

Towards a Fairness-Aware Scoring System for Algorithmic Decision-Making

Yi Yang*

Ying Wu[†]Xiangyu Chang[‡]Mei Li[§]

Abstract

Scoring systems, as simple classification models, have significant advantages in interpretability and transparency when making predictions. They facilitate humans' decision-making by allowing them to make a quick prediction by hand through adding and subtracting a few point scores and thus have been widely used in various fields such as medical diagnosis of Intensive Care Units. However, (un)fairness issues in these models have long been criticized, and the use of biased data in the construction of score systems heightens this concern. In this paper, we propose a general framework to create data-driven fairness-aware scoring systems. Our approach is first to develop a social welfare function that incorporates both efficiency and equity. Then, we translate the social welfare maximization problem in economics into the empirical risk minimization task of the machine learning community to derive a fairness-aware scoring system with the help of mixed integer programming. We show that the proposed framework provides practitioners or policymakers great flexibility to select their desired fairness requirements and also allows them to customize their own requirements by imposing various operational constraints. Experimental evidence on several real data sets verifies that the proposed scoring system can achieve the optimal welfare of stakeholders and balance the interpretability, fairness, and efficiency issues.

1 Introduction

Prediction models play an essential role in our daily life. They are frequently used to facilitate humans in a variety of decision-making scenarios. A *scoring system*, as a form of an interpretable

*Department of Information Management and E-Business, School of Management, Xi'an Jiaotong University

[†]Department of Information Management and E-Business, School of Management, Xi'an Jiaotong University

[‡]Center for Intelligent Decision-Making and Machine Learning, School of Management, Xi'an Jiaotong University;
email: xiangyuchang@xjtu.edu.cn.

[§]Department of Marketing and Supply Chain Management, Price College of Business, University of Oklahoma

predictive model, is a sparse linear model whose coefficients are small integers. These coefficients could be directly transferred to point scores for a scorecard which allows users to predict by only adding/subtracting or multiplying a few small numbers (Ustun and Rudin, 2016). These tools are convenient and practical since they allow practitioners to make quick predictions by hand, without a computer or calculator, and without extensive training (Zeng et al., 2017).

The applications of scoring systems can be traced back at least to Burgess (1928)’s research on a parole violation. Nowadays, scoring systems are still widely used in many fields, from medical diagnosis and criminal justice to financial loans and humanitarian aid. As the vast majority of predictive models in the healthcare and justice systems, scoring systems have been studied extensively over the years, and several well-known scorecards have been developed. Examples from the medical field include APACHE I, II, and III (Knaus et al., 1981, 1985, 1991); SAPS I, II, and III (Le Gall et al., 1984, 1993; Moreno et al., 2005) to predict ICU mortality risk; and SIRS (Bone et al., 1992) to detect system inflammatory response syndrome. Instances from criminal justice, Salient Factor Score (Hoffman and Adelberg, 1980), Offense Gravity Score (Kramer and Scirica, 1986), Criminal History Score (Hoffman and Beck, 1997), and COMPAS (Correctional Offender Management Profiling System for Alternative Sanctions) score (Northpointe, 2015) are adopted at various stages of the criminal justice system, such as at pretrial, parole, probation, or even sentencing in some states. In the field of finance and business, as one of the earliest risk management tools, scoring systems are used not only for credit assessment (Capon, 1982; Tsaih et al., 2004; Karlan and Zinman, 2011; Li et al., 2020) and fraud prevention (Dionne et al., 2009; Vona, 2012; Halvaiee and Akbari, 2014; Gómez et al., 2018), but also for direct marketing (Malthouse, 1999; Bose and Chen, 2009) and insurance risk evaluation (Coutts, 1984; Frees et al., 2011). In addition, several scoring systems are also developed to estimate poverty in the population and prioritize aid resources allocation (Hernandez and Torero, 2018; Skoufias et al., 2020).

Generally, there are multiple ways to construct a scoring system. In some cases, it is hand-crafted by a panel of experts using only their domain expertise (e.g. the APACHE I by Knaus et al. (1981)). Nevertheless, some scoring systems are data-driven, usually in the sense that they are derived using regression models followed by the rounding of coefficients to obtain integer-valued point scores (e.g. the SPAS II by Le Gall et al. (1993)). In addition to traditional statistical approaches, more and more machine learning techniques are also introduced to construct scoring systems based on data. The standard procedures to develop a data-driven scoring system are illustrated in a flowchart in Figure 1. Typically, data set construction is the first activity, which starts with data collection of historical cases followed by the data processing procedures such as data cleansing and feature engineering. On the basis of the processed data set, the scorecard construction could be carried out. This phase relies on the chosen algorithm and the requirements

of application scenarios. It conducts model fitting with the training set and scales the model into a scorecard. After that, the derived scorecard will be evaluated on the test set to provide an overview of its predicted performance. Once the scorecard is validated, it will be implemented in practice, and a monitoring and tracking procedure will be conducted to give the flag of update or redevelopment (Thomas et al., 2017).

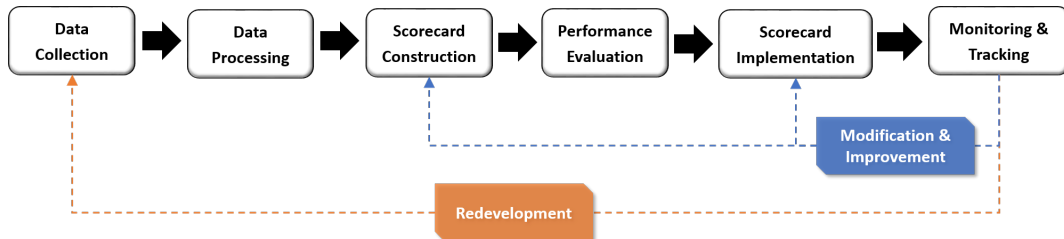


Figure 1: General steps in developing a data-driven scoring system.

As indicated in Figure 1, data plays a vital role in the development of scoring systems and is the foundation of the whole process. The quality of the input data set primarily determines the quality of system output. Although nowadays large-scale data collection is becoming more accessible, affordable, and effective, it has created many societal issues. One of the most sensitive issues is fairness concerns. Data, especially big data, is frequently heterogeneous, generated by subgroups with their own traits and behaviors. These heterogeneities can bias the data. A model built on biased data might result in unfair predictions (Mehrabi et al., 2021). Empirical work is increasingly lending support to these concerns. In criminal justice, a well-known example is the COMPAS scandal that a nonprofit organization named ProPublica argued that the COMPAS criminal recidivism scoring system is biased against African-American defendants (Chouldechova, 2017). Angwin et al. (2016) showed that this system skewed towards labeling black defendants as high risk whereas white defendants as low risk. Coincidentally, in healthcare, a widely used algorithm that allocates resources like care management produces scores that exhibit significant racial bias: black patients are considerably sicker than white patients at a given risk score (Obermeyer et al., 2019; Obermeyer and Mullainathan, 2019). Note that this health system now drives important healthcare decisions for over 70 million people in the US, and the disparity of access to medical resources affects a number of lives in the underrepresented minorities. Similar unfairness is also demonstrated to exist in gender in healthcare delivery (Bierman, 2007; Chen et al., 2008). In the financial field, big data credit-scoring models pose significant risks to transparency and fairness. They can place groups of individuals that share a sensitive feature like gender, race, or religion at a systematic disadvantage in terms of rejection rate or interest rate. For example, Apple Pay has

received criticism for setting credit limits for female customers at a much lower level than for those of comparable male customers (Vigdor, 2019). Besides, Fuster et al. (2021) demonstrate that the use of machine-learning algorithms enlarges the disparity in credit market outcomes across different groups of borrowers.

The advent of evidence in all these fields heightens people’s concerns regarding fairness and gives rise to growing public distrust in the decision-making systems that impact our daily lives. Several laws and regulations have been established to protect fairness based on sensitive features for some high-stakes domains like credit, housing, education, healthcare, and employment. Examples include the Equal Credit Opportunity Act (ECOA), Equal Employment Opportunity Act (EEOA), Fair Housing Act (FHA), Section 1557 of the Affordable Care Act (ACA), and General Data Protection Regulation (GDPR). Most of the legislation prohibits *disparate treatment* to address procedural discrimination: treating individuals differently on the basis of membership in a certain class (e.g., race or gender) and intent to discriminate (Barocas and Selbst, 2016). This indicates that any explicit use of sensitive features in constructing algorithmic predictions is strictly prohibited by laws. However, even if the sensitive features are excluded from inputs, the prediction results can still be biased toward or against individuals with specific sensitive features since the other input attributes usually are correlated with the sensitive ones. Sophisticated algorithms may combine these facially neutral features and treat them as proxies for sensitive characteristics, thereby circumventing existing non-discrimination laws and systematically still denying resource access to certain groups (Hurley and Adebayo, 2017). Thus, considering this “unintentional” discrimination as well as the outcome (un)fairness are crucial when developing data-driven scoring systems.

Therefore, in this paper, we mainly focus on two major doctrines of outcome (un)fairness: *disparate impact* and *disparate mistreatment*. *Disparate impact* recognizes a liability for the case where a system adversely affects the members from one group more than another, even if it appears to be neutral. *Disparate mistreatment* recognizes a liability for the case where a system produces the misclassification rates differ for groups of individuals with different memberships. To respond to these two concepts of unfairness, a great deal of efforts in the machine learning community has been invested in defining the corresponding notions/criteria of algorithmic fairness, such as *statistical parity* (Corbett-Davies et al., 2017) and *equality of opportunity* (Hardt et al., 2016). These notions are proposed to accommodate different application scenarios and equity goals. Unfortunately, there is no universally accepted notions of fairness, since they all have their own advantages and disadvantages. Most of the existing predictive systems or algorithms are specifically designed for only one of the notions, and thus, they are difficult to accommodate more than one notion simultaneously. As a result, their application is greatly limited to a certain scenario. Moreover, these works applying in-process interventions for fairness are usually based on an empirical risk minimization problem

that incorporates a fairness constraint with the unfairness tolerance level ϵ . Since utilizing hard loss in this kind of framework results in a difficult nonconvex and nonsmooth problem, researchers usually apply convex surrogate losses such as the cross entropy loss instead of 0-1 loss to construct the relaxed versions both of the objective function (i.e., error rate) and the fairness constraint (Hu and Chen, 2020; Donini et al., 2018; Zafar et al., 2017a, 2019). However, these approximations may lead to a poor trade-off between accuracy and sparsity, the sub-optimality of the solutions with respect to the original problem, and produce systems that are not robust to outliers (Ustun and Rudin, 2016). Furthermore, it is difficult to directly make use of these techniques to create fair scoring systems because the former usually assume continuous coefficients and are hard to control the sparsity, while the latter need to be accurate, sparse, and use small coprime integer coefficients. It is also noteworthy that these works rarely provide guidelines regarding selecting the appropriate value of ϵ . They either vary the value of ϵ in the range $[0, 1]$ (for the purpose to see its impact on accuracy, for example) or set ϵ to a value arbitrarily (e.g., $\epsilon = 0$ for strict elimination of group difference). However, the smaller difference between groups can constitute adverse impact and, greater differences may not, depending on circumstances (Barocas and Selbst, 2016). Thus, finding a proper fairness level to balance efficiency and equity is also of great importance since it has legal, ethical, economic, and regulatory implications.

Because the scoring systems are used in many socially sensitive environments to make important and life-changing decisions, their performance and outcomes directly bear on individuals' well-being. Thus, our research views prediction results as resource allocations awarded to individuals and, by extension, to various social groups. This work proposes a flexible framework to design a fairness-aware scoring system from a welfare-centric perspective, which does not suffer from the above limitations. More specifically, first, we formalize an individual utility function depended on the system outcome and its level of (un)fairness. Then, we formulate them into a traditional social welfare maximization framework prevalent in economics. This welfare perspective allows us to directly engage both needs of efficiency and equity. After that, we cast the welfare maximization problem as the empirical risk minimization (ERM) task at the center of supervised learning. By solving the corresponding mixed integer programming, a fairness-aware scoring system is developed as well as the optimal fairness level. Unlike the previous literature, the 0-1 hard loss is directly incorporated in the proposed framework to avoid the flaws derived by surrogate loss. Finally, the applications of our system on several real data sets from different fields are provided for illustrating its efficacy and efficiency.

1.1 Related Literature

This section briefly discusses the literature concerning (un)fairness issues in machine learning, operation management, and economics.

1.1.1 Fairness Research in Machine Learning

Recently, the fairness issues of machine learning algorithms have attracted more and more attention from researchers in that community. Much fair machine learning literature focuses on classification scenarios where a disadvantaged group suffers from discrimination through a classifier. Plenty of works have been conducted to formalize the concept of fairness, such as statistical parity (Dwork et al., 2012; Corbett-Davies et al., 2017), conditional statistical parity (Corbett-Davies et al., 2017), equality of opportunity and equalized odds (Hardt et al., 2016), individual fairness (Dwork et al., 2012), and representational fairness (Zafar et al., 2017c), etc. Based on these notions, various algorithmic interventions are designed to implement the fairness requirements. Previous works on this topic can be mainly categorized into three groups: modify the learning procedure whether in pre-processing, in-processing training, or post-processing stages (Barocas et al., 2017). Most of the existing algorithmic or in-processing approaches mainly aim at solving a constrained optimization problem by imposing a constraint on the fairness level with various fairness measures while optimizing the learning objective like accuracy. However, because most fairness metrics are non-convex due to the use of the indicator function, it is difficult to solve the master optimization problem. A widely-used strategy to achieve convexity is to adopt surrogate functions for both objectives and constraints. Examples of this scheme include Woodworth et al. (2017); Quadrianto and Sharmanska (2017); Zafar et al. (2017a,b); Donini et al. (2018); Zafar et al. (2019); Hossain et al. (2020). Most of these works are limited to a single notion of fairness or support only a single sensitive attribute, which limits their generality (Kozodoi et al., 2021). Although several attempts have been made to develop a unified framework that can handle more than one fairness notion (Quadrianto and Sharmanska, 2017; Zafar et al., 2017a, 2019), they still utilize surrogate functions instead of hard loss to avoid non-convex optimization. This may lead to sub-par fairness and the sub-optimality of the produced classifier (Lohaus et al., 2020). Besides, these methods usually assume the classifier coefficients are continuous and hard to control the sparsity, which becomes an obstacle to using these techniques to create scoring systems. The proposed framework directly applies 0-1 hard loss without approximation to construct scoring systems that tackle the above flaws. Similar to our approach, Lawless and Günlük (2020) formulate optimal decision rules subject to explicit constraints on fairness. However, unlike our approach, which applies the social welfare maximization to produce the optimal fairness level and scoring system simultaneously, their approach aims to

find predictive rules only maximizing the accuracy for a given fairness level. Besides, since they finally produce boolean rule sets in disjunctive normal form, their approach is unable to provide “a qualitative understanding of the relationship between joint values of the input variables and the resulting predicted response value” (Hastie et al., 2009). Thus, their models cannot help users gauge the influence of each input variable with respect to the final output.

1.1.2 Fairness Research in Operation Management

In this paper, we view the prediction results to some extent as “resource allocations” awarded to individuals, furthermore to various social groups. While fairness is a relatively recent criterion in the areas of machine learning decision-making, it has a long history of study in the literature on resource allocation (Steinhaus, 1948). In the context of resource allocation in which a set of goods or chores must be distributed among several agents, fairness is desirable in these cases and indicates that each agent gets a fair share. In recent decades, a number of works have been conducted to study the *fair division*, spanning settings with both divisible (Brams et al., 1996; Robertson and Webb, 1998; Abdulkadiroğlu et al., 2004; Procaccia, 2013, 2015; Gal et al., 2017) and indivisible items (Steinhaus, 1948; Lipton et al., 2004; Aziz et al., 2015; Lang and Rothe, 2016; Cole and Gkatzelis, 2018; Aziz et al., 2019). They focus on computing the allocations to produce agents’ utilities satisfying certain established concepts of fairness, such as *envy-freeness* (Foley, 1967) — no agent should prefer another’s allocation to his own. It is noteworthy that this stream of researches usually considers fairness at an individual level. Besides, several studies have also addressed the problem of scarce resource fair allocation such as organ transplantation (Alagoz et al., 2009; Bertsimas et al., 2013; Zou et al., 2020) and social services (Zardari et al., 2010; Azizi et al., 2018). These works usually construct allocation models to maximize certain objectives (e.g., overall life years from transplant) while maintaining fairness and produce the participants’ rank ordering. Then, the resource is assigned to the participants on the waiting list according to their priority positions.

Different from works in resource allocation, we mainly focus on the group-based fairness notions in this paper, since group-based unfairness may have more deleterious consequences due to discrimination introducing additional implications for one’s group and one’s sense of rights and opportunities (Dover et al., 2015). Besides, we focus on the case of the classification model, which aims to eliminate the “misallocations of resources” (i.e., a person should not be misclassified). The prediction result for each individual is not affected by the others unlike in traditional resource allocation.

1.1.3 Fairness Research in Economics

Fairness has also been well-studied in several branches of economics. Usually the traditional economic models assume an agent is rational and self-interested. However, many field and experimental results show that the concern for fairness does effect people’s decisions (Kahneman et al., 1986; Babcock and Loewenstein, 1996; Charness and Rabin, 2002; Bandiera et al., 2005; Benjamin et al., 2010). The *social preference* branch of behavioral economics captures this phenomena and indicates that the self-interest must sometimes be appended to account for interdependent preferences such as fairness. Then, several researches in this field focus on developing economic models to better illustrate the non purely self-interested behaviors. For example, Becker (1974) formulates utility functions in a general way in two persons version for describing the deviations from self-interest. Then, Fehr and Schmidt (1999) propose utility functions of linear form for multiple agents, which incorporates the inequality aversion. They indicate agents may differ with respect to the disutility from inequality. At a close time, Bolton and Ockenfels (2000) consider the relative share instead of payoff difference for each individual in utility function. Charness and Rabin (2002) develop models that consider not only the distribution of outcomes but also the intentions of others’ decisions. Some works also provide direct neurobiological evidence in support of the existence of fairness considerations of social preferences in the human brain. Tricomi et al. (2010) shows that the brain’s reward circuitry is sensitive to both advantageous and disadvantageous inequality. On the other hand, many researchers in welfare economics also have long considered issues of fairness to be important in evaluating the desirability of different economic outcomes (Rabin, 1993). Note that in addition to studies of fair allocation in economic models, another branch in welfare economics deals with social welfare functions considering fairness requirements, such as Rawlsian (or max-min) (Rawls, 1999) fairness and α -fairness (Atkinson et al., 1970).

1.2 Contributions

We propose a paradigm to develop a fairness-aware scoring system to tackle the unfairness issues for the existing scoring systems in this article. We highlight our primary contributions as follows:

1.2.1 Problem Formulation

- Most in-processing fairness machine learning approaches directly modify standard machine learning models by adding fairness constraints. However, they usually assume the desired fairness level is pre-specified and show little clues for balancing the two goals: efficiency and equity (or accuracy and fairness as said in the field of machine learning). Unlike the previous researches, in this paper, we develop fairness-aware models through a welfare maximization

perspective. By constructing a social welfare function combining both equity and efficiency, we transfer the classical welfare maximization problem in economics into an empirical risk minimization framework in machine learning to derive a data-driven fairness-aware scoring system as well as the optimal fairness level. This approach explicitly considers the well-being of people affected by the system decisions and provides a scheme for achieving a better trade-off between efficiency and equity.

