On the Maximal Local Disparity of Fairness-Aware Classifiers

Jinqiu Jin [1] **Haoxuan Li** [2] **Fuli Feng** [1]

Abstract

Fairness has become a crucial aspect in the development of trustworthy machine learning algorithms. Current fairness metrics to measure the violation of demographic parity have the following drawbacks: (i) the *average difference* of model predictions on two groups cannot reflect their *distribution disparity*, and (ii) the *overall* calculation along all possible predictions conceals the *extreme local disparity* at or around certain predictions. In this work, we propose a novel fairness metric called **M**aximal **C**umulative ratio **D**isparity along varying **P**redictions' neighborhood (MCDP), for measuring the maximal local disparity of the fairness-aware classifiers. To accurately and efficiently calculate the MCDP, we develop a provably exact and an approximate calculation algorithm that greatly reduces the computational complexity with low estimation error. We further propose a bi-level optimization algorithm using a differentiable approximation of the MCDP for improving the algorithmic fairness. Extensive experiments on both tabular and image datasets validate that our fair training algorithm can achieve superior fairness-accuracy trade-offs.

1. Introduction

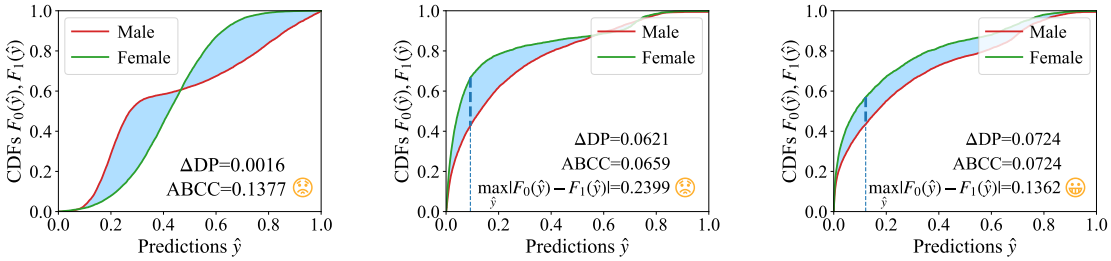
Nowadays, machine learning algorithms have been widely-used in high-stake applications such as loan management (Mukerjee et al., 2002), job-hiring (Faliagka et al., 2012), and recidivism prediction (Berk et al., 2021). Nonetheless, these algorithms are prone to exhibit discrimination against particular groups, leading to unfair decision-making results (Tolan et al., 2019; Raghavan et al., 2020; Mehrabi et al., 2021). To address this issue, growing attentions have been paid to developing comprehensive fairness criterions (Jacobs & Wallach, 2021; Han et al., 2023a) and effective fair learning algorithms (Wan et al., 2023; Han et al., 2024).

Existing group fairness notions require algorithms to treat different groups equally, and the degree of fairness violation is usually measured via the dissimilarity of model predictions. For example, **D**emographic **P**arity (DP) requires model predictions to be independent of sensitive attributes (Dwork et al., 2012; Kamishima et al., 2012; Jiang et al., 2020). To measure the violation of DP, most of existing works adopt $\Delta\mathrm{DP}$ metric, which calculates the **difference in average** predictions between the two demographic groups (Zemel et al., 2013; Chuang & Mroueh, 2021; Li et al., 2023b). However, since having the same values in average predictions between the two groups cannot ensure that the distributions are also the same, we argue that the widely used $\Delta\mathrm{DP}$ may fail to detect the violation of demographic parity. Figure 1a illustrates a toy example where the red and green curves represent the **C**umulative **D**istribution **F**unction (CDF) for the male and female groups, respectively. Despite the small $\Delta\mathrm{DP}$, it is clear that the prediction distribution is not independent of gender as a sensitive attribute, with males more likely to be assigned extremely small and large predictive probabilities, while females are more concentrated in the middle-sized probabilities. To show the limitation of $\Delta\mathrm{DP}$, we further calculate the **A**rea **B**etween **C**DF **C**urves (ABCC) to measure the total variation between CDFs (Han et al., 2023a), which demonstrates significantly greater differences between groups.

Moreover, as most of the existing metrics measure the **overall** disparity along all possible predictions, they fail to capture the **local** disparity at or around certain predictions. As a matter of fact, different metric values do not imply the relative magnitude of extreme local disparities. For example, although both $\Delta\mathrm{DP}$ and ABCC values of predictions in Figure 1b is better than those in Figure 1c, its CDF disparity along predictions is less uniform, resulting in much more serious maximal disparity ($0.2399 > 0.1362$). Generally, a pre-defined threshold is used to make a binary decision based on model predictions, where varying thresholds lead to changing proportions of positive decisions of different groups (Chen & Wu, 2020). If the decision threshold takes the value where the maximal CDF disparity is achieved, the group unfairness would be seriously exacerbated (*e.g.,* the difference of group positive proportion in Figure 1b will be up to nearly 24%). Therefore, it's important to capture and

[1]University of Science and Technology of China [2]Peking University. Correspondence to: Haoxuan Li <hxli@stu.pku.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).



(a) Small ΔDP, but large ABCC (b) Large maximal local disparity (c) Small maximal local disparity

Figure 1. The empirical distribution functions of model predictions over male and female groups. (a) shows a toy example, while (b) and (c) are the testing results of FairMixup (Chuang & Mroueh, 2021) and our proposed algorithm on the Adult dataset, respectively.

measure the extreme local disparity, which is yet prone to be "averaged" by overall variation of previous fairness metrics.

To address the limitations of previous fairness metrics, we propose a novel fairness metric called **M**aximal **C**umulative ratio **D**isparity along varying **P**redictions' ϵ-neighborhood, denoted as $\mathrm{MCDP}(\epsilon)$, whose core idea is calculating the maximal local disparity of the CDFs of two demographic groups. Considering that vast of predictions falling in a small prediction interval may result in sharp distribution changes, we firstly replace the exact CDF disparity at each prediction as the minimal disparity within its ϵ-neighborhood. In this way, the group disparity along varying predictions becomes smoother, thus the metric value becomes more robust to sharp distributions. We also theoretically prove several properties of the proposed $\mathrm{MCDP}(\epsilon)$ metric, including but not limited to its monotonicity *w.r.t.* ϵ, and its relationship with previous fairness metrics.

Furthermore, given the model predictions for real instances in two demographic groups, we adopt empirical distribution function as the estimated CDF, and further propose two algorithms which exactly and approximately calculate the empirical metric value, respectively. Specifically, the approximation algorithm can greatly reduce the computational complexity with low value error. To train a fair classifier in view of maximal local disparity, we firstly adopt a differentiable approximation of CDF disparity based on temperature sigmoid function (Han & Moraga, 1995), and then minimize the maximal estimated CDF disparity using the bi-level optimization approach (Ji et al., 2021). In this way, the distribution disparity becomes more uniform (Figure 1c), further approaching demographic parity.

The contributions of this paper are summarized as follows:

- We propose a novel fairness metric called $\mathrm{MCDP}(\epsilon)$ to measure the maximal local disparity of a classifier, and derive several theoretical properties about the metric.
- To empirically estimate $\mathrm{MCDP}(\epsilon)$ with finite instances, we propose an exact and an approximate calculation algorithm, where the approximate algorithm greatly reduces computational complexity with low estimation error.
- We further develop a bi-level optimization algorithm using differentiable approximation of $\mathrm{MCDP}(0)$ to train a fair classifier which minimizes the maximal local disparity.
- Experiments on tabular and image datasets (Adult, Bank and CelebA) demonstrate that our learning algorithm can effectively achieve better fairness-accuracy trade-offs.

2. Preliminaries

2.1. Demographic Parity and Measurements

Without loss of generality, we consider the binary classification task where each instance consists of an input $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$, a class label $y \in \mathcal{Y} = \{0, 1\}$ and a group label $s \in \{0, 1\}$, which is defined by sensitive attributes such as gender, age or race. We focus on demographic parity which requires the model's predictive probabilities $\hat{y} = f(\boldsymbol{x})$ to be independent of the group label s, where $f : \mathcal{X} \to [0, 1]$ is a classifier with model parameter $\boldsymbol{\theta}$. To measure the model's fairness violation, the following metric calculates the average prediction difference of two groups

$$\Delta\mathrm{DP}(f) = \left|\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_0} f(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_1} f(\boldsymbol{x})\right|,$$

where $\mathcal{P}_a = \mathbb{P}(\boldsymbol{x}|s = a), a \in \{0, 1\}$ denotes the distributions of instances in two groups. In addition, given a finite dataset $\mathcal{D} = \{\boldsymbol{x}_i, y_i, s_i\}_{i=1}^N$ with N samples and model predictions $\{\hat{y}_i\}_{i=1}^N$ with $\hat{y}_i = f(\boldsymbol{x}_i)$, the empirical estimation of $\Delta\mathrm{DP}(f)$ can be obtained as

$$\widehat{\Delta\mathrm{DP}}(f, \mathcal{D}) = \left|\frac{1}{|\mathcal{S}_0|}\sum_{i\in\mathcal{S}_0}\hat{y}_i - \frac{1}{|\mathcal{S}_1|}\sum_{i\in\mathcal{S}_1}\hat{y}_i\right|,$$

where $\mathcal{S}_a = \{i : s_i = a\}, a \in \{0, 1\}$ is the index set of instances in two groups. Apart from the widely-used ΔDP ($\widehat{\Delta\mathrm{DP}}$) metric[1] which measures unfairness in expectation-level, recent work (Han et al., 2023a) proposes a distribution-level metric called ABCC as follows

$$\mathrm{ABCC}(f) = \int_0^1 |F_0(\hat{y}) - F_1(\hat{y})|\,\mathrm{d}\hat{y},$$

[1]To improve the notation briefness, we omit the arguments f and $\mathcal{D}$, *e.g.,* $\Delta\mathrm{DP}, \widehat{\Delta\mathrm{DP}}$ is short for $\Delta\mathrm{DP}(f), \widehat{\Delta\mathrm{DP}}(f, \mathcal{D})$.

where $F_0(\hat{y})$ and $F_1(\hat{y})$ ($\hat{y} \in [0,1]$) are the CDFs of model predictions of instances from $\mathcal{P}_0$ and $\mathcal{P}_1$

$$F_a(\hat{y}) = \mathbb{P}(f(\boldsymbol{x}) \leq \hat{y}),\ \boldsymbol{x} \sim \mathcal{P}_a,\ a \in \{0,1\}.$$

Similar to $\widehat{\Delta\mathrm{DP}}$, the estimated ABCC value using $\{\hat{y}_i\}_{i=1}^N$ can be computed as follows

$$\widehat{\mathrm{ABCC}}(f, \mathcal{D}) = \int_0^1 \left| \hat{F}_0(\hat{y}) - \hat{F}_1(\hat{y}) \right| \mathrm{d}\hat{y},$$

where $\hat{F}_0(\hat{y})$ and $\hat{F}_1(\hat{y})$ are the empirical distribution functions of model predictions of instances in two groups

$$\hat{F}_a(\mathcal{D}; \hat{y}) = \frac{1}{|\mathcal{S}_a|} \sum_{i \in \mathcal{S}_a} \mathbb{I}(\hat{y}_i \leq \hat{y}),\ a \in \{0,1\}, \tag{1}$$

and $\mathbb{I}(\cdot)$ denotes the indicator function.

2.2. Discussions about Previous Metrics

Drawbacks of ΔDP. Although ΔDP has become the *de facto* fairness criterion in previous literatures, it is insufficient to measure the violation of demographic parity. The reason is that $\Delta\mathrm{DP} = 0$ does not indicate identical distributions of group prediction, thus the independency of predictions and group labels cannot be guaranteed. As shown in Figure 1a, the distribution gap between the two groups is evident despite ΔDP is very close to 0.

Drawbacks of ABCC. Unlike ΔDP, $\mathrm{ABCC} = 0$ is a necessary and sufficient condition for establishing demogarphic parity. However, as ABCC value cannot reflect the *local* distribution disparity, it fails to accurately measure the degree of unfairness in cases where extreme local disparity is emphasized. For example, although the maximal disparity in Figure 1c (0.1362) is much smaller than that of Figure 1b (0.2399), its ABCC value is misleadingly larger (*i.e.,* $0.0724 > 0.0659$).

Summary. Previous expectation-level and distribution-level metrics tend to average the extreme but important local disparity through overall calculation, thus their values cannot accurately measure the fairness violation in certain cases. This enlighten us to develop new fairness metrics to capture the maximal local disparity of classifiers.

3. The Proposed Metric

In this section, we propose a novel fairness metric called $\mathrm{MCDP}(\epsilon)$ to measure the maximal local disparity of classifiers, and provide theoretical properties, estimation algorithms and an optimization framework about the metric. All the proofs of the theorems can be referred in Appendix A.

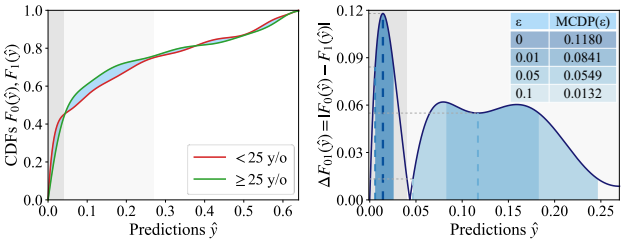


Figure 2. Predictions of FairMixup on the Bank dataset. The neighborhood hyper-parameter ϵ decides the manner of calculating local disparity, leading to different $\mathrm{MCDP}(\epsilon)$ values.

3.1. $\mathrm{MCDP}(\epsilon)$ Metric

Denote the absolute difference of the two groups' CDFs of model predictions as $\Delta F(\hat{y}) = |F_0(\hat{y}) - F_1(\hat{y})|$. To capture the worst case that a classifier violates demographic parity, an intuitive idea is to calculate the **M**aximal **C**umulative ratio **D**isparity along varying **P**redictions using

$$\mathrm{MCDP}(f) = \max_{\hat{y} \in [0,1]} \Delta F(\hat{y}). \tag{2}$$

As MCDP only focuses on the maximal disparity, it serves as a more rigorous fairness measurement compared with previous metrics. However, it is susceptible to extremely sharp distributions within certain intervals. As an example shown in Figure 2, the MCDP value is large (0.1180) due to a large number of instances with predictions around 0.02, which may misleadingly reflect the unfairness. To address this issue, we introduce a local measurement $\epsilon \geq 0$ to smooth $\Delta F(\hat{y})$. For a specific prediction point y_0, we take the minimal disparity within its ϵ-neighborhood $[y_0 - \epsilon, y_0 + \epsilon]$ as the maximal local disparity instead of the exact disparity $\Delta F(y_0)$. In this way, we slightly modify Eq. (2) to compute the **M**aximal **C**umulative ratio **D**isparity along varying **P**redictions' ϵ-neighborhood as

$$\mathrm{MCDP}(f; \epsilon) = \max_{y_0 \in [0,1]} \min_{\hat{y}: |\hat{y} - y_0| \leq \epsilon} \Delta F(\hat{y}). \tag{3}$$

$\mathrm{MCDP}(\epsilon)$ can also be interpreted as the maximum of the minimal CDF disparity of any prediction intervals with length 2ϵ (or $\geq \epsilon$ if endpoints 0 or 1 is included). In particular, as the zero length interval degenerates to a point, $\mathrm{MCDP}(0)$ degenerates to MCDP in Eq. (2).

3.2. Theoretical Analysis of $\mathrm{MCDP}(\epsilon)$

We derive several properties of $\mathrm{MCDP}(\epsilon)$ metric as below:

Theorem 3.1 (Properties of $\mathrm{MCDP}(\epsilon)$)**.** *The proposed* $\mathrm{MCDP}(\epsilon)$ *metric has the following desired properties:* ① $\mathrm{MCDP}(\epsilon)$ *has a range of* $[0,1]$. ② $\mathrm{MCDP}(0) = 0$ *holds if and only if demographic parity is established.* ③ $\mathrm{MCDP}(0)$ *is invariant to any monotone and invertible transformation* $T : [0,1] \to [0,1]$. ④ $\mathrm{MCDP}(\epsilon)$ *is a monotonically decreasing function* w.r.t. ϵ. ⑤ *Assume* $\Delta F(\hat{y})$ *is continuous*

Algorithm 1 Exact Calculation of $\widehat{\mathrm{MCDP}}(\epsilon)$

Input: Dataset $\mathcal{D} = \{\boldsymbol{x}_i, y_i, s_i\}_{i=1}^N$, model predictions $\{\hat{y}_i\}_{i=1}^N$, local measurement $\epsilon \geq 0$.
1: Calculate $\Delta\hat{F}(\hat{y})$ using $\{\hat{y}_i, s_i\}_{i=1}^N$ by Eq. (1);
2: Set $\hat{y}_0 = 0,\ \hat{y}_{N+1} = 1$;
3: **if** $\epsilon = 0$ **then**
4: $\quad\widetilde{\mathrm{MCDP}}(\epsilon) = \max_i \Delta\hat{F}(\hat{y}_i),\ i \in [0, N]$;
5: **else**
6: $\quad\widetilde{\mathrm{MCDP}}(\epsilon) = \min_{i:\hat{y}_i \leq \epsilon} \Delta\hat{F}(\hat{y}_i)$;
7: $\quad\mathcal{I} = \{i \in [0, N+1] : \hat{y}_i \leq 1 - \epsilon\}$;
8: $\quad$**for** $i \in \mathcal{I}$ **do**
9: $\quad\quad\mathcal{J} = \{j \in [0, N+1] : \hat{y}_j \in [\hat{y}_i, \hat{y}_i + 2\epsilon]\}$;
10: $\quad\quad M = \min_j \Delta\hat{F}(\hat{y}_j),\ j \in \mathcal{J}$;
11: $\quad\quad\widetilde{\mathrm{MCDP}}(\epsilon) = \max\{\widetilde{\mathrm{MCDP}}(\epsilon), M\}$;
12: $\quad$**end for**
13: **end if**
Output: $\widetilde{\mathrm{MCDP}}(\epsilon)$.

on $[0, 1]$ *with Lipschitz constant* L *(Goldstein, 1977), then*

$$\mathrm{ABCC} \leq \begin{cases} \mathrm{MCDP}(\epsilon) + \frac{\epsilon L}{2}, & \textit{if } L \leq 2, \\ \mathrm{MCDP}(\epsilon) + 2\epsilon\left(1 - \frac{1}{L}\right), & \textit{if } L > 2. \end{cases}$$

Remarks. Properties ② and ③ show that $\mathrm{MCDP}(0)$ satisfies the sufficiency and fidelity criteria for fairness measurement (Han et al., 2023a). A visualized interpretation of property ④ can be referred in Figure 2, where wider intervals (*i.e.,* larger 2ϵ values) lead to smaller $\mathrm{MCDP}(\epsilon)$ values. Property ⑤ provides an upper bound of the ABCC metric with given $\mathrm{MCDP}(\epsilon)$ values, which also suggests that the distribution-level disparity can be controlled by minimizing the maximal local disparity.

3.3. Estimating the $\mathrm{MCDP}(\epsilon)$ Metric

In real-world settings where algorithmic fairness is evaluated on finite data samples, we adopt the empirical distribution function as the estimated CDF. Formally, the empirical $\mathrm{MCDP}(\epsilon)$ metric estimated over $\mathcal{D}$ can be written as

$$\widehat{\mathrm{MCDP}}(f, \mathcal{D}; \epsilon) = \max_{y_0 \in [0,1]} \min_{\hat{y}:|\hat{y} - y_0| \leq \epsilon} \Delta\hat{F}(\hat{y}), \tag{4}$$

where $\Delta\hat{F}(\hat{y}) = |\hat{F}_0(\hat{y}) - \hat{F}_1(\hat{y})|$. By the Glivenko-Cantelli theorem (Tucker, 1959), the estimated metric value above converges to the true metric value in Eq. (3) almost surely with increasing sample size N (see Appendix B for formal proofs), which demonstrates that $\widehat{\mathrm{MCDP}}(\epsilon)$ serves as a preferable estimation of $\mathrm{MCDP}(\epsilon)$.

