Happiness as a Measure of Fairness

Georg Pichler
Institute of Telecommunications
TU Wien

Marco Romanelli
Computer Science Dept.
Hofstra University

Pablo Piantanida
ILLS & Mila - Quebec AI Institute
CNRS, CentraleSupélec - Univ. Paris-Saclay

Abstract

In this paper, we propose a novel fairness framework grounded in the concept of *happiness*, a measure of the utility each group gains from decision outcomes. By capturing fairness through this intuitive lens, we not only offer a more human-centered approach, but also one that is mathematically rigorous: In order to compute the optimal, fair post-processing strategy, only a linear program needs to be solved. This makes our method both efficient and scalable with existing optimization tools. Furthermore, it unifies and extends several well-known fairness definitions, and our empirical results highlight its practical strengths across diverse scenarios.

1 INTRODUCTION

In classification problems where the considered dataset can be naturally divided into several groups, such as by a specific attribute, it is often desirable for the classifier to treat each group "equally". However, unfair results are often observed and may arise from a variety of sources, such as imbalances in the training set, biases introduced during the learning process, or existing biases in the training data itself. These unfair outcomes typically mean that groups are not treated equally, receiving systematically better or worse outcomes compared to others.

Many techniques have been developed to enhance fairness in Machine Learning (ML) systems, e.g., (Hardt et al., 2016; Jiang et al., 2022; Agarwal et al., 2018; Caton and Haas, 2024; Gohar and Cheng, 2023). These methods can be applied at different stages of the ML pipeline: pre-processing techniques adjust the training data before learning begins, in-processing methods guide the training to promote fair outcomes, and post-processing approaches modify the model's outputs to reduce unfairness after training is complete. In the latter case, the output of a soft-classifier is post-processed, ensuring fair classification in the process, typically at the expense of the overall accuracy.

Popular post-processing techniques aim to equalize statistical measures involving only labels and group membership Berk et al. (2021); Bharti et al. (2023); Hardt et al. (2016); Jiang et al. (2022); Tang and Zhang (2022). In doing so, fairness is often framed through the lens of group-wise performance metrics (e.g., accuracy, false positive rates, or precision) valuated separately for each demographic. When these metrics are aligned across groups, the model is considered fair. However, this performance-centric perspective, while intuitive, can obscure deeper forms of unfairness particularly when the training data already embeds historical or structural biases. By focusing solely on output parity, these approaches risk masking disparities in how different groups experience the outcomes, leading to fairness definitions that are technically satisfied but practically insufficient. Our work challenges this narrow perspective and proposes an alternative one that captures the utility derived by each group, offering a richer and potentially more just notion of fairness.

To illustrate the limitations of current fairness metrics, consider a ML system designed to determine whether to grant a loan to an individual. The input X includes features such as the borrower's profile and the amount of credit requested, and the output Y is a binary decision indicating loan approval or rejection. Now suppose we have two demographic groups, Group 0 (G_0) and Group 1 (G_1), that are identical in all features except that, on average, individuals from G_0 request twice as much credit as those from G_1. A classifier that approves loans at the same rate for both groups, thus appearing fair under standard group-based metrics, would in fact result in an unequal allocation of

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

resources: group G_1 would receive significantly less total credit than group G_0, despite the groups being otherwise indistinguishable. This outcome is intuitively unfair, but standard fairness definitions based on performance metrics (e.g., Equalized Odds or demographic parity) would fail to capture it, since they only consider prediction correctness or distribution, not the magnitude or utility of the outcome. Moreover, if this imbalance is already embedded in the training data, the classifier may simply learn to replicate it, and fairness constraints applied post hoc will not rectify the core issue.

1.1 Contributions

- We introduce the concept of **happiness** as a tool reflecting how satisfied an individual is with the output of a classifier. Fairness can then be measured by how evenly happiness is allocated between groups–where perfect equality in allocation corresponds to equal happiness across groups (cf. Section 2), e.g. the loan amount in the example above.
- We apply our definition to find the optimal post-processing strategy, resulting in a linear program which can be solved efficiently (cf. Section 3).
- Importantly, we show how this definition encompasses other popular definitions of fairness as special cases (cf. Section 4).
- We provide numerical experiments showcasing the utility of this approach (cf. Section 5). The code to reproduce our results is available at `https://github.com/g-pichler/HappinessAsAMeasureOfFairness`.

1.2 Related Work

The problem of data bias and its impact on the fairness of ML models is well documented in the literature Barocas et al. (2023); Cerrato et al. (2024). Over the years, alongside the development of dedicated datasets for fairness evaluation Le Quy et al. (2022), several notions of fairness have been proposed, as well as algorithms to mitigate bias in ML models.

Fairness notions in Machine Learning

Different definitions of fairness, often based on specific metrics, have been proposed in the literature. Among the most prominent are *Overall Accuracy Equality* Berk et al. (2021), *Equalized Odds* Hardt et al. (2016); Bharti et al. (2023); Tang and Zhang (2022), *Equal Opportunity* Hardt et al. (2016), *Demographic Parity* Jiang et al. (2022), and *Statistical Parity* Dwork et al. (2012), to name a few. These fairness criteria guide prediction correction by balancing the distribution of predicted labels with respect to the sensitive attributes which characterize different groups Verma and Rubin (2018).

A variety of extensions have been proposed to generalize these definitions to multi-class problems Alghamdi et al. (2022); Liu et al. (2023); Rouzot et al. (2023), multi-group settings Dwork et al. (2023), and even regression tasks Taturyan et al. (2024). Other lines of work have explored entirely different directions, such as leveraging the theory of calibration Pleiss et al. (2017), applying tools from optimal transport Gordaliza et al. (2019); Silvia et al. (2020); Buyl and De Bie (2022); Wang et al. (2023), or using cryptographic primitives Yadav et al. (2024). Perhaps more aligned with the work presented in this paper, Agarwal et al. (2018); Liu et al. (2023); Woodworth et al. (2017); Liu et al. (2019); Kim et al. (2018); Perdomo et al. (2020); Zafar et al. (2019) offer various attempts at establishing a unified framework to obtain and or evaluate fairness. However, a key distinction lies in the fact that none of these approaches are equipped to incorporate a general notion of happiness, which introduced in this work. As a result, they are not well-suited for settings where capturing individual and then by extension group-level utility is essential.

Algorithms to mitigate bias in ML models

Three main categories of approaches have been proposed to mitigate bias in ML models: (i) pre-processing approaches, which aim to modify the training data to reduce or eliminate bias; (ii) in-processing approaches, which seek to adjust the learning algorithm itself to address biases caused by dominant features or other distributional effects; and (iii) post-processing approaches, which apply transformations to the model's output to enhance fairness in predictions (Gohar and Cheng, 2023; Mehrabi et al., 2021; Fabris et al., 2022; Caton and Haas, 2024).

Post-processing methods are popular in the literature due to their *flexibility*: they do not explicitly modify the underlying model and therefore do not require access to the training algorithms or the models themselves. They are also valued for their *lightweight* nature, as they only require access to the model's predictions and sensitive attribute information. Furthermore, their *ease of use* makes them appealing in scenarios where modifying the data or the model may have legal implications or compromise their interpretability.

More closely aligned with our framework are the works of Liu et al. (2018); Weber et al. (2022), which in-

troduce the notion of delayed impact and long-term fairness effects. These approaches are orthogonal to ours, as they examine how conventional fairness constraints on decision policies affect long-term outcomes. To this end, repeated classification is considered where the classifier optimizes a utility function tied to institutional gain, such as loan repayment probability for a bank or loan office. In these settings, fairness impacts emerge through changes in external variables rather than the label itself. A complementary perspective is offered in Kasy and Abebe (2021), where fairness constraints are modeled as institutional costs, particularly when treatment of a subpopulation is boosted to ensure merit-based recognition. This can introduce new disparities, whose impact can be characterized by influence functions, which are not captured by standard fairness metrics. These works require a definition quantifying how well a group performs at each point in time. Commonly, average credit score, when dealing with loan approval, or educational success metrics, when using admissions data, are used. This is akin to the definition of "happiness" introduced in this work. While our work does not include temporal modeling, in essence we propose to directly use the relevant metric when defining the fairness of a classifier. This is also mentioned in (Liu et al., 2018, Sec. 4.3), but not expanded upon.

