Enforcing Fair Predicted Scores on Intervals of Percentiles by Difference-of-Convex Constraints

Yutian He
Department of Mathematics
University of Iowa
yutian-he@uiowa.edu

Yankun Huang
Department of Information Systems
Arizona State University
yankun.huang@asu.edu

Yao Yao
School of Mathematics
University of Minnesota Twin Cities
yaoxx323@umn.edu

Qihang Lin
Department of Business Analytics
University of Iowa
qihang-lin@uiowa.edu

Abstract

Fairness in machine learning has become a critical concern. Existing approaches often focus on achieving full fairness across all score ranges generated by predictive models, ensuring fairness in both high- and low-percentile populations. However, this stringent requirement can compromise predictive performance and may not align with the practical fairness concerns of stakeholders. In this work, we propose a novel framework for building partially fair machine learning models that enforce fairness only within a specific percentile interval of interest while maintaining flexibility in other regions. We introduce statistical metrics to evaluate partial fairness within a given percentile interval. To achieve partial fairness, we propose an in-processing method by formulating the model training problem as constrained optimization with difference-of-convex constraints, which can be solved by an inexact difference-of-convex algorithm (IDCA). We provide the complexity analysis of IDCA for finding a nearly KKT point. Through numerical experiments on real-world datasets, we demonstrate that our framework achieves high predictive performance while enforcing partial fairness where it matters most.

1 INTRODUCTION

Fairness in machine learning has emerged as a critical issue, especially when AI systems are employed to assist in high-stakes decision-making (Barocas et al., 2023). Existing research on fairness has introduced multiple criteria to evaluate and enforce fairness of machine learning models among different groups (Feldman et al., 2015; Dwork et al., 2012; Hardt et al., 2016; Chzhen et al., 2020b; Vogel et al., 2021). However, most of these metrics assume that fairness is equally important at all levels of the output predicted scores-a requirement that may impose unnecessary constraints and compromise models' predictive performance. In some applications, fairness considerations are instead localized within specific percentile intervals of the score distribution (e.g., the top/bottom 0–20% or regions near decision boundaries), where outcomes are more contested and socially consequential.

For example, in school admissions, machine learning models may be used to rank applicants by predicted admissibility/qualification scores. Fairness is especially critical for those near the admission threshold, where small disparities in prediction can determine acceptance or rejection and are most likely to be scrutinized. By contrast, applicants with very high scores are automatically admitted, making fairness concerns at the top percentiles less contentious. Ensuring fairness around the threshold both reduces complaints and promotes equal educational opportunities.

The uneven importance of fairness across score percentiles is supported by findings in social psychology and human factors. According to the *uncertainty management theory* (van den Bos, 2001), individuals rely more on fairness judgments when facing uncertainty. Those with scores in the medium percentiles, e.g., near decision boundaries, experience more uncertainty regarding the decisions made about them than those in high or low percentiles. van den Bos (2001) shows through experiments that participants exposed to uncertainty manipulations reacted more strongly to unfair procedures, highlighting that uncertainty amplifies sensitivity to fairness. Additionally, Wang et al. (2020)

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

found that, in algorithmic decision-making, individuals with favorable outcomes perceived algorithms as fairer than those with unfavorable ones. These findings suggest that when outcomes are near borderline or unfavorable, people increasingly rely on fairness perceptions for evaluating the legitimacy of the decision. Consequently, fairness concerns are more pronounced within certain percentile ranges of the score distribution. This implies that enforcing fairness uniformly across all percentiles, as many existing techniques do, may impose unnecessary restrictions.

To address this gap, we propose the concept of *partial fairness*, which ensures fairness within specific percentile intervals of predicted scores, providing a practical alternative to full fairness by focusing on contested cases while allowing flexibility elsewhere. We propose multiple metrics to quantify partial fairness and adopt an *in-processing* approach, formulating training as a constrained optimization problem that maximizes predictive performance subject to partial fairness constraints based on the proposed metrics. Because partial fairness targets only critical score percentiles, it reduces unnecessary restrictions, enabling models to maintain stronger predictive performance.

Following standard modeling practices in machine learning, we formulate the constrained optimization problems using empirical approximations and surrogate functions. We show that each proposed fairness constraint can be expressed as the difference of two convex (possibly non-smooth) functions. Embedding these constraints into the training process leads to a *difference-of-convex* (DC) constrained optimization problem.

DC optimization has been studied extensively for several decades (Pham Dinh and Le Thi, 1997; Hiriart-Urruty, 1985; Le Thi and Pham Dinh, 2018; Pham Dinh and Le Thi, 2014; Le Thi and Pham Dinh, 2024). The best-known algorithm for DC optimization is the *difference-of-convex algorithm* (DCA), which iteratively updates the solution by solving a sequence of convex subproblems obtained by linearizing the concave component of each DC function (Pham Dinh and El Bernoussi, 1986; Pham Dinh and Le Thi, 2014). When these subproblems are computationally challenging, they are often solved approximately, leading to the *inexact DCA* (IDCA), which is more practical and widely applicable than the standard DCA (Cruz Neto et al., 2020; Ferreira et al., 2026). Although the asymptotic convergence theory of DCA for DC-constrained optimization has been well established (Pang et al., 2017; Sriperumbudur and Lanckriet, 2012; Pham Dinh and Le Thi, 2014), the convergence rate and the complexity of IDCA for DC-constrained optimization has not been analyzed.

To bridge this gap, we study the IDCA for general non-smooth DC-constrained optimization and provide a theoretical analysis of its oracle complexity for computing a nearly ϵ-Karush–Kuhn–Tucker (KKT) point. To the best of our knowledge, this is the first complexity analysis of IDCA for DC-constrained problems, making it an important contribution in its own right.

2 RELATED WORKS

Various fairness metrics have been studied in the literature, including demographic parity (Calders et al., 2009; Zafar et al., 2017; Feldman et al., 2015; Dwork et al., 2012), equality of opportunity (Hardt et al., 2016), equality of odds (Hardt et al., 2016), predictive quality parity (Chouldechova, 2017), counterfactual fairness (Kusner et al., 2017) and predicted parity (calibration) (Chouldechova, 2017).

Most of the above metrics are tailored to classification tasks, as they evaluate fairness based on predicted class labels. In contrast, fairness metrics have also been proposed directly on predicted scores/probabilities, so they are applicable to both classification and regression tasks. For examples, fairness metrics based on pairwise comparison (AUC-based metrics) have been studied in Borkan et al. (2019); Kallus and Zhou (2019); Dixon et al. (2018); Vogel et al. (2021); Narasimhan et al. (2020); Beutel et al. (2019); Yao et al. (2022); Yang et al. (2023). Other score-based fairness metrics include the ones based on Wasserstein distance (Miroshnikov et al., 2022, 2021; Chzhen et al., 2020b,a; Chzhen and Schreuder, 2022), ROC (Vogel et al., 2021), Gini coefficient (Chen et al., 2024), and the Kolmogorov-Smirnov test (Chen et al., 2024).

The fairness metrics mentioned above aim to achieve fairness over the entire distribution of scores generated by predictive models. In contrast, the partial fairness metrics proposed in this work focus on measuring fairness over a specific percentile interval, where prediction outcomes are more contested and fairness concerns are more frequently raised.

The ROC-based fairness notion proposed by Vogel et al. (2021) requires $\mathrm{ROC}(\alpha) = \alpha$ for all $\alpha \in [0, 1]$, where $\mathrm{ROC}(\alpha)$ denotes the ROC curve comparing the predicted scores of two groups. Although the authors do not explicitly introduce the concept of partial fairness, one can restrict α to a subinterval of $[0, 1]$ to recover a form of partial statistical parity (Definition 4.2) considered in our work; see Section 3.3 of Vogel et al. (2021) for details. That said, our notion of partial fairness is more general, encompassing not only partial statistical parity but also partial demographic parity (Definition 21) and partial group AUC fairness (Definition 21). Moreover, the numerical approach used to enforce ROC-based fairness in Vogel et al. (2021) (Algorithm 2) is heuristic and lacks theoretical guarantees,

whereas our method is grounded in DC optimization and admits rigorous convergence analysis.

Fairness metrics have also been studied for selection tasks (Khalili et al., 2021). These tasks typically require a model that predicts qualification scores, which are then used to make selection decisions. The techniques proposed in this paper can further promote fairness in selection by producing fair qualification scores.

The three main approaches for building a fairness-aware machine learning model include the pre-processing (Dwork et al., 2012), post-processing (Hardt et al., 2016; Chzhen et al., 2020b; Chzhen and Schreuder, 2022; Xian et al., 2023), and in-processing methods (Agarwal et al., 2018; Goh et al., 2016; Wan et al., 2023). The methods in this paper are the in-processing methods, where we optimize a model's prediction performance subject to partial fairness constraints, which can be formulated as a DC-constrained optimization problem.

DCA is the most studied numerical algorithm for DC optimization (Hoang, 1984; Pham Dinh and El Bernoussi, 1986; Sriperumbudur and Lanckriet, 2012; Pham Dinh and Le Thi, 1997; Le Thi et al., 2018; Pham Dinh and Le Thi, 2014; Abbaszadehpeivasti et al., 2024). Each iteration of DCA requires solving a convex subproblem exactly, which can be challenging. To address this, we employ IDCA, a variant of DCA that allows for inexact solutions to the subproblems (Xu et al., 2019; Cruz Neto et al., 2020; Ferreira et al., 2026; Kanzow and Neder, 2024; Moudafi and Maingé, 2006; Liu and Takeda, 2022). In addition to traditional DCA and IDCA, there exist many other numerical algorithms for DC optimization, including proximal DCA (Cruz Neto et al., 2020; Souza et al., 2016; Sun et al., 2003; Moudafi and Maingé, 2006; Pang et al., 2017; Moudafi, 2024), boosted DCA (Aragón-Artacho et al., 2018; Aragón-Artacho and Phan, 2020; Aragón-Artacho et al., 2022; Zhang and Niu, 2024; Ferreira et al., 2024, 2026), inertial DCA (Maingé and Moudafi, 2008; de Oliveira and Tcheou, 2019), DC composite algorithm (Le Thi et al., 2024; Hu et al., 2024), gradient descent method (Khamaru and Wainwright, 2019), proximal gradient method (Khamaru and Wainwright, 2019), Frank-Wolfe method (Khamaru and Wainwright, 2019), smoothing method (Sun and Sun, 2023; Moudafi, 2021; Yao et al., 2022; Hu et al., 2024), augmented Lagrangian method (Pang and Tao, 2018; Sun and Sun, 2023), and codifferential method (Bagirov and Ugon, 2011).

However, most of the aforementioned works focus exclusively on unconstrained DC optimization, making their results inapplicable to our problems. The convergence of DCA and its variants in the constrained setting has been established by Sriperumbudur and Lanckriet (2012); Pham Dinh and Le Thi (2014); Pang et al. (2017); Pang and Tao (2018); Le Thi et al. (2024). Notably, Pang et al. (2017); Pang and Tao (2018) proved convergence to a directional stationary point, which is a stronger notion of stationarity than the nearly KKT point considered in this paper. However, for our problems, those results of DCA are not applicable due to the difficulty of solving the subproblem exactly. To the best of our knowledge, no existing work has established convergence results for IDCA in constrained settings. Our work, therefore, contributes to the convergence and complexity analysis of IDCA.

The theoretical analysis in this work is motivated by recent developments in the complexity analysis of first-order methods for weakly convex non-smooth constrained optimization (Ma et al., 2020; Boob et al., 2023; Huang and Lin, 2023; Jia and Grimmer, 2025). Assuming different constraint qualifications on the subproblems, those works establish a complexity of $\mathcal{O}(\epsilon^{-4})$ for finding a nearly ϵ-KKT point, which matches the complexity bound achieved in this work. While a weakly convex function is a DC function, the converse does not hold. Hence, the results from these prior works cannot be directly applied to IDCA.

3 PRELIMINARY AND NOTATIONS

We consider a prediction task in which the goal is to develop a model that predicts a continuous or discrete target variable $\zeta \in \mathbb{R}$ using a feature vector $\boldsymbol{\xi} \in \mathbb{R}^d$. Each data point has a categorical sensitive attribute (e.g., gender and race), denoted by $\gamma \in \mathcal{G} := \{1, 2, \ldots, G\}$, which may or may not be a component of $\boldsymbol{\xi}$. A data point is represented as a random triplet $(\boldsymbol{\xi}, \zeta, \gamma) \in \mathbb{R}^{d+2}$ following a ground truth distribution.

Let $h_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}$ be a predictive model that assigns a predicted score $h_{\mathbf{w}}(\boldsymbol{\xi})$ to a feature vector $\boldsymbol{\xi}$. When ζ is continuous, this score directly predicts ζ. When ζ is binary, the score can be converted into a prediction of ζ after thresholding. We are interested in the fairness of $h_{\mathbf{w}}(\boldsymbol{\xi})$ among the groups of data defined by γ. Here, $h_{\mathbf{w}}$ depends on a parameter vector $\mathbf{w}$ that belongs to a convex and closed set $\mathcal{W} \subset \mathbb{R}^n$. We denote the gradient of $h_{\mathbf{w}}$ with respect to $\mathbf{w}$ by $\nabla_{\mathbf{w}} h_{\mathbf{w}}(\boldsymbol{\xi})$ and make the following assumption throughout the paper.

Assumption 3.1. For any $\boldsymbol{\xi}$, $h_{\mathbf{w}}(\boldsymbol{\xi})$ is differentiable and $\nabla_{\mathbf{w}} h_{\mathbf{w}}(\boldsymbol{\xi})$ is L_h-Lipschitz continuous in $\mathbf{w}$.

Let $\Pr(\mathcal{A})$ and $\Pr(\mathcal{A}|\mathcal{B})$ denote, respectively, the probability of $\mathcal{A}$ and the conditional probability of $\mathcal{A}$ conditioning on $\mathcal{B}$. Let $\mathbf{1}_{\mathcal{A}}$ be the 0-1 indicator function of event $\mathcal{A}$. Let

$$F_{\mathbf{w},k}(\theta) := \Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) \leq \theta | \gamma = k), \ \forall k \in \mathcal{G}, \qquad (1)$$

i.e., the cumulative distribution function (CDF) of the predicted scores in group k. Let $[m] := \{1, \ldots, m\}$. For a convex function $f(\cdot)$, let $\partial f(\cdot)$ be its subdifferential. Let $\|\cdot\|$ be the Euclidean norm and $\mathcal{N}_{\mathcal{W}}(\mathbf{w})$ be the normal cone of $\mathcal{W}$ at $\mathbf{w} \in \mathcal{W}$.

We say a model satisfies *statistical parity* (also called *distributional parity*) when the distribution of $h_{\mathbf{w}}(\boldsymbol{\xi})$ is independent of γ, which means

$$\begin{aligned} &\mathrm{SP}_{k,k'}(\mathbf{w}) \\ &:= \max_{\theta} |\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta | \gamma = k) - \Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta | \gamma = k')| \\ &= 0, \ \forall k, k' \in \mathcal{G}. \end{aligned} \tag{2}$$

For a binary classification task where $\zeta \in \{1, -1\}$, a prediction $\widehat{\zeta}$ can be generated with a predetermined threshold $\widehat{\theta}$ (typically $\widehat{\theta} = 0$) with $\widehat{\zeta} = 1$ if $h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta}$ and $\widehat{\zeta} = -1$, otherwise. In this case, (2) is often relaxed to the *demographic parity*, i.e.,

$$\begin{aligned} &\mathrm{DP}_{k,k'}(\mathbf{w}) \\ &:= \left|\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta} | \gamma = k) - \Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta} | \gamma = k')\right| \\ &= 0, \ \forall k, k' \in \mathcal{G}. \end{aligned} \tag{3}$$

4 PARTIAL FAIRNESS

We propose new criteria of fairness where the fairness is only enforced among the instances within a certain percentile interval in different groups. For simplicity, in this and the next sections, we assume that $h_{\mathbf{w}}(\boldsymbol{\xi})$ has an absolutely continuous distribution conditioning on $\gamma = k$ for each $k \in \mathcal{G}$.

Definition 4.1. For $k \in \mathcal{G}$ and $\theta \in \mathbb{R}$, we call $\bar{F}_{\mathbf{w},k}(\theta)$ the *rank* of score θ in group k, where

$$\bar{F}_{\mathbf{w},k}(\theta) := \Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta | \gamma = k) = 1 - F_{\mathbf{w},k}(\theta).$$

Note that $\bar{F}_{\mathbf{w},k}(\theta)$ is also called *the complementary cumulative distribution function* (CCDF) of $h_{\mathbf{w}}(\boldsymbol{\xi})$ over group k. By definition, $\bar{F}_{\mathbf{w},k}(\theta) = p$ means θ is the $100(1-p)$th percentile of the scores in group k. Then, we propose a generalization of (2) as follows.

Definition 4.2. We say $h_{\mathbf{w}}(\boldsymbol{\xi})$ satisfies the *partial statistical parity (pSP) with respect to percentile interval* $\mathcal{I} = [\alpha, \beta) \subset [0, 1]$ if

$$\begin{aligned} &\mathrm{SP}^{\mathcal{I}}_{k,k'}(\mathbf{w}) \\ &:= \max_{\theta} \left| \begin{array}{l} \Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta \big| \gamma = k, \bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\big) \\ -\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta \big| \gamma = k', \bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\big) \end{array} \right| \\ &= 0, \ \forall k, k' \in \mathcal{G}. \end{aligned} \tag{4}$$

For example, when $\mathcal{I} = [0\%, 20\%)$, the condition $\bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}$ means score $h_{\mathbf{w}}(\boldsymbol{\xi})$ is among the top 20% scores in group k, and partial statistical parity requires the top 20% scores in each group to have the same distribution but puts no restriction on the bottom 80% scores in any group. When $\mathcal{I} = [0, 1]$, (4) is reduced to (2). Similarly, (3) can also be generalized for partial fairness.

Definition 4.3. We say $h_{\mathbf{w}}(\boldsymbol{\xi})$ satisfies the *partial demographic parity (pDP) with respect to percentile interval* $\mathcal{I} = [\alpha, \beta) \subset [0, 1]$ and threshold $\widehat{\theta}$ if

$$\begin{aligned} &\mathrm{DP}^{\mathcal{I}}_{k,k'}(\mathbf{w}) \\ &:= \left| \begin{array}{l} \Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta} \big| \gamma = k, \bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\big) \\ -\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta} \big| \gamma = k', \bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\big) \end{array} \right| \\ &= 0, \ \forall k, k' \in \mathcal{G}. \end{aligned} \tag{5}$$

We target on an in-processing method to produce a $\mathbf{w} \in \mathcal{W}$ that satisfies condition (4) or (5) by solving the following optimization problems subject to fairness constraints:

$$\min_{\mathbf{w} \in \mathcal{W}} f_0(\mathbf{w}) \ \text{ s.t. } \ \mathrm{SP}^{\mathcal{I}}_{k,k'}(\mathbf{w}) \le \kappa, \ \forall k, k' \in \mathcal{G}, \tag{6}$$

$$\min_{\mathbf{w} \in \mathcal{W}} f_0(\mathbf{w}) \ \text{ s.t. } \ \mathrm{DP}^{\mathcal{I}}_{k,k'}(\mathbf{w}) \le \kappa, \ \forall k, k' \in \mathcal{G}, \tag{7}$$

where $f_0(\mathbf{w})$ represents the loss function defined on training data, which is designed such that minimizing $f_0(\mathbf{w})$ optimizes the predictive performance of $h_{\mathbf{w}}(\boldsymbol{\xi})$. The right-hand side $\kappa \in [0, 1]$ is the allowed violation of the fairness constraints, which is specified by the users of our techniques based on their tolerance to unfairness according to the applications. Our model can be extended to the partial variants of other fairness metrics, e.g., the **group AUC fairness** (Vogel et al., 2021; Yao et al., 2022), **equality of opportunity** (Hardt et al., 2016), and **equality of odds** (Hardt et al., 2016). See Section A.

Parameter κ and interval $\mathcal{I}$ are chosen by users of our technique based on their specific fairness concerns while the algorithm and theoretical results we develop are applicable to any choice. When users do not have a targeted interval, we provide some practical guidelines for classification problem in Section B.

5 OPTIMIZATION MODELS

There are two challenges for solving (6) and (7). First, the probabilities in (4) and (5) are defined on the unknown population distribution. In practice, we must approximate them using a training dataset, denoted by $\mathcal{D} = \{(\boldsymbol{\xi}_i, \zeta_i, \gamma_i)\}_{i=1}^n$. Second, approximating the probabilities using the empirical distribution over $\mathcal{D}$ will introduce discontinuous 0-1 indicator functions in (6) and (7). To obtain a continuous optimization problem, we have to further approximate the indicator

function by a continuous surrogate, which often leads to non-convexity and non-smoothness. In this section, we will follow these two steps to derive a data-driven continuous approximation of (6) and (7).

We first provide an equivalent formulation of the constraint in (6) whose proof is in Section C.

Lemma 5.1. *A solution $\mathbf{w} \in \mathcal{W}$ is feasible to (6) if and only if, for any $p \in [\alpha, \beta - \kappa(\beta - \alpha))$, there exists $\theta_p \in \mathbb{R}$ such that*

$$\begin{aligned} &\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta_p | \gamma = k) \geq p \quad \textit{and} \\ &\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta_p | \gamma = k) \leq p + \kappa(\beta - \alpha), \quad \forall k \in \mathcal{G}. \end{aligned} \tag{8}$$

However, there are uncountably many constraints in (8), which are computationally intractable. To derive a tractable approximation of (8), we first partition the training data into subsets based on γ, namely, $\mathcal{D} = \cup_{k \in \mathcal{G}} \mathcal{D}_k$, where

$$\mathcal{D}_k = \{(\boldsymbol{\xi}, \zeta, \gamma) \in \mathcal{D} | \gamma = k\} = \{(\boldsymbol{\xi}_i^k, \zeta_i^k, k)\}_{i=1}^{n_k},$$

and n_k is the size of data from group k. Then, we approximate the probabilities in (8) using the expectation of the indicator $\mathbf{1}_{h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta}$ over the empirical distribution over $\mathcal{D}_k$. Additionally, we create a finite subset $\widehat{\mathcal{I}} \subset [\alpha, \beta - \kappa(\beta - \alpha))$ and only impose the constraints in (8) with $p \in \widehat{\mathcal{I}}$. This leads to an approximation of (8) that requires, for any $p \in \widehat{\mathcal{I}}$, there exists $\theta_p \in \mathbb{R}$ such that

$$\begin{aligned} &\tfrac{1}{n_k} \textstyle\sum_{i=1}^{n_k} \mathbf{1}_{h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) > \theta_p} \geq p \quad \text{and} \\ &\tfrac{1}{n_k} \textstyle\sum_{i=1}^{n_k} \mathbf{1}_{h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) > \theta_p} \leq p + \kappa(\beta - \alpha), \quad \forall k \in \mathcal{G}. \end{aligned}$$

Next, we approximate the indicator function above by a continuous surrogate $\sigma(x) \approx \mathbf{1}_{x>0}$ and obtain the final approximation of (6):

$$\begin{aligned} &\min_{\mathbf{w} \in \mathcal{W},\ (\theta_p)_{p \in \widehat{\mathcal{I}}}} f_0(\mathbf{w}) \quad \text{s.t.} \\ &\frac{1}{n_k} \sum_{i=1}^{n_k} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - \theta_p) \geq p, \forall p \in \widehat{\mathcal{I}}, k \in \mathcal{G}, \\ &\frac{1}{n_k} \sum_{i=1}^{n_k} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - \theta_p) \leq p + \kappa(\beta - \alpha), \forall p \in \widehat{\mathcal{I}}, k \in \mathcal{G}, \end{aligned} \tag{9}$$

where $(\theta_p)_{p \in \widehat{\mathcal{I}}} \in \mathbb{R}^{|\widehat{\mathcal{I}}|}$ is the auxiliary decision variable. The examples of σ include $\sigma(x) = \max\{\min\{x + 0.5, 1\}, 0\}$ and $\sigma(x) = \exp(x)/(1 + \exp(x))$.

We next provide an equivalent formulation of the constraint in (7) whose proof is in Section C also.

Lemma 5.2. *A solution $\mathbf{w} \in \mathcal{W}$ is feasible to (7) if and only if, for any k and k' in $\mathcal{G}$,*

$$\left| \begin{array}{l} \min\left\{\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta} \big| \gamma = k\big), \beta\right\} \\ -\min\left\{\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta} \big| \gamma = k\big), \alpha\right\} \\ -\min\left\{\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta} \big| \gamma = k'\big), \beta\right\} \\ +\min\left\{\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta} \big| \gamma = k'\big), \alpha\right\} \end{array} \right| \leq \kappa(\beta - \alpha), \tag{10}$$

Following the same steps, we approximate the probabilities in (10) using data $\mathcal{D}$ and approximate the indicator function using a surrogate $\sigma(\cdot)$. The final approximation of (7) is as follows.

$$\begin{aligned} &\min_{\mathbf{w} \in \mathcal{W}} f_0(\mathbf{w}) \\ &\text{s.t.} \begin{array}{l} \min\left\{\frac{1}{n_k}\sum_{i=1}^{n_k} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - \widehat{\theta}), \beta\right\} \\ -\min\left\{\frac{1}{n_k}\sum_{i=1}^{n_k} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - \widehat{\theta}), \alpha\right\} \\ -\min\left\{\frac{1}{n_{k'}}\sum_{i=1}^{n_{k'}} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k'}) - \widehat{\theta}), \beta\right\} \\ +\min\left\{\frac{1}{n_{k'}}\sum_{i=1}^{n_{k'}} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k'}) - \widehat{\theta}), \alpha\right\} \leq \kappa(\beta - \alpha), \end{array} \\ &\forall k, k' \in \mathcal{G}. \end{aligned} \tag{11}$$

Although (9) and (11) are formulated very differently, we can show that, under mild assumptions on f_0, σ and $h_{\mathbf{w}}(\cdot)$, both problems can be formulated as the following optimization model.

$$\begin{aligned} &\min_{\mathbf{w} \in \mathcal{W}} f_0(\mathbf{w}) := f_0^+(\mathbf{w}) - f_0^-(\mathbf{w}) \\ &\text{s.t.} \ \ f_i(\mathbf{w}) := f_i^+(\mathbf{w}) - f_i^-(\mathbf{w}) \leq 0, \ i \in [m], \end{aligned} \tag{12}$$

where functions in $\{f_i^+\}_{i=0}^m$ and $\{f_i^-\}_{i=0}^m$ are real-valued, convex but potentially non-smooth. Since each f_i is the difference of two convex functions, (12) is a difference-of-convex (DC) constrained optimization problem. In Section D, we present some specific examples of (9) and (11) that are instances of (12).

Although the primary focus of this work is on optimization methods for enforcing partial fairness, we also provide statistical guarantees for models obtained from (9). In particular, we analyze the generalization error arising from three sources: the discretization of $\mathcal{I}$, the surrogate approximation $\sigma(x) \approx \mathbf{1}_{x>0}$, and sampling error. The results are presented in Section F.