- Unlike most fairness machine learning methods applying surrogate loss to make optimization problems convex, this paper provides a way to utilize 0-1 hard loss to encode both the welfare objective as well as the fairness constraints and then produce a fair scoring system without the rounding procedure. This allows us to avoid the sub-optimality problem due to approximations. Furthermore, the proposed framework also affords the practitioners great flexibility to select the fairness criteria that they wish to enforce and customize their application requirements into a scoring system.

1.2.2 Theoretical Analysis

In this study, we drive several theoretical bounds on the total welfare of the proposed models. These results can provide useful and effective suggestions on the coefficients setting to achieve a better performance of derived scoring system. Besides, they also show an at-a-glance view of the relationship between fairness and total welfare.

1.2.3 Empirical Analysis

By evaluating the existing medical diagnosis scoring systems on a real data set, we unmasked the existence of unfairness between different genders, even though the disparate treatment is guaranteed. Afterward, we develop several novel scorecards (e.g., FASS and FASS7) for mortality prediction on Sepsis, which achieves the balance of fairness and efficiency. We present the performance of the proposed systems compared with existing medical scorecards on real Sepsis data. With the embedding of fairness considerations and achieving optimal welfare, the developed scorecards can capture risk-predictive rules and helps reveal the complexity of Sepsis by discovering promising interactions between these variables, which may give insights to further medical research and decision making. We also provide a detailed experimental comparison between our method and popular classification methods on several data sets. The results suggest that the proposed framework can produce scoring systems that achieve optimal social welfare and guarantee fairness with various fairness measures.

1.3 Organizations

The remainder of this paper is organized as follows. In Section 2, we motivate our research using a real-life example in medical diagnose to provide detailed illustrations regarding *disparate impact* and *disparate mistreatment* phenomena. Then, we demonstrate unfairness problems with the current medical scoring systems. In Section 3, we develop a general framework to construct a fairness-aware scoring system with the help of a social welfare function and mixed integer programming. We then present in Section 4 how different fairness measures are formulated and incorporated into our framework for different application contexts. Section 5 describes several theoretical bounds for the proposed method. In Section 6, the experimental study of our approach is carried out on several real data sets. All the technical proofs can be found in Appendix section.

2 Motivating Example

In this section, we first provide a real example of patient mortality prediction to illustrate different notions of fairness commonly used in machine learning literature. Then, we show that there still exists a disparity in the prediction results between people from different social groups for existing medical scoring systems. In such cases, there is an urgent need for developing a fairness-aware scoring system to aid human decision-making.

We consider a Sepsis mortality prediction example. Sepsis, as organ dysfunction caused by a dysregulated host response to infection, is a leading cause of mortality and morbidity worldwide (Sweeney et al., 2018). Thus, rapid identification and urgent treatment of Sepsis patients with a high risk of in-hospital death can significantly improve the outcome of Sepsis. To solve this problem, several scoring systems (e.g., SAPS II) have been applied to assess the illness severity of patients with Sepsis. These tools adopt patient vital signs, laboratory results, and demographic statistics as risk factors and output the severity assessments. However, their results might be biased against certain demographic groups. These biased predictions stem from the hidden or neglected biases in data for model construction or the algorithms themselves (Mehrabi et al., 2021), which finally leads to an unfair medical resources allocation among patients with different traits.

Figure 2 demonstrates three medical scoring systems with and without different unfair phenomena in this application on an example data set containing information of 6 patients. Here, the medical scoring systems are adopted to decide whether a patient is at a high mortality risk and needs urgent treatment based on a set of features, including some sensitive features like gender and other non-sensitive features like age and body temperature. The true label on whether a patient actually died is also shown. In what follows, we present different unfairness through the performance of three scoring systems D1, D2 and D3.

	Individual Features					True Label	System Prediction (Decision to Urgent Treatment)			Gender Gap				
#	Sensitive	Non-sensitive					Is Dead	D1	D2	D3	Fairness Notion	D1	D2	D3
	Gender	Age >75	Temp. >38	Arterial pH < 7.35									
1	M	1	1	1	1	1	0	1	SP	0	0	1/3	
2	M	1	0	0	0	1	1	1					
3	M	0	0	0	1	0	1	0	EO	0	1/2	1/2	
4	F	1	1	1	1	1	1	1					
5	F	1	1	0	1	0	1	1	OMR	0	2/3	1/3	
6	F	0	0	1	0	1	0	1					

Figure 2: Decisions of three medical scoring systems (denoted by D1, D2 and D3) for sepsis mortality prediction example.

Disparate Impact Disparate impact problem arises if a decision-making system outputs the results which benefit (or hurt) a group of people with the same values of sensitive features more frequently than other groups (Barocas and Selbst, 2016). Disparate impact elimination reflects the ability of a decision-making system to achieve *statistical parity* (Corbett-Davies et al., 2017) or also known as *demographic parity* (Dwork et al., 2012).

We assume that Sepsis patients benefit from a decision of urgent treatment since this indicates they will be allocated more medical resources. Under this assumption, we deem system D3 to be unfair due to disparate impact. As shown in Figure 2, the fraction of males and females that were predicted to be at high risk by D3 are different (2/3 and 1, respectively). In this case, there exists a treatment rate gap of 1/3 between two gender groups and thus, D3 does not satisfy *statistical parity*. In contrast, D1 and D2 guarantee the fairness with this notion because the treatment rate of males and females are the same and equals to 2/3 (i.e., treatment rate gaps are 0).

Disparate mistreatment Disparate mistreatment exists if a decision-making system achieves different misclassification error rates for groups of people with different values of sensitive features (Zafar et al., 2019). In addition to the overall error rate, this has been extended to different misclassifications such as false negatives and false positives. Here, we consider two commonly-used fairness notions in this category, namely *equal overall misclassification rate* (Zafar et al., 2017a) and *equality of opportunity* (Hardt et al., 2016). The former one eliminates disparate mistreatment by ensuring the same error rate among different groups, while the latter ensures the same false negative rate.

In Figure 2, only D1 is free from disparate mistreatment because it has the same false negative and overall error rates for two groups. On the other hand, both D2 and D3 are unfair due to disparate mistreatment since their rates of erroneous decisions for males and females are different. D2 and D3 both achieve different false negative rates ($1/2$ and 1) for males and females. D2 also has different error rates ($2/3$ and $1/3$) for males and females, whereas D3 has rates of $2/3$ and 0 .

The above example shows an intuitive illustration of different notions for algorithmic fairness. Now, we will showcase the existence of these unfairness phenomena in the existing medical scoring systems for real-life practice. In our focal context, we consider five commonly-used scoring systems for Sepsis mortality prediction: SAPS II, LODS, SOFA, qSOFA, and SIRS (Ribas et al., 2012; Sweeney et al., 2018). We evaluate their performance on fairness through a real Sepsis data set extracted from Medical Information Mart for Intensive Care database (MIMIC-III). The detailed information regarding the scoring systems and the data set will be presented in Section 6.1.

Table 1: Fairness checks of the existing medical scoring systems on Sepsis data set.

Scoring System	Accuracy	Fairness Level		
		SP	EO	OMR
SAPS II	0.7442	0.0064	0.1135	0.0235
LODS	0.7363	0.0323	0.0772	0.0111
SOFA	0.7209	0.0329	0.0351	0.0081
qSOFA	0.6304	0.0422	0.0530	0.0052
SIRS	0.5567	0.0011	0.1031	0.0409

Table 1 displays the results of fairness checks for these systems with respect to different fairness notions. Column (3)-(5) of Table 1 gives the absolute value of rate difference between two genders which is calculated based on different fairness notions, including the statistical parity (SP), equality of opportunity (EO), and equal overall misclassification rate (OMR). It demonstrates that there indeed exist disparities between males and females. In particular, the phenomenon of disparate mistreatment measured by EO is most evident. The gap of true positive rate between two groups can reach up to 11.35% (achieved by SAPS II). This indicates that more than 10% more male sepsis patients with a high mortality risk are ignored at the population level than females. In comparison, the disparate mistreatment by OMR and disparate impact by SP are less severe. However, the absolute values of rate difference for them are still up to 4.22% (by qSOFA) and 4.09% (by qSOFA), respectively.

Note that all scoring systems we checked above do not use the sensitive attribute (i.e., gender) as an input. Even if in this case, their prediction results can still be biased against people with specific sensitive attribute values. This will lead to the problem of unfair allocation of medical resources among patients from different groups. Lots of literature attests this bias against certain groups (e.g., the minority) does exist in many healthcare delivery fields. For example, several empirical studies show black patients are less likely to be selected for organ transplants, survive cardiac episodes, or receive high-cost lifesaving procedures, etc., compared to white patients (Becker et al., 1993; Jha et al., 2005; Rubineau and Kang, 2012; Ganju et al., 2020). Thus, developing a scoring system that considers fairness and then achieves global optimal decision-making is becoming crucial in these applications. Our approach will address this issue by constructing a general framework of fairness-aware scoring system. Then, we will show subsequently that it leads to a better overall welfare performance than existing techniques.

3 Model Formulation

In this section, we first formalize a utilitarian social welfare function. After that, we translate a social welfare maximization task into a constrained loss minimization problem at the center of supervised learning and then develop a scoring system achieving the optimal social welfare.

Suppose that $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denotes a data set with n i.i.d. observations, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^{d+1}$ is the i th individual’s feature vector with the form of $[1, x_{i,1}, \dots, x_{i,d}]^T$ and $y_i \in \mathcal{Y} = \{-1, 1\}$ is the i th’s class label. Moreover, s_i represents the sensitive feature (e.g., race, gender) of individual i with $s_i \in \mathcal{S} = \{a_1, a_2, \dots, a_c\}$, and thus there are c subgroups in population according to the sensitive feature. Then, in this research, we focus on the group-based unfairness which considers the disparities among members of different social groups featured by s_i . We note that \mathbf{x}_i may further contain or not the sensitive feature s_i in it for real applications. Here, we assume s_i is contained in the feature vector \mathbf{x}_i for simplicity. Although individuals in real life might be coded with multiple sensitive class features, we will consider only a single sensitive feature of focus in this work. However, we will show that the proposed framework will be easily extended to the case where more than one sensitive feature needs to be considered. In this work, we focus on the linear models of the form $\hat{y} = \text{sign}[\mathbf{w}^T \mathbf{x}]$, where $\mathbf{w} = [w_0, w_1, \dots, w_d]^T \in \mathcal{W}$ represents a vector of coefficients and w_0 is an intercept term.

Traditional economics has a long history of using models of *homo economicus*. That is, they assume that an individual is entirely rational and cares about his own payoffs but is indifferent about

the outcomes of others (Cox and Sadiraj, 2012). However, there is a large body of experimental and field evidence showing that lots of people in reality are not narrow self-interest (Kahneman et al., 1986; Babcock and Loewenstein, 1996; Charness and Rabin, 2002; Bandiera et al., 2005; Benjamin et al., 2010). They also concern about others and are driven by fairness considerations (Fehr and Schmidt, 1999; Dawes et al., 2007; Tricomi et al., 2010). Thus, we assume that an individual receives utilities not only depended on the system outcome regulated by \mathbf{w} but also on its level of (un)fairness δ ¹, which is represented as $U_i(\mathbf{w}, \delta)$. Then, we assume a decision maker wishes to maximize the following social welfare function given as a weighted sum of individual utilities,

$$SWF(\mathbf{w}, \delta) = \sum_{i=1}^n \zeta_i U_i(\mathbf{w}, \delta), \quad (1)$$

where $\zeta_i \in [0, 1]$ is the social weight that represents the value placed by society on i th individual's welfare and is normalized so that $\sum_{i=1}^n \zeta_i = 1$.

Normally, *social preferences* capture such departures from narrow self-interest. They refer to the phenomena that people seem to care about certain social goals, such as fairness in addition to their own benefits (Li, 2008). Their distributional preferences models usually assume that people are self-interested but are also concerned about the inequity between theirs and others' outcomes, and an individual's utility function is a linear combination of these two parts (Fehr and Schmidt, 1999; Charness and Rabin, 1999, 2002). Inspired by the above core idea, we consider a setting where the utility of an individual i is additively separable into data utility and fairness utility:

$$U_i(\mathbf{w}, \delta) = u_i(\mathbf{w}) + v_i(\delta), \quad (2)$$

where u_i is data utility that reflects an individual i 's own payoff owing to system output, and v_i is fairness utility that represents the influence of (un)fairness on i . For simplicity, let us consider the following case with linear utility functions:

$$u_i(\mathbf{w}) = a_i - b_i \mathbb{1} [y_i \mathbf{w}^T \mathbf{x}_i \leq 0], \quad (3)$$

$$v_i(\delta) = -\rho_i \delta, \quad (4)$$

where both $a_i \geq 0$ and $b_i \geq 0$.

Let us discuss the Eqs. (3) and (4) for deeply understanding this framework.

- **Data Utility:** We first introduce individual data utility. For an individual i , his data utility depends on whether the system classifies him correctly or not, and we assume misclassification

¹Here, δ actually shows the maximal level of unfairness of the system outcomes. Thus, a smaller value of δ implies a higher fairness level.

usually reduces a person's data utility. Take the medical diagnose as an example, misdiagnosing a patient may lead to premature death due to a lack of appropriate treatment. On the other hand, incorrectly diagnosing a healthy person will also cost his time and money to take unnecessary examinations and treatments, or even brings him side effects from invasive care. Specifically, as shown in (3), if i is correctly classified, he will receive the positive data utility with a_i . If i is misclassified, the data utility decreases to $a_i - b_i$.

- **Fairness Utility:** Next, we specify individual fairness utility. In equation (4), ρ_i is the individual preference weight placed on the fairness. It reflects a person's attitude towards unfair phenomena among groups. When $\rho_i = 0$ for all i , it falls on the basic assumption of purely self-interested homo economicus as in traditional economic models where people only care about their own outcomes. When i is an individual of advantageous group, if $\rho_i > 0$, it is a weight reflects his "kindness" towards the less advantageous. In this case, i would like to sacrifice his own data utility to help others even though i is in a favorable position. On the contrary, $\rho_i < 0$ represents i 's sense of "competition". It means i prefers a bigger gap among groups. In the case where i belongs to the less advantageous group, if $\rho_i > 0$, it is a weight that reflects his "hostility" towards the advantageous. In other words, i is unwilling to see the inequity among outcomes and hopes to reduce the gap among groups. If $\rho_i < 0$, it shows the "generosity" of i . That means even though i is in a less favorable position, he is willing to see the advantageous to get more benefits.

Afterward, combining (1)-(4) leads to

$$SWF(\mathbf{w}, \delta) = \sum_{i=1}^n \zeta_i \left[a_i - b_i \mathbb{1} [y_i \mathbf{w}^T \mathbf{x}_i \leq 0] - \rho_i \delta \right] \quad (5)$$

$$= \sum_{i=1}^n \zeta_i a_i - \sum_{i=1}^n \zeta_i b_i \mathbb{1} [y_i \mathbf{w}^T \mathbf{x}_i \leq 0] - \delta \sum_{i=1}^n \zeta_i \rho_i, \quad (6)$$

where δ denotes the unfairness level achieved by the classifier \mathbf{w} . We further apply the utilitarian social welfare function in which all people are treated the same and social weights are equal across all individuals: $\zeta_i = \frac{1}{n}$ for all i . Hence, finding a classifier to maximize social welfare in this case is equivalent to solving the following optimization problem

$$\min_{\mathbf{w}, \delta} \quad \frac{1}{n} \sum_{i=1}^n b_i \mathbb{1} [y_i \mathbf{w}^T \mathbf{x}_i \leq 0] + \frac{1}{n} \delta \sum_{i=1}^n \rho_i, \quad (7)$$

$$\text{s.t.} \quad g(\mathbf{w}, \mathcal{D}) \leq \delta, \quad (8)$$

$$\mathbf{w} \in \mathcal{W},$$

where (8) is the fairness constraint with $g(\mathbf{w}, \mathcal{D})$ encoding a specific fairness measure, δ is the maximal unfairness level to be tolerated, and \mathcal{W} encodes hard qualities that must be satisfied by the coefficients. It is noteworthy that in problem (7), the objective function consists of two parts. The first part is the average value of the weighted 0 – 1 loss which penalizes misclassification, and could be regarded as weighted error rate over data set. The second part reflects the “penalty” for unfairness. Thus, the above optimization problem could be adapted into a regularized empirical risk minimization (ERM) framework in the machine learning community as follows:

$$\min_{\mathbf{w}, \delta} \quad \frac{1}{n} \sum_{i=1}^n b_i \mathbb{1} [y_i \mathbf{w}^T \mathbf{x}_i \leq 0] + \bar{\rho} \delta + \lambda_0 \|\mathbf{w}\|_0 + \epsilon \|\mathbf{w}\|_1, \quad (9)$$

$$\begin{aligned} \text{s.t.} \quad & g(\mathbf{w}, \mathcal{D}) \leq \delta, \\ & \mathbf{w} \in \mathcal{W}, \end{aligned} \quad (10)$$

where $\bar{\rho} = \frac{\sum_{i=1}^n \rho_i}{n}$ is the average preference for fairness in the population. Usually, the constraints (10) restrict coefficients to a finite set of discrete values such as $\mathcal{W} = \{-10, \dots, 10\}^{d+1}$ to output an integer score. In addition to the original objective in problem (7), two more penalties are added into the problem (9). Specifically, ℓ_0 -penalty is applied to control the sparsity of the model where $\|\mathbf{w}\|_0 = \sum_{j=1}^d \mathbb{1} [w_j \neq 0]$ is the number of non-zero coefficients. The classifier tends to include more coefficients if its weight λ_0 becomes bigger. The ℓ_1 -penalty in the objective is used to obtain the coprime coefficients to reduce redundancy, and the ℓ_1 -penalty parameter ϵ should be set small enough to avoid ℓ_1 -regularization.

Now, we have cast the social welfare maximization problem prevalent in economics as a constrained loss minimization problem and adapted it into a regularized ERM framework to derive an optimal scoring system. In the following, we show that the proposed framework could degenerate to some common-used classification approaches in the machine learning field with some choices of model parameters.

- **Degeneration to Classification Model for Specified δ**

Now let us consider a case where the value of δ is pre-specified, for example, by laws or standards as δ^s . In this situation, the optimization problem (9) is equivalent to

$$\min_{\mathbf{w}} \quad \frac{1}{N} \sum_{i=1}^n b_i \mathbb{1} [y_i \mathbf{w}^T \mathbf{x}_i \leq 0] + \lambda_0 \|\mathbf{w}\|_0 + \epsilon \|\mathbf{w}\|_1 \quad (11)$$

$$\begin{aligned} \text{s.t.} \quad & g(\mathbf{w}, \mathcal{D}) \leq \delta^s, \\ & \mathbf{w} \in \mathcal{W}. \end{aligned} \quad (12)$$

This will train a fairness-guarantee classifier following most of the existing algorithmic or in-process approaches, which mainly aim at solving a constrained optimization problem by imposing a constraint on a certain level of fairness while optimizing the accuracy (Zafar et al., 2017b; Donini et al., 2018; Zafar et al., 2019). Unlike the existing approaches, framework (11) directly optimizes the (example-weighted) error rate by 0-1 loss as well as the model sparsity without making approximations that other methods make for scalability. As a result, it avoids these approximations and will normally achieve better classification performance while guaranteeing fairness.