Thus, our framework introduces a general class of fairness constraints that subsume common definitions, while even enabling the optimization of long-term effects, within a setting where classification accuracy remains the core objective, relevant to both model developers and end users.

2 HAPPINESS AS A CRITERION FOR EVALUATING FAIRNESS

While the scenario discussed in Section 1, involving two identical groups where unfair behavior arises from biased training data, is intentionally contrived, it showcases the need for a framework to address the issue of disparate resource allocation. In this paper, we focus on a post-processing strategy that can be applied to any soft classifier, resulting in a fair classification. Our method naturally allows for a trade-off between accuracy and fairness. In the following, we will provide a brief outline of our method and its advantages.

Owing to the fact that performance metrics alone can be insufficient for determining fairness, for the first time, we take a more general approach and define a *happiness function* $\boldsymbol{\eta}$, which, in general, takes all features, the group index characterized by the sensitive feature(s), the ground truth as well as the (hard) classifier output label as inputs. It produces a real number as output, which quantifies the happiness of an individual with the classifier output. If – on average – the happiness of individuals in the two groups are close, a classifier is considered fair. Note that the happiness can really be an arbitrary function of all features, and $\boldsymbol{\eta}$ is only required to output a real number.

The choice of the happiness function must be tailored to the specific problem setting and the intended use of the classifier. Notably, the same classifier may require distinct happiness formulations depending on its application context. For example, consider a classifier trained to estimate (or quantize) individual income. The perceived utility (or *happiness*) associated with a given prediction would naturally differ depending on whether the system is employed for assessing credit risk or determining taxation. This highlights the importance of aligning the happiness function with the operational goals and stakeholder perspectives inherent to each use case.

2.1 Proposed Framework

We restrict ourselves to linear, group-dependent post-processing of soft classifiers. Thus, our post-processing step can be described by a conditional probability distribution on the finite label space for each group.

The resulting trade-off between classification accuracy and fairness naturally yields a linear optimization problem: the accuracy of the classification is a linear function of these conditional distributions. The expected happiness (conditioned on the group) is also a linear function of the conditional post-processing probability distributions. This holds for arbitrary happiness functions. Thus, finding the optimal post-processing rule is equivalent to solving a linear program. Our method is introduced in detail in Section 3.

It was already pointed out in Agarwal et al. (2018) that many criteria for fairness can be phrased as linear constraints. We can use this fact and recover these fairness notions using a happiness function if we allow vector-valued $\boldsymbol{\eta}$. This does not create any complications in the optimization procedure, as it merely introduces additional linear constraints. In Section 4 we formally show, how $\boldsymbol{\eta}$ has to be chosen to recover "Statistical parity", "Overall accuracy" and "Equalized odds" from our definitions, showcasing the flexibility and generality of the proposed framework.

A major advantage of defining *fairness* in terms of *equal happiness of all groups*, is that it can be *tailored towards the particular application* of the classification system. It also allows for *additional features beyond the classification result and ground-truth* to be used in the computation of happiness. In case of a system designed for credit approval, e.g., the size of the loan

can also be considered in addition to the binary decision of whether to approve the loan. Yet, our method retains the advantageous property that the computational problem is still a linear program.

Although our methods can be readily extended to accommodate more than two groups introducing additional constraints, such an extension is not pursued in the present work in the interest of brevity and clarity.

2.2 Illustrative Example

To motivate our method, we provide an example, comparing it to other methods of post-processing with different fairness metrics.

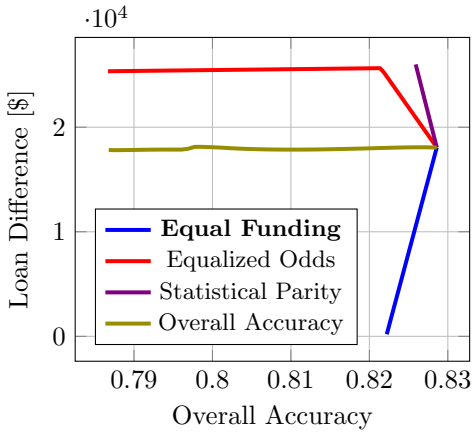


Figure 1: Our post-processing method (Equal Funding) guarantees any target accuracy level (up to 83.3%) while minimizing funding disparities between groups. Notably, it achieves perfect fairness, i.e., zero difference in loan allocations between G_0 and G_1 with less than a one percentage point loss in accuracy w.r.t. the baseline unfair classifier.

In this example, an ML system is tasked with approving (or denying) a loan. We use a completely synthetic dataset inspired by ADULT (Becker and Kohavi, 1996), divided in two groups, namely G_0 and G_1. All features are independently generated. The annual income is drawn from the Gaussian distribution (μ, σ^2) with mean $\mu_0 = \$50{,}000$ and standard deviation $\sigma_0 = \$1{,}000$. The *base* loan amount is an independently drawn Gaussian with standard deviation $\$10{,}000$, and mean $\mu_1 = \$500{,}000$. A loan is granted if the loan amount is less or equal than 10 times the annual income. Thus, a loan is granted with a probability of 0.5. However, G_0 requests the base loan amount, while for G_1, $\$50{,}000$ are added to the loan, while the probability of approval remains the same. This results in the allocation of additional funds to G_1 in the training data.

Subsequently, we train a random forest classifier on this linear classification problem, achieving an expected soft accuracy of 0.97 on the training set and 0.833 on the test set. Given any accuracy $\alpha \leq 0.833$, a post-processing strategy is applied to the classifier, that, while guaranteeing accuracy α maximizes fairness. E.g., in the case of Overall Accuracy Berk et al. (2021) this procedure minimizes the absolute difference between the accuracy on the two groups, while maximizing overall accuracy. For each value α we then compute the resulting difference in funding allocated to the two groups on average, when this post-processing step is applied. The plot in Figure 1 shows the difference in loan amount as a function of accuracy α.

Note that our method, dubbed "Equal funding" for this specific application, allows us to directly constrain the difference in funding while maintaining high accuracy. It is noteworthy that the other three methods, in this case, never achieve a meaningful reduction of the difference between the two groups. At best, in the case of Overall Accuracy, the difference remains unchanged, while optimization for Equalized Odds and Statistical Parity even increase the imbalance between the groups.

In the context of Figure 1, it is worth to point out, that the Equalized Odds predictor as defined in Hardt et al. (2016) corresponds to a single point : it is the left-most point of the "Equalized Odds" line, where equal odds for both groups correspond to about 0.795 accuracy and a funding gap of $25,384.

3 MAIN DEFINITIONS AND THEORETICAL RESULTS

We consider a standard classification problem, where $X \in \mathcal{X} = \mathbb{R}^d$ is the (random) feature vector, $Y \in \mathcal{Y}$ is the label in a finite space $\mathcal{Y}$, and in addition, we use $Z \in \{0, 1\}$ to denote the group characterized by the sensitive feature(s).

A **soft classifier** is a function of $\hat{Y}(X)$, taking the features X as input and producing a probability distribution on the labels $\mathcal{Y}$ as its output. We can interpret such a classifier as a random variable $\hat{Y}$ which depends on (Y, Z) only through X, i.e., $(Y, Z) - X - \hat{Y}$ form a Markov chain. This reflects the fact that in general, the classifier does not have access to Y and Z directly. We interpret $\hat{Y}$ as a (random) estimator, estimating Y from X. Performance of the estimator is judged by a **loss function**. For simplicity, we will use the probability of incorrect classification:

$$\ell(Y, \hat{Y}) := \mathbb{P}\{Y \neq \hat{Y}\} = \mathbb{E}[\mathbb{1}\{Y \neq \hat{Y}\}]. \qquad (1)$$

However, the results of this paper also hold for other loss functions.