Remark 5.3. Definitions 4.2 and 4.3 can be easily extended to settings where the percentile interval $\mathcal{I}$ varies across groups, as well as to the case where $\mathcal{I}$ is not a single interval but a union of multiple disjoint intervals. In both cases, the corresponding optimization models (9) and (11) can be naturally extended, formulated as DC programs, and efficiently solved within the IDCA framework.

6 INEXACT DCA

Since (12) is non-convex, finding an optimal or ϵ-optimal solution of (12) is intractable in general. Therefore, the recent studies on non-convex optimization are mainly motivated to find an ϵ-Karush–Kuhn–Tucker (ϵ-KKT) solution.

Definition 6.1. A point $\mathbf{w} \in \mathcal{W}$ is an *ϵ-KKT point* of (12) if there exist $\mathbf{n}_{\mathcal{W}} \in \mathcal{N}_{\mathcal{W}}(\mathbf{w})$, $\mathbf{f}_i^+ \in \partial f_i^+(\mathbf{w})$,

$\mathbf{f}_i^- \in \partial f_i^-(\mathbf{w})$, and $\lambda_i \geq 0$ for $i \in [m]$ satisfying

$$\left\|\mathbf{f}_0^+ - \mathbf{f}_0^- + \sum_{i=1}^m \lambda_i(\mathbf{f}_i^+ - \mathbf{f}_i^-) + \mathbf{n}_{\mathcal{W}}\right\| \leq \epsilon, \quad (13)$$

$$f_i^+(\mathbf{w}) - f_i^-(\mathbf{w}) \leq \epsilon,\ i \in [m], \quad (14)$$

$$|\lambda_i(f_i^+(\mathbf{w}) - f_i^-(\mathbf{w}))| \leq \epsilon,\ i \in [m]. \quad (15)$$

However, when a problem is non-smooth, even finding an ϵ-KKT point is intractable in general as the set of ϵ-KKT points often has a zero measure (e.g., $\min_{\mathbf{w}} \|\mathbf{w}\|$). Following the recent literature on non-convex non-smooth optimization, we consider the following weaker class of solutions.

Definition 6.2. A point $\mathbf{w} \in \mathcal{W}$ is a *nearly ϵ-KKT point* of (12) if there exist $\widehat{\mathbf{w}} \in \mathcal{W}$, $\mathbf{n}_{\mathcal{W}} \in \mathcal{N}_{\mathcal{W}}(\widehat{\mathbf{w}})$, $\mathbf{f}_i^+ \in \partial f_i^+(\widehat{\mathbf{w}})$, $\mathbf{f}_i^- \in \partial f_i^-(\mathbf{w})$, and $\lambda_i \geq 0$ for $i \in [m]$ satisfying $\|\widehat{\mathbf{w}} - \mathbf{w}\| \leq \epsilon$, (13), (14) and (15).

The only difference between an ϵ-KKT point and a nearly ϵ-KKT point is that the former satisfies the stationarity condition (13) only with its own subgradients while the latter satisfies (13) with its own subgradients and the subgradients of a nearby point (i.e., $\widehat{\mathbf{w}}$).

We use an inexact diffference-of-convex algorithm (IDCA) to find a nearly ϵ-KKT point of (12). At the k-th iteration of IDCA, if the current solution is $\mathbf{w}^{(k)} \in \mathcal{W}$, we linearize the concave component of each of the objective and constraint functions in (12) at $\mathbf{w}^{(k)}$ to obtain the convex approximation of (12) below

$$\begin{aligned} \min_{\mathbf{v}\in\mathcal{W}}\ & g_0(\mathbf{v}) := f_0^+(\mathbf{v}) - \widehat{f}_0^-(\mathbf{v};\mathbf{w}^{(k)}) \quad (16)\\ \text{s.t.}\ & g_i(\mathbf{v}) := f_i^+(\mathbf{v}) - \widehat{f}_i^-(\mathbf{v};\mathbf{w}^{(k)}) \leq 0,\ i \in [m], \end{aligned}$$

where

$$\widehat{f}_i^-(\mathbf{v};\mathbf{w}) := f_i^-(\mathbf{w}) + (\mathbf{f}_i^-)^\top(\mathbf{v} - \mathbf{w}) \quad (17)$$

and $\mathbf{f}_i^-$ is an arbitrary subgradient in $\partial f_i^-(\mathbf{w})$. Note that, by the convexity of f_i^-, we have $f_i(\cdot) \leq g_i(\cdot)$ for $i = 0, 1, \ldots, m$. Then IDCA proceeds to generate the next iterate $\mathbf{w}^{(k+1)}$ as an ϵ_k-optimal solution of (16), meaning that

$$\max\left\{g_0(\mathbf{w}^{(k+1)}) - g_0(\widehat{\mathbf{w}}^{(k+1)}), g(\mathbf{w}^{(k+1)})\right\} \leq \epsilon_k, \quad (18)$$

where $g(\cdot) := \max_{i\in[m]} g_i(\cdot)$ and $\widehat{\mathbf{w}}^{(k+1)}$ is the optimal solution of (16). It is formally presented in Algorithm 1.

Since (16) is convex, we will directly apply the switching subgradient (SSG) method (Nesterov, 2018, (3.2.24)) to (16) to find $\mathbf{w}^{(k+1)}$. This method is presented in Algorithm 2. One of the inputs of SSG is an infeasibility tolerance, denoted by ϵ. At iteration t of SSG, if the current solution $\mathbf{v}^{(t)}$ is nearly feasible, i.e., $g_i(\mathbf{v}^{(t)}) \leq \epsilon$, $\mathbf{v}^{(t)}$ will be updated along the subgradient of g_0. Otherwise, $\mathbf{v}^{(t)}$ will be updated along the subgradient of $\max_{i\in[m]} g_i$. The final output is the nearly feasible solution that has the smallest objective value $g_0(\mathbf{v}^{(t)})$.

Algorithm 1 Inexact Difference-of-Convex Algorithm (IDCA)

Input: $\mathbf{w}^{(0)}$ feasible to (12), the number of iterations K, precision ϵ_k, $k = 0, 1, \ldots, K-1$.
for $k = 0$ **to** $K - 1$ **do**
  Find $\mathbf{w}^{(k+1)}$ that satisfies (18).
end for

Algorithm 2 Switching Subgradient (SSG) Method for (16): SSG($\mathbf{v}^{(0)}$,ϵ, T)

Input: Infeasibility tolerance $\epsilon > 0$, initial solution $\mathbf{v}^{(0)} \in \mathcal{W}$ with $g_i(\mathbf{v}^{(0)}) \leq \epsilon$ for $i \in [m]$ and the number of iterations T.
$\mathcal{T} \leftarrow \emptyset$.
for $t = 0$ **to** $T - 1$ **do**
  if $g(\mathbf{v}^{(t)}) \leq \epsilon$ **then**
    $\mathbf{v}^{(t+1)} \leftarrow \mathrm{Proj}_{\mathcal{W}}(\mathbf{v}^{(t)} - \frac{\epsilon}{\|\mathbf{g}_0^{(t)}\|^2}\mathbf{g}_0^{(t)})$ for any $\mathbf{g}_0^{(t)} \in \partial g_0(\mathbf{v}^{(t)})$. $\mathcal{T} \leftarrow \mathcal{T} \cup \{t\}$.
  else
    $\mathbf{v}^{(t+1)} \leftarrow \mathrm{Proj}_{\mathcal{W}}(\mathbf{v}^{(t)} - \frac{g(\mathbf{v}^{(t)})}{\|\mathbf{g}^{(t)}\|^2}\mathbf{g}^{(t)})$ for any $\mathbf{g}^{(t)} \in \partial g(\mathbf{v}^{(t)})$.
  end if
end for
Output: $\mathbf{v}^{(\tau)}$ with $\tau = \arg\min_{t\in\mathcal{T}} g_0(\mathbf{v}^{(t)})$

The efficiency of Algorithm 1 is measured by its *oracle complexity*, which is the total number of subgradients computed before finding a nearly ϵ-KKT point of (12). In order to establish the oracle complexity of Algorithm 1, we make the following assumptions on (12).

Assumption 6.3. The following statements hold.

A. Functions in $\{f_i^+\}_{i=0}^m$ and $\{f_i^-\}_{i=0}^m$ are real-valued and μ-strongly convex on $\mathbb{R}^n$.

B. There exists constant $\nu > 0$ such that, if $f_i(\mathbf{w}) \leq 0$ for $i \in [m]$, there exists $\mathbf{v} \in \mathcal{W}$ such that $f_i^+(\mathbf{v}) - \widehat{f}_i^-(\mathbf{v};\mathbf{w}) \leq -\nu,\ \forall i \in [m]$.

C. $f_{\mathrm{lb}} := \min_{\mathbf{w}\in\mathcal{W}} f_0(\mathbf{w}) \geq -\infty$.

D. $f_i^+(\mathbf{w})$ and $f_i^-(\mathbf{w})$ are M-Lipchitz continuous on $\mathcal{W}$ for $i = 0, 1, \ldots, m$.

E. A solution $\mathbf{w}_{\mathrm{feas}}$ satisfying $f_i(\mathbf{w}_{\mathrm{feas}}) \leq 0$ for $i \in [m]$ is accessible.

The strong convexity assumption in Assumption 6.3A can be easily satisfied by replacing $f_i^+(\mathbf{w})$ and $f_i^-(\mathbf{w})$ with $f_i^+(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{w}\|^2$ and $f_i^-(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{w}\|^2$, respectively, if necessary. Assumption 6.3B is the generalization of

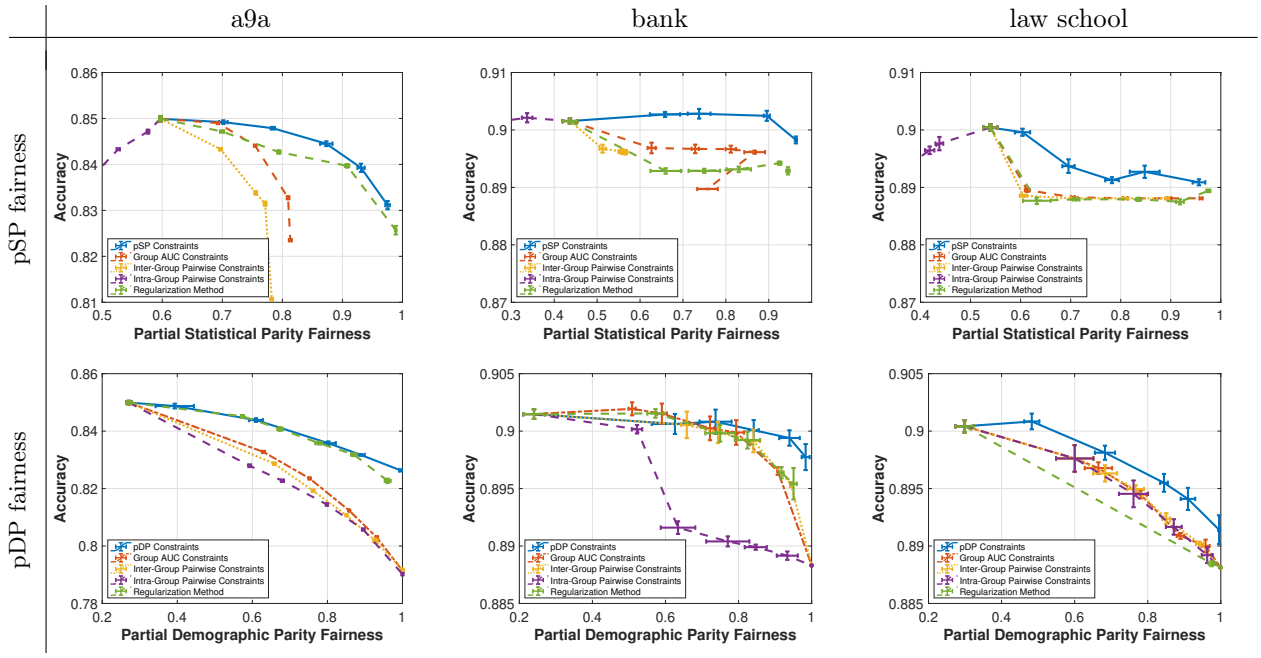

Figure 1: Pareto frontiers showing the trade-off between classification accuracy and pSP fairness (the first row) and the trade-off between classification accuracy and pDP fairness (the second row) by each method. $\mathcal{I}$ is $[5\%, 30\%]$ for *a9a*, $[0\%, 25\%]$ for *bank*, and $[70\%, 100\%]$ for *law school*.

the uniform Slater's condition by Ma et al. (2020); Huang and Lin (2023). Two real-world examples satisfying this assumption, including a fair classification problem, are given by Huang and Lin (2023), where $f_i^-(\mathbf{w}) = \frac{\rho}{2}\|\mathbf{w}\|^2$ for some $\rho > 0$. Assumption 6.3C and D are common in the literature. Assumption 6.3E is not restricted because any $\mathbf{w}_{\text{feas}}$ making $h_{\mathbf{w}_{\text{feas}}}(\boldsymbol{\xi}) \equiv 0$ is a trivially fair model and thus satisfies most fairness constraints, including (9) and (11). Under Assumption 6.3B, subproblem (16) is feasible as long as $\mathbf{w}^{(k)}$ is feasible to (12).

By a proof similar to Nesterov (2018, Theorem 3.2.3), the oracle complexity of Algorithm 2 for finding $\mathbf{w}^{(k+1)}$ in (18) is known as follows. The proof is provided in Section E.1.

Proposition 6.4. *For any k, if $f_i(\mathbf{w}^{(k)}) \le 0$ for $i \in [m]$ and $\mathbf{w}^{(k+1)} = \text{SSG}(\mathbf{w}^{(k)}, \epsilon_k, T_k)$ with $T_k \ge \frac{M^2}{\epsilon_k \mu} \ln\left(1 + \frac{\mu\|\mathbf{w}^{(k)} - \widehat{\mathbf{w}}^{(k+1)}\|^2}{\epsilon_k}\right)$, $\mathbf{w}^{(k+1)}$ satisfies (18).*

The oracle complexity of Algorithm 1 for finding a nearly ϵ-KKT point of (12) is as follows. The proof is in Section E.5.

Theorem 6.5. *Suppose*

$$K = \left\lceil 2\max\left\{1, 4M^2, \frac{8M^4}{\mu\nu}\right\} \frac{\left(f_0(\mathbf{w}^{(0)}) - f_{\text{lb}}\right)}{\mu\epsilon^2} \right\rceil, \tag{19}$$

$$\epsilon_k = \frac{\mu}{8} \min\left\{1, \frac{1}{4M^2}, \frac{\mu\nu}{8M^4}\right\} \cdot \epsilon^2, \tag{20}$$

and set $\mathbf{w}^{(k+1)} = \text{SSG}(\mathbf{w}^{(k)}, \epsilon_k, T_k)$ along with $T_k = \frac{M^2}{\epsilon_k \mu} \ln\left(1 + \frac{\mu\|\mathbf{w}^{(k)} - \widehat{\mathbf{w}}^{(k+1)}\|^2}{\epsilon_k}\right)$. Algorithm 1 finds a nearly ϵ-KKT point for (12) with oracle complexity $\tilde{O}(\epsilon^{-4})$.

7 NUMERICAL EXPERIMENTS

In this section, we demonstrate the effectiveness of the proposed approach for training partially fair models for binary classification. The experiments are conducted on a MacBook Air (Apple M1 chip, 8-core CPU and 8GB of memory) in MATLAB using three publicly available datasets: *a9a* (Chang and Lin, 2011; Kohavi, 1996), *bank* (Chang and Lin, 2011; Moro et al., 2014), *law school* (Wightman, 1998) and *American Community Survey (ACS)* dataset (Ding et al., 2021). More details of the datasets can be found in Section G.1.

We compare our methods with other in-processing methods that solve constrained optimization problems with the same objective function as our approach but different fairness constraints, including the constraints that enforce **Group AUC Fairness** (Yao et al., 2023; Vogel et al., 2021), **Inter-Group Pairwise Fairness** (Kallus and Zhou, 2019; Beutel et al., 2019), and **Intra-Group Pairwise Fairness** (Beutel et al., 2019). We present the optimization models of these methods and their training procedures in Section G.3. Additionally, we compare our model with a regularization method whose details are given in Section G.4.

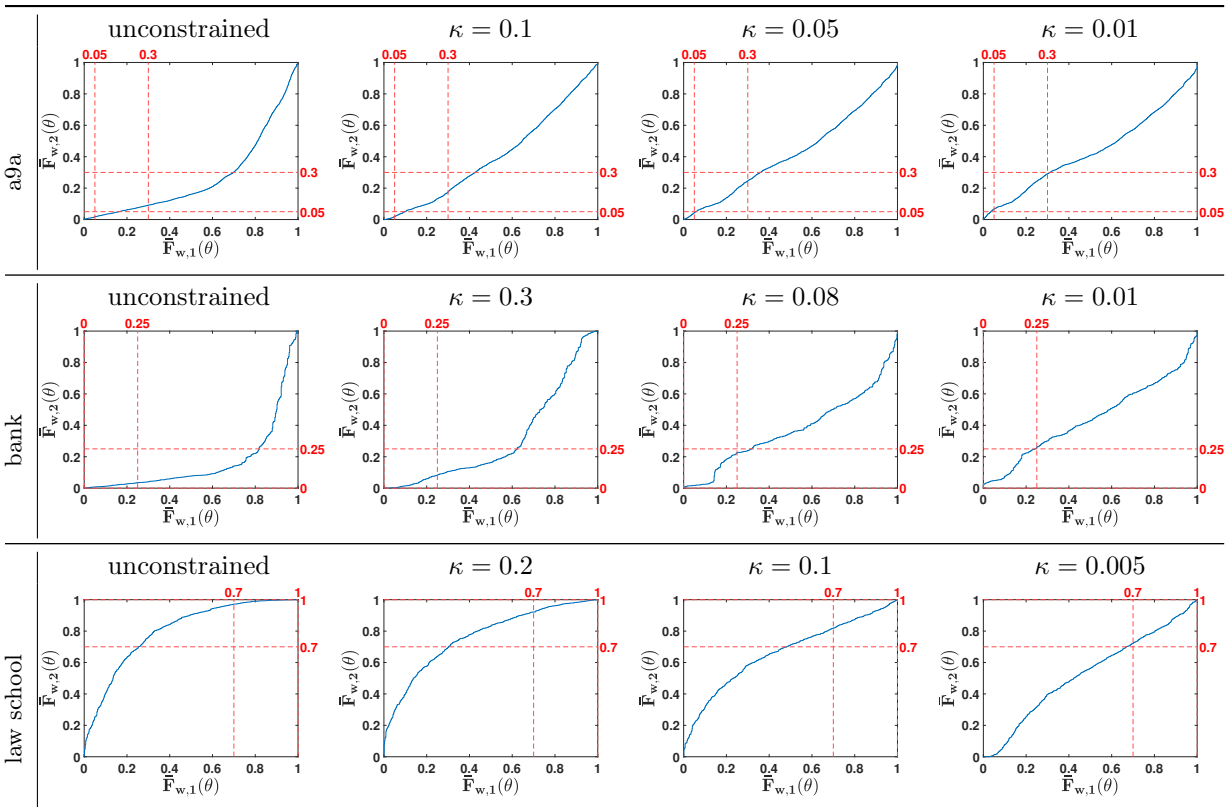

Figure 2: How the predicted positive rates in both groups ($\bar{F}_{\mathbf{w},1}(\theta)$ and $\bar{F}_{\mathbf{w},2}(\theta)$) change with threshold θ when using different κ in (9) by the proposed method. $\mathcal{I}$ is $[5\%, 30\%]$ for *a9a*, $[0\%, 25\%]$ for *bank*, and $[70\%, 100\%]$ for *law school* (between the two red numbers marked on both axises).

Using Pareto frontiers, we compare the efficiency of these methods in trading off prediction performance and partial fairness of the output models. Additional numerical results are presented in Sections G.5, G.6 and G.7.

7.1 Partial Statistical Parity

We consider a linear model with cross terms, i.e., $h_{\mathbf{w}}(\boldsymbol{\xi}) = \langle \mathbf{w}, (1, \boldsymbol{\xi}, \gamma, \gamma \cdot \boldsymbol{\xi}) \rangle$, and optimize its classification accuracy subject to **partial statistical parity (pSP) constraints** with different tolerances to violation. In particular, we apply IDCA (Algorithm 1) to (9) with $f_0(\mathbf{w})$ being the empirical loss in (48), ℓ being the logistic loss in Example D.1 and $\mathcal{W} = \mathbb{R}^{2d+2}$. We choose $\sigma(x) = \max\{x + \frac{1}{2}, 0\} - \max\{x - \frac{1}{2}, 0\}$ in (9) just as Example D.3. The interval $\mathcal{I} = [\alpha, \beta)$ is chosen for each dataset in the way described in Section 4 with $p = \Pr(h_{\mathbf{w}^*}(\boldsymbol{\xi}) > 0)$, and is stated in the caption of each figure. We set $\widehat{\mathcal{I}}$ to be a set of ten equally spaced values in $\mathcal{I}$. To create the Pareto frontier, we choose different values for κ in (9) (see Section G.2.1). The implementation details, such as choices of tuning parameters, for IDCA are presented in Section G.2.2.

For each model returned by each method, we evaluate its classification accuracy and fairness in terms of $1 - \max_{k,k} \mathrm{SP}^{\mathcal{I}}_{k,k'}(\mathbf{w})$ on the testing set. We randomly split the data into training, validation, and testing sets with five random seeds, and evaluate the means of both performance metrics along with their 95% confidence intervals (error bars). The results are presented in Figure 1 (pSP fairness), where the y-axis represents accuracy and the x-axis represents fairness. Note that the Pareto frontier of each method starts from the same unconstrained solution, i.e., $\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathcal{W}} f_0(\mathbf{w})$. On all three datasets, our method achieves a higher accuracy than all the baselines at all levels of fairness, showing the effectiveness of our methods in trading accuracy for partial fairness. The same experiment is also conducted on the ACS dataset and the results are presented in Section G.8.

We also evaluate the effectiveness and robustness of our method on a deep learning task. Specifically, we consider an image classification problem formulated as (9) on the CelebA dataset (Liu et al., 2015). In this setting, $h_{\mathbf{w}}(\boldsymbol{\xi})$ is set as a ResNet-18 model (a CNN pretrained on ImageNet) and fine-tuned to predict the "Attractive" attribute, while enforcing a partial statistical parity constraint with gender as the sensitive attribute. Experimental details and results are pro-

vided in Section G.9.

To demonstrate the effectiveness of our method in enforcing similar distributions of scores between groups, we compare the predicted positive rates between the groups at different thresholds θ. Specifically, similar to an ROC curve, we plot $\bar{F}_{\mathbf{w},2}(\theta)$ against $\bar{F}_{\mathbf{w},1}(\theta)$ (see Definition 4.1) in Figure 2 with θ changing between $+\infty$ and $-\infty$. We show the curves of the unconstrained model $\mathbf{w}^*$ and the models developed by our method with different κ in (9). According to Figure 2, the unconstrained model is highly unfair as $\bar{F}_{\mathbf{w},1}(\theta)$ and $\bar{F}_{\mathbf{w},2}(\theta)$ increase at very different speeds within the targeted interval $\mathcal{I}$ (between the two red numbers marked on both axises). In our method, as κ decreases, $\bar{F}_{\mathbf{w},1}(\theta)$ and $\bar{F}_{\mathbf{w},2}(\theta)$ become closer and eventually increase at a similar speed, producing a curve of slope close to one (45-degree angle) within $\mathcal{I}$, which indicates the pSP fairness is mostly achieved.

7.2 Partial Demographic Parity

We consider the same linear model as in Section 7.1. This time, we want to optimize its classification accuracy subject to **partial demographic parity (pDP) constraints** with different tolerances to violation. In particular, we apply IDCA (Algorithm 1) to (11) with $\widehat{\theta} = 0$, where $f_0(\cdot)$, $\sigma(\cdot)$, $\mathcal{I}$, and $\mathcal{W}$ are the same as in Section 7.1. The values of κ used to create the Pareto frontiers can be found in Section G.2.1 and the implementation details of IDCA are given in Section G.2.2.

For each model returned by each method, we evaluate its classification accuracy and fairness in terms of $1 - \max_{k,k} \mathrm{DP}^{\mathcal{I}}_{k,k'}(\mathbf{w})$ on the testing set. The data is split in the same way as in Section 7.1 to create the error bars. The results are presented in Figure 1 (pDP fairness), where the y-axis represents accuracy and the x-axis represents fairness. We observe that, on the *law school* dataset, our method achieves a higher accuracy than all the baselines for all levels of fairness. On the *a9a* and *bank* datasets, our method achieves the highest accuracy at the high levels of fairness (e.g., ≥ 0.85) and achieves similar accuracy to the baselines at the medium and low levels of fairness.

8 CONCLUSION

We introduced a novel in-processing framework for achieving partial fairness in machine learning only in critical regions, providing more flexibility than enforcing fairness across the entire score range. We formulate the problem as DC optimization with partial fairness constraints. The IDCA is introduced as a training algorithm, for which we prove novel theoretical guarantees and demonstrate strong empirical performance.

Acknowledgment

The authors would like to thank all the anonymous reviewers for their valuable comments and suggestions. This work was supported by the National Science Foundation under Grant No. 2147253 and No. 2246757.

References

Abbaszadehpeivasti, H., de Klerk, E., and Zamani, M. (2024). On the rate of convergence of the difference-of-convex algorithm (DCA). *Journal of Optimization Theory and Applications*, 202(1):475–496.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 60–69. PMLR.

Aragón-Artacho, F. J., Campoy, R., and Phan, T. V. (2022). The boosted DC algorithm for linearly constrained DC programming. *Set-Valued and Variational Analysis*, 30(4):1265–1289.

Aragón-Artacho, F. J., Fleming, R. M., and Phan, T. V. (2018). Accelerating the DC algorithm for smooth functions. *Mathematical Programming*, 169(1):95–118.

Aragón-Artacho, F. J. and Phan, T. V. (2020). The boosted difference of convex functions algorithm for nonsmooth functions. *SIAM Journal on Optimization*, 30(1):980–1006.

Bagirov, A. M. and Ugon, J. (2011). Codifferential method for minimizing nonsmooth DC functions. *Journal of Global Optimization*, 50(1):3–22.

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

Beutel, A., Chen, J., Doshi, T., Qian, H., et al. (2019). Fairness in recommendation ranking through pairwise comparisons. In *KDD'15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2212–2220.

Boob, D., Deng, Q., and Lan, G. (2023). Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming*, 197(1):215–279.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of the 2019 World Wide Web Conference*, pages 491–500.

Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building classifiers with independency constraints. In

2009 IEEE International Conference on Data Mining Workshops, pages 13–18. IEEE.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, no. 27.

Chen, Y., Giudici, P., Liu, K., and Raffinetti, E. (2024). Measuring fairness in credit scoring. *Expert Systems with Applications*, 258:1–13, no. 125184.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2020a). Fair regression via plug-in estimator and recalibration with statistical guarantees. *Advances in Neural Information Processing Systems*, 33:19137–19148.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2020b). Fair regression with Wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33:7321–7331.

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

Cruz Neto, J., Oliveira, P. R., Soubeyran, A., and Souza, J. C. O. (2020). A generalized proximal linearized algorithm for DC functions with application to the optimal size of the firm problem. *Annals of Operations Research*, 289(2):313–339.

de Oliveira, W. and Tcheou, M. P. (2019). An inertial algorithm for DC programming. *Set-Valued and Variational Analysis*, 27(4):895–919.