However, it is intractable to traverse all possible y_0 values in $[0, 1]$ and $\hat{y}$ values in y_0's ϵ-neighborhood, which poses a challenge to compute the metric in Eq. (4). Nevertheless,

Algorithm 2 Approximate Calculation of $\widehat{\mathrm{MCDP}}(\epsilon)$

Input: Dataset $\mathcal{D} = \{\boldsymbol{x}_i, y_i, s_i\}_{i=1}^N$, model predictions $\{\hat{y}_i\}_{i=1}^N$, local measurement $\epsilon > 0$, sampling frequency $K \in \mathbb{N}_+$.
1: Calculate $\Delta\hat{F}(\hat{y})$ using $\{\hat{y}_i, s_i\}_{i=1}^N$ by Eq. (1);
2: Set the step-size $\delta = \frac{\epsilon}{K}$;
3: $\widetilde{\mathrm{MCDP}}(\epsilon) = \min_j \Delta\hat{F}(j\delta),\ j \in \{0, \cdots, K\}$;
4: **for** $j = 1, \cdots, \lceil \frac{1}{\delta} \rceil - 2K$ **do**
5: $\quad M = \min_k \Delta\hat{F}(k\delta),\ k \in \{j, \cdots, j + 2K - 1\}$;
6: $\quad\widetilde{\mathrm{MCDP}}(\epsilon) = \max\{\widetilde{\mathrm{MCDP}}(\epsilon), M\}$;
7: **end for**
Output: $\widetilde{\mathrm{MCDP}}(\epsilon; K)$.

thanks to the step-like pattern's property of the empirical distribution function, we can traverse finite y_0 and $\hat{y}$ values. Based on different traversing strategies, we develop two algorithms which calculates the exact and approximate value of $\widehat{\mathrm{MCDP}}(\epsilon)$, respectively. The exact algorithm only traverses predictions of instances in $\mathcal{D}$ (lines 7-10 in Algorithm 1). In contrast, the approximate algorithm firstly samples prediction points that are equally spaced by $\frac{\epsilon}{K}$ on $[0, 1]$, where $K \in \mathbb{N}^+$ is a pre-defined hyper-parameter to control the sampling frequency. Afterwards, it traverses consecutive $2K$ sampled points to estimate the maximal local disparity (lines 4-5 in Algorithm 2). Notably, the two algorithms have the following properties:

Theorem 3.2 (Exactness)**.** *The* $\widetilde{\mathrm{MCDP}}(\epsilon)$ *value returned by Algorithm 1 equals to the true value in Eq.* (4)*,* i.e., *it calculates* $\widehat{\mathrm{MCDP}}(\epsilon)$ *exactly.*

Theorem 3.3 (Over-estimation)**.** *The* $\widetilde{\mathrm{MCDP}}(\epsilon)$ *value returned by Algorithm 2 never underestimates the true metric value,* i.e., $\widetilde{\mathrm{MCDP}}(\epsilon) \geq \widehat{\mathrm{MCDP}}(\epsilon)$.

Theorem 3.4 (Monotonicity *w.r.t.* sampling frequency)**.** *Denote* $\widetilde{\mathrm{MCDP}}(\epsilon; K)$ *as the* $\widetilde{\mathrm{MCDP}}(\epsilon)$ *value returned by Algorithm 2 with sampling frequency* K*. For any* $p > q \geq 0,\ p, q \in \mathbb{N}$*, we have* $\widetilde{\mathrm{MCDP}}(\epsilon; 2^p) \leq \widetilde{\mathrm{MCDP}}(\epsilon; 2^q)$.

Computational Complexity. The calculation process of Algorithms 1 and 2 are mainly based on specific traverse strategies on y_0 and $\hat{y}$ in Eq. (4). As the exact algorithm needs to traverse instances in $\mathcal{D}$ twice (lines 7-9 in Algorithm 1), its computational complexity is $\mathcal{O}(N^2)$. As to the approximate algorithm, the complexity of traversing sampled prediction points (line 4 in Algorithm 2) is $\mathcal{O}(\frac{K}{\epsilon})$, and the complexity of finding the minimal CDF disparity in each consecutive $2K$ points is $\mathcal{O}(K)$. Therefore, the overall computational complexity is $\mathcal{O}(\frac{K^2}{\epsilon})$. More detailed analysis can be referred in Appendix C. In practice, as K can be set with very small values (*i.e.,* $K \ll N$), the computational

Algorithm 3 DiffMCDP: Bi-Level Optimization Algorithm

Input: Training data $\mathcal{D} = \{\boldsymbol{x}_i, y_i, s_i\}_{i=1}^N$, classifier f parameterized by $\boldsymbol{\theta}$, regularization strength λ, temperature τ, epoch number E, batch size B, learning rate η.

Randomly initialize $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$;
for $e = 0, \cdots, E-1$ **do**
  Draw a mini-batch $\mathcal{D}_e = \{\boldsymbol{x}'_i, y'_i, s'_i\}_{i=1}^B$ from $\mathcal{D}$;
  // The following calculations are based on $\mathcal{D}_e$
  $\hat{y}'_i = f(\boldsymbol{x}'_i),\ i = 1, \cdots, B$;
  $\hat{y}^* = \arg\max_{\hat{y}\in[0,1]} \Delta\tilde{F}_\tau(\hat{y})$;
  $\mathcal{L} = \frac{1}{B}\sum_{i=1}^B \ell(\boldsymbol{\theta}; \hat{y}'_i, y'_i) + \lambda\Delta\tilde{F}_\tau(\hat{y}^*)$;
  $\boldsymbol{\theta}_{e+1} = \boldsymbol{\theta}_e - \eta\nabla_{\boldsymbol{\theta}}\mathcal{L}$;
end for

Output: A fair classifier f parameterized by $\boldsymbol{\theta}_E$.

complexity of the approximate algorithm is greatly reduced compared to the exact algorithm.

More Discussions about K. As discussed before, increasing K values would boost the computational complexity of Algorithm 2. On the flip side, according to Theorems 3.3 and 3.4, the estimation error would keep decreasing as K increases, indicating a trade-off between efficiency and accuracy. In practice, both estimation accuracy and efficiency can achieve promising results with varying K values.

3.4. DiffMCDP: Bi-Level Optimization Algorithm

Based on previous analysis, it is essential to train a fair classifier which minimizes the maximal local disparity to approach demographic parity. According to property ⑤ in Theorem 3.1, we can minimize $\mathrm{MCDP}(0)$ as an upper bound of $\mathrm{MCDP}(\epsilon)$ for any $\epsilon > 0$. A natural idea is to impose the $\widehat{\mathrm{MCDP}}(0)$ metric as a regularization term on the classification loss. However, as the empirical distribution functions are not differentiable *w.r.t.* model parameter $\boldsymbol{\theta}$, directly regularizing $\widehat{\mathrm{MCDP}}(0)$ is implausible. To address this issue, we firstly estimate $\Delta\hat{F}(\hat{y})$ in a differentiable way

$$\Delta\tilde{F}_\tau(\hat{y}) = \left| \frac{1}{|\mathcal{S}_0|}\sum_{i\in\mathcal{S}_0} \sigma_\tau(\hat{y} - \hat{y}_i) - \frac{1}{|\mathcal{S}_1|}\sum_{i\in\mathcal{S}_1} \sigma_\tau(\hat{y} - \hat{y}_i) \right|,$$

where $\sigma_\tau(x) = \frac{1}{1+\exp(-\tau x)}$ is a variant of sigmoid function with temperature $\tau > 0$ as a hyper-parameter (Han & Moraga, 1995). Notably, when the temperature tends to infinity, we have $\Delta\tilde{F}_\tau(\hat{y})$ converges to the $\Delta\hat{F}(\hat{y})$ as follows.

Theorem 3.5. *$\Delta\tilde{F}_\tau(\hat{y}) \xrightarrow{a.e.} \Delta\hat{F}(\hat{y})$ as $\tau \to \infty$.*

With the differentiable estimation above, the fairness-regularized objective function can be written as

$$\min_{\boldsymbol{\theta}} \left(\frac{1}{|\mathcal{D}|}\sum_{i=1}^N \ell(\boldsymbol{\theta}; \hat{y}_i, y_i) + \lambda \cdot \max_{\hat{y}\in[0,1]} \Delta\tilde{F}_\tau(\hat{y}) \right), \quad (5)$$

where $\ell(\boldsymbol{\theta}; \hat{y}_i, y_i)$ denotes the classification loss for $\boldsymbol{x}_i$, and $\lambda > 0$ controls the trade-off between accuracy and fairness. To solve Eq. (5), we adopt the bi-level optimization approach – firstly find the prediction which achieves the maximal CDF disparity as $\hat{y}^* = \arg\max_{\hat{y}\in[0,1]} \Delta\tilde{F}_\tau(\hat{y})$, and then find the optimal model parameter by $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \frac{1}{|\mathcal{D}|}\sum_{i=1}^N \ell(\boldsymbol{\theta}; \hat{y}_i, y_i) + \lambda\Delta\tilde{F}_\tau(\hat{y}^*)$. We provide the detailed learning algorithm in Algorithm 3.

4. Experiments

Datasets and Backbone Models. In our experiments, we adopt two tabular datasets and one image dataset for evaluation: ① **Adult** (Kohavi, 1996) is a popular UCI dataset which contains personal information of over 40K individuals from US 1994 census data. The task is to predict whether a person's annual income is over \$50K or not, and we treat *gender* as the sensitive attribute. ② **Bank** (Moro et al., 2014) dataset is collected from a Portuguese banking institution's marketing campaigns, and its goal is to predict whether a client will make a deposit subscription or not. We take *age* as the sensitive attribute (whether the age is over 25 or not). ③ **CelebA** (Liu et al., 2015) dataset contains over 20K face images of celebrities, where each image has 40 human-labeled binary face attributes. We use *gender* as the sensitive attribute, and choose *attractive face* and *wavy hair* as target attributes to create two meta-datasets, denoted as **CelebA-A** and **CelebA-W** respectively. For tabular datasets, we adopt a two-layer multi-layer perceptron as the backbone model. For CelebA-A and CelebA-W, we use ResNet-18 (He et al., 2016) initialized with pretrained weights.

Baselines and Evaluation Protocols. We compare our proposed method (denoted as **DiffMCDP**) with the following baselines: ① **ERM** trains the model with the vanilla classification loss without fairness objectives. ② **AdvDebias** (Zhang et al., 2018) minimizes the adversary's ability of inferring sensitive attributes from model representations. ③ **DiffDP** imposes the $\Delta\mathrm{DP}$ metric as a regularization term on the empirical risk. ④ **FairMixup** (Chuang & Mroueh, 2021) regularizes models on interpolated distributions between two groups. ⑤ **DRAlign** (Li et al., 2023b) uses gradient-guided parity alignment to encourage gradient-weighted consistency of neurons across groups. ⑥ **DiffABCC** regularizes the ABCC metric on the empirical risk. We use average precision (AP) to evaluate the classification accuracy, and use $\Delta\mathrm{DP}$, ABCC, and $\mathrm{MCDP}(\epsilon)$ to measure the algorithmic fairness. More implementation details can be referred in Appendix D.

4.1. Performance Comparison

Figure 3 shows the trade-off relationships between AP and $\mathrm{MCDP}(0)$ of baselines and our proposed method. We can observe that DiffMCDP consistently outperforms other

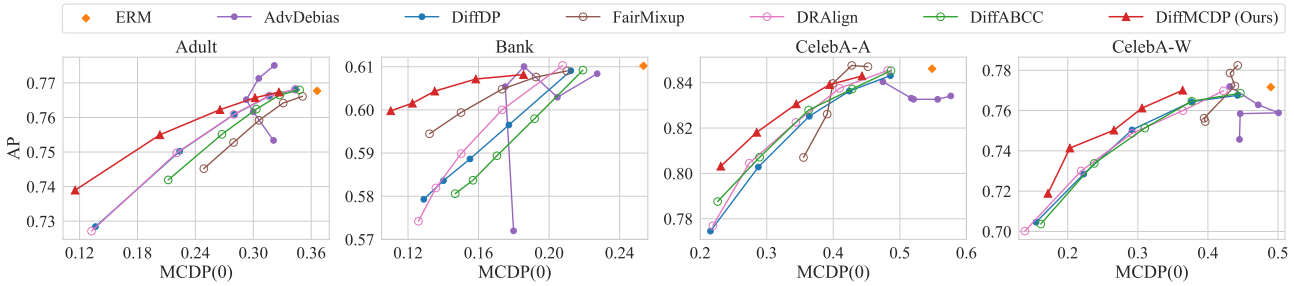


Figure 3. Trade-offs between AP and $\mathrm{MCDP}(0)$ of baselines and the proposed method. Each marker represents the average testing result of 5 runs with a specific fairness-accuracy trade-off coefficient. The curves closer to the upper-left corners indicate better performances.

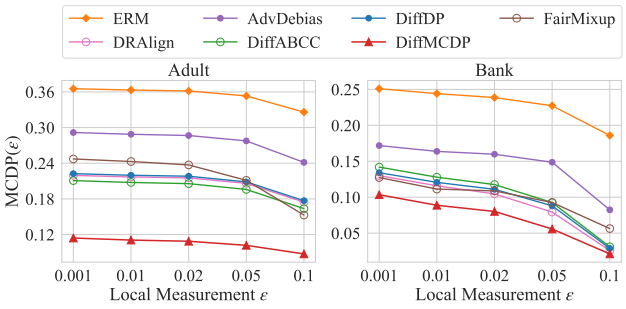
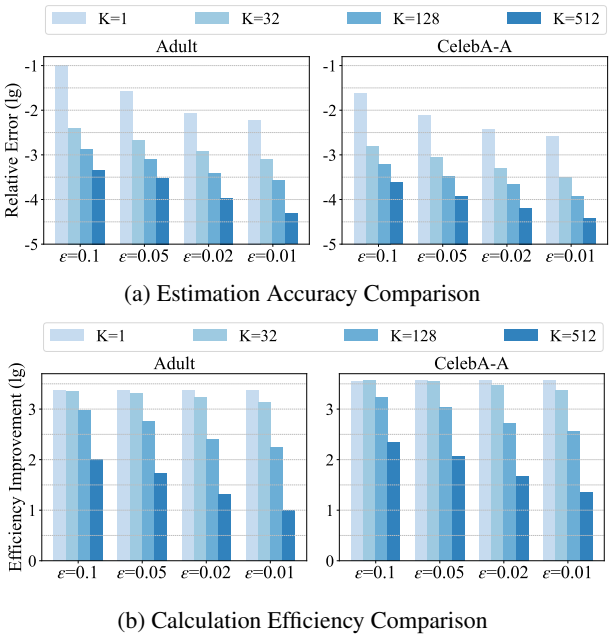


Figure 4. Comparison of $\mathrm{MCDP}(\epsilon)$ results with varying ϵ. The metric values are calculated by the exact algorithm (Algorithm 1).



(a) Estimation Accuracy Comparison

(b) Calculation Efficiency Comparison

Figure 5. Varying K and ϵ in $\widehat{\mathrm{MCDP}}(\epsilon)$ calculation algorithms.

baselines in terms of fairness-accuracy trade-offs across all datasets, which suggests that our proposed fair training algorithm can effectively reduce maximal local disparity to improve fairness. In addition, traditional fair training algorithms can achieve desired trade-offs between accuracy and our proposed fairness metric. For example, the $\mathrm{MCDP}(0)$ values of both DiffDP and DiffABCC decrease with increasing regularization strengths, which demonstrates that solely optimizing expectation-level or distribution-level metrics do contribute to minimizing the maximal local disparity. Nevertheless, they obtain inferior performance compared to DiffMCDP which directly regularizes $\mathrm{MCDP}(0)$, indicating their limitation in approaching demographic parity.

To explore the performance of our method under different fairness measurements, we report the values of $\Delta\mathrm{DP}$, ABCC and $\mathrm{MCDP}(0)$ metrics in Tables 1 and 2 (the trade-off curves can be referred in Appendix E.1). While DiffMCDP achieves the optimal $\mathrm{MCDP}(0)$ results across various datasets, it also obtains comparable even superior $\Delta\mathrm{DP}$ or ABCC values compared to other baselines. For example, DiffMCDP achieves the lowest ABCC values in tabular datasets, and its $\Delta\mathrm{DP}$ values are optimal in two image datasets. These results illustrate that optimizing the maximal local disparity is also beneficial to improve both expectation-level and distribution-level fairness metrics.

Lastly, we plot the changes of $\mathrm{MCDP}(\epsilon)$ as varying local measurements ϵ in Figure 4 (the results of image datasets are deferred in Appendix E.2). From the figure, we can observe that DiffMCDP still consistently outperforms other baselines in terms of different variants of the metric. This validates the effectiveness of regularizing the upper bound of $\mathrm{MCDP}(\epsilon)$ by $\mathrm{MCDP}(0)$ in the training objective, demonstrating that our proposed framework is applicable to various scenarios with varying ϵ values. Additionally, it is noteworthy that the relative performance of other baselines changes with increasing ϵ values (*e.g.,* there exists many intersection lines between 0.05 and 0.1 in Adult). This indicates that the relative fairness performance of various algorithms may change based on different manners of calculating local disparity, thus it's essential to select varying ϵ values in $\mathrm{MCDP}(\epsilon)$ for more comprehensive evaluation.

4.2. Exact and Approximate Calculation

To explore the estimation accuracy and efficiency of the approximate algorithm on real-world data, we run both the exact and approximate algorithms with varying K and ϵ values to compute $\widehat{\mathrm{MCDP}}(\epsilon)$ of the testing results in Figure 3. Then we calculate the *relative error* by $(V_a - V_e)/V_e$, where V_e and V_a are the estimated $\widehat{\mathrm{MCDP}}(\epsilon)$ metric value

Table 1. The performance *w.r.t.* different fairness metrics of baselines and our proposed DiffMCDP algorithm on tabular datasets. Following previous work (Jung et al., 2023), we select the fairest model which achieves at least 95% of the vanilla model's accuracy (*i.e.,* AP of ERM) on validation set, and report the average and standard deviation of metric values of 5 independent runs with different seeds on testing set. The optimal and sub-optimal results are highlighted with **bold** and underline, respectively.

	Adult				Bank			
	AP ↑	ΔDP ↓	ABCC ↓	MCDP(0) ↓	AP ↑	ΔDP ↓	ABCC ↓	MCDP(0) ↓
ERM	**76.77**±0.54	19.67±0.94	18.89±0.29	36.62±0.53	**61.02**±1.12	9.85±1.28	9.83±0.98	25.36±1.81
AdvDebias	76.52±0.62	12.73±1.81	12.95±1.35	29.28±2.26	60.54±1.02	1.95±2.76	3.95±0.97	17.52±4.12
DiffDP	75.02±0.28	7.77±1.20	8.97±0.73	22.37±1.11	58.36±1.67	1.03±0.23	2.13±0.20	14.01±1.34
FairMixup	74.52±0.40	**3.87**±1.20	7.34±0.62	24.87±1.53	59.45±1.94	1.28±0.44	2.95±0.91	13.22±2.91
DRAlign	74.98±0.28	7.64±1.22	8.86±0.75	22.08±1.19	58.19±1.54	1.15±0.41	1.98±0.23	13.58±1.26
DiffABCC	74.19±0.24	5.90±1.28	7.83±0.74	21.18±1.23	58.06±1.60	**0.99**±0.44	2.23±0.19	14.67±1.35
DiffMCDP	73.90±0.29	6.63±0.85	**6.09**±0.59	**11.53**±1.12	59.98±1.80	2.13±0.72	**1.83**±0.28	**11.02**±1.00

Table 2. Performance comparison on the CelebA-A and CelebA-W datasets. All details are the same as Table 1.