- **Degeneration to Classic Classification Model**

Especially with a proper choice of the value of δ^s (e.g., $\delta^s = 1$) such that the constraint (12) ceases to bind, (11) could further degenerate to the ordinary classification model which only focuses on the (example weighted) accuracy (it could be viewed as maximizing only the total data utility from the decision maker’s perspective):

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{N} \sum_{i=1}^n b_i \mathbb{1} [y_i \mathbf{w}^T \mathbf{x}_i \leq 0] + \lambda_0 \|\mathbf{w}\|_0 + \epsilon \|\mathbf{w}\|_1 \\ \text{s.t.} \quad & \mathbf{w} \in \mathcal{W}. \end{aligned} \tag{13}$$

Note that if the heterogeneity of data utility preference b_i is further ignored, then (13) degrades to the classic ERM framework as in Ustun and Rudin (2016), which directly applies 0 – 1 loss instead of convex surrogate functions. This will produce scoring systems that are robust to outliers and attain the learning-theoretic guarantee on predictive accuracy (Brooks, 2011; Nguyen and Sanner, 2013; Ustun and Rudin, 2016).

3.1 Formulation of Mixed Integer Programming

Unfortunately, solving the problem given by (9) is very challenging. The indicator functions in (9) are non-continuous and non-convex functions of the classifier coefficient \mathbf{w} , therefore leading to non-convex formulations, which are difficult to solve directly. To figure this out, we reformulate the problem (9) into the following mixed integer programming task to recover the optimal fairness-aware

scoring systems:

$$\begin{aligned}
& \min_{\mathbf{w}, \psi, \Phi, \alpha, \beta, \delta} \frac{1}{n} \sum_{i=1}^n b_i \psi_i + \sum_{j=1}^d \Phi_j + \bar{\rho} \delta, \\
& \text{s.t.} \quad M_i \psi_i \geq \gamma - \sum_{j=0}^d y_i w_j x_{i,j} \quad i = 1, \dots, N \quad \text{0-1 loss,} \quad (14a) \\
& \quad G(\psi, \mathcal{D}_p, \mathcal{D}_q) \leq \delta \quad p, q = 1, \dots, c \quad \text{fairness,} \quad (14b) \\
& \quad \Phi_j = \lambda_0 \alpha_j + \epsilon \beta_j \quad j = 1, \dots, d \quad \text{coef. penalty,} \quad (14c) \\
& \quad -\Omega_j \alpha_j \leq w_j \leq \Omega_j \alpha_j \quad j = 1, \dots, d \quad \ell_0\text{-norm,} \quad (14d) \\
& \quad -\beta_j \leq w_j \leq \beta_j \quad j = 1, \dots, d \quad \ell_1\text{-norm,} \quad (14e) \\
& \quad w_j \in \mathcal{W}_j \quad j = 0, \dots, d \quad \text{coefficient set,} \\
& \quad \psi_i \in \{0, 1\} \quad i = 1, \dots, N \quad \text{loss variables,} \\
& \quad \Phi_j \in \mathbb{R}_+ \quad j = 1, \dots, d \quad \text{penalty variables,} \\
& \quad \alpha_j \in \{0, 1\} \quad j = 1, \dots, d \quad \ell_0 \text{ variables,} \quad (14f) \\
& \quad \beta_j \in \mathbb{R}_+ \quad j = 1, \dots, d \quad \ell_1 \text{ variables,} \\
& \quad \delta \in [0, 1] \quad \text{(un)fairness level.}
\end{aligned}$$

Here, $\mathcal{D}_p = \{(\mathbf{x}_i, y_i)\}_{s_i=a_p}$ and $\mathcal{D}_q = \{(\mathbf{x}_i, y_i)\}_{s_i=a_q}$ are individuals from any two different groups p and q , respectively.

In this formulation, constraint set (14a) uses Big-M constraints for 0-1 loss to set the loss variables $\psi_i = \mathbb{1}[y_i \mathbf{w}^T \mathbf{x}_i \leq 0]$ to 1 if the i th example is misclassified by the classifier \mathbf{w} . The Big-M constant (Wolsey, 1998) M_i can be set as $M_i = \max_{\mathbf{w} \in \mathcal{W}} (\gamma - y_i \mathbf{w}^T \mathbf{x}_i)$, and its computation is simple since \mathbf{w} is restricted to a finite set. The value of γ could be set to a small positive number which is not greater than a lower bound on $|y_i \mathbf{w}^T \mathbf{x}_i|$ (i.e. $0 < \gamma \leq \min_i |y_i \mathbf{w}^T \mathbf{x}_i|$). When the features are binary, γ can be set to any value between 0 and 1 since the coefficients are all integers. In other cases, γ might be set arbitrarily according to an implicit assumption on the values of the features (Ustun and Rudin, 2016; Zeng et al., 2017). With this setting on hand, if example i is misclassified, the value of right-hand side of the inequality (14a) is positive. Thus, ψ_i has to be 1 to satisfy the inequality. On the contrary, if i is classified correctly, we have $\gamma - \sum_{j=0}^d y_i w_j x_{i,j} \leq 0$. In this case, the value of ψ_i could be 0 or 1. However, since the bigger value of ψ_i results in more penalty in the objective, ψ_i will be forced to equal to 0 in this case. Therefore, ψ_i will work as an indicator to show whether the i th example is misclassified or not.

To evaluate the unfairness level achieved by \mathbf{w} in classification settings, we will focus on several generally-known fairness measures proposed via the machine learning community, which are

calculated over sub-population groups. Constraint set (14b) encodes the fairness assessment as inequalities among any two different groups p and q in society. Its explicit expressions $G(\cdot)$ depends on the given fairness notion and will be presented detailedly in Section 4. Besides, constraint set (14c) represents the total penalty assigned to each coefficient, where $\alpha_j = \mathbb{1}[w_j \neq 0]$ defined by (14d) encodes the ℓ_0 -penalty and $\beta_j = |w_j|$ defined by (14e) encodes the ℓ_1 -penalty. In (14d), $\Omega_j = \max_{w_j \in \mathcal{W}_j} |w_j|$ is defined as the largest absolute value of each coefficient.

Although the proposed framework mainly focuses on the considerations regarding fairness, it also allows people to implement a variety of operational constraints into its mixed integer programming formulation. Remark 1 shows some examples of operational constraints that can be encoded into this method. Note that our framework could also handle multiple operational constraints at the same time. Thus, this framework provides decision makers with great flexibility for their model customization in a simple way.

Remark 1 *The mixed integer programming formulation ensures that several types of operational constraints could be implemented. We specify here some common choices for different applications.*

- 1) *Model Size Control: we could limit the number of input features with the help of the indicator variables α_j by adding the constraint: $A_l \leq \sum_{j=1}^d \alpha_j \leq A_u$, where A_l is the lower bound and A_u is the upper bound of the model size, respectively.*
- 2) *Logical Relationship: some logical structures such as “if-then” constraints to ensure that a classifier will contain features α_j and α_k only if it also contains the feature α_l . This could be encoded as $\alpha_j + \alpha_k \leq \alpha_l$.*
- 3) *Domain Knowledge: some established relationships between input features and the outcome could be pre-specified with sign constraints in this model. For example, if the feature j is a well-known factor for specific outcomes (e.g., excess body weight usually cause a higher risk of type 2 diabetes (Organization et al., 2015)), this positive or negative relationship could be set by adding $w_j > 0$ or $w_j < 0$, respectively.*
- 4) *Preference for Feature Selection: practitioners may have soft preferences between different features. This could be realized by adjusting the value of λ_0 for different features. For example, if we prefer feature j to feature k to some extent, we can express this requirement as $\lambda_{0,k} = \lambda_{0,j} + \Lambda$, where $\Lambda > 0$ shows the maximal additional social welfare loss we could tolerate for using feature j instead of feature k . In this way, the feature k will be used only if it brings additional welfare gain greater than Λ . This approach can also be used to deal with the problem of missing values in dataset (Ustun and Rudin, 2016).*

4 Fairness Constrains

In this paper, we focus on two of the most popular doctrines of fairness used in the machine learning literature: *disparate impact* and *disparate mistreatment* (Zafar et al., 2019).

In Section 2, we have already shown their intuitive illustrations through a Sepsis example. In the following, we will show how these notions are formalized mathematically to develop fairness metrics into a binary classification problem and how to incorporate them into the proposed framework for deriving various fairness-aware scoring systems.

4.1 No Disparate Impact

As mentioned previously, a decision-making system suffers from disparate impact if its outcomes disproportionately hurt (or benefit) people with certain sensitive attribute values (e.g., females, blacks). In the algorithmic decision making context, even though the procedural discrimination (i.e. the decisions of systems are (partly) based on a subject’s sensitive feature) is strictly prohibited by laws, algorithms or decision-making procedures satisfying this criteria frequently produce different outcomes across groups based on the sensitive attributes, thus resulting in disparate impact (Fu et al., 2021). In response, *statistical parity*, as one of the first fairness notions suggested in machine learning field, is developed. *Statistical parity* simply requires the independence of the sensitive feature s and the decision \hat{y} . In other words, the system decisions should achieve the same distributions across all demographic groups. Thinking of the event $\hat{y} = 1$ as “acceptance” in binary classification scenario, this notion requires the acceptance rate to be identical for all groups, i.e.,

$$P(\hat{y} = 1 \mid s = a_p) = P(\hat{y} = 1 \mid s = a_q)$$

for any two different groups p and q . Then, a property that a decision system satisfies *statistical parity* between two groups p and q up to bias δ could be expressed as:

$$\left| P(\hat{y} = 1 \mid s = a_p) - P(\hat{y} = 1 \mid s = a_q) \right| \leq \delta$$

for any $p, q = 1, \dots, c$.

Representing this inequality empirically leads to

$$\left| \frac{1}{N_{a_p}} \left[\sum_{i \in I_{a_p}^-} \mathbb{1}[y_i \mathbf{w}^T \mathbf{x}_i \leq 0] + N_{a_p}^+ - \sum_{i \in I_{a_p}^+} \mathbb{1}[y_i \mathbf{w}^T \mathbf{x}_i \leq 0] \right] - \frac{1}{N_{a_q}} \left[\sum_{i \in I_{a_q}^-} \mathbb{1}[y_i \mathbf{w}^T \mathbf{x}_i \leq 0] + N_{a_q}^+ - \sum_{i \in I_{a_q}^+} \mathbb{1}[y_i \mathbf{w}^T \mathbf{x}_i \leq 0] \right] \right| \leq \delta, \quad (15)$$

where $I_{a_p}^+ = \{i \in \{1, 2, \dots, n\} | s_i = a_p, y_i = 1\}$, $I_{a_p}^- = \{i \in \{1, 2, \dots, n\} | s_i = a_p, y_i = -1\}$, $N_{a_p} = |I_{a_p}^+ \cup I_{a_p}^-|$ and $N_{a_p}^+ = |I_{a_p}^+|$ for any $p = 1, \dots, c$.

For any two different groups $p, q = 1, \dots, c$, the left-hand side of (15) could be re-expressed by the indicator variables in Section 3.1 as follows:

$$G_{SP} = \left| \left(\frac{N_{a_p}^+}{N_{a_p}} - \frac{N_{a_q}^+}{N_{a_q}} \right) + \frac{1}{N_{a_p}} \left[\sum_{i \in I_{a_p}^-} \psi_i - \sum_{i \in I_{a_p}^+} \psi_i \right] - \frac{1}{N_{a_q}} \left[\sum_{i \in I_{a_q}^-} \psi_i - \sum_{i \in I_{a_q}^+} \psi_i \right] \right|.$$

Afterwards, we can rewrite the fairness constraint (15) for *statistical parity* as $G_{SP} \leq \delta$. Incorporating this inequality into (14b), then we can derive a fairness-aware scoring system based on this notion.

Note that *statistical parity* is well-suited to contexts such as employment or school admissions, where it may be desirable or required by laws or regulations for diversity or affirmative action (Chouldechova, 2017; Lohaus et al., 2020). In these situations, selecting individuals equally across racial, gender, or geographical groups might be necessary. Moreover, because this notion is independent of the target value y , it is also appealing in applications where there does not exist the ground-truth information for decisions or the historical decisions used for training are biased themselves and thus cannot be trusted (Zafar et al., 2019). Implementing *statistical parity* will aid the prevention of discrimination based on redundant encoding (Dwork et al., 2012). It may also benefit building up the reputation of the disadvantageous minority group in the long term (Hu and Chen, 2018). However, this fairness notion might be inadequate in some cases. When disproportionality is truly present and independent from a sensitive feature, enforcing *statistical parity* requires us to reject qualified candidates from one group and/or approve unqualified candidates from the other group. This risks introducing reverse discrimination against qualified individuals. In addition, since this notion ignores any possible correlation between y and s , it may reject the optimal classifier $\hat{y} = y$ when base rates are different (i.e. $P(y = 1 | s = a_p) \neq P(y = 1 | s = a_q)$).

4.2 No Disparate Mistreatment

An algorithmic decision-making system has disparate mistreatment if it achieves unequal misclassification error rate (or conversely, accuracy) for groups of people with different values of sensitive features. This notion has been also extended to different types of misclassifications such as false negatives and false positives. In this category, we consider two frequently-used fairness notions in machine learning community: *equal overall misclassification rate (OMR)* (Zafar et al., 2017a) and *equality of opportunity (EO)* (Hardt et al., 2016).

4.2.1 Equal Overall Misclassification Rate

This notion is also known as *accuracy parity* (Zhao and Gordon, 2019), which requires the error rate to be same among all groups. It can be expressed as

$$P(\hat{y} \neq y \mid s = a_p) = P(\hat{y} \neq y \mid s = a_q)$$

for any $p, q = 1, \dots, c$.

Then, an algorithmic decision-making system satisfies equal OMR between any two groups p and q up to bias δ could be expressed as follows:

$$\left| P(\hat{y} \neq y \mid s = a_p) - P(\hat{y} \neq y \mid s = a_q) \right| \leq \delta.$$

Rewriting this inequality with the indicator variables ψ_i used in the mixed integer programming formulation gives the fairness constraint for equal OMR as

$$G_{OMR} = \left| \frac{1}{N_{a_p}} \sum_{i \in I_{a_p}} \psi_i - \frac{1}{N_{a_q}} \sum_{i \in I_{a_q}} \psi_i \right| \leq \delta, \quad (16)$$

where $I_{a_p} = \{i \in \{1, 2, \dots, n\} \mid s_i = a_p\}$ for any $p = 1, \dots, c$.

4.2.2 Equality of Opportunity

The second fairness notion EO desires to ensure that the true positive rate of each sensitive group is same. i.e., $TP_p = P(\hat{y} = 1 \mid s = a_p, y = 1)$ is the same for $\forall p \in \{1, \dots, c\}$. We also slack this requirement with the maximal unfairness level to be tolerated (i.e., δ) and re-express it with ψ_i in a way similar to the previous notions. Then, the fairness constraint for EO is given by

$$G_{EO} = \left| \frac{1}{N_{a_p}^+} \sum_{i \in I_{a_p}^+} \psi_i - \frac{1}{N_{a_q}^+} \sum_{i \in I_{a_q}^+} \psi_i \right| \leq \delta$$

for any $p, q = 1, \dots, c$.

Remark 2 Some other kinds of fairness notions related to absence of disparate mistreatment such as predictive equality (Corbett-Davies et al., 2017) and equalized odds (Hardt et al., 2016) could be also implemented directly by constructing corresponding fairness constraint function $G(\cdot)$ with the help of ψ_i . In this paper, we mainly choice the aforementioned instants (SP, equal OMR, and EO) as the demonstrated fairness measurements. The others are left for interested readers.

Since no disparate mistreatment slightly relaxes the requirement that \hat{y} is independent of s , it will not rule out the prefect predictor $\hat{y} = y$ even when the base rates differ across groups. In the scenarios where ground truth information for decisions is accessible and reliable, it would be possible to distinguish disproportionality in decision outcomes among groups that result from candidates' qualifications as well as the discrimination against certain groups. Thus, it will effectively avoid reverse-discrimination and is widely discussed in healthcare (Rajkomar et al., 2018), criminal justice (Chouldechova, 2017), and credit fields (Lohaus et al., 2020). However, it may also be insufficient from certain perspectives. For example, Berk et al. (2021) argues in settings where a cost-weighted approach is required, *equal overall misclassification rate* might be inadequacy. Zhang et al. (2019) also suggests enforcing *equality of opportunity* can make the outcomes seem fairer in a short time but lead to undesirable results in the long run.

Now, we have derived the specific expressions of fairness constraints for different fairness notions and discussed their suitable application contexts. Applying them in the fairness constraint set (14b), we can develop the optimal scoring system that maximizing social welfare based on a specific fairness notion.

It's worth noting that even though the fairness constraint set (14b) is defined on a specific fairness notion, the proposed framework could be easily extended for satisfying multiple fairness notions simultaneously. In certain application scenarios, it might be desirable to evaluate the level of (un)fairness on more than one notion of fairness defined above (e.g., measure the (un)fairness on both disparate impact and disparate mistreatment). In this case, a desirable scoring system could be conducted by including the corresponding constraints simultaneously. Furthermore, the proposed framework can also incorporate fairness with respect to multiple sensitive features (e.g., race, gender, religion, disability) simultaneously by including constraints for each sensitive feature separately. This indicates the high flexibility of the proposed framework.

5 Theoretical Analysis

Some theoretical bounds of proposed scoring systems are presented in this section. First, we show that although a finite discrete set \mathcal{W} is used to construct the scoring system, the total social welfare of the proposed method is not worse than a baseline classifier with real-valued coefficients $\theta \in \mathbb{R}^{d+1}$. Afterward, we show the relationship between the maximum social welfare and the optimal (un)fairness level. This may provide us a quick approach to roughly forecast the range of the optimal fairness level for our scoring system. Note that all the technical proofs of these theorems are provided in the Appendix section.

In the first place, Theorem 1 indicates that we can always generate a finite discrete coefficients

set such that the social welfare of the proposed method with discrete coefficients is even better than the social welfare of a baseline linear classifier with real-valued coefficients. This phenomenon can be found in the experimental study (see Section 6).

Theorem 1 *Let $\boldsymbol{\theta} = [\theta_0, \theta_1, \dots, \theta_d]^T \in \mathbb{R}^{d+1}$ denote the real-valued coefficients of any linear classifier which is trained with a data set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and achieves a (un)fairness level δ with respect to a given fairness requirement $G(\cdot)$. Besides, denote $\eta_{(k)}$ as the value of the k^{th} smallest margin achieved by training examples. Especially when $k = 1$, $\eta_{(1)} = \min_i \frac{|\boldsymbol{\theta}^T \mathbf{x}_i|}{\|\boldsymbol{\theta}\|_2}$. Let $\mathcal{I}_{(k)} = \left\{ i \in \{1, 2, \dots, n\} \mid \frac{|\boldsymbol{\theta}^T \mathbf{x}_i|}{\|\boldsymbol{\theta}\|_2} < \eta_{(k)} \right\}$ denote the set of training examples whose margin is smaller than $\eta_{(k)}$, and $X_{(k)} = \max_{i \notin \mathcal{I}_{(k)}} \|\mathbf{x}_i\|_2$ denote the largest magnitude of training example $\mathbf{x}_i \in \mathcal{D}$ for $i \notin \mathcal{I}_{(k)}$.*

Fitting a linear classifier via the proposed framework (9) and restricting coefficients to $\mathcal{W} = \{-\Omega, \dots, \Omega\}^{d+1}$, then suppose that we obtain the coefficients $\mathbf{w}^ = [w_0^*, w_1^*, \dots, w_d^*]^T$. If the resolution parameter Ω satisfies*

$$\Omega > \frac{X_{(k)} \sqrt{d+1}}{2\eta_{(k)}},$$

then the difference of total social welfare between the two classifiers is bounded as

$$SWF(\mathbf{w}^*) - SWF(\boldsymbol{\theta}) \geq (1 - k) \max_{i \in \mathcal{I}_{(k)}} b_i - N\bar{\rho}\Delta_F(k), \quad (17)$$

where $\Delta_F(k)$ is a function of k , whose concrete expression depends on the given fairness definition.