Ding, F., Hardt, M., Miller, J. P., and Schmidt, L. (2021). Retiring Adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34:6478–6490.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. (2018). Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems*, 31:2791–2801.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *KDD'15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268.

Ferreira, O. P., Mordukhovich, B. S., Santos, W., and Souza, J. C. O. (2026). An inexact boosted difference of convex algorithm for nondifferentiable functions. *Journal of Optimization Theory and Applications*, 208(2):1–27, no. 71.

Ferreira, O. P., Santos, E. M., and Souza, J. C. O. (2024). A boosted DC algorithm for nondifferentiable DC components with non-monotone line search. *Computational Optimization and Applications*, 88(3):783–818.

Goh, G., Cotter, A., Gupta, M., and Friedlander, M. P. (2016). Satisfying real-world goals with dataset constraints. *Advances in Neural Information Processing Systems*, 29:2415–2423.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29:3315–3323.

Hiriart-Urruty, J.-B. (1985). Generalized differentiability/duality and optimization for problems dealing with differences of convex functions. In Ponstein, J., editor, *Convexity and Duality in Optimization: Proceedings, Groningen, The Netherlands, 1984*, volume 256 of *Lecture Notes in Economics and Mathematical Systems*, pages 37–70. Springer-Verlag.

Hoang, T. (1984). Global minimization of a difference of two convex functions. In Hammer, G. and Pallaschke, D., editors, *Selected Topics in Operations Research and Mathematical Economics*, volume 226 of *Lecture Notes in Economics and Mathematical Systems*, pages 98–118. Springer-Verlag.

Hu, Q., Qi, Q., Lu, Z., and Yang, T. (2024). Single-loop stochastic algorithms for difference of max-structured weakly convex functions. *Advances in Neural Information Processing Systems*, 37:56738–56765.

Huang, Y. and Lin, Q. (2023). Oracle complexity of single-loop switching subgradient methods for nonsmooth weakly convex functional constrained optimization. *Advances in Neural Information Processing Systems*, 36:61327–61340.

Jia, Z. and Grimmer, B. (2025). First-order methods for nonsmooth nonconvex functional constrained optimization with or without Slater points. *SIAM Journal on Optimization*, 35(2):1300–1329.

Kallus, N. and Zhou, A. (2019). The fairness of risk scores beyond classification: Bipartite ranking and the xAUC metric. *Advances in Neural Information Processing Systems*, 32:3438–3448.

Kanzow, C. and Neder, T. (2024). A bundle-type method for nonsmooth DC programs. *Journal of Global Optimization*, 88(2):285–326.

Khalili, M. M., Zhang, X., Abroshan, M., and Sojoudi, S. (2021). Improving fairness and privacy in selection problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8092–8100.

Khamaru, K. and Wainwright, M. J. (2019). Convergence guarantees for a class of non-convex and non-smooth optimization problems. *Journal of Machine Learning Research*, 20:1–52, no. 154.

Kohavi, R. (1996). Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *KDD'96: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 202–207.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30:4066–4076.

Le Thi, H. A. and Pham Dinh, T. (2018). DC programming and DCA: Thirty years of developments. *Mathematical Programming*, 169(1):5–68.

Le Thi, H. A. and Pham Dinh, T. (2024). Open issues and recent advances in DC programming and DCA. *Journal of Global Optimization*, 88(3):533–590.

Le Thi, H. A., Van Ngai, H., and Pham Dinh, T. (2018). Convergence analysis of difference-of-convex algorithm with subanalytic data. *Journal of Optimization Theory and Applications*, 179(1):103–126.

Le Thi, H. A., Van Ngai, H., and Pham Dinh, T. (2024). Minimizing compositions of differences-of-convex functions with smooth mappings. *Mathematics of Operations Research*, 49(2):1140–1168.

Liang, P. (2016). CS229T/STAT231: Statistical learning theory. Lecture Notes. `https://web.stanford.edu/class/cs229t/notes.pdf`.

Liu, T. and Takeda, A. (2022). An inexact successive quadratic approximation method for a class of difference-of-convex optimization problems. *Computational Optimization and Applications*, 82(1):141–173.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738.

Ma, R., Lin, Q., and Yang, T. (2020). Quadratically regularized subgradient methods for weakly convex optimization with weakly convex constraints. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6554–6564. PMLR.

Maingé, P.-E. and Moudafi, A. (2008). Convergence of new inertial proximal methods for DC programming. *SIAM Journal on Optimization*, 19(1):397–413.

Miroshnikov, A., Kotsiopoulos, K., Franks, R., and Kannan, A. R. (2021). Model-agnostic bias mitigation methods with regressor distribution control for Wasserstein-based fairness metrics. *arXiv preprint arXiv:2111.11259*.

Miroshnikov, A., Kotsiopoulos, K., Franks, R., and Ravi Kannan, A. (2022). Wasserstein-based fairness interpretability framework for machine learning models. *Machine Learning*, 111(9):3307–3357.

Moro, S., Cortez, P., and Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31.

Moudafi, A. (2021). Few remarks on a complete smooth regularization of DC optimization problems. ResearchGate Preprint. `https://www.researchgate.net/publication/352689832`.

Moudafi, A. (2024). DC and split composite minimization problems. ResearchGate Preprint. `https://www.researchgate.net/publication/382703993`.

Moudafi, A. and Maingé, P.-E. (2006). On the convergence of an approximate proximal method for DC functions. *Journal of Computational Mathematics*, 24(4):475–480.

Narasimhan, H., Cotter, A., Gupta, M., and Wang, S. (2020). Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5248–5255.

Nesterov, Y. (2018). *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, 2nd edition.

Pang, J.-S., Razaviyayn, M., and Alvarado, A. (2017). Computing B-stationary points of nonsmooth DC programs. *Mathematics of Operations Research*, 42(1):95–118.

Pang, J.-S. and Tao, M. (2018). Decomposition methods for computing directional stationary solutions of a class of nonsmooth nonconvex optimization problems. *SIAM Journal on Optimization*, 28(2):1640–1669.

Pham Dinh, T. and El Bernoussi, S. (1986). Algorithms for solving a class of nonconvex optimization problems. Methods of subgradients. In Hiriart-Urruty, J.-B., editor, *Fermat Days 85: Mathematics for Optimization*, volume 129 of *North-Holland Mathematics Studies*, pages 249–271. North-Holland.

Pham Dinh, T. and Le Thi, H. A. (1997). Convex analysis approach to DC programming: Theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1):289–355.

Pham Dinh, T. and Le Thi, H. A. (2014). Recent advances in DC programming and DCA. In Le-Thi, H. A. and Nguyen, N.-T., editors, *Transactions on Computational Intelligence XIII*, volume 8342 of *Lecture Notes in Computer Science*, pages 1–37. Springer.

Souza, J. C. O., Oliveira, P. R., and Soubeyran, A. (2016). Global convergence of a proximal linearized algorithm for difference of convex functions. *Optimization Letters*, 10(7):1529–1539.

Sriperumbudur, B. K. and Lanckriet, G. R. (2012). A proof of convergence of the concave-convex procedure using Zangwill's theory. *Neural Computation*, 24(6):1391–1407.

Sun, K. and Sun, X. A. (2023). Algorithms for difference-of-convex programs based on difference-of-Moreau-envelopes smoothing. *INFORMS Journal on Optimization*, 5(4):321–339.

Sun, W.-Y., Sampaio, R. J., and Candido, M. (2003). Proximal point algorithm for minimization of DC function. *Journal of Computational Mathematics*, 21(4):451–462.

van den Bos, K. (2001). Uncertainty management: The influence of uncertainty salience on reactions to perceived procedural fairness. *Journal of Personality and Social Psychology*, 80(6):931–941.

Vogel, R., Bellet, A., and Clémençon, S. (2021). Learning fair scoring functions: Bipartite ranking under ROC-based fairness constraints. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 784–792. PMLR.

Wan, M., Zha, D., Liu, N., and Zou, N. (2023). In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, 17(3):1–27, no. 35.

Wang, R., Harper, F. M., and Zhu, H. (2020). Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, no. 684.

Wightman, L. F. (1998). LSAC national longitudinal bar passage study. *LSAC Research Report Series*. `https://archive.lawschooltransparency.com/reform/projects/investigations/2015/documents/NLBPS.pdf`.

Xian, R., Yin, L., and Zhao, H. (2023). Fair and optimal classification via post-processing. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 37977–38012. PMLR.

Xu, Y., Qi, Q., Lin, Q., Jin, R., and Yang, T. (2019). Stochastic optimization for DC functions and non-smooth non-convex regularizers with non-asymptotic convergence. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 6942–6951. PMLR.

Yang, Z., Ko, Y. L., Varshney, K. R., and Ying, Y. (2023). Minimax AUC fairness: Efficient algorithm with provable convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11909–11917.

Yao, W., Zhou, Z., Li, Z., Han, B., and Liu, Y. (2024). Understanding fairness surrogate functions in algorithmic fairness. *Transactions on Machine Learning Research*, pages 1–49. `https://openreview.net/forum?id=iBgmoMTlaz`.

Yao, Y., Lin, Q., and Yang, T. (2022). Large-scale optimization of partial AUC in a range of false positive rates. *Advances in Neural Information Processing Systems*, 35:31239–31253.

Yao, Y., Lin, Q., and Yang, T. (2023). Stochastic methods for AUC optimization subject to AUC-based fairness constraints. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 10324–10342. PMLR.

Yu, H., Neely, M., and Wei, X. (2017). Online convex optimization with stochastic constraints. *Advances in Neural Information Processing Systems*, 30:1427–1437.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 962–970. PMLR.

Zhang, H. and Niu, Y.-S. (2024). A boosted-DCA with power-sum-DC decomposition for linearly constrained polynomial programs. *Journal of Optimization Theory and Applications*, 201(2):720–759.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

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

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

Enforcing Fair Predicted Scores on Intervals of Percentiles by Difference-of-Convex Constraints: Supplementary Materials

Contents

1 INTRODUCTION **1**

2 RELATED WORKS **2**

3 PRELIMINARY AND NOTATIONS **3**

4 PARTIAL FAIRNESS **4**

5 OPTIMIZATION MODELS **4**

6 INEXACT DCA **5**

7 NUMERICAL EXPERIMENTS **7**

7.1 Partial Statistical Parity 8

7.2 Partial Demographic Parity 9

8 CONCLUSION **9**

A EXTENSIONS TO GROUP AUC FAIRNESS, EQUALITY OF OPPORTUNITY, AND EQUALITY OF ODDS **15**

B HOW TO CHOOSE $\mathcal{I}$ FOR CLASSIFICATION PROBLEM: **17**

C EQUIVALENT FORMULATIONS OF STATISTICAL PARITY AND DEMOGRAPHIC PARITY PARTIAL FAIRNESS CONSTRAINTS **17**

D REFORMULATE TRAINING PROBLEMS WITH PARTIAL FAIRNESS CONSTRAINTS AS DC OPTIMIZATION **19**

D.1 DC Objective Function in (9) and (11) 20

D.2 DC Constraints in (9) 21

D.3 DC Constraints in (11) 22

D.4 DC Constraints in (30), (31) and (32) 23

E PROOFS FOR COMPLEXITY ANALYSIS **25**

E.1 Proof of Proposition 6.4 25

E.2 Distance Between $\mathbf{w}^{(k)}$ and $\widehat{\mathbf{w}}^{(k+1)}$ . . . 26
E.3 Bounded Lagrangian Multipliers . . . 27
E.4 Near Feasibility of Solutions . . . 27
E.5 Proof of Theorem 6.5 . . . 28

F GENERALIZATION BOUNDS FOR MODELS SOLVED FROM (9) 29

G ADDITIONAL DETAILS OF NUMERICAL EXPERIMENTS 31
G.1 Details of Datasets . . . 31
G.2 Implementation Details of the Proposed Methods . . . 32
G.2.1 Choices of κ in (9) and (11) . . . 32
G.2.2 Choices of Tuning Parameters in IDCA . . . 32
G.3 Implementation Details of the In-processing Baselines in Comparison . . . 32
G.3.1 Constrained Optimization Models for the In-processing Baselines . . . 32
G.3.2 Training Algorithm for the In-processing Baselines . . . 34
G.4 Implementation Details for the Regularization Method in Comparison . . . 35
G.4.1 Optimization Model for the Regularization Method . . . 35
G.4.2 Training Algorithm for the Regularization Method . . . 36
G.5 Changes of Score Densities with κ . . . 37
G.6 Robustness to the Choice of $\mathcal{I}$ . . . 37
G.7 Optimization Performance of IDCA . . . 39
G.8 Numerical Results on the ACS Dataset . . . 41
G.9 Numerical Results on Deep Learning . . . 42
G.10 Comparison Between Partial Statistical Parity and Full Statistical Parity . . . 43
G.11 Performances of Models from (9) with Different $|\hat{\mathcal{I}}|$ . . . 44

A EXTENSIONS TO GROUP AUC FAIRNESS, EQUALITY OF OPPORTUNITY, AND EQUALITY OF ODDS

Suppose the data points are ranked in descending order based on their scores $h_{\mathbf{w}}(\boldsymbol{\xi})$. *Group AUC fairness* (Yao et al., 2023; Vogel et al., 2021) requires that, with a probability of 50%, a random data point sampled from any group is not ranked lower than a random data point sampled from any other group, that is,

$$\mathrm{GAUC}_{k,k'}(\mathbf{w}) := \left|\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) \geq h_{\mathbf{w}}(\boldsymbol{\xi}')|\gamma = k, \gamma' = k') - 0.5\right| = 0, \ \forall k, k' \in \mathcal{G}, \tag{21}$$

where $(\boldsymbol{\xi}, \zeta, \gamma)$ and $(\boldsymbol{\xi}', \zeta', \gamma')$ are two i.i.d. random data points.

Similar to Definitions 4.2 and 4.3, we can also generalize group AUC fairness in (21) by only considering the comparison between random data points sampled from some rank ranges of different groups.

Definition A.1. We say $h_{\mathbf{w}}(\boldsymbol{\xi})$ satisfies the *partial group AUC fairness with respect to percentile interval* $\mathcal{I} := [\alpha, \beta) \subset [0, 1]$ if

$$\mathrm{GAUC}^{\mathcal{I}}_{k,k'}(\mathbf{w}) := \left|\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi}) \geq h_{\mathbf{w}}(\boldsymbol{\xi}')\middle|\gamma = k, \gamma' = k', \bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}, \bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\boldsymbol{\xi}')) \in \mathcal{I}\right) - 0.5\right| = 0, \tag{22}$$

for any k and k' in $\mathcal{G}$, where $(\boldsymbol{\xi}, \zeta, \gamma)$ and $(\boldsymbol{\xi}', \zeta', \gamma')$ are two i.i.d. random data points.

When $\mathcal{I} = [0, 1]$ for each $k \in \mathcal{G}$, (22) is reduced to (21).

Consider a binary classification task where $\zeta \in \{1, -1\}$. Similar to demographic parity in (3), we say **equality of opportunity** (Hardt et al., 2016) holds with respect to a decision threshold $\widehat{\theta}$ if

$$\mathrm{EOP}_{k,k'}(\mathbf{w}) = \left|\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta}|\gamma = k, \zeta = 1) - \Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta}|\gamma = k', \zeta = 1)\right| = 0, \ \forall k, k' \in \mathcal{G}. \tag{23}$$

Similarly, we say **equality of odds** (Hardt et al., 2016) holds with respect to a decision threshold $\widehat{\theta}$ if

$$\mathrm{EOD}_{k,k'}(\mathbf{w}) := \left|\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta}|\gamma = k, \zeta = c) - \Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta}|\gamma = k', \zeta = c)\right| = 0, \ \forall k, k' \in \mathcal{G}, \ \forall c \in \{1, -1\}. \tag{24}$$

Once again, similar to Definitions 4.2 and 4.3, we can also extend equality of opportunity and equality of odds above to partial fairness metrics as follows.

Definition A.2. We say $h_{\mathbf{w}}(\boldsymbol{\xi})$ satisfies the *partial equality of opportunity with respect to percentile interval* $\mathcal{I} = [\alpha, \beta) \subset [0, 1]$ and threshold $\widehat{\theta}$ if

$$\mathrm{EOP}^{\mathcal{I}}_{k,k'}(\mathbf{w}) := \left| \begin{array}{l} \Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta}\big|\gamma = k, \zeta = 1, \bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\big) \\ -\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta}\big|\gamma = k', \zeta = 1, \bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\big) \end{array} \right| = 0, \ \forall k, k' \in \mathcal{G}. \tag{25}$$

Definition A.3. We say $h_{\mathbf{w}}(\boldsymbol{\xi})$ satisfies the *partial equality of odd with respect to percentile interval* $\mathcal{I} = [\alpha, \beta) \subset [0, 1]$ and threshold $\widehat{\theta}$ if

$$\mathrm{EOD}^{\mathcal{I}}_{k,k'}(\mathbf{w}) := \max_{c \in \{-1,1\}} \left| \begin{array}{l} \Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta}\big|\gamma = k, \zeta = c, \bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\big) \\ -\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta}\big|\gamma = k', \zeta = c, \bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\big) \end{array} \right| = 0, \ \forall k, k' \in \mathcal{G}. \tag{26}$$

Just like (6) and (7), an in-processing method to produce a $\mathbf{w} \in \mathcal{W}$ satisfying (22), (25) or (26) can be implemented by solving the following optimization problems subject to partial fairness constraints:

$$\min_{\mathbf{w} \in \mathcal{W}} f_0(\mathbf{w}) \ \text{ s.t. } \ \mathrm{GAUC}^{\mathcal{I}}_{k,k'}(\mathbf{w}) \leq \kappa, \ \forall k, k' \in \mathcal{G}, \tag{27}$$

$$\min_{\mathbf{w} \in \mathcal{W}} f_0(\mathbf{w}) \ \text{ s.t. } \ \mathrm{EOP}^{\mathcal{I}}_{k,k'}(\mathbf{w}) \leq \kappa, \ \forall k, k' \in \mathcal{G}, \tag{28}$$

$$\min_{\mathbf{w} \in \mathcal{W}} f_0(\mathbf{w}) \ \text{ s.t. } \ \mathrm{EOD}^{\mathcal{I}}_{k,k'}(\mathbf{w}) \leq \kappa, \ \forall k, k' \in \mathcal{G}, \tag{29}$$

where, just as in (6) and (7) , $f_0(\mathbf{w})$ is the loss function measuring the predictive performance of $h_{\mathbf{w}}(\boldsymbol{\xi})$.

Following a process similar to the one that derives (9) and (11), we can construct the approximation of the constraints in (27). Let $n_k^{\alpha} = \lceil \alpha n_k \rceil$ and $n_k^{\beta} = \lceil \beta n_k \rceil$ so $n_k^{\beta} - n_k^{\alpha}$ represents the number of data points in $\mathcal{D}_k$ whose ranks are in $\mathcal{I}$. Let $h_{\mathbf{w}}(\boldsymbol{\xi}^k_{[i]})$ be the i-th largest score in $\{h_{\mathbf{w}}(\boldsymbol{\xi}^k_i)|i = 1, \ldots, n_k\}$ with ties broken in an arbitrary way. Then, for any k and k' in $\mathcal{G}$, $\mathrm{GAUC}^{\mathcal{I}}_{k,k'}(\mathbf{w})$ defined in (22) can be approximated as

$$\left| \frac{\sum_{i=n_k^{\alpha}+1}^{n_k^{\beta}} \sum_{j=n_{k'}^{\alpha}+1}^{n_{k'}^{\beta}} \mathbf{1}_{h_{\mathbf{w}}(\boldsymbol{\xi}^k_{[i]}) \geq h_{\mathbf{w}}(\boldsymbol{\xi}^{k'}_{[j]})}}{(n_k^{\beta} - n_k^{\alpha})(n_{k'}^{\beta} - n_{k'}^{\alpha})} - 0.5 \right|.$$

Using a surrogate σ to approximate the indicator functions above, we obtain the final approximation of (27):

$$\begin{aligned} &\min_{\mathbf{w} \in \mathcal{W}} f_0(\mathbf{w}) \\ &\text{s.t. } \frac{\sum_{i=n_k^{\alpha}+1}^{n_k^{\beta}} \sum_{j=n_{k'}^{\alpha}+1}^{n_{k'}^{\beta}} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}^k_{[i]}) - h_{\mathbf{w}}(\boldsymbol{\xi}^{k'}_{[j]}))}{(n_k^{\beta} - n_k^{\alpha})(n_{k'}^{\beta} - n_{k'}^{\alpha})} \leq 0.5 + \kappa, \ \forall k, k' \in \mathcal{G}. \end{aligned} \tag{30}$$

Note that, depending on σ, the constraint in (30) may not necessarily be symmetric in k and k', so there can be G^2 constraints in general.

Following a process similar to the one we used to derive the approximation (11) of (7), we can construct the approximation of (28). In particular, we first introduce the partition $\mathcal{D} = (\cup_{k\in\mathcal{G}}\mathcal{D}_{k,+}) \cup (\cup_{k\in\mathcal{G}}\mathcal{D}_{k,-})$, where

$$\begin{aligned}\mathcal{D}_{k,+} =&\{(\boldsymbol{\xi},\zeta,\gamma)\in\mathcal{D}|\gamma=k,\zeta=1\} = \{(\boldsymbol{\xi}_i^{k,+},1,k)\}_{i=1}^{n_{k,+}},\\ \mathcal{D}_{k,-} =&\{(\boldsymbol{\xi},\zeta,\gamma)\in\mathcal{D}|\gamma=k,\zeta=-1\} = \{(\boldsymbol{\xi}_i^{k,-},-1,k)\}_{i=1}^{n_{k,-}},\end{aligned}$$

and $n_{k,+}$ and $n_{k,-}$ are the corresponding group sizes ($n_{k,+}+n_{k,-}=n_k$). Then given the percentile interval $\mathcal{I}=[\alpha,\beta)\subset[0,1]$ and threshold $\widehat{\theta}$ in (25) and (26), we obtain the following approximation of (28):

$$\begin{aligned}&\min_{\mathbf{w}\in\mathcal{W}} f_0(\mathbf{w})\\ \text{s.t.}\quad &\min\left\{\tfrac{1}{n_{k,+}}\textstyle\sum_{i=1}^{n_{k,+}}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k,+})-\widehat{\theta}),\beta\right\}\\ &-\min\left\{\tfrac{1}{n_{k,+}}\textstyle\sum_{i=1}^{n_{k,+}}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k,+})-\widehat{\theta}),\alpha\right\}\\ &-\min\left\{\tfrac{1}{n_{k',+}}\textstyle\sum_{i=1}^{n_{k',+}}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k',+})-\widehat{\theta}),\beta\right\}\\ &+\min\left\{\tfrac{1}{n_{k',+}}\textstyle\sum_{i=1}^{n_{k',+}}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k',+})-\widehat{\theta}),\alpha\right\}\le\kappa(\beta-\alpha),\\ &\forall k,k'\in\mathcal{G}.\end{aligned}\tag{31}$$

Similarly, we can obtain the following approximation of (29):

$$\begin{aligned}&\min_{\mathbf{w}\in\mathcal{W}} f_0(\mathbf{w})\\ \text{s.t.}\quad &\max_{c\in\{+,-\}}\Bigg\{\min\left\{\tfrac{1}{n_{k,c}}\textstyle\sum_{i=1}^{n_{k,c}}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k,c})-\widehat{\theta}),\beta\right\}\\ &-\min\left\{\tfrac{1}{n_{k,c}}\textstyle\sum_{i=1}^{n_{k,c}}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k,c})-\widehat{\theta}),\alpha\right\}\\ &-\min\left\{\tfrac{1}{n_{k',c}}\textstyle\sum_{i=1}^{n_{k',c}}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k',c})-\widehat{\theta}),\beta\right\}\\ &+\min\left\{\tfrac{1}{n_{k',c}}\textstyle\sum_{i=1}^{n_{k',c}}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k',c})-\widehat{\theta}),\alpha\right\}\Bigg\}\le\kappa(\beta-\alpha),\\ &\forall k,k'\in\mathcal{G}.\end{aligned}\tag{32}$$

B HOW TO CHOOSE $\mathcal{I}$ FOR CLASSIFICATION PROBLEM:

The algorithm and theoretical results we develop are applicable to any interval $\mathcal{I}$ that users choose based on their specific fairness concerns. When users do not have a targeted interval, we recommend selecting $\mathcal{I}$ to cover an estimated positive decision rate p. For instance, in applications such as school admissions, p can be the historical admission rate. Then one may choose $\mathcal{I}$ such that $p\in\mathcal{I}$, as this score region is likely the most contested for fairness issues. To refine $\mathcal{I}$, one can first train a model without fairness constraints, i.e., $\mathbf{w}^*=\arg\min_{\mathbf{w}\in\mathcal{W}} f_0(\mathbf{w})$, and generate a candidate set of intervals that all cover p. Then, among these candidates, one may select an interval $\mathcal{I}$ over which the model $h_{\mathbf{w}^*}(\boldsymbol{\xi})$ exhibits the most significant unfairness. If the historical positive decision rate p is not available, one may specify a decision threshold $\widehat{\theta}$ for the predicted scores (e.g., $\widehat{\theta}=0$ for binary classification) according to the application context, and define $p=\Pr(h_{\mathbf{w}^*}(\boldsymbol{\xi})>\widehat{\theta})$.

C EQUIVALENT FORMULATIONS OF STATISTICAL PARITY AND DEMOGRAPHIC PARITY PARTIAL FAIRNESS CONSTRAINTS

In this section, we present the proofs of Lemma 5.1 and Lemma 5.2. Before that, we first present a few observations that will be used in the proofs.