	CelebA-A				CelebA-W			
	AP ↑	ΔDP ↓	ABCC ↓	MCDP(0) ↓	AP ↑	ΔDP ↓	ABCC ↓	MCDP(0) ↓
ERM	**84.62**±0.38	54.32±0.77	37.20±0.69	54.85±0.92	77.16±0.68	34.02±2.24	28.94±1.36	48.96±2.16
AdvDebias	84.04±0.51	46.90±1.89	30.72±2.05	47.45±1.83	**77.19**±0.47	31.14±1.44	25.27±2.48	43.11±5.20
DiffDP	82.52±0.42	36.03±1.77	21.33±1.40	36.40±1.61	75.04±1.60	23.01±0.95	16.99±1.29	29.17±3.00
FairMixup	80.70±1.03	34.64±3.23	**14.15**±1.63	35.52±2.67	75.45±0.72	26.49±4.48	17.96±1.04	39.61±3.01
DRAlign	80.45±1.05	26.36±2.48	15.25±1.71	27.38±2.46	74.86±1.49	22.20±1.54	16.95±1.11	29.16±2.76
DiffABCC	80.71±0.86	28.05±1.83	15.70±1.67	28.94±1.89	73.38±1.69	19.93±1.85	**13.67**±0.88	23.77±3.44
DiffMCDP	80.32±0.67	**21.10**±1.94	16.03±1.61	**23.10**±1.66	74.14±1.26	**19.58**±1.91	15.35±1.26	**20.29**±1.78

returned by the exact and approximate algorithms, respectively. Moreover, we measure the *efficiency improvement* by T_e/T_a, where T_e and T_a represents the time of a single run of the exact and approximate algorithms.

We plot the results across different settings in Figure 5 (results on more datasets are deferred in Appendix E.3), where we have the following observations. ① With increasing K values, the relative error decreases as Theorems 3.3 and 3.4 suggests. Meanwhile, the calculation time also increases, which is consistent with the positive correlation between computational complexity and sampling frequency. It is noteworthy that selecting moderate K values can achieve promising performance. For example, when $K = 32$, the average estimation error and efficiency improvement are $0.03\% - 3\%$ and over 1000 times, respectively. ② As ϵ increases, the calculation efficiency improvement over the exact algorithm keeps increasing, which can be explained by the inverse correlation between computational complexity and ϵ. Meantime, the relative error also increases, and a potential reason is that estimating the minimal CDF value with $2K$ sampled prediction points within a larger interval tends to be more inaccurate. ③ The efficiency improvements are more obvious in CelebA-A dataset compared with Adult dataset, which is attributed to that the number of instances in CelebA-A is larger than Adult, and the computational complexity of the exact algorithm is quadratically proportional to the sample size. This indicates that the efficiency superiority of approximation algorithm becomes more evident when adopted to predictions on more instances.

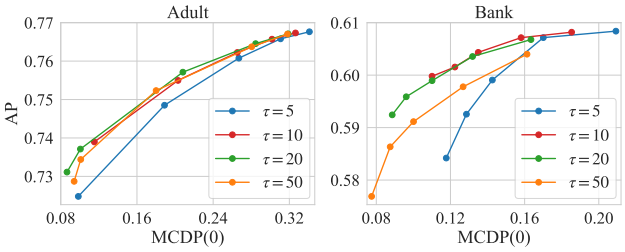


Figure 6. Trade-offs with varying temperature τ.

4.3. In-Depth Analysis

Effect of Temperature τ. The temperature τ in Algorithm 3 is crucial to approximate the true maximal CDF disparity. On the one hand, as shown in Theorem 3.5, the estimation error vanishes when the temperature τ grows. On the other hand, in practice, a large τ value may arise the gradient vanishing problem (Roodschild et al., 2020), which limits the learning capacity and optimization convergence. To verify this point, we tune τ in $\{5, 10, 20, 50\}$ and plot the fairness-accuracy trade-off curves as shown in Figure 6 (the results of image datasets are deferred in Appendix E.4). We find that very small or large temperatures ($\tau = 5, 50$) may lead to dissatisfactory results, thus it's better to adopt moderate temperatures ($\tau = 10, 20$) to effectively trade-off the estimation accuracy and the gradient magnitude.

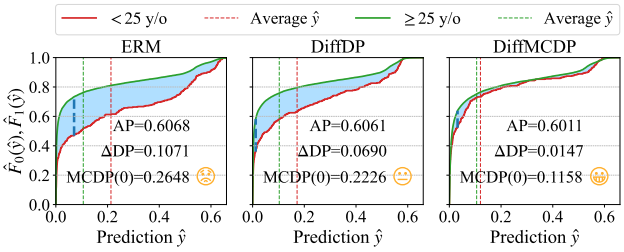


Figure 7. Visualizations of empirical distribution functions.

Visualizations of Model Predictions. Figure 7 visualizes the empirical distribution functions of model predictions for two age groups in Bank dataset, where the training methods include ERM, DiffDP and DiffMCDP. As ERM optimizes the vanilla classification loss without fairness objectives, its prediction distribution gap is the most evident. Taking a step further, DiffDP regularizes the $\widehat{\Delta\mathrm{DP}}$ metric on empirical risk, thus the average prediction gap is much smaller than ERM. Nevertheless, it leaves a lot to be desired in terms of reducing the maximal local disparity. By contrast, our proposed DiffMCDP method is much more effective in narrowing the maximal gap between prediction distributions, *i.e.,* the $\mathrm{MCDP}(0)$ value of DiffMCDP declined by 47.98% compared to DiffDP, which is also beneficial for minimizing the difference of average predictions.

5. Related Work

Fairness Notions and Metrics. While machine learning algorithms are widely-used in high-stake applications with broad social impacts, algorithmic fairness has been a crucial requirement in accessing and regulating the models (Yeung, 2018; Chen et al., 2019). Generally, algorithmic fairness notions can be categorized into group fairness (Feldman et al., 2015; Shui et al., 2022; Sun et al., 2023; Jin et al., 2024), individual fairness (Dwork et al., 2012; Biega et al., 2018; Li et al., 2023a; Wicker et al., 2023) and counterfactual fairness (Kusner et al., 2017; Chiappa, 2019; Wu et al., 2019; Ma et al., 2022; Han et al., 2023b; Rosenblatt & Witter, 2023). Group fairness requires the model to treat different groups specified by certain sensitive attributes equally without discrimination. For example, demographic parity (Zemel et al., 2013; Jiang et al., 2020; Fukuchi & Sakuma, 2023) defines fairness as the independence of model predictions and sensitive attributes, while equalized odds (Hardt et al., 2016) requires the prediction and group membership to be independent conditioned on the target label.

To measure the degree of fairness violation, various fairness metrics have been adopted for model evaluation (Garg et al., 2020; Franklin et al., 2022; Han et al., 2024). Specifically, to measure the violation of demographic parity, $\Delta\mathrm{DP}$ (Zemel et al., 2013) calculates the difference of average model predictions of instances in two demographic groups; p-Rule (Zafar et al., 2017b) computes the ratio between probabilities of two groups assigned the positive decision outcome; $\mathrm{SDD}/\mathrm{SPDD}$ (Jiang et al., 2020) averages the binary group prediction's disparity over 100 uniformly-spaced thresholds; and ABCC (Han et al., 2023a) measures the difference between the prediction distribution for different groups. In this work, we point out that these expectation-level or distribution-level metrics may fail to measure unfairness in certain cases, since the extreme but important local disparity is prone to be averaged by overall variation.

Fair Machine Learning Algorithms. To alleviate unfairness of machine learning systems, various fairness-aware algorithms have been proposed, which can be generally divided into three categories: pre-processing methods (Kamiran & Calders, 2012; Calmon et al., 2017; Zhang et al., 2018; Biswas & Rajan, 2021) attempt to adjust the training data distribution to remove the underlying data bias; in-processing methods (Kamishima et al., 2012; Zafar et al., 2017a;b; Chuang & Mroueh, 2021; Roh et al., 2021) aim to reduce model's intrinsic discrimination during the training stage; and post-processing (Hardt et al., 2016; Pleiss et al., 2017; Noriega-Campero et al., 2019; Chen et al., 2024; Yin et al., 2024) methods perform calibration on model predictions after the training. Among these three categories, in-processing methods fundamentally improve the fairness of both model outputs and representations, which can also lead to the highest utility (Barocas et al., 2023; Wan et al., 2023). Recently, there is also a line of works focusing on how to achieve algorithmic fairness in special cases, for instance, fairness under distributional shift (Chai & Wang, 2022; Jiang et al., 2023; Roh et al., 2023) or with missing attribute values (Zhao et al., 2022; Feng et al., 2023; Zhu et al., 2023). Nevertheless, it is unknown that whether these algorithms can improve fairness in terms of reducing maximal local disparity or not, and this work re-evaluates some of the methods on benchmark datasets.

6. Conclusions and Future Work

In this work, we reveal that previous expectation-level fairness metrics ($\Delta\mathrm{DP}$) cannot accurately measure the violation of demographic parity. Meanwhile, the overall variation of distribution-level metrics (ABCC) may conceal the extreme but important local disparity. Upon this understanding, we propose a novel metric called $\mathrm{MCDP}(\epsilon)$, which calculates the maximal local disparity (defined by the minimal CDF disparity of a prediction's ϵ-neighborhood) of two demographic groups. Accordingly, we propose two algorithms to estimate $\mathrm{MCDP}(\epsilon)$ with finite samples, where the approximate algorithm greatly reduces the computational complexity with high accuracy compared to the exact one. Furthermore, we develop a fair model learning framework which regularizes a differentiable estimation of $\mathrm{MCDP}(0)$. Extensive experiments are conducted on both tabular and image datasets demonstrate the effectiveness of our proposal.

In the future, we plan to extend the $\mathrm{MCDP}(\epsilon)$ metric to measure the unfairness in terms of other fairness notions, such as equalized odds (Hardt et al., 2016) and demographic parity over multiple or continuous attributes (Jiang et al., 2022; Grari et al., 2024). Moreover, designing more effective fair learning algorithms to control the maximal local disparity is also a interesting and promising research direction.

Acknowledgement

This work was supported in part by National Natural Science Foundation of China (No. 623B2002 and 62272437) and the CCCD Key Lab of Ministry of Culture and Tourism.

Impact Statement

Fairness is a crucial issue when designing trustworthy machine learning systems. Recent years have witnessed many impressive works which aims at developing fair machine learning algorithms under various scenarios. In this paper, by noticing the gap between overall distribution disparity (which is the core of traditional fairness measurements) and local disparity, we propose a novel fairness metric called $\mathrm{MCDP}(\epsilon)$, which captures the maximal local disparity of model predictions on two demographic groups. Compared with previous metrics such as $\Delta\mathrm{DP}$ and ABCC, $\mathrm{MCDP}(\epsilon)$ is more strict and conservative in measuring the violation of demographic parity, which is beneficial for reducing the risk of discrimination and improving algorithmic fairness when developing high-stake applications. We call on the research community to investigate more rigorous and theoretically-guaranteed fairness evaluation protocols, which can advance the field of trustworthy machine learning.

References

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

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 50 (1):3–44, 2021.

Biega, A. J., Gummadi, K. P., and Weikum, G. Equity of Attention: Amortizing Individual Fairness in Rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 405–414, 2018.

Bishop, C. M. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

Biswas, S. and Rajan, H. Fair Preprocessing: Towards Understanding Compositional Fairness of Data Transformers in Machine Learning Pipeline. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 981–993, 2021.

Boyd, S. P. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. Optimized Pre-Processing for Discrimination Prevention. *Advances in Neural Information Processing Systems*, 30, 2017.

Chai, J. and Wang, X. Fairness with Adaptive Weights. In *International Conference on Machine Learning*, pp. 2853–2866. PMLR, 2022.

Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. Fairness under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 339–348, 2019.

Chen, M. and Wu, M. Towards Threshold Invariant Fair Classification. In *Conference on Uncertainty in Artificial Intelligence*, pp. 560–569. PMLR, 2020.

Chen, W., Klochkov, Y., and Liu, Y. Post-hoc Bias Scoring Is Optimal For Fair Classification. In *International Conference on Learning Representations*, 2024.

Chiappa, S. Path-Specific Counterfactual Fairness. *Proceedings of the AAAI conference on artificial intelligence*, 33 (01):7801–7808, 2019.

Chuang, C.-Y. and Mroueh, Y. Fair Mixup: Fairness via Interpolation. In *International Conference on Learning Representations*, 2021.

Chung, K. L. and Erdös, P. On the Application of the Borel-Cantelli Lemma. *Transactions of the American Mathematical Society*, 72(1):179–186, 1952.

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., and Zemel, R. Flexibly Fair Representation Learning by Disentanglement. In *International Conference on Machine Learning*, pp. 1436–1445. PMLR, 2019.

Dai, E. and Wang, S. Say No to the Discrimination: Learning Fair Graph Neural Networks with Limited Sensitive Attribute Information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 680–688, 2021.

Dutta, S., Wei, D., Yueksel, H., Chen, P.-Y., Liu, S., and Varshney, K. Is There a Trade-off between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing. In *International Conference on Machine Learning*, pp. 2803–2813. PMLR, 2020.

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

Faliagka, E., Ramantas, K., Tsakalidis, A., and Tzimas, G. Application of Machine Learning Algorithms to an Online Recruitment System. In *International Conference on Internet and Web Applications and Services*, pp. 215–220, 2012.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, 2015.

Feng, R., Calmon, F. P., and Wang, H. Adapting Fairness Interventions to Missing Values. *Advances in Neural Information Processing Systems*, 2023.

Franklin, J. S., Bhanot, K., Ghalwash, M., Bennett, K. P., McCusker, J., and McGuinness, D. L. An Ontology for Fairness Metrics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 265–275, 2022.

Fukuchi, K. and Sakuma, J. Demographic Parity Constrained Minimax Optimal Regression under Linear Model. *Advances in Neural Information Processing Systems*, 2023.

Garg, P., Villasenor, J., and Foggo, V. Fairness Metrics: A Comparative Analysis. In *2020 IEEE International Conference on Big Data*, pp. 3662–3666. IEEE, 2020.

Goldstein, A. Optimization of Lipschitz Continuous Functions. *Mathematical Programming*, 13:14–22, 1977.

Graham, R. L., Knuth, D. E., Patashnik, O., and Liu, S. Concrete Mathematics: A Foundation for Computer Science. *Computers in Physics*, 3(5):106–107, 1989.

Grari, V., Laugel, T., Hashimoto, T., Lamprier, S., and Detyniecki, M. On the Fairness ROAD: Robust Optimization for Adversarial Debiasing. In *International Conference on Learning Representations*, 2024.

Han, J. and Moraga, C. The Influence of the Sigmoid Function Parameters on the Speed of Backpropagation Learning. In *International Workshop on Artificial Neural Networks*, pp. 195–201. Springer, 1995.

Han, X., Jiang, Z., Jin, H., Liu, Z., Zou, N., Wang, Q., and Hu, X. Retiring δDP: New Distribution-Level Metrics for Demographic Parity. *Transactions on Machine Learning Research*, 2023a.

Han, X., Zhang, L., Wu, Y., and Yuan, S. Achieving Counterfactual Fairness for Anomaly Detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 55–66. Springer, 2023b.

Han, X., Chi, J., Chen, Y., Wang, Q., Zhao, H., Zou, N., and Hu, X. Ffb: A Fair Fairness Benchmark for In-Processing Group Fairness Methods. In *International Conference on Learning Representations*, 2024.

Hardt, M., Price, E., and Srebro, N. Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, 29, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Jacobs, A. Z. and Wallach, H. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385, 2021.

Ji, K., Yang, J., and Liang, Y. Bilevel Optimization: Convergence Analysis and Enhanced Design. In *International Conference on Machine Learning*, pp. 4882–4892. PMLR, 2021.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. Wasserstein Fair Classification. In *Uncertainty in Artificial Intelligence*, pp. 862–872. PMLR, 2020.

Jiang, Z., Han, X., Fan, C., Yang, F., Mostafavi, A., and Hu, X. Generalized Demographic Parity for Group Fairness. In *International Conference on Learning Representations*, 2022.

Jiang, Z., Han, X., Jin, H., Wang, G., Chen, R., Zou, N., and Hu, X. Chasing Fairness under Distribution Shift: A Model Weight Perturbation Approach. *Advances in Neural Information Processing Systems*, 2023.

Jin, J., Li, H., Feng, F., Ding, S., Wu, P., and He, X. Fairly Recommending with Social Attributes: A Flexible and Controllable Optimization Approach. *Advances in Neural Information Processing Systems*, 36, 2024.

Jung, S., Park, T., Chun, S., and Moon, T. Re-Weighting Based Group Fairness Regularization via Classwise Robust Optimization. In *International Conference on Learning Representations*, 2023.

Kamiran, F. and Calders, T. Data Preprocessing Techniques for Classification without Discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference*, pp. 35–50. Springer, 2012.

Kohavi, R. 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*, volume 96, pp. 202–207, 1996.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual Fairness. *Advances in Neural Information Processing Systems*, 30, 2017.

Li, P., Xia, E., and Liu, H. Learning Antidote Data to Individual Unfairness. In *International Conference on Machine Learning*, pp. 20168–20181. PMLR, 2023a.

Li, T., Guo, Q., Liu, A., Du, M., Li, Z., and Liu, Y. FAIRER: Fairness as Decision Rationale Alignment. In *International Conference on Machine Learning*, pp. 19471–19489. PMLR, 2023b.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep Learning Face Attributes in the Wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Ma, J., Guo, R., Wan, M., Yang, L., Zhang, A., and Li, J. Learning Fair Node Representations with Graph Counterfactual Fairness. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*, pp. 695–703, 2022.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Moro, S., Cortez, P., and Rita, P. A Data-Driven Approach to Predict the Success of Bank Telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Mukerjee, A., Biswas, R., Deb, K., and Mathur, A. P. Multi-Objective Evolutionary Algorithms for the Risk-Return Trade-off in Bank Loan Management. *International Transactions in Operational Research*, 9(5):583–597, 2002.

Nair, V. and Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. In *International Conference on Machine Learning*, pp. 807–814. PMLR, 2010.

Noriega-Campero, A., Bakker, M. A., Garcia-Bulle, B., and Pentland, A. Active Fairness in Algorithmic Decision Making. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 77–83, 2019.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32, 2019.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On Fairness and Calibration. *Advances in Neural Information Processing Systems*, 30, 2017.

Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 469–481, 2020.

Roh, Y., Lee, K., Whang, S. E., and Suh, C. FairBatch: Batch Selection for Model Fairness. In *International Conference on Learning Representations*, 2021.

Roh, Y., Lee, K., Whang, S. E., and Suh, C. Improving Fair Training under Correlation Shifts. In *International Conference on Machine Learning*, pp. 29179–29209. PMLR, 2023.

Roodschild, M., Gotay Sardiñas, J., and Will, A. A New Approach for the Vanishing Gradient Problem on Sigmoid Activation. *Progress in Artificial Intelligence*, 9(4): 351–360, 2020.

Rosen, K. H. *Discrete Mathematics and Its Applications*. The McGraw Hill Companies, 2007.

Rosenblatt, L. and Witter, R. T. Counterfactual Fairness Is Basically Demographic Parity. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14461–14469, 2023.

Seabold, S. and Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the 9th Python in Science Conference*, volume 57, pp. 10–25080. Austin, TX, 2010.

Shui, C., Xu, G., Chen, Q., Li, J., Ling, C. X., Arbel, T., Wang, B., and Gagné, C. On Learning Fairness and Accuracy on Multiple Subgroups. *Advances in Neural Information Processing Systems*, 35:34121–34135, 2022.

Sun, R., Zhou, F., Dong, Z., Xie, C., Hong, L., Li, J., Zhang, R., Li, Z., and Li, Z. Fair-CDA: Continuous and Directional Augmentation for Group Fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8): 9918–9926, 2023.

Tolan, S., Miron, M., Gómez, E., and Castillo, C. Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia. In *Proceedings of the 17th International Conference on Artificial Intelligence and Law*, pp. 83–92, 2019.

Tucker, H. G. A Generalization of the Glivenko-Cantelli Theorem. *The Annals of Mathematical Statistics*, 30(3): 828–830, 1959.

Wan, M., Zha, D., Liu, N., and Zou, N. In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Transactions on Knowledge Discovery from Data*, 17(3):1–27, 2023.

