilde{y}_{ik} \in [0, 1]$ by $\tilde{y}^{\mathrm{d}}_{ik} \in \{0, 1\}$ based on the binary decisions, which results in a 0-1 integer optimization problem, more difficult to solve than Problem (8).

To address this issue, we design the following swapping mechanism. It starts with the initial decisions of the classifier $\hat{\mathbf{Y}}^{\mathrm{d}}$, then swapping the class of a protected instance i ($z_i = 1$) initially classified in a negative class ($\hat{y}_{i1} = 0$) to the positive class ($\tilde{y}_{i1} = 1$), or swapping the class of an unprotected instance i ($z_i = 0$) initially classified in the positive class ($\hat{y}_{i1} = 1$) to a negative class ($\tilde{y}_{i1} = 0$). Thus, the method involves two types of swaps that adjust the decision to reach a desired level of fairness,

- s_1: classifying an instance i of the protected group from their original negative class to the positive class:

$$s_1: \quad \hat{y}^{\mathrm{d}}_{i1} = 0 \longmapsto \tilde{y}^{\mathrm{d}}_{i1} = 1 \quad \text{when } z_i = 1 \tag{9}$$

- s_0: classifying an instance i of the unprotected class from their original positive class to the negative class whose prediction score is highest (the class with the second-best membership for instance i):

$$s_0: \quad \hat{y}^{\mathrm{d}}_{i1} = 1 \longmapsto \tilde{y}^{\mathrm{d}}_{i1} = 0 \quad \text{when } z_i = 0 \tag{10}$$

We avoid changing the decision of positively classified instances from the protected group and of negatively classified instances from the unprotected group because this would increase discrimination, as defined by DP and DI. Note that, unlike (Kamiran and Calders 2009), the proposed mechanism can change the number of positively classified instances.

Similarly to Equation (7), and considering now the decisions, the loss function remains additive. The increase in loss when modifying the decision (by swapping) of a single instance i is therefore computed using the same L_1/L_2 interpolated loss as before, controlled by the parameter γ:

$$\begin{aligned} \Delta\ell'_i(s) &= \ell'(\tilde{\mathbf{Y}}^{\mathrm{d}}(s(i)), \hat{\mathbf{Y}}) - \ell(\hat{\mathbf{Y}}^{\mathrm{d}}, \hat{\mathbf{Y}}) \\ &= \ell'_i(\tilde{\mathbf{y}}^{\mathrm{d}}_i(s), \hat{\mathbf{y}}_i) - \ell_i(\hat{\mathbf{y}}^{\mathrm{d}}_i, \hat{\mathbf{y}}_i) \end{aligned} \tag{11}$$

where $s \in \{s_0, s_1\}$ and $\ell'(\tilde{\mathbf{Y}}^{\mathrm{d}}(s(i)), \hat{\mathbf{Y}})$ is the loss after moving the instance i from one class to another according to swap s. The difference with the distance of the original assignment to classes, $\hat{\mathbf{Y}}^{\mathrm{d}}$, from the probabilistic predictions, $\hat{\mathbf{Y}}$, is taken in order to obtain a 0 contribution when there is no swap, but other cost functions can be used as well. Of course, it is assumed that this original assignment to classes is optimal so that any swap will lead to a non-negative cost. The objective is now to determine the sequence of swaps minimizing the loss changes until reaching a state (we define a 'state' as a pair (n_1, n_0) – see Equation (12)) satisfying the fairness constraint.

Defining the Fairness Frontier For illustration, we built a grid (see Figure 1), where the x-axis represents the number of protected instances classified in the positive class, n_1, and the y-axis represents the number of unprotected instances classified in the positive class, n_0,

$$\begin{aligned} n_1 &\triangleq N\big(Y^{\mathrm{d}}_1 = 1 \wedge Z = 1\big) \\ n_0 &\triangleq N\big(Y^{\mathrm{d}}_1 = 1 \wedge Z = 0\big) \end{aligned} \tag{12}$$

where $N(\cdot)$ is the number of instances in the dataset verifying the condition expressed through the random variables, but computed empirically on the dataset (here, a test set).

Each black point on the grid[2] represents a (n_1, n_0) state, and the horizontal as well as vertical distance between two

[2]Note that the origin of the grid is not $(0, 0)$.

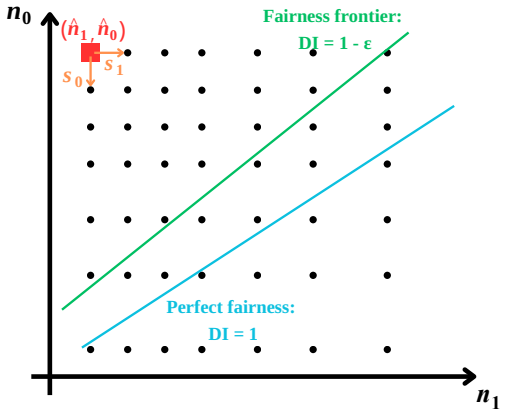

Figure 1: Swapping grid. The initial state $(\hat{n}_1, \hat{n}_0)$ is computed from the original predictions of the classifier.