Given an individual in group z with features x and true label y, we seek to quantify how happy they are with a classification results $\hat{y}$. This notion is captured by a **happiness** function $\boldsymbol{\eta}\colon \mathcal{Y}\times\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}\to\mathbb{R}^n$, where $n\in\mathbb{N}$, which gives the happiness score $\boldsymbol{\eta}(\hat{y},x,y,z)$. When $n>1$, happiness is simply measured by multiple scalar happiness functions simultaneously.

Given a trained estimator $\hat{Y}$, we want to perform demography-dependent post-processing, increasing the fairness of the estimator, while maintaining high accuracy. This can be achieve by another soft classifier $\tilde{Y}$, which is given $(\hat{Y},Z)$, the (random) output of the original estimator and the group index. The soft classifier $\tilde{Y}$ produces another probability distribution at its output. Thus, $\tilde{Y}$ can be interpreted as another random variable, which depends on all other random variables only through $\hat{Y}$ and Z, i.e., it is demography-dependent and we have the Markov chain $(X,Y)-(\hat{Y},Z)-\tilde{Y}$. This leads to the following factorization of the complete probability distribution:

$$\begin{aligned}p_{XYZ\hat{Y}\tilde{Y}}(x,y,z,\hat{y},\tilde{y}) &= p_{XYZ}(x,y,z)p_{\hat{Y}|XYZ}(\hat{y}|x,y,z)\\ &\quad\cdot p_{\tilde{Y}|\hat{Y}XYZ}(\tilde{y}|\hat{y},x,y,z) \qquad (2)\\ &= p_{XYZ}(x,y,z)p_{\hat{Y}|X}(\hat{y}|x)p_{\tilde{Y}|\hat{Y},Z}(\tilde{y}|\hat{y},z). \qquad (3)\end{aligned}$$

To complete the problem setup, we need to define fairness in terms of happiness and specify the trade-off between fairness and accuracy resulting from that definition. We say that a classifier is ε-fair if the average happiness of individuals in G_0 and G_1 are no more than ε apart. This results in a trade-off between fairness and accuracy, where we are interested in minimizing the loss $\ell(Y,\tilde{Y})$ among all ε-fair classifiers $\tilde{Y}$ for some fixed $\varepsilon>0$. This is formalized in the following definition.

Definition 1. *For a happiness function $\boldsymbol{\eta}$ and $\varepsilon\geq 0$, an estimator $\hat{Y}$ is ε-**fair** if*

$$\begin{aligned}\boldsymbol{\phi}(\boldsymbol{\eta},\hat{Y}) := \big|&\mathbb{E}[\boldsymbol{\eta}(\hat{Y},X,Y,Z)|Z=0]\\ &-\mathbb{E}[\boldsymbol{\eta}(\hat{Y},X,Y,Z)|Z=1]\big|\leq\varepsilon, \quad (4)\end{aligned}$$

where an inequality such as (4) is to be understood as component-wise[1] when relating vector-valued quantities.

*For a happiness function $\boldsymbol{\eta}$ and a soft classifier $\hat{Y}$, a pair (ε,L) is **achievable** if there exists a demographic-dependent post-processing, i.e., an estimator $\tilde{Y}$ satisfying (3), which is ε-fair and satisfies $\ell(Y,\tilde{Y})\leq L$.*

Fortunately, though $\boldsymbol{\eta}$ may depend on arbitrary features, the resulting problem is still a linear program.

[1]For a vector $\mathbf{a}=(a_1,a_2,\dots a_n)$ and a scalar b, we use the notation $\mathbf{a}\leq b$ for component-wise inequality $a_i\leq b$ for all $i\in\{1,\dots,n\}$.

In practical applications the happiness function should be chosen in such a way that it measures the satisfaction of an individual with the decision made by the model in the best possible way. Examples are included in Section 5.

Theorem 1. *For a fixed $\varepsilon\geq 0$ we can find the minimum L, such that (ε,L) is achievable by solving*

$$\min_{p_{\tilde{Y}|\hat{Y}Z}} \ell(Y,\tilde{Y}) \qquad (5)$$

$$s.t.\ \boldsymbol{\phi}(\boldsymbol{\eta},\tilde{Y})\leq\varepsilon. \qquad (6)$$

This is a linear programming problem.

The proof and further discussion of this result can be found in Appendix A.

In practice, however, the joint distribution of (X,Y,Z) will not be available. Thus, we need to replace the expectations by empirical averages. However, this will be reasonably accurate, even with modest training set sizes, as only the expected value of $2(n+1)|\mathcal{Y}|^2$ random variables needs to be approximated. Note that n is the dimension of the happiness function $\boldsymbol{\eta}$. Details are given in Appendix A.1, in particular Lemma 5.

In line with prior literature, we aim to optimize the trade-off between classification accuracy and equality, i.e., parity of our generalized fairness metric, between the two groups of interest. The constraint introduced in (4) and applied in (6) bounds the difference between the two expectations and consequently ensures (approximately) equal happiness across the two groups, simultaneously creating tension with the objective of minimal loss in (5). Notably, an alternative approach bounding both expected values individually from below, thus guaranteeing a minimum happiness level for both groups, is also possible. In fact, the resulting minimization problem remains a linear program and could be solved using the same techniques. However, this could lead to solutions where one group's happiness level is much higher than the other's, and, consequently, it is not pursued in this work.

4 RECOVERING OTHER FAIRNESS CRITERIA

In this section, we show that many previously established criteria of fairness can be recovered using the setup presented in Section 3. We will showcase this using "Statistical Parity", "Overall Accuracy" and "Equalized Odds". We take advantage of the fact that these definitions can be phrased as linear constraints, which was already observed in a different context in Agarwal et al. (2018).

The definitions are taken from Rouzot et al. (2023, Table II). While the original definitions ask for exact

equality, we will include $\varepsilon \geq 0$. Thus, the original definitions from Rouzot et al. (2023, Table II) can be recovered by setting $\varepsilon = 0$ in what follows.

Definition 2 (Statistical Parity)**.** *A classifier $\hat{Y}$ has* ε Statistical Parity *(or* demographic parity*) if*

$$\left|p_{\hat{Y}|Z}(\cdot|0) - p_{\hat{Y}|Z}(\cdot|1)\right| \leq \varepsilon. \tag{7}$$

Lemma 1. *A classifier $\hat{Y}$ has ε Statistical Parity if and only if it is ε-fair w.r.t. the $n = |\mathcal{Y}|$ dimensional happiness function $\boldsymbol{\eta}(y, \hat{y}, z) = (\mathbb{1}_{\tilde{y}}(\hat{y}))_{\tilde{y}\in\mathcal{Y}}$.*

Proof. Substituting the given happiness function $\boldsymbol{\eta}$ in (4), $\hat{Y}$ is ε-fair if for all $\tilde{y} \in \mathcal{Y}$ we have

$$\left|\mathbb{E}[\mathbb{1}_{\tilde{y}}(\hat{Y})|Z = 0] - \mathbb{E}[\mathbb{1}_{\tilde{y}}(\hat{Y})|Z = 1]\right| \leq \varepsilon, \tag{8}$$

which is equivalent to (7). □

Definition 3 (Overall Accuracy)**.** *A classifier $\hat{Y}$ has* ε equal Overall Accuracy *if*

$$\left|P(Y = \hat{Y}|Z = 0) - P(Y = \hat{Y}|Z = 1)\right| \leq \varepsilon. \tag{9}$$

Lemma 2. *A classifier $\hat{Y}$ has ε equal Overall Accuracy if and only if it is ε-fair w.r.t. the $n = 1$ dimensional happiness function $\boldsymbol{\eta}(y, \hat{y}, z) = \mathbb{1}_{\hat{y}}(y)$.*

Proof. The result follows from substituting the given happiness function in (4). □

Definition 4 (Equalized Odds)**.** *A classifier $\hat{Y}$ has* ε Equalized Odds *if for all $y \in \mathcal{Y}$,*

$$\left|p_{\hat{Y}|ZY}(\cdot|0, y) - p_{\hat{Y}|ZY}(\cdot|1, y)\right| \leq \varepsilon. \tag{10}$$