Wick, M., Tristan, J.-B., et al. Unlocking Fairness: A Trade-off Revisited. *Advances in Neural Information Processing Systems*, 32, 2019.

Wicker, M., Piratia, V., and Weller, A. Certification of Distributional Individual Fairness. *Advances in Neural Information Processing Systems*, 2023.

Wu, Y., Zhang, L., and Wu, X. Counterfactual Fairness: Unidentification, Bound and Algorithm. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019.

Yeung, K. Algorithmic Regulation: A Critical Interrogation. *Regulation & Governance*, 12(4):505–523, 2018.

Yin, T., Ton, J.-F., Guo, R., Yao, Y., Liu, M., and Liu, Y. Fair Classifiers that Abstain without Harm. In *International Conference on Learning Representations*, 2024.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180, 2017a.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness Constraints: Mechanisms for Fair Classification. In *Artificial Intelligence and Statistics*, pp. 962–970. PMLR, 2017b.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning Fair Representations. In *International Conference on Machine Learning*, pp. 325–333. PMLR, 2013.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Zhao, T., Dai, E., Shu, K., and Wang, S. Towards Fair Classifiers without Sensitive Attributes: Exploring Biases in Related Features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pp. 1433–1442, 2022.

Zhu, Z., Yao, Y., Sun, J., Li, H., and Liu, Y. Weak Proxies Are Sufficient and Preferable for Fairness with Missing Sensitive Attributes. In *International Conference on Machine Learning*, pp. 43258–43288. PMLR, 2023.

Appendix

A. Proofs

A.1. Proofs of Theorem 3.1

Theorem 3.1 (Properties of $\mathrm{MCDP}(\epsilon)$). *The proposed* $\mathrm{MCDP}(\epsilon)$ *metric has the following desired properties:* ① $\mathrm{MCDP}(\epsilon)$ *has a range of* $[0,1]$. ② $\mathrm{MCDP}(0)=0$ *holds if and only if demographic parity is established.* ③ $\mathrm{MCDP}(0)$ *is invariant to any monotone and invertible transformation* $T:[0,1]\to[0,1]$ *on* $\hat{y}$. ④ $\mathrm{MCDP}(\epsilon)$ *is a monotonically decreasing function* w.r.t. ϵ. ⑤ *Assume* $\Delta F(\hat{y})$ *is continuous on* $[0,1]$ *with Lipschitz constant* L *(Goldstein, 1977), then*

$$\mathrm{ABCC} \leq \begin{cases} \mathrm{MCDP}(\epsilon)+\frac{\epsilon L}{2}, & \text{if } L\leq 2,\\ \mathrm{MCDP}(\epsilon)+2\epsilon\left(1-\frac{1}{L}\right), & \text{if } L>2. \end{cases}$$

Proof. ① If demographic parity is established, *i.e.,* $\forall\hat{y}\in[0,1],\ F_0(\hat{y})=F_1(\hat{y})$, then $\mathrm{MCDP}(\epsilon)=0$. In addition, let $F_0(\hat{y})=\mathbb{I}(\hat{y}=1)$ and $F_1(\hat{y})=1$, we have $\mathrm{MCDP}(\epsilon)=1$. This proves that $\mathrm{MCDP}(\epsilon)\in[0,1]$.

② According to Eq. (3), $\mathrm{MCDP}(0)$ can be written as

$$\mathrm{MCDP}(0)=\max_{\hat{y}\in[0,1]}\Delta F(\hat{y})=\max_{\hat{y}\in[0,1]}\left|F_0(\hat{y})-F_1(\hat{y})\right|.$$

Thus the following holds

$$\mathrm{MCDP}(0)=0 \iff \forall\hat{y}\in[0,1],\ F_0(\hat{y})=F_1(\hat{y}),$$

which indicates that $\mathrm{MCDP}(0)=0$ and demographic parity are equivalent.

③ Denote the prediction which achieves the largest $\mathrm{MCDP}(0)$ value as $\hat{y}^*=\arg\max_{\hat{y}\in[0,1]}\Delta F(\hat{y})$. Moreover, denote the CDFs for the transformed predictions in two groups as $F_0^T(\hat{y})$ and $F_1^T(\hat{y})$. As T is a monotone and invertible transformation on $[0,1]$, for any $\hat{y}'\in[0,1]$, there exists $\hat{y}=T^{-1}(\hat{y}')$, such that $F_a^T(\hat{y}')=F_a(\hat{y}),\ a\in\{0,1\}$. Thus we have

$$\left|F_0^T(\hat{y}')-F_1^T(\hat{y}')\right|=\left|F_0(\hat{y})-F_1(\hat{y})\right|\leq\left|F_0(\hat{y}^*)-F_1(\hat{y}^*)\right|=\mathrm{MCDP}(0),\ \forall\hat{y}'\in[0,1],$$

and the equal sign holds if and only if $\hat{y}'=T(\hat{y}^*)$. This indicates that $\mathrm{MCDP}(0)$ is invariant to T.

④ For any $y_0\in[0,1]$ and $0\leq\epsilon_a<\epsilon_b\leq 1$, we have

$$\{\hat{y}:|\hat{y}-y_0|\leq\epsilon_b\}\subset\{\hat{y}:|\hat{y}-y_0|\leq\epsilon_a\}\implies\max_{y_0\in[0,1]}\min_{|\hat{y}-y_0|\leq\epsilon_b}\Delta F(\hat{y})\geq\max_{y_0\in[0,1]}\min_{|\hat{y}-y_0|\leq\epsilon_a}\Delta F(\hat{y}),$$

which yields that $\mathrm{MCDP}(\epsilon_a)\leq\mathrm{MCDP}(\epsilon_b)$. □

⑤ To complete the proof of Property ⑤, we firstly present and prove the following two lemmas:

Lemma A.1. *If* $\Delta F(\hat{y})$ *is continuous on* $[0,1]$ *with Lipschitz constant* $L\leq 2$*, then* $\mathrm{MCDP}(\epsilon)=0\Rightarrow\mathrm{ABCC}\leq\frac{\epsilon L}{2}$.

Lemma A.2. *If* $\Delta F(\hat{y})$ *is continuous on* $[0,1]$ *with Lipschitz constant* $L>2$*, then* $\mathrm{MCDP}(\epsilon)=0\Rightarrow\mathrm{ABCC}\leq\frac{2\epsilon(L-1)}{L}$.

To complete the proof of Lemma A.1, we first give some useful lemmas (**which is based on** $\mathrm{MCDP}(\epsilon)=0$) below.

Lemma A.3. *Let* $\tilde{y}_1=\min\{\hat{y}\in[0,1]:\Delta F(\hat{y})=0\}$*, then* $\tilde{y}_1\leq\epsilon$ *and* $\int_0^{\tilde{y}_1}\Delta F(\hat{y})\mathrm{d}\hat{y}\leq\frac{\epsilon L\tilde{y}_1}{2}$.

Proof. Firstly, we note that the following holds

$$\min_{\hat{y}\leq\epsilon}\Delta F(\hat{y})\leq\max_{y_0\in[0,1]}\min_{|\hat{y}-y_0|\leq\epsilon}\Delta F(\hat{y})=\mathrm{MCDP}(\epsilon)=0\Longrightarrow\min_{\hat{y}\leq\epsilon}\Delta F(\hat{y})=0.$$

Thus $\tilde{y}_1\leq\arg\min_{\hat{y}\leq\epsilon}\Delta F(\hat{y})=0\leq\epsilon$. Furthermore, for any $\hat{y}\in[0,\tilde{y}_1]$, by Lipschitz continuity, we have

$$\left|\frac{\Delta F(\hat{y})-\Delta F(\tilde{y}_1)}{\hat{y}-\tilde{y}_1}\right|\leq L\Longrightarrow\Delta F(\hat{y})\leq L(\tilde{y}_1-\hat{y}),\tag{6}$$

which yields that

$$\int_0^{\tilde{y}_1} \Delta F(\hat{y})\mathrm{d}\hat{y} \leq \int_0^{\tilde{y}_1} L(\tilde{y}_1 - \hat{y})\mathrm{d}\hat{y} = \frac{L\tilde{y}_1^2}{2} \leq \frac{\epsilon L \tilde{y}_1}{2}.$$

□

Lemma A.4. *For any* $0 \leq a < b \leq 1$ *such that* $\Delta F(a) = \Delta F(b) = 0$, $\int_a^b \Delta F(\hat{y})\mathrm{d}\hat{y} \leq \frac{(b-a)^2 L}{4}$.

Proof. For any $\hat{y} \in [a, b]$, similar to the derivation in Eq. (6), we have

$$\Delta F(\hat{y}) \leq \min\{(\hat{y} - a)L, (b - \hat{y})L\} = \begin{cases} (\hat{y} - a)L, & \text{if } \hat{y} \leq \frac{a+b}{2}, \\ (b - \hat{y})L, & \text{if } \hat{y} \geq \frac{a+b}{2}. \end{cases}$$

Thus we have

$$\int_a^b \Delta F(\hat{y})\mathrm{d}\hat{y} \leq \int_a^{\frac{a+b}{2}} (\hat{y} - a)L\mathrm{d}\hat{y} + \int_{\frac{a+b}{2}}^b (b - \hat{y})L\mathrm{d}\hat{y} = \frac{(b-a)^2 L}{4}.$$

This completes the proof of Lemma A.4. □

Lemma A.5. *For any* $\hat{y} \in (\tilde{y}_1, 1)$ *such that* $\Delta F(\hat{y}) = 0$ *(where* $\tilde{y}_1$ *is defined as Lemma A.3 stated), there exists* $\hat{y}_l, \hat{y}_r$ *such that* $\max\{0, \hat{y} - 2\epsilon\} \leq \hat{y}_l \leq \hat{y} \leq \hat{y}_r \leq \min\{\hat{y} + 2\epsilon, 1\}$ *and* $\Delta F(\hat{y}_l) = \Delta F(\hat{y}_r) = 0$.

Proof. Assume to the contrary that there exists $y_0 \in (\tilde{y}_1, 1)$, such that $\Delta F(y_0) = 0$ and one of the following satisfies. (i) If $\min_{\hat{y} \in [y_0 - 2\epsilon, y_0)} \Delta F(\hat{y}) > 0$, let $\tilde{y}_l = \max\{\hat{y} < y_0 : \Delta F(\hat{y}) = 0\}$ and $y_l = \frac{\tilde{y}_l + y_0}{2}$, we have $\tilde{y}_l \in [\tilde{y}_1, y_0 - 2\epsilon)$ and $[y_l - \epsilon, y_l + \epsilon] \subset (\tilde{y}_l, y_0)$. Thus the $\mathrm{MCDP}(\epsilon)$ value should satisfy that

$$\mathrm{MCDP}(\epsilon) = \max_{y_0 \in [0,1]} \min_{|\hat{y} - y_0| \leq \epsilon} \Delta F(\hat{y}) \geq \min_{|\hat{y} - y_l| \leq \epsilon} \Delta F(\hat{y}) \geq \min_{\hat{y} \in (\tilde{y}_l, y_0)} \Delta F(\hat{y}) > 0. \tag{7}$$

(ii) If $\min_{\hat{y} \in (y_0, y_0 + 2\epsilon]} \Delta F(\hat{y}) > 0$, let $\tilde{y}_r = \min\{\hat{y} > y_0 : \Delta F(\hat{y}) = 0\}$ and $y_r = \frac{y_0 + \tilde{y}_r}{2}$. Similar to (i), we have

$$\mathrm{MCDP}(\epsilon) \geq \min_{|\hat{y} - y_r| \leq \epsilon} \Delta F(\hat{y}) \geq \min_{\hat{y} \in (y_0, \tilde{y}_r)} \Delta F(\hat{y}) > 0. \tag{8}$$

The fact that at least one of Eq. (7) and Eq. (8) holds derives that $\mathrm{MCDP}(\epsilon) > 0$, which contradicts with $\mathrm{MCDP}(\epsilon) = 0$. Therefore, the original assumption must be false, which completes the proof of Lemma A.5. □

Proof of Lemma A.1. According to Lemma A.5, we can construct a sequence of predictions $0 \leq \tilde{y}_1 < \cdots < \tilde{y}_M < \tilde{y}_{M+1} = 1$, such that $\tilde{y}_1 = \min\{\hat{y} \in [0, 1] : \Delta F(\hat{y}) = 0\}$ and for any $i = 1, \cdots, M$, $\Delta F(\tilde{y}_i) = 0$ and $\tilde{y}_{i+1} - \tilde{y}_i \leq 2\epsilon$. By Lemmas A.3, A.4 and A.5, the value of ABCC should satisfy that

$$\begin{aligned} \mathrm{ABCC} = \int_0^1 \Delta F(\hat{y})\mathrm{d}\hat{y} &= \int_0^{\tilde{y}_1} \Delta F(\hat{y})\mathrm{d}\hat{y} + \sum_{i=1}^M \int_{\tilde{y}_i}^{\tilde{y}_{i+1}} \Delta F(\hat{y})\mathrm{d}\hat{y} \\ &\leq \frac{\epsilon L \tilde{y}_1}{2} + \sum_{i=1}^M \frac{(\tilde{y}_{i+1} - \tilde{y}_i)^2 L}{4} \leq \frac{\epsilon L \tilde{y}_1}{2} + \sum_{i=1}^M \frac{(\tilde{y}_{i+1} - \tilde{y}_i) \cdot 2\epsilon L}{4} \\ &= \frac{\epsilon L}{2}(\tilde{y}_1 + \tilde{y}_{M+1} - \tilde{y}_1) = \frac{\epsilon L}{2}. \end{aligned} \tag{9}$$

This completes the proof of Lemma A.1.

To complete the proof of Lemma A.2, we first give the following lemmas (**which is based on** $\mathrm{MCDP}(\epsilon) = 0$ **and** $L > 2$).

Lemma A.6 (Monotonicity of CDFs)**.** *For any* $0 \leq a < b \leq 1$, $F_0(a) < F_0(b)$ *and* $F_1(a) < F_1(b)$.

Lemma A.7. *Let* $\tilde{y}_1 = \min\{\hat{y} \in [0, 1] : \Delta F(\hat{y}) = 0\}$, *then the following holds: (i)* $h = \max_{\hat{y} \in [0, \tilde{y}_1]} \Delta F(\hat{y}) \leq F_0(\tilde{y}_1)$, *and (ii)* $\int_0^{\tilde{y}_1} \Delta F(\hat{y})\mathrm{d}\hat{y} \leq \left(\tilde{y}_1 - \frac{h}{2L}\right) h$.

Proof. Let $\hat{y}^* = \arg\max_{\hat{y}\in[0,\tilde{y}_1]} \Delta F(\hat{y})$. By Lemma A.6 and $F_0(\tilde{y}_1) = F_1(\tilde{y}_1)$, we have

$$h = \Delta F(\hat{y}^*) \leq \max\{F_0(\hat{y}^*) - F_1(\hat{y}^*), F_1(\hat{y}^*) - F_0(\hat{y}^*)\} \leq \max\{F_0(\tilde{y}_1) - F_1(0), F_1(\tilde{y}_1) - F_0(0)\} \leq F_0(\tilde{y}_1).$$

Moreover, by Lipschitz continuity, h should also satisfy that $h = \Delta F(\hat{y}^*) \leq L(\tilde{y}_1 - \hat{y}^*) \leq L\tilde{y}_1$. Denote $\tilde{y}_r = \tilde{y}_1 - \frac{h}{L} \geq 0$. Combining Eq. (6) and $h \geq \Delta F(\hat{y}), \forall \hat{y} \in [0, \tilde{y}_1]$, we can control the bound of $\int_0^{\tilde{y}_1} \Delta F(\hat{y})\mathrm{d}\hat{y}$ as follows

$$\int_0^{\tilde{y}_1} \Delta F(\hat{y})\mathrm{d}\hat{y} = \int_0^{\tilde{y}_r} \Delta F(\hat{y})\mathrm{d}\hat{y} + \int_{\tilde{y}_r}^{\tilde{y}_1} \Delta F(\hat{y})\mathrm{d}\hat{y} \leq \int_0^{\tilde{y}_r} h\mathrm{d}\hat{y} + \int_{\tilde{y}_r}^{\tilde{y}_1} L(\tilde{y}_1 - \hat{y})\mathrm{d}\hat{y} = \tilde{y}_1 h - \frac{h^2}{2L}.$$

This completes the proof of Lemma A.7. □

Lemma A.8. *For any $0 \leq a < b \leq 1$ such that $\Delta F(a) = \Delta F(b) = 0$, the following holds: (i) $h = \max_{\hat{y}\in[a,b]} \Delta F(\hat{y}) \leq F_0(b) - F_0(a)$, and (ii) $\int_a^b \Delta F(\hat{y})\mathrm{d}\hat{y} \leq \left(b - a - \frac{h}{L}\right) h$.*

Proof. Denote $\hat{y}^* = \arg\max_{\hat{y}\in[a,b]} \Delta F(\hat{y})$. Combining Lemma A.6 and $\Delta F(a) = \Delta F(b) = 0$, we have $F_0(a) = F_1(a) \leq \min\{F_0(\hat{y}^*), F_1(\hat{y}^*)\}$ and $F_0(b) = F_1(b) \geq \max\{F_0(\hat{y}^*), F_1(\hat{y}^*)\}$. Thus the following holds

$$h = \Delta F(\hat{y}^*) = \max\{F_0(\hat{y}^*), F_1(\hat{y}^*)\} - \min\{F_0(\hat{y}^*), F_1(\hat{y}^*)\} \leq F_0(b) - F_0(a).$$

Furthermore, according to Lipschitz continuity, for any $\hat{y} \in [a, b]$, we have

$$\Delta F(\hat{y}) \leq \min\{(\hat{y} - a)L, (b - \hat{y})L\} \leq \frac{(b-a)L}{2}, \tag{10}$$

which also implies that $h \leq \frac{(b-a)L}{2}$. Let $\tilde{y}_l = a + \frac{h}{L}$ and $\tilde{y}_r = b - \frac{h}{L}$, we have $a \leq \tilde{y}_l \leq \tilde{y}_r \leq b$. Combining Eq. (10) and $h \geq \Delta F(\hat{y}), \forall \hat{y} \in [a, b]$, we obtain an upper bound of $\int_a^b \Delta F(\hat{y})\mathrm{d}\hat{y}$ as follows

$$\begin{aligned}
\int_a^b \Delta F(\hat{y})\mathrm{d}\hat{y} &= \int_a^{\tilde{y}_l} \Delta F(\hat{y})\mathrm{d}\hat{y} + \int_{\tilde{y}_l}^{\tilde{y}_r} \Delta F(\hat{y})\mathrm{d}\hat{y} + \int_{\tilde{y}_r}^{b} \Delta F(\hat{y})\mathrm{d}\hat{y} \\
&\leq \int_a^{\tilde{y}_l} (\hat{y} - a)L\mathrm{d}\hat{y} + \int_{\tilde{y}_l}^{\tilde{y}_r} h\mathrm{d}\hat{y} + \int_{\tilde{y}_r}^{b} (b - \hat{y})L\mathrm{d}\hat{y} \\
&= \left(b - a - \frac{h}{L}\right) h.
\end{aligned}$$