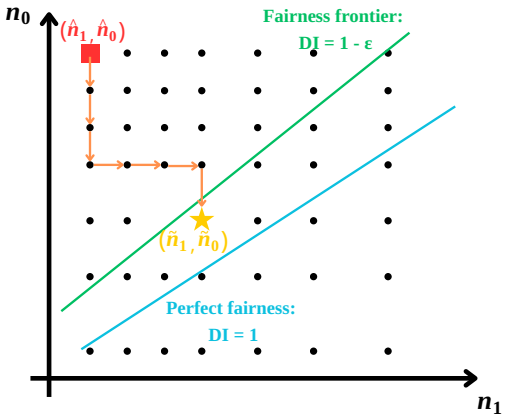

Figure 2: Swapping grid for the greedy swapping post-processing method. The final state is $(\tilde{n}_1, \tilde{n}_0)$ and we also show the swaps done to reach it.

states correspond to the cost of performing the swap from one state to the other one, $\Delta\ell'_i(s)$ (see Equation (11)). When the classifier determines the probability predictions $\hat{\mathbf{Y}}$, the corresponding state on the grid of Figure 1 is the large red square at the top left, $(\hat{n}_1, \hat{n}_0)$. Starting from that state, our OS post-processing method involves a mix of s_0 and s_1 swaps. A s_0 swap will decrease n_0 of one unit (vertical move) while a s_1 swap will increase n_1 of one unit (horizontal move).

To determine which and how many swaps to perform, we define the fairness frontier based on the DI[3] fairness measure (DI). By rewriting Equation (2) in terms of $N(\cdot)$, we obtain

$$\text{DI} = \frac{N\big(\tilde{Y}_1^{\text{d}} = 1 \wedge Z = 1\big)/N(Z=1)}{N\big(\tilde{Y}_1^{\text{d}} = 1 \wedge Z = 0\big)/N(Z=0)} = \frac{n_1/N(Z=1)}{n_0/N(Z=0)} \tag{13}$$

This work refers to 'perfect fairness' as the scenario where DI $= 1$, equivalent to DP $= 0$. However, it is essential to acknowledge that the term 'perfect fairness' may be questionable, as it depends on the specific fairness measure applied. Given the multitude of definitions of fairness, this interpretation may not fully capture all dimensions of fairness, or all possible applications. As before, we define a relaxation of it by fixing DI $\geq 1 - \varepsilon$ where ε is an adjustable fairness threshold. We will refer to this relaxation as the 'fairness frontier'. Of course, it is assumed that the initial solution $(\hat{n}_1, \hat{n}_0)$ does not satisfy this fairness constraint; if it does, no fairness improvement mechanism is needed. Because $N(Z = 1)$ and $N(Z = 0)$ are constant for a given test set, equations DI $= 1$ and DI $= 1 - \varepsilon$ correspond to straight lines on the (n_1, n_0) plane. Based on that, we draw on the swapping grid in Figure 1 the perfect fairness in blue (DI $= 1$) and the fairness frontier in green (DI $= 1 - \varepsilon$).

Greedy Swapping On the grid represented in Figure 1, several paths from the initial state exist to reach the fairness frontier, on which our fairness objective is satisfied. Inspired by (Kamiran and Calders 2009), the greedy swapping (shortened later as GS) method consists of selecting the lowest-cost swap starting from $(\hat{n}_1, \hat{n}_0)$ (the state predicted by the classifier), then the next lowest-cost swap, and so on until reaching the fairness frontier. As a special example based on post-processing methods using classification threshold to improve fairness (see among others (Jang, Shi, and Wang 2022; Kamiran, Karim, and Zhang 2012)), we consider here that the cost of performing a swap (see Equation (11)) is provided by the distance between the probabilistic prediction and the decision threshold (which usually corresponds to 0.5), $|\hat{y}_{i1} - 0.5|$, in a binary classification problem.

Figure 2 illustrates the GS post-processing method. It is defined on two-classes problems and applies the following procedure:

1. Select all candidates for swaps of type s_1 or s_0 (see Equations (9) and (10)).
2. Order those candidates (i.e., candidates for s_1 are protected instances with $\hat{y}_{i1} < 0.5$, and candidates for s_0 are unprotected instances with $\hat{y}_{i1} \geq 0.5$) closest to 0.5 first in a joint list.
3. Swap the first element of the list not yet switched, corresponding to the current most uncertain instance.
4. Repeat Step 3 till the state is just below the fairness frontier. The end state $(\tilde{n}_1, \tilde{n}_0)$ is the yellow star in Figure 2.

Optimal Swapping The non-greedy version of our swapping post-processing, heuristic, method (i.e., the optimal swapping, OS) is illustrated in Figure 3. Inspired by the ranking approach developed in (Kleinberg et al. 2018), it consists of determining the state on the grid just below the fairness frontier (in green) that is closest to the initial state $(\hat{n}_1, \hat{n}_0)$ in terms of the cumulated swapping cost $\Delta\ell'_i(s)$ defined in Equation (11). Instead of sequentially looking to the next best, local, move as in the greedy approach, the idea here is to efficiently identify all the candidate solutions along

[3]The same procedure can straightforwardly be adapted to DP.

