Stochastic Methods for AUC Optimization subject to AUC-based Fairness Constraints

Yao Yao
University of Iowa

Qihang Lin
University of Iowa

Tianbao Yang
Texas A&M University

Abstract

As machine learning being used increasingly in making high-stakes decisions, an arising challenge is to avoid unfair AI systems that lead to discriminatory decisions for protected population. A direct approach for obtaining a fair predictive model is to train the model through optimizing its prediction performance subject to fairness constraints. Among various fairness constraints, the ones based on the area under the ROC curve (AUC) are emerging recently because they are threshold-agnostic and effective for unbalanced data. In this work, we formulate the problem of training a fairness-aware predictive model as an AUC optimization problem subject to a class of AUC-based fairness constraints. This problem can be reformulated as a min-max optimization problem with min-max constraints, which we solve by stochastic first-order methods based on a new Bregman divergence designed for the special structure of the problem. We numerically demonstrate the effectiveness of our approach on real-world data under different fairness metrics.

1 INTRODUCTION

AI systems have been increasingly used to assist in making high-stakes decisions such as lending decision (Addo et al., 2018), bail and parole decision (Dressel and Farid, 2018), resource allocation (Davahli et al., 2021) and so on. Along with this trend, a question arising is how to ensure AI systems are fair and do not produce discriminatory decisions for protected groups defined by some sensitive variables (e.g., age, race, gender, etc.). To answer this question, the first step is to define and quantitatively measure fairness of AI systems, which is itself an active research area.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

For a classification task, a variety of fairness metrics have been studied including demographic parity (Beutel et al., 2019b; Calders et al., 2009; Gajane and Pechenizkiy, 2017), equality of opportunity (Hardt et al., 2016), equality of odds (Hardt et al., 2016), predictive quality parity (Chouldechova, 2017) and counter factual fairness (Kusner et al., 2017). All of these fairness metrics are formulated based on statistical relationships between predicted class labels and sensitive variables. However, many predictive models only generate a predicted risk score and a predicted class label is obtained afterwards by comparing the score with a threshold. A good threshold is not always easy to choose in practice and may vary with datasets and applications. In fact, it is likely that a model satisfies a fairness criterion with one threshold but violates the same fairness criterion with another threshold. Moreover, the threshold is often chosen to achieve a targeted predicted positive rate. When doing so, it is not easy to ensure a targeted fairness criterion is satisfied at the same threshold.

With these drawbacks in the threshold-dependent fairness metrics, there have been growing efforts on developing threshold independent fairness metrics, among which the fairness metrics based on AUC, or equivalently, pairwise comparison, are prevalent (Beutel et al., 2019a; Borkan et al., 2019; Dixon et al., 2018; Kallus and Zhou, 2019; Narasimhan et al., 2020; Vogel et al., 2021; Yang et al., 2022c). These metrics are directly defined based on statistical relationships between predicted risk scores and sensitive variables and thus do not require a predetermined threshold.

Regardless of the fairness metric applied, training a fair predictive model requires balancing the model's prediction performance and fairness, two potentially conflicting targets. Hence, it is naturally to formulate this problem as constrained optimization where the model's prediction performance is optimized subject to some fairness constraints. This approach has been studied with constraints based on threshold-dependent fairness metrics (Agarwal et al., 2018; Cotter et al., 2018, 2019; Cruz et al., 2022; Diana et al., 2021; Dwork et al., 2012; Goh et al., 2016; Kearns et al., 2018; Woodworth et al., 2017) and threshold-agnostic fairness metrics (Narasimhan et al., 2020; Vogel et al., 2021; Zafar et al., 2017) with different optimization algorithms

applied during training. In Narasimhan et al. (2020), a proxy-Lagrangian method from Cotter et al. (2018, 2019) is applied to optimization with fairness constraints while regularization methods are applied by Beutel et al. (2017); Vogel et al. (2021) to optimize a weighted sum of prediction performance and fairness metrics.

Online learning is a common setting in machine learning where data becomes available sequentially and the model needs to be updated by the latest data. When learning a fair model online, the methods in Narasimhan et al. (2020); Vogel et al. (2021) need to compute stochastic gradients of the constraint functions. However, due to the pairwise comparison involved in their optimization models, computing one online stochastic gradient requires processing a pair of data points, one from the protected group and the other from the unprotected group. This requires that data points always arrive in pairs, which is not always guaranteed in practice. For the similar reason, when training models off-line using an existing dataset, the methods by Narasimhan et al. (2020); Vogel et al. (2021) require processing all pairs of data points and thus need a computational cost quadratic in data size, which is prohibited for large datasets.

In this paper, we focus on developing efficient numerical methods for training a classification model under AUC-based threshold-agnostic fairness constraints by addressing the computational issues mentioned above. The main contribution of this paper is formulating the aforementioned problem into a stochastic optimization problem subject to min-max constraints. Although the min-max constraints are new and challenging structures, we propose a special Bregman divergence after changing variables such that the problem can be solved efficiently by the existing stochastic first-order methods for constrained stochastic optimization such as Boob et al. (2022); Lin et al. (2020); Ma et al. (2020). Compared to Narasimhan et al. (2020); Vogel et al. (2021), the main advantage of our approach is that it supports model training in an online setting with one data point, instead of a data pair, arriving each time in any sequence. Moreover, when applied under the off-line setting, our approach only has a computational cost linear in data size. One limitation of our approach is that we must use a quadratic surrogate loss to approximate the AUCs in the objective and constraint functions. However, the numerical results on real-world datasets show that the models found by our methods trade off classification performance and fairness more effectively than existing techniques.

2 RELATED WORKS

A survey of prevalent fairness metrics, including some discussed in the previous section, is provided by Verma and Rubin (2018). However, most metrics discussed in Verma and Rubin (2018) are based on predicted class labels and thus threshold dependent. The threshold-agnostic fairness metrics based on AUC (see examples in Section 3) have been proposed in Borkan et al. (2019); Dixon et al. (2018); Kallus and Zhou (2019); Vogel et al. (2021). They have been extended to a broader class of metrics based on pairwise comparison, so the target variable can be continuous or ordinal (e.g., in a regression or ranking problem) (Beutel et al., 2019a; Narasimhan et al., 2020). The class of fairness metrics we consider in this paper is more general than Borkan et al. (2019); Dixon et al. (2018); Kallus and Zhou (2019) and has similar generality to Beutel et al. (2019a); Narasimhan et al. (2020). A ROC-based fairness metric is proposed by Vogel et al. (2021) which is threshold-agnostic and stronger than the AUC-based ones in this paper. However, their optimization algorithms do not have theoretical convergence guarantees and require processing data points in pairs per iteration, which leads to a quadratic computational cost and is not ideal for training online.

The three main approaches for building a fairness-aware machine learning model include the pre-processing, post-processing, and in-processing methods. The pre-processing method reduce machine bias by re-sampling and balancing training data (Dwork et al., 2012). The post-processing method adjusts the prediction results after to ensure fairness (Hardt et al., 2016). The methods in this paper are the in-processing methods, which enforce fairness of a model during training by adding constraints or regularization to the optimization problem (Agarwal et al., 2018; Goh et al., 2016; Yang et al., 2022c).

Most in-processing methods are based on threshold-dependent fairness metrics (Agarwal et al., 2018; Cotter et al., 2018, 2019; Cruz et al., 2022; Diana et al., 2021; Dwork et al., 2012; Goh et al., 2016; Kearns et al., 2018; Woodworth et al., 2017) while this work considers threshold-independent metrics. The unconstrained optimization approach by Yang et al. (2022c) minimizes the maximum of four different AUC scores to achieve a balance between classification performance and fairness, while we ensure fairness by constraints. Although a constrained optimization approach is also presented in the appendix of Yang et al. (2022c), no convergence analysis is provided. Fairness constrained optimization is an important application of stochastic constrained optimization for which many effective algorithms have been developed under the convex setting (Boob et al., 2022; Lin et al., 2020; Yan and Xu, 2022; Yang et al., 2022a) and the non-convex setting (Boob et al., 2022; Ma et al., 2020). A proxy Lagrangian method has been developed for optimization subject to a class of rate constraints (Cotter et al., 2018, 2019; Narasimhan et al., 2020), which include almost all fairness constraints we discussed above. The theoretical complexity of the proxy Lagrangian method has been analyzed (Cotter et al., 2019) when the objective function is convex or non-convex (Cotter et al., 2019) although a strong Bayesian optimization oracle is assumed in the non-convex case. Unconstrained

optimization has also been considered for building fairness-aware models where fairness is enforced through a penalty term (Beutel et al., 2017, 2019a; Vogel et al., 2021).

When directly applied to the AUC-based fairness constraints, the optimization algorithms mentioned above all need to request a pair of data points, one from the protected group and one from the opposite group, to construct the stochastic gradients. This is not ideal for online learning because data may not always arrive in pairs. On the contrary, our method is developed by first reformulating the AUC-based fairness constraints into min-max constraints using a quadratic loss (Ying et al., 2016). The stochastic gradient of this formulation can be computed using only one data point each time with any order of arrivals. Min-max stochastic constraints are new in optimization literature, so we develop a new Bregmen divergence by changing variables so that the existing algorithms like Boob et al. (2022); Lin et al. (2020) and their convergence analysis can be applied. Yang et al. (2022b) develop an algorithm for stochastic compositional optimization subject to compositional constraints which can be applied to our problem with the same computational complexity. This is because our min-max constraints can be also viewed as compositional constraints. They focus on the convex case but we also consider the non-convex case under some additional assumption (see Assumption 4).

3 PRELIMINARIES

Consider a binary classification problem, where the goal is to build a model that predicts a binary label $\zeta \in \{1, -1\}$ based on a feature vector $\boldsymbol{\xi} \in \mathbb{R}^p$. The sensitive feature of a data point is denoted by $\gamma \in \{1, -1\}$, which may or may not be a coordinate of $\boldsymbol{\xi}$. This feature divides the data into a protected group ($\gamma = 1$) and an unprotected group ($\gamma = -1$). We denote a data point by a triplet $\mathbf{z} = (\boldsymbol{\xi}, \zeta, \gamma) \in \mathbb{R}^{p+2}$ which is a random vector. We say $\mathcal{G} \subset \mathbb{R}^{p+2}$ has a *positive measure w.r.t.* $\mathbf{z}$ if $\Pr(\mathbf{z} \in \mathcal{G}) > 0$. Let $h_{\mathbf{w}} : \mathbb{R}^p \to \mathbb{R}$ be the predictive model that produces a score $h_{\mathbf{w}}(\boldsymbol{\xi})$ for $\boldsymbol{\xi}$. Function $h_{\mathbf{w}}$ is parameterized by a vector $\mathbf{w}$ from a convex compact set $\mathcal{W} \subset \mathbb{R}^d$. We assume $h_{\mathbf{w}}(\cdot)$ is differentiable and consider threshold-agnostic fairness metrics defined based on the join distribution of $h_{\mathbf{w}}(\boldsymbol{\xi})$, ζ and γ.

Definition 1 (AUC defined by subsets) *Let $\mathbf{z} = (\boldsymbol{\xi}, \zeta, \gamma)$ and $\mathbf{z}' = (\boldsymbol{\xi}', \zeta', \gamma')$ be i.i.d random data points. Given two sets $\mathcal{G}$ and $\mathcal{G}'$ in $\mathbb{R}^{p+2}$ with positive measures w.r.t. $\mathbf{z}$, the AUC w.r.t. $\mathcal{G}$ and $\mathcal{G}'$ is*

$$AUC_{\mathbf{w}}(\mathcal{G}, \mathcal{G}') := Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > h_{\mathbf{w}}(\boldsymbol{\xi}') | \mathbf{z} \in \mathcal{G}, \mathbf{z}' \in \mathcal{G}').$$

When $\mathcal{G} = \mathcal{D}_+ := \{\mathbf{z} | \zeta = 1\}$ and $\mathcal{G}' = \mathcal{D}_- := \{\mathbf{z} | \zeta = -1\}$, $\text{AUC}_{\mathbf{w}}(\mathcal{G}, \mathcal{G}')$ is reduced to the standard AUC for a binary classification problem.

Definition 2 (AUC-based fairness metric) *Given sets $\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{G}_1'$ and $\mathcal{G}_2'$ in $\mathbb{R}^{p+2}$ with positive measures w.r.t. $\mathbf{z}$, the AUC-based fairness metric w.r.t. $\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{G}_1'$ and $\mathcal{G}_2'$ is*

$$|AUC_{\mathbf{w}}(\mathcal{G}_1, \mathcal{G}_1') - AUC_{\mathbf{w}}(\mathcal{G}_2, \mathcal{G}_2')| \in [0, 1], \tag{1}$$

where $AUC_{\mathbf{w}}(\cdot, \cdot)$ follows Definition 1.

We say model $h_{\mathbf{w}}$ is unfair if the value of (1) is close to one and is fair if close to zero. This class of fairness metrics contains several existing threshold-agnostic metrics in literature, including the inter-group pairwise fairness (Beutel et al., 2019a; Kallus and Zhou, 2019), the intra-group pairwise fairness (Beutel et al., 2019a), the positive/negative average equality gap (Borkan et al., 2019) and the fairness metric based on background-subgroup AUCs (Borkan et al., 2019). In Appendix A, we discuss how Definition 2 is reduced to these metrics by setting $\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{G}_1'$ and $\mathcal{G}_2'$ to be different sets.

Besides fairness, we are also interested in the performance of the model as a classifier. In this paper, we also use the AUC, namely, $\text{AUC}_{\mathbf{w}}(\mathcal{D}_+, \mathcal{D}_-)$, as the performance metric and optimize it subject to fairness constraints. This choice is made only to obtain a uniform structure in the objective and constraint functions. The numerical methods we presented in this paper can be also applied when the classification performance is optimized by a traditional method, e.g., minimizing the empirical logistic loss.

The general formulation of our problem can be written as

$$\begin{aligned} \max_{\mathbf{w} \in \mathcal{W}} \quad & \text{AUC}_{\mathbf{w}}(\mathcal{D}_+, \mathcal{D}_-), \\ \text{s.t.} \quad & |\text{AUC}_{\mathbf{w}}(\mathcal{G}_1, \mathcal{G}_1') - \text{AUC}_{\mathbf{w}}(\mathcal{G}_2, \mathcal{G}_2')| = 0. \end{aligned}$$

The equality constraint used here may be too restrict because an absolutely fair model may have a poor prediction performance and may be unnecessarily overly fair for users. To provide some flexibility to users, we replace the equality constraint to two inequalities after introducing a targeted *level of fairness*, denoted by $\kappa \geq 0$, on the right-hand sides:

$$\begin{aligned} \max_{\mathbf{w} \in \mathcal{W}} \quad & \text{AUC}_{\mathbf{w}}(\mathcal{D}_+, \mathcal{D}_-), \\ \text{s.t.} \quad & \text{AUC}_{\mathbf{w}}(\mathcal{G}_1, \mathcal{G}_1') - \text{AUC}_{\mathbf{w}}(\mathcal{G}_2, \mathcal{G}_2') \leq \kappa, \\ & \text{AUC}_{\mathbf{w}}(\mathcal{G}_2, \mathcal{G}_2') - \text{AUC}_{\mathbf{w}}(\mathcal{G}_1, \mathcal{G}_1') \leq \kappa. \end{aligned} \tag{2}$$