Note that when $k = 1$, $\Delta_F(k)$ equals to zero for all three fairness notions. In this case, according to the proof of Theorem 1, the coefficient set \mathcal{W} contains a classifier with discrete coefficients \mathbf{w} that achieves the same classification results (and thus the same social welfare) as the baseline classifier with real coefficients $\boldsymbol{\theta}$. Since our linear classifier with \mathbf{w}^* is optimal over \mathcal{W} , it may attain a better social welfare than \mathbf{w} , and thus is better than the real-valued classifier $\boldsymbol{\theta}$.

Consequently, Theorem 1 shows that if we select an appropriate coefficient set \mathcal{W} , i.e., choose a resolution parameter Ω large enough, the proposed scoring system could achieve even larger total social welfare than real-valued classifiers. This may provide practitioners some clues for model parameters setting in practice.

Afterward, in Theorem 2, we establish the relationship between the optimal fairness level of the proposed scoring system and the social welfare it can achieve.

Theorem 2 *Let \mathbf{w}^* and δ^* denote the optimal classifier maximizing total social welfare and its corresponding fairness level, respectively. Then, it holds that*

$$SWF(\mathbf{w}^*) > \Delta^*(\delta^*) - \mathbb{1}[\bar{\rho} > 0] \bar{\rho}, \quad (18)$$

where Δ^* is a function of δ^* and its detailed expression depends on the selected fairness notion.

Theorem 2 gives the lower bound of the social welfare for the proposed scoring system, as a function of its optimal fairness level δ^* . The proof of Theorem 2 in Appendix B presents the concrete expression of Δ^* for all three fairness notions discussed previously. It is noteworthy that for these fairness notions, Δ^* is a linear function of δ^* . This enables rapid estimation of the optimal (un)fairness level for the scoring system. Specifically, we consider a case where practitioners want to rough estimate the maximal unfairness level that people may need to tolerate for a certain value of social welfare (e.g., SWF_1). By setting the left-hand side of inequality (18) equal to SWF_1 , they can simply deduce the upper bound of the corresponding unfairness level δ^* . This will provide them a quick overview of the system’s fairness performance, which may further allow them to make rapid adjustments to relevant policies or system information publishing.

6 Experimental Study

In this section, we present experiments with the proposed method on several real data sets. We show that our approach is effective in developing a fairness-aware scoring system that maximizes total welfare. Besides, we highlight the advantages of our framework in flexibility and interpretability.

6.1 Application to Sepsis Mortality Prediction

Sepsis is a common syndrome that has posed a significant threat to healthcare systems worldwide. It has become one of the leading causes of death and most costly conditions in Intensive Care Units (ICUs) of American hospitals (Angus et al., 2001; Torio and Andrews, 2013). The early and accurate prediction of clinical outcomes such as in-hospital mortality of Sepsis patients aids clinical decision-making through fast response to patients at greater risk (Mukherjee and Evans, 2017).

By the motivating example in Section 2, we show the existence of unfairness in Sepsis mortality prediction when employing existing medical scoring systems. In the following part, we will construct fairness-aware scoring systems for this task and compare them with the existing scorecards to demonstrate their effectiveness.

6.1.1 Data and Processing

In the following experiments, we test the efficacy of the proposed method on the Sepsis data extracted from the Medical Information Mart for Intensive Care database (MIMIC-III). This data repository has been widely used for medical model development and validation (Henry et al., 2015; Nemati et al., 2018). The data set includes 2021 patients with 19 variables (the worst value of each

variable within 24hr of ICU admission). The outcome of interest is in-hospital death: 1 indicates death and 0 otherwise.

We convert the raw variables in the Sepsis data set into rule-based binary-coded data where each column represents whether the attributes satisfy a specific rule. We directly refer to the 77 rules discovered and analyzed in Wu et al. (2021) where the Rulefit method (Friedman et al., 2008) is used to produce informative rules that are able to predict the outcome of interest: in-hospital death accurately. More details about the data extraction from the MIMIC-III database and data pre-processing by RuleFit are presented in the Appendix section.

6.1.2 Model Setting and Baselines

In this experiment, we train fairness-aware scoring systems (FASSs) for different fairness measures as mentioned in Section 4. To further show the interpretability of our approach, we also consider the FASSs model with an additional operational constraint that limits the model size (see Remark 1). Usually, humans can only handle a few cognitive entities at once (seven plus or minus two according to Miller (1956)). Thus, we set the model size to be at most 7 (denoted by FASS7) so that it could be explained and understood by medical practitioners in a short time.

For all our methods, the coefficient set is chosen as $\mathcal{W} = \{-10, \dots, 10\}^{d+1}$ and $a_i = b_i = 1$ for $\forall i = 1, \dots, n$ for simplicity. Note with this setting, the data utility is reduced to the accuracy. In addition, we set $\lambda_0 \in [7 \times 10^{-5}, 9 \times 10^{-4}]$ and $\epsilon = 0.01$ so that the proposed method will sacrifice little welfare for sparsity. The CPLEX 12.6.3 is employed to solve the final mixed integer programming.

The experiments also compare our approaches to several commonly used medical scoring systems for in-hospital mortality prediction of Sepsis patients in ICUs as discussed in Section 2. More specifically, we consider the following baseline scoring systems:

SAPS II: Simplified Acute Physiology Score (Le Gall et al., 1993) is developed based on medical data from 137 ICUs of 12 countries in Europe/North America. The scoring system measures 17 variables of ICU patients and assigns different points to different variables. A regression-based model is provided to convert a total score to a mortality probability.

LODS: Logistic Organ Dysfunction System (Le Gall et al., 1996) uses the same database as SAPS II for model development but aims to assess the dysfunction levels of 6 human organ systems among ICU patients. The total dysfunction score ranges

from 0 to 22 and can also be converted to a mortality probability by a logistic regression model.

SOFA: Developed through expert consensus, Sepsis-related Organ Failure Assessment (Vincent et al., 1996) measures the degree of organ failure and evaluates morbidity of septic patients. The total score ranges from 0 to 24, with 0 to 4 for each of the six organ systems. Although the score is not initially designed to predict patient survival, it has become a basic variable in many mortality prediction models due to a high correlation between organ failure and survival.

qSOFA: The quick SOFA (Singer et al., 2016) is a simplified and quick version of SOFA, often used at the bedside to identify patients with suspected infection who are at greater risk of bad clinical outcomes outside the ICU. It only consists of three criteria (1 point for each) about blood pressure, respiratory rate, and central nervous system status. A score \geq two is usually considered a Sepsis case and is associated with at least a threefold increase in in-hospital mortality.

SIRS: Systemic Inflammatory Response Syndrome (Bone et al., 1992) is actually not a scoring system for Sepsis mortality, but we include it in our experiments as it forms an essential part of the initial definition of Sepsis-a host's SIRS to infection. The manifestation by two or more of the four conditions of SIRS is considered to be a SIRS case. We manually assign 1 point to each condition and assume a score of > 3 to be a positive case.

The above scoring systems have been widely used and validated by medical centers and researchers worldwide (Minne et al., 2008; Arabi et al., 2003). All the systems can be used at ICU admission or 24hrs after admission for disease severity evaluation and mortality prediction. In general, higher scores indicate more severe health conditions and hence a higher risk of mortality. Our experiments assume a risk probability higher than 0.5 to be positive for prediction models like SAPS II, LODS and a score greater than 12, 2, and 3 for SOFA, qSOFA, and SIRS, respectively. Note that all the baseline scores are computed with the raw variables rather than the rule-based binary variables.

With this set-up, we randomly partition the Sepsis data into a training set (70%) and test set (30%) and repeat the partition randomly 5 times to evaluate the average performance of all models unless otherwise stated. After that, FASS scorecards with all of the data are produced to show the interpretability.

6.1.3 Results

- **Social Welfare Maximization**

We first compare the performance of our methods in terms of social welfare maximization with the baseline scoring systems mentioned above. Table 2 provides an overview of average values of social welfare for all scoring systems on Sepsis data set. This chart shows clearly that the proposed FASS model consistently achieves the optimal social welfare with all fairness metrics. Especially when fairness is measured by equal OMR, FASS can bring total social welfare gains up to 10.12% compared to SAPS II (the best one among baselines). For EO, our method increases total welfare by nearly 4% compared with the optimal baseline scoring system. As for SP, FASS has a 1.55% increase in welfare compared to the best baseline (SAPS II). However, it significantly dominates the second-best baseline with around 13% increase.

In summary, all these results indicate that the proposed scoring system performs effectively in achieving optimal social welfare with different fairness metrics. It outperforms the existing medical scoring systems in Sepsis mortality prediction.

Table 2: The average values of total social welfare for all scoring systems on Sepsis data set.

Dataset	Fairness	$\bar{\rho}$	Baselines					Ours	
	Notions		SAPS II	LODS	SOFA	qSOFA	SIRS	FASS	FASS7
Sepsis	Train set								
	SP	5	0.7024	0.5904	0.5540	0.3805	0.5015	0.7124	0.7060
	EO	0.2	0.7269	0.7217	0.7158	0.6167	0.5355	0.7665	0.7462
	OMR	5	0.6298	0.6601	0.6400	0.5871	0.3281	0.7442	0.7339
	Test set								
	SP	5	0.6761	0.5368	0.5638	0.5089	0.4520	0.6916	0.6869
	EO	0.2	0.7102	0.7198	0.7096	0.6278	0.5373	0.7550	0.7443
	OMR	5	0.5768	0.6393	0.5787	0.5509	0.4022	0.7405	0.7223

- **Interpretability and Flexibility**

To show interpretability, we also develop a FASS model whose model size is no greater than 7 (denoted as FASS7 for short) and present its results in the last column of Table 2. The table shows that FASS7 achieves just slightly lower welfare than FASS without this operational constraint. Nevertheless, it still outperforms all baselines for three fairness metrics, even though its model size is limited.

Figure 3 shows the final scorecards produced by our method on the whole Sepsis data set.

For brevity of illustration, we only discuss the EO scorecard here and leave out the unit of measurement for variables within the rules.

The EO scorecard identifies two risk-increasing rules (rules assigned to positive points) and five risk-decreasing rules (rules assigned to negative points). Each rule consists of two or three conditions, and each condition consists of a variable and a related cut-off value (e.g., 7.2 for pH.art in the first rule). If all the conditions are satisfied, a rule is endorsed, and we plus or minus the corresponding score. A higher total score indicates a greater risk of in-hospital death. We find that the risk tendency of these score factors/criteria from the proposed scorecards is in line with the results in the Rulefit prediction model in [Wu et al. \(2021\)](#).

The EO scorecard showed in Figure 3 is able to identify many important variables and informative cut-off values of variables associated with in-hospital death of Sepsis. For example, the most involved risk variables in the scorecard are FiO2 (Fraction of inspired oxygen) and pH.art (arterial pH) with cut-off values at around 0.8 (0.75,0.85) and 7.1 (7.2), respectively. FiO2 is routinely measured in ICUs to assess patient pulmonary function and the presence or severity degree of Sepsis-related respiratory dysfunction ([Santana et al., 2013](#)). The scorecard indicates a FiO2 level higher than 0.8 to be at greater risk, which is consistent with a piece of recent evidence: [Dahl et al. \(2015\)](#) find in their subjects that the relative risk of mortality is 2.1 in patients with an average $\text{FiO}_2 \geq 0.80$ as compared to patients with an average $\text{FiO}_2 \leq 0.40$. The identified 7.1 of arterial pH (below its normal range: 7.35-7.45) implies the potential presence of acidosis that leads to unfavorable outcomes for ICU patients 7.1 is also the recommended treatment threshold of acute metabolic acidosis in severe Sepsis and septic shock from the Survival Sepsis Campaign ([Dellinger et al., 2013](#)).

Other variables like GCS (Glasgow Coma Scale) at 9, age at 80, and potassium at 4.25 also play roles in the scorecard, and their effect on the prediction of ICU mortality has been demonstrated in ([Kurowski et al., 2016](#); [Martin-Loeches et al., 2019](#); [Solinger and Rothman, 2013](#); [Gogos et al., 2003](#)). For example, GCS reflects a patient's degree of disturbance of consciousness. A GCS below nine is well-acknowledged as severe disturbance and is associated with higher death risk ([Kurowski et al., 2016](#)).

Obviously, the scorecard identifies multiple cut-off values for some yet recognized variables, such as GCS at 5 and 7, bilirubin at 2.3 and 7.3, pH.art at 7.1 and 7.2. This may reveal possible and complicated interactions between these Sepsis-related variables, given that Sepsis is a rather complex syndrome with unclear pathology and multiple comorbidities ([Singer et al., 2016](#); [Iskander et al., 2013](#)). The interaction between GCS and bilirubin, as suggested in the

fifth rule, may affect the risk thresholds of both variables. Recent evidence like Wang et al. (2020) argues that the level of bilirubin correlates with mortality in patients with traumatic brain injury who always have lower GCS. Further, Sedlak and Snyder (2004); Marconi et al. (2018) point that a high bilirubin level sometimes confers various health benefits due to its antioxidant activity. For pH.art, the study of Kraut and Madias (2010) suggests that metabolic acidosis might be beneficial for oxygen delivery and metabolism. Thus a slight upward adjustment of pH cut-off from 7.1 to 7.2 may not always be harmful, as in the fourth rule.

In conclusion, the proposed scorecard can capture risk-predictive rules with meaningful thresholds for the informative variables within the rules. Besides, it also helps reveal the complexity of Sepsis by discovering promising interactions between these variables, which may give insights to further medical research and decision-making.

Note that for all fairness metrics, the developed scorecards satisfy the operational constraints regarding the model size. Sparsity and small integer coefficients in our models help practitioners make quick predictions without a calculator or a computer. Significantly, the proposed methods could also help them understand how the model works and how each input affects the final output. These transparency and interpretability benefits facilitate its adoption in a real-life decision process. It is also noteworthy that these results indicate that FASS could handle the operational constraints effectively. This demonstrates that the proposed framework provides decision makers a great deal of flexibility. It allows practitioners to customize their requirements into operational constraints and develop an application-specific scoring system. This shows the great advantages of our models for practical applications.

6.2 Other Numerical Experiments

In this section, we conduct several numerical experiments to compare the performance of our scoring systems to other popular classification models in machine learning.

6.2.1 Data Sets and Experimental Setup

We conduct numerical experiments with two real-life datasets from UCI Machine Learning Repository: The *Adult* income data set and the *German* credit data set (Dua and Graff, 2017).

The original *Adult* data set contains 48,842 observations with 14 features in total. Moreover, it has a binary class label which indicates whether an individual makes over 50,000 dollars a year. The original *German* data set contains 1,000 observations with 20 features, and a binary class label indicates whether a customer’s credit is good or not. We delete all data points with missing

PREDICT +1 IF SCORE > 0		
Factors	Scores	Patient's score
1. creatine > 1.2 sbp ≤ 99.8 mean.bp ≤ 58	5 points
2. GCS ≤ 9.65 FiO2 > 0.65	5 points
3. pH.art ≤ 7.2 age > 51.3	5 points
4. potassium.serum ≤ 4.25 bilirubin ≤ 7.3 GCS > 5.2	-2 points
5. hematocrit > 24.5 age ≤ 60.8 FiO2 ≤ 0.8	-3 Points
6. bilirubin ≤ 7.2 GCS > 5.1	-5 points
7. potassium.serum > 4.1 sbp > 94.1 MAP > 42	-9 points
ADD POINTS FROM ROWS 1-7	TOTAL SCORE =

(a) Statistical Parity

PREDICT +1 IF SCORE > 0		
Factors	Scores	Patient's score
1. pH.art ≤ 7.2 age > 45.7	5 points
2. albumin ≤ 2.8 FiO2 > 0.75 GCS ≤ 9.2	4 points
3. potassium.serum ≤ 5.05 heart rate ≤ 133 FiO2 ≤ 0.75	-1 points
4. bilirubin ≤ 2.3 pH.art ≤ 7.2 FiO2 ≤ 0.85	-1 points
5. potassium.serum ≤ 4.25 bilirubin ≤ 7.3 GCS > 5.2	-1 Points
6. sodium > 130 pH.art > 7.1 FiO2 ≤ 0.8	-2 points
7. temperature > 35.4 age ≤ 80.2 GCS > 7.5	-3 points
ADD POINTS FROM ROWS 1-7	TOTAL SCORE =

(b) Equality of Opportunity

PREDICT +1 IF SCORE >= 0		
Factors	Scores	Patient's score
1. potassium.serum > 4.25 MAP ≤ 59	1 points
2. pH.art ≤ 7.2 bilirubin > 2.3	1 points
3. pH.art ≤ 7.2 age > 45.7	1 points
4. pH.art > 7.1 sodium > 130 FiO2 ≤ 0.8	-1 points
5. temperature > 35.4 age ≤ 80.2 GCS > 7.5	-1 Points
ADD POINTS FROM ROWS 1-5	TOTAL SCORE =

(c) Equal Overall Misclassification Rate

Figure 3: Scorecards developed by FASS7 for Sepsis prediction. The training welfare of the SP scorecard and EO scorecard is 0.6896 and 0.7466, respectively. The scorecard derived for equal OMR achieves a training welfare of 0.7277.

Table 3: Summary of real UCI data sets

Data set	$N_{original}$	$d_{original}$	N	d	Sensitive feature
<i>Adult</i>	48,842	14	2,000	36	Gender (binary)
<i>German</i>	1,000	20	3,000	65	

values and processed each data set by binarizing all input features. Because both two data sets are very imbalanced (the prior positive rate is 24% for *Adult* and 30% for *German*), some sampling methods in imbalanced learning are applied to eliminate its impact on the results. We used random undersampling to *Adult* data set and SMOTENC (Chawla et al., 2002) to *German* data set to create balanced data sets for the purpose to compare the relative performance between classifiers. The final data sets as shown in Table 3, where the gender (with feature values: male and female) is used as a sensitive feature. In each experiment, we randomly partition the data into training set (70%) and test set (30%) and repeat the partition randomly 5 times to evaluate the average performance of models unless otherwise stated. As a comparison, 6 baseline scoring system and linear classifiers (Lasso, Ridge, Elastic Net, SVM, Huberized SVM and SLIM) are also conducted in all the experiments, and the hyperparameters are selected via 5-fold cross-validation. For the proposed fairness-aware scoring system, the coefficient set is chosen as $\mathcal{W} = \{-10, \dots, 10\}^{d+1}$ and $a_i = b_i = 1$ for all $i = 1, \dots, n$ for simplicity. In the case, the data utility is reduced to the accuracy. In addition, we set $\lambda_0 < \frac{1}{nd}$ and $\epsilon = 0.01$ so that the proposed method will not sacrifice the total welfare for sparsity. The CPLEX 12.6.3 is employed to solve the final mixed integer programming.

6.2.2 Results

• Classification Model for Specified δ^s

We start with a simpler case where the fairness level δ^s is pre-specified by decision makers in advance as discussed previously in (11). In this situation, the data utility optimizing scoring system which controls the fairness level is developed. Tables 4 and 5 contain the results of the experiments on two data sets.

In the *Adult* data set, we first experiment with baseline classifiers. As shown in Table 4, all baselines lead to competitive performance on overall accuracy. Especially, Ridge achieves the best accuracy and hence the maximum data utility on training and test sets, whose value is 0.8149 and 0.8067, respectively. However, these classifiers result in highly disparate impact and disparate mistreatment for male and female subgroups. Specifically, the disparate impact is exceptionally high since the minimum SP rate difference between two groups, achieved by SLIM, is 0.3724 on the training set and 0.3959 on the test set. Moreover, the disparate

Table 4: The average values of data utility and fairness level for all methods on *Adult* data set with $\delta^s = 0.05$.