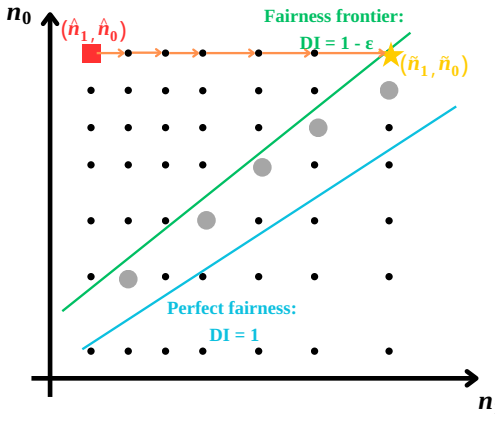

Figure 3: Swapping grid for the optimal swapping post-processing method. Big gray points correspond to candidates satisfying DI $> 1 - \varepsilon$, and the final state is $(\tilde{n}_1, \tilde{n}_0)$.

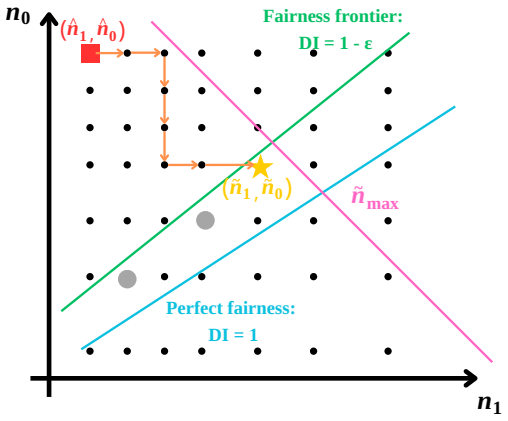

Figure 4: Swapping grid for the optimal swapping post-processing method with a constraint on the number of instances classified in the positive class, $\tilde{n}_0 + \tilde{n}_1 \leq \tilde{n}_{\max}$.

the fairness frontier – together with their loss – and then select the best one (see Figure 3).

Because the different costs $\Delta\ell_i'(s)$ related to different instances i have an independent impact on the total loss function, the order in which the moves are made to reach a given state $(\tilde{n}_1, \tilde{n}_0)$ from the initial state does not matter. In addition, the number of s_0 (vertical) moves to reach a state $(\tilde{n}_1, \tilde{n}_0)$ from some given state is always the same, regardless of the path. The same holds for s_1 (horizontal) moves. This means that to calculate the minimal cost to reach a given state $(\tilde{n}_1, \tilde{n}_0)$ from initial state $(\hat{n}_1, \hat{n}_0)$, we must choose the $(\hat{n}_0 - \tilde{n}_0)$ least-cost s_0 swaps and the $(\tilde{n}_1 - \hat{n}_1)$ least-cost s_1 swaps. This is, in fact, closely related to the theorem in (Kleinberg et al. 2018) reinterpreted in our special case. The resulting loss is then the sum of the costs of the s_0 swaps and the s_1 swaps.

It is therefore advantageous to pre-compute the swapping costs, $\Delta\ell_i'(s_0)$ and $\Delta\ell_i'(s_1)$ (see Equation (11)), and sort the resulting lists by increasing loss (smallest loss indexed first). After having sorted the two lists by increasing cost, the optimal sequence of swaps is obtained by applying the following procedure:

1. Compute the cumulated loss for both types of swaps from the state predicted by the classifier, $(\hat{n}_1, \hat{n}_0)$. That is, for s_0, $c\ell(\tilde{n}_0) = \sum_{n_0=\hat{n}_0-1}^{\tilde{n}_0} \Delta\ell'_{\hat{n}_0-n_0}(s_0)$ (by steps of -1), and, for s_1, $c\ell(\tilde{n}_1) = \sum_{n_1=\hat{n}_1+1}^{\tilde{n}_1} \Delta\ell'_{\hat{n}_1-n_1}(s_1)$ (by steps of $+1$). Then, the cumulated loss for an ending state $(\tilde{n}_1, \tilde{n}_0)$ is $c\ell(\tilde{n}_1, \tilde{n}_0) = c\ell(\tilde{n}_1) + c\ell(\tilde{n}_0)$.

2. Identify all fair candidates just below the fairness frontier (big gray points in Figure 3) and compute their cumulated cost from the starting state $(\hat{n}_1, \hat{n}_0)$ (big red square in Figure 3). To do so, enumerate all the potential end state $\tilde{n}_1$ coordinates going from $\hat{n}_1$ to $N(Z=1)$ and, for each of these coordinates, compute the corresponding $\tilde{n}_0$ on the frontier from Equation (13) and DI $\geq 1 - \varepsilon$,

$$\tilde{n}_0 = \left\lfloor \frac{1}{(1-\varepsilon)} \frac{N(Z=0)}{N(Z=1)} \tilde{n}_1 \right\rfloor \text{ for } \tilde{n}_1 = \hat{n}_1 + 1 \text{ to } N(Z=1) \tag{14}$$

   where $\lfloor x \rfloor$ returns the closest integer lesser or equal to x. Then, retrieve the cumulated cost of $(\tilde{n}_1, \tilde{n}_0)$ from point 1.