Let

$$\bar{F}^{-1}_{\mathbf{w},k}(p) := \inf\{\theta\in\mathbb{R}|\bar{F}_{\mathbf{w},k}(\theta)\le p\},$$

which is the inverse CCDF of $h_{\mathbf{w}}(\boldsymbol{\xi})$ for group k. Basically, $\bar{F}^{-1}_{\mathbf{w},k}(p)$ represents the cutoff value above which the top $100p\%$ scores fall, or equivalently, $\bar{F}^{-1}_{\mathbf{w},k}(p)$ is the $(1-p)$-quantile of the conditional distribution of $h_{\mathbf{w}}(\boldsymbol{\xi})$

conditioning on $\gamma = k$. Therefore, we can then rewrite

$$\begin{aligned}
&\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta | \gamma = k, \bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\right) \\
=& \frac{\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta, \bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I} | \gamma = k\right)}{\Pr\left(\bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I} | \gamma = k\right)} \\
=& \frac{\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta, h_{\mathbf{w}}(\boldsymbol{\xi}) \in \left[\bar{F}^{-1}_{\mathbf{w},k}(\beta), \bar{F}^{-1}_{\mathbf{w},k}(\alpha)\right] \big| \gamma = k\big)}{\beta - \alpha} \\
=& \frac{\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta, h_{\mathbf{w}}(\boldsymbol{\xi}) > \bar{F}^{-1}_{\mathbf{w},k}(\beta) \big| \gamma = k\big)}{\beta - \alpha} - \frac{\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta, h_{\mathbf{w}}(\boldsymbol{\xi}) > \bar{F}^{-1}_{\mathbf{w},k}(\alpha) \big| \gamma = k\big)}{\beta - \alpha} \\
=& \frac{\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \max\left\{\theta, \bar{F}^{-1}_{\mathbf{w},k}(\beta)\right\} \big| \gamma = k\big)}{\beta - \alpha} - \frac{\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \max\left\{\theta, \bar{F}^{-1}_{\mathbf{w},k}(\alpha)\right\} \big| \gamma = k\big)}{\beta - \alpha} \\
=& \frac{\min\left\{\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta \big| \gamma = k\big), \beta\right\}}{\beta - \alpha} - \frac{\min\left\{\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta \big| \gamma = k\big), \alpha\right\}}{\beta - \alpha}.
\end{aligned} \tag{33}$$

Since we have assumed that $h_{\mathbf{w}}(\boldsymbol{\xi})$ has an absolutely continuous distribution in each group, (33) increases monotonically and continuously from 0 to 1 as θ changes from $+\infty$ to $-\infty$. Note that, when $\bar{F}_{\mathbf{w},k}(\theta) = \Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta \big| \gamma = k\big) \in [\alpha, \beta]$, (33) can be further simplified to

$$\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta | \gamma = k, \bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\right) = \frac{\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta | \gamma = k\right) - \alpha}{\beta - \alpha}. \tag{34}$$

Proof of Lemma 5.1. Suppose $\mathbf{w} \in \mathcal{W}$ is feasible to (6), or equivalently,

$$\left| \begin{array}{l} \Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta \big| \gamma = k, \bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\right) \\ -\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta \big| \gamma = k', \bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\right) \end{array} \right| \leq \kappa, \ \forall \theta \in \mathbb{R}, \ \forall k, k' \in \mathcal{G}. \tag{35}$$

Consider any $p \in [\alpha, \beta - \kappa(\beta - \alpha))$. Let $k \in \mathcal{G}$ be the group such that $\bar{F}^{-1}_{\mathbf{w},k}(p) \leq \bar{F}^{-1}_{\mathbf{w},k'}(p)$ for any $k' \in \mathcal{G}$. Let $\theta_p = \bar{F}^{-1}_{\mathbf{w},k}(p)$ so

$$\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta_p \big| \gamma = k\big) = p \in [\alpha, \beta]. \tag{36}$$

By (33), it holds that

$$\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta_p | \gamma = k, \bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\right) = \frac{p - \alpha}{\beta - \alpha}, \tag{37}$$

$$\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta_p | \gamma = k', \bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\right) \geq \frac{p - \alpha}{\beta - \alpha}, \quad \forall k' \neq k. \tag{38}$$

Combining (35) with $\theta = \theta_p$ and the two inequalities above gives

$$\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta_p | \gamma = k, \bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\right) \leq \frac{p - \alpha}{\beta - \alpha} + \kappa, \tag{39}$$

$$\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta_p | \gamma = k', \bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\right) \leq \frac{p - \alpha}{\beta - \alpha} + \kappa, \quad \forall k' \neq k. \tag{40}$$

Consider any $k' \neq k$. By (33), (38) and the fact that $\frac{p-\alpha}{\beta-\alpha} + \kappa < 1$, we must have $\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta_p \big| \gamma = k'\big) < \beta$. If $p = \alpha$, we have $\theta_p = \bar{F}^{-1}_{\mathbf{w},k}(\alpha) \leq \bar{F}^{-1}_{\mathbf{w},k'}(\alpha)$ so $\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta_p \big| \gamma = k'\big) \geq \alpha$. If $p > \alpha$, by (33) and (38), we must have $\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta_p \big| \gamma = k'\big) > \alpha$. In any case, we have

$$\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta_p \big| \gamma = k'\big) \in [\alpha, \beta], \quad \forall k' \neq k. \tag{41}$$

Applying (33), (36) and (41) to (37), (38), (39) and (40) gives (8).

Suppose (8) holds. We will prove (35) by contradiction. Suppose there exist $\bar{\theta}$ and k and k' in $\mathcal{G}$ such that

$$\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi}) > \bar{\theta} | \gamma = k, \bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\right) - \Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi}) > \bar{\theta} | \gamma = k', \bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\right) > \kappa. \tag{42}$$

Let

$$p = 0.5(\beta-\alpha)\cdot\left(\begin{array}{l}\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi})>\bar{\theta}\middle|\gamma=k,\bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi}))\in\mathcal{I}\right)\\+\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi})>\bar{\theta}\middle|\gamma=k',\bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\boldsymbol{\xi}))\in\mathcal{I}\right)\end{array}\right)-0.5\kappa(\beta-\alpha)+\alpha.$$

By (42), we have

$$\begin{aligned}\alpha &= 0.5(\beta-\alpha)\kappa-0.5\kappa(\beta-\alpha)+\alpha\\ &< p < 0.5(\beta-\alpha)(2-\kappa)-0.5\kappa(\beta-\alpha)+\alpha=\beta-\kappa(\beta-\alpha).\end{aligned}$$

By (8), there exists θ_p such that

$$\Pr(h_{\mathbf{w}}(\boldsymbol{\xi})>\theta_p|\gamma=k)\geq p>\alpha, \tag{43}$$
$$\Pr(h_{\mathbf{w}}(\boldsymbol{\xi})>\theta_p|\gamma=k')\geq p>\alpha, \tag{44}$$
$$\Pr(h_{\mathbf{w}}(\boldsymbol{\xi})>\theta_p|\gamma=k)\leq p+\kappa(\beta-\alpha)<\beta, \tag{45}$$
$$\Pr(h_{\mathbf{w}}(\boldsymbol{\xi})>\theta_p|\gamma=k')\leq p+\kappa(\beta-\alpha)<\beta, \tag{46}$$

so (34) holds for k and k'. Applying (34) to (44) and (45) leads to (38) and (39).

By (38), (42), and the definition of p, we can show that

$$\begin{aligned}&\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi})>\bar{\theta}|\gamma=k',\bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\boldsymbol{\xi}))\in\mathcal{I}\right)\\ <\ &0.5\left(\begin{array}{l}\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi})>\bar{\theta}\middle|\gamma=k,\bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi}))\in\mathcal{I}\right)\\+\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi})>\bar{\theta}\middle|\gamma=k',\bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\boldsymbol{\xi}))\in\mathcal{I}\right)\end{array}\right)-0.5\kappa\\ =\ &\frac{p-\alpha}{\beta-\alpha}\\ \leq\ &\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi})>\theta_p|\gamma=k',\bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\boldsymbol{\xi}))\in\mathcal{I}\right),\end{aligned}$$

which implies $\bar{\theta}>\theta_p$ and thus

$$\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi})>\theta_p|\gamma=k,\bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi}))\in\mathcal{I}\right)\geq\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi})>\bar{\theta}|\gamma=k,\bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi}))\in\mathcal{I}\right).$$

This inequality, (39), and the definition of p together imply

$$\begin{aligned}&0.5\left(\begin{array}{l}\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi})>\bar{\theta}\middle|\gamma=k,\bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi}))\in\mathcal{I}\right)\\+\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi})>\bar{\theta}\middle|\gamma=k',\bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\boldsymbol{\xi}))\in\mathcal{I}\right)\end{array}\right)+0.5\kappa\\ =\ &\frac{p-\alpha}{\beta-\alpha}+\kappa\\ \geq\ &\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi})>\theta_p|\gamma=k,\bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi}))\in\mathcal{I}\right)\\ \geq\ &\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi})>\bar{\theta}|\gamma=k,\bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi}))\in\mathcal{I}\right),\end{aligned}$$

which further contradicts with (42). Suppose there exists $\bar{\theta}$ such that

$$\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi})>\bar{\theta}|\gamma=k,\bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi}))\in\mathcal{I}\right)-\Pr\left(h_{\mathbf{w}}(\boldsymbol{\xi})>\bar{\theta}|\gamma=k',\bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\boldsymbol{\xi}))\in\mathcal{I}\right)<-\kappa.$$

A similar contradiction can be derived by following a similar argument. This indicates that (35) must hold, which completes the proof.

□

Proof of Lemma 5.2. The conclusion is a direct consequence of (33) with $\theta=\widehat{\theta}$. □

D REFORMULATE TRAINING PROBLEMS WITH PARTIAL FAIRNESS CONSTRAINTS AS DC OPTIMIZATION

This section is organized as follows. In Section D.1, we first present multiple examples of the DC objective function in (9), (11), (27), (28) and (29). Next, in Section D.2, we present some examples of (9) whose constraints

are DC. Then, in Section D.3, we present some examples of (11) whose constraints are DC. Finally, in Section D.4, we present some examples of (30), (31) and (32).

Before that, we first present some useful facts. When each $f_i(\mathbf{w})$ in (12) is ρ-weakly convex, meaning that $f_i(\mathbf{w}) + \frac{\rho}{2}\|\mathbf{w}\|^2$ is convex, $f_i(\mathbf{w})$ is a DC function because

$$f_i(\mathbf{w}) = \underbrace{f_i(\mathbf{w}) + \frac{\rho}{2}\|\mathbf{w}\|^2}_{=f_i^+(\mathbf{w})} - \underbrace{\frac{\rho}{2}\|\mathbf{w}\|^2}_{=f_i^-(\mathbf{w})}. \tag{47}$$

Therefore, the IDCA in this work can be applied to weakly convex constrained optimization with oracle complexity of $\tilde{O}(\epsilon^{-4})$ as shown in Theorem 6.5. This complexity matches the best complexity of first-order methods in the literature for weakly convex constrained optimization up to a logarithmic factor. Moreover, if $f_i(\mathbf{w})$ is differentiable and $\nabla f_i(\mathbf{w})$ is L-Lipschitz continuous, $f_i(\mathbf{w})$ is L-weakly convex and thus also DC. However, for differentiable problems, there exist more efficient first-order algorithms in the literature with oracle complexity lower than $O(\epsilon^{-4})$, so we do not recommend using IDCA for differentiable problems in general. In the following subsections, we present the scenarios where (9) and (11) are instances of (12).

D.1 DC Objective Function in (9) and (11)

Given a training set $\mathcal{D} = \{(\boldsymbol{\xi}_i, \zeta_i, \gamma_i)\}_{i=1}^n$, the parameter $\mathbf{w}$ in model $h_{\mathbf{w}}(\cdot)$ is typically obtained by minimizing the following empirical loss

$$f_0(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \ell(h_{\mathbf{w}}(\boldsymbol{\xi}_i), \zeta_i), \tag{48}$$

where $\ell(h_{\mathbf{w}}(\boldsymbol{\xi}), \zeta)$ is a loss function measuring the discrepancy between the predicted score $h_{\mathbf{w}}(\boldsymbol{\xi})$ and the corresponding target variable ζ.

Example D.1. Suppose Assumption 3.1 holds. In the following three scenarios:

- Quadratic loss: $\ell(h_{\mathbf{w}}(\boldsymbol{\xi}_i), \zeta_i) = \frac{1}{2}(h_{\mathbf{w}}(\boldsymbol{\xi}_i) - \zeta_i)^2$ and there exists $B \geq 0$ such that $|h_{\mathbf{w}}(\boldsymbol{\xi}_i)| \leq B$ for $i = 1, \ldots, n$ for any $\mathbf{w} \in \mathcal{W}$;
- Hinge loss: $\ell(h_{\mathbf{w}}(\boldsymbol{\xi}_i), \zeta_i) = (1 - \zeta_i \cdot h_{\mathbf{w}}(\boldsymbol{\xi}_i))_+$ and $\zeta_i \in \{1, -1\}$;
- Logistic loss: $\ell(h_{\mathbf{w}}(\boldsymbol{\xi}_i), \zeta_i) = \ln(1 + \exp(-\zeta_i \cdot h_{\mathbf{w}}(\boldsymbol{\xi}_i)))$ and $\zeta_i \in \{1, -1\}$;

there exists $\rho \geq 0$ such that $f_0(\mathbf{w})$ in (48) is ρ-weakly convex and thus is DC in the form of (47).

Consider a binary classification problem where $\zeta \in \{1, -1\}$. Suppose the goal is to optimize the AUC of $h_{\mathbf{w}}(\boldsymbol{\xi})$, that is,

$$\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) \geq h_{\mathbf{w}}(\boldsymbol{\xi}') \big| \zeta = 1, \zeta' = -1\big). \tag{49}$$

We partition the $\mathcal{D}$ into positive and negative sets, namely, $\mathcal{D} = \mathcal{D}_+ \cup \mathcal{D}_-$, where

$$\mathcal{D}_+ = \{(\boldsymbol{\xi}, \zeta, \gamma) \in \mathcal{D} | \zeta = 1\} = \{(\boldsymbol{\xi}_i^+, 1, \gamma_i^+)\}_{i=1}^{n_+}$$
$$\mathcal{D}_- = \{(\boldsymbol{\xi}, \zeta, \gamma) \in \mathcal{D} | \zeta = -1\} = \{(\boldsymbol{\xi}_i^-, -1, \gamma_i^-)\}_{i=1}^{n_-}.$$

To maximize AUC in (49), one can minimize an objective function $f_0(\mathbf{w})$ defined with pairwise losses, i.e.,

$$f_0(\mathbf{w}) = \frac{1}{n_+ n_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-} \ell(h_{\mathbf{w}}(\boldsymbol{\xi}_i^+) - h_{\mathbf{w}}(\boldsymbol{\xi}_j^-)), \tag{50}$$

where $\ell(h_{\mathbf{w}}(\boldsymbol{\xi}_i^+) - h_{\mathbf{w}}(\boldsymbol{\xi}_i^-))$ is a surrogate loss approximating the indicator $\mathbf{1}_{h_{\mathbf{w}}(\boldsymbol{\xi}_i^+) \leq h_{\mathbf{w}}(\boldsymbol{\xi}_i^-)}$. In some applications, the goal is to maximize the partial AUC of $h_{\mathbf{w}}(\boldsymbol{\xi})$ with the false positive rate restricted in an interval $\mathcal{I} = [\alpha, \beta]$, that is,

$$\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) \geq h_{\mathbf{w}}(\boldsymbol{\xi}') \big| \zeta = 1, \zeta' = -1, \bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\boldsymbol{\xi}')) \in \mathcal{I}\big), \tag{51}$$

where $(\boldsymbol{\xi}, \zeta, \gamma)$ and $(\boldsymbol{\xi}', \zeta', \gamma')$ are two i.i.d. random data points. Similar to (50), let $n_-^\alpha = \lceil \alpha n_- \rceil$ and $n_-^\beta = \lceil \beta n_- \rceil$ so $n_-^\beta - n_-^\alpha$ represents the number of data points in $\mathcal{D}_-$ whose ranks are in $\mathcal{I}$. Let $h_{\mathbf{w}}(\boldsymbol{\xi}_{[i]}^-)$ be the i-th largest score in $\{h_{\mathbf{w}}(\boldsymbol{\xi}_i^-) | i = 1, \ldots, n_-\}$ with ties broken in an arbitrary way. Then, to maximize the partial AUC in (51), one can minimize the following objective function

$$
\begin{aligned}
f_0(\mathbf{w}) &= \frac{1}{n_+(n_-^\beta - n_-^\alpha)} \sum_{i=1}^{n_+} \sum_{j=n_-^\alpha+1}^{n_-^\beta} \ell(h_{\mathbf{w}}(\boldsymbol{\xi}_i^+) - h_{\mathbf{w}}(\boldsymbol{\xi}_{[j]}^-)) \\
&= \frac{1}{n_+(n_-^\beta - n_-^\alpha)} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-^\beta} \ell(h_{\mathbf{w}}(\boldsymbol{\xi}_i^+) - h_{\mathbf{w}}(\boldsymbol{\xi}_{[j]}^-)) - \frac{1}{n_+(n_-^\beta - n_-^\alpha)} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-^\alpha} \ell(h_{\mathbf{w}}(\boldsymbol{\xi}_i^+) - h_{\mathbf{w}}(\boldsymbol{\xi}_{[j]}^-)). \qquad (52)
\end{aligned}
$$

Example D.2. Suppose Assumption 3.1 holds. In the following three scenarios:

- Quadratic loss: $\ell(h_{\mathbf{w}}(\boldsymbol{\xi}_i^+) - h_{\mathbf{w}}(\boldsymbol{\xi}_j^-)) = \frac{1}{2}(1 - h_{\mathbf{w}}(\boldsymbol{\xi}_i^+) + h_{\mathbf{w}}(\boldsymbol{\xi}_j^-))^2$ and there exists $B \geq 0$ such that $|h_{\mathbf{w}}(\boldsymbol{\xi}_i)| \leq B$ for $i = 1, \ldots, n$ for any $\mathbf{w} \in \mathcal{W}$;
- Hinge loss: $\ell(h_{\mathbf{w}}(\boldsymbol{\xi}_i^+) - h_{\mathbf{w}}(\boldsymbol{\xi}_j^-)) = (1 - h_{\mathbf{w}}(\boldsymbol{\xi}_i^+) + h_{\mathbf{w}}(\boldsymbol{\xi}_j^-))_+$;
- Logistic loss: $\ell(h_{\mathbf{w}}(\boldsymbol{\xi}_i^+) - h_{\mathbf{w}}(\boldsymbol{\xi}_j^-)) = \ln(1 + \exp(-h_{\mathbf{w}}(\boldsymbol{\xi}_i^+) + h_{\mathbf{w}}(\boldsymbol{\xi}_j^-)))$;

there exists $\rho \geq 0$ such that the two terms on the right-hand side of (52) are both ρ-weakly convex and thus $f_0(\mathbf{w})$ in (52) is DC in the form of

$$
\begin{aligned}
f_0(\mathbf{w}) &= f_0^+(\mathbf{w}) - f_0^-(\mathbf{w}), \\
f_0^+(\mathbf{w}) &= \frac{1}{n_+(n_-^\beta - n_-^\alpha)} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-^\beta} \ell(h_{\mathbf{w}}(\boldsymbol{\xi}_i^+) - h_{\mathbf{w}}(\boldsymbol{\xi}_{[j]}^-)) + \frac{\rho}{2}\|\mathbf{w}\|^2, \\
f_0^-(\mathbf{w}) &= \frac{1}{n_+(n_-^\beta - n_-^\alpha)} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-^\alpha} \ell(h_{\mathbf{w}}(\boldsymbol{\xi}_i^+) - h_{\mathbf{w}}(\boldsymbol{\xi}_{[j]}^-)) + \frac{\rho}{2}\|\mathbf{w}\|^2.
\end{aligned}
$$

Note that f_0 in (50) is DC in the three scenarios above also because it is a special case of (52) with $\mathcal{I} = [0, 1]$, i.e., $n_-^\alpha = 0$ and $n_-^\beta = n_-$.

D.2 DC Constraints in (9)

Example D.3. Suppose Assumption 3.1 holds and $\sigma(x) = \max\{\min\{x + 0.5, 1\}, 0\} = \sigma^+(x) - \sigma^-(x)$ in (9), where

$$
\sigma^+(x) = \max\{x + 0.5, 0\} \text{ and } \sigma^-(x) = \max\{x - 0.5, 0\}.
$$

Since $\sigma^+(x)$ and $\sigma^-(x)$ are convex and 1-Lipschitz continuous, there exists $\rho \geq 0$ such that

$$
\frac{1}{n_k} \sum_{i=1}^{n_k} \sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - \theta_p) \text{ and } \frac{1}{n_k} \sum_{i=1}^{n_k} \sigma^-(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - \theta_p)
$$

are both ρ-weakly convex jointly in $\mathbf{w}$ and θ_p for any $p \in \widehat{\mathcal{I}}$ and any $k \in \mathcal{G}$. Therefore, the constraints in (9) are instances of the constraints in (12) in the form of

$$
f_i(\mathbf{w}) = \underbrace{\frac{1}{n_k} \sum_{i=1}^{n_k} \sigma^-(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - \theta_p) + \frac{\rho}{2}\|\mathbf{w}\|^2}_{f_i^+(\mathbf{w})} - \underbrace{\left[\frac{1}{n_k} \sum_{i=1}^{n_k} \sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - \theta_p) - p + \frac{\rho}{2}\|\mathbf{w}\|^2\right]}_{f_i^-(\mathbf{w})} \leq 0
$$

and

$$f_i(\mathbf{w}) = \underbrace{\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\theta_p)+\frac{\rho}{2}\|\mathbf{w}\|^2}_{f_i^+(\mathbf{w})} - \underbrace{\left[\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma^-(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\theta_p)+p+\kappa(\beta-\alpha)+\frac{\rho}{2}\|\mathbf{w}\|^2\right]}_{f_i^-(\mathbf{w})} \le 0$$

for any $p \in \widehat{\mathcal{I}}$ and any $k \in \mathcal{G}$.

Example D.4. Suppose Assumption 3.1 holds and $\sigma(x) = \exp(x)/(1+\exp(x))$ in (9). Also, suppose there exists $B \ge 0$ such that $\|\nabla h_{\mathbf{w}}(\boldsymbol{\xi}_i)\| \le B$ for $i = 1, \ldots, n$ for any $\mathbf{w} \in \mathcal{W}$. In this case, $\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\theta_p)$, as a function of $\mathbf{w}$ and θ_p, is differentiable with a Lipschitz continuous gradient on $\mathcal{W}$. Hence, there exists $\rho \ge 0$ such that $\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\theta_p)$ is ρ-weakly convex jointly in $\mathbf{w}$ and θ_p for any $p \in \widehat{\mathcal{I}}$ and any $k \in \mathcal{G}$. Therefore, the constraints in (9) are DC constraints in the form of

$$f_i(\mathbf{w}) = \underbrace{\frac{\rho}{2}\|\mathbf{w}\|^2}_{f_i^+(\mathbf{w})} - \underbrace{\left[\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\theta_p)-p+\frac{\rho}{2}\|\mathbf{w}\|^2\right]}_{f_i^-(\mathbf{w})} \le 0$$

and

$$f_i(\mathbf{w}) = \underbrace{\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\theta_p)+\frac{\rho}{2}\|\mathbf{w}\|^2}_{f_i^+(\mathbf{w})} - \underbrace{\left[p+\kappa(\beta-\alpha)+\frac{\rho}{2}\|\mathbf{w}\|^2\right]}_{f_i^-(\mathbf{w})} \le 0$$

for any $p \in \widehat{\mathcal{I}}$ and any $k \in \mathcal{G}$.

D.3 DC Constraints in (11)

Suppose $\sigma(x)$, $\sigma^+(x)$ and $\sigma^-(x)$ are the same as in Example D.3. We first observe that

$$\begin{aligned}
&\min\left\{\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta}),\alpha\right\}\\
&= \frac{1}{n_k}\sum_{i=1}^{n_k}\sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta})+\alpha-\max\left\{\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma^-(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta})+\alpha,\sum_{i=1}^{n_k}\sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta})\right\},\\
&\min\left\{\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta}),\beta\right\}\\
&= \frac{1}{n_k}\sum_{i=1}^{n_k}\sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta})+\beta-\max\left\{\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma^-(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta})+\beta,\sum_{i=1}^{n_k}\sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta})\right\},
\end{aligned} \tag{53}$$

where the right-hand sides of both equalities are differences of two weakly convex functions and thus are DC.

Example D.5. Suppose Assumption 3.1 holds and $\sigma(x)$, $\sigma^+(x)$ and $\sigma^-(x)$ in (11) are the same as in Example D.3. For the same reason as in Example D.3,

$$\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta}) \text{ and } \frac{1}{n_k}\sum_{i=1}^{n_k}\sigma^-(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta})$$

are both weakly convex for any $k \in \mathcal{G}$. So are

$$\max\left\{\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma^-(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta})+\alpha,\sum_{i=1}^{n_k}\sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta})\right\}$$

and

$$\max\left\{\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma^-(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta})+\beta,\sum_{i=1}^{n_k}\sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta})\right\}$$

for any $k \in \mathcal{G}$. Therefore, by (53), there exists $\rho \geq 0$ such that the constraints in (11) are instances of the constraints in (12) in the form of

$$\begin{aligned}
f_i(\mathbf{w}) &= f_i^+(\mathbf{w}) - f_i^-(\mathbf{w}),\\
f_i^+(\mathbf{w}) &= \frac{1}{n_k}\sum_{i=1}^{n_k}\sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta}) + \max\left\{\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma^-(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta})+\alpha, \sum_{i=1}^{n_k}\sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta})\right\}\\
&\quad + \frac{1}{n_{k'}}\sum_{i=1}^{n_{k'}}\sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k'})-\widehat{\theta}) + \max\left\{\frac{1}{n_{k'}}\sum_{i=1}^{n_{k'}}\sigma^-(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k'})-\widehat{\theta})+\beta, \sum_{i=1}^{n_{k'}}\sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k'})-\widehat{\theta})\right\}\\
&\quad + \frac{\rho}{2}\|\mathbf{w}\|^2,\\
f_i^-(\mathbf{w}) &= \frac{1}{n_{k'}}\sum_{i=1}^{n_{k'}}\sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k'})-\widehat{\theta}) + \max\left\{\frac{1}{n_{k'}}\sum_{i=1}^{n_{k'}}\sigma^-(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k'})-\widehat{\theta})+\alpha, \sum_{i=1}^{n_{k'}}\sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k'})-\widehat{\theta})\right\}\\
&\quad + \frac{1}{n_k}\sum_{i=1}^{n_k}\sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k})-\widehat{\theta}) + \max\left\{\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma^-(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k})-\widehat{\theta})+\beta, \sum_{i=1}^{n_k}\sigma^+(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k})-\widehat{\theta})\right\}\\
&\quad + \frac{\rho}{2}\|\mathbf{w}\|^2 + \kappa(\beta-\alpha),
\end{aligned}$$

for any $k, k' \in \mathcal{G}$.

Example D.6. Suppose Assumption 3.1 holds and $\sigma(x) = \exp(x)/(1+\exp(x))$ in (11). Also, suppose there exists $B \geq 0$ such that $\|\nabla h_{\mathbf{w}}(\boldsymbol{\xi}_i)\| \leq B$ for $i = 1, \ldots, n$ for any $\mathbf{w} \in \mathcal{W}$. In this case, $-\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta})$ is differentiable with a Lipschitz continuous gradient on $\mathcal{W}$. Hence, $-\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta})$ is weakly convex for any $k \in \mathcal{G}$. So are

$$-\min\left\{\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta}), \beta\right\} = \max\left\{-\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta}), -\beta\right\}$$

and

$$-\min\left\{\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta}), \alpha\right\} = \max\left\{-\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta}), -\alpha\right\}$$

for any $k \in \mathcal{G}$. Therefore, there exists $\rho \geq 0$ such that the constraints in (11) are DC constraints in the form of

$$\begin{aligned}
f_i(\mathbf{w}) &= f_i^+(\mathbf{w}) - f_i^-(\mathbf{w}),\\
f_i^+(\mathbf{w}) &= \max\left\{-\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta}), -\alpha\right\} + \max\left\{-\frac{1}{n_{k'}}\sum_{i=1}^{n_{k'}}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k'})-\widehat{\theta}), -\beta\right\} + \frac{\rho}{2}\|\mathbf{w}\|^2,\\
f_i^-(\mathbf{w}) &= \max\left\{-\frac{1}{n_k}\sum_{i=1}^{n_k}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k)-\widehat{\theta}), -\beta\right\} + \max\left\{-\frac{1}{n_{k'}}\sum_{i=1}^{n_{k'}}\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k'})-\widehat{\theta}), -\alpha\right\} + \frac{\rho}{2}\|\mathbf{w}\|^2\\
&\quad + \kappa(\beta-\alpha),
\end{aligned}$$

for any $k, k' \in \mathcal{G}$.