Lemma 3. *A classifier $\hat{Y}$ has ε Equalized Odds if and only if it is ε-fair w.r.t. the $n = |\mathcal{Y}|^2$ dimensional happiness function $\boldsymbol{\eta}(y, \hat{y}, z) = \left(\frac{\mathbb{1}_{y',\tilde{y}}(y,\hat{y})}{p_{Y|Z}(y'|z)}\right)_{(y',\tilde{y})\in\mathcal{Y}^2}$.*

Proof. Note that

$$\mathbb{E}[\boldsymbol{\eta}_{y',\tilde{y}}(Y, \hat{Y}, z)|Z = z] = \frac{\mathbb{E}[\mathbb{1}_{y',\tilde{y}}(Y, \hat{Y})|Z = z]}{p_{Y|Z}(y'|z)} \tag{11}$$

$$= \frac{p_{\hat{Y}Y|Z}(\tilde{y}, y'|z)}{p_{Y|Z}(y'|z)} = p_{\hat{Y}|YZ}(\tilde{y}|y', z), \tag{12}$$

where it is understood that $\boldsymbol{\eta}_{y',\tilde{y}}(Y, \hat{Y}, z)$ is an $|\mathcal{Y}|^2$-dimensional, vector-valued happiness function indexed by $(y', \tilde{y}) \in \mathcal{Y}^2$. Thus, substituting the given happiness function in (4) is equivalent to (10) for all $y \in \mathcal{Y}$. □

5 DEMONSTRATIVE CASE STUDIES

In this section we report three studies which showcase the application of the proposed post-processing framework using the synthetic dataset mentioned in Section 2.2, the `Adult` dataset Becker and Kohavi (1996), and the `Financial Risk for Loan Approval` dataset (Zoppelletto, 2024), respectively. All our experiments in this section follow the same procedure, which we will describe in the following.

Dataset and Baseline Classifier. All datasets contain a total of 48,842 samples. Given a dataset with features X, labels Y, and group labels Z, we split the dataset using a fixed random seed into training, validation, and test sets, containing 20%, 16%, and 64% of the data, respectively. We then train a simple random forest baseline classifier $\hat{Y}$ on the training data with accuracy $1 - \ell(Y, \hat{Y})$.

Outline of Experiments. We define a custom happiness function $\boldsymbol{\eta}$ and solve the linear program (5) for different values $\varepsilon > 0$. For each ε, we thereby obtain a ε-fair classifier $\tilde{Y}$ with accuracy $A(\varepsilon) = 1 - \ell(Y, \tilde{Y})$. Note that $A(\varepsilon)$ is monotonically increasing, thus $\varepsilon(A)$ exists. For comparison, we also post-process $\hat{Y}$ using "Statistical Parity", "Overall Accuracy" and "Equalized Odds". This can be achieved within our framework, but repeating the same process, replacing $\boldsymbol{\eta}$ with a happiness function $\boldsymbol{\eta}'$ for the corresponding method introduced in Section 4. Each method yields a family of post-processed classifiers $\tilde{Y}'$, for $\varepsilon' > 0$. We report the difference in happiness $\varepsilon(A') = \boldsymbol{\phi}(\boldsymbol{\eta}, \tilde{Y}')$ between the two groups measured by $\boldsymbol{\eta}$, as a function of accuracy $A' = 1 - \ell(Y, \tilde{Y}')$, where $\ell(\cdot, \cdot)$ is defined in (1).

For all classifiers, we perform the minimization (5) by using the validation set for the empirical approximation of the expectation operator, as outlined in Appendix A.1. The values $\varepsilon(A)$, i.e., the happiness gap as a function of accuracy, is subsequently computed using the test and validation sets. Statistical robustness is evidenced by the fact that the results for the test and validation set are close.

Computational Resources. Each experiment was performed in less than three minutes on a *AMD Ryzen 7 5700X* without GPU support. The memory required was less than 1 GB. This includes dataset generation and training of the random forest classifier.

5.1 Case Study with Synthetic Data

In this section, we provide a detailed explanation of the experiment introduced in Section 2.2.

We apply our method to a synthetic dataset, inspired by the `Adult` dataset Becker and Kohavi (1996). All data in this dataset is randomly drawn. For the column $Z = X_{\texttt{sex}}$, which we use as group label, we keep the original imbalance of about 1/3 female and 2/3 male from `Adult`. Values for the following features are drawn uniformly, independently at random: $X_{\texttt{age}}$, $X_{\texttt{hours\_per\_week}}$, $X_{\texttt{education}}$, $X_{\texttt{workclass}}$ and $X_{\texttt{race}}$. These features are completely independent of the classification task. They merely ensure that the linear classification problem introduced next is not learned perfectly by the baseline classifier.

Values for a new feature $X_{\texttt{yearly\_salary}}$ are normally distributed with mean $\mu_0 = \$50{,}000$ and standard deviation $\sigma_0 = \$1{,}000$. The *base* loan amount U is an independently drawn Gaussian with standard deviation \$10,000, and mean $\mu_1 = \$500{,}000$. A loan is granted if the *base* loan amount is less or equal than 10 times the annual income, i.e., $Y = \mathbb{1}\{10 \cdot X_{\texttt{yearly\_salary}} \geq U\}$. Thus, a loan is granted with a probability of 0.5. However, while male applicants request the base loan amount $X_{\texttt{loan\_requested}} = U$, for female applicants, \$50,000 are added to the loan, $X_{\texttt{loan\_requested}} = U + \$50{,}000$. The decision of acceptance is based on the *base* loan amount U, skewing the approval and resulting in the allocation additional funds to female applicants. This bias is already present in the training data.

Our baseline classifier is a simple random forest model trained on the training set, achieving approximately 82% accuracy on the test set.

The considered happiness function, based on the prediction and the requested loan, is defined as

$$\boldsymbol{\eta}(\hat{Y}, X, Y, Z) = \hat{Y} \cdot X_{\texttt{loan\_requested}},$$

where happiness is set to zero for rejected loan requests and to the loan amount for approved ones. Although simplistic, this example effectively illustrates the flexibility of our proposed framework. It demonstrates that ensuring equalized false and true positive rates across groups does not necessarily guarantee fairness in the allocation of resources.

In Figure 1, we observe optimizing for any of "Statistical Parity", "Overall Accuracy" and "Equalized Odds" does not improve fairness in terms of allocated funding measured by $\boldsymbol{\eta}$. Indeed "Statistical Parity" and "Equalized Odds" even amplify the bias in the training data, leading to an even larger difference after post-processing. All methods maintain reasonable classification performance, but only directly minimizing the difference in funding allows for a reduction of the imbalances in monetary allocation across groups. Note that this is possible while sacrificing less than one percentage point in accuracy. Figure 1 shows the test set results; the validation set plots are in Appendix B.1.

5.2 Case Study with `Adult` Data

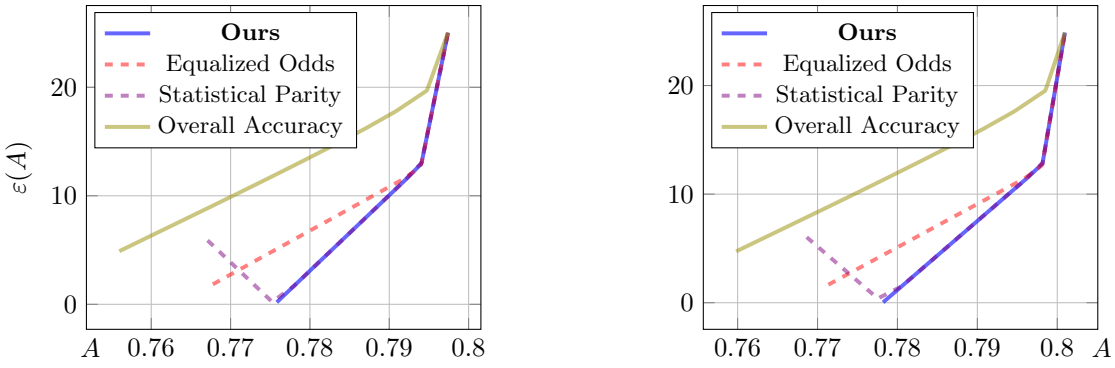
In this case study, we focus on the `Adult` dataset (Becker and Kohavi, 1996). While we retain the standard classification task of predicting whether an individual's income exceeds \$50,000, i.e., the hard decision is 1 if the predicted income is at least \$50,000 and 0 otherwise. We expand the analysis to incorporate a broader perspective on individual well-being.