3. Select the fair candidate with the lowest cumulated cost (corresponding to the ending state $(\tilde{n}_1, \tilde{n}_0)$, which is the yellow star in Figure 3).

4. Identify the lowest-cost swaps to reach the ending state $(\tilde{n}_1, \tilde{n}_0)$.

This results in a computational complexity dominated by the fact that the data have to be sorted, the other operations being linear.

Alternatively, instead of selecting the least cumulated loss state in step (3), the procedure could return the list of candidate states $(\tilde{n}_1, \tilde{n}_0)$ together with their corresponding cumulated loss and DP, allowing the analyst to choose the most suitable solution (balancing divergence from the original classifier and achieved fairness measure). It would also allows us to plot the Pareto frontiers.

Beyond DP and DI, other fairness definitions could be integrated by modifying the fairness frontier or by adding supplementary constraints. The grid-based formulation should make it easy to identify the feasible points that satisfy such constraints, which will be investigated in the next subsection.

Additional Constraints In the following, we illustrate the flexibility of the OS method by introducing an additional constraint on the total number of instances classified in the preferred (positive) class ($\tilde{y}_{i1} = 1$). The additional constraint states that the total number of selected instances must be less or equal to a given threshold ($\tilde{n}_{\max}$), such that $\tilde{n}_0 + \tilde{n}_1 \leq \tilde{n}_{\max}$.

Therefore, only potential states satisfying this constraint are eligible in the above-mentioned procedure (point 3 of

the OS procedure). In Figure 4, we represent this additional constraint by a pink straight line, corresponding to $\tilde{n}_{\max}$, the maximum number of instances classified in the positive class. The fair candidates (big gray points in Figure 4) are those who satisfy both the fairness constraint (just below the fairness frontier) and the maximum number of selected instances (below the line $\tilde{n}_{\max}$). Then, the algorithm must select the fair candidate with the lowest cumulative cost (yellow star in Figure 4). Other linear constraints could also be incorporated into the grid, depending on the study case.

Experiments

In this section, we describe the experimental methodology we used to evaluate our fairness-aware post-processing methods and discuss the results of the experiments.

Experimental Methodology

For the experimental comparisons, we run a standard five-fold cross-validation. This subsection contains the experimental details, including: the datasets used, the classifiers applied before the post-processing phase, the post-processing methods from the literature that we compare to ours, the measures used to evaluate the performance of the methods, and the parameters selection phase.

Datasets We used five datasets described in (Le Quy et al. 2022): Adult, Bank marketing, COMPAS, German credit, and Law school. Although the datasets Adult, COMPAS, and German credit are commonly used in fairness-aware machine learning research (Delobelle et al. 2021; Fish, Kun, and Lelkes 2016; Krasanakis et al. 2018), they have recently been criticized for their limitations (Fabris et al. 2022). To ensure comparability with existing research, we still decided to incorporate these datasets, along with two additional datasets also recognized in the literature (Bank marketing and Law school) (Le Quy et al. 2022) to provide a broader evaluation. Each dataset consists of instances with a set of feature variables X_f, a binary sensitive variable Z, and a binary target variable Y.

Classifiers Post-processing methods may be applied to any classifier that outputs a prediction matrix $\hat{\mathbf{Y}}$. To assess the robustness and generalization of the post-processing methods across various classification scenarios, we selected five different classifiers representing a range of modeling approaches and complexities: logistic regression, decision tree, random forest, bagging, and boosting.

Comparison to Post-Processing Methods from Literature We compare our fairness-aware post-processing methods with the 'massaging' method of (Kamiran and Calders 2009), and the 'reject option based classification' method of (Kamiran, Karim, and Zhang 2012), both designed for binary classification problems. We chose these benchmark methods to ensure fair comparisons with our methods, as they operate in the same setting where true target values Y are not used at test time. In contrast, many recent studies – such as those based on calibration and equalized odds – require access to Y (see Related Work for more details about it), which makes direct comparison less meaningful in real-world post-deployment scenarios. Furthermore, since our focus is on group fairness, we excluded methods that incorporate individual fairness, ensuring a consistent and meaningful evaluation.

Massaging The first method that we compare to ours is the 'massaging' method (shortened later as MASS) proposed by Kamiran and Calders (Kamiran and Calders 2009). This approach is initially a pre-processing technique consisting in modifying the target labels $\mathbf{Y}^{\mathrm{d}}$ in the initial dataset. The method identifies two groups in the training data: (1) instances i disadvantaged, $z_i = 1$, initially assigned to the negative class, $y^{\mathrm{d}}_{i1} = 0$; and (2) instances i advantaged, $z_i = 0$, initially assigned to the positive class, $y^{\mathrm{d}}_{i1} = 1$. Based on a ranking function to assess probabilities for target class membership, instances in each group are ordered and the M[4] first instances of each group have their target labels modified. The value of M is the same for each group to maintain the same distribution between target labels. The dataset with the modified target labels is then used by the classifier to be trained.