Metric	Baselines						Ours		
	Ridge	Lasso	Elasticnet	SVM	Huberized SVM	SLIM	FASS-EO	FASS-OMR	FASS-SP
Train set									
Data utility	0.8149	0.8129	0.8141	0.8020	0.8040	0.8114	0.7986	0.7996	0.7681
EO rate	0.1976	0.2139	0.2042	0.2106	0.2458	0.1444	0.0389		
OMR rate	0.0715	0.0684	0.0703	0.0925	0.0815	0.0740		0.0440	
SP rate	0.4033	0.4098	0.4065	0.4548	0.4615	0.3724			0.0414
Test set									
Data utility	0.8067	0.8050	0.8047	0.7866	0.7876	0.7930	0.7927	0.7880	0.7593
EO rate	0.1897	0.1897	0.1941	0.1956	0.2180	0.1210	0.0495		
OMR rate	0.0828	0.0817	0.0822	0.1086	0.1073	0.0966		0.0599	
SP rate	0.4259	0.4258	0.4278	0.4771	0.4872	0.3959			0.0783

Table 5: The average values of data utility and fairness level for all methods on *German* data set with $\delta^s = 0.01$.

Metric	Baselines						Ours		
	Ridge	Lasso	Elasticnet	SVM	Huberized SVM	SLIM	FASS-EO	FASS-OMR	FASS-SP
Train set									
Data utility	0.8261	0.8251	0.8257	0.8232	0.8178	0.8161	0.7694	0.8035	0.7969
EO rate	0.0227	0.0249	0.0203	0.0242	0.0215	0.0371	0.0035		
OMR rate	0.0202	0.0210	0.0193	0.0208	0.0221	0.0171		0.0096	
SP rate	0.1494	0.1294	0.1280	0.1269	0.1272	0.1115			0.0076
Test set									
Data utility	0.8187	0.8196	0.8207	0.8144	0.8114	0.8076	0.7522	0.8002	0.7993
EO rate	0.0325	0.0372	0.0366	0.0341	0.0523	0.0480	0.0188		
OMR rate	0.0293	0.0346	0.0347	0.0321	0.0297	0.0427		0.0282	
SP rate	0.1624	0.1369	0.1384	0.1455	0.1488	0.1115			0.0334

mistreatment based on EO is also significant since the minimum EO fairness level, again achieved by SLIM, is 0.1444 (resp. 0.1210) on the training (resp. test) set. In comparison, the disparate mistreatment based on OMR is relatively milder, whose best level is 0.0684 and 0.0817 on training and test sets, respectively. To harness these disparities, we set $\delta^s = 0.05$ and develop our FASS scoring systems with different fairness notions. As can be seen from Table 4, all our methods sacrifice only a little accuracy to strictly limit the unfairness level less than 0.05 on the training set and significantly reduce the disparity levels on the test set. Especially for the disparate impact on the training set, the accuracy of FASS-SP is 0.7681. However, its disparity is significantly reduced from 0.3724 to 0.0414 compared to SLIM, and it achieves the best fairness performance on the test set. In addition, both FASS-EO and FASS-OMR lead to the best fairness levels for disparate mistreatment while maintaining high accuracy close to the baselines (it even outperforms two SVMs on the test set on accuracy besides fairness level).

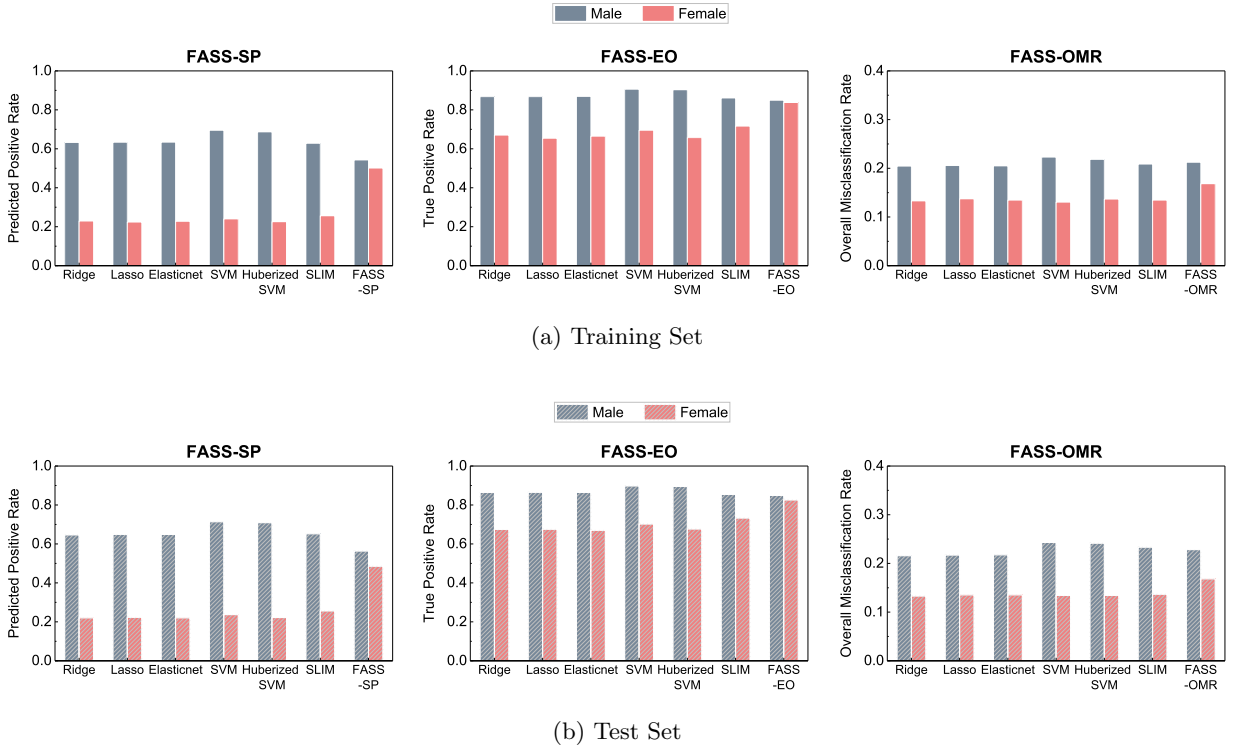


Figure 4: Illustration of disparate impact and disparate mistreatment for male and female subgroups on *Adult* data set.

The details regarding disparities between the two groups are presented in Figure 4. As can be seen from this figure, the results of all baseline classifiers indicate that the male group

significantly dominates the female group in all fairness notions. In comparison, the proposed framework achieves more similar rates among the two groups. For SP and EO, our methods reduce the predicted positive rate and the true positive rate for males and increase that for females. Although the overall misclassification rate of FASS-OMR in the male group has not decreased significantly, the one in the female group has been improved a lot. As a result, the rate gap between the two groups is greatly narrowed, and our scoring system achieves a better fairness level with respect to all fairness measures.

For the *German* data, the phenomenon of disparity is less severe compared to the *Adult* data, thus δ^s is set as 0.01 in this scenario. As shown in Table 5, the disparate impact, again, is the most significant since the minimum SP rate difference of baselines is 0.1115 by SLIM both on training and test sets. However, our approach can limit this value to $0.0076 < 0.01$ on the training set and reduce it to 0.0334 on the test set while guaranteeing a competitive accuracy. In contrast to the disparate impact, the disparate mistreatment issue is slight on *German* data, and our framework still outperforms the baselines on fairness levels for both EO and equal OMR notions.

Overall, the experimental results on the two UCI data sets show that the proposed methods can achieve a superb fairness level for all fairness notions, with the price of only a moderate loss in accuracy (data utility). This demonstrates the effectiveness of the proposed method when the maximal fairness level is specified in advance.

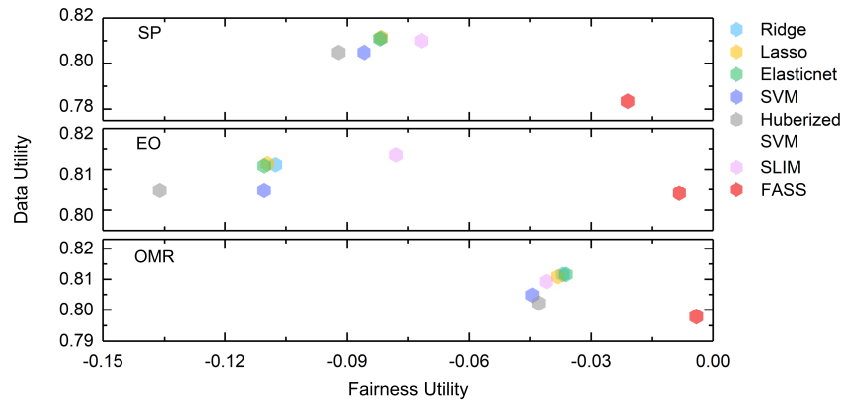
• Social Welfare Maximization

Next, we examine the effectiveness of the proposed methods in achieving the optimal social welfare as defined in Section 3. Note that under this condition, the goal is to develop a scoring system which maximizes the sum of data utility and fairness utility over population.

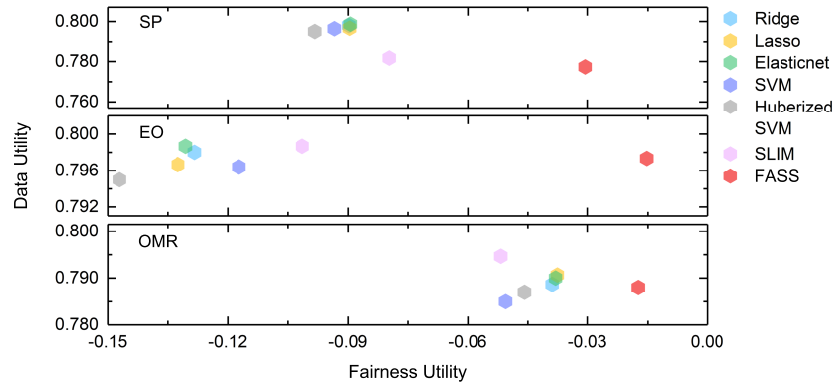
Table 6 summarizes the results we got from applying baseline models and our approaches when incorporating different fairness measures. These results clearly show that the proposed method works well and yields the maximum total welfare in all data sets. Figure 5 provides more insights regarding the trade-offs between data utility and fairness utility for welfare maximization on *Adult* data set. It can be seen from this figure that our framework attains competitive data utility compared to the baselines while gaining more fairness utility. Hence, the proposed fairness-aware scoring system improves the final social welfare significantly. Due to the lack of space, only *Adult* results are presented here. Complete graphs of *German* data set can be found in the Supplementary Material, which also achieves similar results as in Figure 5. Note that these results also provide evidence for our theoretical analysis in Theorem 1.

Table 6: The average values of total social welfare for all methods on UCI data sets.

Dataset	Fairness	$\bar{\rho}$	Baselines						Ours
	Notions		Ridge	Lasso	Elasticnet	SVM	Huberized SVM	SLIM	FASS
Adult	Train set								
	SP	0.2	0.7296	0.7299	0.7290	0.7190	0.7127	0.7384	0.7626
	EO	0.5	0.7036	0.7018	0.7005	0.6944	0.6686	0.7357	0.7959
	OMR	0.5	0.7747	0.7728	0.7754	0.7604	0.7594	0.7684	0.7938
	Test set								
	SP	0.2	0.7082	0.7070	0.7092	0.7030	0.6968	0.7020	0.7468
	EO	0.5	0.6698	0.6643	0.6682	0.6791	0.6479	0.6973	0.7821
	OMR	0.5	0.7499	0.7532	0.7521	0.7345	0.7412	0.7430	0.7707
	German	Train set							
SP		0.2	0.7961	0.7998	0.7996	0.7938	0.7938	0.7902	0.8105
EO		5	0.6818	0.6915	0.6845	0.6835	0.6904	0.6824	0.7604
OMR		5	0.6721	0.6875	0.6825	0.6818	0.6705	0.6680	0.7591
Test set									
SP		0.2	0.7886	0.7938	0.7965	0.7872	0.7876	0.7801	0.8089
EO		5	0.5980	0.6004	0.6009	0.6061	0.5994	0.4537	0.6659
OMR		5	0.6389	0.6565	0.6318	0.6534	0.6350	0.6057	0.7240



(a) Training Set



(b) Test Set

Figure 5: Trade-offs between data utility and fairness utility with different fairness measures on *Adult* data set.

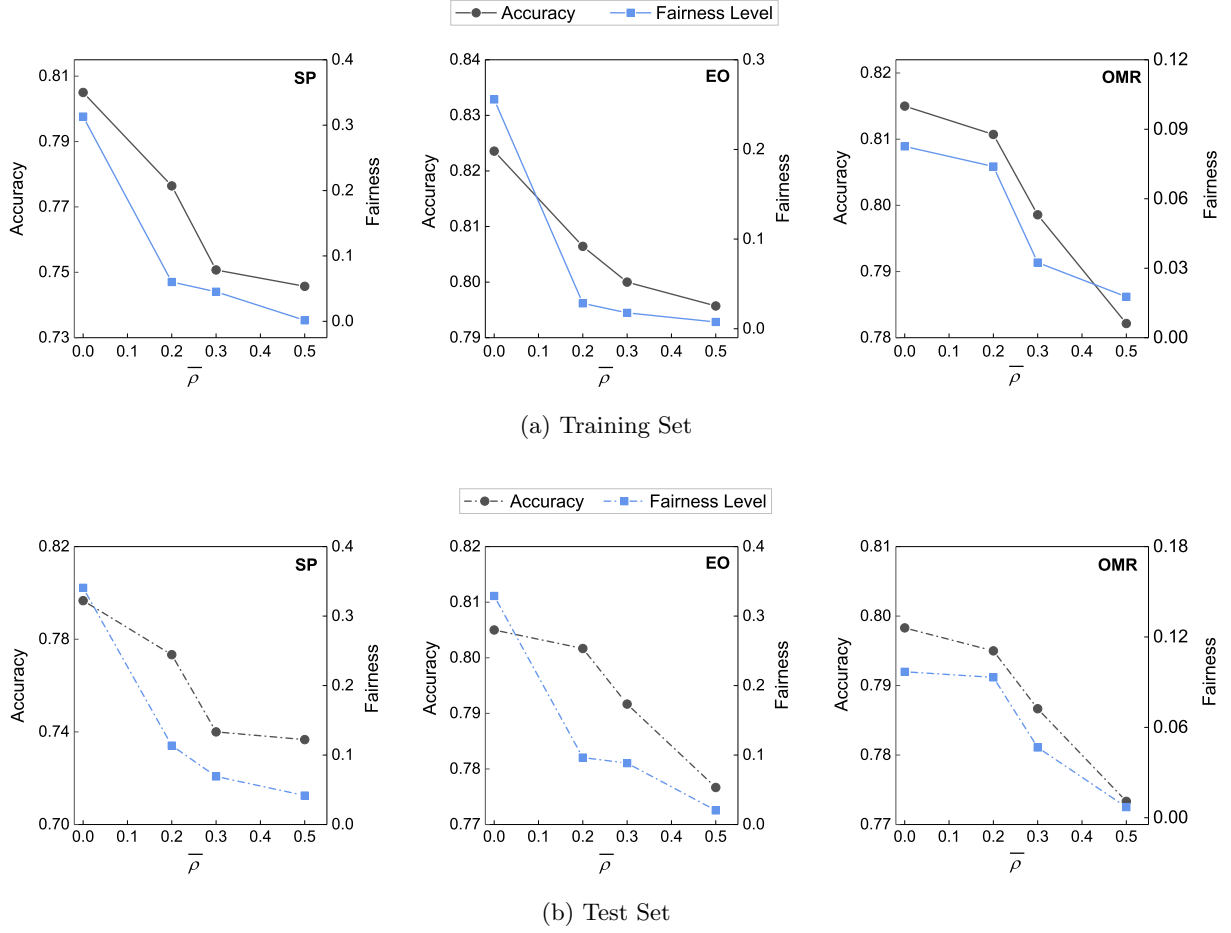


Figure 6: Trade-offs between data utility and fairness utility with different fairness measures on *Adult* data set in a randomly selected run.

- **Impact of Average Preference for Fairness**

In the previous study, we assume the average value of fairness preference is set arbitrarily. In this experiment, we study the impact of varying the average preference for fairness on the performance of the proposed system. We start by varying the value of $\bar{\rho}$ associated with the fairness level. For each value of $\bar{\rho}$, we report the accuracy, which reflects data utility, and the (un)fairness level δ between two groups. The resulting graphs for our approach can be found in Figure 6. It is apparent from this figure that both accuracy and (un)fairness level decreases as $\bar{\rho}$ increases in all fairness notions. Note that $\bar{\rho}$ controls the trade-off between accuracy and (un)fairness level. When $\bar{\rho}$ becomes larger, the scoring system produced by (7) would like to sacrifice more classification accuracy to attain a lower (un)fairness level since the latter one will bring more benefits for the objective function. Thus, as $\bar{\rho}$ increases, we transition from an unfair model with the best accuracy to a fairer model but with poorer accuracy. For the lack of space, the complete results of *German* data set are presented in the Supplementary Material, which shows a similar tendency as in Figure 6.

7 Conclusions

In this research, we proposed a general framework for developing data-driven scoring systems that incorporate fairness considerations. We first constructed a social welfare function that allows practitioners to balance efficiency and equity. Then, we transferred the welfare maximization problem into the classic empirical risk minimization framework to develop a fairness-aware scoring system. By using the 0-1 hard loss directly and mixed integer programming techniques, this framework allows practitioners to develop systems with respect to different fairness measures and also allows them to customize their own requirements by adding other operational constraints. Experiments on several real data sets confirm the interpretability and effectiveness of our approach.

There are several directions for future research.

- Construction of fairness-aware scoring systems on big data: If a huge data set is used for model training, the computation time might be concerned in some applications. Then, some data reduction algorithms or speed-up techniques commonly used in the mixed integer programming field could be incorporated into our framework to save time.
- Maintenance of fairness-aware scoring systems: This paper mainly focuses on developing a fairness-aware scoring system with a fixed-size data set. In the real world, however, historical data is fast growing with time. To redevelop an entirely new fair scorecard accordingly is not likely to be cost-effective. Thus, update the fair scorecard based on the existing one and new

coming data by online learning may be of interest since it potentially provides a way to save a great deal of time and money.

8 Appendix

A. Proof of Theorem 1

Proof 1 Applying the Theorem 1 of [Ustun and Rudin \(2016\)](#), it is easy to deduce that for a baseline classifier with real coefficients θ , there exists the coefficient set \mathcal{W} contains a classifier with discrete coefficients \mathbf{w} that assigns the exactly same label for any example i as the baseline classifier with θ . i.e., for all $i \in \{1, 2, \dots, n\}$, there exist $\mathbf{w} \in \mathcal{W}$ where $\mathcal{W} = \{-\Omega, \dots, \Omega\}$ with $\Omega > \frac{X_1 \sqrt{d+1}}{2\eta_{(1)}}$, we have $\mathbb{1}[y_i \mathbf{w}^T \mathbf{x}_i \leq 0] = \mathbb{1}[y_i \theta^T \mathbf{x}_i \leq 0]$.