Specifically, we use our framework to highlight that an individual who earns at least \$50,000 without working excessive hours may experience greater happiness. We posit that holding a job that yields higher income for fewer hours per week contributes positively to overall well-being, as the saved time can be reallocated to additional income-generating activities or personal pursuits.

To formalize this intuition, we define a happiness function as $\boldsymbol{\eta}(\hat{Y}, X, Y, Z) = 100 \cdot \hat{Y} - X_{\texttt{hours\_per\_week}}$, where $\hat{Y}$ is the predicted label, and the feature $X_{\texttt{hours\_per\_week}}$, indicating the number of hours an individual works each week has support is $\mathcal{X}_{\texttt{hours\_per\_week}} = \{x \in \mathbb{Z} \mid 1 \leq x \leq 99\}$. This function captures the idea that achieving a high income with fewer working hours leads to increased happiness.

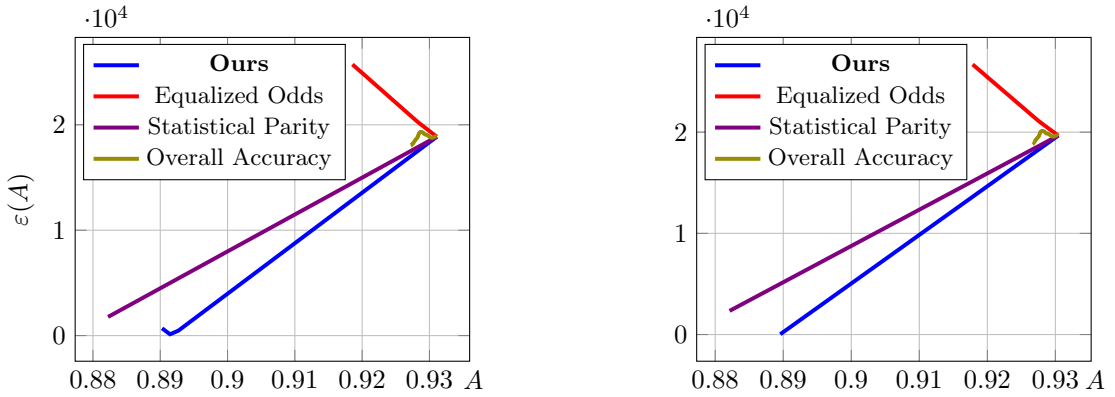
Consistent with the observations reported in Section 5.1, Figure 2 illustrates that our post-processing method achieves the most favorable trade-off between accuracy and happiness when the decision-making process accounts for the average number of working hours across different demographic groups. While the scenario depicted is simplified, it is representative of practical settings where one may wish to deploy our method to promote equitable working conditions for different groups. For instance, it could be applied to foster a work environment in which no group is disproportionately required to work overtime in order to receive comparable benefits, here represented by earning a salary above or below \$50,000. More broadly, our approach may be viewed as a scalable tool for addressing structural disparities in the workplace, with potential to mitigate issues such as high employee turnover.

Interestingly, while the "Statistical Parity" baseline achieves similar performance to our method in this particular case, it does not explicitly optimize for happiness. This limitation is reflected in the behavior of the growth function: the disparity in happiness between the two groups increases at both ends of the accuracy spectrum, with a localized point where the difference is negligible. Figure 2b further showcases



(a) Test (b) Validation

Figure 2: Experiment on `Adult` dataset.



(a) Test (b) Validation

Figure 3: Experiment on `Financial Risk for Loan Approval` dataset.

how our proposed solution achieves minimal $\varepsilon(A)$ on the validation dataset, in stark contrast with "Statistical Parity" and the other criteria.

5.3 Case Study with `Financial Risk for Loan Approval` Data

The `Financial Risk for Loan Approval` dataset (Zoppelletto, 2024) is used in classification tasks aimed at identifying individuals who pose a low risk of defaulting on a mortgage. Unlike `Adult`, this dataset includes more detailed information about the applicant's financial and employment status, factors commonly used by financial institutions when evaluating loan applications. These include, among others, employment or unemployment status, duration in that status, credit score, and the stated reason for requesting a loan.

Beyond risk assessment, such indicators can also be leveraged to model a user's perceived utility or happiness as a tradeoff between securing a mortgage and managing the repayment burden under the constraints of their economic situation as understood by financial institutions. In particular, we define a happiness function that models the potential impact of loan approval decisions on an individual's well-being, under the assumption that a loan can lead to long-term financial gains (e.g., home ownership, business investment), but also comes with repayment obligations (i.e., future costs due to interest).

We define the happiness function as $\boldsymbol{\eta}(\hat{Y}, X, Y, Z) = \hat{Y} \cdot \big(X_{\texttt{loan\_requested}} \cdot R(X) - C(X)\big)$, where $C(X) = X_{\texttt{loan\_requested}} \cdot \rho(X_{\texttt{credit\_score}}) \cdot X_{\texttt{duration}}$ is the total interest cost; $\rho(X_{\texttt{credit\_score}})$ is the interest rate determined by the credit score using the step function reported in Appendix B.2; $R(X)$ is the estimated return on investment as defined in Appendix B.2.

The function outputs zero if the loan is not approved ($\hat{Y} = 0$). This utility-based formulation captures both the benefit of approval and the financial burden of repayment, enabling evaluations beyond predictive accuracy.

Figure 3 reports the trade-off between the accuracy

of the classifier and the happiness function defined above. In particular, similarly to the case shown in Section 5.2, our proposed solution achieves higher accuracy among all the criteria for minimal $\varepsilon(A)$ (cf. Figure 3b). As previously disclosed, this is slightly different for the test data (cf. Figure 3a), where the optimal value for $\varepsilon(A)$ is not achieved by any criterion, though ours achieves the best accuracy overall when $\varepsilon(A)$ approaches 0.

6 SUMMARY AND CONCLUDING REMARKS

We have introduced a novel post-processing criterion for fairness in ML, through the lens of happiness, a measure of group satisfaction with a classifier's output. The proposed post-processing strategy is formulated as a linear program and thus, it allows for efficient computation while begin general enough to encompass many existing fairness criteria as particular instances.

Finally, we demonstrated the practicality of our approach through a series of case studies. Importantly, our method is readily extensible to fairness across multiple groups by incorporating additional constraints into the problem formulation.

This represents a novel perspective on the challenge of ensuring equal treatment of diverse populations, accounting not only for the classifier's output, but also for how this output translates into individual happiness. For example, this applies to scenarios involving the allocation of resources across groups, highlighting just one facet of the broader impact our framework can have in promoting equitable outcomes.

Limitations

Our framework, in line with comparable criteria, focuses on group-level fairness and does not account for individual happiness, which means that the interests of outliers or individuals whose preferences significantly diverge from the group may be overlooked.

Additionally, the derivation of our algorithm relies crucially on two assumptions: that the label space is finite and that the classifier outputs a soft prediction, i.e., a probability distribution over the label space. As a result, the current formulation is not directly applicable to regression problems or settings where predictions are continuous rather than categorical.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. M. (2018). A reductions approach to fair classification. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR.

Alghamdi, W., Hsu, H., Jeong, H., Wang, H., Michalak, P. W., Asoodeh, S., and Calmon, F. (2022). Beyond adult and COMPAS: Fair multi-class prediction via information projection. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

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

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

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.

Bharti, B., Yi, P., and Sulam, J. (2023). Estimating and controlling for equalized odds via sensitive attribute predictors. *Advances in neural information processing systems*, 36:37173–37192.

Buyl, M. and De Bie, T. (2022). Optimal transport of classifiers to fairness. *Advances in Neural Information Processing Systems*, 35:33728–33740.

Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7).