With the intuition of the swapping mechanism, the concept of massaging can also be adapted as a post-processing method. In this variant, the post-processing MASS method is applied to the classifier's predictions $\hat{\mathbf{Y}}^{\mathrm{d}}$ instead of applying it to the target label of the dataset $\mathbf{Y}^{\mathrm{d}}$. The MASS post-processing method consists of ordering the s_0 and s_1 swaps according to their increasing costs and swapping the M least expensive s_0 and the M least expensive s_1.

The MASS method is closely related to our OS method but differs in some key aspects. First, the number of swaps in MASS is fixed, and it enforces an equal number of s_0 and s_1 swaps. This constraint simplifies the optimization problem, making it computationally efficient but limiting its applicability to specific scenarios where such a balance is appropriate. Second, MASS adopts a greedy strategy for selecting swaps, similar to the GS method. Unlike the MASS method, we proposed a non-greedy, optimal variant capable of exploring a broader range of solutions, not restricted to a fixed number of instances classified on each class.

Reject option based classification The second post-processing method that we compare to ours is the 'reject option based classification' (shortened later as ROC), developed by Kamiran et al. (Kamiran, Karim, and Zhang 2012), also for supervised binary classification problems. It consists of building a critical region around the decision boundary, regarding the classifier's probabilistic predictions. By fixing the decision boundary at 0.5, which is the case most often, the critical region corresponds to $[0.5 - \theta, 0.5 + \theta]$. All protected instances i (i.e., where $z_i = 1$) that have a probability of belonging to the positive class in this critical region ($\hat{y}_{i1} \in [0.5 - \theta, 0.5 + \theta]$) are classified in the positive class ($\tilde{y}^{\mathrm{d}}_{i1} = 1$), while all unprotected instances i (i.e., where $z_i = 1$) that have a probability of belonging to the positive class in this critical region ($\hat{y}_{i1} \in [0.5 - \theta, 0.5 + \theta]$) are classified in the negative class ($\tilde{y}^{\mathrm{d}}_{i1} = 0$).

[4] See in (Kamiran and Calders 2009) how the value of M is computed.

Paralleling our GS method, the ROC method realizes s_0 and s_1 swaps for all instances predicted in the critical region. The size of the region depends on the margin parameter θ, which is chosen to ensure the results satisfy the specified fairness conditions. To make comparisons possible with our swapping methods, θ is determined by trying to enforce DI $= 1 - \varepsilon$, which corresponds to our fairness frontier in our swapping methods[5]. Therefore, the ROC method differs from our swapping methods in its approach to fairness adjustments. While our methods focus on targeted swaps between protected and unprotected instances, ROC modifies predictions within a predefined critical region around the decision boundary.

Measures To evaluate the performance of the fairness-aware post-processing methods studied, we based our comparisons on three measures: the accuracy (ACC), which refers to the percentage of correctly classified instances on the test data; the demographic parity (DP), as defined by Equation (1); and the F1 demographic parity (F1DP), as defined by Equation (4).

Parameters γ and ε We must tune the parameter γ for COV and OS methods, balancing L_1 ($\gamma = 1$) and L_2 ($\gamma = 0$) norms. Therefore, we have tested three options for each: $\gamma = 1$ (L_1), $\gamma = 0.5$ (L_1L_2), and $\gamma = 0$ (L_2).

We also need to choose the parameter ε, quantifying the relaxation in the fairness constraints. We preliminarily tested several variations of ε, demonstrating that all problems can be solved with $\varepsilon = 0$ in our methods, which corresponds to perfect fairness as defined by Equations (1) and (2). Therefore, we apply $\varepsilon = 0$ for COV, GS, and OS. To ensure fair comparisons with ROC, which also requires fixing ε, we apply $\varepsilon = 0$ for ROC.

Results

Table 2 contains the results of our experiments for each dataset. Each number in the table corresponds to the mean of the measure across the five folds and the five classifiers for the specified post-processing method. For clarity, we highlight the highest ACC and F1DP and the closest to zero DP in bold for each dataset. In addition to the per-dataset evaluation presented in Table 2, we also examined the results separately for each classifier. The trends remained consistent across all classifiers, with no major shifts in the relative performance of the methods.

It can be observed that our OS post-processing method always offers the best results in terms of F1DP and DP, regardless of the gamma value (i.e., varying between the L_1 and L_2 norms). Note that in OS, the fairness frontier is continuous, while the reachable states are discrete (on a grid); therefore, even with $\varepsilon = 0$, the closest feasible state may still result in a DP slightly different from zero. With the best results on F1DP, the OS method successfully improves fairness while maintaining high accuracy on the investigated datasets.

However, regarding accuracy alone, the ROC method demonstrates the highest performance on all datasets. Despite this, it exhibits a higher level of DP than our post-processing methods, resulting in a lower F1DP score, indicating that the ROC method is less effective at improving fairness. The higher level of DP for ROC may be explained by the fact of imposing $\varepsilon = 0$, which can lead to overly restrictive adjustments. To investigate this possibility further, we analyzed the sign of DP (considered in absolute value in the results) and observed evidence of reverse discrimination caused by switching too many instances between groups. To ensure that the choice of ε does not affect the results, we conducted additional comparisons between our OS method and the ROC method across several ε values. The results consistently showed higher DP and F1DP scores for OS, regardless of the ε value tested. This highlights the robustness of the method across different fairness-accuracy trade-offs.