This completes the proof of Lemma A.8. □

Proof of Lemma A.2. According to Lemma A.5, for any $\Delta F(\hat{y}) = |F_0(\hat{y}) - F_1(\hat{y})|$, we can construct a sequence of predictions $0 = \tilde{y}_0 \leq \tilde{y}_1 < \cdots < \tilde{y}_M < \tilde{y}_{M+1} = 1$, such that (i) $\tilde{y}_1 = \min\{\hat{y} \in [0, 1] : \Delta F(\hat{y}) = 0\} \leq \epsilon$, and (ii) for any $i = 1, \cdots, M$, $\Delta F(\tilde{y}_i) = 0$ and $\tilde{y}_{i+1} - \tilde{y}_i \leq 2\epsilon$. Denote δ_i and h_i ($i = 0, \cdots, M$) as $\delta_i = \tilde{y}_{i+1} - \tilde{y}_i$ and $h_i = \max_{\hat{y}\in[\tilde{y}_i,\tilde{y}_{i+1}]} \Delta F(\hat{y})$, respectively. By Lemmas A.7 and A.8, the following holds

$$\begin{gathered}
0 \leq h_0 \leq \delta_0 L, \qquad 0 \leq h_i \leq \frac{\delta_i L}{2}, \text{ for any } i = 1, \cdots, M, \\
\sum_{i=0}^{M} h_i \leq F_0(\tilde{y}_1) + \sum_{i=1}^{M} (F_0(\tilde{y}_{i+1}) - F_0(\tilde{y}_i)) = F_0(\tilde{y}_{i+1}) = 1, \\
\int_0^{\tilde{y}_1} \Delta F(\hat{y})\mathrm{d}\hat{y} \leq \left(\delta_0 - \frac{h_0}{2L}\right) h_0, \qquad \int_{\tilde{y}_i}^{\tilde{y}_{i+1}} \Delta F(\hat{y})\mathrm{d}\hat{y} \leq \left(\delta_i - \frac{h_i}{L}\right) h_i, \text{ for any } i = 1, \cdots, M.
\end{gathered} \tag{11}$$

Combining (i) $\sum_{i=0}^{M} \delta_i = \tilde{y}_{M+1} - \tilde{y}_0 = 1$, (ii) $\delta_0 \in [0, \epsilon]$, (iii) $\delta_i \in (0, 2\epsilon]$, $\forall i = 1, \cdots, M$, (iv) ABCC =

$\sum_{i=0}^{M}\int_{\tilde{y}_i}^{\tilde{y}_{i+1}}\Delta F(\hat{y})\mathrm{d}\hat{y}$, and Eq. (11), for a given M, we can construct the following optimization problem

$$\begin{aligned}
&\max_{\delta_i,h_i}\left(\delta_0-\frac{h_0}{2L}\right)h_0+\sum_{i=1}^{M}\left(\delta_i-\frac{h_i}{L}\right)h_i,\\
&\textit{s.t. } 0\le\delta_0\le\epsilon,\quad 0\le h_0\le\delta_0 L,\\
&\qquad 0\le\delta_i\le 2\epsilon,\quad 0\le h_i\le\frac{\delta_i L}{2},\qquad \forall i=1,\cdots,M,\\
&\qquad \sum_{i=0}^{M}\delta_i=1,\quad \sum_{i=0}^{M}h_i\le 1.
\end{aligned}\tag{12}$$

As discussed before, for any $\Delta F(\hat{y})$, we can construct a feasible solution of the optimization problem in Eq. (12); moreover, the ABCC value should less than or equal to the corresponding objective value. Therefore, the optimal objective value of Eq. (12) should be an upper bound of ABCC. To solve the problem, we firstly consider the following problem

$$\begin{aligned}
&\min_{\delta_i,h_i}\left(\frac{h_0}{2L}-\delta_0\right)h_0+\sum_{i=1}^{M}\left(\frac{h_i}{L}-\delta_i\right)h_i,\\
&\textit{s.t. } \sum_{i=1}^{M}\delta_i=1,\quad \sum_{i=1}^{M}h_i\le 1.
\end{aligned}\tag{13}$$

Note that the problem in Eq. (13) is a convex relaxation of the original problem in Eq. (12), thus the (negative of) optimal value of Eq. (13) is upper bound on the optimal value of Eq. (12) and ABCC (Boyd & Vandenberghe, 2004). The Lagrange function of the above optimization problem is

$$\mathcal{L}(\delta_i,h_i;\alpha,\beta)=\left(\frac{h_0}{2L}-\delta_0\right)h_0+\sum_{i=1}^{M}\left(\frac{h_i}{L}-\delta_i\right)h_i+\alpha\left(\sum_{i=1}^{M}\delta_i-1\right)+\beta\left(\sum_{i=1}^{M}h_i-1\right),$$

where $\beta\ge 0,\alpha$ are the Lagrange multipliers. By requiring the derivatives of $\mathcal{L}(\delta_i,h_i;\alpha,\beta)$ *w.r.t.* δ_i,h_i to be zero, we have

$$\delta_0=h_0=0,\quad \delta_i=h_i=\frac{1}{M},\ i=1,\cdots,M.\tag{14}$$

It's easy to validate that Eq. (14) also satisfies all constraints in Eq. (12) (as $L>2$), indicating that Eq. (14) is the optimal solution of the problem in Eq. (12). Note that $1=\sum_{i=0}^{M}\delta_i\le M\cdot 2\epsilon$ holds, thus $M\ge\frac{1}{2\epsilon}$. Therefore, we can compute an upper bound of ABCC as below

$$\mathrm{ABCC}\le 0+\sum_{i=1}^{M}\left(\frac{1}{M}-\frac{1}{ML}\right)\frac{1}{M}=\frac{1}{M}\left(1-\frac{1}{L}\right)\le 2\epsilon\left(1-\frac{1}{L}\right).$$

This completes the proof of Lemma A.2. **Note that** the whole proof process does not require $L\le 2$, *i.e.,* Lemmas A.3-A.8 applies to all L values (including $L>2$).

At last, we restate Property ⑤ as the following theorem, and complete its proof accordingly.

Theorem A.9 (An upper bound of ABCC)**.** *Assume $\Delta F(\hat{y})$ is continuous on $[0,1]$ with Lipschitz constant L, then*

$$\mathrm{ABCC}\le\begin{cases}\mathrm{MCDP}(\epsilon)+\frac{\epsilon L}{2}, & \textit{if } L\le 2,\\ \mathrm{MCDP}(\epsilon)+2\epsilon\left(1-\frac{1}{L}\right), & \textit{if } L>2.\end{cases}$$

Proof. For any $\Delta F(\hat{y})=|F_0(\hat{y})-F_1(\hat{y})|$ such that $\mathrm{MCDP}(\epsilon)=M>0$, construct $\tilde{F}_0(\hat{y}),\tilde{F}_1(\hat{y})$ as follows

$$\tilde{F}_0(\hat{y})=\begin{cases}\max\{F_0(\hat{y})-M,F_1(\hat{y})\}, & \text{if } F_0(\hat{y})\ge F_1(\hat{y})\\ \min\{F_0(\hat{y})+M,F_1(\hat{y})\}, & \text{if } F_0(\hat{y})<F_1(\hat{y})\end{cases},\quad \tilde{F}_1(\hat{y})=F_1(\hat{y}),\tag{15}$$

and denote $\Delta\tilde{F}(\hat{y})=|\tilde{F}_0(\hat{y})-\tilde{F}_1(\hat{y})|$. We show that $\tilde{F}_0(\hat{y}),\tilde{F}_1(\hat{y})$, and $\Delta\tilde{F}(\hat{y})$ have the following properties.

Lemma A.10. *The value of $\tilde{F}_0(\hat{y})$ satisfies that (i) $\tilde{F}_0(0) \geq 0$, (ii) $\tilde{F}_0(1) = 1$, (iii) for any $0 \leq a < b \leq 1$, $\tilde{F}_0(a) \leq \tilde{F}_0(b)$, and (iv) $|\tilde{F}_0(\hat{y}) - F_0(\hat{y})| \leq M$.*

Proof. (i)(ii) By Eq. (15), the following holds

$$\tilde{F}_0(0) = \begin{cases} \max\{F_0(0) - M, F_1(0)\} \geq F_1(0) \geq 0, & \text{if } F_0(\hat{y}) \geq F_1(\hat{y}) \\ \min\{F_0(0) + M, F_1(0)\} \geq \min\{M, 0\} \geq 0, & \text{if } F_0(\hat{y}) < F_1(\hat{y}) \end{cases}, \quad \tilde{F}_0(1) = \tilde{F}_1(1) = 1.$$

(iii) By the monotonicity of CDF, for any $0 \leq a < b \leq 1$, we have $\tilde{F}_0(b) \geq F_1(b) \geq F_1(a) \geq \tilde{F}_0(a)$.

(iv) By Eq. (15), $|\tilde{F}_0(\hat{y}) - F_0(\hat{y})| \leq M$ satisfies that

$$|\tilde{F}_0(\hat{y}) - F_0(\hat{y})| = \begin{cases} |\max\{-M, F_1(\hat{y}) - F_0(\hat{y})\}| = |\min\{M, F_0(\hat{y}) - F_1(\hat{y})\}| \leq M, & \text{if } F_0(\hat{y}) \geq F_1(\hat{y}) \\ |\min\{M, F_1(\hat{y}) - F_0(\hat{y})\}| \leq M, & \text{if } F_0(\hat{y}) < F_1(\hat{y}). \end{cases}$$

This completes the proof of Lemma A.10. □

Lemma A.11. *The value of $\Delta\tilde{F}(\hat{y})$ can be computed by $\Delta\tilde{F}(\hat{y}) = \max\{0, \Delta F(\hat{y}) - M\}$.*

Proof. By Eq. (15), we have

$$\begin{aligned} \Delta\tilde{F}(\hat{y}) = |\tilde{F}_0(\hat{y}) - F_1(\hat{y})| &= \begin{cases} |\max\{F_0(\hat{y}) - F_1(\hat{y}) - M, 0\}| = |\max\{\Delta F(\hat{y}) - M, 0\}|, & \text{if } F_0(\hat{y}) \geq F_1(\hat{y}) \\ |\min\{F_0(\hat{y}) - F_1(\hat{y}) + M, 0\}| = |\min\{M - \Delta F(\hat{y}), 0\}|, & \text{if } F_0(\hat{y}) < F_1(\hat{y}) \end{cases} \\ &= \begin{cases} |\Delta F(\hat{y}) - M|, & \text{if } \Delta F(\hat{y}) \geq M \\ 0, & \text{if } \Delta F(\hat{y}) < M \end{cases}. \end{aligned}$$

which yields that $\Delta\tilde{F}(\hat{y}) = \max\{0, \Delta F(\hat{y}) - M\}$. □

Lemma A.12. *$\Delta\tilde{F}(\hat{y})$ and $\Delta F(\hat{y})$ exhibit the same monotonic behavior.*

Proof. One the one hand, for any $a, b \in [0, 1]$ such that $\Delta F(a) \leq \Delta F(b)$, by Lemma A.11, we have

$$\begin{aligned} \Delta\tilde{F}(a) - \Delta\tilde{F}(b) &= \max\{0, \Delta F(a) - M\} - \max\{0, \Delta F(b) - M\} \\ &= \begin{cases} 0, & \text{if } M \geq \Delta F(b) \\ M - \Delta F(b) < 0, & \text{if } \Delta F(a) < M < \Delta F(b) \\ \Delta F(a) - \Delta F(b) \leq 0, & \text{if } M \leq \Delta F(a) \end{cases}, \end{aligned} \tag{16}$$

which yields that $\Delta\tilde{F}(a) \leq \Delta\tilde{F}(b)$. On the other hand, for any $a, b \in [0, 1]$ such that $\Delta F(a) \geq \Delta F(b)$, $\Delta\tilde{F}(a) \geq \Delta\tilde{F}(b)$ can be proved in a similar manner. This yields that $\Delta\tilde{F}(\hat{y})$ and $\Delta F(\hat{y})$ have the identical monotonic trend. □

Lemma A.13. *For any $a, b \in [0, 1]$, $|\Delta\tilde{F}(a) - \Delta\tilde{F}(b)| \leq |\Delta F(a) - \Delta F(b)|$.*

Proof. If $\Delta F(a) \leq \Delta F(b)$, by Eq. (16), we have

$$|\Delta\tilde{F}(a) - \Delta\tilde{F}(b)| = \begin{cases} 0 \leq \Delta F(b) - \Delta F(a), & \text{if } M \geq \Delta F(b) \\ \Delta F(b) - M < \Delta F(b) - \Delta F(a), & \text{if } \Delta F(a) < M < \Delta F(b) \\ \Delta F(b) - \Delta F(a), & \text{if } M \leq \Delta F(a) \end{cases},$$

which yields that $|\Delta\tilde{F}(a) - \Delta\tilde{F}(b)| \leq |\Delta F(a) - \Delta F(b)|$. When $\Delta F(a) \geq \Delta F(b)$, $|\Delta\tilde{F}(a) - \Delta\tilde{F}(b)| \leq |\Delta F(a) - \Delta F(b)|$ can be proved in a similar way. This completes the proof of Lemma A.13. □

Next we continue the proof of Theorem A.9. By (i),(ii) and (iii) in Lemma A.10, $\tilde{F}_0(\hat{y})$ and $\tilde{F}_1(\hat{y})$ in Eq. (15) are valid CDFs. Moreover, by Lemma A.13 and Lipschitz continuity of $\Delta F(\hat{y})$, the following holds

$$\left|\frac{\Delta\tilde{F}(a)-\Delta\tilde{F}(b)}{a-b}\right| \leq \left|\frac{\Delta F(a)-\Delta F(b)}{a-b}\right| \leq L, \quad \text{for any } a,b\in[0,1],$$

which demonstrates that $\Delta\tilde{F}(\hat{y})$ is also continuous on $[0,1]$ with Lipschitz constant L. Furthermore, as Lemma A.12 illustrates that the monotonicity of $\tilde{F}_0(\hat{y})$ and $\tilde{F}_1(\hat{y})$ is consistent, the following holds

$$\begin{aligned}
&\arg\min_{|\hat{y}-y_0|\leq\epsilon}\Delta F(\hat{y}) = \arg\min_{|\hat{y}-y_0|\leq\epsilon}\Delta\tilde{F}(\hat{y}), \quad \text{for any } y_0\in[0,1],\\
\Longrightarrow &\arg\max_{y_0\in[0,1]}\Delta F\left(\arg\min_{|\hat{y}-y_0|\leq\epsilon}\Delta F(\hat{y})\right) = \arg\max_{y_0\in[0,1]}\Delta\tilde{F}\left(\arg\min_{|\hat{y}-y_0|\leq\epsilon}\Delta\tilde{F}(\hat{y})\right),\\
\Longrightarrow &\max_{y_0\in[0,1]}\min_{|\hat{y}-y_0|\leq\epsilon}\Delta\tilde{F}(\hat{y}) = \Delta\tilde{F}(\hat{y}^*) \overset{(*)}{=} \max\{0,\Delta F(\hat{y}^*)-M\} = 0,
\end{aligned}$$

where $\hat{y}^*\in[0,1]$ satisfies that $\Delta F(\hat{y}^*)=M$, and equality (*) holds due to Lemma A.11. To conclude, $\tilde{F}_0(\hat{y})$ and $\tilde{F}_1(\hat{y})$ in Eq. (15) satisfies all conditions of Lemmas A.1 and A.2. Denote A and $\tilde{A}$ as the ABCC values based on $\Delta F(\hat{y})$ and $\Delta\tilde{F}(\hat{y})$, respectively. By (iv) in Lemma A.10 and $||a-b|-|c-d||\leq|a-c|+|b-d|$, we have

$$\begin{aligned}
A-\tilde{A} &= \int_0^1 |F_0(\hat{y})-F_1(\hat{y})|\,\mathrm{d}\hat{y} - \int_0^1 \left|\tilde{F}_0(\hat{y})-\tilde{F}_1(\hat{y})\right|\mathrm{d}\hat{y}\\
&\leq \int_0^1 \left|F_0(\hat{y})-\tilde{F}_0(\hat{y})\right|\mathrm{d}\hat{y} + \int_0^1 \left|F_1(\hat{y})-\tilde{F}_1(\hat{y})\right|\mathrm{d}\hat{y}\\
&\leq \int_0^1 M\mathrm{d}\hat{y} + 0 = M.
\end{aligned}$$

By Lemmas A.1 and A.2, we have

$$A\leq\tilde{A}+M\leq\begin{cases} M+\frac{\epsilon L}{2}, & \text{if } L\leq 2,\\ M+2\epsilon\left(1-\frac{1}{L}\right), & \text{if } L>2.\end{cases}$$

This completes the proof of Theorem A.9.

□

A.2. Proofs of Theorems 3.2, 3.3 and 3.4

To complete the proofs of theorems about Algorithms 1 and 2, we start with giving some important lemmas and corollaries.

Lemma A.14 (Subset min-max property (Rosen, 2007))**.** *For any two finite ordered sets A, B, if $A\subseteq B$, then* $\min A\geq\min B$ *and* $\max A\leq\max B$.

Lemma A.15. *The range of $\Delta\hat{F}(\hat{y})=|\hat{F}_0(\hat{y})-\hat{F}_1(\hat{y})|$ over the interval $[y_1,y_2]\subseteq[0,1]$ is a finite set*

$$\left\{\Delta\hat{F}(\hat{y}):\hat{y}\in[y_1,y_2]\right\} = \left\{\Delta\hat{F}(\hat{y}_i):\hat{y}_i\in[\hat{y}_l,\hat{y}_r]\right\},\ l,r\in[0,N+1],$$

where $\hat{y}_0=0,\ \hat{y}_{N+1}=1,\ \hat{y}_l=\max_i\{\hat{y}_i:\hat{y}_i\leq y_1\},\ \hat{y}_r=\max_i\{\hat{y}_i:\hat{y}_i\leq y_2\},\ i\in[0,N+1]$.

Proof. **In the following, for the simplicity of exposition, the default value domain of subscript indices i,j in model predictions $\hat{y}_i,\hat{y}_j$ is $\{0,1,\cdots,N+1\}$, unless we point out different value domains in equations.**

According to Eq. (1), for any $\hat{y}\in[y_1,y_2]$, there exists an instance's prediction $\hat{y}_j=\max_i\{\hat{y}_i:\hat{y}_i\leq\hat{y}\}$ such that $\hat{F}_a(\hat{y})=\hat{F}_a(\hat{y}_j),\ a\in\{0,1\}$, which implies that $\Delta\hat{F}(\hat{y})=\Delta\hat{F}(\hat{y}_j)$. As $\hat{y}\geq y_1$ holds, $\{\hat{y}_i:\hat{y}_i\leq y_1\}$ should be a subset of $\{\hat{y}_i:\hat{y}_i\leq\hat{y}\}$. According to Lemma A.14, we have $\hat{y}_j\geq\hat{y}_l$. Similarly, we can prove that $\hat{y}_j\leq\hat{y}_r$. Therefore,

$$\{\Delta\hat{F}(\hat{y}):\hat{y}\in[y_1,y_2]\}\subseteq\{\Delta\hat{F}(\hat{y}_i):\hat{y}_i\in[\hat{y}_l,\hat{y}_r]\}. \tag{17}$$

Meanwhile, for any $\hat{y}_i \in [\hat{y}_l, \hat{y}_r]$, we have $\hat{F}_a(\hat{y}_i) = \hat{F}_a(y'),\ a \in \{0,1\}$ and $\Delta\hat{F}(\hat{y}_i) = \Delta\hat{F}(y')$, where $y' = \max\{y_1, \hat{y}_i\} \in [y_1, y_2]$. Thus the following holds

$$\{\Delta\hat{F}(\hat{y}) : \hat{y} \in [y_1, y_2]\} \supseteq \{\Delta\hat{F}(\hat{y}_i) : \hat{y}_i \in [\hat{y}_l, \hat{y}_r]\}. \tag{18}$$

Combining Eq. (17) and Eq. (18) completes the proof of Lemma A.15. □

According to Lemma A.14 and Lemma A.15, we have the following corollary:

Corollary A.16 (An over-estimation of the minimal value of $\Delta\hat{F}(\hat{y})$)**.** *The minimal value of $\Delta\hat{F}(\hat{y})$ over the interval $[y_1, y_2] \subseteq [0,1],\ y_2 \geq y_1 + \delta$ can be over-estimated by the minimal value of*

$$\left\{\Delta\hat{F}(k\delta) : k = \left\lceil \frac{y_1}{\delta} \right\rceil, \left\lceil \frac{y_1}{\delta} + 1 \right\rceil, \cdots, \left\lfloor \frac{y_2}{\delta} \right\rfloor \right\},$$

where $\delta \in (0,1)$ is the step-size hyper-parameter.