Cerrato, M., Köppel, M., Wolf, P., and Kramer, S. (2024). 10 years of fair representations: Challenges and opportunities.

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*, pages 214–226.

Dwork, C., Lee, D., Lin, H., and Tankala, P. (2023). From pseudorandomness to multi-group fairness and back.

Fabris, A., Messina, S., Silvello, G., and Susto, G. A. (2022). Tackling documentation debt: A survey on algorithmic fairness datasets. EAAMO '22, New York, NY, USA. Association for Computing Machinery.

Gohar, U. and Cheng, L. (2023). A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 6619–6627. ijcai.org.

Gordaliza, P., Barrio, E. D., Fabrice, G., and Loubes, J.-M. (2019). Obtaining fairness using optimal transport theory. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2357–2365. PMLR.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323.

Jiang, Z., Han, X., Fan, C., Yang, F., Mostafavi, A., and Hu, X. (2022). Generalized demographic parity for group fairness. In *International Conference on Learning Representations*.

Kasy, M. and Abebe, R. (2021). Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 576–586.

Kim, M., Reingold, O., and Rothblum, G. (2018). Fairness through computationally-bounded awareness. *Advances in neural information processing systems*, 31.

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):e1452.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. (2018). Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR.

Liu, L. T., Simchowitz, M., and Hardt, M. (2019). The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pages 4051–4060. PMLR.

Liu, T., Wang, H., Wang, Y., Wang, X., Su, L., and Gao, J. (2023). Simfair: A unified framework for fairness-aware multi-label classification. In Williams, B., Chen, Y., and Neville, J., editors, *Thirty-Seventh AAAI Conference on Artificial Intelligence*, pages 14338–14346. AAAI Press.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020). Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR.

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

Rouzot, J., Ferry, J., and Huguet, M.-J. (2023). Learning optimal fair scoring systems for multi-class classification. *ICTAI 2022 - The 34th IEEE International Conference on Tools with Artificial Intelligence, Oct 2022, Virtual, United States*.

Silvia, C., Ray, J., Tom, S., Aldo, P., Heinrich, J., and John, A. (2020). A general approach to fairness with optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3633–3640.

Tang, Z. and Zhang, K. (2022). Attainability and optimality: The equalized odds fairness revisited. In *Conference on Causal Learning and Reasoning*, pages 754–786. PMLR.

Taturyan, G., Chzhen, E., and Hebiri, M. (2024). Regression under demographic parity constraints via unlabeled post-processing. *Advances in Neural Information Processing Systems*, 37:117917–117953.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA. Association for Computing Machinery.

Wang, H., He, L., Gao, R., and Calmon, F. (2023). Aleatoric and epistemic discrimination: Fundamental limits of fairness interventions. *Advances in Neural Information Processing Systems*, 36:27040–27062.

Weber, A., Metevier, B., Brun, Y., Thomas, P. S., and da Silva, B. C. (2022). Enforcing delayed-impact fairness guarantees. *arXiv preprint arXiv:2208.11744*.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. (2017). Learning non-discriminatory predictors. In *Conference on learning theory*, pages 1920–1953. PMLR.

Yadav, C., Roy Chowdhury, A., Boneh, D., and Chaudhuri, K. (2024). FairProof : Confidential and certifiable fairness for neural networks. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 55682–55705. PMLR.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2019). Fairness constraints: a flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42.

Zoppelletto, L. (2024). Financial Risk for Loan Approval. Kaggle.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
   Yes. See Section 3.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   Not Applicable. Though we do not provide an explicit algorithm, the studied problem is a linear program see Theorem 1.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
   Yes. The code is provided as supplementary material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
   Yes.
   (b) Complete proofs of all theoretical results.
   Yes. Either provided in the main text or in Appendix A.
   (c) Clear explanations of any assumptions.
   Yes.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
   Yes. The code can be found in the supplementary material, while Section 5 contains details about the experimental setting.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
   Yes. See Section 5 and Appendix B.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
   Not Applicable. Our setup consists of solving a linear program, which is deterministic.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
   Yes. See the paragraph on "Computational Resources" in Section 5.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets.
   Yes.
   (b) The license information of the assets, if applicable.
   Not Applicable. The datasets used are publicly available.
   (c) New assets either in the supplemental material or as a URL, if applicable.
   Not Applicable.
   (d) Information about consent from data providers/curators.
   Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.
   Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots.
   Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
   Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
   Not Applicable.

Happiness as a Measure of Fairness: Supplementary Materials

A PROOF AND DISCUSSION OF THEOREM 1

Theorem 1. *For a fixed $\varepsilon \geq 0$ we can find the minimum L, such that (ε, L) is achievable by solving*

$$\min_{p_{\tilde{Y}|\hat{Y}Z}} \ell(Y, \tilde{Y}) \tag{5}$$

$$s.t.\ \boldsymbol{\phi}(\boldsymbol{\eta}, \tilde{Y}) \leq \varepsilon. \tag{6}$$

This is a linear programming problem.

Proof of Theorem 1. The constraint ensures that $\tilde{Y}$ is ε-fair, while maximizing the accuracy of $\tilde{Y}$. It remains to show that the optimization problem (5) is a linear program.

First, we argue that $\ell(Y, \tilde{Y})$ is an affine function of $p_{\tilde{Y}|\hat{Y}Z}$, as

$$\ell(Y, \tilde{Y}) = P\{Y \neq \tilde{Y}\} = 1 - P\{Y = \tilde{Y}\} \tag{13}$$

$$= 1 - \sum_{y,\hat{y},z} p_{\hat{Y}YZ}(\hat{y}, y, z) p_{\tilde{Y}|\hat{Y}Z}(y|\hat{y}, z), \tag{14}$$

where we used (3). The implicit constraints, ensuring that $p_{\tilde{Y}|\hat{Y}Z}$ is a valid probability mass function can be written as linear inequalities.

Finally, we can complete the proof by showing that $\mathbb{E}[\boldsymbol{\eta}(\tilde{Y}, X, Y, Z)|Z = z]$ can be written as a linear function

$$\mathbb{E}[\boldsymbol{\eta}(\tilde{Y}, X, Y, Z)|Z = z] = \sum_{\tilde{y},\hat{y}} \boldsymbol{\xi}(\tilde{y}, \hat{y}, z) p(\tilde{y}|\hat{y}, z), \tag{15}$$

where the coefficients are given by $\boldsymbol{\xi}(\tilde{y}, \hat{y}, z) = p_{\hat{Y}|Z}(\hat{y}|z)\, \mathbb{E}\big[\boldsymbol{\eta}(\tilde{y}, X, Y, z)\big|\hat{Y} = \hat{y}, Z = z\big]$, which yields

$$\sum_{\tilde{y},\hat{y}} \boldsymbol{\xi}(\tilde{y}, \hat{y}, z) p(\tilde{y}|\hat{y}, z) = \sum_{\tilde{y},\hat{y}} p_{\hat{Y}|Z}(\hat{y}|z)\, \mathbb{E}\big[\boldsymbol{\eta}(\tilde{y}, X, Y, z)\big|\hat{Y} = \hat{y}, Z = z\big] p_{\tilde{Y}|\hat{Y}Z}(\tilde{y}|\hat{y}, z) \tag{16}$$

$$= \sum_{\tilde{y},\hat{y}} \mathbb{E}\big[\boldsymbol{\eta}(\tilde{y}, X, Y, z)\big|\hat{Y} = \hat{y}, Z = z\big] p_{\tilde{Y}\hat{Y}|Z}(\tilde{y}, \hat{y}|z) \tag{17}$$

$$= \mathbb{E}\Big[\mathbb{E}\big[\boldsymbol{\eta}(\tilde{Y}, X, Y, z)\big|\hat{Y}, Z = z\big]\Big|Z = z\Big] \tag{18}$$

$$= \mathbb{E}\big[\boldsymbol{\eta}(\tilde{Y}, X, Y, Z)\big|Z = z\big]. \tag{19}$$

□

A.1 Empirical Approximation

When replacing expectation with empirical expectation in Theorem 1, an accurate result can be obtained if the loss function $\ell(Y, \tilde{Y})$ as well as $\boldsymbol{\xi}(\tilde{y}, \hat{y}, z)$ can be well approximated.