Solving (2) directly is challenging because the objective and constraint functions involve indicator functions which are discontinuous. A common solution is to introduce a surrogate loss to approximate the indicator function. In particular, focusing on the objective function first, we have

$$\begin{aligned} & \max_{\mathbf{w} \in \mathcal{W}} \text{AUC}_{\mathbf{w}}(\mathcal{D}_+, \mathcal{D}_-) \\ = \quad & \max_{\mathbf{w} \in \mathcal{W}} \Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > h_{\mathbf{w}}(\boldsymbol{\xi}') | \zeta = 1, \zeta' = -1) \\ \Leftrightarrow \quad & \min_{\mathbf{w} \in \mathcal{W}} \Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) \leq h_{\mathbf{w}}(\boldsymbol{\xi}') | \zeta = 1, \zeta' = -1) \\ = \quad & \min_{\mathbf{w} \in \mathcal{W}} \mathbb{E}\big[\mathbb{I}_{(h_{\mathbf{w}}(\boldsymbol{\xi}) - h_{\mathbf{w}}(\boldsymbol{\xi}') \leq 0)} | \zeta = 1, \zeta' = -1\big] \\ \approx \quad & \min_{\mathbf{w}} \mathbb{E}\big[\ell(h_{\mathbf{w}}(\boldsymbol{\xi}) - h_{\mathbf{w}}(\boldsymbol{\xi}')) | \zeta = 1, \zeta' = -1\big], \end{aligned} \tag{3}$$

where $\ell(\cdot)$ is a continuous surrogate loss function that approximates the indicator functions $\mathbb{I}_{(\cdot\leq 0)}$ and $\mathbb{I}_{(\cdot<0)}$. Similar to (3), we approximate the left-hand side of the first constraint in (2) as follows

$$\begin{aligned}
&\text{AUC}_{\mathbf{w}}(\mathcal{G}_1, \mathcal{G}_1') - \text{AUC}_{\mathbf{w}}(\mathcal{G}_2, \mathcal{G}_2') \\
=&\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > h_{\mathbf{w}}(\boldsymbol{\xi}')|\mathbf{z} \in \mathcal{G}_1, \mathbf{z}' \in \mathcal{G}_1') \\
&- \Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > h_{\mathbf{w}}(\boldsymbol{\xi}')|\mathbf{z} \in \mathcal{G}_2, \mathbf{z}' \in \mathcal{G}_2') \\
=&\Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > h_{\mathbf{w}}(\boldsymbol{\xi}')|\mathbf{z} \in \mathcal{G}_1, \mathbf{z}' \in \mathcal{G}_1') \\
&+ \Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) \leq h_{\mathbf{w}}(\boldsymbol{\xi}')|\mathbf{z} \in \mathcal{G}_2, \mathbf{z}' \in \mathcal{G}_2') - 1 \\
\approx&\mathbb{E}\big[\ell(h_{\mathbf{w}}(\boldsymbol{\xi}') - h_{\mathbf{w}}(\boldsymbol{\xi}))|\mathbf{z} \in \mathcal{G}_1, \mathbf{z}' \in \mathcal{G}_1'\big] \\
&+ \mathbb{E}\big[\ell(h_{\mathbf{w}}(\boldsymbol{\xi}) - h_{\mathbf{w}}(\boldsymbol{\xi}'))|\mathbf{z} \in \mathcal{G}_2, \mathbf{z}' \in \mathcal{G}_2'\big] - 1. \quad (4)
\end{aligned}$$

Similarly, we approximate the left-hand side of the second constraint in (2) as

$$\begin{aligned}
&\text{AUC}_{\mathbf{w}}(\mathcal{G}_2, \mathcal{G}_2') - \text{AUC}_{\mathbf{w}}(\mathcal{G}_1, \mathcal{G}_1') \\
\approx&\mathbb{E}\big[\ell(h_{\mathbf{w}}(\boldsymbol{\xi}') - h_{\mathbf{w}}(\boldsymbol{\xi}))|\mathbf{z} \in \mathcal{G}_2, \mathbf{z}' \in \mathcal{G}_2'\big] \\
&+ \mathbb{E}\big[\ell(h_{\mathbf{w}}(\boldsymbol{\xi}) - h_{\mathbf{w}}(\boldsymbol{\xi}'))|\mathbf{z} \in \mathcal{G}_1, \mathbf{z}' \in \mathcal{G}_1'\big] - 1. \quad (5)
\end{aligned}$$

Using (3) as the objective function and (4) and (5) as the left-hand sides of the inequality constraints. We obtain the following approximation to (2).

$$\begin{aligned}
\min_{\mathbf{w}\in\mathcal{W}}\ &\mathbb{E}\big[\ell(h_{\mathbf{w}}(\boldsymbol{\xi}) - h_{\mathbf{w}}(\boldsymbol{\xi}'))|\zeta = 1, \zeta' = -1\big] \quad (6)\\
\text{s.t. }&\mathbb{E}\big[\ell(h_{\mathbf{w}}(\boldsymbol{\xi}') - h_{\mathbf{w}}(\boldsymbol{\xi}))|\mathbf{z} \in \mathcal{G}_1, \mathbf{z}' \in \mathcal{G}_1'\big], \\
&+ \mathbb{E}\big[\ell(h_{\mathbf{w}}(\boldsymbol{\xi}) - h_{\mathbf{w}}(\boldsymbol{\xi}'))|\mathbf{z} \in \mathcal{G}_2, \mathbf{z}' \in \mathcal{G}_2'\big] \leq 1 + \kappa, \\
&\mathbb{E}\big[\ell(h_{\mathbf{w}}(\boldsymbol{\xi}') - h_{\mathbf{w}}(\boldsymbol{\xi}))|\mathbf{z} \in \mathcal{G}_2, \mathbf{z}' \in \mathcal{G}_2'\big] \\
&+ \mathbb{E}\big[\ell(h_{\mathbf{w}}(\boldsymbol{\xi}) - h_{\mathbf{w}}(\boldsymbol{\xi}'))|\mathbf{z} \in \mathcal{G}_1, \mathbf{z} \in \mathcal{G}_1'\big] \leq 1 + \kappa.
\end{aligned}$$

Although (6) have continuous objective and constraint functions, it is still computationally challenging in general because each expectation in (6) is taken over a pair of random data points from two different subsets. When formulated using the empirical distribution over n data points, each expectation becomes double summations which have $O(n^2)$ computational cost. Moreover, (6) is not suitable for online learning as computing its stochastic gradient requires data arriving in pairs (one from $\mathcal{G}_i$ and one from $\mathcal{G}_i'$), which is not always the case. Fortunately, when the loss function is quadratic, more specifically, when $\ell(\cdot) = c_1(\cdot - c_2)^2$ with $c_1, c_2 > 0$, it is shown by Ying et al. (2016) that each expected loss in (6) can be reformulated as the optimal value of a min-max optimization problem whose objective function can be computed in $O(n)$ cost under the empirical distribution. The new formulation also supports online learning since its stochastic gradient can be computed even with one data point (see Lemma 1 below). To derive the reformulation of (6) with quadratic loss functions, we need the following lemma by (Ying et al., 2016) whose proof is provided in Appendix B just for completeness.

Lemma 1 *Let $\mathbf{z} = (\boldsymbol{\xi}, \zeta, \gamma)$ and $\mathbf{z}' = (\boldsymbol{\xi}', \zeta', \gamma')$ be i.i.d random data points. Given any two sets $\mathcal{G}$ and $\mathcal{G}'$ in $\mathbb{R}^{p+2}$ with positive measures w.r.t. $\mathbf{z}$,*

$$\begin{aligned}
&\mathbb{E}\left[c_1(h_{\mathbf{w}}(\boldsymbol{\xi}) - h_{\mathbf{w}}(\boldsymbol{\xi}') - c_2)^2|\mathbf{z} \in \mathcal{G}, \mathbf{z}' \in \mathcal{G}'\right] = \\
&\min_{a,b\in\mathcal{I}_{\mathcal{G},\mathcal{G}'}} \max_{\alpha\in\mathcal{I}_{\mathcal{G},\mathcal{G}'}} \mathbb{E}\left\{F_{\mathcal{G},\mathcal{G}'}(\mathbf{w}, a, b; \mathbf{z}) + \alpha G_{\mathcal{G},\mathcal{G}'}(\mathbf{w}; \mathbf{z}) - \alpha^2\right\}, \quad (7)
\end{aligned}$$

where

$$\begin{aligned}
&F_{\mathcal{G},\mathcal{G}'}(\mathbf{w}, a, b; \mathbf{z}) := \\
&\quad c_1c_2^2 - \frac{2c_1c_2h_{\mathbf{w}}(\boldsymbol{\xi})\mathbb{I}_{\mathcal{G}}(\mathbf{z})}{Pr(\mathbf{z} \in \mathcal{G})} + \frac{2c_1c_2h_{\mathbf{w}}(\boldsymbol{\xi})\mathbb{I}_{\mathcal{G}'}(\mathbf{z})}{Pr(\mathbf{z} \in \mathcal{G}')} \\
&\quad + \frac{c_1(h_{\mathbf{w}}(\boldsymbol{\xi}) - a)^2\mathbb{I}_{\mathcal{G}}(\mathbf{z})}{Pr(\mathbf{z} \in \mathcal{G})} + \frac{c_1(h_{\mathbf{w}}(\boldsymbol{\xi}) - b)^2\mathbb{I}_{\mathcal{G}'}(\mathbf{z})}{Pr(\mathbf{z} \in \mathcal{G}')}, \quad (8)\\
&G_{\mathcal{G},\mathcal{G}'}(\mathbf{w}; \mathbf{z}) := \frac{2c_1h_{\mathbf{w}}(\boldsymbol{\xi})\mathbb{I}_{\mathcal{G}}(\mathbf{z})}{Pr(\mathbf{z} \in \mathcal{G})} - \frac{2c_1h_{\mathbf{w}}(\boldsymbol{\xi})\mathbb{I}_{\mathcal{G}'}(\mathbf{z})}{Pr(\mathbf{z} \in \mathcal{G}')},
\end{aligned}$$

and $\mathcal{I}_{\mathcal{G},\mathcal{G}'} \subset \mathbb{R}$ is the smallest interval such that

$$\begin{aligned}
&0, \pm\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi})|\mathbf{z} \in \mathcal{G}\big], \pm\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi}')|\mathbf{z}' \in \mathcal{G}'\big], \\
&\pm\left(\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi})|\mathbf{z} \in \mathcal{G}\big] - \mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi}')|\mathbf{z}' \in \mathcal{G}'\big]\right) \quad \in \mathcal{I}_{\mathcal{G},\mathcal{G}'}
\end{aligned}$$

for any $\mathbf{w} \in \mathcal{W}$.

According to Lemma 1, the new formulation (7) needs three auxiliary variables, a, b and α in a large enough interval $\mathcal{I}_{\mathcal{G},\mathcal{G}'}$. We then apply Lemma 1 to each conditional expected loss in (6) with $\ell(\cdot) = c_1(\cdot - c_2)^2$. To do so, we first define $\mathcal{I}$ as any bounded interval such that

$$\mathcal{I}_{\mathcal{D}_+,\mathcal{D}_-}, \mathcal{I}_{\mathcal{G}_1,\mathcal{G}_1'}, \mathcal{I}_{\mathcal{G}_2,\mathcal{G}_2'} \subset \mathcal{I}, \quad (9)$$

where $\mathcal{I}_{\mathcal{G},\mathcal{G}'}$ is defined as in Lemma 1. We then introduce fifteen auxiliary variables a_i, b_i and α_i in $\mathcal{I}$ for $i = 0, \ldots, 4$. Here, (a_i, b_i, α_i) for each i corresponds to one conditional expected loss in (6) (there are five of them). In additional, we define the primal variable $\mathbf{x} = (\mathbf{w}, (a_i, b_i)_{i=0}^4) \in \mathcal{X} := \mathcal{W} \times \mathcal{I}^{10}$ and the dual variable $\boldsymbol{\alpha} = (\alpha_i)_{i=0}^4 \in \mathcal{I}^5$. With these notations, we apply Lemma 1 and reformulate (6) as

$$f^* := \min f_0(\mathbf{x}) \text{ s.t. } f_1(\mathbf{x}) \leq 1 + \kappa,\ f_2(\mathbf{x}) \leq 1 + \kappa, \quad (10)$$

where

$$f_0(\mathbf{x}) := \max_{\alpha_0\in\mathcal{I}} \mathbb{E}\big[F_{\mathcal{D}_+,\mathcal{D}_-}(\mathbf{x}; \mathbf{z}) + \alpha_0 G_{\mathcal{D}_+,\mathcal{D}_-}(\mathbf{w}; \mathbf{z}) - \alpha_0^2\big] \quad (11)$$

$$f_1(\mathbf{x}) := \max_{\alpha_1,\alpha_2\in\mathcal{I}} \mathbb{E}\begin{bmatrix} F_{\mathcal{G}_1',\mathcal{G}_1}(\mathbf{x}; \mathbf{z}) + \alpha_1 G_{\mathcal{G}_1',\mathcal{G}_1}(\mathbf{w}; \mathbf{z}) - \alpha_1^2 \\ +F_{\mathcal{G}_2,\mathcal{G}_2'}(\mathbf{x}; \mathbf{z}) + \alpha_2 G_{\mathcal{G}_2,\mathcal{G}_2'}(\mathbf{w}; \mathbf{z}) - \alpha_2^2 \end{bmatrix} \quad (12)$$

$$f_2(\mathbf{x}) := \max_{\alpha_3,\alpha_4\in\mathcal{I}} \mathbb{E}\begin{bmatrix} F_{\mathcal{G}_2',\mathcal{G}_2}(\mathbf{x}; \mathbf{z}) + \alpha_3 G_{\mathcal{G}_2',\mathcal{G}_2}(\mathbf{w}; \mathbf{z}) - \alpha_3^2 \\ +F_{\mathcal{G}_1,\mathcal{G}_1'}(\mathbf{x}; \mathbf{z}) + \alpha_4 G_{\mathcal{G}_1,\mathcal{G}_1'}(\mathbf{w}; \mathbf{z}) - \alpha_4^2 \end{bmatrix} \quad (13)$$

and

$$\begin{aligned}
F_{\mathcal{D}_+,\mathcal{D}_-}(\mathbf{x}; \mathbf{z}) &= F_{\mathcal{D}_+,\mathcal{D}_-}(\mathbf{w}, a_0, b_0; \mathbf{z}) \\
F_{\mathcal{G}_1',\mathcal{G}_1}(\mathbf{x}; \mathbf{z}) &= F_{\mathcal{G}_1',\mathcal{G}_1}(\mathbf{w}, a_1, b_1; \mathbf{z}) \\
F_{\mathcal{G}_2,\mathcal{G}_2'}(\mathbf{x}; \mathbf{z}) &= F_{\mathcal{G}_2,\mathcal{G}_2'}(\mathbf{w}, a_2, b_2; \mathbf{z}) \\
F_{\mathcal{G}_2',\mathcal{G}_2}(\mathbf{x}; \mathbf{z}) &= F_{\mathcal{G}_2',\mathcal{G}_2}(\mathbf{w}, a_3, b_3; \mathbf{z}) \\
F_{\mathcal{G}_1,\mathcal{G}_1'}(\mathbf{x}; \mathbf{z}) &= F_{\mathcal{G}_1,\mathcal{G}_1'}(\mathbf{w}, a_4, b_4; \mathbf{z})
\end{aligned}$$

with $F_{\mathcal{G},\mathcal{G}'}(\mathbf{w}, a, b; \mathbf{z})$ and $G_{\mathcal{G},\mathcal{G}'}(\mathbf{w}; \mathbf{z})$ defined in (8).

4 CONVEX CASE

In this section, we introduce the stochastic feasible level-set (SFLS) method by Lin et al. (2020) for solving (10) when the problem is convex. We make the following assumptions in this section.

Assumption 1 $\mathbb{E}[F_{\mathcal{G},\mathcal{G}'}(\mathbf{x};\mathbf{z})]+\alpha\mathbb{E}[G_{\mathcal{G},\mathcal{G}'}(\mathbf{w};\mathbf{z})]$ *is convex in* $\mathbf{x}$ *for any sets* $\mathcal{G}$ *and* $\mathcal{G}'$ *and any* $\alpha \in \mathbb{R}$.

This assumption holds when $h_{\mathbf{w}}(\boldsymbol{\xi}) = \mathbf{w}^\top \boldsymbol{\xi}$.

Assumption 2 *There exists* $\sigma > 0$ *such that*

$$\mathbb{E}\left[\exp(|F_{\mathcal{G},\mathcal{G}'}(\mathbf{x};\mathbf{z})|^2/\sigma^2)\right] \le \exp(1), \quad (14)$$

$$\mathbb{E}\left[\exp(|G_{\mathcal{G},\mathcal{G}'}(\mathbf{w};\mathbf{z})|^2/\sigma^2)\right] \le \exp(1), \quad (15)$$

$$\mathbb{E}\left[\exp(\|\nabla F_{\mathcal{G},\mathcal{G}'}(\mathbf{x};\mathbf{z})\|_2^2/\sigma^2)\right] \le \exp(1), \quad (16)$$

$$\mathbb{E}\left[\exp(\|\nabla G_{\mathcal{G},\mathcal{G}'}(\mathbf{w};\mathbf{z})\|_2^2/\sigma^2)\right] \le \exp(1) \quad (17)$$

for any sets $\mathcal{G}$ *and* $\mathcal{G}'$ *and any* $\mathbf{x} \in \mathcal{X}$, *where* $\nabla F_{\mathcal{G},\mathcal{G}'}(\mathbf{x};\mathbf{z})$ *and* $\nabla G_{\mathcal{G},\mathcal{G}'}(\mathbf{w};\mathbf{z})$ *are the gradients of* $F_{\mathcal{G},\mathcal{G}'}$ *and* $F_{\mathcal{G},\mathcal{G}'}$ *with respect to* $\mathbf{x}$ *and* $\mathbf{w}$, *respectively.*

Assumption 3 (Strict Feasibility) *There exists* $\tilde{\mathbf{x}} \in \mathcal{X}$ *such that* $\max\{f_1(\tilde{\mathbf{x}}), f_2(\tilde{\mathbf{x}})\} < 1+\kappa$.