We apply the above results to the reduced data set $D \setminus \mathcal{I}_{(k)}$, it follows that

$$\begin{aligned} & \sum_{i=1}^n b_i \mathbb{1}[y_i \mathbf{w}^T \mathbf{x}_i \leq 0] - \sum_{i=1}^n b_i \mathbb{1}[y_i \theta^T \mathbf{x}_i \leq 0] \\ &= \sum_{i \in \mathcal{I}_{(k)}} b_i \left\{ \mathbb{1}[y_i \mathbf{w}^T \mathbf{x}_i \leq 0] - \mathbb{1}[y_i \theta^T \mathbf{x}_i \leq 0] \right\} + \sum_{i \notin \mathcal{I}_{(k)}} b_i \left\{ \mathbb{1}[y_i \mathbf{w}^T \mathbf{x}_i \leq 0] - \mathbb{1}[y_i \theta^T \mathbf{x}_i \leq 0] \right\} \\ &= \sum_{i \in \mathcal{I}_{(k)}} b_i \left\{ \mathbb{1}[y_i \mathbf{w}^T \mathbf{x}_i \leq 0] - \mathbb{1}[y_i \theta^T \mathbf{x}_i \leq 0] \right\} \end{aligned} \quad (19)$$

$$\leq \sum_{i \in \mathcal{I}_{(k)}} b_i \quad (20)$$

$$\leq (k-1) \max_{i \in \mathcal{I}_{(k)}} b_i. \quad (21)$$

The equation (19) is due to the fact that the classifier with \mathbf{w} assigns the exactly same label as the classifier with θ for any example $i \in D \setminus \mathcal{I}_{(k)}$. The inequality in (20) results from the most extreme case where all examples in $\mathcal{I}_{(k)}$ are misclassified by \mathbf{w} but correctly classified by θ . The inequality in (21) follows from the fact that there exist $k-1$ elements in $\mathcal{I}_{(k)}$.

Next, we discuss the fairness level of the discrete linear classifier \mathbf{w} . Note that for convenience, we abuse notation somewhat and use $\psi_i(\mathbf{w}) = \mathbb{1}[y_i \mathbf{w}^T \mathbf{x}_i \leq 0]$ as in Section 3.1 to indicate whether an example i is misclassified or not by the classifier with coefficients \mathbf{w} . Now, we consider the following three fairness definitions mentioned in Section 4.

I. Equal Overall Misclassification Rate

We first consider the case where the equal overall misclassification rate is used to measure fairness level. Note that c is the number of different groups, and since the classifier with θ satisfies a given fairness requirement $G(\cdot)$ and δ , we have

$$G_{OMR}(\theta) = \left| \frac{1}{N_{a_p}} \sum_{i \in I_{a_p}} \psi_i(\theta) - \frac{1}{N_{a_q}} \sum_{i \in I_{a_q}} \psi_i(\theta) \right| \leq \delta \quad (22)$$

for any $p, q = 1, \dots, c$.

Recall that the classifier with \mathbf{w} assigns the exactly same label as the classifier with θ for any example $i \in D \setminus \mathcal{I}_{(k)}$. Thus, $\psi_i(\theta) = \psi_i(\mathbf{w})$ for all $i \in D \setminus \mathcal{I}_{(k)}$. Besides, we denote $z_p = \sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}} (\psi_i(\mathbf{w}) - \psi_i(\theta))$ for

$p = 1, 2, \dots, c$. Summing $|z_p|$ over p , it follows that

$$\begin{aligned} \sum_{p=1}^c |z_p| &= \sum_{p=1}^c \left| \sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}} \psi_i(\mathbf{w}) - \psi_i(\boldsymbol{\theta}) \right| \\ &\leq \sum_{p=1}^c \sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}} \left| \psi_i(\mathbf{w}) - \psi_i(\boldsymbol{\theta}) \right| \end{aligned} \quad (23)$$

$$\leq |\mathcal{I}_{(k)}| = k - 1. \quad (24)$$

Since $\sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}} \psi_i(\mathbf{w}) = z_p + \sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}} \psi_i(\boldsymbol{\theta})$, we have

$$\begin{aligned} G_{OMR}(\mathbf{w}) &= \left| \frac{1}{N_{a_p}} \sum_{i \in I_{a_p}} \psi_i(\mathbf{w}) - \frac{1}{N_{a_q}} \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}) \right| \\ &= \left| \frac{1}{N_{a_p}} \left(\sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}} \psi_i(\mathbf{w}) + \sum_{i \in I_{a_p} - \mathcal{I}_{(k)}} \psi_i(\mathbf{w}) \right) - \frac{1}{N_{a_q}} \left(\sum_{i \in \mathcal{I}_{(k)} \cap I_{a_q}} \psi_i(\mathbf{w}) + \sum_{i \in I_{a_q} - \mathcal{I}_{(k)}} \psi_i(\mathbf{w}) \right) \right| \\ &= \left| \frac{1}{N_{a_p}} \left(\sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}} \psi_i(\boldsymbol{\theta}) + z_p + \sum_{i \in I_{a_p} - \mathcal{I}_{(k)}} \psi_i(\boldsymbol{\theta}) \right) - \frac{1}{N_{a_q}} \left(\sum_{i \in \mathcal{I}_{(k)} \cap I_{a_q}} \psi_i(\boldsymbol{\theta}) + z_q + \sum_{i \in I_{a_q} - \mathcal{I}_{(k)}} \psi_i(\boldsymbol{\theta}) \right) \right| \\ &= \left| \frac{1}{N_{a_p}} \left(\sum_{i \in I_{a_p}} \psi_i(\boldsymbol{\theta}) + z_p \right) - \frac{1}{N_{a_q}} \left(\sum_{i \in I_{a_q}} \psi_i(\boldsymbol{\theta}) + z_q \right) \right| \\ &= \left| \frac{1}{N_{a_p}} \sum_{i \in I_{a_p}} \psi_i(\boldsymbol{\theta}) - \frac{1}{N_{a_q}} \sum_{i \in I_{a_q}} \psi_i(\boldsymbol{\theta}) + \frac{z_p}{N_{a_p}} - \frac{z_q}{N_{a_q}} \right| \\ &\leq \left| \frac{1}{N_{a_p}} \sum_{i \in I_{a_p}} \psi_i(\boldsymbol{\theta}) - \frac{1}{N_{a_q}} \sum_{i \in I_{a_q}} \psi_i(\boldsymbol{\theta}) \right| + \left| \frac{z_p}{N_{a_p}} - \frac{z_q}{N_{a_q}} \right| \\ &\leq \delta + \left| \frac{z_p}{N_{a_p}} - \frac{z_q}{N_{a_q}} \right| \\ &\leq \delta + (k - 1) \max \left\{ \frac{1}{N_{a_p}}, \frac{1}{N_{a_q}} \right\}, \end{aligned} \quad (25)$$

for any $p, q = 1, \dots, c$. The equation (25) follows from the fact that $\psi_i(\boldsymbol{\theta}) = \psi_i(\mathbf{w})$ for all $i \in D_N \setminus \mathcal{I}_{(k)}$, and the last inequality is due to the inequality (24). Hence, the maximal increment of the tolerance level of unfairness among all groups is $\Delta_F(k) = (k - 1) \max \left\{ \frac{1}{N_{a_1}}, \frac{1}{N_{a_2}}, \dots, \frac{1}{N_{a_c}} \right\}$ for \mathbf{w} .

II. Equality of Opportunity

Recall that if the equality of opportunity is used, we have

$$G_{EO}(\boldsymbol{\theta}) = \left| \frac{1}{N_{a_p}^+} \sum_{i \in I_{a_p}^+} \psi_i(\boldsymbol{\theta}) - \frac{1}{N_{a_q}^+} \sum_{i \in I_{a_q}^+} \psi_i(\boldsymbol{\theta}) \right| \leq \delta$$

for any $p, q = 1, \dots, c$. We denote $z_p^+ = \sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}^+} (\psi_i(\mathbf{w}) - \psi_i(\boldsymbol{\theta}))$ for $p = 1, 2, \dots, c$. Similar in Case I, it

follows that

$$\begin{aligned}
\sum_{p=1}^c |z_p^+| &= \sum_{p=1}^c \left| \sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}^+} \psi_i(\mathbf{w}) - \psi_i(\boldsymbol{\theta}) \right| \\
&\leq \sum_{p=1}^c \sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}^+} \left| \psi_i(\mathbf{w}) - \psi_i(\boldsymbol{\theta}) \right| \\
&\leq \sum_{p=1}^c \sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}^+} \left| \psi_i(\mathbf{w}) - \psi_i(\boldsymbol{\theta}) \right| \\
&\leq |\mathcal{I}_{(k)}| = k - 1.
\end{aligned}$$

Afterwards, for any two groups $p, q = 1, \dots, c$, we have

$$\begin{aligned}
G_{EO}(\mathbf{w}) &= \left| \frac{1}{N_{a_p}^+} \sum_{i \in I_{a_p}^+} \psi_i(\mathbf{w}) - \frac{1}{N_{a_q}^+} \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}) \right| \\
&= \left| \frac{1}{N_{a_p}^+} \left(\sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}^+} \psi_i(\mathbf{w}) + \sum_{i \in I_{a_p}^+ - \mathcal{I}_{(k)}} \psi_i(\mathbf{w}) \right) - \frac{1}{N_{a_q}^+} \left(\sum_{i \in \mathcal{I}_{(k)} \cap I_{a_q}^+} \psi_i(\mathbf{w}) + \sum_{i \in I_{a_q}^+ - \mathcal{I}_{(k)}} \psi_i(\mathbf{w}) \right) \right| \\
&= \left| \frac{1}{N_{a_p}^+} \left(\sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}^+} \psi_i(\boldsymbol{\theta}) + z_p^+ + \sum_{i \in I_{a_p}^+ - \mathcal{I}_{(k)}} \psi_i(\boldsymbol{\theta}) \right) - \frac{1}{N_{a_q}^+} \left(\sum_{i \in \mathcal{I}_{(k)} \cap I_{a_q}^+} \psi_i(\boldsymbol{\theta}) + z_q^+ + \sum_{i \in I_{a_q}^+ - \mathcal{I}_{(k)}} \psi_i(\boldsymbol{\theta}) \right) \right| \\
&= \left| \frac{1}{N_{a_p}^+} \left(\sum_{i \in I_{a_p}^+} \psi_i(\boldsymbol{\theta}) + z_p^+ \right) - \frac{1}{N_{a_q}^+} \left(\sum_{i \in I_{a_q}^+} \psi_i(\boldsymbol{\theta}) + z_q^+ \right) \right| \\
&\leq \left| \frac{1}{N_{a_p}^+} \sum_{i \in I_{a_p}^+} \psi_i(\boldsymbol{\theta}) - \frac{1}{N_{a_q}^+} \sum_{i \in I_{a_q}^+} \psi_i(\boldsymbol{\theta}) \right| + \left| \frac{z_p^+}{N_{a_p}^+} - \frac{z_q^+}{N_{a_q}^+} \right| \\
&\leq \delta + \left| \frac{z_p^+}{N_{a_p}^+} - \frac{z_q^+}{N_{a_q}^+} \right| \\
&\leq \delta + (k-1) \max \left\{ \frac{1}{N_{a_p}^+}, \frac{1}{N_{a_q}^+} \right\}.
\end{aligned}$$

Hence, for \mathbf{w} the maximal increment of the tolerance level of unfairness among all groups is

$$\Delta_F(k) = (k-1) \max \left\{ \frac{1}{N_{a_1}^+}, \frac{1}{N_{a_2}^+}, \dots, \frac{1}{N_{a_c}^+} \right\}.$$

III. Statistical Parity

Now, we consider a relative complex case where the given fairness notion is statistical parity. Recall that for any two groups $p, q = 1, \dots, c$, it follows that

$$G_{SP}(\boldsymbol{\theta}) = \left| \left(\frac{N_{a_p}^+}{N_{a_p}} - \frac{N_{a_q}^+}{N_{a_q}} \right) + \frac{1}{N_{a_p}} \left[\sum_{i \in I_{a_p}^-} \psi_i(\boldsymbol{\theta}) - \sum_{i \in I_{a_p}^+} \psi_i(\boldsymbol{\theta}) \right] - \frac{1}{N_{a_q}} \left[\sum_{i \in I_{a_q}^-} \psi_i(\boldsymbol{\theta}) - \sum_{i \in I_{a_q}^+} \psi_i(\boldsymbol{\theta}) \right] \right| \leq \delta.$$

Again, we denote $z_p^+ = \sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}^+} (\psi_i(\mathbf{w}) - \psi_i(\boldsymbol{\theta}))$ and $z_p^- = \sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}^-} (\psi_i(\mathbf{w}) - \psi_i(\boldsymbol{\theta}))$ for $p = 1, 2, \dots, c$.

Then, it is easy to deduce that

$$\begin{aligned}
\sum_{p=1}^c |z_p^+| + |z_p^-| &= \sum_{p=1}^c \left(\left| \sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}^+} (\psi_i(\mathbf{w}) - \psi_i(\boldsymbol{\theta})) \right| + \left| \sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}^-} (\psi_i(\mathbf{w}) - \psi_i(\boldsymbol{\theta})) \right| \right) \\
&\leq \sum_{p=1}^c \left(\sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}^+} |\psi_i(\mathbf{w}) - \psi_i(\boldsymbol{\theta})| + \sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}^-} |\psi_i(\mathbf{w}) - \psi_i(\boldsymbol{\theta})| \right) \\
&\leq \sum_{p=1}^c \left(\sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}^+} 1 + \sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}^-} 1 \right) \\
&= \sum_{p=1}^c \sum_{i \in \mathcal{I}_{(k)} \cap I_{a_p}} 1 \\
&= |\mathcal{I}_{(k)}| = k - 1.
\end{aligned} \tag{26}$$

Similar in the above two cases, the value of $G_{SP}(\mathbf{w})$ is upper bounded as

$$\begin{aligned}
G_{SP}(\mathbf{w}) &= \left| \left(\frac{N_{a_p}^+}{N_{a_p}} - \frac{N_{a_q}^+}{N_{a_q}} \right) + \frac{1}{N_{a_p}} \left[\sum_{i \in I_{a_p}^-} \psi_i(\mathbf{w}) - \sum_{i \in I_{a_p}^+} \psi_i(\mathbf{w}) \right] - \frac{1}{N_{a_q}} \left[\sum_{i \in I_{a_q}^-} \psi_i(\mathbf{w}) - \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}) \right] \right| \\
&= \left| \left(\frac{N_{a_p}^+}{N_{a_p}} - \frac{N_{a_q}^+}{N_{a_q}} \right) + \frac{1}{N_{a_p}} \left[\sum_{i \in I_{a_p}^-} \psi_i(\boldsymbol{\theta}) + z_p^- - \sum_{i \in I_{a_p}^+} \psi_i(\boldsymbol{\theta}) - z_p^+ \right] \right. \\
&\quad \left. - \frac{1}{N_{a_q}} \left[\sum_{i \in I_{a_q}^-} \psi_i(\boldsymbol{\theta}) + z_q^- - \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}) - z_q^+ \right] \right| \\
&= \left| \left(\frac{N_{a_p}^+}{N_{a_p}} - \frac{N_{a_q}^+}{N_{a_q}} \right) + \frac{1}{N_{a_p}} \left[\sum_{i \in I_{a_p}^-} \psi_i(\boldsymbol{\theta}) - \sum_{i \in I_{a_p}^+} \psi_i(\boldsymbol{\theta}) \right] - \frac{1}{N_{a_q}} \left[\sum_{i \in I_{a_q}^-} \psi_i(\boldsymbol{\theta}) - \sum_{i \in I_{a_q}^+} \psi_i(\boldsymbol{\theta}) \right] \right. \\
&\quad \left. + \left(\frac{z_p^- - z_p^+}{N_{a_p}} - \frac{z_q^- - z_q^+}{N_{a_q}} \right) \right| \\
&\leq \delta + \left| \frac{z_p^- - z_p^+}{N_{a_p}} - \frac{z_q^- - z_q^+}{N_{a_q}} \right| \\
&\leq \delta + (k - 1) \max \left\{ \frac{1}{N_{a_p}}, \frac{1}{N_{a_q}} \right\}
\end{aligned}$$

for any $p, q = 1, \dots, c$. The last inequality follow from the inequality (26). Therefore, the maximal increment of the tolerance level of unfairness among all groups is $\Delta_F(k) = (k - 1) \max \left\{ \frac{1}{N_{a_1}}, \frac{1}{N_{a_2}}, \dots, \frac{1}{N_{a_c}} \right\}$ for \mathbf{w} .

With $\Delta_F(k)$ in hand, we can then calculate the bound of social welfare difference between two classifiers. For the classifier with \mathbf{w} and the classifier with $\boldsymbol{\theta}$, we have

$$\begin{aligned}
SWF(\mathbf{w}) &= \sum_{i=1}^n a_i - \sum_{i=1}^n b_i \mathbb{1} [y_i \mathbf{w}^T \mathbf{x}_i \leq 0] - \delta_{\mathbf{w}} \sum_{i=1}^n \rho_i \\
SWF(\boldsymbol{\theta}) &= \sum_{i=1}^n a_i - \sum_{i=1}^n b_i \mathbb{1} [y_i \boldsymbol{\theta}^T \mathbf{x}_i \leq 0] - \delta_{\boldsymbol{\theta}} \sum_{i=1}^n \rho_i,
\end{aligned}$$

where $\delta_{\mathbf{w}}$ (resp. $\delta_{\boldsymbol{\theta}}$) is the tolerance level of unfairness that the classifier with \mathbf{w} (resp. $\boldsymbol{\theta}$) can achieve. Note that $\delta_{\boldsymbol{\theta}} = \delta$ according to the assumption. Then, we have

$$\begin{aligned}
SWF(\mathbf{w}) - SWF(\boldsymbol{\theta}) &= - \left[\sum_{i=1}^n b_i \mathbb{1} [y_i \mathbf{w}^T \mathbf{x}_i \leq 0] - \sum_{i=1}^n b_i \mathbb{1} [y_i \boldsymbol{\theta}^T \mathbf{x}_i \leq 0] \right] - N\bar{\rho}(\delta_{\mathbf{w}} - \delta_{\boldsymbol{\theta}}) \\
&\geq (1-k) \max_{i \in \mathcal{I}_{(k)}} b_i - N\bar{\rho}(\delta + \Delta_F(k) - \delta) \\
&= (1-k) \max_{i \in \mathcal{I}_{(k)}} b_i - N\bar{\rho}\Delta_F(k).
\end{aligned} \tag{27}$$

Recall that according to the definition of \mathbf{w}^* , it is the discrete classifier maximizing the social welfare. Based on the inequality (27), we have

$$\begin{aligned}
SWF(\mathbf{w}^*) - SWF(\boldsymbol{\theta}) &\geq SWF(\mathbf{w}) - SWF(\boldsymbol{\theta}) \\
&= (1-k) \max_{i \in \mathcal{I}_{(k)}} b_i - N\bar{\rho}\Delta_F(k).
\end{aligned} \tag{28}$$

B. Proof of Theorem 2

Proof 2 Let $\mathcal{V}(\mathbf{w}) = \sum_{i=1}^n v_i$ be the overall data utility of \mathbf{w} , and $\mathbf{w}^{data} = \operatorname{argmax}_{\mathbf{w} \in \mathcal{W}} \mathcal{V}(\mathbf{w})$ denote the classifier only focus on the data utility. We first prove

$$\mathcal{V}(\mathbf{w}^{data}) \geq \max \{ \Delta^*, \mathcal{V}(\mathbf{w}^*) \}, \tag{29}$$

where Δ^* is a parameter related to δ^* and the corresponding fairness definition. According to the definition of \mathbf{w}^{data} , we have $\mathcal{V}(\mathbf{w}^{data}) \geq \mathcal{V}(\mathbf{w}^*)$ directly.