On the contrary, the COV method appears to be much less effective in DP reduction[6] because the fairness constraints are imposed on the soft DP (based on the probabilistic predictions instead of the decisions). While COV is less effective in reducing DP, it outputs revised probabilities, which are valuable for tasks requiring soft predictions (e.g., risk scoring or threshold tuning). Moreover, it can also be used with numerical protected variables such as, e.g., the age, which makes it useful in some situations.

To complete this first analysis, we performed Friedman-Nemenyi statistical tests and Wilcoxon signed-rank tests to compare the post-processing methods (Demšar 2006). We limit our tests to the F1DP measure, providing a trade-off between accuracy and fairness. To compare a single version of each post-processing method (COV, OS, GS, MASS, and ROC), we selected the L_1L_2 version ($\gamma = 0.5$) of COV and OS methods, interpolating between L_1 and L_2 norms for the loss function. This choice was motivated by our first results (see Table 2) which indicated that the performance differences among L_1, L_1L_2, and L_2 versions were minimal. We computed a global comparison of the methods, resulting in 125 comparison points for the statistical tests (5 datasets $\times$ 5 folds $\times$ 5 classifiers). The p-value of the Friedman test is 3.3×10^{-67}, which is highly significant. This means that at least one method is significantly different from the others. Based on that, we performed a Nemenyi test to compare the average ranks of the methods. The Nemenyi test is similar to an ANOVA test, but it compares all methods to each other and ranks them (Demšar 2006). Figure 5 represents the mean rank of each method with 95% Nemenyi confidence intervals. The higher the rank, the better the method in terms of F1DP, and methods with non-overlapping critical distance intervals are considered significantly different. In our analysis, the OS method, highlighted in blue, achieved the best average rank and was significantly better than all other post-processing methods, highlighted in red. However, it can also be observed that the MASS and GS methods still

[5]Code for the ROC method and determination of θ based on a fairness measure is available at https://github.com/Trusted-AI/AIF360/blob/main/aif360/algorithms/postprocessing/reject_option_classification.py.

[6]Note however that additional results (not reported here) show that COV is quite effective in reducing the soft DP based on the probabilistic predictions because its fairness constraint is based on this quantity.

Measure → **Method ↓**	ACC	F1DP	DP	ACC	F1DP	DP	ACC	F1DP	DP	ACC	F1DP	DP	ACC	F1DP	DP
	Adult			**Bank marketing**			**COMPAS**			**German credit**			**Law school**		
COV L_1	0.8514	0.8965	0.0532	0.8988	0.9299	0.0365	0.6670	0.7780	0.0662	0.7322	0.8232	0.0578	0.9496	0.9729	0.0025
COV L_1L_2	0.8512	0.9012	0.0423	0.8999	0.9366	0.0235	0.6665	0.7788	0.0631	0.7350	0.8289	0.0476	0.9499	0.9716	0.0057
COV L_2	0.8500	0.9062	0.0295	0.9002	0.9343	0.0288	0.6692	0.7788	0.0682	0.7338	0.8279	0.0479	0.9504	0.9665	0.0167
OS L_1	0.8474	**0.9174**	**0.0000**	0.8997	0.9471	0.0001	0.6631	**0.7973**	**0.0002**	0.7320	0.8397	0.0140	0.9492	**0.9739**	0.0002
OS L_1L_2	0.8473	0.9173	**0.0000**	0.8997	0.9471	0.0001	0.6629	0.7972	**0.0002**	0.7316	0.8396	0.0137	0.9492	**0.9739**	0.0002
OS L_2	0.8465	0.9169	**0.0000**	0.9002	**0.9474**	**0.0000**	0.6624	0.7968	**0.0002**	0.7326	**0.8405**	**0.0128**	0.9485	0.9735	**0.0001**
GS	0.8460	0.9166	0.0001	0.8963	0.9449	0.0010	0.6635	**0.7973**	0.0010	0.7284	0.8365	0.0161	0.9465	0.9723	0.0004
MASS	0.8464	0.9167	0.0002	0.8999	0.9468	0.0012	0.6629	0.7967	0.0017	0.7292	0.8360	0.0190	0.9474	0.9726	0.0008
ROC	**0.8516**	0.8901	0.0675	**0.9015**	0.8815	0.1365	**0.6777**	0.7189	0.2341	**0.7364**	0.8133	0.0872	**0.9511**	0.9528	0.0453

Table 2: Mean of ACC, F1DP, and DP for the various fairness-aware post-processing methods obtained on the five datasets. The best results for each dataset and measure are highlighted in bold.