We use empirical approximations of $p_{\hat{Y}YZ}(\hat{y}, y, z)$ as well as $\boldsymbol{\xi}(\tilde{y}, \hat{y}, z)$ defined as

$$\hat{p}_{\hat{Y}YZ}(\hat{y}, y, z) := \frac{1}{|\mathcal{D}|} \sum_{(x', y', p_{\hat{Y}}, z') \in \mathcal{D}} p_{\hat{Y}}(\hat{y}) \mathbb{1}_{z'}(z) \mathbb{1}_{y'}(y), \tag{20}$$

$$\hat{\boldsymbol{\xi}}(\tilde{y}, \hat{y}, z) := \frac{1}{\hat{p}_Z(z)\,|\mathcal{D}|} \sum_{(x, y, p_{\hat{Y}}, z') \in \mathcal{D}} \mathbb{1}_{z'}(z) p_{\hat{Y}}(\hat{y}) \boldsymbol{\eta}(\tilde{y}, x, y, z), \tag{21}$$

where $\mathcal{D}$ is a validation dataset and $\mathbb{1}$ denotes the indicator function. If the approximations are tight, a bound on the solution of (5) can readily be obtained.

Lemma 4. *Let $L(\varepsilon)$ be the solution of (5) and $\hat{L}(\varepsilon)$ the solution when substituting with (20) and (21). Assume that $|\hat{p}_{\hat{Y}YZ}(\hat{y},y,z)-p_{\hat{Y}YZ}(\hat{y},y,z)|\leq\delta$ and $|\hat{\boldsymbol{\xi}}(\tilde{y},\hat{y},z)-\boldsymbol{\xi}(\tilde{y},\hat{y},z)|\leq\delta$ for all $\tilde{y},\hat{y},y,z$. Then, $L(\varepsilon)\leq\hat{L}(\varepsilon-2\delta)+\delta$.*

Proof. Let $p_{\tilde{Y}|\hat{Y}Z}$ achieve $\hat{L}(\varepsilon-2\delta)$. By the assumption, substituting $\hat{p}$ and $\hat{\boldsymbol{\xi}}$ in (14) and (15), results in an error of at most δ, thus yielding $\ell(Y,\tilde{Y})\leq\hat{L}(\varepsilon-2\delta)+\delta$ and

$$|\mathbb{E}[\boldsymbol{\eta}(\tilde{Y},X,Y,Z)|Z=0]-\mathbb{E}[\boldsymbol{\eta}(\tilde{Y},X,Y,Z)|Z=1]|\leq\varepsilon. \tag{22}$$

□

Noting that $\mathbb{E}[\hat{p}_{\hat{Y}YZ}(\hat{y},y,z)]=p_{\hat{Y}YZ}(\hat{y},y,z)$ and $\mathbb{E}[\hat{\boldsymbol{\xi}}(\tilde{y},\hat{y},z)]=\boldsymbol{\xi}(\tilde{y},\hat{y},z)$, there are a total of $2(1+n)|\mathcal{Y}|^2$ random variables, which are required to be within δ of their respective expected value for the conditions of Lemma 4 to be satisfied. Note in particular, that this only depends on the size of the label space, not on the feature spaces. For known $\boldsymbol{\xi}$, a concentration result can be used to obtain bounds for the necessary size of the validation set to guarantee adequate approximation with high probability:

Lemma 5. *Let $\gamma,\delta>0$ and assume $A\leq\boldsymbol{\eta}\leq B$ with $1\leq C:=B-A$. Furthermore, let the validation set $\mathcal{D}$ contain at least D samples from each group. Then, if*

$$D\geq\frac{C^2}{2\delta^2}\log\frac{4(n+1)|\mathcal{Y}|^2}{\gamma}, \tag{23}$$

with probability at least $1-\gamma$, we have that the assumptions of Lemma 4 are satisfied, i.e., $|\hat{p}_{\hat{Y}YZ}(\hat{y},y,z)-p_{\hat{Y}YZ}(\hat{y},y,z)|\leq\delta$ and $|\hat{\boldsymbol{\xi}}(\tilde{y},\hat{y},z)-\boldsymbol{\xi}(\tilde{y},\hat{y},z)|\leq\delta$ for all $\tilde{y},\hat{y},y,z$.

Proof. For each $\hat{y},y,z\in\mathcal{Y}^2\times\{0,1\}$ let $\mathcal{E}^0_{\hat{y},y,z}$ be the event that $|\hat{p}_{\hat{Y}YZ}(\hat{y},y,z)-p_{\hat{Y}YZ}(\hat{y},y,z)|\geq\delta$. Similarly, let $\mathcal{E}^i_{\tilde{y},\hat{y},z}$ be the event that $|\hat{\boldsymbol{\xi}}_i(\tilde{y},\hat{y},z)-\boldsymbol{\xi}_i(\tilde{y},\hat{y},z)|\geq\delta$ for all $\tilde{y},\hat{y},z,i\in\mathcal{Y}^2\times\{0,1\}\times\{1,\dots,n\}$. Using Hoeffding's inequality we obtain

$$\begin{aligned}
P\{\mathcal{E}^i_{\tilde{y},\hat{y},z}\}&\leq 2\exp\Big(-\frac{2\delta^2D}{C^2}\Big)\qquad i\in\{1,\dots,n\}\\
P\{\mathcal{E}^0_{\hat{y},y,z}\}&\leq 2\exp\big(-4\delta^2D\big)\leq 2\exp\Big(-\frac{2\delta^2D}{C^2}\Big).
\end{aligned}$$

Thus, bounding the probability of the union,

$$\begin{aligned}
P\Bigg\{\bigcup_{y\in\mathcal{Y},y'\in\mathcal{Y},z\in\mathcal{Z},i\in\{0,\dots,n\}}\mathcal{E}^i_{y,y',z}\Bigg\}&\leq 2(n+1)|\mathcal{Y}|^2 2\exp\Big(-\frac{2\delta^2D}{C^2}\Big)\\
&=4(n+1)|\mathcal{Y}|^2\exp\Big(-\frac{2\delta^2D}{C^2}\Big)
\end{aligned}$$

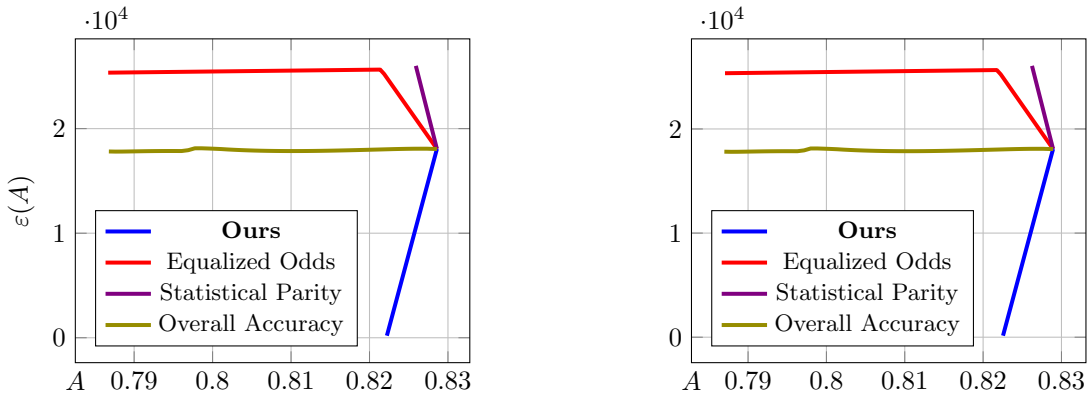
and the desired bounds hold with probability at least $1-\gamma$ if (23) holds. □

Note that the bound (23) yields achievable quantities in typical situations. E.g., for any of the criteria in Section 4 and a binary classification problem, we can take $C=1$ and $n=|\mathcal{Y}|=2$. If we require an accuracy of $\delta=0.02$ with a probability of at least $1-\gamma=0.99$, the bound (23) requires $D\geq 10{,}596$, which is satisfied in the experiments in Section 5.