D.4 DC Constraints in (30), (31) and (32)

Example D.7. Suppose Assumption 3.1 holds and $\sigma(x) = \exp(x)/(1+\exp(x))$ in (11). Also, suppose there exists $B \geq 0$ such that $\|\nabla h_{\mathbf{w}}(\boldsymbol{\xi}_i)\| \leq B$ for $i = 1, \ldots, n$ for any $\mathbf{w} \in \mathcal{W}$. We first show that the constraints in (30) can be formulated as DC constraints.

Let $n_k^\alpha = \lceil \alpha n_k \rceil$ and $n_k^\beta = \lceil \beta n_k \rceil$ so $n_k^\beta - n_k^\alpha$ represents the number of data points in $\mathcal{D}_k$ whose ranks are in $\mathcal{I}$. We define $[i]$ is the index of the i-th largest coordinate in vector $(h_{\boldsymbol{w}}(\boldsymbol{\xi}_i^k))_{i=1}^{n_k}$ with ties broken arbitrarily. Given

any k and k' in $\mathcal{G}$, we first define

$$\begin{aligned} G_j &:= \sum_{i=n_k^\alpha+1}^{n_k^\beta} \sigma\left(h_{\boldsymbol{w}}(\boldsymbol{\xi}_{[i]}^k) - h_{\boldsymbol{w}}(\boldsymbol{\xi}_j^{k'})\right) \\ &= \underbrace{\sum_{i=1}^{n_k^\beta} \sigma\left(h_{\boldsymbol{w}}(\boldsymbol{\xi}_{[i]}^k) - h_{\boldsymbol{w}}(\boldsymbol{\xi}_j^{k'})\right)}_{=:G_j^+} - \underbrace{\sum_{i=1}^{n_k^\alpha} \sigma\left(h_{\boldsymbol{w}}(\boldsymbol{\xi}_{[i]}^k) - h_{\boldsymbol{w}}(\boldsymbol{\xi}_j^{k'})\right)}_{=:G_j^-}, \quad \forall j = 1, \ldots, n_{k'}. \end{aligned}$$

Next, we define $\langle j \rangle$ as the index of the j-th smallest coordinate in vector $(G^-{}_j)_{j=1}^{n_{k'}}$ with ties broken arbitrarily, so we can write

$$\sum_{i=n_k^\alpha+1}^{n_k^\beta} \sum_{j=n_{k'}^\alpha+1}^{n_{k'}^\beta} \sigma\left(h_{\boldsymbol{w}}(\boldsymbol{\xi}_{[i]}^k) - h_{\boldsymbol{w}}(\boldsymbol{\xi}_{[j]}^{k'})\right) = \sum_{j=n_{k'}^\alpha+1}^{n_{k'}^\beta} G_{\langle j \rangle} = \sum_{j=1}^{n_{k'}^\beta} G_{\langle j \rangle} - \sum_{j=1}^{n_{k'}^\alpha} G_{\langle j \rangle}. \tag{54}$$

For any positive integers n and N with $n \le N$, we define

$$\Delta_n^N := \left\{ \mathbf{p} \in \mathbb{R}^N \,\middle|\, \sum_{j=1}^N p_j = n, 0 \le p_j \le 1, \forall j = 1, \ldots, N \right\}.$$

Also, let $G^+ := \sum_{j=1}^{n_{k'}} G_j^+$. Then we can write

$$\begin{aligned} \sum_{j=1}^{n_{k'}^\beta} G_{\langle j \rangle} &= \min_{\mathbf{p} \in \Delta_{n_{k'}^\beta}^{n_{k'}}} \sum_{j=1}^{n_{k'}} p_j (G_j^+ - G_j^-) \\ &= \min_{\mathbf{p} \in \Delta_{n_{k'}^\beta}^{n_{k'}}} \sum_{j=1}^{n_{k'}} p_j (G_j^+ - G_j^- + G^+ - G^+) \\ &= n_{k'}^\beta G^+ + \min_{\mathbf{p} \in \Delta_{n_{k'}^\beta}^{n_{k'}}} \sum_{j=1}^{n_{k'}} p_j (-(G^+ - G_j^+) - G_j^-) \\ &= n_{k'}^\beta G^+ - \max_{\mathbf{p} \in \Delta_{n_{k'}^\beta}^{n_{k'}}} \sum_{j=1}^{n_{k'}} p_j ((G^+ - G_j^+) + G_j^-), \end{aligned}$$

which is a difference of two weakly convex functions. Similarly, we can write $\sum_{j=1}^{n_{k'}^\alpha} G_{\langle j \rangle} = n_{k'}^\alpha G^+ - \max_{\mathbf{p} \in \Delta_{n_{k'}^\alpha}^{n_{k'}}} \sum_{j=1}^{n_{k'}} p_j ((G^+ - G_j^+) + G_j^-)$, which is also a difference of two weakly convex functions. Therefore, (54) is also a difference of two weakly convex functions, so there exists $\rho \ge 0$ such that the constraints in (30) are DC constraints in the form of

$$\begin{aligned} f_i(\mathbf{w}) = &\underbrace{\frac{(n_{k'}^\beta - n_{k'}^\alpha) G^+ + \max_{\mathbf{p} \in \Delta_{n_{k'}^\alpha}^{n_{k'}}} \sum_{j=1}^{n_{k'}} p_j ((G^+ - G_j^+) + G_j^-)}{(n_k^\beta - n_k^\alpha)(n_{k'}^\beta - n_{k'}^\alpha)} + \frac{\rho}{2} \|\mathbf{w}\|^2}_{f_i^+(\mathbf{w})} \\ &- \underbrace{\frac{\max_{\mathbf{p} \in \Delta_{n_{k'}^\beta}^{n_{k'}}} \sum_{j=1}^{n_{k'}} p_j ((G^+ - G_j^+) + G_j^-)}{(n_k^\beta - n_k^\alpha)(n_{k'}^\beta - n_{k'}^\alpha)} + \frac{\rho}{2} \|\mathbf{w}\|^2 + 0.5 + \kappa}_{f_i^-(\mathbf{w})} \le 0 \end{aligned}$$

for any $k, k' \in \mathcal{G}$.

For DC constraints in (31) and (32), their structures are almost identical to that of (11). Hence, they can be formulated by the same procedure as in Example D.5 and D.6.

E PROOFS FOR COMPLEXITY ANALYSIS

In this section, we present the proofs for the lemmas, propositions, and theorems related to the complexity analysis of IDCA. All proofs are under Assumption 6.3.

E.1 Proof of Proposition 6.4

Proof. Since $f_i(\mathbf{w}^{(k)}) \le 0$ for $i \in [m]$, (16) is feasible. Recall that $\widehat{\mathbf{w}}^{(k+1)}$ is the optimal solution of (16). We prove this proposition by contradiction. Suppose $T_k \ge \frac{M^2}{\epsilon_k \mu} \ln\left(1 + \frac{\mu \|\mathbf{w}^{(k)} - \widehat{\mathbf{w}}^{(k+1)}\|^2}{\epsilon_k}\right)$ but $g_0(\mathbf{w}^{(k+1)}) - g_0(\widehat{\mathbf{w}}^{(k+1)}) > \epsilon_k$, or equivalently, $g_0(\mathbf{v}^{(t)}) - g_0(\widehat{\mathbf{w}}^{(k+1)}) > \epsilon_k$ for all $t \in \mathcal{T}$, after Algorithm 2 terminates.

If $t \in \mathcal{T}$, we have

$$\begin{aligned}
\|\mathbf{v}^{(t+1)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 &\le \|\mathbf{v}^{(t)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 - \frac{2\epsilon_k}{\|\mathbf{g}_0^{(t)}\|^2} \left\langle \mathbf{g}_0^{(t)}, \mathbf{v}^{(t)} - \widehat{\mathbf{w}}^{(k+1)} \right\rangle + \frac{\epsilon_k^2}{\|\mathbf{g}_0^{(t)}\|^2} \\
&\le \left(1 - \frac{\epsilon_k \mu}{\|\mathbf{g}_0^{(t)}\|^2}\right) \|\mathbf{v}^{(t)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 - \frac{2\epsilon_k}{\|\mathbf{g}_0^{(t)}\|^2} \left(g_0(\mathbf{v}^{(t)}) - g_0(\widehat{\mathbf{w}}^{(k+1)})\right) + \frac{\epsilon_k^2}{\|\mathbf{g}_0^{(t)}\|^2} \\
&< \left(1 - \frac{\epsilon_k \mu}{\|\mathbf{g}_0^{(t)}\|^2}\right) \|\mathbf{v}^{(t)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 - \frac{\epsilon_k^2}{\|\mathbf{g}_0^{(t)}\|^2} \\
&\le \left(1 - \frac{\epsilon_k \mu}{M^2}\right) \|\mathbf{v}^{(t)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 - \frac{\epsilon_k^2}{M^2},
\end{aligned}$$

where the second inequality is by the μ-strong convexity of g_0, the third is because $g_0(\mathbf{v}^{(t)}) - g_0(\widehat{\mathbf{w}}^{(k+1)}) > \epsilon_k$ and the last is because of Assumption 6.3D.

If $t \notin \mathcal{T}$, by similar arguments, we have

$$\begin{aligned}
\|\mathbf{v}^{(t+1)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 &\le \|\mathbf{v}^{(t)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 - \frac{2\epsilon_k}{\|\mathbf{g}^{(t)}\|^2} \left\langle \mathbf{g}^{(t)}, \mathbf{v}^{(t)} - \widehat{\mathbf{w}}^{(k+1)} \right\rangle + \frac{\epsilon_k^2}{\|\mathbf{g}^{(t)}\|^2} \\
&\le \left(1 - \frac{\epsilon_k \mu}{\|\mathbf{g}^{(t)}\|^2}\right) \|\mathbf{v}^{(t)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 - \frac{2\epsilon_k}{\|\mathbf{g}^{(t)}\|^2} \left(g(\mathbf{v}^{(t)}) - g(\widehat{\mathbf{w}}^{(k+1)})\right) + \frac{\epsilon_k^2}{\|\mathbf{g}^{(t)}\|^2} \\
&< \left(1 - \frac{\epsilon_k \mu}{\|\mathbf{g}^{(t)}\|^2}\right) \|\mathbf{v}^{(t)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 - \frac{\epsilon_k^2}{\|\mathbf{g}^{(t)}\|^2} \\
&\le \left(1 - \frac{\epsilon_k \mu}{M^2}\right) \|\mathbf{v}^{(t)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 - \frac{\epsilon_k^2}{M^2}.
\end{aligned}$$

The two inequalities above together imply

$$\|\mathbf{v}^{(t+1)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 < \left(1 - \frac{\epsilon_k \mu}{M^2}\right) \|\mathbf{v}^{(t)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 - \frac{\epsilon_k^2}{M^2},$$

for $t = 0, 1, \ldots, T_k - 1$. For brevity, for $t = 0, 1, \ldots, T_k - 1$, set

$$a_t := \|\mathbf{v}^{(t)} - \widehat{\mathbf{w}}^{(k+1)}\|^2, \quad q := 1 - \frac{\epsilon_k \mu}{M^2}, \quad c := \frac{\epsilon_k^2}{M^2}.$$

So the inequality is $a_{t+1} < q a_t - c$. By fully applying this recursively, we get

$$\begin{aligned}
&a_1 < q a_0 - c, \\
&a_2 < q a_1 - c < q(q a_0 - c) - c = q^2 a_0 - c(q+1), \\
&a_3 < q a_2 - c < q(q^2 a_0 - c(q+1)) - c = q^3 a_0 - c(q^2 + q + 1),
\end{aligned}$$

and in general

$$a_{T_k} < q^{T_k} a_0 - c \sum_{j=0}^{T_k - 1} q^j = q^{T_k} a_0 - c \frac{1 - q^{T_k}}{1 - q} = q^{T_k} \left(a_0 + \frac{c}{1 - q}\right) - \frac{c}{1 - q}.$$

by noting that $q \in (0,1)$. Substituting back a_t, q and c, we get

$$\begin{aligned}
\|\mathbf{v}^{(T_k)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 &< \left(1 - \frac{\epsilon_k \mu}{M^2}\right)^{T_k} \left(\|\mathbf{v}^{(0)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 + \frac{\epsilon_k^2}{M^2} \Big/ \left(\frac{\epsilon_k \mu}{M^2}\right)\right) - \frac{\epsilon_k^2}{M^2} \Big/ \left(\frac{\epsilon_k \mu}{M^2}\right) \\
&= \left(1 - \frac{\epsilon_k \mu}{M^2}\right)^{T_k} \left(\|\mathbf{v}^{(0)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 + \frac{\epsilon_k}{\mu}\right) - \frac{\epsilon_k}{\mu} \\
&\leq \exp\left(-\frac{\epsilon_k \mu}{M^2} T_k\right) \left(\|\mathbf{v}^{(0)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 + \frac{\epsilon_k}{\mu}\right) - \frac{\epsilon_k}{\mu} \\
&\leq 0,
\end{aligned}$$

where the last inequality is by the definition of T_k, which ensures

$$\exp\left(-\frac{\epsilon_k \mu}{M^2} T_k\right) \leq \frac{\epsilon_k/\mu}{\|\mathbf{v}^{(0)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 + \epsilon_k/\mu} \iff T_k \geq \frac{M^2}{\epsilon_k \mu} \ln\left(1 + \frac{\mu \|\mathbf{w}^{(k)} - \widehat{\mathbf{w}}^{(k+1)}\|^2}{\epsilon_k}\right),$$

and the fact that $\mathbf{v}^{(0)} = \mathbf{w}^{(k)}$. This contradiction indicates that $g_0(\mathbf{w}^{(k+1)}) - g_0(\widehat{\mathbf{w}}^{(k+1)}) \leq \epsilon_k$. Note that $g(\mathbf{w}^{(k+1)}) \leq \epsilon_k$ since $\mathbf{w}^{(k+1)} = \mathbf{v}^{(\tau)}$ with $\tau \in \mathcal{T}$. □

E.2 Distance Between $\mathbf{w}^{(k)}$ and $\widehat{\mathbf{w}}^{(k+1)}$

As follows, we present subproblem (16) defined by a generic solution $\mathbf{w} \in \mathcal{W}$ instead of just $\mathbf{w}^{(k)}$.

$$\begin{aligned}
&\min_{\mathbf{v} \in \mathcal{W}} \; g_0(\mathbf{v}) := f_0^+(\mathbf{v}) - \widehat{f}_0^-(\mathbf{v}; \mathbf{w}) \\
&\text{s.t.} \;\; g_i(\mathbf{v}) := f_i^+(\mathbf{v}) - \widehat{f}_i^-(\mathbf{v}; \mathbf{w}) \leq 0, \; i \in [m],
\end{aligned} \tag{55}$$

where

$$\widehat{f}_i^-(\mathbf{v}; \mathbf{w}) := f_i^-(\mathbf{w}) + (\mathbf{f}_i^-)^\top (\mathbf{v} - \mathbf{w}), \tag{56}$$

and $\mathbf{f}_i^-$ is an arbitrary subgradient in $\partial f_i^-(\mathbf{w})$. Like $\widehat{\mathbf{w}}^{(k+1)}$, let $\widehat{\mathbf{w}}$ be the optimal solution of (55). Furthermore, when $f_i(\mathbf{w}) \leq 0$ for $i \in [m]$, by Assumption 6.3B, the constraints in (55) satisfy the Slater's condition, so there exists a vector of Lagrangian multipliers $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top \in \mathbb{R}_+^m$ such that the following KKT conditions of (55) hold.

$$\mathbf{f}_0^+ - \mathbf{f}_0^- + \textstyle\sum_{i=1}^m \lambda_i (\mathbf{f}_i^+ - \mathbf{f}_i^-) + \mathbf{n}_{\mathcal{W}} = \mathbf{0}, \tag{57}$$

$$g_i(\widehat{\mathbf{w}}) \leq 0, \; i \in [m], \tag{58}$$

$$\lambda_i g_i(\widehat{\mathbf{w}}) = 0, \; i \in [m]. \tag{59}$$

where $\mathbf{n}_{\mathcal{W}} \in \mathcal{N}_{\mathcal{W}}(\widehat{\mathbf{w}})$, $\mathbf{f}_i^+ \in \partial f_i^+(\widehat{\mathbf{w}})$, $\mathbf{f}_i^- \in \partial f_i^-(\mathbf{w})$ for $i = 0, 1, \ldots, m$. Using the three conditions above, we can show that $\mathbf{w}$ is a nearly ϵ-KKT point when $\|\widehat{\mathbf{w}} - \mathbf{w}\| \leq O(\epsilon)$.

Lemma E.1. *Suppose $f_i(\mathbf{w}) \leq 0$ for $i \in [m]$ and $\|\widehat{\mathbf{w}} - \mathbf{w}\| \leq \min\{1, \frac{1}{2M}, \frac{1}{2M \max_{i \in [m]} \lambda_i}\} \cdot \epsilon$,[1] where $\widehat{\mathbf{w}}$ is the optimal solution of (55) and λ_i for $i \in [m]$ are the Lagrangian multipliers corresponding to $\widehat{\mathbf{w}}$. Then $\mathbf{w}$ is a nearly ϵ-KKT solution of (12).*

Proof. Since $f_i(\mathbf{w}) \leq 0$ for $i \in [m]$, we have (57), which means (13) holds with $\|\widehat{\mathbf{w}} - \mathbf{w}\| \leq \epsilon$. Moreover, since g_i is $2M$-Lipschitz continuous and $f_i(\mathbf{w}) = g_i(\mathbf{w})$, (58) and (59) imply that

$$f_i(\mathbf{w}) = g_i(\mathbf{w}) \leq g_i(\widehat{\mathbf{w}}) + 2M\|\widehat{\mathbf{w}} - \mathbf{w}\| \leq \epsilon$$

and

$$|\lambda_i f_i(\mathbf{w})| = |\lambda_i g_i(\mathbf{w})| \leq |\lambda_i g_i(\widehat{\mathbf{w}})| + 2M\lambda_i \|\widehat{\mathbf{w}} - \mathbf{w}\| \leq \epsilon.$$

This means $\mathbf{w}$ is a nearly ϵ-KKT solution of (12). □

[1] Here, we follow the convention that $\frac{1}{\max_{i \in [m]} \lambda_i} = +\infty$ if $\max_{i \in [m]} \lambda_i = 0$.

E.3 Bounded Lagrangian Multipliers

Suppose $f_i(\mathbf{w}^{(k)}) \le 0$ for $i \in [m]$. Let $\boldsymbol{\lambda}^{(k)} = (\lambda_1^{(k)}, \ldots, \lambda_m^{(k)})^\top \in \mathbb{R}_+^m$ be the Lagrangian multipliers corresponding to $\widehat{\mathbf{w}}^{(k+1)}$. In other words, (57), (58) and (59) hold with $\mathbf{w} = \mathbf{w}^{(k)}$, $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(k)}$ and $\widehat{\mathbf{w}} = \widehat{\mathbf{w}}^{(k+1)}$. According to Lemma E.1, to analyze the oracle complexity of IDCA, it suffices to analyze the complexity for finding $\mathbf{w}^{(k)}$ satisfying $\|\mathbf{w}^{(k)} - \widehat{\mathbf{w}}^{(k+1)}\| \le \min\{1, \frac{1}{2M}, \frac{1}{2M \max_{i\in[m]} \lambda_i^{(k)}}\} \cdot \epsilon$. However, this inequality will be difficult to guarantee if $\max_{i\in[m]} \lambda_i^{(k)}$ increases to infinity as k increases. Fortunately, we can provide a uniform upper bound for $\max_{i\in[m]} \lambda_i^{(k)}$ for each k as long as $f_i(\mathbf{w}^{(k)}) \le 0$ for $i \in [m]$ for each k.

Lemma E.2. *Suppose $f_i(\mathbf{w}) \le 0$ for $i \in [m]$. Let $\widehat{\mathbf{w}}$ be the optimal solution of* (55) *and λ_i for $i \in [m]$ be the Lagrangian multipliers corresponding to $\widehat{\mathbf{w}}$. We have*

$$\|\widehat{\mathbf{w}} - \mathbf{w}\| \le 2M/\mu \quad \text{and} \quad \textstyle\sum_{i=1}^m \lambda_i \le \Lambda := 2M/\sqrt{2\mu\nu}. \tag{60}$$

Proof. Since $f_i(\mathbf{w}) \le 0$ for $i \in [m]$. According to Assumption 6.3B, there exists $\mathbf{v} \in \mathcal{W}$ such that

$$g_i(\mathbf{v}) \le -\nu, \ \forall i \in [m],$$

so the conditions (57), (58) and (59) hold.

Choose any $\widehat{\mathbf{f}}_i^- \in \partial f_i^-(\widehat{\mathbf{w}})$ for $i = 0, 1, \ldots, m$. Since $\mathbf{n}_{\mathcal{W}} \in \mathcal{N}_{\mathcal{W}}(\widehat{\mathbf{w}})$, we have $0 \ge -\langle \mathbf{n}_{\mathcal{W}}, \widehat{\mathbf{w}} - \mathbf{w}\rangle$. This inequality and (57) imply

$$\begin{aligned}
0 \ge& \left\langle \mathbf{f}_0^+ - \mathbf{f}_0^- + \textstyle\sum_{i=1}^m \lambda_i(\mathbf{f}_i^+ - \mathbf{f}_i^-), \widehat{\mathbf{w}} - \mathbf{w}\right\rangle \\
=& \left\langle \mathbf{f}_0^+ - \widehat{\mathbf{f}}_0^- + \textstyle\sum_{i=1}^m \lambda_i(\mathbf{f}_i^+ - \widehat{\mathbf{f}}_i^-), \widehat{\mathbf{w}} - \mathbf{w}\right\rangle + \left\langle \widehat{\mathbf{f}}_0^- - \mathbf{f}_0^- + \textstyle\sum_{i=1}^m \lambda_i(\widehat{\mathbf{f}}_i^- - \mathbf{f}_i^-), \widehat{\mathbf{w}} - \mathbf{w}\right\rangle \\
\ge& - \|\mathbf{f}_0^+ - \widehat{\mathbf{f}}_0^- + \textstyle\sum_{i=1}^m \lambda_i(\mathbf{f}_i^+ - \widehat{\mathbf{f}}_i^-)\| \cdot \|\widehat{\mathbf{w}} - \mathbf{w}\| + \mu(1 + \textstyle\sum_{i=1}^m \lambda_i)\|\widehat{\mathbf{w}} - \mathbf{w}\|^2 \\
\ge& - 2(1 + \textstyle\sum_{i=1}^m \lambda_i)M\|\widehat{\mathbf{w}} - \mathbf{w}\| + \mu(1 + \textstyle\sum_{i=1}^m \lambda_i)\|\widehat{\mathbf{w}} - \mathbf{w}\|^2,
\end{aligned}$$

where the second inequality is because of the μ-strongly convexity of f_i^- for $i \in [m]$ and the Cauchy-Schwarz inequality, and the last inequality is by Assumption 6.3D. Reorganizing the inequality above gives the first inequality in (60).

If $\max_{i\in[m]} \lambda_i = 0$, the conclusion holds trivially. Suppose $\max_{i\in[m]} \lambda_i > 0$. Since $\mathbf{f}_i^+ - \mathbf{f}_i^- \in \partial g_i(\widehat{\mathbf{w}})$, $g_i(\cdot)$ is μ-strongly convex by Assumption 6.3A and $\mathbf{n}_{\mathcal{W}} \in \mathcal{N}_{\mathcal{W}}(\widehat{\mathbf{w}})$, we have

$$-\nu \ge g_i(\mathbf{v}) \ge g_i(\widehat{\mathbf{w}}) + \left\langle \mathbf{f}_i^+ - \mathbf{f}_i^- + \mathbf{n}_{\mathcal{W}} / \left(\textstyle\sum_{i=1}^m \lambda_i\right), \mathbf{v} - \widehat{\mathbf{w}}\right\rangle + \frac{\mu}{2}\|\mathbf{v} - \widehat{\mathbf{w}}\|^2.$$

Multiplying both sides of this inequality by λ_i and summing up for $i \in [m]$, we obtain

$$\begin{aligned}
-\nu\left(\textstyle\sum_{i=1}^m \lambda_i\right) \ge& \textstyle\sum_{i=1}^m \lambda_i g_i(\widehat{\mathbf{w}}) + \left\langle \sum_{i=1}^m \lambda_i(\mathbf{f}_i^+ - \mathbf{f}_i^-) + \mathbf{n}_{\mathcal{W}}, \mathbf{v} - \widehat{\mathbf{w}}\right\rangle + \frac{\mu}{2}\left(\textstyle\sum_{i=1}^m \lambda_i\right)\|\mathbf{v} - \widehat{\mathbf{w}}\|^2 \\
\ge& - \frac{\|\sum_{i=1}^m \lambda_i(\mathbf{f}_i^+ - \mathbf{f}_i^-) + \mathbf{n}_{\mathcal{W}}\|^2}{2\mu(\sum_{i=1}^m \lambda_i)} \\
=& - \frac{\|\mathbf{f}_0^+ - \mathbf{f}_0^-\|^2}{2\mu(\sum_{i=1}^m \lambda_i)} \\
\ge& - \frac{2M^2}{\mu(\sum_{i=1}^m \lambda_i)},
\end{aligned}$$

where the second inequality is because of (59) and the Young's inequality, the equality is by (57), and the last inequality is by Assumption 6.3D. Reorganizing the inequality above gives the second inequality in (60). □

E.4 Near Feasibility of Solutions

In order to apply Lemma E.2 and Lemma E.1 to each iteration of Algorithm 1, we need to prove that the following result, which is motivated by Lemma 3.5 in Jia and Grimmer (2025) for weakly convex problems.