As the following lemma shows, this assumption holds if $h_{\mathbf{w}}(\cdot)$ becomes a constant mapping for some $\mathbf{w} \in \mathcal{W}$. The proof is provided in Appendix B.

Lemma 2 *Assumption 3 holds if* $c_1 c_2^2 \le 0.5$ *and there exists* $\mathbf{w} \in \mathcal{W}$ *such that* $h_{\mathbf{w}}(\cdot)$ *is a constant mapping.*

The SFLS method relies on the following *level-set function*

$$H(r) := \min_{\mathbf{x}\in\mathcal{X}} \mathcal{P}(r,\mathbf{x}), \quad (18)$$

where $r \in \mathbb{R}$ is a *level parameter* and

$$\mathcal{P}(r,\mathbf{x}) = \max\{f_0(\mathbf{x}) - r, f_1(\mathbf{x}) - 1 - \kappa, f_2(\mathbf{x}) - 1 - \kappa\}.$$

By lemmas 2.3.4 and 2.3.6 in Nesterov (2003) and Lemma 1 in Lin et al. (2018b), $H(r)$ is non-increasing and convex and has an unique root at $r = f^*$. The SFLS method is essentially a root-finding procedure that generates a sequence of $r^{(k)}$, $k = 0, 1, \ldots$, approaching f^* from the right. The update of $r^{(k)}$ requires the knowledge of $H(r)$ which is unknown. Typically, another algorithm is applied to (18) to obtain an upper bound estimation of $H(r)$. This algorithm is called a stochastic oracle of $H(r)$ defined below.

Definition 3 *Given* $r > f^*$, $\epsilon_{\mathcal{A}} > 0$, *and* $\delta \in (0,1)$, *a* ***stochastic oracle*** $\mathcal{A}(r, \epsilon_{\mathcal{A}}, \delta)$ *returns* $U(r) \in \mathbb{R}$ *and* $\bar{\mathbf{x}} \in \mathcal{X}$ *that satisfy the inequalities* $\mathcal{P}(r,\bar{\mathbf{x}}) - H(r) \le \epsilon_{\mathcal{A}}$ *and* $|U(r) - H(r)| \le \epsilon_{\mathcal{A}}$ *with a probability of at least* $1-\delta$.

Suppose a stochastic oracle $\mathcal{A}$ exists, the SFLS method by Lin et al. (2020) is presented in Algorithm 1 with its convergence property given in Proposition 1.

Algorithm 1 Stochastic Feasible Level-Set Method (SFLS)

Inputs: A stochastic oracle $\mathcal{A}$, a level parameter $r^{(0)} > f^*$, an optimality tolerance $\epsilon_{\text{opt}} > 0$, an oracle error $\epsilon_{\mathcal{A}} > 0$, a probability $\delta \in (0,1)$, and a step length parameter $\theta > 1$.
for $k = 0, 1 \cdots$, **do**
  $\delta^{(k)} = \frac{\delta}{2^k}$
  $(U(r^{(k)}), \mathbf{x}^{(k)}) = \mathcal{A}(r^{(k)}, \epsilon_{\mathcal{A}}, \delta^{(k)})$
  if $U(r^{(k)}) \ge -\epsilon_{opt}$ **then**
    Halt and return $\mathbf{x}^{(k)}$
  $r^{(k+1)} \leftarrow r^{(k)} + U(r^{(k)})/(2\theta)$ and $k \leftarrow k+1$
end for

Proposition 1 (Theorem 5 in Lin et al. (2020)) *Suppose* $\epsilon_{opt} = -\frac{1}{\theta}H(r^{(0)})\epsilon$ *and* $\epsilon_{\mathcal{A}} = -\frac{\theta-1}{2\theta^2(\theta+1)}H(r^{(0)})\epsilon$ *for* $\epsilon \in (0,1)$. *Algorithm 1 generates a feasible solution at each iteration with a probability of at least* $1-\delta$. *Moreover, it returns an* $\mathbf{x}^{(k)}$ *that is feasible and relative* ϵ*-optimal, i.e.,* $(f_0(\mathbf{x}^{(k)}) - f^*)/(f_0(\mathbf{x}^{(0)}) - f^*) \le \epsilon$ *with this probability in at most* $\tilde{O}(\frac{1}{\epsilon^2})$ *iterations.* [1]

The remaining question is how to design an stochastic oracle $\mathcal{A}(r,\epsilon,\delta)$ satisfying Defintion 3. Let $\tilde{\mathbf{y}} = (\tilde{y}_0, \tilde{y}_1, \tilde{y}_2) \in \Delta_3 := \{\tilde{\mathbf{y}} \in \mathbb{R}^3_+ | \sum_{i=0}^2 \tilde{y}_i = 1\}$. With (11), (12) and (13), we can reformulate (18) into

$$H(r) := \min_{\mathbf{x}\in\mathcal{X}} \max_{\tilde{\mathbf{y}}\in\Delta_3, \boldsymbol{\alpha}\in\mathcal{I}^5} \tilde{\phi}(\mathbf{x}, \tilde{\mathbf{y}}, \boldsymbol{\alpha}) \quad (19)$$

where the definition of $\tilde{\phi}(\mathbf{x}, \tilde{\mathbf{y}}, \boldsymbol{\alpha})$ is in Appendix D. This min-max optimization problem is not jointly concave in $\tilde{\mathbf{y}}$ and $\boldsymbol{\alpha}$ due to their product terms. As a result, the standard stochastic mirror descent method, e.g., Nemirovski et al. (2009), does not necessarily converge in theory if applied directly to (19). Motivated by Lin et al. (2018a), we equivalently convert this min-max problem above into a convex-concave min-max problem by changing variables. In particular, we define variables $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_i)_{i=0}^4$ where $\tilde{\alpha}_0 = \tilde{y}_0\alpha_0, \tilde{\alpha}_1 = \tilde{y}_1\alpha_1, \tilde{\alpha}_2 = \tilde{y}_1\alpha_2, \tilde{\alpha}_3 = \tilde{y}_2\alpha_3, \tilde{\alpha}_4 = \tilde{y}_2\alpha_4$ and define $\mathbf{y} = (\tilde{\mathbf{y}}, \tilde{\boldsymbol{\alpha}})$ and

$$\mathcal{Y} := \left\{ \mathbf{y} = (\tilde{\mathbf{y}}, \tilde{\boldsymbol{\alpha}}) \middle| \begin{array}{c} \tilde{\mathbf{y}} \in \Delta_3, \tilde{\alpha}_0 \in \tilde{y}_0 \cdot \mathcal{I}, \\ \tilde{\alpha}_1, \tilde{\alpha}_2 \in \tilde{y}_1 \cdot \mathcal{I}, \tilde{\alpha}_3, \tilde{\alpha}_4 \in \tilde{y}_2 \cdot \mathcal{I} \end{array} \right\}.$$

Eliminating $\boldsymbol{\alpha}$ by $\tilde{\boldsymbol{\alpha}}$ in (19) gives

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \{\phi(\mathbf{x},\mathbf{y}) - d_y(\mathbf{y})\}, \quad (20)$$

where $\phi(\mathbf{x},\mathbf{y}) := \mathbb{E}[\Phi(\mathbf{x},\mathbf{y},\mathbf{z})]$,

$$d_y(\mathbf{y}) := \frac{\tilde{\alpha}_0^2}{\tilde{y}_0} + \frac{\tilde{\alpha}_1^2}{\tilde{y}_1} + \frac{\tilde{\alpha}_2^2}{\tilde{y}_1} + \frac{\tilde{\alpha}_3^2}{\tilde{y}_2} + \frac{\tilde{\alpha}_4^2}{\tilde{y}_2}, \quad (21)$$

$$\Phi(\mathbf{x},\mathbf{y},\mathbf{z}) := \tilde{\mathbf{y}}^\top \mathbf{F}(\mathbf{x},\mathbf{z}) + \tilde{\boldsymbol{\alpha}}^\top \mathbf{G}(\mathbf{w},\mathbf{z}), \quad (22)$$

[1] Here and in the rest of the paper, $\tilde{O}$ suppresses the logarithmic factors in the order of magnitude.

$$\mathbf{F}(\mathbf{x},\mathbf{z}) = \begin{pmatrix} F_{\mathcal{D}_+,\mathcal{D}_-}(\mathbf{x};\mathbf{z}) - r \\ F_{\mathcal{G}_1',\mathcal{G}_1}(\mathbf{x};\mathbf{z}) + F_{\mathcal{G}_2,\mathcal{G}_2'}(\mathbf{x};\mathbf{z}) - 1 - \kappa \\ F_{\mathcal{G}_2',\mathcal{G}_2}(\mathbf{x};\mathbf{z}) + F_{\mathcal{G}_1,\mathcal{G}_1'}(\mathbf{x};\mathbf{z}) - 1 - \kappa \end{pmatrix}$$

and

$$\mathbf{G}(\mathbf{w},\mathbf{z}) = \begin{pmatrix} G_{\mathcal{D}_+,\mathcal{D}_-}(\mathbf{w};\mathbf{z}) \\ G_{\mathcal{G}_1,\mathcal{G}_1'}(\mathbf{w};\mathbf{z}) \\ G_{\mathcal{G}_2',\mathcal{G}_2}(\mathbf{w};\mathbf{z}) \\ G_{\mathcal{G}_2,\mathcal{G}_2'}(\mathbf{w};\mathbf{z}) \\ G_{\mathcal{G}_1',\mathcal{G}_1}(\mathbf{w};\mathbf{z}) \end{pmatrix}.$$

We also slightly generalize (23) to the following problem

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \{\phi(\mathbf{x},\mathbf{y}) - d_y(\mathbf{y}) + d_x(\mathbf{x})\}, \tag{23}$$

where $d_x(\mathbf{x}) = \frac{\hat{\rho}}{2}\|\mathbf{x}-\tilde{\mathbf{x}}\|_2^2$ for some $\hat{\rho} \geq 0$ and some $\tilde{\mathbf{x}} \in \mathcal{X}$. In this section, we focus on the convex case and only need to solve (23) with $\hat{\rho} = 0$. When we solve the weakly convex case later, we will set $\hat{\rho} > 0$ and choose some $\tilde{\mathbf{x}}$.

Note that (23) is a convex-concave min-max problem. In fact, except the term $d_y(\mathbf{y})$, the objective function is linear in $\mathbf{y}$, which allows us to apply stochastic mirror descent (SMD) method. The SMD method requires some distance generating function on $\mathcal{X}$ and $\mathcal{Y}$ and their corresponding *Bregman divergences*. In our problem, the distance generating functions on $\mathcal{X}$ and $\mathcal{Y}$ are chosen as $\omega_x(\mathbf{x}) := \frac{1}{2}\|\mathbf{x}\|_2^2$ and $\omega_y(\mathbf{y}) := 2(1+\sqrt{2}I)^2\left(\sum_{i=0}^{2}\tilde{y}_i \ln \tilde{y}_i + \ln 3\right) + d_y(\mathbf{y})$ respectively, where $I := \max_{\alpha\in\mathcal{I}}|\alpha|$. Function $\omega_y(\mathbf{y})$ is specially designed for the set $\mathcal{Y}$ so, as we will show below, the iterates in the SMD method can be updated in closed-forms. Note that we can always choose $\mathcal{I}$ such that it satisfies (9) and is bounded. In fact, since $\mathcal{W}$ is compact and $\mathbb{E}[h_{\mathbf{w}}(\boldsymbol{\xi})|\mathcal{G}]$ is continuous in $\mathbf{w}$, the intervals $\mathcal{I}_{\mathcal{D}_+,\mathcal{D}_-}$, $\mathcal{I}_{\mathcal{G}_1,\mathcal{G}_1'}$ and $\mathcal{I}_{\mathcal{G}_2,\mathcal{G}_2'}$ are all bounded, so we can also set $\mathcal{I}$ to be bounded. This ensures $I < +\infty$.

Let $\|\mathbf{x}\|_x := \|\mathbf{x}\|_2$ and $\|\mathbf{y}\|_y := \|\mathbf{y}\|_{1,2} := \sqrt{\|\tilde{\mathbf{y}}\|_1^2 + \|\tilde{\boldsymbol{\alpha}}\|_2^2}$. It is clear that $\omega_x(\mathbf{x})$ is 1-strongly convex on $\mathcal{X}$ with respect to $\|\mathbf{x}\|_x$. It is shown by Lemma 2 in Lin et al. (2018a) that $\omega_y(\mathbf{y})$ is 1-strongly convex on $\mathcal{Y}$ with respect to $\|\mathbf{y}\|_y$. Hence, we can use them to define Bregman divergence $V_x(\mathbf{x},\mathbf{x}') = \omega_x(\mathbf{x}) - [\omega_x(\mathbf{x}') + \nabla\omega_x(\mathbf{x}')^\top(\mathbf{x}-\mathbf{x}')] = \frac{1}{2}\|\mathbf{x}-\mathbf{x}'\|_2^2$ and

$$\begin{aligned} V_y(\mathbf{y},\mathbf{y}') &:= \omega_y(\mathbf{y}) - [\omega_y(\mathbf{y}') + \nabla\omega_y(\mathbf{y}')^\top(\mathbf{y}-\mathbf{y}')] \\ &= 2(1+\sqrt{2}I)^2\sum_{i=0}^{2}\tilde{y}_i \ln\left(\frac{\tilde{y}_i}{\tilde{y}_i'}\right) + \tilde{y}_0\left(\frac{\tilde{\alpha}_0}{\tilde{y}_0} - \frac{\tilde{\alpha}_0'}{\tilde{y}_0'}\right)^2 \\ &\quad + \tilde{y}_1\sum_{i=1}^{2}\left(\frac{\tilde{\alpha}_i}{\tilde{y}_1} - \frac{\tilde{\alpha}_i'}{\tilde{y}_1'}\right)^2 + \tilde{y}_2\sum_{j=3}^{4}\left(\frac{\tilde{\alpha}_j}{\tilde{y}_2} - \frac{\tilde{\alpha}_j'}{\tilde{y}_2'}\right)^2. \end{aligned} \tag{24}$$

With these Bregman divergences, we describe the SMD method in Algorithm 2. The subproblems (25) and (26) have closed-form solutions, which are characterized in Lemma 3 in Appendix B.2.

Algorithm 2 Stochastic Mirror Descent for (23)

1: **Input:** Level parameter $r \in \mathbb{R}$, number of iterations T, step size η_t and τ_t, $\hat{\rho} \geq 0$ and $\tilde{\mathbf{x}}$.
2: Set $\mathbf{x}^{(0)} = \mathbf{0}$, $\tilde{\mathbf{y}}^{(0)} = (\frac{1}{3},\frac{1}{3},\frac{1}{3})^\top$, $\tilde{\boldsymbol{\alpha}}^{(0)} = \mathbf{0}$ and $\mathbf{y}^{(0)} = (\tilde{\mathbf{y}}^{(0)}, \tilde{\boldsymbol{\alpha}}^{(0)})$
3: **for** $t = 0$ **to** $T-1$ **do**
4: Sample $\mathbf{z}^{(t)}$.
5: Compute stochastic gradients:

$$\mathbf{g}_{\mathbf{x}}^{(t)} = \nabla_x\Phi(\mathbf{x}^{(t)},\mathbf{y}^{(t)},\mathbf{z}^{(t)}),\ \mathbf{g}_{\mathbf{y}}^{(t)} = \nabla_y\Phi(\mathbf{x}^{(t)},\mathbf{y}^{(t)},\mathbf{z}^{(t)})$$

6: Primal-dual stochastic mirror descent:

$$\begin{aligned} \mathbf{x}^{(t+1)} &= \arg\min_{\mathbf{x}\in\mathcal{X}} \left\langle \mathbf{g}_{\mathbf{x}}^{(t)}, \mathbf{x}\right\rangle + \frac{\|\mathbf{x}-\mathbf{x}^{(t)}\|_2^2}{2\eta_t} + d_x(\mathbf{x}) \\ \mathbf{y}^{(t+1)} &= \arg\min_{\mathbf{y}\in\mathcal{Y}} -\left\langle \mathbf{g}_{\mathbf{y}}^{(t)}, \mathbf{y}\right\rangle + \frac{V_y(\mathbf{y},\mathbf{y}^{(t)})}{\tau_t} + d_y(\mathbf{y}) \end{aligned} \tag{25}$$

7: Compute a stochastic upper bound

$$U(r) := \max_{\mathbf{y}\in\mathcal{Y}} \left\{ \frac{\sum_{t=0}^{T-1}\tau_t\left[\Phi(\mathbf{x}^{(t)},\mathbf{y},\mathbf{z}^{(t)}) - d_y(\mathbf{y}) + d_x(\mathbf{x}^{(t)})\right]}{\sum_{t=0}^{T-1}\tau_t} \right\}. \tag{26}$$

8: **Output:** $U(r)$ and $(\bar{\mathbf{x}},\bar{\mathbf{y}}) = \frac{1}{T}\sum_{t=0}^{T-1}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})$.