B CASE STUDY DETAILS

B.1 Case Study with Synthetic Data

Figure 4 shows the results on the problem introduced in Section 2.2. Values for the test and validation set are very close.



(a) Test (b) Validation

Figure 4: Experiment on `Synthetic` dataset.

B.2 Case Study with `Financial Risk for Loan Approval` Data

In this section we report the details for the reproducibility of the experiment in Section 5.3 where:

- The step function $\rho(X_{\texttt{credit\_score}})$ is defined as:

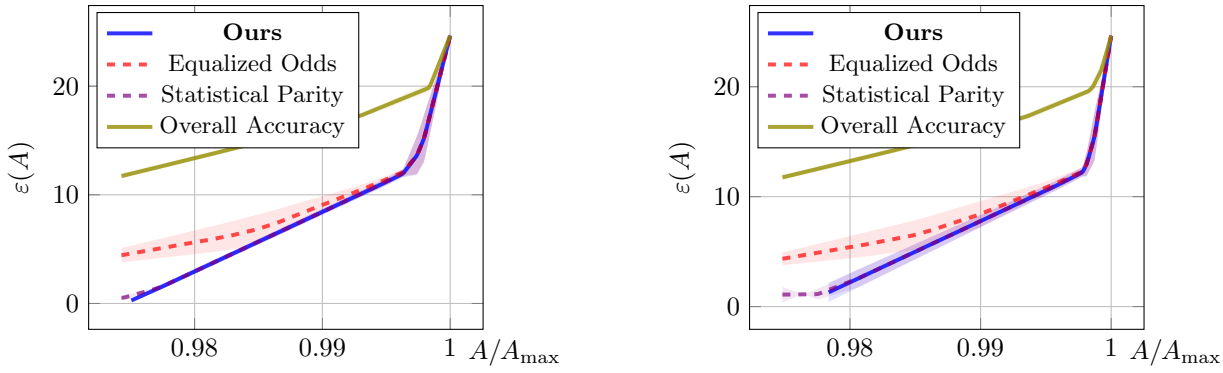
$$\rho(X_{\texttt{credit\_score}}) = \begin{cases} 0.04 & \text{if } X_{\texttt{credit\_score}} \geq 750 \\ 0.06 & \text{if } 700 \leq X_{\texttt{credit\_score}} < 750 \\ 0.08 & \text{if } 650 \leq X_{\texttt{credit\_score}} < 700 \\ 0.12 & \text{if } 600 \leq X_{\texttt{credit\_score}} < 650 \\ 0.18 & \text{if } X_{\texttt{credit\_score}} < 600 \end{cases}$$

- The return of interest is defined as

$$R(X) = \sum_{j \in \{\texttt{loan\_purpose},\texttt{education},\texttt{employment},\texttt{tenure}\}} \beta(X_j),$$

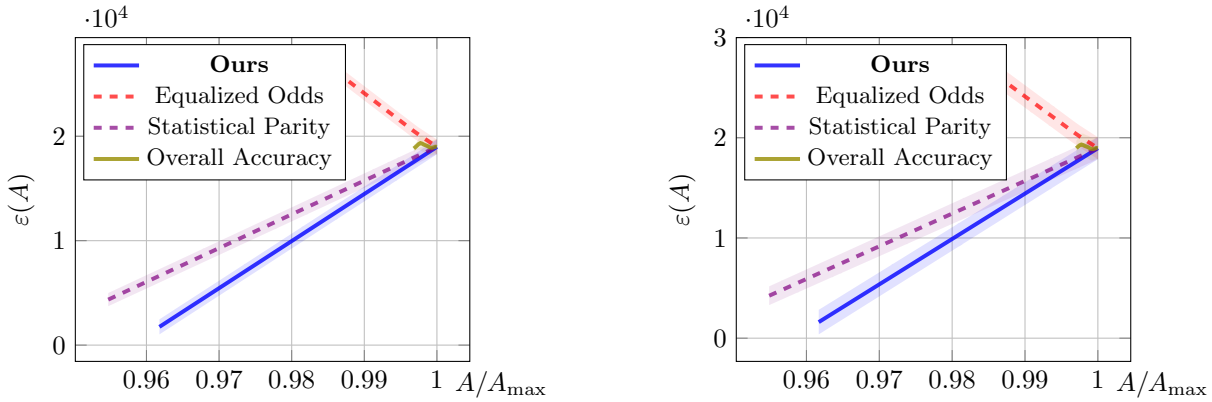
  where $\beta(x)$ are domain-informed weights, assigned as reported below.

- The Bonuses β are assigned according to the following criteria:
  - Loan Purpose Bonuses
    1. $\beta(\texttt{Home}) = 0.08$,
    2. $\beta(\texttt{Auto}) = 0.02$,
    3. $\beta(\texttt{Education}) = 0.12$,
    4. $\beta(\texttt{Debt Consolidation}) = 0.04$,
    5. $\beta(\texttt{Other}) = 0.05$.
  - Education Level Bonuses
    1. $\beta(\texttt{Master}) = 0.01$,
    2. $\beta(\texttt{Doctorate}) = 0.02$.
  - Employment Status Bonuses
    1. $\beta(\texttt{Employed}) = \beta(\texttt{Self-Employed}) = 0.01$.
  - Tenure Bonus
    1. $\beta(n) = 0.01 \cdot \mathbb{1}\{n > 5 \text{ years}\}$.



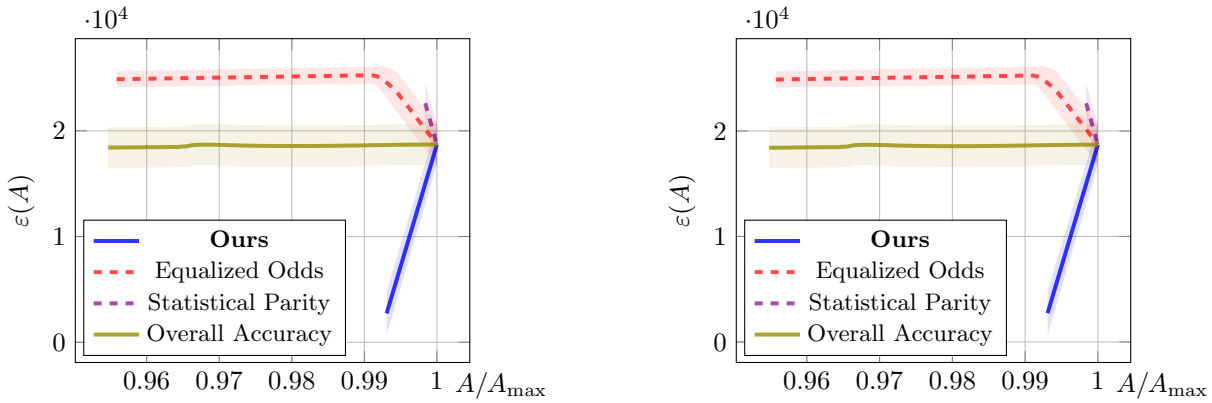
(a) Test (b) Validation

Figure 5: Experiment on `Adult` dataset, with different seeds.



(a) Test (b) Validation

Figure 6: Experiment on `Financial` dataset, with different seeds.



(a) Test (b) Validation

Figure 7: Experiment on `Synthetic` dataset, with different seeds.

C ADDITIONAL NUMERICAL RESULTS

To illustrate the robust nature of our approach, we performed additional experiments with varying random seeds on all three datasets, Synthetic, Adult, and Financial Risk. The split ratios have been kept consistent across all seeds, in line with the original experimental setup in Section 5. We used seeds 0, 1, and 10. The results, reported in Appendix C, demonstrate that the performance of our method is stable across different random seeds,

consistently achieving favorable trade-offs between accuracy and happiness. Notably, the values on the x-axis have been normalized by dividing by the maximum accuracy value, accounting for different accuracy values across seeds, thus attaining better readability. Note that the curves are only depicted for points in regions where $\varepsilon(A)$ is well-defined for all seeds.