Lemma E.3. *Suppose* $\mathbf{w}^{(k)}$ *is not a nearly* ϵ*-KKT solution of* (12)*,* $f_i(\mathbf{w}^{(k)}) \le 0$ *for* $i \in [m]$ *and*

$$\epsilon_k \le \frac{\mu}{8} \min\left\{1, \frac{1}{4M^2}, \frac{\mu\nu}{8M^4}\right\} \cdot \epsilon^2 \tag{61}$$

in the k*-th iteration of Algorithm 1. Then* $f_i(\mathbf{w}^{(k+1)}) \le 0$ *for* $i \in [m]$*.*

Proof. Let $g_i(\cdot)$ be defined as in (16). Let $\boldsymbol{\lambda}^{(k)} = (\lambda_1^{(k)}, \ldots, \lambda_m^{(k)})^\top \in \mathbb{R}_+^m$ be the vector of Lagrangian multipliers corresponding to $\widehat{\mathbf{w}}^{(k+1)}$. Since the Lagrangian function $g_0(\mathbf{v}) + \sum_{i=1}^m \lambda_i^{(k)} g_i(\mathbf{v})$ is $\mu(1 + \sum_{i=1}^m \lambda_i^{(k)})$-strongly convex and is minimized at $\widehat{\mathbf{w}}^{(k+1)}$ over $\mathcal{W}$, we have

$$\begin{aligned} & g_0(\mathbf{w}^{(k+1)}) + \textstyle\sum_{i=1}^m \lambda_i^{(k)} g_i(\mathbf{w}^{(k+1)}) \\ \ge\ & g_0(\widehat{\mathbf{w}}^{(k+1)}) + \textstyle\sum_{i=1}^m \lambda_i^{(k)} g_i(\widehat{\mathbf{w}}^{(k+1)}) + \frac{\mu}{2}\left(1 + \textstyle\sum_{i=1}^m \lambda_i^{(k)}\right) \|\widehat{\mathbf{w}}^{(k+1)} - \mathbf{w}^{(k+1)}\|^2. \end{aligned} \tag{62}$$

Since $\lambda_i^{(k)} g_i(\widehat{\mathbf{w}}^{(k+1)}) = 0$ for $i \in [m]$ and $\mathbf{w}^{(k+1)}$ satisfies (18), the inequality above implies

$$\|\widehat{\mathbf{w}}^{(k+1)} - \mathbf{w}^{(k+1)}\| \le \sqrt{\frac{2\epsilon_k}{\mu}}. \tag{63}$$

If $\|\mathbf{w}^{(k)} - \widehat{\mathbf{w}}^{(k+1)}\| \le \min\{1, \frac{1}{2M}, \frac{\sqrt{2\mu\nu}}{4M^2}\} \cdot \epsilon$ holds, then we must have $\|\mathbf{w}^{(k)} - \widehat{\mathbf{w}}^{(k+1)}\| \le \min\{1, \frac{1}{2M}, \frac{1}{2M \max_{i\in[m]} \lambda_i^{(k)}}\} \cdot \epsilon$ by Lemma E.2 and thus $\mathbf{w}^{(k)}$ is a nearly ϵ-KKT solution of (12) by Lemma E.1. Since we assume $\mathbf{w}^{(k)}$ is not a nearly ϵ-KKT solution, we must have

$$\|\mathbf{w}^{(k)} - \widehat{\mathbf{w}}^{(k+1)}\| > \min\left\{1, \frac{1}{2M}, \frac{\sqrt{2\mu\nu}}{4M^2}\right\} \cdot \epsilon. \tag{64}$$

By (63) and (64),

$$\begin{aligned} \|\mathbf{w}^{(k)} - \mathbf{w}^{(k+1)}\|^2 &\ge \frac{1}{2}\|\mathbf{w}^{(k)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 - \|\mathbf{w}^{(k+1)} - \widehat{\mathbf{w}}^{(k+1)}\|^2 \\ &> \frac{1}{2} \min\left\{1, \frac{1}{4M^2}, \frac{\mu\nu}{8M^4}\right\} \cdot \epsilon^2 - \frac{2\epsilon_k}{\mu}. \end{aligned} \tag{65}$$

The μ-strong convexity of f_i^-, (18) and condition (61) thus implies

$$\begin{aligned} f_i(\mathbf{w}^{(k+1)}) &\le g_i(\mathbf{w}^{(k+1)}) - \frac{\mu}{2}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2 \\ &\le \epsilon_k - \frac{\mu}{4} \min\left\{1, \frac{1}{4M^2}, \frac{\mu\nu}{8M^4}\right\} \cdot \epsilon^2 + \epsilon_k \le 0. \end{aligned} \tag{66}$$

□

E.5 Proof of Theorem 6.5

Proof of Theorem 6.5. Recall that ϵ_k satisfies (61) for $k = 0, 1, \ldots, K-1$ and $f_i(\mathbf{w}^{(0)}) \le 0$ for $i \in [m]$. If $\mathbf{w}^{(0)}$ is not a nearly ϵ-KKT solution of (12), we must have $f_i(\mathbf{w}^{(1)}) \le 0$ for $i \in [m]$ according to Lemma E.3. Repeating this argument for $k = 1, \ldots, K-1$, we can show that, if $\mathbf{w}^{(k)}$ is not a nearly ϵ-KKT solution for $k = 0, 1, \ldots, K-1$, it must hold that $f_i(\mathbf{w}^{(k+1)}) \le 0$ for $i \in [m]$ and for $k = 0, 1, \ldots, K-1$. This allows us to apply Lemma E.2 and Lemma E.1 to each iteration of Algorithm 1.

Similar to (62), we have

$$\begin{aligned}
& f_0(\mathbf{w}^{(k)}) + \textstyle\sum_{i=1}^m \lambda_i^{(k)} f_i(\mathbf{w}^{(k)}) \\
= \; & g_0(\mathbf{w}^{(k)}) + \textstyle\sum_{i=1}^m \lambda_i^{(k)} g_i(\mathbf{w}^{(k)}) \\
\geq \; & g_0(\widehat{\mathbf{w}}^{(k+1)}) + \textstyle\sum_{i=1}^m \lambda_i^{(k)} g_i(\widehat{\mathbf{w}}^{(k+1)}) + \frac{\mu}{2}\left(1 + \textstyle\sum_{i=1}^m \lambda_i^{(k)}\right) \|\widehat{\mathbf{w}}^{(k+1)} - \mathbf{w}^{(k)}\|^2 \\
\geq \; & g_0(\mathbf{w}^{(k+1)}) - \epsilon_k + \frac{\mu}{2}\left(1 + \textstyle\sum_{i=1}^m \lambda_i^{(k)}\right) \|\widehat{\mathbf{w}}^{(k+1)} - \mathbf{w}^{(k)}\|^2 \\
\geq \; & f_0(\mathbf{w}^{(k+1)}) + \frac{\mu}{2}\|\mathbf{w}^{(k)} - \mathbf{w}^{(k+1)}\|^2 - \epsilon_k + \frac{\mu}{2}\left(1 + \textstyle\sum_{i=1}^m \lambda_i^{(k)}\right) \|\widehat{\mathbf{w}}^{(k+1)} - \mathbf{w}^{(k)}\|^2 \\
> \; & f_0(\mathbf{w}^{(k+1)}) + \frac{\mu}{2} \min\left\{1, \frac{1}{4M^2}, \frac{\mu\nu}{8M^4}\right\} \cdot \epsilon^2,
\end{aligned} \tag{67}$$

where the third inequality is by the μ-strong convexity of f_0^-. Recall that $f_i(\mathbf{w}^{(k)}) \leq 0$. Summing up both sides of (67) for $k = 0, 1, \ldots, K-1$ and dividing them by K gives

$$\frac{\mu}{2} \min\left\{1, \frac{1}{4M^2}, \frac{\mu\nu}{8M^4}\right\} \cdot \epsilon^2 < \frac{1}{K}\left(f_0(\mathbf{w}^{(0)}) - f_0(\mathbf{w}^{(K)})\right) \leq \frac{1}{K}\left(f_0(\mathbf{w}^{(0)}) - f_{\text{lb}}\right),$$

contradicting with (19). In fact, with

$$K = \left\lceil 2 \max\left\{1, 4M^2, \frac{8M^4}{\mu\nu}\right\} \frac{\left(f_0(\mathbf{w}^{(0)}) - f_{\text{lb}}\right)}{\mu\epsilon^2} \right\rceil,$$

we have

$$\begin{aligned}
\frac{\mu}{2} \min\left\{1, \frac{1}{4M^2}, \frac{\mu\nu}{8M^4}\right\} \cdot \epsilon^2 &= \frac{\mu\epsilon^2}{2\max\left\{1, 4M^2, \frac{8M^4}{\mu\nu}\right\}} \\
&< \frac{\mu\epsilon^2 (f_0(\mathbf{w}^{(0)}) - f_0(\mathbf{w}^{(K)}))}{2\max\left\{1, 4M^2, \frac{8M^4}{\mu\nu}\right\}(f_0(\mathbf{w}^{(0)}) - f_{\text{lb}})} \leq \frac{1}{K}\left(f_0(\mathbf{w}^{(0)}) - f_0(\mathbf{w}^{(K)})\right).
\end{aligned}$$

This contradiction means one of $\mathbf{w}^{(k)}$ is a nearly ϵ-KKT solution of (12). □

F GENERALIZATION BOUNDS FOR MODELS SOLVED FROM (9)

In this section, we provide statistical guarantees for the fairness constraints and the objective in problem (9). Since the objective corresponds to a standard expected loss (e.g., logistic or squared loss), its generalization properties follow from classical results. Therefore, we focus on the generalization behavior of the fairness constraints.

We analyze the statistical guarantees of the fairness constraints for solutions of (9) in three steps. For simplicity, we assume n_k is the same for each $k \in \mathcal{G}$.

Step 1: Discretization error. By Lemma 5.1, the constraints in (6) are equivalent to (8). We approximate (8) using a finite index set $\widehat{\mathcal{I}} \subset \mathcal{I}$. Assume that the points in $\widehat{\mathcal{I}}$ are equally spaced with spacing $(\beta - \alpha)/|\widehat{\mathcal{I}}|$. Then, by Lipchitz continuity, if $\mathbf{w}$ satisfies (8) for all $p \in \widehat{\mathcal{I}}$, it also satisfies the constraints for all $p \in \mathcal{I}$ up to an error of order $O\left(\frac{\beta-\alpha}{|\widehat{\mathcal{I}}|}\right)$. Consequently, by Lemma 5.1, the original constraints in (6) are satisfied up to an error of order $O\left(\frac{1}{|\widehat{\mathcal{I}}|}\right)$.

Step 2: Statistical estimation error. For each $p \in \widehat{\mathcal{I}}$, we approximate the conditional probabilities in (8) using empirical averages of indicator functions. More specifically, for each $p \in \widehat{\mathcal{I}}$, we approximate (8) by

$$\begin{aligned}
& \tfrac{1}{n_k} \textstyle\sum_{i=1}^{n_k} \mathbf{1}_{h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) > \theta_p} \geq p \quad \text{and} \\
& \tfrac{1}{n_k} \textstyle\sum_{i=1}^{n_k} \mathbf{1}_{h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) > \theta_p} \leq p + \kappa(\beta - \alpha), \quad \forall k \in \mathcal{G}.
\end{aligned} \tag{68}$$

Standard uniform convergence results (see, e.g., Donini et al. (2018); Liang (2016)) imply that, with probability at least $1-\delta$,

$$\sup_{\mathbf{w}\in\mathcal{W},\ \theta_p\in\mathbb{R}} \left| \Pr\left(h_{\mathbf{w}}(\xi) > \theta_p \mid \gamma = k\right) - \frac{1}{n_k}\sum_{i=1}^{n_k} \mathbf{1}\{h_{\mathbf{w}}(\xi_i^k) > \theta_p\} \right| \le E(n_k),$$

for all $k \in \mathcal{G}$ and $p \in \widehat{\mathcal{I}}$, where

$$E(n_k) := \sqrt{\frac{8\,VC\left(\log n_k + 1\right)}{n_k}} + \sqrt{\frac{2\log(4|\widehat{\mathcal{I}}||\mathcal{G}|/\delta)}{n_k}},$$

and VC denotes the VC-dimension of the function class

$$\{\xi \mapsto \mathbf{1}\{h_{\mathbf{w}}(\xi) > \theta_p\} \mid \mathbf{w}\in\mathcal{W},\ \theta_p \in \mathbb{R}\}.$$

This implies that any $\mathbf{w}$ satisfying the constraints in (68) for each $p \in \widehat{\mathcal{I}}$ also satisfies the population constraints in (8) $p \in \widehat{\mathcal{I}}$ up to an error of $O(E(n_k))$.

Step 3: Surrogate approximation error. In (9), the indicator function $\mathbf{1}\{h_{\mathbf{w}}(\xi) > \theta_p\}$ in (68) has been replaced by a smooth surrogate $\sigma(h_{\mathbf{w}}(\xi) - \theta_p)$. The resulting approximation error, known as the *surrogate fairness gap*, has been studied in Yao et al. (2024). Following arguments similar to Theorem 2 in Yao et al. (2024), we obtain that if $\mathbf{w}$ satisfies a surrogate constraint in (9) for some $k \in \mathcal{G}$ and some $p \in \widehat{\mathcal{I}}$, then for any $\gamma \in (0,1)$, it also satisfies the constraint in (68) for the same $k \in \mathcal{G}$ and $p \in \widehat{\mathcal{I}}$ up to an error of

$$O\left(\gamma + \frac{n_k^\gamma}{n_k}\right),$$

where n_k^γ denotes the number of training samples in group k such that

$$\sigma\big(|h_{\mathbf{w}}(\xi) - \theta_p|\big) \le 1-\gamma.$$

Combining the three steps with the triangle inequality easily yields the following generalization guarantees.

Theorem F.1 (Generalization bound for fairness constraints)**.** *Let* $\mathbf{w}$ *be a feasible solution of* (9) *with parameter* κ*. Then, with probability at least* $1-\delta$*,* $\mathbf{w}$ *satisfies the constraints of* (6) *with* κ *there replaced by* $\kappa + O\left(\frac{1}{|\widehat{\mathcal{I}}|} + \frac{1}{\beta-\alpha}\max_{k\in\mathcal{G}}\left(\gamma + \frac{n_k^\gamma}{n_k} + E(n_k)\right)\right)$*. Equivalently,* $\mathbf{w}$ *satisfies partial statistical parity up to an error of* $\kappa + O\left(\frac{1}{|\widehat{\mathcal{I}}|} + \frac{1}{\beta-\alpha}\max_{k\in\mathcal{G}}\left(\gamma + \frac{n_k^\gamma}{n_k} + E(n_k)\right)\right)$*.*

Let $f_0(\mathbf{w}) = \mathbb{E}[\ell(h_{\mathbf{w}}(\xi), \zeta)]$ denote the population risk, where $\ell(\cdot,\cdot)$ is a loss function, and let

$$\hat{f}_0(\mathbf{w}) = \frac{1}{n_0}\sum_{i=1}^{n_0} \ell(h_{\mathbf{w}}(\xi_i^0), \zeta_i^0)$$

be the empirical risk. Let $\widehat{\mathbf{w}}$ be the optimal solution of the following optimization problem

$$\begin{aligned} \min_{\mathbf{w}\in\mathcal{W},\ (\theta_p)_{p\in\widehat{\mathcal{I}}}} \quad & \hat{f}_0(\mathbf{w}) \\ \text{s.t.} \quad & \frac{1}{n_k}\sum_{i=1}^{n_k} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - \theta_p) \ge p,\ \forall p \in \widehat{\mathcal{I}}, k \in \mathcal{G}, \\ & \frac{1}{n_k}\sum_{i=1}^{n_k} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - \theta_p) \le p + \kappa(\beta-\alpha),\ \forall p \in \widehat{\mathcal{I}}, k \in \mathcal{G}, \end{aligned} \tag{69}$$

and let $\mathbf{w}^*$ be the optimal solution of the following problem

$$\min_{\mathbf{w}\in\mathcal{W}} f_0(\mathbf{w}) \ \text{ s.t. } \ \mathrm{SP}_{k,k'}^{\mathcal{I}}(\mathbf{w}) \le \kappa',\ \forall k, k' \in \mathcal{G}, \tag{70}$$

which is the same as (6) except that κ in (6) is replaced by

$$\kappa' = \kappa - O\left(\frac{1}{|\widehat{\mathcal{I}}|} + \frac{1}{\beta - \alpha}\max_{k\in\mathcal{G}}\left(\gamma + \frac{n_k^\gamma}{n_k} + E(n_k)\right)\right).$$

According to the proof of Theorem F.1, we can show that, with probability at least $1-\delta$, $\mathbf{w}^*$ must be a feasible solution of (69) and thus $\hat{f}_0(\widehat{\mathbf{w}}) \le \hat{f}_0(\mathbf{w}^*)$.

Therefore, we have

$$f_0(\widehat{\mathbf{w}}) - f_0(\mathbf{w}^*) = \left[f_0(\widehat{\mathbf{w}}) - \hat{f}_0(\widehat{\mathbf{w}})\right] + \left[\hat{f}_0(\widehat{\mathbf{w}}) - \hat{f}_0(\mathbf{w}^*)\right] + \left[\hat{f}_0(\mathbf{w}^*) - f_0(\mathbf{w}^*)\right],$$

and the second term here is less than or equal to zero. The first and third terms can be bounded by the standard uniform convergence results (see, e.g., Donini et al. (2018); Liang (2016)). Therefore, we have the following statistical guarantee on the objective value of $\widehat{\mathbf{w}}$ relative to the optimal value of (70).

Theorem F.2 (Generalization bound for the objective)**.** *Let $\widehat{\mathbf{w}}$ be the optimal solution of* (69) *and $\mathbf{w}^*$ be the optimal solution of* (70)*. Then, with probability at least $1-\delta$,*

$$f_0(\widehat{\mathbf{w}}) - f_0(\mathbf{w}^*) \le \mathcal{R}_{n_0} + \sqrt{\frac{2\log(4|\widehat{\mathcal{I}}||\mathcal{G}|/\delta)}{n_0}},$$

where $\mathcal{R}_{n_0}$ denotes the Rademacher complexity of the function class

$$\{(\xi,\zeta) \mapsto \ell(h_{\mathbf{w}}(\xi),\zeta) \mid \mathbf{w}\in\mathcal{W}\}.$$

A similar argument yields generalization guarantees for the partial demographic parity formulation in (11).

G ADDITIONAL DETAILS OF NUMERICAL EXPERIMENTS

In this section, additional details and results are provided for the numerical experiments in Section 7.

G.1 Details of Datasets

The details of the three public datasets we used in our numerical experiments are as follows.

- With *a9a* dataset, the task is to predict if the annual income of an individual exceeds \$50K. Gender is the sensitive attribute that splits the data into female ($\gamma = 1$) and male ($\gamma = 2$) groups.
- With *bank* dataset, the task is to predict if a client will subscribe a term deposit with a bank. Age is the sensitive attribute that splits the data into the clients with ages between 25 and 60 ($\gamma = 1$) and clients with other ages ($\gamma = 2$).
- With *law school* dataset, the task is to predict if a student will pass the bar exam of law schools on the first try. Race is the sensitive attribute that splits the data into white ($\gamma = 1$) or non-white ($\gamma = 2$) groups.
- With *ACS* dataset, the task is to predict whether a US working adults' yearly income is above $50,000$. Gender is the sensitive attribute that splits the data into female ($\gamma = 1$) and male ($\gamma = 2$) groups.

Some statistics of these datasets are given in Table 1. Data *a9a* was provided with separated training and testing sets, so we further randomly split the original training data into a new training set (90%) and a validation set (10%). For *bank* dataset, we randomly split it into training (60%), validation (20%) and testing (20%) sets. For *law school* dataset, we randomly split it into training (56.25%), validation (18.75%) and testing (25%) sets. For *ACS* dataset, we randomly split it into training (60%), validation (20%) and testing (20%) sets.

The training sets are used to build the models. The validation sets are only used for tuning hyper-parameters. The testing sets are used to evaluate the prediction performance and the fairness of the obtained models. The variation of the performance of each method caused by the aforementioned random data splitting is displayed by the error bars (95% confidence intervals) in Figure 1.

Table 1: Statistics of the datasets.

Datasets	#Instances	#Attributes	Class Label	Sensitive Attribute
a9a	48,842	123	Income	Gender
bank	41,188	54	Subscription	Age
law school	20,798	12	Passing exam	Race
ACS	1,664,500	813	Income	Gender

G.2 Implementation Details of the Proposed Methods

In this section, we present the implementation details of the proposed methods that produce the numerical results in Section 7, including the choices of κ in (9) and (11) and the choices of tuning parameters in IDCA (Algorithm 1).

G.2.1 Choices of κ in (9) and (11)

Through the experiments, the prediction performance and the fairness of a model solved from (9) or (11) do not change with κ at a consistent pace. Therefore, the points in the Pareto frontier will not be nicely spaced if we choose κ from an equally spaced sequence within $[0, 1]$. For a better visualization, we choose κ manually for different datasets and different fairness metrics such that the points along the Pareto frontiers in Figure 1 are reasonably spaced. Also, we always make sure a small enough κ is included such that the frontiers can approach a high level of fairness (towards the right end of the x-axis).

The values of κ we used in different cases are presented below.

- For problem (9) (pSP fairness):
  a9a: $\kappa \in \{0.01, 0.05, 0.1, 0.2, 0.28\}$.
  bank: $\kappa \in \{0.01, 0.08, 0.22, 0.3\}$.
  law school: $\kappa \in \{0.005, 0.08, 0.1, 0.15, 0.2\}$.
- For problem (11) (pDP fairness):
  a9a: $\kappa \in \{0.005, 0.03, 0.05, 0.1, 0.15\}$.
  bank: $\kappa \in \{0.01, 0.03, 0.05, 0.07, 0.1\}$.
  law school: $\kappa \in \{0.005, 0.04, 0.06, 0.1, 0.15\}$.

G.2.2 Choices of Tuning Parameters in IDCA

For each κ listed in Section G.2.1, we apply IDCA to (9) and (11) with the initial solution $\mathbf{w}^{(0)} = \mathbf{0}$. For each κ and each random splitting of training, validation, and testing sets, we select T from $\{150, 200\}$ and set $\epsilon_k = \epsilon$ with ϵ selected from $\{5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}\}$. We first select the combination of T and ϵ that produces the highest classification accuracy on the validation set after 50 outer iterations. Then we use the best combination of T and ϵ to run IDCA with the total number of outer iterations K selected from $\{100, 150, 200, 250, 300, 350, 400\}$ that produces the highest accuracy on the validation set. Then we plot the classification accuracies and fairness values of the final solution on the testing set to create the Pareto frontiers in the figures.

G.3 Implementation Details of the In-processing Baselines in Comparison

In this section, we present the optimization models of the in-processing methods we compare with in Section 7 and describe the numerical methods used to solve these models.

G.3.1 Constrained Optimization Models for the In-processing Baselines

We compare our methods with other in-processing methods that solve constrained optimization problems with the same objective function as our approach but different fairness constraints. More specifically, we obtain a model by

minimizing the same $f_0(\mathbf{w})$ as in (9) and (11) subject to the constraints that enforce Group AUC Fairness (Yao et al., 2023; Vogel et al., 2021), Inter-Group Pairwise Fairness (Kallus and Zhou, 2019; Beutel et al., 2019), and Intra-Group Pairwise Fairness (Beutel et al., 2019). The three models are defined below.

Baseline G.1. **Group AUC Fairness**. Suppose the data points are ranked in descending order based on their scores $h_{\mathbf{w}}(\boldsymbol{\xi})$. Group AUC fairness requires that, with probability 50%, a random data point sampled from any group is not ranked lower than a random data point sampled from a different group, that is, (21) holds. Recall that we have partitioned the training data into subsets based on γ, namely, $\mathcal{D} = \cup_{k\in\mathcal{G}}\mathcal{D}_k$, where

$$\mathcal{D}_k = \{(\boldsymbol{\xi}, \zeta, \gamma) \in \mathcal{D} | \gamma = k\} = \{(\boldsymbol{\xi}_i^k, \zeta_i^k, k)\}_{i=1}^{n_k},$$

and n_k is the size of data from group k. Similar to (50), we will approximate the probability above using the sample average of indicator $\mathbf{1}_{h_{\mathbf{w}}(\boldsymbol{\xi})-h_{\mathbf{w}}(\boldsymbol{\xi}')\geq 0}$ over all data pairs between groups k and k'. Then we approximate the indicator function by a surrogate $\sigma(h_{\mathbf{w}}(\boldsymbol{\xi}) - h_{\mathbf{w}}(\boldsymbol{\xi}'))$. This leads to the following optimization problem similar to (9) and (11),

$$\begin{aligned} &\min_{\mathbf{w}\in\mathcal{W}} f_0(\mathbf{w}) \\ &\text{s.t.} \quad \frac{1}{n_k n_{k'}} \sum_{i=1}^{n_k}\sum_{j=1}^{n_{k'}} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - h_{\mathbf{w}}(\boldsymbol{\xi}_j^{k'})) - 0.5 \leq \kappa, \ \forall k, k' \in \mathcal{G}. \end{aligned} \tag{71}$$

Following Yao et al. (2023), we choose $\sigma(\cdot)$ to be the quadratic surrogate in the numerical experiments, namely,

$$\sigma(x) = \frac{1}{2}(1+x)^2. \tag{72}$$

Baseline G.2. **Inter-Group Pairwise Fairness**. Suppose the data points are ranked in descending order based on their scores $h_{\mathbf{w}}(\boldsymbol{\xi})$. Inter-Group Pairwise Fairness requires that the probability of ranking a random positive data point from group k higher than a random negative data point from group k' must be the same as the probability of ranking a random positive data point from group k' higher than a random negative data point from group k. In other words, it requires

$$\begin{aligned} &\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) \geq h_{\mathbf{w}}(\boldsymbol{\xi}') | \gamma = k, \gamma' = k', \zeta = 1, \zeta' = -1) \\ =& \Pr(h_{\mathbf{w}}(\boldsymbol{\xi}') \geq h_{\mathbf{w}}(\boldsymbol{\xi}) | \gamma = k, \gamma' = k', \zeta = -1, \zeta' = 1), \ \forall k, k' \in \mathcal{G}, \end{aligned}$$

where $(\boldsymbol{\xi}, \zeta, \gamma)$ and $(\boldsymbol{\xi}', \zeta', \gamma')$ are two i.i.d. random data points. It can be reformulated equivalently as

$$\begin{aligned} &\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) \geq h_{\mathbf{w}}(\boldsymbol{\xi}') | \gamma = k, \gamma' = k', \zeta = 1, \zeta' = -1) \\ +& \Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) \geq h_{\mathbf{w}}(\boldsymbol{\xi}') | \gamma = k, \gamma' = k', \zeta = -1, \zeta' = 1) = 1, \ \forall k, k' \in \mathcal{G}. \end{aligned} \tag{73}$$

To derive a tractable approximation of (73), we first partition the training data into subsets based on γ and ζ, namely, $\mathcal{D} = (\cup_{k\in\mathcal{G}}\mathcal{D}_{k,+}) \cup (\cup_{k\in\mathcal{G}}\mathcal{D}_{k,-})$, where

$$\mathcal{D}_{k,+} = \{(\boldsymbol{\xi}, \zeta, \gamma) \in \mathcal{D} | \gamma = k, \zeta = 1\} = \{(\boldsymbol{\xi}_i^{k,+}, 1, k)\}_{i=1}^{n_{k,+}}$$

and

$$\mathcal{D}_{k,-} = \{(\boldsymbol{\xi}, \zeta, \gamma) \in \mathcal{D} | \gamma = k, \zeta = -1\} = \{(\boldsymbol{\xi}_i^{k,-}, -1, k)\}_{i=1}^{n_{k,-}},$$

where $n_{k,+}$ and $n_{k,-}$ are the group sizes ($n_{k,+} + n_{k,-} = n_k$). Similar to (71), using sample approximation and surrogates, we obtain the following optimization problem based on fairness metric (73),

$$\begin{aligned} &\min_{\mathbf{w}\in\mathcal{W}} f_0(\mathbf{w}) \\ &\text{s.t.} \quad \frac{1}{n_{k,+} n_{k',-}} \sum_{i=1}^{n_{k,+}}\sum_{j=1}^{n_{k',-}} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k,+}) - h_{\mathbf{w}}(\boldsymbol{\xi}_j^{k',-})) \\ &\quad + \frac{1}{n_{k,-} n_{k',+}} \sum_{i=1}^{n_{k,-}}\sum_{j=1}^{n_{k',+}} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_j^{k,-}) - h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k',+})) - 1 \leq \kappa, \ \forall k, k' \in \mathcal{G}. \end{aligned} \tag{74}$$

We choose $\sigma(\cdot)$ to be (72) also.