The convergence property of Algorithm 2 is well known (see, e.g., Lin et al. (2020); Nemirovski et al. (2009)) and, in combination with Proposition 1, it implies the total complexity of Algorithm 1 as stated in the following theorem.

Theorem 1
Let $D_x := \sqrt{\max_{\mathbf{x}\in\mathcal{X}}\omega_x(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{X}}\omega_x(\mathbf{x})}$ and $D_y := \sqrt{\max_{\mathbf{y}\in\mathcal{Y}}\omega_y(\mathbf{y}) - \min_{\mathbf{y}\in\mathcal{Y}}\omega_y(\mathbf{y})}$. There exists a constant M depending on σ, $\hat{\rho}$, I, D_x and D_y such that:

- *Algorithm 2 is a stochastic oracle by Definition 3 if $T \geq \tilde{O}\left(\frac{1}{\epsilon_{\mathcal{A}}^2}\ln(\frac{1}{\delta})\right)$ and*

$$\eta_t = 2D_x^2/(M\sqrt{t+1}), \tau_t = 2D_y^2/(M\sqrt{t+1}). \tag{27}$$

- *Suppose ϵ_{opt} and $\epsilon_{\mathcal{A}}$ are defined the same as in Proposition 1. If Algorithm 2 is used as the stochastic oracle $\mathcal{A}$ with η_t and τ_t defined as in (27) and $T \geq \tilde{O}\left(\frac{1}{\epsilon_{\mathcal{A}}^2}\ln(\frac{1}{\delta^{(k)}})\right)$ with $\delta^{(k)}$ defined in Algorithm 1. Algorithm 1 returns a relative ϵ-optimal and feasible solution with probability of at least $1-\delta$ after running at most $\tilde{O}\left(\frac{1}{\epsilon^2}\ln(\frac{1}{\delta})\right)$ stochastic mirror descent steps across all calls of $\mathcal{A}$.*

In Appendix C, we provide the definition of M and the exact value of T and we also give a brief discussion on how

this theorem is obtained by applying the convergence results in Lin et al. (2020); Nemirovski et al. (2009) to Algorithm 2.

5 WEAKLY-CONVEX CASE

In this section, we apply the proximal point techniques by Boob et al. (2022); Jia and Grimmer (2022); Ma et al. (2020) to extend the approach to the case where the objective and constraint functions in (10) are weakly convex.

Definition 4 *Given $h : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$, we say h is μ-strongly convex for $\mu \geq 0$ if*

$$h(\mathbf{x}) \geq h(\mathbf{x}') + \mathbf{g}^\top(\mathbf{x} - \mathbf{x}') + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}'\|_2^2$$

for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and any $\mathbf{g} \in \partial h(\mathbf{x})$, and we say h is ρ-weakly convex for $\rho \geq 0$ if

$$h(\mathbf{x}) \geq h(\mathbf{x}') + \mathbf{g}^\top(\mathbf{x} - \mathbf{x}') - \frac{\rho}{2}\|\mathbf{x} - \mathbf{x}'\|_2^2$$

for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and any $\mathbf{g} \in \partial h(\mathbf{x})$. Here, $\partial h(\mathbf{x})$ is the subdifferential of h at $\mathbf{x}$.

In this section, we do not assume Assumption 1 but assume Assumptions 2 and 3 and the following assumption.

Assumption 4 *The following statements hold:*

1. *$\mathbb{E}[F_{\mathcal{G},\mathcal{G}'}(\mathbf{x}; z)] + \alpha\mathbb{E}[G_{\mathcal{G},\mathcal{G}'}(\mathbf{w}; z)]$ is ρ-weakly convex in $\mathbf{x}$ for any sets $\mathcal{G}$ and $\mathcal{G}'$ and any $\alpha \in \mathbb{R}$.*
2. *There exist $\sigma_\epsilon > 0$ and $\rho_\epsilon > 0$ such that*

$$\min_{\mathbf{x}' \in \mathcal{X}} \left\{ \max_{i=1,2} f_i(\mathbf{x}') - 1 - \kappa + \frac{\rho + \rho_\epsilon}{2}\|\mathbf{x}' - \mathbf{x}\|^2 \right\} < -\sigma_\epsilon$$

for any $\mathbf{x} \in \mathcal{X}$ with $\max_{i=1,2} f_i(\mathbf{x}) - 1 - \kappa \leq \epsilon^2$.

3. *$\|\mathbf{g}\| \leq G$ for a constant G for any $\mathbf{g} \in \partial f_i(\mathbf{x})$ for $i = 0, 1, 2$ and $\mathbf{x} \in \mathcal{X}$.*

In Appendix B.3, we will provide a sufficient condition for Assumption 4.2 to hold. In this case, the objective or the constraint functions can be non-convex, so finding an ϵ-optimal solution is challenging in general. Hence, we target at finding a nearly ϵ-stationary point defined below.

Definition 5 *A point $\mathbf{x} \in \mathcal{X}$ is called a **ϵ-Karush-Kuhn-Tucker (KKT) point** of (10) if there exist Lagrangian multiplies $\lambda_i \geq 0$ and $\mathbf{g}_i \in \partial f_i(\mathbf{x})$ for $i = 1$ and 2 such that*

$$\begin{aligned} &Dist(\mathbf{g}_0 + \lambda_1\mathbf{g}_1 + \lambda_2\mathbf{g}_2, -\mathcal{N}_{\mathcal{X}}(\mathbf{x})) \leq \epsilon, \\ &|\lambda_i(f_i(\mathbf{x}) - 1 - \kappa)| \leq \epsilon^2,\ f_i(\mathbf{x}) \leq 1 + \kappa + \epsilon,\ i = 1, 2, \end{aligned}$$

*where $\mathcal{N}_{\mathcal{X}}(\mathbf{x})$ is the normal cone of $\mathcal{X}$ at $\mathbf{x}$. Let $\hat{\rho} > \rho$. A point $\tilde{\mathbf{x}} \in \mathcal{X}$ is called a **nearly ϵ-stationary point** of (10) if $\|\widehat{\mathbf{x}} - \tilde{\mathbf{x}}\| \leq \epsilon$ where*

$$\widehat{\mathbf{x}} \equiv \arg\min_{\mathbf{x}' \in \mathcal{X}} f_0(\mathbf{x}') + \frac{\hat{\rho}}{2}\|\mathbf{x}' - \tilde{\mathbf{x}}\|_2^2, \quad (28)$$
$$s.t.\ f_i(\mathbf{x}') + \frac{\hat{\rho}}{2}\|\mathbf{x}' - \tilde{\mathbf{x}}\|_2^2 \leq 1 + \kappa, i = 1, 2.$$

Algorithm 3 Inexact Quadratically Regularized Constrained Method

Input: An ϵ^2-feasible solution $\tilde{\mathbf{x}}^{(0)}$, $\rho + \rho_\epsilon \geq \hat{\rho} > \rho$, $\delta \in (0, 1)$, $\hat{\epsilon} = \min\left\{1, \sqrt{\frac{\hat{\rho}-\rho}{4}}\left(\frac{G+2\hat{\rho}D_x}{\sqrt{2\sigma_\epsilon(\hat{\rho}-\rho)}} + 1\right)^{-\frac{1}{2}}\right\}\epsilon$, and the number of iterations S.
for $s = 0, \ldots, S-1$ **do**
  Compute $\tilde{\mathbf{x}}^{(s+1)} = \mathcal{B}(\tilde{\mathbf{x}}^{(s)}, \hat{\rho}, \hat{\epsilon}, \frac{\delta}{S})$
Output: $\tilde{\mathbf{x}}^{(R)}$ where R is a random index uniformly sampled from $\{0, \ldots, S\}$.

Remark 1 *Since $\widehat{\mathbf{x}}$ is optimal for (28), there exist Lagrangian multiplies $\widehat{\lambda}_i \geq 0$ and $\widehat{\mathbf{g}}_i \in \partial f_i(\widehat{\mathbf{x}})$ for $i = 1$ and 2 such that*

$$\begin{aligned} &Dist(\widehat{\mathbf{g}}_0 + \widehat{\lambda}_1\widehat{\mathbf{g}}_1 + \widehat{\lambda}_2\widehat{\mathbf{g}}_2, -\mathcal{N}_{\mathcal{X}}(\widehat{\mathbf{x}})) \\ \leq &\hat{\rho}(1 + \widehat{\lambda}_1 + \widehat{\lambda}_2)\|\widehat{\mathbf{x}} - \tilde{\mathbf{x}}\|_2, \\ &|\widehat{\lambda}_i(f_i(\widehat{\mathbf{x}}) - 1 - \kappa)| \leq \frac{\widehat{\lambda}_i\hat{\rho}}{2}\|\widehat{\mathbf{x}} - \tilde{\mathbf{x}}\|_2^2, \\ &f_i(\widehat{\mathbf{x}}) \leq 1 + \kappa,\ i = 1, 2. \end{aligned}$$

As discussed in Boob et al. (2022); Jia and Grimmer (2022); Ma et al. (2020), when $\widehat{\lambda}_i$ for $i = 1$ and 2 are bounded, a nearly ϵ-stationary point $\tilde{\mathbf{x}}$ is no more than ϵ away from $\widehat{\mathbf{x}}$, which is an $O(\epsilon)$-KKT point of (10). This justifies why a nearly ϵ-stationary point is a reasonable target for solving (10) when the problem is non-convex. Different assumptions are considered in Boob et al. (2022); Jia and Grimmer (2022); Ma et al. (2020) to ensure the boundness of $\widehat{\lambda}_i$. This paper follows Ma et al. (2020) by assuming Assumption 4.2 and the boundness of $\widehat{\lambda}_i$ under this assumption follows Lemma 1 in Ma et al. (2020).

Next we apply the inexact quadratically regularized constrained (IQRC) method by Ma et al. (2020) to (10), which is given in Algorithm 3. This algorithm requires an oracle define below.

Definition 6 *Given $\tilde{\mathbf{x}} \in \mathcal{X}$, $\hat{\rho} > 0$, $\hat{\epsilon} > 0$, $\delta \in (0, 1)$, a **stochastic oracle** $\mathcal{B}(\tilde{\mathbf{x}}, \hat{\rho}, \hat{\epsilon}, \delta)$ returns $\mathbf{x}' \in \mathcal{X}$ such that, with a probability of at least $1 - \delta$, $\mathbf{x}'$ is an $\hat{\epsilon}^2$-feasible and $\hat{\epsilon}^2$-optimal solution of (28).*

According to the definition of this oracle, in its iteration t, Algorithm 3 needs to find $\hat{\epsilon}^2$-feasible and $\hat{\epsilon}^2$-optimal solution of subproblem (28) with $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}^{(s)}$. Since (28) is convex when $\hat{\rho} > \rho$, Algorithm 1 can be used as an oracle $\mathcal{B}$. To do so, we need to derive and solve the level-set subproblem (18) corresponding to (28), which is

$$\widehat{H}(r) := \min_{\mathbf{x} \in \mathcal{X}} \left\{ \mathcal{P}(r, \mathbf{x}) + \frac{\hat{\rho}}{2}\|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2 \right\}. \quad (29)$$

Following the same step as in Section 4, (29) can be reformulated as (23). Recall that we set $\hat{\rho} = 0$ in Section 4 when

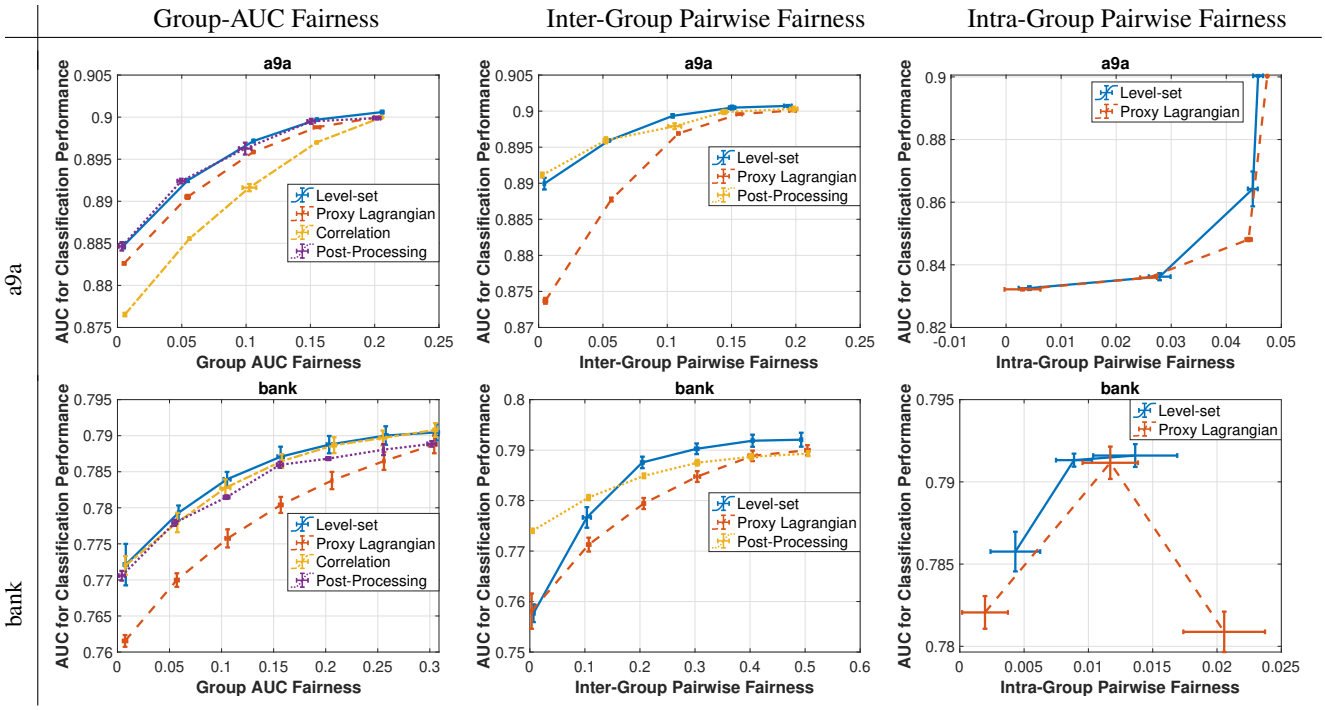

Figure 1: Pareto frontiers by each method on testing set in convex case (see Appendix E.4 for COMPAS dataset).

the problem is convex, but here we set $\hat{\rho} > \rho$ because of non-convexity.

According to Theorem 1, when Algorithm 2 is used as the oracle $\mathcal{A}$ in Algorithm 1, Algorithm 1 becomes an oracle $\mathcal{B}$ for Algorithm 3 with an iteration complexity of $\tilde{O}(\frac{1}{\hat{\epsilon}^4})$=$\tilde{O}(\frac{1}{\epsilon^4})$. According to Theorem 1 in Ma et al. (2020), Algorithm 3 finds a nearly ϵ-stationary point of (10) in $O(\frac{1}{\epsilon^2})$ iterations with $\mathcal{B}$ called once in each iteration. Combining these two results, we know that the total iteration complexity of Algorithm 3 is $\tilde{O}(\frac{1}{\epsilon^4}) \times O(\frac{1}{\epsilon^2}) = \tilde{O}(\frac{1}{\epsilon^6})$. This is formally stated in the following theorem. The proof is omitted since this theorem can be easily obtained from the existing results according to the discussion above.

Theorem 2 *Suppose Algorithm 3 uses Algorithm 1 as oracle $\mathcal{B}$ and ϵ_{opt} and $\epsilon_{\mathcal{A}}$ in Algorithm 1 are set as in Proposition 1 except that H is replaced by $\widehat{H}$ in (29). Also, suppose Algorithm 1 uses Algorithm 2 as oracle $\mathcal{A}$ and η_t, τ_t and T are set as in Theorem 1. Algorithm 3 returns $\tilde{\mathbf{x}}^{(R)}$ as a nearly ϵ-stationary point of (10) within $\tilde{O}(\frac{1}{\epsilon^6})$ stochastic mirror descent steps across all calls of $\mathcal{B}$.*

6 NUMERICAL EXPERIMENTS

In this section, we demonstrate the effectiveness of the proposed approaches for AUC optimization subject to the AUC-based fairness constraints given in Examples 1, 2 and 3 in Section 3. All experiments are conducted on a computer with the CPU 2GHz Quad-Core Intel Core i5 and the GPU NVIDIA GeForce RTX 2080 Ti.