Theorem 3.2 (Exactness). *The $\widetilde{\mathrm{MCDP}}(\epsilon)$ value returned by Algorithm 1 equals to the $\widehat{\mathrm{MCDP}}(\epsilon)$ value in Eq. (4),* i.e., *Algorithm 1 calculates $\widehat{\mathrm{MCDP}}(\epsilon)$ exactly without error.*

Proof. According to Lemma A.15, when $\epsilon = 0$, we have

$$\begin{aligned} &\{\Delta\hat{F}(\hat{y}) : \hat{y} \in [0,1]\} = \{\Delta\hat{F}(\hat{y}_i) : \hat{y}_i \in [0,1]\} = \{\Delta\hat{F}(\hat{y}_i)\}, \\ \Rightarrow\ &\widehat{\mathrm{MCDP}}(0) = \max_{\hat{y}\in[0,1]} \Delta\hat{F}(\hat{y}) = \max_i \Delta\hat{F}(\hat{y}_i) = \widetilde{\mathrm{MCDP}}(0). \end{aligned} \tag{19}$$

When $\epsilon > 0$, for any $y_0 \in [0, \epsilon)$, as $[0,\epsilon] \subseteq [0, y_0 + \epsilon] \subset [0, 2\epsilon]$, by Lemma A.14 and A.15, we have

$$\max_{y_0 \in [0,\epsilon)} \min_{|\hat{y} - y_0| \leq \epsilon} \Delta\hat{F}(\hat{y}) = \min_{\hat{y} \leq \epsilon} \Delta\hat{F}(\hat{y}) = \min_{i:\hat{y}_i \leq \epsilon} \Delta\hat{F}(\hat{y}_i), \tag{20}$$

where $\min_{i:\hat{y}_i \leq \epsilon} \Delta\hat{F}(\hat{y}_i)$ equals to the initial $\widetilde{\mathrm{MCDP}}(\epsilon)$ value in line 6 of Algorithm 1. For any $y_0 \in [\epsilon, 1]$, let $\hat{y}_{0l} = \max_i\{\hat{y}_i : \hat{y}_i \leq y_0 - \epsilon\}$ and $\hat{y}_{0r} = \max_i\{\hat{y}_i : \hat{y}_i \leq y_0 + \epsilon\}$. By Lemmas A.15 and A.14, we have

$$\begin{aligned} &\{\Delta\hat{F}(\hat{y}) : |\hat{y} - y_0| \leq \epsilon\} = \{\Delta\hat{F}(\hat{y}_i) : \hat{y}_i \in [\hat{y}_{0l}, \hat{y}_{0r}]\} \supseteq \{\Delta\hat{F}(\hat{y}_i) : \hat{y}_i \in [\hat{y}_{0l}, \hat{y}_{0l} + 2\epsilon]\}, \\ \Rightarrow\ &\min_{i:\hat{y}_i \in [\hat{y}_{0l}, \hat{y}_{0l}+2\epsilon]} \Delta\hat{F}(\hat{y}_i) \geq \min_{|\hat{y} - y_0| \leq \epsilon} \Delta\hat{F}(\hat{y}), \\ \Rightarrow\ &\max_{i:\hat{y}_i \leq 1-\epsilon} \min_{j:\hat{y}_j \in [\hat{y}_i, \hat{y}_i + 2\epsilon]} \Delta\hat{F}(\hat{y}_j) \geq \max_{y_0 \in [\epsilon, 1]} \min_{|\hat{y} - y_0| \leq \epsilon} \Delta\hat{F}(\hat{y}). \end{aligned} \tag{21}$$

where the last step uses the fact that $\{\hat{y}_{0l} : y_0 \in [\epsilon, 1]\} = \{\hat{y}_i : \hat{y}_i \leq 1 - \epsilon\}$.

On the other hand, for any i such that $\hat{y}_i \leq 1 - \epsilon$, by Lemma A.15, there exists $y'_i = \hat{y}_i + \epsilon$ (thus $y'_i \in [\epsilon, 1]$) such that

$$\begin{aligned} \{\Delta\hat{F}(\hat{y}) : |\hat{y} - y'_i| \leq \epsilon\} &= \{\Delta\hat{F}(\hat{y}) : \hat{y} \in [y'_i - \epsilon, \min(1, y'_i + \epsilon)]\} = \{\Delta\hat{F}(\hat{y}_j) : \hat{y}_j \in [y'_i - \epsilon, \min(1, y'_i + \epsilon)]\} \\ &= \{\Delta\hat{F}(\hat{y}_j) : \hat{y}_j \in [\hat{y}_i, \hat{y}_i + 2\epsilon]\}. \end{aligned} \tag{22}$$

Note that for any i such that $\hat{y}_i \leq 1 - \epsilon$, the following also holds

$$\left\{\min_{|\hat{y} - y'_i| \leq \epsilon} \Delta\hat{F}(\hat{y}) : \hat{y}_i \leq 1 - \epsilon\right\} \subseteq \left\{\min_{|\hat{y} - y_0| \leq \epsilon} \Delta\hat{F}(\hat{y}) : y_0 \in [\epsilon, 1]\right\}. \tag{23}$$

By Lemma A.14, Eq. (22) and Eq. (23), we have

$$\max_{i:\hat{y}_i \leq 1-\epsilon} \min_{\hat{y}_j \in [\hat{y}_i, \hat{y}_i + 2\epsilon]} \Delta\hat{F}(\hat{y}_j) = \max_{i:\hat{y}_i \leq 1-\epsilon} \min_{|\hat{y} - y'_i| \leq \epsilon} \Delta\hat{F}(\hat{y}) \leq \max_{y_0 \in [\epsilon, 1]} \min_{|\hat{y} - y_0| \leq \epsilon} \Delta\hat{F}(\hat{y}). \tag{24}$$

Taking a step further, combining Eq. (21) and Eq. (24), yields that

$$\max_{i:\hat{y}_i \le 1-\epsilon} \min_{\hat{y}_j \in [\hat{y}_i, \hat{y}_i + 2\epsilon]} \Delta\hat{F}(\hat{y}_j) = \max_{y_0 \in [\epsilon, 1]} \min_{|\hat{y}-y_0| \le \epsilon} \Delta\hat{F}(\hat{y}). \tag{25}$$

At last, according to Eq. (20) and Eq. (25), we can derive the relationship between $\widehat{\mathrm{MCDP}}(\epsilon)$ (Eq. (4)) and $\widetilde{\mathrm{MCDP}}(\epsilon)$ (Algorithm 1) as follows

$$\begin{aligned}\widehat{\mathrm{MCDP}}(\epsilon) &= \max_{y_0 \in [0,1]} \min_{|\hat{y}-y_0| \le \epsilon} \Delta\hat{F}(\hat{y}) = \max\left\{ \max_{y_0 \in [0,\epsilon)} \min_{|\hat{y}-y_0| \le \epsilon} \Delta\hat{F}(\hat{y}), \max_{y_0 \in [\epsilon,1]} \min_{|\hat{y}-y_0| \le \epsilon} \Delta\hat{F}(\hat{y}) \right\} \\ &= \max\left\{ \min_{i:\hat{y}_i \le \epsilon} \Delta\hat{F}(\hat{y}_i), \max_{i:\hat{y}_i \le 1-\epsilon} \min_{\hat{y}_j \in [\hat{y}_i, \hat{y}_i + 2\epsilon]} \Delta\hat{F}(\hat{y}_j) \right\} = \widetilde{\mathrm{MCDP}}(\epsilon),\ \epsilon > 0.\end{aligned} \tag{26}$$

Combining Eq. (19) and Eq. (26) completes the proof of Theorem 3.2.

□

Theorem 3.3 (Over-estimation). The $\widetilde{\mathrm{MCDP}}(\epsilon)$ value returned by the approximate algorithm satisfies that $\widetilde{\mathrm{MCDP}}(\epsilon) \ge \widehat{\mathrm{MCDP}}(\epsilon)$, *i.e.,* Algorithm 2 never underestimates $\widehat{\mathrm{MCDP}}(\epsilon)$.

Proof. For any $y_0 \in [0, \epsilon],\ \epsilon > 0$, according to Eq. (20) and Corollary A.16, we have

$$\max_{y_0 \in [0,\epsilon)} \min_{|\hat{y}-y_0| \le \epsilon} \Delta\hat{F}(\hat{y}) = \min_{\hat{y} \le \epsilon} \Delta\hat{F}(\hat{y}) \le \min_{j \in \{0, \cdots, K\}} \Delta\hat{F}(j\delta), \tag{27}$$

where $\min_j \Delta\hat{F}(j\delta), j \in \{0, \cdots, K\}$ is the initial value of $\widetilde{\mathrm{MCDP}}(\epsilon)$ in Algorithm 2 (line 3).

For any $y_0 \in (\epsilon, 1-\epsilon)$, by Corollary A.16 and Corollary A.14, we have

$$\begin{aligned}&\min_{|\hat{y}-y_0| \le \epsilon} \Delta\hat{F}(\hat{y}) \le \min_{k \in \{l, \cdots, r\}} \Delta\hat{F}(k\delta) \le \min_{k \in \{l, \cdots, l+2K-1\}} \Delta\hat{F}(k\delta), \\ &\text{where } l = \left\lceil \frac{y_0 - \epsilon}{\delta} \right\rceil = \left\lceil \frac{y_0}{\delta} \right\rceil - K, \quad r = \left\lfloor \frac{y_0 + \epsilon}{\delta} \right\rfloor = \left\lfloor \frac{y_0}{\delta} \right\rfloor + K \ge l + 2K - 1.\end{aligned} \tag{28}$$

Note that $\{\lceil \frac{y_0}{\delta} \rceil - K : y_0 \in (\epsilon, 1-\epsilon)\} = \{1, \cdots, \lceil \frac{1}{\delta} \rceil - 2K\}$. Taking a step further, Eq. (28) derives that

$$\max_{y_0 \in (\epsilon, 1-\epsilon)} \min_{|\hat{y}-y_0| \le \epsilon} \Delta\hat{F}(\hat{y}) \le \max_{l \in \{1, \cdots, \lceil \frac{1}{\delta} \rceil - 2K\}} \min_{k \in \{l, \cdots, l+2K-1\}} \Delta\hat{F}(k\delta). \tag{29}$$

For any $y_0 \in [1-\epsilon, 1]$, $\min_{|\hat{y}-y_0| \le \epsilon} \Delta\hat{F}(\hat{y}) = \Delta\hat{F}(1) = 0$ holds, which yields that

$$\max_{y_0 \in [1-\epsilon, 1]} \min_{|\hat{y}-y_0| \le \epsilon} \Delta\hat{F}(\hat{y}) = \Delta\hat{F}(1) = 0. \tag{30}$$

Combining Eq. (27), Eq. (29) and Eq. (30), the $\widetilde{\mathrm{MCDP}}(\epsilon)$ value returned by Algorithm 2 satisfies that

$$\begin{aligned}\widetilde{\mathrm{MCDP}}(\epsilon) &= \max\left\{ \min_{j \in \{0, \cdots, K\}} \Delta\hat{F}(j\delta), \max_{l \in \{1, \cdots, \lceil \frac{1}{\delta} \rceil - 2K\}} \min_{k \in \{l, \cdots, l+2K-1\}} \Delta\hat{F}(k\delta) \right\} \\ &\ge \max\left\{ \max_{y_0 \in [0,\epsilon)} \min_{|\hat{y}-y_0| \le \epsilon} \Delta\hat{F}(\hat{y}), \max_{y_0 \in (\epsilon, 1-\epsilon)} \min_{|\hat{y}-y_0| \le \epsilon} \Delta\hat{F}(\hat{y}), \max_{y_0 \in [1-\epsilon, 1]} \min_{|\hat{y}-y_0| \le \epsilon} \Delta\hat{F}(\hat{y}) \right\} \\ &= \max_{y_0 \in [0,1]} \min_{|\hat{y}-y_0| \le \epsilon} \Delta\hat{F}(\hat{y}) = \widehat{\mathrm{MCDP}}(\epsilon),\ \epsilon > 0.\end{aligned}$$

This completes the proof of Theorem 3.3.

□

Theorem 3.4 (Monotonicity *w.r.t.* sampling frequency K). Denote $\widetilde{\mathrm{MCDP}}(\epsilon; K)$ as the $\widehat{\mathrm{MCDP}}(\epsilon)$ value returned by Algorithm 2 with sampling frequency K. For any $p > q \geq 0,\ p, q \in \mathbb{N}$, we have $\widetilde{\mathrm{MCDP}}(\epsilon; 2^p) \leq \widetilde{\mathrm{MCDP}}(\epsilon; 2^q)$.

Proof. Let $K_p = 2^p,\ K_q = 2^q,\ \delta_p = \frac{\epsilon}{K_p},\ \delta_q = \frac{\epsilon}{K_q}$, and

$$\mathcal{Y}_K^{(j)} = \{j \cdot \frac{\epsilon}{K}, \cdots, (j + 2K - 1) \cdot \frac{\epsilon}{K}\}, \quad \text{for all } j = 1, \cdots, \left\lceil \frac{K}{\epsilon} \right\rceil - 2K,\ K \in \mathbb{N}_+.$$

Before giving the proof, we first introduce two lemmas which will be used later.

Lemma A.17. *(Graham et al., 1989) If $m \in \mathbb{N}_+,\ n \in \mathbb{Z}$, then the equation below can be used to convert ceilings to floors*

$$\left\lceil \frac{n}{m} \right\rceil = \left\lfloor \frac{n + m - 1}{m} \right\rfloor = \left\lfloor \frac{n - 1}{m} \right\rfloor + 1.$$

Lemma A.18. *For any $j_p = 1, \cdots, \left\lceil \frac{1}{\delta_p} \right\rceil - 2K_p$, there exists $j_q \in \{1, \cdots, \left\lceil \frac{1}{\delta_q} \right\rceil - 2K_q\}$, such that $\mathcal{Y}_{K_q}^{(j_q)} \subseteq \mathcal{Y}_{K_p}^{(j_p)}$.*

Proof. Let $j_q = \lceil j_p \cdot 2^{q-p} \rceil \geq \lceil 1 \cdot 2^{-1} \rceil = 1$. By lemma A.17, we have

$$\begin{aligned} j_q &= \lfloor (j_p - 1) \cdot 2^{q-p} \rfloor \\ &\leq \left\lfloor (\lceil \frac{1}{\delta_p} \rceil - 2K_p - 1) \cdot 2^{q-p} \right\rfloor = \left\lfloor (\lceil \frac{1}{\delta_p} \rceil - 1) \cdot 2^{q-p} \right\rfloor - 2K_q \\ &\leq \left\lfloor \frac{1}{\delta_p} \cdot 2^{q-p} \right\rfloor - 2K_q = \left\lfloor \frac{1}{\delta_q} \right\rfloor - 2K_q \leq \left\lceil \frac{1}{\delta_q} \right\rceil - 2K_q, \end{aligned}$$

which suggests that $j_q \in \{1, \cdots, \left\lceil \frac{1}{\delta_q} \right\rceil - 2K_q\}$. For any $y_q = (j_q + l_q) \cdot \delta_q \in \mathcal{Y}_{K_q}^{(j_q)},\ l_q = 0, \cdots, 2K_q - 1$, we denote y_p as $y_p = (j_p + l_p) \cdot \delta_p$, where $l_p = (j_q + l_q) \cdot 2^{p-q} - j_p \in \mathbb{Z}$. On the one hand, we have $l_p \geq 0$ since the following holds

$$\frac{(j_q + l_q) \cdot 2^{p-q}}{j_p} \geq \frac{(j_p \cdot 2^{q-p} + 0) \cdot 2^{p-q}}{j_p} = 1.$$

On the other hand, according to Lemma A.17, we have $j_q = \lfloor (j_p - 1) \cdot 2^{q-p} \rfloor \leq (j_p - 1) \cdot 2^{q-p}$. Note that

$$\begin{aligned} l_p - (2K_p - 1) &= \big((j_q + l_q) \cdot 2^{p-q} - j_p\big) - (2K_p - 1) \\ &\leq \big((j_p - 1) \cdot 2^{q-p} + 1 + 2K_q - 1\big) \cdot 2^{p-q} - j_p - 2K_p + 1 \\ &= j_p - 1 + 2 \cdot 2^q \cdot 2^{p-q} - j_p - 2 \cdot 2^p + 1 = 0. \end{aligned}$$

Thus $l_p \in \{0, \cdots, 2K_p - 1\}$. Moreover, note that

$$y_q - y_p = (j_q + l_q) \cdot \frac{\epsilon}{2^q} - \big(j_p + (j_q + l_q) \cdot 2^{p-q} - j_p\big) \cdot \frac{\epsilon}{2^p} = 0,$$

in other words, $y_p = y_q$ and $y_p \in \mathcal{Y}_{K_p}^{(j_p)}$. This yields that $\mathcal{Y}_{K_q}^{(j_q)} \subseteq \mathcal{Y}_{K_p}^{(j_p)}$. □

Proof of Theorem 3.4. Rewrite $\widetilde{\mathrm{MCDP}}(\epsilon; K)$ as the following form:

$$\begin{aligned} \widetilde{\mathrm{MCDP}}(K; \epsilon) &= \max\{\underbrace{\min_{j \in \{0, \cdots, K\}} \Delta\hat{F}(j \cdot \frac{\epsilon}{K})}_{M_1(K)}, \underbrace{\max_{l \in \{1, \cdots, \lceil \frac{K}{\epsilon} \rceil - 2K\}} \min_{k \in \{l, \cdots, l + 2K - 1\}} \Delta\hat{F}(k \cdot \frac{\epsilon}{K})}_{M_2(K)}\} \\ &= \max\left(M_1(K), M_2(K)\right). \end{aligned} \tag{31}$$

To prove the theorem, we show that for any $p > q \geq 0,\ p, q \in \mathbb{N}$, $M_1(K_p) \leq M_1(K_q)$ and $M_2(K_p) \leq M_2(K_q)$ holds.

① *Proof of* $M_1(K_p) \leq M_1(K_q)$: For any $j_q = 0, \cdots, K_q$, there exists $j_p = j_q \cdot 2^{p-q} \in [0, K_p],\ j_p \in \mathbb{N}$, such that $j_q\delta_q = j_p\delta_p$. By Lemma A.14, we have

$$\{\Delta\hat{F}(j\delta_q)\}_{j=0}^{K_q} \subseteq \{\Delta\hat{F}(j\delta_p)\}_{j=0}^{K_p} \ \Rightarrow\ M_1(K_q) \geq M_1(K_p). \tag{32}$$

② *Proof of* $M_2(K_p) \leq M_2(K_q)$: According to Lemma A.18, for any $j_p = 1, \cdots, \left\lceil \frac{1}{\delta_p} \right\rceil - 2K_p$, there exists $j_q^{(p)} = \lceil j_p \cdot 2^{q-p} \rceil \in \{1, \cdots, \left\lceil \frac{1}{\delta_q} \right\rceil - 2K_q\}$, such that $\mathcal{Y}_{K_q}^{(j_q)} \subseteq \mathcal{Y}_{K_p}^{(j_p)}$. By Lemma A.14, we have

$$\min_{k \in \{j_q^{(j_p)}, \cdots, j_q^{(j_p)} + 2K_q - 1\}} \Delta\hat{F}(k\epsilon_q) \geq \min_{k \in \{j_p, \cdots, j_p + 2K_p - 1\}} \Delta\hat{F}(k\epsilon_p).$$

Note that $\left\{ j_q^{(j_p)} : j_p \in \left\{1, \cdots, \left\lceil \frac{1}{\delta_p} \right\rceil - 2K_p \right\} \right\} \subseteq \left\{1, \cdots, \left\lceil \frac{1}{\delta_q} \right\rceil - 2K_q \right\}$. Using Lemma A.14 again, we have

$$\begin{aligned} \max_{j_q \in \{1, \cdots, \left\lceil \frac{1}{\delta_q} \right\rceil - 2K_q\}} \min_{k \in \{j_q, \cdots, j_q + 2K_q - 1\}} \Delta\hat{F}(k\epsilon_q) &\geq \max_{j_p \in \{1, \cdots, \left\lceil \frac{1}{\delta_p} \right\rceil - 2K_p\}} \min_{k \in \{j_p, \cdots, j_p + 2K_p - 1\}} \Delta\hat{F}(k\epsilon_p), \\ \Rightarrow \quad M_2(K_q) &\geq M_2(K_p). \end{aligned} \tag{33}$$

Combining Eq. (31), Eq. (32), and Eq. (33), we obtain that

$$\widetilde{\mathrm{MCDP}}(\epsilon; K_p) = \max\{M_1(K_p), M_2(K_p)\} \leq \max\{M_1(K_q), M_2(K_q)\} \leq \widetilde{\mathrm{MCDP}}(\epsilon; K_q).$$

This completes the proof of Theorem 3.4.