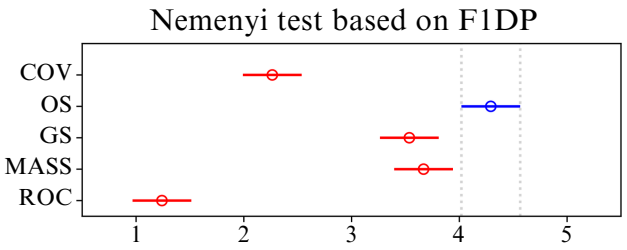

Figure 5: Mean ranks and 95% Nemenyi confidence intervals on the F1DP measure. The x-axis shows the mean rank of the methods. The higher the rank, the better the method in terms of F1DP. The best method, highlighted in blue, is the OS method, whose results are significantly different from all other methods.

	COV	OS	GS	MASS	ROC
COV	\	***	***	***	***
OS	***	\	***	***	***
GS	***	***	\	n.s.	***
MASS	***	***	n.s.	\	***
ROC	***	***	***	***	\

Table 3: Significance level of Wilcoxon signed-rank tests to compare the fairness-aware post-processing methods on F1DP. Significance levels: *** ($p < 0.001$) and n.s. = not significant ($p \geq 0.05$).

obtain close results.

We rely on Wilcoxon signed-rank tests to further support these first statistical tests to measure the pairwise difference between the fairness-aware post-processing methods (see Table 3). With a level of confidence of 99.9% ($\alpha = 0.001$), we can see on Table 3 that the OS method is significantly better, in terms of F1DP, than the other fairness-aware post-processing methods on the investigated datasets – all *p*-values for this method are significant.

To go deeper into the analyses, we performed the same statistical tests for each dataset separately, which showed the same pattern observed in general, though with lower significance levels for some datasets. Further studies could investigate the dataset characteristics that would justify these observations.

Conclusion and Future Work

In this paper, we proposed and studied two fairness-aware post-processing methods, *least L_1/L_2 norm with covariance constraints* and *optimal swapping*, to address group fairness in supervised classification. Both methods are designed to balance fairness (in terms of DP or DI) and accuracy while remaining applicable to any classifier that computes class membership probabilities. Experimental validation on five datasets and five classifiers, including comparisons with related methods from the literature, demonstrated that the optimal swapping consistently achieves an effective trade-off between demographic parity and accuracy, as measured by the F1 demographic parity. The simplicity of the optimal swapping method, in terms of ease of implementation, and its flexibility to incorporate additional linear constraints make it practical and adaptable to diverse applications.

While the results are promising, several avenues for future research remain. First, extending our methods to handle multi-class classification problems with several ordered, preferred categories would significantly enhance their applicability. We could also extend our methods to support sensitive attributes with more than two categories, or even multiple sensitive variables simultaneously. Additionally, future work may explore incorporating alternative fairness metrics beyond DP and DI, for instance conditional DP, in our methods by modifying the constraint formulation in COV or redefining the frontier in OS to reflect the target fairness metric. Moreover, our methods could be further compared to other types of post-processing methods, more recent, such as those based on calibration or equalized odds, which require the true target values Y, and to optimal transport approaches.

Acknowledgements

We thank Alexia Kneip, Eve Beghein and François Gouverneur for working on preliminary investigations with their master's thesis (Kneip and Beghein 2022; Gouverneur 2023). We also thank the reviewers for their relevant remarks allowing to improve the manuscript.

References

Alves, G.; Bernier, F.; Couceiro, M.; Makhlouf, K.; Palamidessi, C.; and Zhioua, S. 2023. Survey on fairness notions and related tensions. *EURO journal on decision processes*, 11: 100033.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1): 3–44.

Bishop, C. 2006. *Pattern recognition and machine learning*. Springer.

Castelnovo, A.; Crupi, R.; Greco, G.; Regoli, D.; Penco, I. G.; and Cosentini, A. C. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1): 1–21.

Caton, S.; and Haas, C. 2024. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7): 1–38.

Chen, P.; Wu, L.; and Wang, L. 2023. AI fairness in data management and analytics: A review on challenges, methodologies and applications. *Applied Sciences*, 13(18): 10258.

Delobelle, P.; Temple, P.; Perrouin, G.; Frénay, B.; Heymans, P.; and Berendt, B. 2021. Ethical adversaries: Towards mitigating unfairness with adversarial machine learning. *ACM SIGKDD Explorations Newsletter*, 23(1): 32–41.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7: 1–30.

Diamond, S.; and Boyd, S. 2016. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83): 1–5.

Duda, R. O.; Hart, P. E.; and Stork, D. G. 2001. *Pattern classification*. Wiley, 2nd edition.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*, 214–226.

Elkan, C. 2001. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI '2001)*, volume 17, 973–978. Lawrence Erlbaum Associates Ltd.

Fabris, A.; Messina, S.; Silvello, G.; and Susto, G. A. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6): 2074–2152.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (KKD '15)*, 259–268.

Ferrara, E. 2023. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1): 3.

Fish, B.; Kun, J.; and Lelkes, A. D. 2016. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 16th SIAM International Conference on Data Mining (SDM '16)*, 144–152.

Golub, G. H.; and Van Loan, C. F. 1996. *Matrix computations*. Johns Hopkins University Press, 3rd edition.

Gordaliza, P.; Del Barrio, E.; Fabrice, G.; and Loubes, J.-M. 2019. Obtaining fairness using optimal transport theory. In *Proceedings of the 36th International Conference on Machine Learning (ICML '19)*, 2357–2365. PMLR.