Datasets Information. The experiments are conducted using three public datasets: *a9a* (Chang and Lin, 2011; Dua and Graff, 2017; Kohavi, 1996), *bank* (Chang and Lin, 2011; Dua and Graff, 2017; Moro et al., 2004) and *COMPAS* (Fabris et al., 2022; J. Angwin and Kirchner, 2016). Details about these datasets can be found in Appendix E.1.

Baselines. We compare our methods with three baselines, the proxy-Lagrangian method (Cotter et al., 2019), the correlation-penalty method (Beutel et al., 2019a) and the post-processing method (Kallus and Zhou, 2019). The description of each baseline is provided in Appendix E.2.

Convex case. For convex case, we consider a linear model, i.e., $h_{\mathbf{w}}(\boldsymbol{\xi}) = \boldsymbol{\xi}^\top \mathbf{w}$. A smaller κ in (2) makes the model more fair in terms of the corresponding fairness metric but may compromise the classification performance in terms of AUC. Hence, we varies κ in (2) so each method in comparison will generate a Pareto frontier, showing the trade-off between performance and fairness.

For the three baselines and our algorithm, the process to tune the hyper-parameters is explained in Appendix E.3. We then evaluate AUC and the fairness metric of the output model on testing set and report the Pareto frontiers by each method in Figure 1. We repeat each experiment five times with different random seeds and report the standard errors of the AUC scores and the fairness metrics through the error bars on each curve. Due to the page limit, we postpone the plots of *COMPAS* dataset to Appendix E.4.

Weakly-convex case. For weakly-convex case, we choose

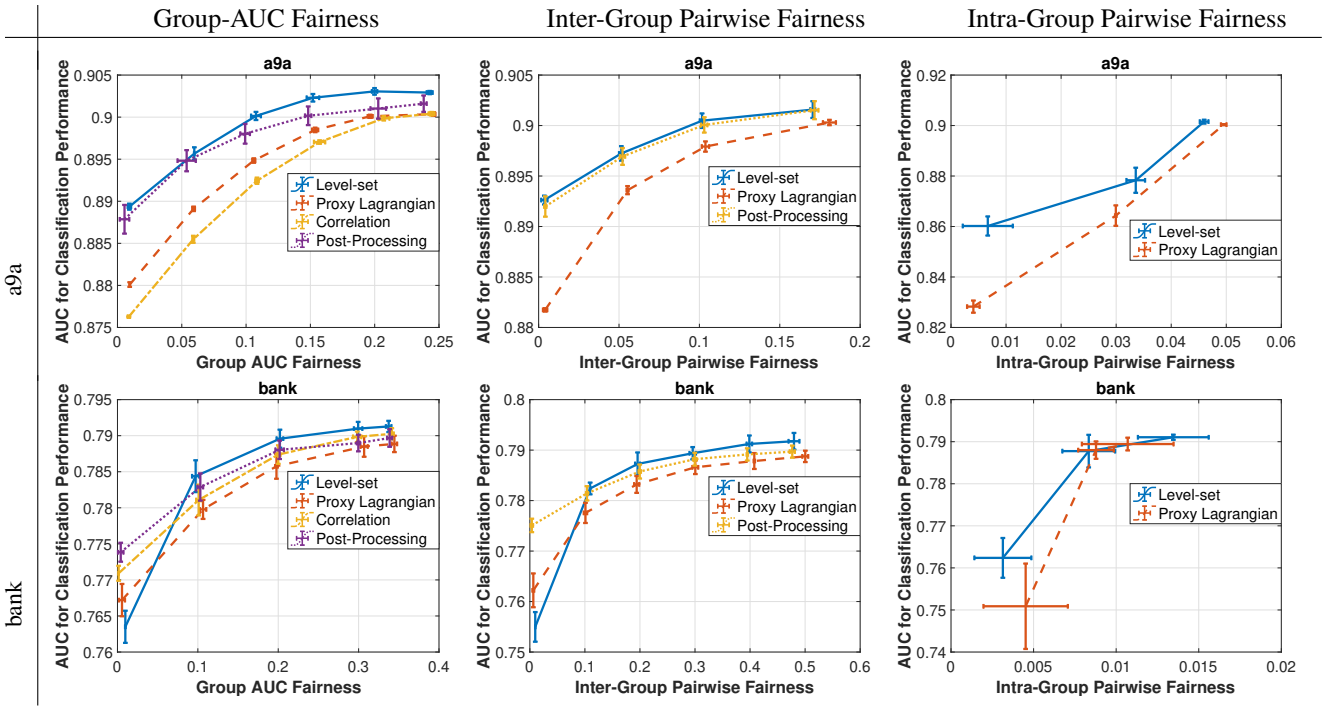

Figure 2: Pareto frontiers by each method on testing set in weakly-convex case (see Appendix E.4 for COMPAS dataset).

$h_{\mathbf{w}}$ to be a two-layer neural network with 10 hidden neurons and the sigmoid activation functions. The process of tuning hyperparameters is in Appendix E.3. In the non-convex case, the original proxy-Lagrangian method in Cotter et al. (2019) updates $\mathbf{w}$ through an approximate Bayesian optimization oracle, which can solve a non-convex problem with a reasonably small optimality gap. Here, we directly perform one stochastic gradient descent step to update $\mathbf{w}$ just as in the convex case because it is unclear how to design such an oracle due to non-convexity. The Pareto frontiers in weakly-convex case are reported with error bars in Figure 2.

It can be observed from Figures 1 and 2 that the level-set method performs better than the other three baselines when κ is not too small. When κ is small, the level-set method is less efficient in trading performance for fairness on the *bank* dataset. This is likely because the approximation gap between (10) and (2) is large on this dataset. As a result, we have to use a very small κ in (10) in order to achieve the targeted fairness level in the original problem (2), which leads to very restrictive constraints in (10) and harms the classification performance in terms of AUC.

7 CONCLUSION and LIMITATION

We consider AUC optimization subject to a class of AUC-based fairness constraints, which includes most of the existing threshold-agnostic and comparison-based fairness metrics in literature. When solving this problem in an online setting where the data arrives sequentially, the existing optimization methods need to receive at least a pair of data points to update the model, which may not be allowed by the order of data's arrivals. In addition, when the original problem is formulated using an empirical distribution in an off-line setting, the computational cost becomes quadratic in data size due to the definition of AUC. This computational cost is too high when the data is large.

To address these computational challenges, we reformulated this problem into a min-max optimization problem subject to min-max constraints using a quadratic loss function to approximate the AUCs in the objective and constraint functions. The new optimization formulation allows the model to be updated in an online fashion with one data point arriving each time. In the off-line setting, the new formulation also reduces the computational cost to only linear in data size. By introducing a novel Bregman divergence after changing variables, we show that existing stochastic optimization algorithms can be applied to the new formulation in the convex and weakly convex cases. In the numerical experiments, we observe an efficient trade off between classification performance and fairness by the models created by our approaches.

However, we acknowledge that our formulation only works for the quadratic loss function. It is our future work to further extend our methods for a general loss function.

Acknowledgements

This work was jointly supported by the University of Iowa Jumpstarting Tomorrow Program and NSF award 2147253.

References

P. M. Addo, D. Guegan, and B. Hassani. Credit risk analysis using machine and deep learning models. *Risks*, 6(2):38, 2018. 1

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018. 1, 2

A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017. 2, 3

A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, et al. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2212–2220, 2019a. 1, 2, 3, 8, 12

A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 453–459, 2019b. 1

D. Boob, Q. Deng, and G. Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming*, pages 1–65, 2022. 2, 3, 7

D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. 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, 2019. 1, 2, 3, 12

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009. 1

C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011. 8

A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017. 1

A. Cotter, M. Gupta, H. Jiang, N. Srebro, K. Sridharan, S. Wang, B. Woodworth, and S. You. Training fairness-constrained classifiers to generalize. In *ICML 2018 workshop:"Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2018. 1, 2

A. Cotter, H. Jiang, M. R. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *J. Mach. Learn. Res.*, 20 (172):1–59, 2019. 1, 2, 8, 9

A. F. Cruz, C. Belém, J. Bravo, P. Saleiro, and P. Bizarro. Fairgbm: Gradient boosting with fairness constraints. *arXiv preprint arXiv:2209.07850*, 2022. 1, 2

M. R. Davahli, W. Karwowski, and K. Fiok. Optimizing covid-19 vaccine distribution across the united states using deterministic and stochastic recurrent neural networks. *Plos one*, 16(7):e0253925, 2021. 1

E. Diana, W. Gill, M. Kearns, K. Kenthapadi, and A. Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76, 2021. 1, 2

L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018. 1, 2

J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018. 1

D. Dua and C. Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`. 8

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012. 1, 2

A. Fabris, S. Messina, G. Silvello, and G. A. Susto. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152, 2022. doi: 10.1007/s10618-022-00854-z. 8

P. Gajane and M. Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017. 1

G. Goh, A. Cotter, M. Gupta, and M. P. Friedlander. Satisfying real-world goals with dataset constraints. *Advances in Neural Information Processing Systems*, 29, 2016. 1, 2

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016. 1, 2

S. M. J. Angwin, J. Larson and L. Kirchner. Machine bias. *ProPublica, May, 23*, 2016. 8

Z. Jia and B. Grimmer. First-order methods for nonsmooth nonconvex functional constrained optimization with or without slater points. *arXiv preprint arXiv:2212.00927*, 2022. 7

N. Kallus and A. Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *Advances in neural information processing systems*, 32, 2019. 1, 2, 3, 8, 12

M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018. 1, 2

R. Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining*, pages 202–207, 1996. 8

M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017. 1

Q. Lin, R. Ma, and T. Yang. Level-set methods for finite-sum constrained convex optimization. In *International conference on machine learning*, pages 3112–3121. PMLR, 2018a. 5, 6

Q. Lin, S. Nadarajah, and N. Soheili. A level-set method for convex optimization with a feasible solution path. *SIAM Journal on Optimization*, 28(4):3290–3311, 2018b. 5

Q. Lin, S. Nadarajah, N. Soheili, and T. Yang. A data efficient and feasible level set method for stochastic convex optimization with expectation constraints. *Journal of machine learning research*, 2020. 2, 3, 5, 6, 7, 15, 16

R. Ma, Q. Lin, and T. Yang. Quadratically regularized subgradient methods for weakly convex optimization with weakly convex constraints. In *International Conference on Machine Learning*, pages 6554–6564. PMLR, 2020. 2, 7, 8

S. Moro, P. Cortez, and P. Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems, 62*, pages 22–31, 2004. doi: 10.1016/j.dss.2014.03.001. 8

H. Narasimhan, A. Cotter, M. Gupta, and S. Wang. Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5248–5255, 2020. 1, 2

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009. 5, 6, 7, 15

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003. 5

S. Verma and J. Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018. 2

R. Vogel, A. Bellet, and S. Clémençon. Learning fair scoring functions: Bipartite ranking under roc-based fairness constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 784–792. PMLR, 2021. 1, 2, 3

B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, 2017. 1, 2

Y. Yan and Y. Xu. Adaptive primal-dual stochastic gradient method for expectation-constrained convex stochastic programs. *Mathematical Programming Computation*, pages 1–45, 2022. 2

S. Yang, X. Li, and G. Lan. Data-driven minimax optimization with expectation constraints. *arXiv preprint arXiv:2202.07868*, 2022a. 2

S. Yang, Z. Zhang, and E. X. Fang. Stochastic compositional optimization with compositional constraints. *arXiv preprint arXiv:2209.04086*, 2022b. 3

Z. Yang, Y. L. Ko, K. R. Varshney, and Y. Ying. Minimax auc fairness: Efficient algorithm with provable convergence. *arXiv preprint arXiv:2208.10451*, 2022c. 1, 2

Y. Ying, L. Wen, and S. Lyu. Stochastic online auc maximization. *Advances in neural information processing systems*, 29, 2016. 3, 4

M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 962–970. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/zafar17a.html. 1

A EXAMPLES of FAIRNESS METRICS SATISFYING DEFINITION 2

In this section, we present five examples of fairness metrics that satisfy Definition 2 and thus can be applied as fairness constraints in (2) and solved by the optimization algorithms in this paper. In the discussion below, we assume all data points are ranked decreasingly in $h_{\mathbf{w}}(\boldsymbol{\xi})$ so $h_{\mathbf{w}}(\boldsymbol{\xi}) > h_{\mathbf{w}}(\boldsymbol{\xi}')$ means z is ranked higher than z'.

Example 1 (Group AUC Fairness) *Let $\mathcal{G}_1 = \{\mathbf{z}|\gamma = 1\}$, $\mathcal{G}_1' = \{\mathbf{z}|\gamma = -1\}$ and $\mathcal{G}_2 = \mathcal{G}_2' = \mathbb{R}^{p+2}$ (so $AUC_{\mathbf{w}}(\mathcal{G}_2, \mathcal{G}_2') = 0.5$). The AUC-based fairness metric becomes $|Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > h_{\mathbf{w}}(\boldsymbol{\xi}')|\gamma = 1, \gamma' = -1) - 0.5|$. When it is small, a random data point from the protected group is ranked above a random data point from the unprotected group with nearly 50% probability. In other words, if we use $h_{\mathbf{w}}(\boldsymbol{\xi})$ to predict sensitive variable γ, it must has a poor prediction performance in terms of AUC w.r.t. γ (instead of ζ).*

Example 2 (Inter-Group Pairwise Fairness) *Let $\mathcal{G}_1 = \{\mathbf{z}|\zeta = 1, \gamma = 1\}$, $\mathcal{G}_1' = \{\mathbf{z}|\zeta = -1, \gamma = -1\}$, $\mathcal{G}_2 = \{\mathbf{z}|\zeta = 1, \gamma = -1\}$ and $\mathcal{G}_2' = \{\mathbf{z}|\zeta = -1, \gamma = 1\}$. In this case, the AUC-based fairness metric becomes the cross-AUC in Kallus and Zhou (2019), which is also called inter-group pairwise fairness (Beutel et al., 2019a). When it is small, the probability of a random positive data point being ranked above a random negative data point from the opposite group is nearly independent of the group.*

Example 3 (Intra-Group Pairwise Fairness) *Let $\mathcal{G}_1 = \{\mathbf{z}|\zeta = 1, \gamma = 1\}$, $\mathcal{G}_1' = \{\mathbf{z}|\zeta = -1, \gamma = 1\}$, $\mathcal{G}_2 = \{\mathbf{z}|\zeta = 1, \gamma = -1\}$ and $\mathcal{G}_2' = \{\mathbf{z}|\zeta = -1, \gamma = -1\}$. In this case, the AUC-based fairness metric becomes the intra-group pairwise fairness introduced by Beutel et al. (2019a). When it is small, the probability of a random positive data point being ranked above a random negative data point from the same group is nearly independent of the group. In other words, the classical AUCs (w.r.t. class labels) evaluated separately on each group are similar.*

Example 4 (Average Equality Gaps) *Let $\mathcal{G}_1 = \{\mathbf{z}|\zeta = 1, \gamma = 1\}$, $\mathcal{G}_1' = \{\mathbf{z}|\zeta = 1\}$ and $\mathcal{G}_2 = \mathcal{G}_2' = \mathbb{R}^{p+2}$. The AUC-based fairness metric becomes the positive average equality gap introduced by Borkan et al. (2019), i.e., $|Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > h_{\mathbf{w}}(\boldsymbol{\xi}')|\gamma = 1, \zeta = 1, \zeta' = 1) - 0.5|$. Similar to Example 1, when this value is small, a random positive data point from the protected group is ranked above a random positive data from the whole dataset with nearly 50% probability. Similarly, the negative average equality gap by Borkan et al. (2019) is obtained when $\mathcal{G}_1 = \{\mathbf{z}|\zeta = -1, \gamma = 1\}$, $\mathcal{G}_1' = \{\mathbf{z}|\zeta = -1\}$ and $\mathcal{G}_2 = \mathcal{G}_2' = \mathbb{R}^{p+2}$. In this case, the AUC-based fairness metric becomes $|Pr(h_{\mathbf{w}}(\boldsymbol{\xi}) > h_{\mathbf{w}}(\boldsymbol{\xi}')|\gamma = 1, \zeta = -1, \zeta' = -1) - 0.5|$. It has the similar interpretation as the positive average equality gap.*

Example 5 (BPSN AUC and BNSP AUC) *When $\mathcal{G}_1 = \{\mathbf{z}|\zeta = 1\}$ and $\mathcal{G}_1' = \{\mathbf{z}|\zeta = -1, \gamma = 1\}$, $AUC_{\mathbf{w}}(\mathcal{G}_1, \mathcal{G}_1')$ becomes the background positive subgroup negative (BPSN) AUC in Borkan et al. (2019). When $\mathcal{G}_2 = \{\mathbf{z}|\zeta = 1, \gamma = 1\}$ and $\mathcal{G}_2' = \{\mathbf{z}|\zeta = -1\}$, $AUC_{\mathbf{w}}(\mathcal{G}_2, \mathcal{G}_2')$ becomes the background negative subgroup positive (BNSP) AUC in Borkan et al. (2019). One fairness metric introduced by Borkan et al. (2019) is the absolute difference between the BPSN AUC and the BNSP AUC, which is exactly (1) w.r.t $\mathcal{G}_1$, $\mathcal{G}_1'$, $\mathcal{G}_2$ and $\mathcal{G}_2'$ chosen above. When this metric is small, the probability of a random positive data point from the whole dataset being ranked above a random negative data point from the protected group is close to the probability of a random positive data point from the protected group being ranked above a random negative data point from the whole dataset.*

B TECHNICAL LEMMAS AND THEIR PROOFS

In this section, we provide some technical lemmas and their proofs.