□

A.3. Proofs of Theorem 3.5

Theorem 3.5. $\Delta\tilde{F}_\tau(\hat{y}) \xrightarrow{\text{a.e.}} \Delta\hat{F}(\hat{y})$ as $\tau \to \infty$, and in particular, $\lim_{\tau\to\infty} \Delta\tilde{F}_\tau(\hat{y}) = \Delta\hat{F}(\hat{y}),\ \forall \hat{y} \notin \{\hat{y}_i\}_{i=1}^N$.

Proof. Note that for any $i = 1, \cdots, N$, we have

$$\lim_{\tau\to\infty} \sigma_\tau(\hat{y} - \hat{y}_i) = \frac{1}{1 + \exp(-\tau(\hat{y} - \hat{y}_i))} = \begin{cases} 1, & \text{if } \hat{y}_i < \hat{y}, \\ \frac{1}{2}, & \text{if } \hat{y}_i = \hat{y}, \\ 0, & \text{if } \hat{y}_i > \hat{y}. \end{cases}$$

Thus for any $\hat{y} \notin \{\hat{y}_i\}_{i=1}^N$, the following holds

$$\lim_{\tau\to\infty} \sum_{i \in \mathcal{S}_a} \sigma_\tau(\hat{y} - \hat{y}_i) = \sum_{i \in \mathcal{S}_a} \lim_{\tau\to\infty} \sigma_\tau(\hat{y} - \hat{y}_i) = \sum_{i \in \mathcal{S}_a} \mathbb{I}(\hat{y}_i \leq \hat{y}),\ a \in \{0, 1\}. \tag{34}$$

Then we have

$$\lim_{\tau\to\infty} \Delta\tilde{F}_\tau(\hat{y}) = \left| \frac{1}{|\mathcal{S}_0|} \sum_{i \in \mathcal{S}_0} \mathbb{I}(\hat{y}_i \leq \hat{y}) - \frac{1}{|\mathcal{S}_1|} \sum_{i \in \mathcal{S}_1} \mathbb{I}(\hat{y}_i \leq \hat{y}) \right| = |\hat{F}_0(\hat{y}) - \hat{F}_1(\hat{y})| = \Delta\hat{F}(\hat{y}), \quad \forall \hat{y} \notin \{\hat{y}_i\}_{i=1}^N. \tag{35}$$

Note that $\{\hat{y}_i\}_{i=1}^N$ has measure zero, which further yields that $\Delta\tilde{F}_\tau(\hat{y}) \xrightarrow{\text{a.e.}} \Delta\hat{F}(\hat{y})$. □

B. Estimation Error Analysis of $\widehat{\mathrm{MCDP}}(\epsilon)$ Metric

To conduct analysis of estimation error of the $\widehat{\mathrm{MCDP}}(\epsilon)$ metric, we start with giving two lemmas that will be used in proofs.

Lemma B.1. *For real numbers* $a, b, c, d \in \mathbb{R}$, $||a - b| - |c - d|| \leq |a - c| + |b - d|$ *holds.*

Lemma B.2. *For* $a, b, c, d \in \mathbb{R}$ *where* $a \leq b$ *and* $c \leq d$, $\max\{|a - c|, |b - d|\} \leq \max\{|b - c|, |a - d|\}$ *holds.*

Proof. If $b \leq d$, then we have $|b-d| = d-b \leq d-a \leq \max\{|b-c|, |a-d|\}$. Likewise, $d \leq b$ derives that $|d-b| = b-d \leq b-c \leq \max\{|b-c|, |a-d|\}$. Thus the Lemma B.2 gets proved. □

Next we provide the statement of the Glivenko–Cantelli theorem as below:

Theorem B.3 (Glivenko-Cantelli)**.** *(Tucker, 1959) Let $X_1, \cdots, X_n$ be* i.i.d. *random variables in $\mathbb{R}$ with common cumulative distribution function $F(x) = \mathbb{P}(X_1 \leq x)$. Let $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x)$ be the empirical distribution function. Then*

$$\mathbb{P}\left\{\sup_{x\in\mathbb{R}} |F(x) - \hat{F}_n(x)| > \delta\right\} \leq 8(n+1)e^{-\frac{n\delta^2}{32}}. \tag{36}$$

In particular, by the Borel-Cantelli Lemma (Chung & Erdös, 1952), $\hat{F}_n(x) \xrightarrow{a.s.} F(x),\ n \to \infty$, i.e.,

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F(x) - \hat{F}_n(x)| = 0.$$

By Theorem B.3, we have the following corollaries in the binary classification task's setting of this work:

Corollary B.4. *Suppose the instances $\{\boldsymbol{x}_i\}_{i=1}^N$ in $\mathcal{D}$ are* i.i.d.*, then*

$$\hat{F}_{a,n_a}(\hat{y}) \xrightarrow{a.s.} F_a(\hat{y}),\ n_a \to \infty,\ a = 0, 1,$$

*where n_a denotes the number of instances with group label a in $\mathcal{D}$ (*i.e.*, $n_a = |\mathcal{S}_a|$), and $\hat{F}_{a,n_a}(\hat{y})$ is the empirical distribution function of predictions of group a when estimated by N_a samples.*

Corollary B.5. *Denote $\Delta\hat{F}_{n_0,n_1}(\hat{y}) = |\hat{F}_{0,n_0}(\hat{y}) - \hat{F}_{1,n_1}(\hat{y})|$, then we have*

$$\Delta\hat{F}_{n_0,n_1}(\hat{y}) \xrightarrow{a.s.} \Delta F(\hat{y}),\ n_0 \to \infty,\ n_1 \to \infty. \tag{37}$$

Proof. According to Lemma B.4, for any $\delta > 0$ and $a \in \{0, 1\}$, there exists $N_a(\delta) > 0$, such that

$$\max_{\hat{y}\in[0,1]} |\hat{F}_{a,n_a}(\hat{y}) - F_a(\hat{y})| \leq \delta, \quad \text{for any } n_a > N_a(\epsilon). \tag{38}$$

Let $n_0 > N_0(\frac{\delta}{2})$ and $n_1 > N_1(\frac{\delta}{2})$, we have

$$\begin{aligned}
&\max_{\hat{y}\in[0,1]} |\Delta\hat{F}_{n_0,n_1}(\hat{y}) - \Delta F(\hat{y})| \\
=& \max_{\hat{y}\in[0,1]} \Big| |\hat{F}_{0,n_0}(\hat{y}) - \hat{F}_{1,n_1}(\hat{y})| - |F_0(\hat{y}) - F_1(\hat{y})| \Big| \\
\overset{(a)}{\leq}& \max_{\hat{y}\in[0,1]} \Big(|\hat{F}_{0,n_0}(\hat{y}) - \hat{F}_0(\hat{y})| + |\hat{F}_{1,n_1}(\hat{y}) - \hat{F}_1(\hat{y})| \Big) \\
\leq& \max_{\hat{y}\in[0,1]} |\hat{F}_{0,n_0}(\hat{y}) - \hat{F}_0(\hat{y})| + \max_{\hat{y}\in[0,1]} |\hat{F}_{1,n_1}(\hat{y}) - \hat{F}_1(\hat{y})| \\
\overset{(b)}{\leq}& \frac{\delta}{2} + \frac{\delta}{2} = \delta,
\end{aligned} \tag{39}$$

where inequality (a) holds due to Lemma B.1, and inequality (b) holds due to Eq. (38). By Eq. (39), we have

$$\lim_{\substack{n_0\to\infty \\ n_1\to\infty}} \sup_{\hat{y}\in[0,1]} \Big| \Delta\hat{F}_{n_0,n_1}(\hat{y}) - \Delta F(\hat{y}) \Big| = 0,$$

which yields that $\Delta\hat{F}_{n_0,n_1}(\hat{y}) \xrightarrow{\text{a.s.}} \Delta F(\hat{y})$. □

For notation clarity, let $M(\epsilon) = \text{MCDP}(\epsilon)$ and $\hat{M}_{n_0,n_1}(\epsilon) = \widehat{\text{MCDP}}(\epsilon) = \max_{y_0\in[0,1]} \min_{|\hat{y}-y_0|\leq\epsilon} \Delta\hat{F}_{n_0,n_1}(\hat{y})$. We have the following theorem about the estimation error convergence rate of the empirical and true metrics:

Theorem B.6 (Convergence of $\widehat{\mathrm{MCDP}}(\epsilon)$)**.** *As the sample size n_0 and n_1 of two demographic groups increase, the value of $\widehat{\mathrm{MCDP}}(\epsilon)$ metric converges with probability one to the true metric value $\mathrm{MCDP}(\epsilon)$,* i.e.,

$$\hat{M}_{n_0,n_1}(\epsilon) \xrightarrow{a.s.} M(\epsilon),\ n_0 \to \infty,\ n_1 \to \infty. \tag{40}$$

Additionally, the estimation error satisfies

$$Err_{emp} = \mathbb{E}_{\mathcal{D}}\left[|\hat{M}_{n_0,n_1}(\epsilon) - M(\epsilon)|^2\right] = \mathcal{O}\left(\frac{\ln n}{n}\right),\ n = \min\{n_0, n_1\}. \tag{41}$$

Proof. Let $y^* = \arg\max_{y^*\in[0,1]} \Delta F(\hat{y})$ and $\hat{y}^* = \arg\max_{y^*\in[0,1]} \Delta\hat{F}_{n_0,n_1}(\hat{y})$, then both ① $\Delta F(\hat{y}^*) \le \Delta F(y^*)$ and ② $\Delta\hat{F}_{n_0,n_1}(y^*) \le \Delta\hat{F}_{n_0,n_1}(\hat{y}^*)$ hold. By Lemma B.2 and Eq. (39), given $\delta > 0$, for any $n_0 > N_0(\delta), n_1 > N_1(\delta)$, we have

$$\begin{aligned}
|M(0) - \hat{M}_{n_0,n_1}(0)| &= |\Delta F(y^*) - \Delta\hat{F}_{n_0,n_1}(\hat{y}^*)| \\
&\le \max\left\{|\Delta F(y^*) - \Delta\hat{F}_{n_0,n_1}(y^*)|, |\Delta F(\hat{y}^*) - \Delta\hat{F}_{n_0,n_1}(\hat{y}^*)|\right\} \\
&\le \max_{\hat{y}\in[0,1]} |\Delta F(\hat{y}) - \Delta\hat{F}_{n_0,n_1}(\hat{y})| \le \delta.
\end{aligned} \tag{42}$$

As to the cases when $\epsilon > 0$, we denote $G(y_0),\ y_0^*,\ \hat{G}_{n_0,n_1}(y_0),\ \widehat{y_0^*}$ as follows

$$\begin{aligned}
G(y_0) &= \min_{|\hat{y}-y_0|\le\epsilon} \Delta F(\hat{y}), & y_0^* &= \arg\max_{y_0\in[0,1]} G(y_0), \\
\hat{G}_{n_0,n_1}(y_0) &= \min_{|\hat{y}-y_0|\le\epsilon} \Delta\hat{F}_{n_0,n_1}(\hat{y}), & \widehat{y_0^*} &= \arg\max_{y_0\in[0,1]} \hat{G}_{n_0,n_1}(y_0).
\end{aligned}$$

Thus we have $G(\widehat{y_0^*}) \le G(y_0^*)$ and $\hat{G}_{n_0,n_1}(y_0^*) \le \hat{G}_{n_0,n_1}(\widehat{y_0^*})$. Moreover, for any $y_0' \in [0,1]$, there exists $\hat{y}', \hat{y}'' \in [y_0' - \epsilon, y_0' + \epsilon] \cap [0,1]$, such that $G(y_0') = \Delta F(\hat{y}') \le \Delta F(\hat{y}'')$ and $\hat{G}(y_0') = \Delta\hat{F}_{n_0,n_1}(\hat{y}'') \le \Delta\hat{F}_{n_0,n_1}(\hat{y}')$. By Lemma B.2 and Eq. (39), given $\delta > 0$ and $\epsilon > 0$, for any $n_0 > N_0(\delta), n_1 > N_1(\delta)$, we have

$$\begin{aligned}
|M(\epsilon) - \hat{M}_{n_0,n_1}(\epsilon)| &= |G(y_0^*) - \hat{G}_{n_0,n_1}(\widehat{y_0^*})| \\
&\le \max\left\{|G(y_0^*) - \hat{G}_{n_0,n_1}(y_0^*)|, |G(\widehat{y_0^*}) - \hat{G}_{n_0,n_1}(\widehat{y_0^*})|\right\} \\
&\le \max_{y_0\in[0,1]} |G(y_0) - \hat{G}_{n_0,n_1}(y_0)| \overset{(c)}{=} |\Delta F(\hat{y}') - \Delta\hat{F}_{n_0,n_1}(\hat{y}'')| \\
&\le \max\left\{|\Delta F(\hat{y}'') - \Delta\hat{F}_{n_0,n_1}(\hat{y}'')|, |\Delta F(\hat{y}') - \Delta\hat{F}_{n_0,n_1}(\hat{y}')|\right\} \\
&\le \max_{\hat{y}\in[0,1]} |\Delta F(\hat{y}) - \Delta\hat{F}_{n_0,n_1}(\hat{y})| \le \delta,
\end{aligned} \tag{43}$$

Equality (c) holds by taking $y_0' = \arg\max_{y_0\in[0,1]} |G(y_0) - \hat{G}_{n_0,n_1}(y_0)|$. Combining Eq. (42) and Eq. (43) yields that $\hat{M}_{n_0,n_1}(\epsilon)$ converges to $M(\epsilon)$ almost surely (*i.e.,* Eq. (40) holds). Furthermore, let Eq. (36) equal to 1, we have $\delta = \mathcal{O}\left(\sqrt{\frac{\ln n}{n}}\right)$. Combining Eq. (42), Eq. (43) and Eq. (39), for any $\epsilon \ge 0$, the estimation error should satisfy that

$$\begin{aligned}
Err_{\text{emp}} &= \mathbb{E}_{\mathcal{D}}\left[|\hat{M}_{n_0,n_1}(\epsilon) - M(\epsilon)|^2\right] \\
&\le \left(\max_{\hat{y}\in[0,1]} |\Delta F(\hat{y}) - \Delta\hat{F}_{n_0,n_1}(\hat{y})|\right)^2 \\
&\le \left(\max_{\hat{y}\in[0,1]} |\hat{F}_{0,n_0}(\hat{y}) - \hat{F}_0(\hat{y})| + \max_{\hat{y}\in[0,1]} |\hat{F}_{1,n_1}(\hat{y}) - \hat{F}_1(\hat{y})|\right)^2 \\
&= \left(\mathcal{O}\left(\sqrt{\frac{\ln n_0}{n_0}}\right) + \mathcal{O}\left(\sqrt{\frac{\ln n_1}{n_1}}\right)\right)^2 \\
&= \mathcal{O}\left(\frac{\ln n}{n}\right),\ n = \min\{n_0, n_1\}.
\end{aligned}$$

Thus Eq. (41) gets proved. This completes the proof of Theorem B.6. □

C. Detailed Analysis about the Computational Complexity

In Section 3.3, we conduct a rough computational complexity analysis of the exact and approximate $\widehat{\mathrm{MCDP}}(\epsilon)$ calculation algorithms, where we only focus on the traverse strategies of y_0 and $\hat{y}$ in Eq. (4). Here we provide a more thorough computational complexity analysis of both algorithms.

Exact Calculation Analysis. In Algorithm 1, the calculation process can be divided into the following steps.

(1) **Calculating $\Delta\hat{F}(\hat{y})$ (line 1).** As $\Delta\hat{F}(\hat{y}) = |\hat{F}_0(\hat{y}) - \hat{F}_1(\hat{y})|$, this step is equivalent to calculating the empirical distribution function $\hat{F}_a(\hat{y})$ using the predictions $\{\hat{y}_i\}_{i\in\mathcal{S}_a}$ in two groups ($a = 0, 1$). Due to its step-like pattern's property, the empirical distribution function is implemented by constructing a sorted array of sample values with the corresponding cumulative ratio in practice (*e.g.,* the Statsmodels library (Seabold & Perktold, 2010) in Python), and then perform binary-search on the array given a specific value. In line 1, the complexity of initializing the sorted array structure is $\mathcal{O}(|\mathcal{S}_0|\log|\mathcal{S}_0|) + \mathcal{O}(|\mathcal{S}_1|\log|\mathcal{S}_1|) = \mathcal{O}(N\log N)$.

(2) **Initializing the $\widetilde{\mathrm{MCDP}}(\epsilon)$ (lines 2-6).** When $\epsilon = 0$, the algorithm returns the maximal $\Delta\hat{F}(\hat{y})$ value on predictions of data samples, and its computational complexity is $\mathcal{O}(N)$. In cases where $\epsilon > 0$, the $\widetilde{\mathrm{MCDP}}(\epsilon)$ value is initialized with the minimal $\Delta\hat{F}(\hat{y})$ prediction smaller than or equal to ϵ, whose complexity is also $\mathcal{O}(N)$.

(3) **Traversing y_0 and $\hat{y}$ when $\epsilon > 0$ (lines 7-12).** For each prediction $\hat{y}_i \leq 1 - \epsilon$, the algorithm finds the minimal $\Delta\hat{F}(\hat{y})$ value of predictions in the 2ϵ interval with left endpoint $\hat{y}_i$, and finally keeps the maximum of the selected minimums. As the above process involves a double traversal of predictions in $\mathcal{D}$, its complexity is $\mathcal{O}(N^2)$.

In summary, the overall computational complexity when $\epsilon = 0$ is $\mathcal{O}(N\log N) + \mathcal{O}(N) = \mathcal{O}(N\log N)$. When $\epsilon > 0$, the total complexity is $\mathcal{O}(N\log N) + \mathcal{O}(N) + \mathcal{O}(N^2) = \mathcal{O}(N^2)$, most of which is made up by the traverse process.

Approximate Calculation Analysis. Similar to the exact calculation, Algorithm 2 firstly calculates $\Delta\hat{F}(\hat{y})$ using the predictions of data samples in two groups (line 1), whose computational complexity is $\mathcal{O}(N\log N)$. Next, the algorithm samples equally-spaced prediction points by step-size $\delta = \frac{\epsilon}{K}$ on $[0, 1]$, where the total number of samples is $\mathcal{O}(\frac{1}{\delta})$. Afterwards, it firstly initializes $\widetilde{\mathrm{MCDP}}(\epsilon)$ with the minimal $\Delta\hat{F}(\hat{y})$ value of the first $K + 1$ sampled points (line 3), and then updates it with the maximum of the minimal $\Delta\hat{F}(\hat{y})$ value of all consecutive $2K$ sampled points (lines 4-7). The computational complexity of initialization and traverse update are $\mathcal{O}(K)$ and $\mathcal{O}(\frac{K}{\delta})$, respectively. To conclude, the total complexity of Algorithm 2 is $\mathcal{O}(N\log N) + \mathcal{O}(\frac{1}{\delta}) + \mathcal{O}(K) + \mathcal{O}(\frac{K}{\delta}) = \max\{\mathcal{O}(N\log N), \mathcal{O}(\frac{K^2}{\epsilon})\}$. Specifically, when the sampling frequency K is set with small values such that $K \ll N$, the calculation of $\Delta\hat{F}(\hat{y})$ will make up more complexity than traversing sampled predictions, and its total complexity will be much smaller than that of exact algorithm.