Gouverneur, F. 2023. *Fairness in machine learning : focus on post-processing methods.* Master's thesis, Université catholique de Louvain. Supervisor: Saerens, Marco.

Hansen, P.; Pereyra, V.; and Scherer, G. 2013. *Least squares data fitting with applications*. John Hopkins University Press.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 29th Neural Information Processing Systems Conference (NIPS '16)*.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2009. *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Hort, M.; Chen, Z.; Zhang, J. M.; Harman, M.; and Sarro, F. 2024. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 1(2): 1–52.

Jang, T.; Shi, P.; and Wang, X. 2022. Group-aware threshold adaptation for fair classification. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI '22)*, volume 36, 6988–6995.

Jui, T. D.; and Rivas, P. 2024. Fairness issues, current approaches, and challenges in machine learning models. *International Journal of Machine Learning and Cybernetics*, 1–31.

Kamiran, F.; and Calders, T. 2009. Classifying without discriminating. In *Proceedings of the 2nd international conference on computer, control and communication (ICCCC '09)*, 1–6. IEEE.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1): 1–33.

Kamiran, F.; Karim, A.; and Zhang, X. 2012. Decision theory for discrimination-aware classification. In *Proceedings of the IEEE 12th international conference on data mining (ICDM '12)*, 924–929. IEEE.

Kim, J.-Y.; and Cho, S.-B. 2022. An information theoretic approach to reducing algorithmic bias for machine learning. *Neurocomputing*, 500: 26–38.

Kleinberg, J.; Ludwig, J.; Mullainathan, S.; and Rambachan, A. 2018. Algorithmic fairness. In *Proceedings of 113th Annual Meeting of the American Economic Association (AEA papers and proceedings '18)*, volume 108, 22–27.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kneip, A.; and Beghein, E. 2022. *Fairness in supervised classification: investigation of three different techniques.* Master's thesis, Louvain School of Management, Université catholique de Louvain. Supervisor: Saerens, Marco.

Krasanakis, E.; Spyromitros-Xioufis, E.; Papadopoulos, S.; and Kompatsiaris, Y. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 28th International Conference on World Wide Web (WWW '18)*, 853–862.

Lazar Reich, C.; and Vijaykumar, S. 2021. A Possibility in Algorithmic Fairness: Can Calibration and Equal Error Rates Be Reconciled? In *2nd Symposium on Foundations of Responsible Computing (FORC 2021)*, 4–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Le Quy, T.; Roy, A.; Iosifidis, V.; Zhang, W.; and Ntoutsi, E. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3): 1–59.

Lohia, P. K.; Ramamurthy, K. N.; Bhide, M.; Saha, D.; Varshney, K. R.; and Puri, R. 2019. Bias mitigation post-processing for individual and group fairness. In *Proceedings of the 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '19)*, 2847–2851. IEEE.

Mardia, K. V.; Kent, J. T.; and Bibby, J. M. 1979. *Multivariate analysis*. Academic Press.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6): 1–35.

Mishler, A.; Kennedy, E. H.; and Chouldechova, A. 2021. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 4th ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 386–400.

Mitchell, S.; Potash, E.; Barocas, S.; D'Amour, A.; and Lum, K. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8: 141–163.

Murphy, K. 2022. *Probabilistic machine learning: an introduction*. MIT Press.

Noriega-Campero, A.; Bakker, M. A.; Garcia-Bulle, B.; and Pentland, A. 2019. Active fairness in algorithmic decision making. In *Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*, 77–83.

Pessach, D.; and Shmueli, E. 2022. A review on fairness in machine learning. *ACM Computing Surveys*, 55(3): 1–44.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On fairness and calibration. *Advances in neural information processing systems*, 30.

Romei, A.; and Ruggieri, S. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5): 582–638.

Santafe, G.; Inza, I.; and Lozano, J. A. 2015. Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review*, 44: 467–508.

Sen, P. C.; Hajra, M.; and Ghosh, M. 2020. Supervised classification algorithms in machine learning: A survey and review. In *Proceedings of the 1st International Conference on Emerging Technology in Modelling and Graphics (IEM-Graph '18)*, 99–111. Springer.

Small, E.; Sokol, K.; Manning, D.; Salim, F. D.; and Chan, J. 2024. Equalised odds is not equal individual odds: Post-processing for group and individual fairness. In *Proceedings of the 7th ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 1559–1578.

Vancompernolle Vromman, F.; Courtain, S.; Leleux, P.; de Schaetzen, C.; Beghein, E.; Kneip, A.; and Saerens, M. 2024. Maximum Entropy Logistic Regression for Demographic Parity in Supervised Classification. In *Proceedings of the 35th Benelux Conference on Artificial Intelligence and Machine Learning (BNAIC/Benelearn '23)*, 189–208. Springer. ISBN 978-3-031-74650-5.

Xian, R.; Yin, L.; and Zhao, H. 2023. Fair and optimal classification via post-processing. In *Proceedings of the 40th International Conference on Machine Learning (ICML '23)*, 37977–38012. PMLR.