B.1 Proofs of Lemma 1 and 2

Proof.[of Lemma 1] For simplicity of notation, we directly use $\mathcal{G}$ and $\mathcal{G}'$ to represent the events $\mathbf{z} \in \mathcal{G}$ and $\mathbf{z}' \in \mathcal{G}'$, respectively, when no confusion can be caused. Because $\mathbf{z} = (\boldsymbol{\xi}, \zeta, \gamma)$ and $\mathbf{z}' = (\boldsymbol{\xi}', \zeta, \gamma)$ are i.i.d. data samples, we have $\mathbb{E}\big[G_1(\mathbf{z}) + G_2(\mathbf{z}')|\mathcal{G}, \mathcal{G}'\big] = \mathbb{E}\big[G_1(\mathbf{z})|\mathcal{G}\big] + \mathbb{E}\big[G_2(\mathbf{z}')|\mathcal{G}'\big]$ and $\mathbb{E}\big[G_1(\mathbf{z})G_2(\mathbf{z}')|\mathcal{G}, \mathcal{G}'\big] = \mathbb{E}\big[G_1(\mathbf{z})|\mathcal{G}\big]\mathbb{E}\big[G_2(\mathbf{z}')|\mathcal{G}'\big]$ for any

measurable functions G_1 and G_2. Based on this fact, we have

$$\begin{aligned}
& \mathbb{E}\big[(h_{\mathbf{w}}(\boldsymbol{\xi}) - h_{\mathbf{w}}(\boldsymbol{\xi}') - c_2)^2 | \mathcal{G}, \mathcal{G}'\big] \\
= \; & c_2^2 - 2c_2\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi})|\mathcal{G}\big] + 2c_2\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi}')|\mathcal{G}'\big] + \mathbb{E}\big[(h_{\mathbf{w}}(\boldsymbol{\xi}))^2|\mathcal{G}\big] + \mathbb{E}\big[(h_{\mathbf{w}}(\boldsymbol{\xi}'))^2|\mathcal{G}'\big] - 2\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi})|\mathcal{G}\big]\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi}')|\mathcal{G}'\big] \\
= \; & c_2^2 - 2c_2\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi})|\mathcal{G}\big] + 2c_2\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi}')|\mathcal{G}'\big] + \mathbb{E}\big[(h_{\mathbf{w}}(\boldsymbol{\xi}))^2|\mathcal{G}\big] - (\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi})|\mathcal{G}\big])^2 + \mathbb{E}\big[(h_{\mathbf{w}}(\boldsymbol{\xi}'))^2|\mathcal{G}'\big] - (\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi}')|\mathcal{G}'\big])^2 \\
& + (\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi})|\mathcal{G}\big])^2 + (\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi}')|\mathcal{G}'\big])^2 - 2\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi})|\mathcal{G}\big]\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi}')|\mathcal{G}'\big] \\
= \; & c_2^2 - 2c_2\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi})|\mathcal{G}\big] + 2c_2\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi}')|\mathcal{G}'\big] + \min_a \mathbb{E}\big[(h_{\mathbf{w}}(\boldsymbol{\xi}) - a)^2|\mathcal{G}\big] + \min_b \mathbb{E}\big[(h_{\mathbf{w}}(\boldsymbol{\xi}') - b)^2|\mathcal{G}'\big] \\
& + \max_\alpha \left\{2\alpha\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi})|\mathcal{G}\big] - 2\alpha\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi}')|\mathcal{G}'\big] - \alpha^2\right\} \\
= \; & c_2^2 - \frac{2c_2\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi})\mathbb{I}_{\mathcal{G}}(\mathbf{z})\big]}{\Pr(\mathbf{z}\in\mathcal{G})} + \frac{2c_2\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi}')\mathbb{I}_{\mathcal{G}'}(\mathbf{z}')\big]}{\Pr(\mathbf{z}'\in\mathcal{G}')} + \min_a \frac{\mathbb{E}\big[(h_{\mathbf{w}}(\boldsymbol{\xi}) - a)^2\mathbb{I}_{\mathcal{G}}(\mathbf{z})\big]}{\Pr(\mathbf{z}\in\mathcal{G})} + \min_b \frac{\mathbb{E}\big[(h_{\mathbf{w}}(\boldsymbol{\xi}') - b)^2\mathbb{I}_{\mathcal{G}'}(\mathbf{z}')\big]}{\Pr(\mathbf{z}'\in\mathcal{G}')} \\
& + \max_\alpha \left\{2\alpha\left(\frac{\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi})\mathbb{I}_{\mathcal{G}}(\mathbf{z})\big]}{\Pr(\mathbf{z}\in\mathcal{G})} - \frac{\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi}')\mathbb{I}_{\mathcal{G}'}(\mathbf{z}')\big]}{\Pr(\mathbf{z}'\in\mathcal{G}')}\right) - \alpha^2\right\}. \qquad (30)
\end{aligned}$$

Additionally, given any $\mathbf{w} \in \mathcal{W}$, the optimal value of a, b and α are $\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi})|\mathcal{G}\big]$, $\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi}')|\mathcal{G}'\big]$ and $\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi})|\mathcal{G}\big] - \mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi}')|\mathcal{G}'\big]$, respectively. By the definition of $\mathcal{I}_{\mathcal{G},\mathcal{G}'}$, we can restrict the decision variables a, b and α in $\mathcal{I}_{\mathcal{G},\mathcal{G}'}$ without changing the optimal objective values within (30). The proof is thus completed by multiplying both sides of (30) by c_1 and observing that $\boldsymbol{\xi}'$ and $\mathbf{z}'$ in (30) can be replaced by $\boldsymbol{\xi}$ and $\mathbf{z}$ because they are i.i.d. random variables.

□

Proof.[of Lemma 2] By the assumptions of this lemma, there exists $\mathbf{w}^\dagger \in \mathcal{W}$ such that $h_{\mathbf{w}^\dagger}(\boldsymbol{\xi}) = c$ for any $\boldsymbol{\xi}$. Let $\mathbf{x}^\dagger$ be a solution in $\mathcal{X}$ whose $\mathbf{w}$-component equals $\mathbf{w}^\dagger$ and its remaining components are $a_1^\dagger = a_2^\dagger = b_1^\dagger = b_2^\dagger = a_3^\dagger = a_4^\dagger = b_3^\dagger = b_4^\dagger = c$.

By the definitions of $F_{\mathcal{G},\mathcal{G}'}(\mathbf{w}, a, b; \mathbf{z})$ and $G_{\mathcal{G},\mathcal{G}'}(\mathbf{w}; \mathbf{z})$ in (8), we have

$$\begin{aligned}
\mathbb{E}[F_{\mathcal{G}_1',\mathcal{G}_1}(\mathbf{x}^\dagger;\mathbf{z})] &= \mathbb{E}\left[c_1c_2^2 - \frac{2c_1c_2c\mathbb{I}_{\mathcal{G}_1'}(\mathbf{z})}{Pr(\mathbf{z}\in\mathcal{G}_1')} + \frac{2c_1c_2c\mathbb{I}_{\mathcal{G}_1}(\mathbf{z})}{Pr(\mathbf{z}\in\mathcal{G}_1)} + \frac{c_1(c-a_1^\dagger)^2\mathbb{I}_{\mathcal{G}_1'}(\mathbf{z})}{Pr(\mathbf{z}\in\mathcal{G}_1')} + \frac{c_1(c-b_1^\dagger)^2\mathbb{I}_{\mathcal{G}_1}(\mathbf{z})}{Pr(\mathbf{z}\in\mathcal{G}_1)}\right] = c_1c_2^2, \\
\mathbb{E}[G_{\mathcal{G}_1',\mathcal{G}_1}(\mathbf{w}^\dagger;\mathbf{z})] &= \mathbb{E}\left[\frac{c_1c\mathbb{I}_{\mathcal{G}_1'}(\mathbf{z})}{Pr(\mathbf{z}\in\mathcal{G}_1')} - \frac{c_1c\mathbb{I}_{\mathcal{G}_1'}(\mathbf{z})}{Pr(\mathbf{z}\in\mathcal{G}_1)}\right] = 0, \\
\mathbb{E}[F_{\mathcal{G}_2,\mathcal{G}_2'}(\mathbf{x}^\dagger;\mathbf{z})] &= \mathbb{E}\left[c_1c_2^2 - \frac{2c_1c_2c\mathbb{I}_{\mathcal{G}_2}(\mathbf{z})}{Pr(\mathbf{z}\in\mathcal{G}_2)} + \frac{2c_1c_2c\mathbb{I}_{\mathcal{G}_2'}(\mathbf{z})}{Pr(\mathbf{z}\in\mathcal{G}_2')} + \frac{c_1(c-a_2^\dagger)^2\mathbb{I}_{\mathcal{G}_2}(\mathbf{z})}{Pr(\mathbf{z}\in\mathcal{G}_2)} + \frac{c_1(c-b_2^\dagger)^2\mathbb{I}_{\mathcal{G}_2'}(\mathbf{z})}{Pr(\mathbf{z}\in\mathcal{G}_2')}\right] = c_1c_2^2, \\
\mathbb{E}[G_{\mathcal{G}_2,\mathcal{G}_2'}(\mathbf{w}^\dagger;\mathbf{z})] &= \mathbb{E}\left[\frac{c_1c\mathbb{I}_{\mathcal{G}_2}(\mathbf{z})}{Pr(\mathbf{z}\in\mathcal{G}_2)} - \frac{c_1c\mathbb{I}_{\mathcal{G}_2'}(\mathbf{z})}{Pr(\mathbf{z}\in\mathcal{G}_2')}\right] = 0.
\end{aligned}$$

Since $c_1c_2^2 \le 0.5$, applying the equations above to the definitions of $f_1(\mathbf{x})$ in (12) and (13) leads to $f_1(\mathbf{x}^\dagger) = 2c_1c_2^2 \le 1 < 1+\kappa$. Similarly, it holds that $f_2(\mathbf{x}^\dagger) < 1+\kappa$. This means $\mathbf{x}^\dagger$ is a strictly feasible solution and Assumption 3 holds. □

B.2 Closed-Form Solutions for (25) and (26)

The closed form of $\mathbf{x}^{(t+1)}$ is obvious so we only show the closed form of $\mathbf{y}^{(t+1)}$ in (25). Given any $\tau > 0$, $\mathbf{v} = (v_0, v_1, v_2, v_3, v_4) \in \mathbb{R}^5$, $\mathbf{u} = (u_0, u_1, u_2) \in \mathbb{R}^3$ and $\mathbf{y}' = (\tilde{\mathbf{y}}', \tilde{\boldsymbol{\alpha}}') \in \mathcal{Y}$, we consider the following problem

$$\mathbf{y}^\# = (\tilde{\mathbf{y}}^\#, \tilde{\boldsymbol{\alpha}}^\#) = \operatorname*{arg\,min}_{\mathbf{y}=(\tilde{\mathbf{y}},\tilde{\boldsymbol{\alpha}})\in\mathcal{Y}} -(\mathbf{u})^\top\tilde{\mathbf{y}} - (\mathbf{v})^\top\tilde{\boldsymbol{\alpha}} + \frac{V_y(\mathbf{y},\mathbf{y}')}{\tau} + d_y(\mathbf{y}), \qquad (31)$$

which becomes (25) after setting $(\mathbf{u}, \mathbf{v}) = \mathbf{g}_{\mathbf{y}}^{(t)}$, $\tau = \tau_t$ and $\mathbf{y}' = \mathbf{y}^{(t)}$. The following lemma characterizes the closed form of $\mathbf{y}^\#$.

Lemma 3 *Let* $\alpha_0' := \frac{\tilde{\alpha}_0'}{\tilde{y}_0'}$, $\alpha_1' := \frac{\tilde{\alpha}_1'}{\tilde{y}_1'}$, $\alpha_2' := \frac{\tilde{\alpha}_2'}{\tilde{y}_1'}$, $\alpha_3' := \frac{\tilde{\alpha}_3'}{\tilde{y}_2'}$, $\alpha_4' := \frac{\tilde{\alpha}_4'}{\tilde{y}_2'}$ *and let*

$$\mu_i := \min_{\alpha_i\in\mathcal{I}} \left\{ -\alpha_i v_i + \alpha_i^2 + \tfrac{1}{\tau}(\alpha_i - \alpha_i')^2 \right\} \quad \textit{and} \quad \alpha_i^\# := \operatorname*{arg\,min}_{\alpha_i\in\mathcal{I}} \left\{ -\alpha_i v_i + \alpha_i^2 + \tfrac{1}{\tau}(\alpha_i - \alpha_i')^2 \right\}. \qquad (32)$$

for $i = 0, 1, \ldots, 4$. *Let* $\pi_0 := (\tilde{y}'_0)\exp\left(-\frac{\mu_0 - u_0}{2(1+\sqrt{2}I)^2(1/\tau)}\right)$, $\pi_1 := (\tilde{y}'_1)\exp\left(-\frac{\mu_1+\mu_2-u_1}{2(1+\sqrt{2}I)^2(1/\tau)}\right)$ *and* $\pi_2 := (\tilde{y}'_2)\exp\left(-\frac{\mu_3+\mu_4-u_2}{2(1+\sqrt{2}I)^2(1/\tau)}\right)$. *Then,* $\mathbf{y}^\# = (\tilde{\mathbf{y}}^\#, \tilde{\boldsymbol{\alpha}}^\#) \in \mathcal{Y}$ *defined as follows is an optimal solution to* (31)*:*

$$\tilde{y}_i^\# := \frac{\pi_i}{\pi_0 + \pi_1 + \pi_2} \quad \textit{for} \quad i = 0, 1, 2.$$
$$\tilde{\alpha}_0^\# := \tilde{y}_0^\# \alpha_0^\#, \quad \tilde{\alpha}_1^\# := \tilde{y}_1^\# \alpha_1^\#, \quad \tilde{\alpha}_2^\# := \tilde{y}_1^\# \alpha_2^\#, \quad \tilde{\alpha}_3^\# := \tilde{y}_2^\# \alpha_3^\#, \quad \tilde{\alpha}_4^\# := \tilde{y}_2^\# \alpha_4^\#.$$

Proof. Recall the definitions of $V_y(\mathbf{y}, \mathbf{y}')$ in (24) and $d_y(\mathbf{y})$ in (21). (31) can be formulated as

$$\min_{\mathbf{y}\in\mathcal{Y}} \left\{ \begin{array}{c} -(\mathbf{u})^\top \tilde{\mathbf{y}} - (\mathbf{v})^\top \tilde{\boldsymbol{\alpha}} + \frac{2(1+\sqrt{2}I)^2}{\tau} \sum_{i=0}^2 \tilde{y}_i \ln(\frac{\tilde{y}_i}{\tilde{y}'_i}) \\ + \frac{\tilde{y}_0}{\tau}(\frac{\tilde{\alpha}_0}{\tilde{y}_0} - \frac{\tilde{\alpha}'_0}{\tilde{y}'_0})^2 + \frac{\tilde{y}_1}{\tau}(\frac{\tilde{\alpha}_1}{\tilde{y}_1} - \frac{\tilde{\alpha}'_1}{\tilde{y}'_1})^2 + \frac{\tilde{y}_1}{\tau}(\frac{\tilde{\alpha}_2}{\tilde{y}_1} - \frac{\tilde{\alpha}'_2}{\tilde{y}'_1})^2 + \frac{\tilde{y}_2}{\tau}(\frac{\tilde{\alpha}_3}{\tilde{y}_2} - \frac{\tilde{\alpha}'_3}{\tilde{y}'_2})^2 + \frac{\tilde{y}_2}{\tau}(\frac{\tilde{\alpha}_4}{\tilde{y}_2} - \frac{\tilde{\alpha}'_4}{\tilde{y}'_2})^2 \\ + \frac{\tilde{\alpha}_0^2}{\tilde{y}_0} + \frac{\tilde{\alpha}_1^2}{\tilde{y}_1} + \frac{\tilde{\alpha}_2^2}{\tilde{y}_1} + \frac{\tilde{\alpha}_3^2}{\tilde{y}_2} + \frac{\tilde{\alpha}_4^2}{\tilde{y}_2} \end{array} \right\}. \tag{33}$$