D. Additional Experimental Settings

Implementation Details. To make our work standardized and extensible, the implementations are based on a latest open-sourced fairness benchmark, FFB (Han et al., 2024). We conduct experiments with a 96-core Intel CPU (Intel(R) Xeon(R) Platinum 8268 @ 2.90GHz * 2) and a Nvidia-2080Ti GPU (11 GB memory). For tabular datasets Adult (Kohavi, 1996) and Bank (Moro et al., 2014), we use a two-layer MLP (Bishop, 1995) with 256 hidden neurons and ReLU activation function (Nair & Hinton, 2010) as the classifier; for the image dataset CelebA (Liu et al., 2015) (which derives two meta-datasets CelebA-A and CelebA-W), we adopt Resnet-18 (He et al., 2016) with as the backbone model, and initialize it with pretrained weights. The batch size for tabular and image datasets are set as 1024 and 128, respectively, and the total training step is set as 150. We use the Adam optimizer with initial learning rate 0.001, which is decayed by the piecewise strategy (*i.e.,* StepLR scheduler in Pytorch (Paszke et al., 2019)) during training. We use average precision (AP) to evaluate classification accuracy, and adopt different metrics ($\Delta\mathrm{DP}$, ABCC, $\mathrm{MCDP}(\epsilon)$) to measure algorithmic fairness. For $\Delta\mathrm{DP}$, we follow Creager et al. (2019); Dai & Wang (2021); Chen et al. (2024) to compute the difference of positive prediction proportion of two groups. The code implementation of this paper is available at `https://github.com/mitao-cat/icml24_mcdp`.

Hyper-Parameters of Baselines. As fairness-accuracy trade-off is a widely-exist phenomenon in various applications (Wick et al., 2019; Dutta et al., 2020), the hyper-parameters of different fairness algorithms should be carefully selected to ensure reliable comparisons. Generally, all hyper-parameters can be divided into two categories based on whether they directly control the fairness and accuracy (*trade-off HPs*) or not (*other HPs*). For each compared method, we firstly perform

Table 3. Hyper-parameter settings for different methods and datasets.

Method	Other HP	Selected 5 Trade-off HP Values on Each Dataset
AdvDebias	N/A	Adult: $[0.2, 0.4, 0.6, 0.8, 1.0]$; Bank: $[0.3, 0.6, 0.9, 1.2, 1.5]$; CelebA-A, CelebA-W: $[0.5, 1.0, 1.5, 2.0, 2.5]$
DiffDP	N/A	All datasets: $[0.1, 0.2, 0.3, 0.4, 0.5]$
FairMixup	N/A	Adult: $[0.1, 0.2, 0.3, 0.4, 0.5]$; Bank: $[0.3, 0.4, 0.5, 0.6, 0.7]$ CelebA-A: $[0.5, 1, 2, 3, 4]$; CelebA-W: $[0.5, 1.0, 1.5, 2.0, 2.5]$
DRAlign	Alignment strength β	All datasets: $[0.1, 0.2, 0.3, 0.4, 0.5]$
DiffABCC	Temperature τ	All datasets: $[0.1, 0.2, 0.3, 0.4, 0.5]$
DiffMCDP	Temperature τ	All datasets: $[0.1, 0.15, 0.2, 0.25, 0.3]$

grid search on a wide value range for both trade-off HPs and other HPs, and then select the optimal other HP value that achieves the best trade-off performance on each dataset. Afterwards, we select 5 trade-off HP values that can well reflect the fairness-accuracy trade-off trend for each algorithm on the validation set. We additionally require that (i) both fairness and accuracy metrics of different algorithms are within a similar value range, and (ii) the classification accuracy should not drop too sharply. When using a single value for model evaluation, following previous work (Jung et al., 2023), we select the optimal trade-off HP value which achieves at least 95% of the vanilla model's accuracy (*i.e.,* AP of ERM) on validation set. We summarize the detailed hyper-parameters in Table 3.

E. Supplementary Experimental Results

E.1. Fairness-Accuracy Trade-offs

In Section 4.1, we show the trade-offs between AP and $\mathrm{MCDP}(0)$ metrics of baselines and our proposed method. To further investigate their performances, we also plot the fairness-accuracy trade-off curves when adopting $\Delta\mathrm{DP}$ and ABCC as fairness metrics in Figures 8 and 9, respectively. We can observe that DiffMCDP can achieve comparable or even better trade-off performances compared with other baselines, and in particular, DiffDP achieves the optimal results in terms of both $\Delta\mathrm{DP}$ and ABCC fairness metrics on Bank dataset. This indicates that optimizing the maximal disparity is also beneficial to minimize the overall disparity, which is consistent with the results in Table 1.

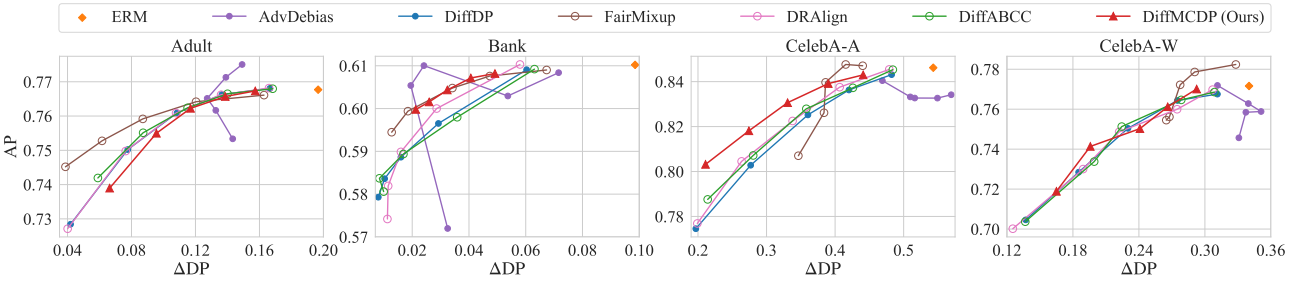


Figure 8. Trade-offs between AP and $\Delta\mathrm{DP}$ of baselines and the proposed method.

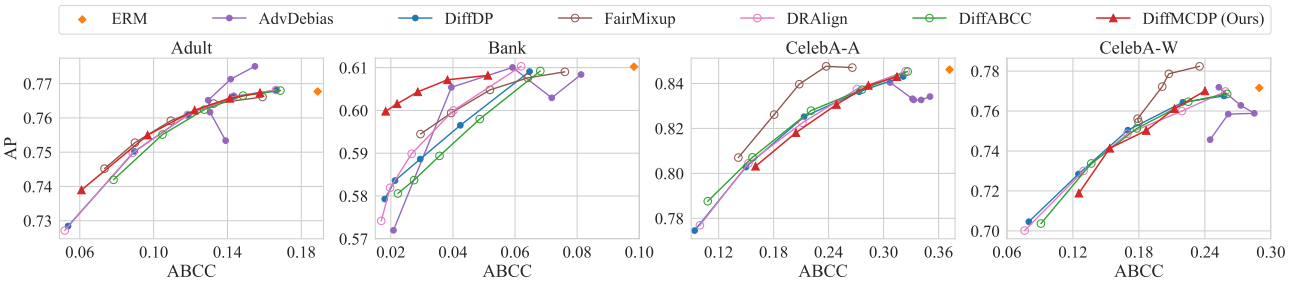


Figure 9. Trade-offs between AP and ABCC of baselines and the proposed method.

E.2. Varying Local Measurements ϵ

In Figure 4 of Section 4.1, we report the different $\mathrm{MCDP}(\epsilon)$ results of each algorithm, in which the relative performance of some baselines changes with increasing ϵ values. To further explore the effect of varying neighborhood hyper-parameter, we plot the fairness-accuracy trade-off curves with ϵ in $\{0.01, 0.05, 0.1\}$. Figures 10 and 11 show the results of Adult and Bank datasets, respectively, from which we observe that DiffMCDP always outperforms other baselines in terms of different $\mathrm{MCDP}(\epsilon)$ metrics. Moreover, the trade-off patterns of some baselines may change with ϵ values, which brings about changing relative performances evaluated by $\mathrm{MCDP}(\epsilon)$. This further reminds practitioners to select varying ϵ values for more comprehensive evaluation of different fair algorithms.

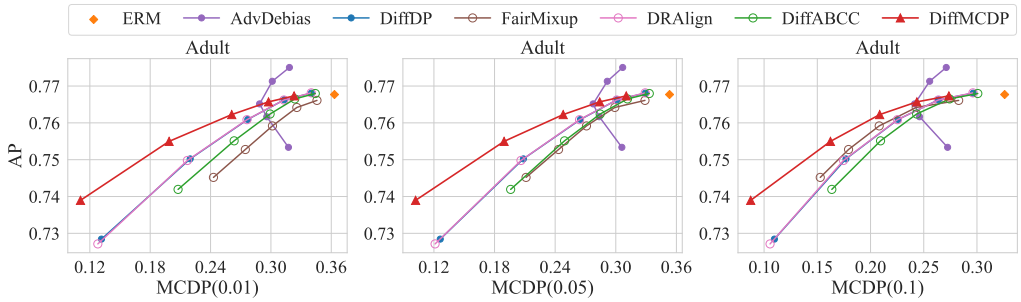


Figure 10. Trade-offs between AP and $\mathrm{MCDP}(\epsilon)$ with varying neighborhood hyper-parameter ϵ on Adult dataset.

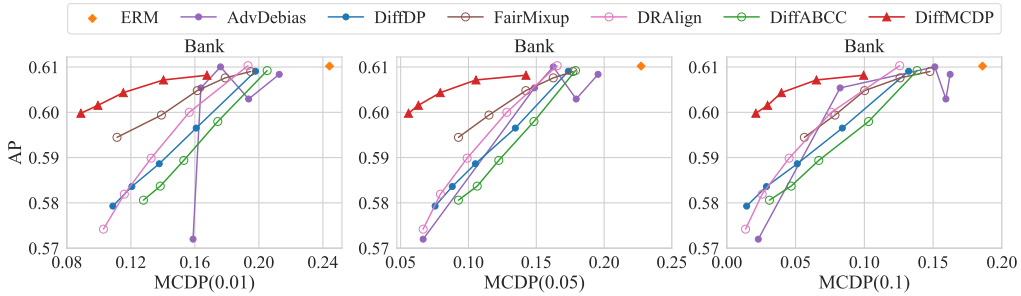


Figure 11. Trade-offs between AP and $\mathrm{MCDP}(\epsilon)$ with varying ϵ values on Bank dataset.

Additionally, we plot the $\mathrm{MCDP}(\epsilon)$ results with varying ϵ values on CelebA dataset in Figure 12, which shows that our proposed DiffMCDP algorithm can consistently achieve lower maximal local disparity than other baselines. Additionally, compared to the results of the tabular data in Figure 4, the differences in the magnitude of $\mathrm{MCDP}(\epsilon)$ values under varying ϵ are much lower in Figure 12, and the relative performance of all methods does not change. We postulate that the CDFs of predictions of two groups on tabular datasets are sharper than those of image datasets, thus the value of neighborhood hyper-parameter ϵ is more prone to affect the value of maximal local disparity.

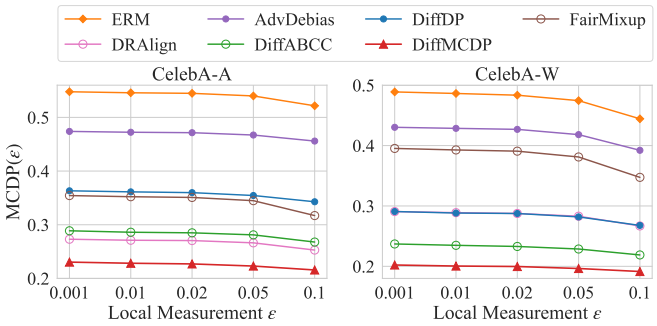


Figure 12. Comparison of $\mathrm{MCDP}(\epsilon)$ results with varying ϵ on image datasets.

E.3. $\widehat{\mathrm{MCDP}}(\epsilon)$ Calculation Algorithms

Continuing from Section 4.2, we show the effect of sampling frequency K and local measurement ϵ on the estimation accuracy and calculation efficiency of $\widehat{\mathrm{MCDP}}(\epsilon)$ calculation algorithms. The performance comparison on Bank and CelebA-W datasets are reported in Figures 13 and 14, respectively. As the exact $\widehat{\mathrm{MCDP}}(0.1)$ values of a small fraction of cases in Bank are very close to zero (< 0.001), which results in extreme relative errors, we report the *absolute error* $V_a - V_e$ in Figure 13 instead. Similar to the observations from Figure 5, the approximate algorithm can greatly improve the calculation efficiency with low error.

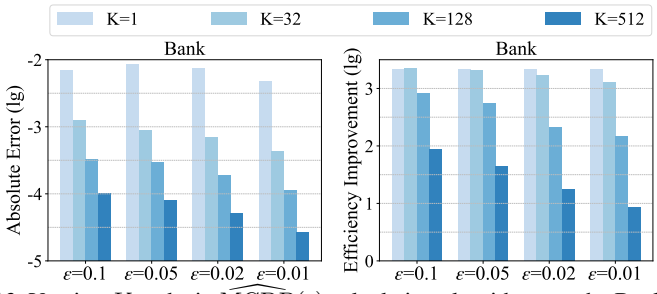


Figure 13. Varying K and ϵ in $\widehat{\mathrm{MCDP}}(\epsilon)$ calculation algorithms on the Bank dataset.

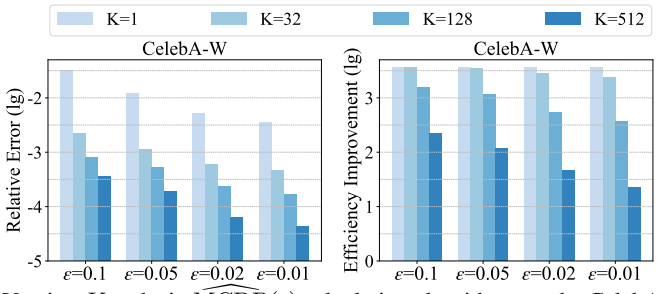


Figure 14. Varying K and ϵ in $\widehat{\mathrm{MCDP}}(\epsilon)$ calculation algorithms on the CelebA-W dataset.

To further illustrate the importance of improving computational efficiency, we report the running time of Algorithm 1 T_e in Table 4. We can observe that the exact calculation algorithm exhibits prolonged execution time in a single run. In research and industry scenarios which involves frequent model evaluations, the total evaluation time for multiple experiments would be unaffordable if the exact algorithm is employed. Nevertheless, the runtime can be shortened to hundreds or thousands of times using the approximate calculation algorithm (*i.e.,* $T_a \ll T_e$), which successfully resolves the efficiency issue.

Table 4. The single-run execution time T_e of the exact calculation algorithm.

Dataset	**Adult**	**Bank**	**CelebA-A**	**CelebA-W**
Time (s)	3.529±0.122	2.900±0.006	7.748±0.248	7.745±0.169

E.4. Varying Temperature τ

Continuing from Section 4.3, we show the effect of varying temperature τ on the performance of DiffMCDP algorithm on CelebA dataset in Figure 15. Similarly, very large or small temperatures ($\tau = 5, 50$) causes sub-optimal results, whereas moderate temperatures ($\tau = 10, 20$) more effectively trade-off the accuracy and fairness.

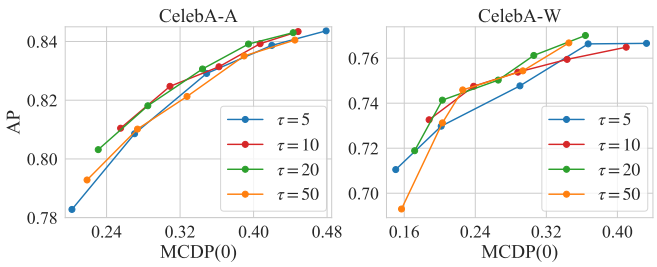


Figure 15. Trade-offs of DiffMCDP with varying temperature τ on image datasets.

F. Implementation of $\widehat{\mathrm{MCDP}}(\epsilon)$ Calculation Algorithms

In this section, we provide the implementations of the exact and approximate calculation of $\widehat{\mathrm{MCDP}}(\epsilon)$ metric in Algorithms 4 and 5, respectively, which shows that our proposed metric can be employed in fairness evaluation of various scenarios.

Algorithm 4 Implementation of the Exact Calculation of $\widehat{\mathrm{MCDP}}(\epsilon)$ in Algorithm 1

```
1  from statsmodels.distributions.empirical_distribution import ECDF
2
3  def MCDP_exact(y_pred, s, epsilon):
4     # y_pred: continuous model predictions in [0,1]
5     # s: binary sensitive attribute in {0,1}
6     # epsilon: neighborhood hyper-parameter
7
8     y_pred, s = y_pred.ravel(), s.ravel()
9     y_pre_1, y_pre_0 = y_pred[s==1], y_pred[s==0]
10    # calculate the empirical distribution functions (line 1)
11    ecdf0, ecdf1 = ECDF(y_pre_0), ECDF(y_pre_1)
12
13    y_pred = np.r_[0,y_pred,1] # line 2
14    if epsilon == 0:
15       # return the maximal deltaF value of predictions (line 4)
16       mcdp = np.max(np.abs(ecdf0(y_pred)-ecdf1(y_pred)))
17    else:
18       y_pred = np.sort(y_pred) # for indice slicing operations
19       # initialize the mcdp value (line 6)
20       init_ypred = y_pred[y_pred <= epsilon]
21       mcdp = np.min(np.abs(ecdf0(init_ypred) - ecdf1(init_ypred)))
22
23       # traverse all predictions on instances in D (lines 7-9)
24       I = y_pred[y_pred <= 1-epsilon].reshape((-1,1)) # line 7
25       # select predictions in 2eps intervals with left endpoints in I (lines 8-9)
26       J = np.where(I<=y_pred,1,0) * np.where(I+2*epsilon>=y_pred,1,0)
27       J = J.reshape(-1) * np.tile(np.arange(len(y_pred))+1,len(I))
28       J = J.reshape((len(I),len(y_pred)))
29
30       # calculate the minimal deltaF values of each 2eps intervals (line 10)
31       delta_ecdf = np.r_[1.1, np.abs(ecdf0(y_pred) - ecdf1(y_pred))]
32       delta_ecdf = delta_ecdf[J] # (len(I),len(y_pred))
33       min_delta_ecdf = np.min(delta_ecdf,axis=1)
34       # keep the maximum of minimal deltaF values (line 11)
35       mcdp = np.max(mcdp, np.max(min_delta_ecdf))*100
36
37    return mcdp
```

Algorithm 5 Implementation of the Approximate Calculation of $\widehat{\mathrm{MCDP}}(\epsilon)$ in Algorithm 2

```
from statsmodels.distributions.empirical_distribution import ECDF

def MCDP_approximate(y_pred, s, epsilon, K):
    # y_pred: continuous model predictions in [0,1]
    # s: binary sensitive attribute in {0,1}
    # epsilon: neighborhood hyper-parameter
    # K: sampling frequency

    assert epsilon>0
    y_pred, s = y_pred.ravel(), s.ravel()
    y_pre_1, y_pre_0 = y_pred[s==1], y_pred[s==0]

    # calculate the empirical distribution functions (line 1)
    ecdf0, ecdf1 = ECDF(y_pre_0), ECDF(y_pre_1)
    delta = epsilon / K      # set the step-size (line 2)
    samples = np.arange(0,1,delta) # sampling prediction points

    # initialize the mcdp value (line 3)
    mcdp = np.min(np.abs(ecdf0(samples[:K+1])-ecdf1(samples[:K+1])))
    # pre-compute the deltaF values for sampled points
    delta_ecdf = np.abs(ecdf0(samples) - ecdf1(samples))

    # traverse all consecutive 2K sampled points (lines 4-5)
    indices = np.arange(1,np.ceil(1/delta)-2*K+1).astype(int) # line 4
    indices = indices + np.arange(2*K).reshape((-1,1)) # line 5
    # keep the maximum of the minimal deltaF value (line 6)
    mcdp = np.max(mcdp, np.max(np.min(delta_ecdf[indices],axis=0)))

    return mcdp
```