Baseline G.3. **Intra-Group Pairwise Fairness**. Suppose the data points are ranked in descending order based on their scores $h_{\mathbf{w}}(\boldsymbol{\xi})$. Intra-Group Pairwise Fairness requires that the probability of ranking a random positive data point from group k higher than a random negative data point from group k must be the same as the probability of ranking a random positive data point from group k' higher than a random negative data point from group k'. In other words, it requires

$$\begin{aligned}&\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) \geq h_{\mathbf{w}}(\boldsymbol{\xi}')|\gamma = \gamma' = k, \zeta = 1, \zeta' = -1)\\ =\ &\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) \geq h_{\mathbf{w}}(\boldsymbol{\xi}')|\gamma = \gamma' = k', \zeta = 1, \zeta' = -1),\ \forall k, k' \in \mathcal{G},\end{aligned}$$

where $(\boldsymbol{\xi}, \zeta, \gamma)$ and $(\boldsymbol{\xi}', \zeta', \gamma')$ are two i.i.d. random data points. It can be reformulated equivalently as

$$\begin{aligned}&\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) \geq h_{\mathbf{w}}(\boldsymbol{\xi}')|\gamma = \gamma' = k, \zeta = 1, \zeta' = -1)\\ +\ &\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}') \geq h_{\mathbf{w}}(\boldsymbol{\xi})|\gamma = \gamma' = k', \zeta = 1, \zeta' = -1) = 1,\ \forall k, k' \in \mathcal{G}.\end{aligned} \tag{75}$$

Similar to (74), we obtain the following optimization problem based on fairness metric (75),

$$\begin{aligned}&\min_{\mathbf{w}\in\mathcal{W}} f_0(\mathbf{w}) \\ &\text{s.t.}\ \frac{1}{n_{k,+}n_{k,-}} \sum_{i=1}^{n_{k,+}} \sum_{j=1}^{n_{k,-}} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k,+}) - h_{\mathbf{w}}(\boldsymbol{\xi}_j^{k,-}))\\ &\quad + \frac{1}{n_{k',-}n_{k',+}} \sum_{i=1}^{n_{k',-}} \sum_{j=1}^{n_{k',+}} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_j^{k',-}) - h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k',+})) - 1 \leq \kappa,\ \forall k, k' \in \mathcal{G}.\end{aligned} \tag{76}$$

We choose $\sigma(\cdot)$ to be (72) once again.

G.3.2 Training Algorithm for the In-processing Baselines

When implementing the three in-processing baselines in the numerical experiments, we set the objective function $f_0(\mathbf{w})$ in (71), (74) and (76) to be the same as the one used in our methods, which is defined in Section 7.1. We then solve (71), (74) and (76) using the projected gradient descent method. In particular, starting with $\mathbf{w}^{(0)} = \mathbf{0}$, we generate $\mathbf{w}^{(k)}$ for $k = 0, 1, \ldots, K-1$ as

$$\mathbf{w}^{(k+1)} = \arg\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{2} \left\| \mathbf{w} - \left(\mathbf{w}^{(k)} - \eta_k \nabla f_0(\mathbf{w}^{(k)})\right) \right\|^2 \ \text{s.t. } \mathbf{w} \text{ satisfies the constraints in (71)}, \tag{77}$$

$$\mathbf{w}^{(k+1)} = \arg\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{2} \left\| \mathbf{w} - \left(\mathbf{w}^{(k)} - \eta_k \nabla f_0(\mathbf{w}^{(k)})\right) \right\|^2 \ \text{s.t. } \mathbf{w} \text{ satisfies the constraints in (74)}, \tag{78}$$

$$\mathbf{w}^{(k+1)} = \arg\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{2} \left\| \mathbf{w} - \left(\mathbf{w}^{(k)} - \eta_k \nabla f_0(\mathbf{w}^{(k)})\right) \right\|^2 \ \text{s.t. } \mathbf{w} \text{ satisfies the constraints in (76)}, \tag{79}$$

for solving (71), (74) and (76), respectively, where η_k is the step size.

Because of (72) and the fact that the model $h_{\mathbf{w}}(\boldsymbol{\xi})$ we choose is linear in $\mathbf{w}$, the constraints in (71), (74) and (76) are all convex quadratic constraints on $\mathbf{w}$ and the Hessian matrices of the constraint functions only need to be computed once and used in each iteration. Also, the projection problems (77), (78) and (79) are convex quadratically constrained quadratic programs (QCQP), which we solve by using `fmincon` function from MATLAB. `fmincon` solves the convex QCQPs using the interior-point method with parameters 'MaxIterations'=1000, 'OptimalityTolerance'=1e-6, and 'ConstraintTolerance'=1e-6. We terminate the training algorithm after $K = 1000$ iterations for the *a9a* and *bank* datasets and $K = 500$ iterations for *the law school* dataset. The numbers of iterations we choose are large enough to ensure the change of the objective value is less than 10^{-5} at termination in all cases.

In order to create the Pareto frontiers by the models solved from (71), (74) and (76), we also need to solve them with different values of κ. Moreover, in order to compare the prediction performances of different models at similar levels of fairness, we choose a sequence of values for κ in (71), (74) and (76) so the models solved from them produce $\max_{k,k} \mathrm{SP}^{\mathcal{I}}_{k,k'}(\mathbf{w})$ and $\max_{k,k} \mathrm{DP}^{\mathcal{I}}_{k,k'}(\mathbf{w})$ similar to our methods on the validation set. The values of κ we used in the baseline optimization models to produce Figure 1 are presented below.

- For problem (71) (pSP fairness):
  a9a: $\kappa \in \{0.7, 1, 2, 3.5\}$.
  bank: $\kappa \in \{0.1, 0.3, 0.56, 0.7, 1\}$.
  law school: $\kappa \in \{0.5, 0.75, 1.15, 1.8, 2.7\}$.
- For problem (71) (pDP fairness):
  a9a: $\kappa \in \{0.3, 0.4, 0.5, 0.7, 1\}$.
  bank: $\kappa \in \{0.06, 0.5, 2.35, 2.62, 3.28, 4\}$.
  law school: $\kappa \in \{1, 3, 4, 5, 6, 10\}$.
- For problem (74) (pSP fairness):
  a9a: $\kappa \in \{1.5, 3, 3.5, 6\}$.
  bank: $\kappa \in \{1.5, 1.8, 2, 3\}$.
  law school: $\kappa \in \{0.4, 1.6, 2.4, 4\}$.
- For problem (74) (pDP fairness):
  a9a: $\kappa \in \{0.9, 1.2, 1.5, 2, 2.7\}$.
  bank: $\kappa \in \{0.3, 1, 5, 5.4, 6\}$.
  law school: $\kappa \in \{2, 6.5, 8, 9, 10, 15\}$.
- For problem (76) (pSP fairness):
  a9a: $\kappa \in \{3, 5, 8\}$.
  bank: $\kappa \in \{2, 3, 4\}$.
  law school: $\kappa \in \{3, 5, 10\}$.
- For problem (76) (pDP fairness):
  a9a: $\kappa \in \{0.7, 1, 1.3, 1.7, 2.2\}$.
  bank: $\kappa \in \{0.1, 0.3, 0.42, 0.5, 1, 2.5\}$.
  law school: $\kappa \in \{0.1, 0.8, 2, 3.5, 10\}$.

G.4 Implementation Details for the Regularization Method in Comparison

In this section, we present the optimization models of the in-processing methods we compare with in Section 7 and describe the numerical methods used to solve these models.

G.4.1 Optimization Model for the Regularization Method

We can also solve (6) using a regularization method (also called a penalty method) using either $\max_{k,k'\in\mathcal{G}} \mathrm{SP}^{\mathcal{I}}_{k,k'}(\mathbf{w})$ in (4) or $\max_{k,k'\in\mathcal{G}} \mathrm{DP}^{\mathcal{I}}_{k,k'}(\mathbf{w})$ in (5) as a regularization term (also called a penalty term). In particular, given a regularization parameter $\lambda \geq 0$, we solve

$$\min_{\mathbf{w}\in\mathcal{W}} f_0(\mathbf{w}) + \lambda \max_{k,k'\in\mathcal{G}} \mathrm{SP}^{\mathcal{I}}_{k,k'}(\mathbf{w}), \tag{80}$$

$$\min_{\mathbf{w}\in\mathcal{W}} f_0(\mathbf{w}) + \lambda \max_{k,k'\in\mathcal{G}} \mathrm{DP}^{\mathcal{I}}_{k,k'}(\mathbf{w}), \tag{81}$$

respectively. We will reformulate (80) and (81) as unconstrained optimization problems by following a procedure similar to the one we used in (9) to derive (9) and (11). We present this procedure in detail below.

First, computing the subgradient of $\max_{k,k'\in\mathcal{G}} \mathrm{SP}^{\mathcal{I}}_{k,k'}(\mathbf{w})$ is difficult because of the two challenges mentioned at the beginning of Section 5 (i.e., how to approximate the probabilities) as well as the challenge in taking the

maximization over θ in (4). According to Lemma 5.1 with $\kappa = 0$, the constraint $\mathrm{SP}^{\mathcal{I}}_{k,k'}(\mathbf{w}) = 0$ holds if and only if, for any $p \in [\alpha, \beta)$, there exists $\theta_p \in \mathbb{R}$ such that

$$\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta_p | \gamma = k) = p, \ \forall k \in \mathcal{G}. \tag{82}$$

Following the same logic and proof as Lemma 5.1, we can show that

$$\max_{k,k' \in \mathcal{G}} \mathrm{SP}^{\mathcal{I}}_{k,k'}(\mathbf{w}) = \frac{2}{\beta - \alpha} \max_{p \in [\alpha,\beta)} \min_{\theta_p} \max_{k \in \mathcal{G}} \left| \Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta_p \big| \gamma = k\big) - p \right|.$$

Similar to Section 5, we approximate each probability by the empirical distribution and a continuous surrogate $\sigma(x) \approx \mathbf{1}_{x>0}$, and approximate the maximization $\max_{p \in [\alpha,\beta)}$ by $\max_{p \in \widehat{\mathcal{I}}}$ where $\widehat{\mathcal{I}}$ is a finite subset satisfying $\widehat{\mathcal{I}} \subset [\alpha, \beta)$. We then obtain

$$\begin{aligned}
\max_{k,k' \in \mathcal{G}} \mathrm{SP}^{\mathcal{I}}_{k,k'}(\mathbf{w}) &\approx \frac{2}{\beta - \alpha} \max_{p \in \widehat{\mathcal{I}}} \min_{\theta_p} \max_{k \in \mathcal{G}} \left| \Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \theta_p \big| \gamma = k\big) - p \right| \\
&\approx \frac{2}{\beta - \alpha} \max_{p \in \widehat{\mathcal{I}}} \min_{\theta_p} \max_{k \in \mathcal{G}} \left| \frac{1}{n_k} \sum_{i=1}^{n_k} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - \theta_p) - p \right|.
\end{aligned}$$

Applying this approximation to (80), we obtain the following regularized variant of (9)

$$\min_{\mathbf{w} \in \mathcal{W},\ (\theta_p)_{p \in \widehat{\mathcal{I}}}} f_0(\mathbf{w}) + \frac{2\lambda}{\beta - \alpha} \max_{p \in \widehat{\mathcal{I}}} \max_{k \in \mathcal{G}} \left| \frac{1}{n_k} \sum_{i=1}^{n_k} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - \theta_p) - p \right|. \tag{83}$$

According to the proof of Lemma 5.2, we have

$$\mathrm{DP}^{\mathcal{I}}_{k,k'}(\mathbf{w}) = \frac{1}{\beta - \alpha} \left| \begin{array}{l} \min\{\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta} | \gamma = k), \beta\} \\ -\min\{\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta} | \gamma = k), \alpha\} \\ -\min\{\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta} | \gamma = k'), \beta\} \\ +\min\{\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta} | \gamma = k'), \alpha\} \end{array} \right|.$$

Similar to Section 5, we approximate each probability by the empirical distribution and a continuous surrogate $\sigma(x) \approx \mathbf{1}_{x>0}$ on the right-hand side above, and obtain

$$\mathrm{DP}^{\mathcal{I}}_{k,k'}(\mathbf{w}) \approx \frac{1}{\beta - \alpha} \left| \begin{array}{l} \min\left\{\frac{1}{n_k} \sum_{i=1}^{n_k} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - \widehat{\theta}), \beta\right\} \\ -\min\left\{\frac{1}{n_k} \sum_{i=1}^{n_k} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - \widehat{\theta}), \alpha\right\} \\ -\min\left\{\frac{1}{n_{k'}} \sum_{i=1}^{n_{k'}} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k'}) - \widehat{\theta}), \beta\right\} \\ +\min\left\{\frac{1}{n_{k'}} \sum_{i=1}^{n_{k'}} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k'}) - \widehat{\theta}), \alpha\right\} \end{array} \right|.$$

Applying this approximation to (81), we obtain the following regularized variant of (11)

$$\min_{\mathbf{w} \in \mathcal{W}} f_0(\mathbf{w}) + \frac{\lambda}{\beta - \alpha} \left| \begin{array}{l} \min\left\{\frac{1}{n_k} \sum_{i=1}^{n_k} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - \widehat{\theta}), \beta\right\} \\ -\min\left\{\frac{1}{n_k} \sum_{i=1}^{n_k} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^k) - \widehat{\theta}), \alpha\right\} \\ -\min\left\{\frac{1}{n_{k'}} \sum_{i=1}^{n_{k'}} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k'}) - \widehat{\theta}), \beta\right\} \\ +\min\left\{\frac{1}{n_{k'}} \sum_{i=1}^{n_{k'}} \sigma(h_{\mathbf{w}}(\boldsymbol{\xi}_i^{k'}) - \widehat{\theta}), \alpha\right\} \end{array} \right|. \tag{84}$$

G.4.2 Training Algorithm for the Regularization Method

We first create a sequence of values of λ (given later). For each value of λ, we apply the standard subgradient descent (SGD) algorithm to (83) and (84) by fine-tuning the step size. In particular, we select the step size in SGD from $\{5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}\}$ and select the one that produces the highest accuracy on the validation set after 10^4 iterations. Then we run SGD using the selected step size for the same total number of

iterations as our method, i.e., TK iterations, and use the last iteration to evaluate the classification accuracies and fairness values on the testing set to create the Pareto frontiers in the figures.

In order to create the Pareto frontiers by the models solved from (83) and (84), we also need to solve them with different values of λ. To compare the prediction performances of different models at similar levels of fairness, we choose a sequence of values for λ in (83) and (84) so the output models produce $\max_{k,k} \mathrm{SP}^{\mathcal{I}}_{k,k'}(\mathbf{w})$ and $\max_{k,k} \mathrm{DP}^{\mathcal{I}}_{k,k'}(\mathbf{w})$ similar to our methods on the validation sets. The values of λ we used in (83) and (84) to produce Figure 1 are presented below.

- For problem (83) (pSP fairness):

  a9a: $\lambda \in \{0.2, 0.3, 0.37, 1\}$.

  bank: $\lambda \in \{0.4, 0.405, 0.41, 1, 1.5\}$.

  law school: $\lambda \in \{0.39, 0.4, 0.5, 0.8, 1.5\}$.

- For problem (84) (pDP fairness):

  a9a: $\lambda \in \{0.13, 0.17, 0.2, 0.23, 1\}$.

  bank: $\lambda \in \{0.005, 0.015, 0.018, 0.1, 2\}$.

  law school: $\lambda \in \{0.001, 0.01, 1\}$.

G.5 Changes of Score Densities with κ

In this section, we present how the densities of the prediction scores produced by our methods change with κ on different datasets. In particular, in Figures 3-8, we plot the estimated densities of $h_{\mathbf{w}}(\boldsymbol{\xi})$ by the models solved from (9) and (11) with different values of κ. The density curves are estimated by the `ksdensity` function from MATLAB using the values of $h_{\mathbf{w}}(\boldsymbol{\xi})$ on the testing sets. For comparisons, the densities of the two sensitive groups are plotted separately, and the areas within the interval $\mathcal{I}$ are highlighted in red.

According to Figures 3-8, we have the following observations:

- When pSP fairness is enforced (Figure 3, Figure 4, and Figure 5), as κ decreases, the distributions of $h_{\mathbf{w}}(\boldsymbol{\xi})$ on $\mathcal{I}$ on the two groups become more and more similar.

- When pDP fairness is enforced (Figure 6, Figure 7, and Figure 8), as κ decreases, the proportions of the red areas above 0, i.e., $\Pr\big(h_{\mathbf{w}}(\boldsymbol{\xi}) > \widehat{\theta} \big| \gamma = k, \bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\boldsymbol{\xi})) \in \mathcal{I}\big)$ with $\widehat{\theta} = 0$, on the two groups become more and more similar.

Both observations are intuitive and consistent with the goals of the two partial fairness metrics we enforce through the constraints in (9) and (11).

G.6 Robustness to the Choice of $\mathcal{I}$

A key remaining question is whether the effectiveness of our method is robust to the choice of interval $\mathcal{I}$. In this section, we provide an affirmative answer to this question by comparing our method with other approaches when different intervals $\mathcal{I}$'s are used. For each dataset, we create a collection of intervals $\mathcal{I}$'s (see below). Then, for each interval, we use the same κ in (9) to train a model, and then evaluate its classification accuracy and pSP fairness. For the other in-processing methods in (71), (74), (76), and (83), we choose the value of κ or λ in their optimization models such that their output solutions achieve an accuracy similar to the one from (9). After this, we evaluate the pSP fairness of their solutions. Finally, we compare the accuracy and pSP fairness of all methods using the mirrored grouped bar charts shown in Figure 9. As expected, the accuracies of all methods are the same for each interval because κ or λ is chosen intentionally to ensure that happens. The differences therefore lie in their pSP fairness, and our method demonstrates the most robust performance across different intervals in the sense of achieve the same highest and most stable fairness as the interval changes.

The interval we choose in this experiment and the values of κ or λ used in each method are given below.

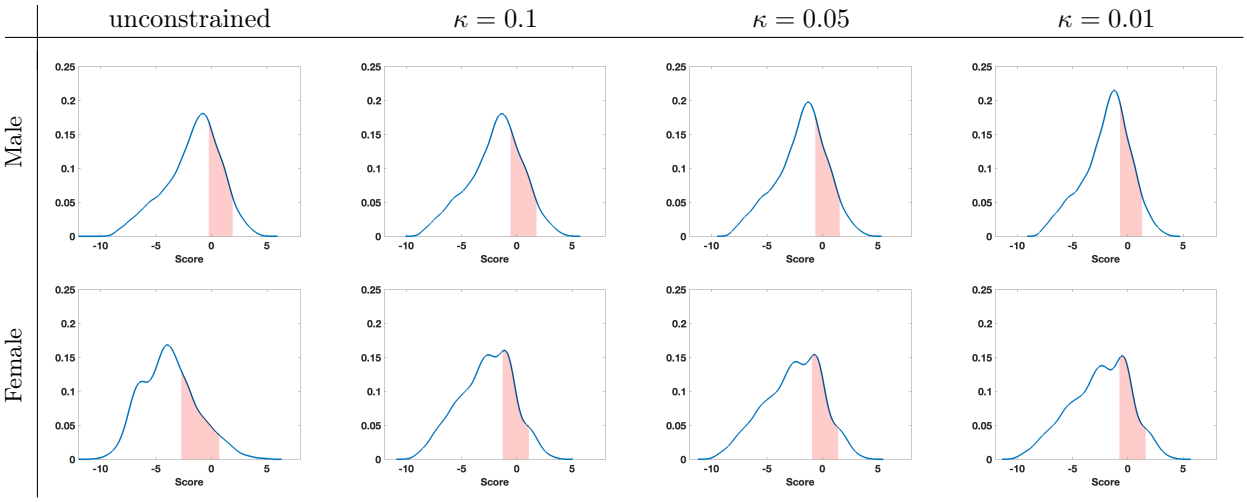

Figure 3: Distributions of predicted scores of different sensitive groups on *a9a* dataset by the unconstrained model and the models solved from (9) (pSP constraints) with different κ's. The interval $\mathcal{I}$ is $[5\%, 30\%]$ and is highlighted in red.

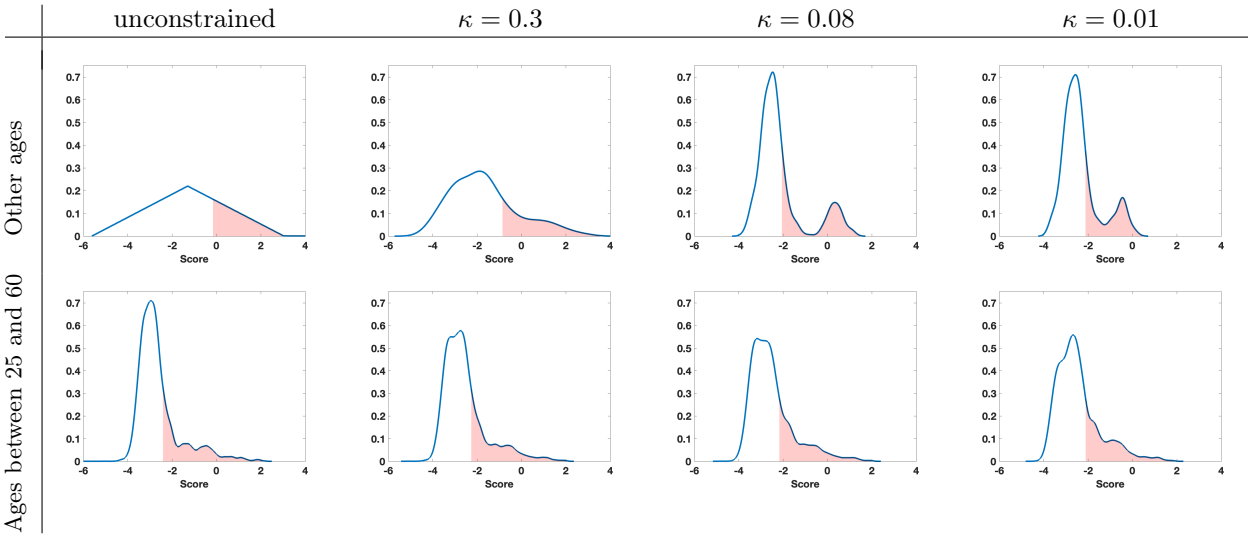

Figure 4: Distributions of predicted scores of different sensitive groups on *bank* dataset by the unconstrained model and the models solved from (9) (pSP constraints) with different κ's. The interval $\mathcal{I}$ is $[0\%, 25\%]$ and is highlighted in red.

- Choices of $\mathcal{I}$'s:

  a9a: $\mathcal{I} \in \{[5\%, 10\%], \mathbf{[5\%,\ 30\%]}, [5\%, 50\%], [5\%, 70\%], [5\%, 90\%]\}$.

  bank: $\mathcal{I} \in \{[0\%, 5\%], \mathbf{[0\%,\ 25\%]}, [0\%, 45\%], [0\%, 65\%], [0\%, 85\%]\}$.

  law school: $\mathcal{I} \in \{[10\%, 100\%], [30\%, 100\%], [50\%, 100\%], \mathbf{[70\%,\ 100\%]}, [90\%, 100\%]\}$.

  where the bold intervals are ones we used for Figure 1.

- For problem (9) (pSP fairness):

  a9a: $\kappa = 0.05$; *bank* : $\kappa = 0.01$; *law school* : $\kappa = 0.005$.

- For problem (71) (pSP fairness):

  a9a: $\kappa = 1.5$; *bank* : $\kappa = 0.7$; *law school* : $\kappa = 2.7$.

- For problem (74) (pSP fairness):

  a9a: $\kappa = 4.6$; *bank* : $\kappa = 3$; *law school* : $\kappa = 4$.

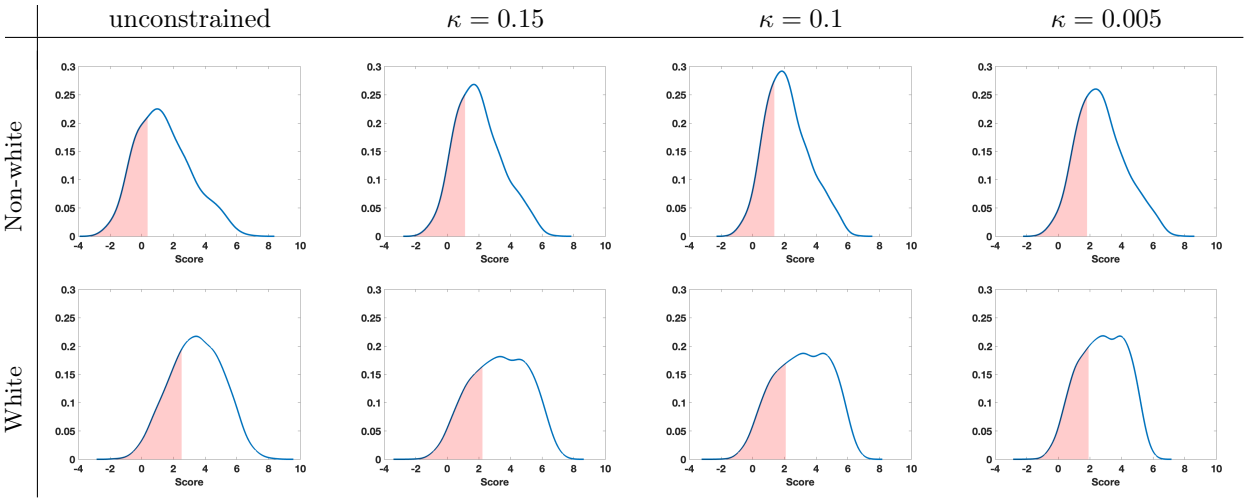

Figure 5: Distributions of predicted scores of different sensitive groups on *law school* dataset by the unconstrained model and the models solved from (9) (pSP constraints) with different κ's. The interval $\mathcal{I}$ is $[70\%, 100\%]$ and is highlighted in red.