We first fix $\tilde{\mathbf{y}} \in \triangle_3$ and only optimize $\tilde{\boldsymbol{\alpha}}$ in (33) subject to constraints $\tilde{\alpha}_0 \in \tilde{y}_0 \cdot \mathcal{I}$, $\tilde{\alpha}_1 \in \tilde{y}_1 \cdot \mathcal{I}$, $\tilde{\alpha}_2 \in \tilde{y}_1 \cdot \mathcal{I}$, $\tilde{\alpha}_3 \in \tilde{y}_2 \cdot \mathcal{I}$ and $\tilde{\alpha}_4 \in \tilde{y}_2 \cdot \mathcal{I}$. By changing variables using $\alpha_0 := \frac{\tilde{\alpha}_0}{\tilde{y}_0}$, $\alpha_1 := \frac{\tilde{\alpha}_1}{\tilde{y}_1}$, $\alpha_2 := \frac{\tilde{\alpha}_2}{\tilde{y}_1}$, $\alpha_3 := \frac{\tilde{\alpha}_3}{\tilde{y}_2}$, $\alpha_4 := \frac{\tilde{\alpha}_4}{\tilde{y}_2}$ and $\alpha'_0 := \frac{\tilde{\alpha}'_0}{\tilde{y}'_0}$, $\alpha'_1 := \frac{\tilde{\alpha}'_1}{\tilde{y}'_1}$, $\alpha'_2 := \frac{\tilde{\alpha}'_2}{\tilde{y}'_1}$, $\alpha'_3 := \frac{\tilde{\alpha}'_3}{\tilde{y}'_2}$, $\alpha'_4 := \frac{\tilde{\alpha}'_4}{\tilde{y}'_2}$, (33) becomes

$$\min_{\tilde{\mathbf{y}}\in\triangle_3} \left\{ \begin{array}{c} -(\mathbf{u})^\top \tilde{\mathbf{y}} + \frac{2(1+\sqrt{2}I)^2}{\tau} \sum_{i=0}^2 \tilde{y}_i \ln(\frac{\tilde{y}_i}{\tilde{y}'_i}) + \tilde{y}_0 \min_{\alpha_0\in\mathcal{I}} \left[-\alpha_0 v_0 + \frac{1}{\tau}(\alpha_0 - \alpha'_0) + \alpha_0^2\right] \\ + \tilde{y}_1 \min_{\alpha_1,\alpha_2\in\mathcal{I}} \left[\sum_{i=1}^2 -\alpha_i v_i + \frac{1}{\tau}(\alpha_i - \alpha'_i) + \alpha_i^2\right] + \tilde{y}_2 \min_{\alpha_3,\alpha_4\in\mathcal{I}} \left[\sum_{i=3}^4 -\alpha_i v_i + \frac{1}{\tau}(\alpha_i - \alpha'_i) + \alpha_i^2\right] \end{array} \right\} \tag{34}$$

$$= \min_{\tilde{\mathbf{y}}\in\triangle_3} \left\{ -(\mathbf{u})^\top \tilde{\mathbf{y}} + \frac{2(1+\sqrt{2}I)^2}{\tau} \sum_{i=0}^2 \tilde{y}_i \ln(\frac{\tilde{y}_i}{\tilde{y}'_i}) + \tilde{y}_0\mu_0 + \tilde{y}_1(\mu_1+\mu_2) + \tilde{y}_2(\mu_3+\mu_4) \right\}, \tag{35}$$

according to the definition of μ_i in (32).

Equality (34) above indicates that the minimization over $\tilde{\boldsymbol{\alpha}}$ in (33) for a given $\tilde{\mathbf{y}}$ is equivalent to the inner minimization over $\boldsymbol{\alpha}$ in (34), which is independent of $\tilde{\mathbf{y}}$ and can be solved for each i separately. Note that the optimal objective value and the solution of the ith inner minimization are μ_i and $\alpha_i^\#$ in (32), where $\alpha_i^\#$ has a closed form. Equality (35) indicates that, after obtaining the optimal α_i, we can solve the optimal $\tilde{\mathbf{y}}$ by solving the outer minimization problem (35) whose solution is exactly $\tilde{\mathbf{y}}^\#$ defined in Lemma 3 which can be verified from the optimality conditions. According to the relationship between α_i and $\tilde{\alpha}_i$, the optimal value of the original variable $\tilde{\alpha}_i$ is exactly $\tilde{\alpha}_i^\#$ defined in Lemma 3.

□

Next, we consider the optimal value $U(r)$ in (26). According to the definition of Φ in (22), (26) can be written as

$$U(r) = \max_{\mathbf{y}=(\tilde{\mathbf{y}},\tilde{\boldsymbol{\alpha}})\in\mathcal{Y}} (\mathbf{u})^\top \tilde{\mathbf{y}} + (\mathbf{v})^\top \tilde{\boldsymbol{\alpha}} - d_y(\mathbf{y}) + \frac{\hat{\rho}}{2} \frac{\sum_{t=0}^{T-1} \|\mathbf{x}^{(t)} - \tilde{\mathbf{x}}\|_2^2}{\sum_{t=0}^{T-1} \tau_t}, \tag{36}$$

where

$$\mathbf{u} = \frac{\sum_{t=0}^{T-1} \mathbf{F}(\mathbf{x}^{(t)}, \mathbf{z}^{(t)})}{\sum_{t=0}^{T-1} \tau_t} \quad \text{and} \quad \mathbf{v} = \frac{\sum_{t=0}^{T-1} \mathbf{G}(\mathbf{w}^{(t)}, \mathbf{z}^{(t)})}{\sum_{t=0}^{T-1} \tau_t}.$$

We denote each component of $\mathbf{u}$ and $\mathbf{v}$ as $\mathbf{v} = (v_0, v_1, v_2, v_3, v_4) \in \mathbb{R}^5$ and $\mathbf{u} = (u_0, u_1, u_2) \in \mathbb{R}^3$. The following lemma provides a closed form to $U(r)$.

Lemma 4 *Let* $U(r)$ *defined in* (26)*, or equivalently, in* (36)*. We have*

$$U(r) := \max\{u_0 + \mu_0, u_1 + \mu_1 + \mu_2, u_2 + \mu_3 + \mu_4\} + \frac{\hat{\rho}}{2} \frac{\sum_{t=0}^{T-1} \|\mathbf{x}^{(t)} - \tilde{\mathbf{x}}\|_2^2}{\sum_{t=0}^{T-1} \tau_t},$$

where $\mu_i := \max_{\alpha_i\in\mathcal{I}} \{\alpha_i v_i - \alpha_i^2\}$ *for* $i = 0, 1, \ldots, 4$.

Proof. Recall the definitions of $d_y(\mathbf{y})$ in (21) and $d_x(\mathbf{x}) = \frac{\hat{\rho}}{2}\|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2$. (36) can be formulated as

$$U(r) = \max_{\mathbf{y}\in\mathcal{Y}} \left\{ \ (\mathbf{u})^\top \tilde{\mathbf{y}} + (\mathbf{v})^\top \tilde{\boldsymbol{\alpha}} - \frac{\tilde{\alpha}_0^2}{\tilde{y}_0} - \frac{\tilde{\alpha}_1^2}{\tilde{y}_1} - \frac{\tilde{\alpha}_2^2}{\tilde{y}_1} - \frac{\tilde{\alpha}_3^2}{\tilde{y}_2} - \frac{\tilde{\alpha}_4^2}{\tilde{y}_2} \ \right\} + \frac{\hat{\rho}}{2} \frac{\sum_{t=0}^{T-1} \|\mathbf{x}^{(t)} - \tilde{\mathbf{x}}\|_2^2}{\sum_{t=0}^{T-1} \tau_t}. \tag{37}$$

Similar to the proof of Lemma 3, we first fix $\tilde{\mathbf{y}} \in \triangle_3$ and only optimize $\tilde{\boldsymbol{\alpha}}$ in (37) subject to constraints $\tilde{\alpha}_0 \in \tilde{y}_0 \cdot \mathcal{I}$, $\tilde{\alpha}_1 \in \tilde{y}_1 \cdot \mathcal{I}$, $\tilde{\alpha}_2 \in \tilde{y}_1 \cdot \mathcal{I}$, $\tilde{\alpha}_3 \in \tilde{y}_2 \cdot \mathcal{I}$ and $\tilde{\alpha}_4 \in \tilde{y}_2 \cdot \mathcal{I}$. By changing variables using $\alpha_0 := \frac{\tilde{\alpha}_0}{\tilde{y}_0}$, $\alpha_1 := \frac{\tilde{\alpha}_1}{\tilde{y}_1}$, $\alpha_2 := \frac{\tilde{\alpha}_2}{\tilde{y}_1}$, $\alpha_3 := \frac{\tilde{\alpha}_3}{\tilde{y}_2}$ and $\alpha_4 := \frac{\tilde{\alpha}_4}{\tilde{y}_2}$, (37) becomes

$$\begin{aligned}
U(r) &= \max_{\tilde{\mathbf{y}}\in\triangle_3} \left\{ (\mathbf{u})^\top \tilde{\mathbf{y}} + \tilde{y}_0 \max_{\alpha_0 \in \mathcal{I}} \left[\alpha_0 v_0 - \alpha_0^2\right] + \tilde{y}_1 \max_{\alpha_1,\alpha_2\in\mathcal{I}^2} \left[\sum_{i=1}^{2} \alpha_i v_i - \alpha_i^2\right] + \tilde{y}_2 \min_{\alpha_3,\alpha_4\in\mathcal{I}^2} \left[\sum_{i=3}^{4} \alpha_i v_i - \alpha_i^2\right] \right\} \\
&\quad + \frac{\hat{\rho}}{2} \frac{\sum_{t=0}^{T-1} \|\mathbf{x}^{(t)} - \tilde{\mathbf{x}}\|_2^2}{\sum_{t=0}^{T-1} \tau_t} \\
&= \max_{\tilde{\mathbf{y}}\in\triangle_3} \left\{ (\mathbf{u})^\top \tilde{\mathbf{y}} + \tilde{y}_0\mu_0 + \tilde{y}_1(\mu_1 + \mu_2) + \tilde{y}_2(\mu_3 + \mu_4) \right\} + \frac{\hat{\rho}}{2} \frac{\sum_{t=0}^{T-1} \|\mathbf{x}^{(t)} - \tilde{\mathbf{x}}\|_2^2}{\sum_{t=0}^{T-1} \tau_t} \\
&= \max\{u_0 + \mu_0, u_1 + \mu_1 + \mu_2, u_2 + \mu_3 + \mu_4\} + \frac{\hat{\rho}}{2} \frac{\sum_{t=0}^{T-1} \|\mathbf{x}^{(t)} - \tilde{\mathbf{x}}\|_2^2}{\sum_{t=0}^{T-1} \tau_t},
\end{aligned}$$

where the second equality is because of the definition of μ_i for $i = 0, \ldots, 4$ and the last equality is because $\tilde{\mathbf{y}} \in \triangle_3$. □

B.3 A Sufficient Condition for Assumption 4.2

In this subsection, we present the following sufficient condition for Assumption 4.2 to hold.

Lemma 5 *Assumption 4.2 holds if* $2c_1c_2^2 - 1 < \kappa$, $\rho < \frac{2(\kappa+1-2c_1c_2^2)}{\max_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} \|\mathbf{x}-\mathbf{x}'\|_2^2}$ *and there exists* $\mathbf{w} \in \mathcal{W}$ *such that* $h_{\mathbf{w}}(\cdot)$ *is a constant mapping.*

Proof. Because $2c_1c_2^2 - 1 < \kappa$ and $\rho < \frac{2(\kappa+1-2c_1c_2^2)}{\max_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} \|\mathbf{x}-\mathbf{x}'\|_2^2}$, there exists ρ_ϵ such that

$$0 < \rho_\epsilon < \frac{2(\kappa + 1 - 2c_1c_2^2)}{\max_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} \|\mathbf{x} - \mathbf{x}'\|_2^2} - \rho. \tag{38}$$

Let $\mathbf{x} \in \mathcal{X}$ be any solution that satisfies $\max_{i=1,2} f_i(\mathbf{x}) - 1 - \kappa \le \epsilon^2$. By the assumptions, there exists $\mathbf{w}^\dagger \in \mathcal{W}$ such that $h_{\mathbf{w}^\dagger}(\boldsymbol{\xi})$ is a constant over $\boldsymbol{\xi}$, denoted by c. Let $\mathbf{x}^\dagger$ be a solution in $\mathcal{X}$ whose $\mathbf{w}$-component equals $\mathbf{w}^\dagger$ and its remaining components are $a_1^\dagger = a_2^\dagger = b_1^\dagger = b_2^\dagger = a_3^\dagger = a_4^\dagger = b_3^\dagger = b_4^\dagger = c$.

According to the proof of Lemma 2 in Section B.1, we have $f_i(\mathbf{x}^\dagger) = 2c_1c_2^2$ for $i = 1$ and 2 and, according to the assumption of this lemma, we have $f_i(\mathbf{x}^\dagger) < 1 + \kappa$ for $i = 1$ and 2. This implies

$$\begin{aligned}
& \min_{\mathbf{x}'\in\mathcal{X}} \left\{ \max_{i=1,2} f_i(\mathbf{x}') - 1 - \kappa + \frac{\rho + \rho_\epsilon}{2} \|\mathbf{x}' - \mathbf{x}\|_2^2 \right\} \\
\le \ & \max_{i=1,2} f_i(\mathbf{x}^\dagger) - 1 - \kappa + \frac{\rho + \rho_\epsilon}{2} \|\mathbf{x}^\dagger - \mathbf{x}\|_2^2 \\
\le \ & 2c_1c_2^2 - 1 - \kappa + \frac{\rho + \rho_\epsilon}{2} \max_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} \|\mathbf{x} - \mathbf{x}'\|_2^2 = -\sigma_\epsilon,
\end{aligned}$$

where $\sigma_\epsilon := \kappa + 1 - 2c_1c_2^2 - \frac{\rho+\rho_\epsilon}{2} \max_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} \|\mathbf{x} - \mathbf{x}'\|_2^2$ is a positive number because of (38). This completes the proof. □

Remark 2 *Condition* $2c_1c_2^2 - 1 < \kappa$ *means the targeted fairness level should not be too small, so there exists a sufficiently feasible solution (see Lemma 2). Condition* $\rho < \frac{2(\kappa+1-2c_1c_2^2)}{\max_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} \|\mathbf{x}-\mathbf{x}'\|_2^2}$ *means the original non-convex problem should not have a high level of non-convexity.*

C DISCUSSION ON THEOREM 1

In this section, we briefly discuss how to directly apply the results from Lin et al. (2020); Nemirovski et al. (2009) to obtain Theorem 1. First, we match our notation to those used in Lin et al. (2020) and instantize the convergence results

in Lin et al. (2020) on (23). Recall that $\|\mathbf{x}\|_x = \|\mathbf{x}\|_2$ and $\|\mathbf{y}\|_y = \sqrt{\|\tilde{\mathbf{y}}\|_1^2 + \|\tilde{\boldsymbol{\alpha}}\|_2^2}$. Their dual norms are $\|\mathbf{x}\|_{*,x} = \|\mathbf{x}\|_2$ and $\|\mathbf{y}\|_{*,y} = \sqrt{\|\tilde{\mathbf{y}}\|_\infty^2 + \|\tilde{\boldsymbol{\alpha}}\|_2^2}$, respectively. The complexity of SMD is known to depend on the diameters of $\mathcal{X}$ and $\mathcal{Y}$ measured by the corresponding distance generating functions, namely,

$$D_x := \sqrt{\max_{\mathbf{x}\in\mathcal{X}} \omega_x(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{X}} \omega_x(\mathbf{x})} \text{ and } D_y := \sqrt{\max_{\mathbf{y}\in\mathcal{Y}} \omega_y(\mathbf{y}) - \min_{\mathbf{y}\in\mathcal{Y}} \omega_y(\mathbf{y})}$$

defined in Theorem 1. Moreover, thanks to Assumption 2, it is not hard to show that there exist constants M_x, M_y and Q, which only depend on σ, I, ρ and D_x, such that

$$\mathbb{E}\left[\exp(\|\nabla_x \Phi(\mathbf{x},\mathbf{y},\mathbf{z})\|_{*,x}^2 / M_x^2)\right] \le \exp(1), \tag{39}$$

$$\mathbb{E}\left[\exp(\|\nabla_y \Phi(\mathbf{x},\mathbf{y},\mathbf{z})\|_{*,y}^2 / M_y^2)\right] \le \exp(1), \tag{40}$$

$$\mathbb{E}\left[\exp(|\Phi(\mathbf{x},\mathbf{y},\mathbf{z}) - \phi(\mathbf{x},\mathbf{y})|^2 / Q^2)\right] \le \exp(1), \tag{41}$$

Additionally, given $\delta \in (0,1)$, we define

$$M := \sqrt{2D_x^2 M_x^2 + 2D_y^2 M_y^2}; \tag{42}$$

$$\Omega(\delta) := \max\left\{\sqrt{12\ln\left(\frac{24}{\delta}\right)}, \frac{4}{3}\ln\left(\frac{24}{\delta}\right)\right\}. \tag{43}$$

With those notations, a brief proof of Theorem 1 is given below.