I. Equal Overall Misclassification Rate

First, we consider an classifier $\mathbf{w}_{uf} \in \mathcal{W}$ that dose not satisfy the equal overall misclassification rate requirement with δ^* for all groups. In other words,

$$G_{OMR}(\mathbf{w}_{uf}) = \left| \frac{1}{N_{a_p}} \sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) - \frac{1}{N_{a_q}} \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) \right| > \delta^* \tag{30}$$

for any $p, q = 1, \dots, c$ and $p \neq q$. From the above equation, it follows that

$$\frac{1}{N_{a_p}} \sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) - \frac{1}{N_{a_q}} \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) < -\delta^* \tag{31}$$

or

$$\frac{1}{N_{a_q}} \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) - \frac{1}{N_{a_p}} \sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) < -\delta^*. \tag{32}$$

Multiplying (31) by N_{a_p} , we can get

$$\sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) - \frac{N_{a_p}}{N_{a_q}} \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) < -N_{a_p} \delta^*. \tag{33}$$

Thus,

$$\begin{aligned}
\sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) &= \sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) - \frac{N_{a_p}}{N_{a_q}} \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) + \frac{N_{a_p} + N_{a_q}}{N_{a_q}} \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) \\
&< N_{a_p} (1 - \delta^*) + N_{a_q}.
\end{aligned} \tag{34}$$

The inequality in (34) results from (33) and the fact that $\frac{1}{N_{a_q}} \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) \leq 1$. Then we multiply (32) by N_{a_q} . Using the similar method, we have

$$\sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) < N_{a_q}(1 - \delta^*) + N_{a_p}.$$

Combining this with (34) gives

$$\sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) < \max \{N_{a_p}(1 - \delta^*) + N_{a_q}, N_{a_q}(1 - \delta^*) + N_{a_p}\}$$

for any $p, q = 1, \dots, c$ and $p \neq q$. Summarizing this inequality over $\binom{c}{2}$ unique pairs of p, q , it is easy to deduce that

$$\sum_{i=1}^n \psi_i(\mathbf{w}_{uf}) = \sum_{p=1}^c \sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) < \frac{1}{c-1} \sum_{p \neq q} \max \{N_{a_p}(1 - \delta^*) + N_{a_q}, N_{a_q}(1 - \delta^*) + N_{a_p}\}.$$

Hence, the lower bound of $\mathcal{V}(\mathbf{w}_{uf})$ is given by

$$\begin{aligned} \mathcal{V}(\mathbf{w}_{uf}) &= \sum_{i=1}^n a_i - \sum_{i=1}^n b_i \psi_i(\mathbf{w}_{uf}) \\ &\geq \sum_{i=1}^n a_i - \max_i b_i \sum_{i=1}^n \psi_i(\mathbf{w}_{uf}) \\ &> \sum_{i=1}^n a_i - \frac{\max_i b_i}{c-1} \sum_{p \neq q} \max \{N_{a_p}(1 - \delta^*) + N_{a_q}, N_{a_q}(1 - \delta^*) + N_{a_p}\}. \end{aligned}$$

Define $\Delta^* = \sum_{i=1}^n a_i - \frac{\max_i b_i}{c-1} \sum_{p \neq q} \max \{N_{a_p}(1 - \delta^*) + N_{a_q}, N_{a_q}(1 - \delta^*) + N_{a_p}\}$. Since $\mathbf{w}^{data} = \arg\max_{\mathbf{w} \in \mathcal{W}} \mathcal{V}(\mathbf{w})$ by definition, it follows that $\mathcal{V}(\mathbf{w}^{data}) \geq \mathcal{V}(\mathbf{w}_{uf}) > \Delta^*$.

II. Equality of Opportunity

Similar in Case I, we first consider an classifier \mathbf{w}_{uf} which dose not satisfy the equality of opportunity requirement with δ^* for all groups. That is,

$$G_{EO}(\mathbf{w}_{uf}) = \left| \frac{1}{N_{a_p}^+} \sum_{i \in I_{a_p}^+} \psi_i(\mathbf{w}_{uf}) - \frac{1}{N_{a_q}^+} \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}) \right| > \delta^*$$

for any two groups $p, q = 1, \dots, c$ and $p \neq q$. Thus, for \mathbf{w}_{uf} , we have

$$\frac{1}{N_{a_p}^+} \sum_{i \in I_{a_p}^+} \psi_i(\mathbf{w}_{uf}) - \frac{1}{N_{a_q}^+} \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}) < -\delta^* \quad (35)$$

or

$$\frac{1}{N_{a_q}^+} \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}) - \frac{1}{N_{a_p}^+} \sum_{i \in I_{a_p}^+} \psi_i(\mathbf{w}_{uf}) < -\delta^*. \quad (36)$$

Multiplying (35) by $N_{a_p}^+$ results in

$$\sum_{i \in I_{a_p}^+} \psi_i(\mathbf{w}_{uf}) - \frac{N_{a_p}^+}{N_{a_q}^+} \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}) < -N_{a_p}^+ \delta^*.$$

Then, we have

$$\begin{aligned} \sum_{i \in I_{a_p}^+} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}) &= \sum_{i \in I_{a_p}^+} \psi_i(\mathbf{w}_{uf}) - \frac{N_{a_p}^+}{N_{a_q}^+} \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}) + \frac{N_{a_p}^+ + N_{a_q}^+}{N_{a_q}^+} \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}) \\ &< N_{a_p}^+(1 - \delta^*) + N_{a_q}^+. \end{aligned} \quad (37)$$

From the inequality (36), using similar methods leads to

$$\sum_{i \in I_{a_p}^+} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}) < N_{a_q}^+(1 - \delta^*) + N_{a_p}^+. \quad (38)$$

Combing (37) and ((38), we obtain

$$\sum_{i \in I_{a_p}^+} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}) < \max \{N_{a_p}^+(1 - \delta^*) + N_{a_q}^+, N_{a_q}^+(1 - \delta^*) + N_{a_p}^+\}$$

for any $p, q = 1, \dots, c$ and $p \neq q$. Summarizing this inequality over $\binom{c}{2}$ unique pairs of p, q , we can easily derive that

$$\sum_{i \in I^+} \psi_i(\mathbf{w}_{uf}) < \frac{1}{c-1} \sum_{p \neq q} \max \{N_{a_p}^+(1 - \delta^*) + N_{a_q}^+, N_{a_q}^+(1 - \delta^*) + N_{a_p}^+\},$$

when $I^+ = \{i \in \{1, 2, \dots, n\} | y_i = 1\}$ is the set of individuals from positive class. Hence,

$$\begin{aligned} \sum_{i=1}^n \psi_i(\mathbf{w}_{uf}) &= \sum_{i \in I^+} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I^-} \psi_i(\mathbf{w}_{uf}) \\ &< \frac{1}{c-1} \sum_{p \neq q} \max \{N_{a_p}^+(1 - \delta^*) + N_{a_q}^+, N_{a_q}^+(1 - \delta^*) + N_{a_p}^+\} + N^-, \end{aligned}$$

when $I^- = \{i \in \{1, 2, \dots, n\} | y_i = -1\}$ is the set of individuals from negative class and $N^- = |I^-|$ is the size of I^- . Then, the lower bound of $\mathcal{V}(\mathbf{w}_{uf})$ is

$$\begin{aligned} \mathcal{V}(\mathbf{w}_{uf}) &= \sum_{i=1}^n a_i - \sum_{i=1}^n b_i \psi_i(\mathbf{w}_{uf}) \\ &\geq \sum_{i=1}^n a_i - \max_i b_i \sum_{i=1}^n \psi_i(\mathbf{w}_{uf}) \\ &> \sum_{i=1}^n a_i - \frac{\max_i b_i}{c-1} \sum_{p \neq q} \max \{N_{a_p}^+(1 - \delta^*) + N_{a_q}^+, N_{a_q}^+(1 - \delta^*) + N_{a_p}^+\} - \max_i b_i N^-. \end{aligned}$$

Here, we define $\Delta^* = \sum_{i=1}^n a_i - \frac{\max_i b_i}{c-1} \sum_{p \neq q} \max \{N_{a_p}^+(1 - \delta^*) + N_{a_q}^+, N_{a_q}^+(1 - \delta^*) + N_{a_p}^+\} - \max_i b_i N^-$. Recall the definition of \mathbf{w}^{data} , it directly follows that $\mathcal{V}(\mathbf{w}^{data}) \geq \mathcal{V}(\mathbf{w}_{uf}) > \Delta^*$.

III. Statistical Parity

When statistical parity is used for fairness requirement, we first consider an classifier \mathbf{w}_{uf} which dose not satisfy this fairness requirement with δ^* for all groups. This leads to

$$\begin{aligned} G_{SP}(\mathbf{w}_{uf}) &= \left| \left(\frac{N_{a_p}^+}{N_{a_p}} - \frac{N_{a_q}^+}{N_{a_q}} \right) + \frac{1}{N_{a_p}} \left[\sum_{i \in I_{a_p}^-} \psi_i(\mathbf{w}_{uf}) - \sum_{i \in I_{a_p}^+} \psi_i(\mathbf{w}_{uf}) \right] - \frac{1}{N_{a_q}} \left[\sum_{i \in I_{a_q}^-} \psi_i(\mathbf{w}_{uf}) - \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}) \right] \right| \\ &> \delta^* \end{aligned}$$

for any two groups $p, q = 1, \dots, c$ and $p \neq q$. For \mathbf{w}_{uf} , this implies that

$$\left(\frac{N_{a_p}^+}{N_{a_p}} - \frac{N_{a_q}^+}{N_{a_q}} \right) + \frac{1}{N_{a_p}} \left[\sum_{i \in I_{a_p}^-} \psi_i(\mathbf{w}_{uf}) - \sum_{i \in I_{a_p}^+} \psi_i(\mathbf{w}_{uf}) \right] - \frac{1}{N_{a_q}} \left[\sum_{i \in I_{a_q}^-} \psi_i(\mathbf{w}_{uf}) - \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}) \right] < -\delta^* \quad (39)$$

or

$$\left(\frac{N_{a_q}^+}{N_{a_q}} - \frac{N_{a_p}^+}{N_{a_p}} \right) + \frac{1}{N_{a_q}} \left[\sum_{i \in I_{a_q}^-} \psi_i(\mathbf{w}_{uf}) - \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}) \right] - \frac{1}{N_{a_p}} \left[\sum_{i \in I_{a_p}^-} \psi_i(\mathbf{w}_{uf}) - \sum_{i \in I_{a_p}^+} \psi_i(\mathbf{w}_{uf}) \right] < -\delta^*. \quad (40)$$

In the case (39) holds, we multiply both sides of it by $N_{a_p}N_{a_q}$. Then, we have

$$N_{a_q} \left[\sum_{i \in I_{a_p}^-} \psi_i(\mathbf{w}_{uf}) - \sum_{i \in I_{a_p}^+} \psi_i(\mathbf{w}_{uf}) \right] - N_{a_p} \left[\sum_{i \in I_{a_q}^-} \psi_i(\mathbf{w}_{uf}) - \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}) \right] < N_{a_p}N_{a_q}^+ - N_{a_q}N_{a_p}^+ - N_{a_p}N_{a_q}\delta^*.$$

Rearrange this inequality, it is easy to derive that

$$\begin{aligned} & N_{a_q} \left[\sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) \right] \\ &= N_{a_q} \left[\sum_{i \in I_{a_p}^-} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_p}^+} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}^-} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}) \right] \\ &< -N_{a_p}N_{a_q}\delta^* + N_{a_p}N_{a_q}^+ + N_{a_q}N_{a_p}^+ + N_{a_p}N_{a_q}^- + N_{a_q}N_{a_q}^- + (N_{a_q} - N_{a_p}) \sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}). \end{aligned} \quad (41)$$

• If $N_{a_q} > N_{a_p}$, based on (41) we have

$$\begin{aligned} & N_{a_q} \left[\sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) \right] \\ &< -N_{a_p}N_{a_q}\delta^* + N_{a_p}N_{a_q}^+ + N_{a_q}N_{a_p}^+ + N_{a_p}N_{a_q}^- + N_{a_q}N_{a_q}^- + (N_{a_q} - N_{a_p})N_{a_q}^+. \end{aligned} \quad (42)$$

The last inequality is due to the fact that $\sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}) \leq N_{a_q}^+$. This finally leads to

$$\sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) < N_{a_p} \left(\frac{N_{a_q}^-}{N_{a_q}} - \delta^* \right) + N_{a_p}^+ + N_{a_q}. \quad (43)$$

• If $N_{a_q} \leq N_{a_p}$, from (41) we have

$$N_{a_q} \left[\sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) \right] < -N_{a_p}N_{a_q}\delta^* + N_{a_p}N_{a_q}^+ + N_{a_q}N_{a_p}^+ + N_{a_p}N_{a_q}^- + N_{a_q}N_{a_q}^-,$$

since $\sum_{i \in I_{a_q}^+} \psi_i(\mathbf{w}_{uf}) \geq 0$. Then,

$$\sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) < (1 - \delta^*)N_{a_p} + N_{a_p}^+ + N_{a_q}^-. \quad (44)$$

Now, we consider the other case where (40) holds. We multiply the both sides of (40) also by $N_{a_p}N_{a_q}$. Applying the similar steps as above, we could derive that

$$\begin{aligned} & N_{a_p} \left[\sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) \right] \\ & < -N_{a_p}N_{a_q}\delta^* + N_{a_q}N_{a_p}^+ + N_{a_p}N_{a_q}^+ + N_{a_q}N_{a_p}^- + N_{a_p}N_{a_q}^- + (N_{a_p} - N_{a_q}) \sum_{i \in I_{a_p}^+} \psi_i(\mathbf{w}_{uf}). \end{aligned}$$

• If $N_{a_q} \geq N_{a_p}$, we can obtain

$$\sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) < (1 - \delta^*)N_{a_q} + N_{a_q}^+ + N_{a_p}^-. \quad (45)$$

• If $N_{a_q} < N_{a_p}$, we have

$$\sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) < N_{a_q} \left(\frac{N_{a_p}^-}{N_{a_p}} - \delta^* \right) + N_{a_q}^+ + N_{a_p}. \quad (46)$$

Combining this with (43), (44) and (45) gives

$$\begin{aligned} & \sum_{i \in I_{a_p}} \psi_i(\mathbf{w}_{uf}) + \sum_{i \in I_{a_q}} \psi_i(\mathbf{w}_{uf}) \\ & < \mathbb{1}[N_{a_q} > N_{a_p}] \max \left\{ (1 - \delta^*)N_{a_q} + N_{a_q}^+ + N_{a_p}^-, N_{a_p} \left(\frac{N_{a_q}^-}{N_{a_q}} - \delta^* \right) + N_{a_p}^+ + N_{a_q} \right\} \\ & \quad + \mathbb{1}[N_{a_q} < N_{a_p}] \max \left\{ (1 - \delta^*)N_{a_p} + N_{a_p}^+ + N_{a_q}^-, N_{a_q} \left(\frac{N_{a_p}^-}{N_{a_p}} - \delta^* \right) + N_{a_q}^+ + N_{a_p} \right\} \\ & \quad + \mathbb{1}[N_{a_q} = N_{a_p}] \max \left\{ (1 - \delta^*)N_{a_q} + N_{a_q}^+ + N_{a_p}^-, (1 - \delta^*)N_{a_p} + N_{a_p}^+ + N_{a_q}^- \right\} \\ & = \mathcal{M}(p, q, \delta^*). \end{aligned} \quad (47)$$

We summarize (47) over $\binom{c}{2}$ unique pairs of p, q . Then, we could easily obtain

$$\sum_{i=1}^n \psi_i(\mathbf{w}_{uf}) < \frac{1}{c-1} \sum_{p \neq q} \mathcal{M}(p, q, \delta^*).$$

Thus,

$$\begin{aligned} \mathcal{V}(\mathbf{w}_{uf}) &= \sum_{i=1}^n a_i - \sum_{i=1}^n b_i \psi_i(\mathbf{w}_{uf}) \\ &\geq \sum_{i=1}^n a_i - \max_i b_i \sum_{i=1}^n \psi_i(\mathbf{w}_{uf}) \\ &> \sum_{i=1}^n a_i - \frac{\max_i b_i}{c-1} \sum_{p \neq q} \mathcal{M}(p, q, \delta^*). \end{aligned}$$

We define $\Delta^* = \sum_{i=1}^n a_i - \frac{\max_i b_i}{c-1} \sum_{p \neq q} \mathcal{M}(p, q, \delta^*)$. Afterwards, it directly follows that $\mathcal{V}(\mathbf{w}^{data}) \geq \mathcal{V}(\mathbf{w}_{uf}) > \Delta^*$ with statistical parity notion.

Consequently, there all exists Δ^* related to δ^* for three different fairness notions such that $\mathcal{V}(\mathbf{w}^{data}) > \Delta^*$. Because of the fact that $\mathcal{V}(\mathbf{w}^{data}) \geq \mathcal{V}(\mathbf{w}^*)$, we could conduct that $\mathcal{V}(\mathbf{w}^{data}) \geq \max \{\Delta^*, \mathcal{V}(\mathbf{w}^*)\}$.

Now, we have shown that in these three fairness notions, $\mathcal{V}(\mathbf{w}^{data}) > \Delta^*$ and $\mathcal{V}(\mathbf{w}^{data}) \geq \mathcal{V}(\mathbf{w}^*)$. Due to the definition of \mathbf{w}^* , we have $SWF(\mathbf{w}^*) \geq SWF(\mathbf{w}^{data})$ which leads to

$$\mathcal{V}(\mathbf{w}^*) - \bar{\rho}\delta^* \geq \mathcal{V}(\mathbf{w}^{data}) - \bar{\rho}\delta^{data} \quad (48)$$

where $\delta^{data} \in [0, 1]$ is the (un)fairness level of \mathbf{w}^{data} . Combing (48) and $\mathcal{V}(\mathbf{w}^{data}) > \Delta^*$, we can obtain

$$SWF(\mathbf{w}^*) = \mathcal{V}(\mathbf{w}^*) - \bar{\rho}\delta^* > \Delta^* - \bar{\rho}\delta^{data}. \quad (49)$$

If $\bar{\rho} > 0$, the right-hand side of the inequality in (49) is greater than $\Delta^* - \bar{\rho}$. If $\bar{\rho} \leq 0$, it is greater than Δ^* . This yields the statement of the theorem.

References

- Abdulkadiroğlu, A., Sönmez, T., and Ünver, M. U. (2004). Room assignment-rent division: A market approach. *Social Choice and Welfare*, 22(3):515–538.
- Alagoz, O., Schaefer, A. J., and Roberts, M. S. (2009). Optimizing organ allocation and acceptance. In *Handbook of Optimization in Medicine*, pages 1–24. Springer.
- Angus, D. C., Linde-Zwirble, W. T., Lidicker, J., Clermont, G., Carcillo, J., and Pinsky, M. R. (2001). Epidemiology of severe sepsis in the united states: Analysis of incidence, outcome, and associated costs of care. *Critical Care Medicine*, 29(7):1303–1310.
- Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. propublica (2016). URL: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Arabi, Y., Al Shirawi, N., Memish, Z., Venkatesh, S., and Al-Shimemeri, A. (2003). Assessment of six mortality prediction models in patients admitted with severe sepsis and septic shock to the intensive care unit: A prospective cohort study. *Critical Care*, 7(5):1–7.
- Atkinson, A. B. et al. (1970). On the measurement of inequality. *Journal of Economic Theory*, 2(3):244–263.
- Aziz, H., Caragiannis, I., Igarashi, A., and Walsh, T. (2019). Fair allocation of indivisible goods and chores. In *International Joint Conference on Artificial Intelligence*, pages 53–59.
- Aziz, H., Gaspers, S., Mackenzie, S., and Walsh, T. (2015). Fair assignment of indivisible objects under ordinal preferences. *Artificial Intelligence*, 227:71–92.