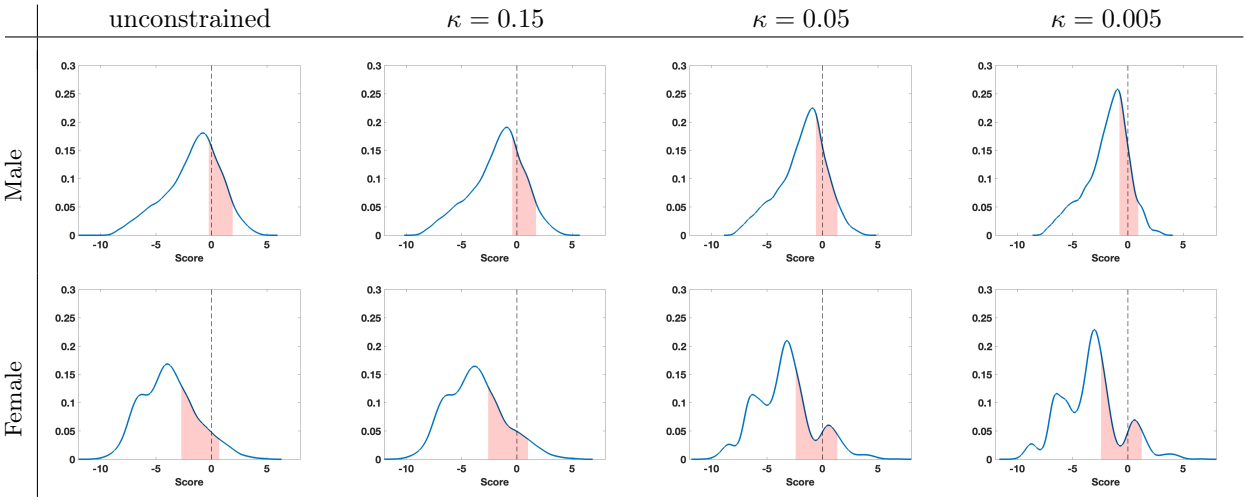

Figure 6: Distributions of predicted scores of different sensitive groups on *a9a* dataset by the unconstrained model and the models solved from (11) (pDP constraints) with different κ's. The interval $\mathcal{I}$ is $[5\%, 30\%]$ and is highlighted in red.

- For problem (76) (pSP fairness):

  a9a: $\kappa = 4.2$; *bank* : $\kappa = 2$; *law school* : $\kappa = 0.9$.

- For problem (83) (pSP fairness):

  a9a: $\lambda = 0.37$; *bank* : $\lambda = 0.3$; *law school* : $\lambda = 0.8$.

G.7 Optimization Performance of IDCA

In this section, we compare the IDCA with the switching subgradient (SSG) method directly applied to (12) (instead of being used as a subroutine), which includes (9) and (11) as special cases. The problem instances we use here are the same as those in Section 7. Note that the complexity of SSG has only been analyzed for convex problems (Nesterov, 2018, (3.2.24)) and weakly convex problems (Huang and Lin, 2023), while its complexity for solving DC optimization is still unknown.

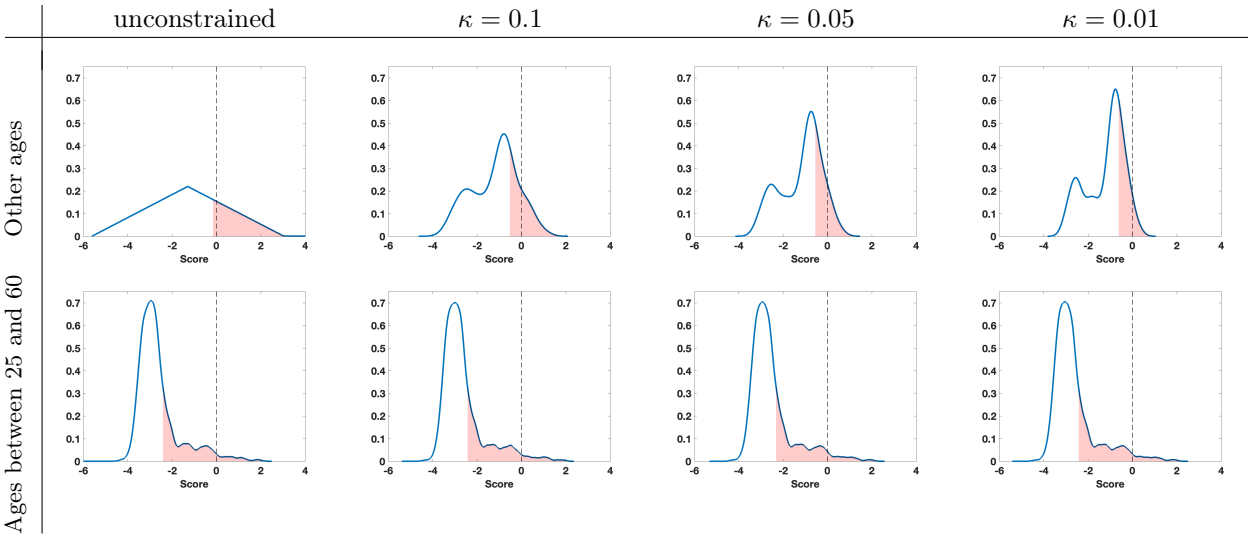

Figure 7: Distributions of predicted scores of different sensitive groups on *bank* dataset by the unconstrained model and the models solved from (11) (pDP constraints) with different κ's. The interval $\mathcal{I}$ is top $[0\%, 25\%]$ and is highlighted in red.

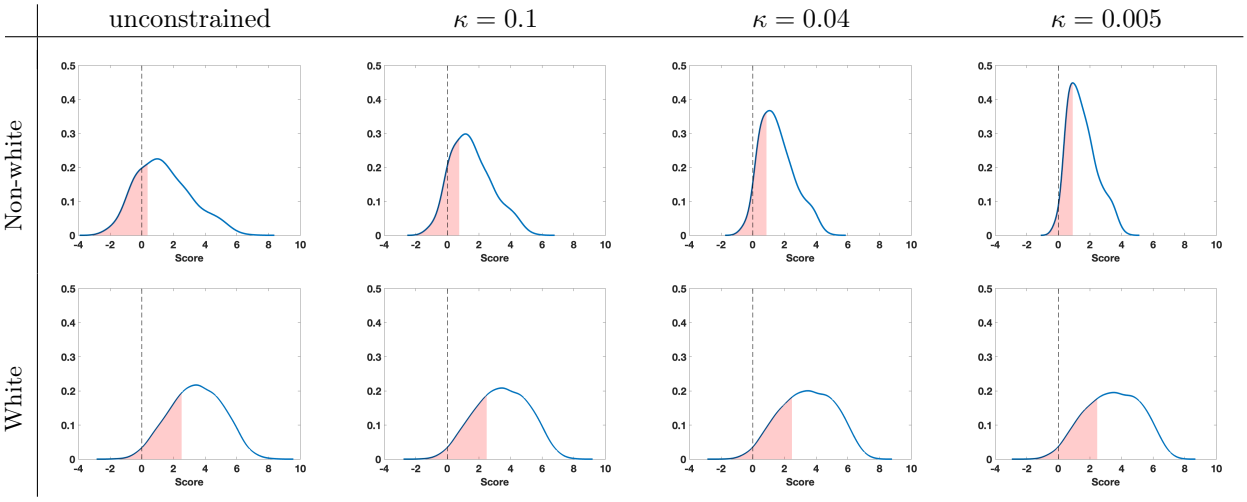

Figure 8: Distributions of predicted scores of different sensitive groups on *law school* dataset by the unconstrained model and the models solved from (11) (pDP constraints) with different κ's. The interval $\mathcal{I}$ is $[70\%, 100\%]$ and is highlighted in red.

We first present the SSG method directly applied to (12) in Algorithm 3, where we denote

$$f(\cdot) := \max_{i \in [m]} f_i(\cdot).$$

The tuning parameter in Algorithm 3 is ϵ_t. Two different ways of setting ϵ_t are considered and the better one is applied in the comparison with our IDCA method. One way is to set a static ϵ_t, where $\epsilon_t = \epsilon$ for some $\epsilon > 0$ for any t. The other way is to set a diminishing ϵ_t, where $\epsilon_t = \frac{c}{t+1}$ for some $c > 0$. We select ϵ in the static setup from $\{10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$ and select c in the dynamic setup from $\{0.2, 0.5, 1, 2\}$. For each of the candidate values of ϵ_t set in both the dynamic and static ways, we run Algorithm 3 for $T = 10000$ iterations and choose ϵ_t that produces the smallest objective value at termination. Then we run Algorithm 3 again for $T = 50000$ iterations using the best ϵ_t.

We compare the objective value $f_0(\mathbf{w})$ and the infeasibility $f(\mathbf{w})$ of the solutions generated by IDCA and SSG as the total number of iterations increases. Their performance for solving (9) and (11) are presented in Figures 10 and 11, respectively. In both figures, the x-axis represents the total number of inner iterations for IDCA and

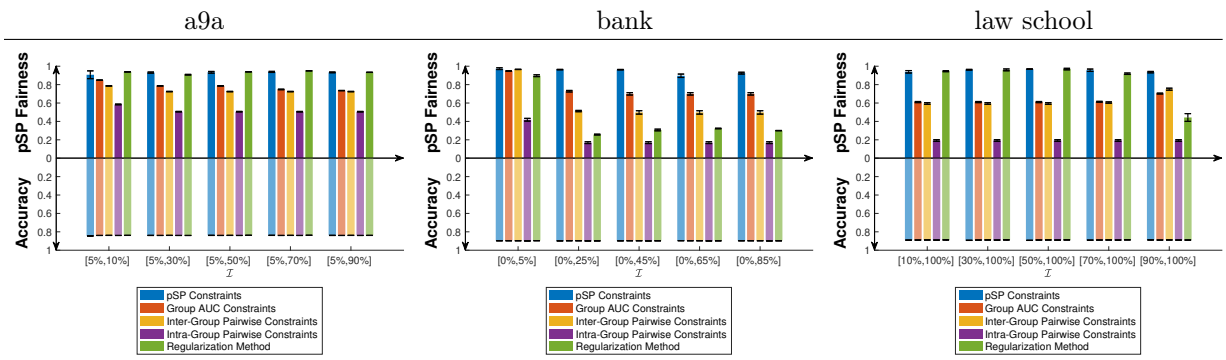

Figure 9: Comparison of different in-processing methods in pSP fairness they guarantee when they achieve the same classification accuracy. Results are presented with different intervals $\mathcal{I}$ on *a9a*, *bank*, and *law school* datasets.

Algorithm 3 Switching Subgradient (SSG) Method for (12) (including (9) and (11) as special cases)

Input: Infeasibility tolerance $\epsilon > 0$, initial solution $\mathbf{w}^{(0)} \in \mathcal{W}$ with $f_i(\mathbf{w}^{(0)}) \leq \epsilon$ for $i \in [m]$ and the number of iterations T.
$\mathcal{T} \leftarrow \emptyset$.
for $t = 0$ **to** $T - 1$ **do**
  if $f(\mathbf{w}^{(t)}) \leq \epsilon_t$ **then**
    Compute $\mathbf{f}_0^{(t)} \in \partial f_0(\mathbf{w}^{(t)})$.
    $\mathbf{w}^{(t+1)} \leftarrow \text{Proj}_{\mathcal{W}}(\mathbf{w}^{(t)} - \frac{\epsilon_t}{\|\mathbf{f}_0^{(t)}\|^2}\mathbf{f}_0^{(t)})$
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{t\}$.
  else
    Compute $\mathbf{f}^{(t)} \in \partial f(\mathbf{w}^{(t)})$.
    $\mathbf{w}^{(t+1)} \leftarrow \text{Proj}_{\mathcal{W}}(\mathbf{w}^{(t)} - \frac{f(\mathbf{w}^{(t)})}{\|\mathbf{f}^{(t)}\|^2}\mathbf{f}^{(t)})$
  end if
end for
Output: $\mathbf{w}^{(\tau)}$ with $\tau = \arg\min_{t \in \mathcal{T}} f_0(\mathbf{w}^{(t)})$

represents the total number of iterations for SSG. The y-axis represents the objective value and the infeasibility achieved by both methods in the first and second rows, respectively.

According to Figure 10, when applied to (9), our IDCA and the SSG method have similar performances on the *a9a* and *bank* datasets. However, IDCA is able to achieve a significantly smaller objective value and smaller infeasibility than SSG on the *law school* dataset. According to Figure 11, when applied to (11), IDCA and the SSG method have similar performances on the *a9a* and *law school* datasets. However, the SSG method has slightly better performance than IDCA on the *bank* dataset. We want to emphasize that the SSG method is only a heuristic method for (9) and (11) (and for (12) in general) as it fails to have any theoretical guarantees. The consequence of the lack of theoretical guarantees is that it may not even return a feasible solution, as shown on the *law school* dataset in Figure 10. In this case, the iterations of the SSG method likely get trapped in an infeasible stationary solution of the constraint functions. On the contrary, we have provided the complexity analysis for IDCA in Theorem 6.5, and IDCA always produces a nearly KKT, and thus, a feasible solution.

G.8 Numerical Results on the ACS Dataset

We also conduct the experiment on the ACS dataset under the same settings and procedures with the same benckmarks as in Section 7.1. The results are shown in Table 2. It can be seen that, on the ACS dataset, our method also attains higher accuracy than every baseline models across all levels of pSP fairness, demonstrating its effectiveness in trading accuracy for partial fairness.

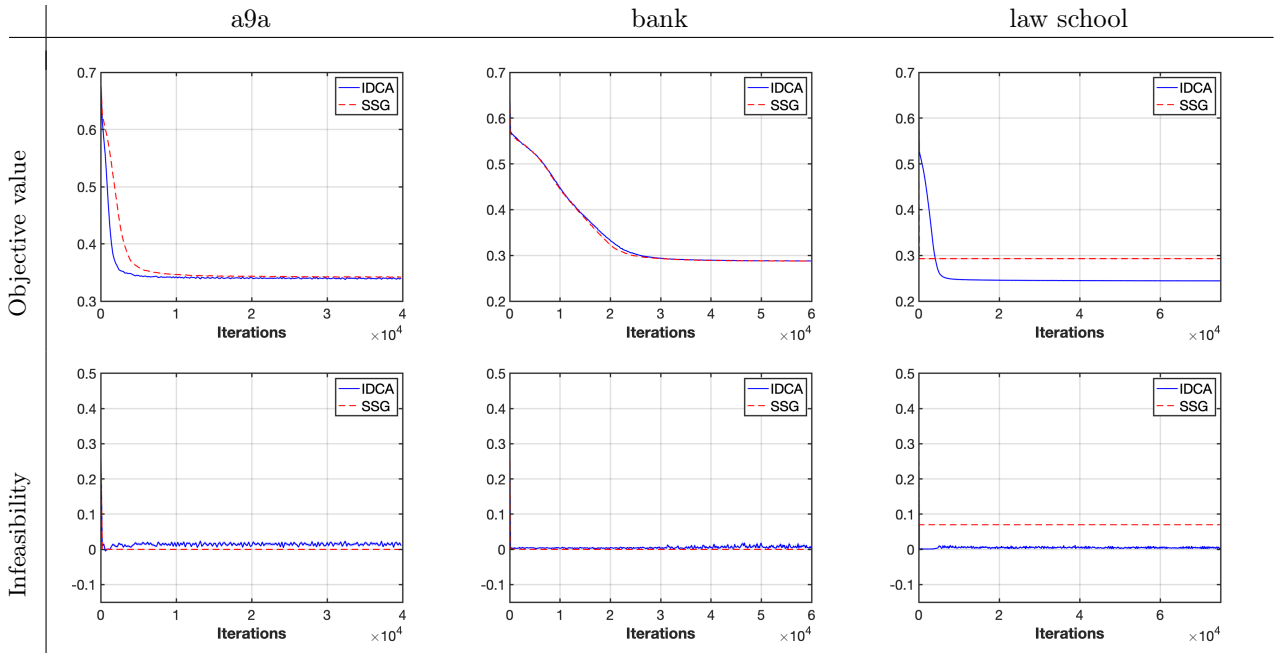

Figure 10: Performances of IDCA and SSG on (9). The interval $\mathcal{I}$ is $[5\%, 30\%]$ for *a9a*, $[0\%, 25\%]$ for *bank*, and $[70\%, 100\%]$ for *law school*, respectively.

Table 2: Performances of the models obtained using different fairness constraints on the ACS dataset. $\mathcal{I}$ is set to $[20\%, 40\%]$

pSP		Group AUC		Inter-group Pairwise		Intra-group Pairwise	
pSP fairness	accuracy	pSP fairness	accuracy	pSP fairness	accuracy	pSP fairness	accuracy
0.7748	0.7986	0.7748	0.7986	0.7748	0.7986	0.7748	0.7986
0.8660	0.8100	0.8078	0.7931	0.8163	0.7931	0.7398	0.7920
0.9090	0.8068	0.8414	0.7927	0.8410	0.7922	0.7026	0.7902
0.9357	0.8043	0.9188	0.7866	0.8821	0.7466	0.6624	0.7848
0.9669	0.8001	0.9362	0.6667	0.8685	0.6322	0.5772	0.7648

G.9 Numerical Results on Deep Learning

The CelebA dataset (Large-scale CelebFaces Attributes Dataset) (Liu et al., 2015) contains 202,599 face images (split into 60% training, 20% validation, and 20% testing sets) with each image annotated by 40 binary attributes (Liu et al., 2015). We consider an image classification problem with a partial statistical parity constraint, formulated as (9), on the CelebA dataset. We instantiate $h_{\mathbf{w}}(\boldsymbol{\xi})$ as a ResNet-18 model (a CNN pretrained on ImageNet) and fine-tune it to predict the "Attractive" attribute, with gender serving as the sensitive attribute in the fairness constraint.

Since SSG (Algorithm 2) is a deterministic method for solving (16), whereas deep learning training relies on stochastic gradients, we instead implement IDCA (Algorithm 1) using the primal–dual stochastic gradient method of Yu et al. (2017) with a mini-batch size of 64 to solve (16), ensuring (18) with high probability.

We evaluate the testing performance of the resulting models under three settings: (i) $\mathcal{I} = [20\%, 40\%]$ with $|\hat{\mathcal{I}}| = 10$, (ii) $\mathcal{I} = [0\%, 100\%]$ (i.e., full statistical parity constraint) with $|\hat{\mathcal{I}}| = 50$, and (iii) $\mathcal{I} = [20\%, 40\%]$ with $|\hat{\mathcal{I}}| = 50$. By varying κ within each setting, we obtain different trade-offs between classification accuracy and partial statistical parity, as shown in Figure 12.

The results indicate that increasing $|\hat{\mathcal{I}}|$ from 10 to 50 has little impact on performance under partial statistical parity constraints. Moreover, for a fixed level of partial statistical parity, enforcing partial constraints yields higher classification accuracy than imposing full statistical parity.

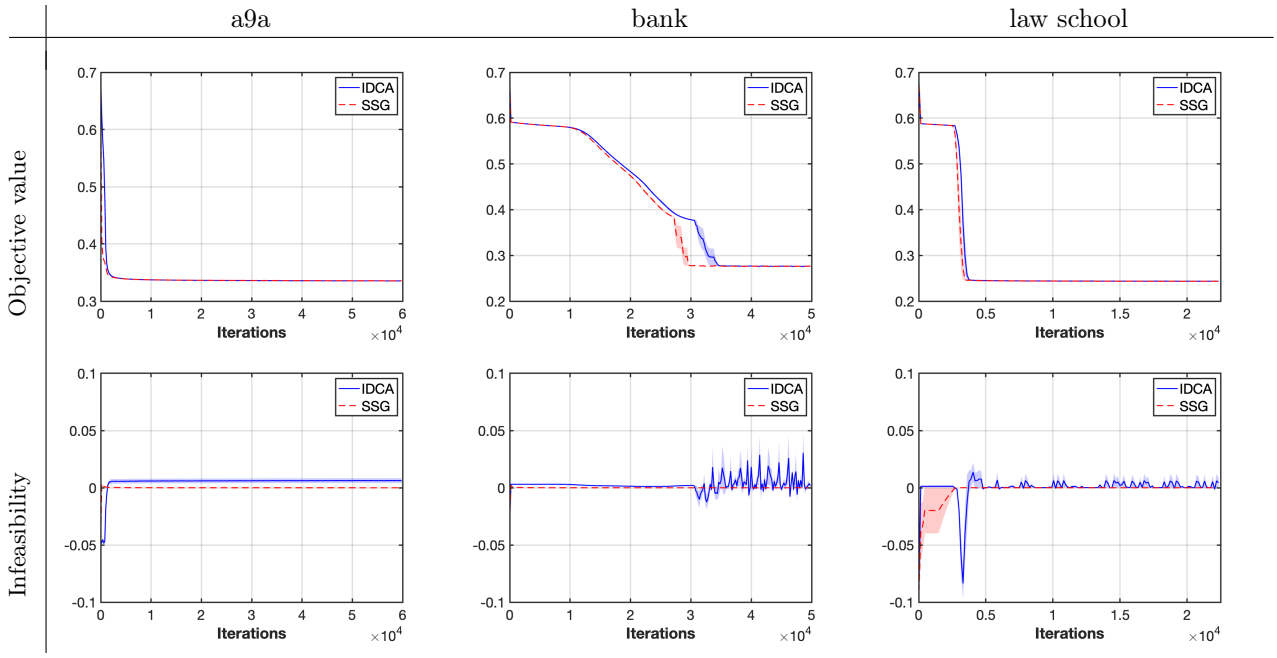

Figure 11: Performances of IDCA and SSG on (11). The interval $\mathcal{I}$ is $[5\%, 30\%]$ for *a9a*, $[0\%, 25\%]$ for *bank*, and $[70\%, 100\%]$ for *law school*, respectively.

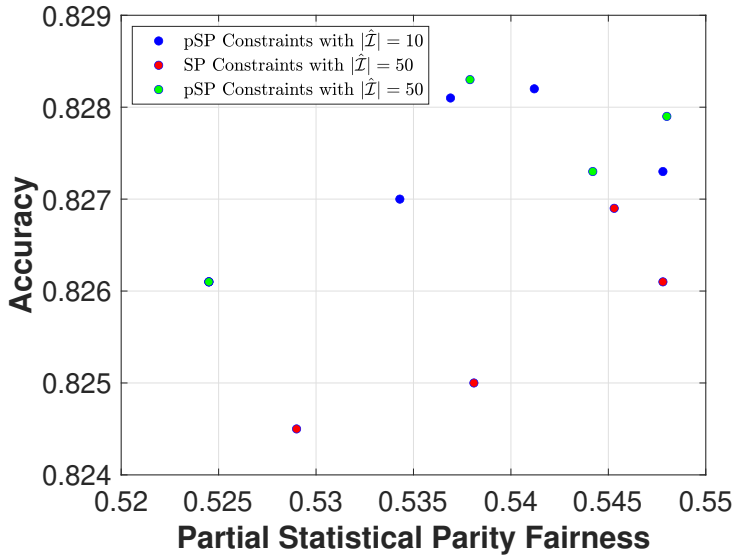

Figure 12: Testing performances of the deep learning models solved from (9) under different settings on CelebA dataset.

G.10 Comparison Between Partial Statistical Parity and Full Statistical Parity

In this section, we compare the performance of models obtained from (9) under full statistical parity ($\mathcal{I} = [0\%, 100\%]$) and partial statistical parity ($\mathcal{I} \neq [0\%, 100\%]$). For the partial statistical parity setting, $\mathcal{I}$ and $\hat{\mathcal{I}}$ are chosen as in Section 7.1 for each dataset (see Figure 1). For the full statistical parity setting, we set $\mathcal{I} = [0\%, 100\%]$ and construct $\hat{\mathcal{I}}$ as a discrete subset with the same spacing as in the corresponding partial-fairness case. Note that when $\mathcal{I} = [0\%, 100\%]$, problem (9) remains a DC program and can therefore be solved using the same IDCA method.

By varying κ in (9), the results for the three datasets are reported in Table 3. Overall, our partial-fairness

approach achieves a superior accuracy–fairness trade-off on the *bank* and *law school* datasets, and a comparable trade-off on the *a9a* dataset.

Table 3: Comparison of the performances of the models solved from (9) under full statistical parity constraint ($\mathcal{I} = [0\%, 100\%]$) and partial statistical parity constraint ($\mathcal{I} \neq [0\%, 100\%]$).

a9a			
$\mathcal{I} = [0\%, 100\%]$		$\mathcal{I} = [5\%, 30\%]$	
pSP fairness	accuracy	pSP fairness	accuracy
0.5970	0.8499	0.5970	0.8499
0.6937	0.8501	0.7017	0.8492
0.7799	0.8485	0.7841	0.8479
0.8873	0.8440	0.8732	0.8445
0.9326	0.8384	0.9310	0.8393
0.9742	0.8322	0.9752	0.8311
bank			
$\mathcal{I} = [0\%, 100\%]$		$\mathcal{I} = [0\%, 25\%]$	
pSP fairness	accuracy	pSP fairness	accuracy
0.4366	0.9015	0.4366	0.9015
0.6451	0.8957	0.6582	0.9027
0.7999	0.8954	0.7376	0.9028
0.8456	0.8959	0.8954	0.9025
0.9356	0.8941	0.9634	0.8983
law school			
$\mathcal{I} = [0\%, 100\%]$		$\mathcal{I} = [70\%, 100\%]$	
pSP fairness	accuracy	pSP fairness	accuracy
0.5397	0.9004	0.5397	0.9004
0.5833	0.8975	0.6038	0.8996
0.6976	0.8903	0.6956	0.8937
0.7855	0.8892	0.7825	0.8913
0.8736	0.8896	0.8479	0.8927
0.9615	0.8881	0.9563	0.8909

G.11 Performances of Models from (9) with Different $|\hat{\mathcal{I}}|$

Theoretically, increasing $|\hat{\mathcal{I}}|$ raises computational cost but can improve the out-of-sample performance of models obtained from (9). To examine this effect, we compare models solved from (9) with $|\hat{\mathcal{I}}| = 10$ and $|\hat{\mathcal{I}}| = 20$ across the three datasets. In particular, we choose $\mathcal{I}$ as in Section 7.1 for each dataset (see Figure 1) and then set $\widehat{\mathcal{I}}$ to be a set of equally spaced values in $\mathcal{I}$. The results, reported in Table 4, show no significant improvement from increasing $|\hat{\mathcal{I}}|$, except on the *law school* dataset.

Table 4: Comparison of the performances of the models solved from (9) with $|\hat{\mathcal{I}}| = 10$ and $|\hat{\mathcal{I}}| = 20$

	$\|\hat{\mathcal{I}}\| = 20$		$\|\hat{\mathcal{I}}\| = 10$	
κ	pSP fairness	accuracy	pSP fairness	accuracy
a9a ($\mathcal{I} = [5\%, 30\%]$)				
Unconstrained	0.5970	0.8499	0.5970	0.8499
0.28	0.6964	0.8491	0.7017	0.8492
0.2	0.7819	0.8481	0.7841	0.8479
0.1	0.8751	0.8446	0.8732	0.8445
0.05	0.9328	0.8387	0.9310	0.8393
0.01	0.9703	0.8325	0.9752	0.8311
bank ($\mathcal{I} = [0\%, 25\%]$)				
Unconstrained	0.4366	0.9015	0.4366	0.9015
0.3	0.6620	0.9026	0.6582	0.9027
0.22	0.7211	0.9030	0.7376	0.9028
0.08	0.8683	0.9025	0.8954	0.9025
0.01	0.9483	0.8985	0.9634	0.8983
law school ($\mathcal{I} = [70\%, 100\%]$)				
Unconstrained	0.5397	0.9004	0.5397	0.9004
0.2	0.6360	0.8980	0.6038	0.8996
0.15	0.7165	0.8929	0.6956	0.8937
0.1	0.7841	0.8922	0.7825	0.8913
0.08	0.8683	0.8932	0.8479	0.8927
0.005	0.9625	0.8910	0.9563	0.8909