Compared with problem (5) in Lin et al. (2020), our problem (23) has the additional terms $d_x(\mathbf{x})$ and $d_y(\mathbf{y})$. However, since we choose the initial solution as $\mathbf{x}^{(0)} = \mathbf{0}$ and $\mathbf{y}^{(0)} = (1/3, 1/3, 1/3, \mathbf{0})$, these additional terms can be eliminated from the proof of any theorems and propositions in Lin et al. (2020), so the convergence results in Lin et al. (2020) also hold for problem (23) and Algorithm 2. Moreover, the algorithm in Lin et al. (2020) is presented using a unified updating scheme for $\mathbf{x}$ and $\mathbf{y}$ with only one step size γ_t while our Algorithm 2 is presented with $\mathbf{x}$ and $\mathbf{y}$ updated separately. However, it is easy to verify that, by choosing $\eta_t = 2D_x^2\gamma_t$ and $\tau_t = 2D_x^2\gamma_t$ with $\gamma_t = 1/(M\sqrt{t+1})$ where M is defined in (42), Algorithm 2 is equivalent to the algorithm in Lin et al. (2020). Hence, according to Theorem 8 in Lin et al. (2020), if

$$T \ge T(\delta,\epsilon_{\mathcal{A}}) := \max\left\{6, \left(\frac{16\left(Q\Omega(\delta) + 10M\Omega(\delta) + 4.5M\right)}{\epsilon_{\mathcal{A}}}\ln\left(\frac{8\left(Q\Omega(\delta) + 10M\Omega(\delta) + 4.5M\right)}{\epsilon_{\mathcal{A}}}\right)\right)^2 - 2\right\}, \tag{44}$$

the outputs $U(r)$ and $\bar{\mathbf{x}}$ by Algorithm 2 satisfy the inequalities $\mathcal{P}(r,\bar{\mathbf{x}}) - H(r) \le \epsilon$ and $|U(r) - H(r)| \le \epsilon$ with a probability of at least $1-\delta$ for any $r > f^*$. Hence, SMD with $T \ge T(\delta,\epsilon_{\mathcal{A}})$ is a valid stochastic oracle defined in Definition 3. Hence, according to Corollary 9 in Lin et al. (2020), SFLS returns a relative ϵ-optimal and feasible solution with probability of at least $1-\delta$ using at most $\tilde{O}\left(\frac{1}{\epsilon^2}\ln(\frac{1}{\delta})\right)$ stochastic mirror descent steps across all calls of SMD. Theorem 1 is thus proved.

D DEFINITION of $\tilde{\phi}$ IN (19) AND TABLE OF NOTATIONS

In Section 4, we can write (18) as

$$H(r) := \min_{\mathbf{x}\in\mathcal{X}} \mathcal{P}(r,\mathbf{x}) = \min_{\mathbf{x}\in\mathcal{X}} \max_{\tilde{\mathbf{y}}\in\Delta_3} \left\{\tilde{y}_0(f_0(\mathbf{x}) - r) + \tilde{y}_1(f_1(\mathbf{x}) - 1 - \kappa) + \tilde{y}_2(f_2(\mathbf{x}) - 1 - \kappa)\right\},$$

where $\Delta_3 := \{\tilde{\mathbf{y}} \in \mathbb{R}_+^3 | \sum_{i=0}^2 \tilde{y}_i = 1\}$. With (11), (12) and (13), we can reformulate the problem above into (19), i.e.,

$$H(r) := \min_{\mathbf{x}\in\mathcal{X}} \max_{\tilde{\mathbf{y}}\in\Delta_3, \boldsymbol{\alpha}\in\mathcal{I}^5} \tilde{\phi}(\mathbf{x},\tilde{\mathbf{y}},\boldsymbol{\alpha}),$$

where

$$\tilde{\phi}(\mathbf{x},\tilde{\mathbf{y}},\boldsymbol{\alpha}) := \mathbb{E}\left\{\begin{array}{c} -r\tilde{y}_0 - (1+\kappa)\tilde{y}_1 - (1+\kappa)\tilde{y}_2 \\ +\tilde{y}_0 F_{\mathcal{D}_+,\mathcal{D}_-}(\mathbf{x};\mathbf{z}) + \tilde{y}_1 F_{\mathcal{G}_1',\mathcal{G}_1}(\mathbf{x};\mathbf{z}) + \tilde{y}_1 F_{\mathcal{G}_2,\mathcal{G}_2'}(\mathbf{x};\mathbf{z}) + \tilde{y}_2 F_{\mathcal{G}_2',\mathcal{G}_2}(\mathbf{x};\mathbf{z}) + \tilde{y}_2 F_{\mathcal{G}_1,\mathcal{G}_1'}(\mathbf{x};\mathbf{z}) \\ +\tilde{y}_0\alpha_0 G_{\mathcal{D}_+,\mathcal{D}_-}(\mathbf{w};\mathbf{z}) + \tilde{y}_1\alpha_1 G_{\mathcal{G}_1',\mathcal{G}_1}(\mathbf{w};\mathbf{z}) + \tilde{y}_1\alpha_2 G_{\mathcal{G}_2,\mathcal{G}_2'}(\mathbf{w};\mathbf{z}) + \tilde{y}_2\alpha_3 G_{\mathcal{G}_2',\mathcal{G}_2}(\mathbf{w};\mathbf{z}) + \tilde{y}_2\alpha_4 G_{\mathcal{G}_1,\mathcal{G}_1'}(\mathbf{w};\mathbf{z}) \\ -\tilde{y}_0\alpha_0^2 - \tilde{y}_1\alpha_1^2 - \tilde{y}_1\alpha_2^2 - \tilde{y}_2\alpha_3^2 - \tilde{y}_2\alpha_4^2 \end{array}\right\}.$$

Since many notations are introduced this paper, we summarize them in Table 1 so readers can find their meanings more easily.

Table 1: Notation used throughout the paper.

Symbol	Definition
$\boldsymbol{\xi}$	Feature vector of a data point.
ζ	Binary label of a data point.
γ	Binary sensitive feature of a data point.
$\mathbf{z} = (\boldsymbol{\xi}, \zeta, \gamma)$	A data point.
$\mathbf{w}$ and $\mathcal{W}$	Parameters of a classification model. It belongs to a convex compact set $\in \mathcal{W}$.
$h_{\mathbf{w}}(\boldsymbol{\xi})$	Predicted score for a data point based its feature $\boldsymbol{\xi}$.
$\mathcal{G}, \mathcal{G}_1, \mathcal{G}_1', \mathcal{G}_2, \mathcal{G}_2'$	Set in $\mathbb{R}^{p+2}$ with positive measures w.r.t. $\mathbf{z}$.
$\mathcal{D}_+$	Positive dataset.
$\mathcal{D}_-$	Negative dataset.
$\ell(\cdot)$	Surrogate loss function that approximates $\mathbb{I}_{(\cdot\leq 0)}$ and $\mathbb{I}_{(\cdot<0)}$.
$c_1(\cdot - c_2)^2$	Quadratic loss function that approximates $\mathbb{I}_{(\cdot\leq 0)}$ and $\mathbb{I}_{(\cdot<0)}$.
a, b, α	Auxiliary variables introduced to formulate the quadratic loss into a min-max problem (7).
$\mathcal{I}_{\mathcal{G},\mathcal{G}'}$	The smallest interval that contains $\left\{ \begin{array}{c} 0, \pm\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi})\|\mathbf{z}\in\mathcal{G}\big], \pm\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi}')\|\mathbf{z}'\in\mathcal{G}'\big], \\ \pm\left(\mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi})\|\mathbf{z}\in\mathcal{G}\big] - \mathbb{E}\big[h_{\mathbf{w}}(\boldsymbol{\xi}')\|\mathbf{z}'\in\mathcal{G}'\big]\right) \end{array} \right\}$.
$\mathcal{I}$ and I	A bounded interval containing $\mathcal{I}_{\mathcal{D}_+,\mathcal{D}_-}, \mathcal{I}_{\mathcal{G}_1,\mathcal{G}_1'}, \mathcal{I}_{\mathcal{G}_2,\mathcal{G}_2'}$ and $I := \max_{\alpha\in\mathcal{I}}\|\alpha\|$.
$\mathcal{X}$	The domain of primal variables.
$\mathcal{Y}$	The domain of dual variables.
Δ_3	The simplex in $\mathbb{R}^3$.
$\omega_x(\mathbf{x})$ and $\omega_y(\mathbf{x})$	Distance generating functions on $\mathcal{X}$ and $\mathcal{Y}$, respectively.
$V_x(\mathbf{x}, \mathbf{x}')$ and $V_y(\mathbf{y}, \mathbf{y}')$	Bregman divergences on $\mathcal{X}$ and $\mathcal{Y}$, respectively.
$H(r)$ and $\widehat{H}(r)$	Level-set functions of (10) and (28), respectively.
r and $r^{(k)}$	Level parameters in the stochastic level-set method.
ρ and $\hat{\rho}$	Weak convexity parameter of (10) and $\hat{\rho} > \rho$.

E ADDITIONAL MATERIALS FOR NUMERICAL EXPERIMENTS

In this section, we present some additional details of our numerical experiments in Section 6.

E.1 Details of Datasets

We provide below some details about the three datasets we used in our numerical experiments.

- The *a9a* dataset is used to predict if the annual income of an individual exceeds \$50K. Gender is the sensitive attribute, i.e., female ($\gamma = 1$) or male ($\gamma = -1$).
- The *bank* dataset is used to predict if a client will subscribe a term deposit. Age is the sensitive attribute, i.e., age between 25 and 60 ($\gamma = 1$) or otherwise ($\gamma = -1$).
- The *COMPAS* dataset is used to predict if a criminal defendant will reoffend. Race is the sensitive attribute, i.e., caucasian ($\gamma = 1$) or non-caucasian ($\gamma = -1$).

Some statistics of these datasets are given in Table 2. Data *a9a* originally has a training set and a testing set, and we further split the training data into a training set (%90) and a validation set (%90). For *bank* and *COMPAS* datasets, we split them into training (%60), validation (%20) and testing (%20) sets. The validation sets are used for tuning hyper-parameters while the testing sets are for performance evaluation.

E.2 Details of Baselines

In this section, we provide the details of three baselines used in our experiments.

- Proxy-Lagrangian is a Lagrangian method for solving (2), where only the indicator function $\mathbb{I}_{(h_{\mathbf{w}}(\boldsymbol{\xi})-h_{\mathbf{w}}(\boldsymbol{\xi}')\leq 0)}$ in the objective function is approximated by a surrogate loss while the indicator functions in the constraints are unchanged.

Table 2: Statistics of the datasets.

Datasets	#Instances	#Attributes	Class Label	Sensitive Attribute
a9a	48,842	123	Income	Gender
bank	41,188	54	Subscription	Age
COMPAS	11,757	14	Recidivism	Race

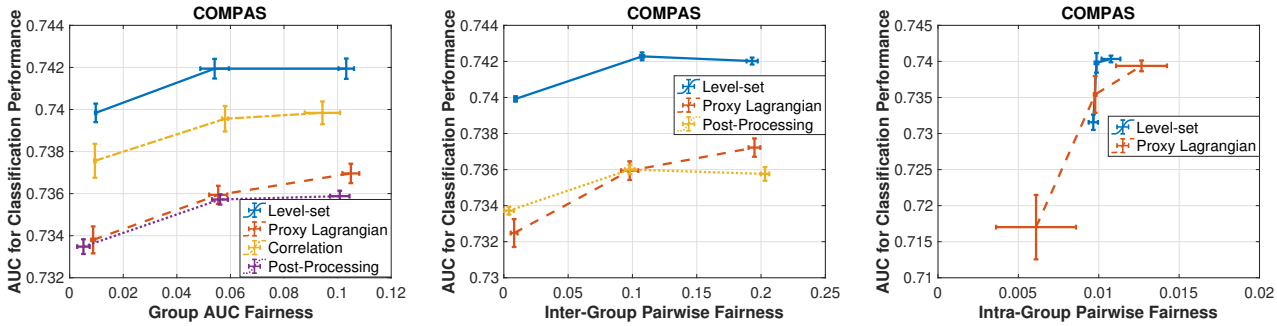

Figure 3: Pareto frontiers by each method for COMPAS dataset in convex case.

- Correlation-penalty is a method that adds the absolute value of the correlation between $h_{\mathbf{w}}(\boldsymbol{\xi})$ and γ in the objective function as a penalty term while optimizing the AUC of $h_{\mathbf{w}}$ for predicting ζ. We are only able to apply this method when the fairness constraints are based on Example 1 because the constraints based on Examples 2 and 3 cannot be equivalently represented as penalty terms of statistical correlations.
- In the post-processing method, we first train a model by optimizing the AUC of $h_{\mathbf{w}}$ for predicting ζ without any constraints. Then we modify the predicted scores on data with $\gamma = 1$ to $\omega_1 h_{\mathbf{w}}(\boldsymbol{\xi}) + \omega_2$ but leave the scores on data with $\gamma = -1$ unchanged. We then tune ω_1 and ω_2 to satisfy the constraints in (2). We are unable to apply post-processing to Example 3 since tuning ω_1 and ω_2 requires knowing the true labels (ζ) of the data, which is impractical.

E.3 Process of Tuning Hyperparameters

In this section, we explain the process to tune the hyper-parameters.

Convex case. For the level-set method and the proxy-Lagrangian method, we solve their constrained optimization problems with different values of κ. For each value of κ, we track the models from all iterations and return the one that is feasible to (2) and reaches the best AUC on the validation set. In the correlation-penalty method, we select λ from a set of candidates, solve the penalized optimization problem by the stochastic gradient descent method, and select the model to return in the same way as the previous two methods. We set $c_2 = 1$ and choose c_1 from 0.5 and 1 for all methods. For the level-set method, we set $\theta = 1$ in Algorithm 1 and $\eta_t = \frac{c}{\sqrt{t+1}}$ in Algorithm 2 with c tuned from $\{10^{-2}, 10^{-1}, 1\}$ based on the AUC of the returned model on the validation set. The learning rates of proxy-Lagrangian and correlation-penalty are tuned in the same way. For post-processing, ω_1 is tuned from a grid in $[0, 5]$ with a gap of 0.05 and ω_2 is tuned from a grid in $[-3, 3]$ with step size 0.1. We use a mini-batch of size 100 in each method when computing stochastic gradients.

Weakly-convex case. The implementation of each method and the process of tuning hyperparameters is the same as the convex case except that we choose $\hat{\rho} = 10^{-5}$ in Algorithm 3.

E.4 Plots of COMPAS Dataset

In this section, we present the Pareto frontier obtained by each method on the COMPAS dataset in Figures 3 and 4 for the convex case and the weakly-convex case, respectively.

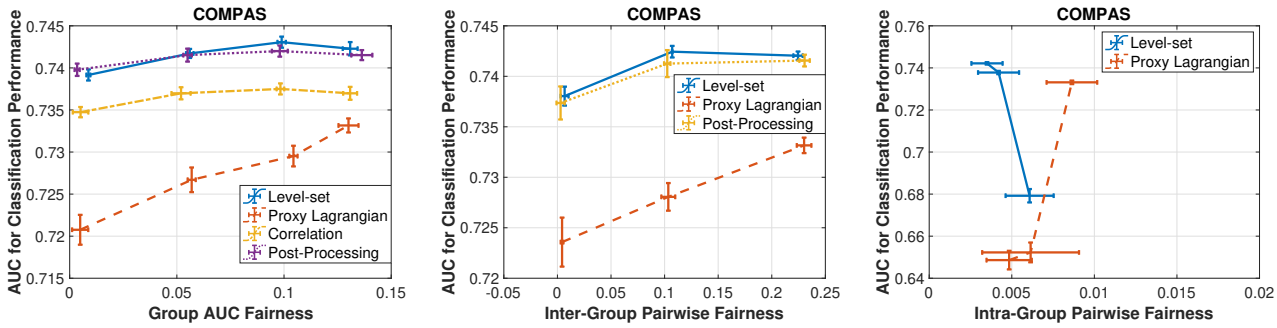

Figure 4: Pareto frontiers by each method for COMPAS dataset in weakly-convex case.