- Azizi, M. J., Vayanos, P., Wilder, B., Rice, E., and Tambe, M. (2018). Designing fair, efficient, and interpretable policies for prioritizing homeless youth for housing resources. In *International Conference on the Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pages 35–51. Springer.
- Babcock, L. and Loewenstein, W. G. (1996). Choosing the wrong pond: Social comparisons in negotiations that reflect a self-serving bias. *The Quarterly Journal of Economics*, 111(1):1–19.
- Bandiera, O., Barankay, I., and Rasul, I. (2005). Social preferences and the response to incentives: Evidence from personnel data. *The Quarterly Journal of Economics*, 120(3):917–962.
- Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1:2017.
- Barocas, S. and Selbst, A. D. (2016). Big data’s disparate impact. *California Law Review*, 104:671.
- Becker, G. S. (1974). A theory of social interactions. *Journal of Political Economy*, 82(6):1063–1093.
- Becker, L. B., Han, B. H., Meyer, P. M., Wright, F. A., Rhodes, K. V., Smith, D. W., Barrett, J., and Project, C. C. (1993). Racial differences in the incidence of cardiac arrest and subsequent survival. *New England journal of medicine*, 329(9):600–606.
- Benjamin, D. J., Choi, J. J., and Strickland, A. J. (2010). Social identity and preferences. *American Economic Review*, 100(4):1913–28.
- 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.
- Bertsimas, D., Farias, V. F., and Trichakis, N. (2013). Fairness, efficiency, and flexibility in organ allocation for kidney transplantation. *Operations Research*, 61(1):73–87.
- Bierman, A. S. (2007). Sex matters: Gender disparities in quality and outcomes of care. *Canadian Medical Association Journal*, 177(12):1520–1521.
- Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193.
- Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A., Schein, R. M., and Sibbald, W. J. (1992). Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*, 101(6):1644–1655.

- Bose, I. and Chen, X. (2009). Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research*, 195(1):1–16.
- Brams, S. J., Brams, S. J., and Taylor, A. D. (1996). *Fair Division: From Cake-cutting to Dispute Resolution*. Cambridge University Press.
- Brooks, J. P. (2011). Support vector machines with the ramp loss and the hard margin loss. *Operations research*, 59(2):467–479.
- Burgess, E. W. (1928). Factors determining success or failure on parole. *The workings of the indeterminate sentence law and the parole system in Illinois*, pages 221–234.
- Capon, N. (1982). Credit scoring systems: A critical analysis. *Journal of Marketing*, 46(2):82–91.
- Charness, G. and Rabin, M. (1999). Social preferences: Some simple tests and a new model. Economics Working Papers 441, Department of Economics and Business, Universitat Pompeu Fabra.
- Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.
- Chen, E. H., Shofer, F. S., Dean, A. J., Hollander, J. E., Baxt, W. G., Robey, J. L., Sease, K. L., and Mills, A. M. (2008). Gender disparity in analgesic treatment of emergency department patients with acute abdominal pain. *Academic Emergency Medicine*, 15(5):414–418.
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.
- Cole, R. and Gkatzelis, V. (2018). Approximating the nash social welfare with indivisible items. *SIAM Journal on Computing*, 47(3):1211–1236.
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’17, page 797–806, New York, NY, USA. Association for Computing Machinery.
- Coutts, S. (1984). Motor insurance rating, an actuarial approach. *Journal of the Institute of Actuaries*, 111(1):87–148.

- Cox, J. C. and Sadiraj, V. (2012). Direct tests of individual preferences for efficiency and equity. *Economic Inquiry*, 50(4):920–931.
- Dahl, R., Grønlykke, L., Haase, N., Holst, L., Perner, A., Wetterslev, J., Rasmussen, B., Meyhoff, C., 6S-Trial, and investigators, T. T. (2015). Variability in targeted arterial oxygenation levels in patients with severe sepsis or septic shock. *Acta Anaesthesiologica Scandinavica*, 59(7):859–869.
- Dawes, C. T., Fowler, J. H., Johnson, T., Mcelreath, R., and Smirnov, O. (2007). Egalitarian motives in humans. *Nature*, 446(7137):794–6.
- Dellinger, R. P., Levy, M. M., Rhodes, A., Annane, D., Gerlach, H., Opal, S. M., Sevransky, J. E., Sprung, C. L., Douglas, I. S., Jaeschke, R., et al. (2013). Surviving sepsis campaign: International guidelines for management of severe sepsis and septic shock, 2012. *Intensive Care Medicine*, 39(2):165–228.
- Dionne, G., Giuliano, F., and Picard, P. (2009). Optimal auditing with scoring: Theory and application to insurance fraud. *Management Science*, 55(1):58–70.
- Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. (2018). Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801.
- Dover, T. L., Major, B., Kunstman, J. W., and Sawyer, P. J. (2015). Does unfairness feel different if it can be linked to group membership? cognitive, affective, behavioral and physiological implications of discrimination and unfairness. *Journal of Experimental Social Psychology*, 56:96–103.
- Dua, D. and Graff, C. (2017). UCI machine learning repository.
- 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*, page 214–226, New York, NY, USA. Association for Computing Machinery.
- Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.
- Foley, D. K. (1967). *Resource Allocation and the Public Sector*, volume 7. Yale Economics Essays.
- Frees, E. W., Meyers, G., and Cummings, A. D. (2011). Summarizing insurance scores using a gini index. *Journal of the American Statistical Association*, 106(495):1085–1098.
- Friedman, J. H., Popescu, B. E., et al. (2008). Predictive learning via rule ensembles. *Annals of Applied Statistics*, 2(3):916–954.

- Fu, R., Huang, Y., and Singh, P. V. (2021). Crowds, lending, machine, and bias. *Information Systems Research*, 32(1):72–92.
- Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. (2021). Predictably unequal? the effects of machine learning on credit markets. *Journal of Finance*, *forthcoming*.
- Gal, Y., Mash, M., Procaccia, A. D., and Zick, Y. (2017). Which is the fairest (rent division) of them all? *Journal of the ACM (JACM)*, 64(6):1–22.
- Ganju, K. K., Atasoy, H., McCullough, J., and Greenwood, B. (2020). The role of decision support systems in attenuating racial biases in healthcare delivery. *Management Science*, 66(11):5171–5181.
- Gogos, C. A., Lekkou, A., Papageorgiou, O., Siagris, D., Skoutelis, A., and Bassaris, H. P. (2003). Clinical prognostic markers in patients with severe sepsis: A prospective analysis of 139 consecutive cases. *Journal of Infection*, 47(4):300–306.
- Gómez, J. A., Arévalo, J., Paredes, R., and Nin, J. (2018). End-to-end neural network architecture for fraud scoring in card payments. *Pattern Recognition Letters*, 105:175–181.
- Halvaiee, N. S. and Akbari, M. K. (2014). A novel model for credit card fraud detection using artificial immune systems. *Applied Soft Computing*, 24:40–49.
- Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS’16, page 3323–3331, Red Hook, NY, USA. Curran Associates Inc.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). Boosting and additive trees. In *The Elements of Statistical Learning*, pages 337–387. Springer.
- Henry, K. E., Hager, D. N., Pronovost, P. J., and Saria, S. (2015). A targeted real-time early warning score (trewscore) for septic shock. *Science Translational Medicine*, 7(299):299ra122.
- Hernandez, M. A. and Torero, M. (2018). A poverty-sensitive scorecard to prioritize lending and grant allocation: Evidence from central america. *Food Policy*, 77:81–90.
- Hoffman, P. and Adelberg, S. (1980). Salient factor score-a nontechnical overview. *Federal Probation*, 44(1):44–52.
- Hoffman, P. B. and Beck, J. L. (1997). The origin of the federal criminal history score. *Federal Sentencing Reporter*, 9(4):192–197.

- Hossain, S., Mladenovic, A., and Shah, N. (2020). Designing fairly fair classifiers via economic fairness notions. In *Proceedings of The Web Conference 2020*, pages 1559–1569.
- Hu, L. and Chen, Y. (2018). A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pages 1389–1398.
- Hu, L. and Chen, Y. (2020). Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 535–545.
- Hurley, M. and Adebayo, J. (2017). Credit scoring in the era of big data. *Yale Journal of Law and Technology*, 18:148–216.
- Iskander, K. N., Osuchowski, M. F., Stearns-Kurosawa, D. J., Kurosawa, S., Stepien, D., Valentine, C., and Remick, D. G. (2013). Sepsis: Multiple abnormalities, heterogeneous responses, and evolving understanding. *Physiological Reviews*, 93(3):1247–1288.
- Jha, A. K., Fisher, E. S., Li, Z., Orav, E. J., and Epstein, A. M. (2005). Racial trends in the use of major procedures among the elderly. *New England Journal of Medicine*, 353(7):683–691.
- Kahneman, D., Knetsch, J. L., and Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *American Economic Review*, 76(4):728–741.
- Karlan, D. and Zinman, J. (2011). Microcredit in theory and practice: Using randomized credit scoring for impact evaluation. *Science*, 332(6035):1278–84.
- Knaus, W. A., Draper, E. A., Wagner, D. P., and Zimmerman, J. E. (1985). Apache ii: A severity of disease classification system. *Critical Care Medicine*, 13(10):818–829.
- Knaus, W. A., Wagner, D. P., Draper, E. A., Zimmerman, J. E., and Damiano, A. M. (1991). The apache iii prognostic system: Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*, 100(6):1619–1636.
- Knaus, W. A., Zimmerman, J. E., Wagner, D. P., Draper, E. A., and Lawrence, D. E. (1981). Apache-acute physiology and chronic health evaluation: A physiologically based classification system. *Critical Care Medicine*, 9(8):591.
- Kozodoi, N., Jacob, J., and Lessmann, S. (2021). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*.
- Kramer, J. H. and Scirica, A. J. (1986). Complex policy choices: The pennsylvania commission on sentencing. *Federal Probation*, 50:15.

- Kraut, J. A. and Madias, N. E. (2010). Metabolic acidosis: Pathophysiology, diagnosis and management. *Nature Reviews Nephrology*, 6(5):274.
- Kurowski, A., Szarpak, L., Frass, M., Samarin, S., and Czyzewski, L. (2016). Gcs scale used as a prognostic factor in unconscious patients following cardiac arrest in prehospital situations: Preliminary data. *American Journal of Emergency Medicine*, 34(6):1178–1179.
- Lang, J. and Rothe, J. (2016). Fair division of indivisible goods. In *Economics and Computation*, pages 493–550. Springer.
- Lawless, C. and Günlük, O. (2020). Fair and interpretable decision rules for binary classification. In *NeurIPS Workshop on Optimization for Machine Learning*, pages 1–26.
- Le Gall, J.-R., Klar, J., Lemeshow, S., Saulnier, F., Alberti, C., Artigas, A., and Teres, D. (1996). The logistic organ dysfunction system: A new way to assess organ dysfunction in the intensive care unit. *Journal of the American Medical Association*, 276(10):802–810.
- Le Gall, J.-R., Lemeshow, S., and Saulnier, F. (1993). A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *Journal of the American Medical Association*, 270(24):2957–2963.
- Le Gall, J.-R., Loirat, P., Alperovitch, A., Glaser, P., Granthil, C., Mathieu, D., Mercier, P., Thomas, R., and Villers, D. (1984). A simplified acute physiology score for icu patients. *Critical Care Medicine*, 12(11):975–977.
- Li, J. (2008). The power of conventions: A theory of social preferences. *Journal of Economic Behavior & Organization*, 65(3-4):489–505.
- Li, Y., Wang, X., Djehiche, B., and Hu, X. (2020). Credit scoring by incorporating dynamic networked information. *European Journal of Operational Research*, 286(3):1103–1112.
- Lipton, R. J., Markakis, E., Mossel, E., and Saberi, A. (2004). On approximately fair allocations of indivisible goods. In *Proceedings of the 5th ACM Conference on Electronic Commerce*, pages 125–131.
- Lohaus, M., Perrot, M., and Von Luxburg, U. (2020). Too relaxed to be fair. In *International Conference on Machine Learning*, pages 6360–6369. PMLR.
- Malthouse, E. C. (1999). Ridge regression and direct marketing scoring models. *Journal of Interactive Marketing*, 13(4):10–23.

- Marconi, V. C., Duncan, M. S., So-Armah, K., Re 3rd, V. L., Lim, J. K., Butt, A. A., Goetz, M. B., Rodriguez-Barradas, M. C., Alcorn, C. W., Lennox, J., et al. (2018). Bilirubin is inversely associated with cardiovascular disease among hiv-positive and hiv-negative individuals in vacs (veterans aging cohort study). *Journal of the American Heart Association*, 7(10):e007792.
- Martin-Loeches, I., Guia, M. C., Vallecoccia, M. S., Suarez, D., Ibarz, M., Irazabal, M., Ferrer, R., and Artigas, A. (2019). Risk factors for mortality in elderly and very elderly critically ill patients with sepsis: A prospective, observational, multicenter cohort study. *Annals of Intensive Care*, 9(1):1–9.
- 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.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.
- Minne, L., Abu-Hanna, A., and de Jonge, E. (2008). Evaluation of sofa-based models for predicting mortality in the icu: A systematic review. *Critical Care*, 12(6):1–13.
- Moreno, R. P., Metnitz, P. G., Almeida, E., Jordan, B., Bauer, P., Campos, R. A., Iapichino, G., Edbrooke, D., Capuzzo, M., and Le Gall, J.-R. (2005). Saps 3-from evaluation of the patient to evaluation of the intensive care unit. part 2: Development of a prognostic model for hospital mortality at icu admission. *Intensive Care Medicine*, 31(10):1345–1355.
- Mukherjee, V. and Evans, L. (2017). Implementation of the surviving sepsis campaign guidelines. *Current Opinion in Critical Care*, 23(5):412–416.
- Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., and Buchman, T. G. (2018). An interpretable machine learning model for accurate prediction of sepsis in the icu. *Critical Care Medicine*, 46(4):547–553.
- Nguyen, T. and Sanner, S. (2013). Algorithms for direct 0–1 loss optimization in binary classification. In *International Conference on Machine Learning*, pages 1085–1093.
- Northpointe, I. (2015). Practitioner’s guide to compas core.
- Obermeyer, Z. and Mullainathan, S. (2019). Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 89–89.

- Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.
- Organization, W. H. et al. (2015). Diabetes, fact sheet n 312. updated january 2015. *The World Health Organization. Cancer. Fact Sheet*, (297).
- Procaccia, A. D. (2013). Cake cutting: Not just child’s play. *Communications of the ACM*, 56(7):78–87.
- Procaccia, A. D. (2015). Cake cutting algorithms. In *Handbook of Computational Social Choice, Chapter 13*. Citeseer.
- Quadrianto, N. and Sharmanska, V. (2017). Recycling privileged learning and distribution matching for fairness. *Advances in Neural Information Processing Systems*, 30:677–688.
- Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, pages 1281–1302.
- Rajkomar, A., Hardt, M., Howell, M., Corrado, G., and Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*.
- Rawls, J. (1999). *A Theory of Justice: Revised Edition*. Harvard university press.
- Ribas, V. J., Vellido, A., Ruiz-Rodríguez, J., and Rello, J. (2012). Severe sepsis mortality prediction with logistic regression over latent factors. *Expert Systems with Applications*, 39(2):1937–1943.
- Robertson, J. and Webb, W. (1998). *Cake-cutting Algorithms: Be Fair if You Can*. CRC Press.
- Rubineau, B. and Kang, Y. (2012). Bias in white: A longitudinal natural experiment measuring changes in discrimination. *Management Science*, 58(4):660–677.
- Santana, A. R., de Sousa, J. L., Amorim, F. F., Menezes, B. M., Araújo, F. V. B., Soares, F. B., de Carvalho Santos, L. C., de Araújo, M. P. B., Rocha, P. H. G., Júnior, P. N. F., et al. (2013). Sao2/fio2 ratio as risk stratification for patients with sepsis. *Critical Care*, 17(4):1–59.
- Sedlak, T. W. and Snyder, S. H. (2004). Bilirubin benefits: Cellular protection by a biliverdin reductase antioxidant cycle. *Pediatrics*, 113(6):1776–1782.
- Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Belomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., et al. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). *Journal of the American Medical Association*, 315:801–810.

- Skoufias, E., Diamond, A., Vinha, K., Gill, M., and Dellepiane, M. R. (2020). Estimating poverty rates in subnational populations of interest: An assessment of the simple poverty scorecard. *World Development*, 129:104887.
- Solinger, A. B. and Rothman, S. I. (2013). Risks of mortality associated with common laboratory tests: A novel, simple and meaningful way to set decision limits from data available in the electronic medical record. *Clinical Chemistry and Laboratory Medicine*, 51(9):1803–1813.
- Steinhaus, H. (1948). The problem of fair division. *Econometrica*, 16:101–104.
- Sweeney, T. E., Perumal, T. M., Henao, R., Nichols, M., Howrylak, J. A., Choi, A. M., Bermejo-Martin, J. F., Almansa, R., Tamayo, E., and Davenport, E. E. (2018). A community approach to mortality prediction in sepsis via gene expression analysis. *Nature Communications*, 9(1):694.
- Thomas, L., Crook, J., and Edelman, D. (2017). *Credit Scoring and Its Applications*. SIAM.
- Torio, C. M. and Andrews, R. M. (2013). National inpatient hospital costs: The most expensive conditions by payer, 2011: Statistical brief #160.
- Tricomi, E., Rangel, A., Camerer, C. F., and O’Doherty, J. P. (2010). Neural evidence for inequality-averse social preferences. *Nature*, 463(feb.25):1089–1091.
- Tsaih, R., Liu, Y.-J., Liu, W., and Lien, Y.-L. (2004). Credit scoring system for small business loans. *Decision Support Systems*, 38(1):91–99.
- Ustun, B. and Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391.
- Vigdor, N. (2019). Apple card investigated after gender discrimination complaints. *The New York Times*.
- Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C., Suter, P., and Thijs, L. G. (1996). The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7):707–710.
- Vona, L. W. (2012). *Fraud Risk Assessment: Building a Fraud Audit Program*. John Wiley & Sons.
- Wang, R., He, M., and Xu, J. (2020). Serum bilirubin level correlates with mortality in patients with traumatic brain injury. *Medicine*, 99(27).
- Wolsey, L. A. (1998). *Integer Programming*, volume 52. John Wiley & Sons.

- Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. (2017). Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR.
- Wu, Y., Huang, S., and Chang, X. (2021). Understanding the complexity of sepsis mortality prediction via rule discovery and analysis: A pilot study. *Medical Decision Making (Major Revision)*.
- Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017a). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180.
- 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.
- Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., and Weller, A. (2017b). From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239.
- Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017c). Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR.
- Zardari, N. u. H., Cordery, I., and Sharma, A. (2010). An objective multiattribute analysis approach for allocation of scarce irrigation water resources. *JAWRA Journal of the American Water Resources Association*, 46(2):412–428.
- Zeng, J., Ustun, B., and Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 3(180):689–722.
- Zhang, X., Khalili, M. M., Tekin, C., and Liu, M. (2019). Group retention when using machine learning in sequential decision making: The interplay between user dynamics and fairness. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.
- Zhao, H. and Gordon, G. (2019). Inherent tradeoffs in learning fair representations. *Advances in neural information processing systems*, 32:15675–15685.
- Zou, J., Lederer, D. J., and Rabinowitz, D. (2020). Efficiency in lung transplant allocation strategies. *Annals of Applied Statistics*, 14(3):1088–1121.