Fairness Uncertainty Quantification: A Constrained Stochastic Optimization Perspective

Abhishek Roy[1] **Prasant Mohapatra**[2]

[1]Department of Statistics, Texas A&M University, College Station, Texas, USA
[2]Department of Computer Science and Engineering, University of South Florida, Tampa, Florida, USA

Abstract

In this paper, we construct confidence interval (CI) for the test-unfairness of a group-fairness aware classifier trained by Stochastic Gradient Descent (SGD) type algorithms. Since these algorithms are almost invariably used to train modern machine learning models, the learned parameters—and consequently their fairness—are inherently random. We quantify this randomness, which can have far-reaching consequences, especially for sensitive applications. Viewing fairness-aware classification through the lens of constrained optimization allows us to leverage tools from stochastic approximation theory to establish the asymptotic normality of the estimated parameters. Existing work typically provides fairness guarantees either in expectation or for the exact minimizer of the constrained objective [del Barrio et al., 2020, Celis et al., 2019], without accounting for the stochasticity induced by the optimization algorithm itself. To construct CIs for fairness in practice, we develop an easily parallelizable online multiplier bootstrap procedure. In doing so, we also extend theoretical guarantees for online bootstrap methods from unconstrained SGD to the constrained optimization setting which may be of independent interest. We illustrate the effectiveness of our approach through extensive simulations and experiments on benchmark datasets.

1 INTRODUCTION

The widespread use of machine learning in high-stakes domains such as medicine Begoli et al. [2019], education Martinez Neda et al. [2021] and judiciary system Iosifidis et al. [2021], has raised serious concerns about the fairness of learned models. It is particularly challenging in policy-driven settings, where decisions must be made online as data arrive while simultaneously enforcing fairness constraints [Sun et al., 2019, Huang et al., 2022]. A range of heuristic methods and optimization-based frameworks have been developed to promote fairness in both offline [Jiang et al., 2020, del Barrio et al., 2020, Kamishima et al., 2011, Woodworth et al., 2017] and online settings [Sun et al., 2019, Iosifidis et al., 2021, Huang et al., 2022, Wang et al., 2023]. Online stochastic optimization algorithms such as SGD are routinely used to train such models because gradient descent is prohibitively expensive for modern large datasets. The randomness of the data distribution propagates through the stochastic training, leading to the randomness of the learned model, and its fairness properties. Existing work typically provides fairness guarantees either in expectation or for the exact minimizer of the loss, without accounting for the stochasticity induced by the training algorithm itself Celis et al. [2019]. Practitioners should account for the statistical uncertainty in the fairness of trained models Moreau and Weber [2024]. In policy-driven settings, real-time confidence intervals for fairness enable more reliable deployment of decisions Ramprasad et al. [2023], Chen et al. [2021], Schifano et al. [2016], Tewari and Murphy [2017].

Standard UQ methods such as Bayesian method, method of Functional Priors, and ensemble method Psaros et al. [2023] are either incapable of quantifying the randomness of the optimization algorithm or require repeated training which is costly, and more importantly, incompatible with real-time decision-making required in the online regime. There is a related line of work in the offline setting that tests if a given trained classifier is unfair with respect to any subpopulation Taskesen et al. [2021], DiCiccio et al. [2020]. In contrast, we quantify the randomness of the fairness of a classifier trained with an online stochastic algorithm. In summary, we answer the following question.

How to quantify the uncertainty in the fairness of a fairness-aware online classifier, i.e., with access to a single stream of data, when trained with an online stochastic optimization algorithm?

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:5845–5865.

Problem Setup To answer the above question, we choose to study the most widely-used group-fairness measures: Statistical Parity Corbett-Davies et al. [2017] or Disparate Impact (DI) and equality of opportunity Hardt et al. [2016] or Disparate Mistreatment (DM) Zafar et al. [2019]. For DI-aware model we use the constrained optimization framework used in [Zhao et al., 2021, Zafar et al., 2019, Radovanović et al., 2020]. For the DM-aware model, we consider a penalized version of the problem similar to [Huang and Vishnoi, 2019, Bechavod and Ligett, 2017] to avoid heuristics required to solve the optimization with convex-concave constraint. As loss functions, we use Cross-Entropy (CE) loss, and squared loss because of its recent success in classification Demirkaya et al. [2020], Hui and Belkin [2020].

Main Contributions.

(a) Our main conceptual novelty lies in viewing fairness-aware classification through the lens of constrained optimization, which allows us to borrow tools from stochastic approximation to establish the asymptotic distribution of both the model parameter and the test fairness (Proposition 4.1, Theorem 4.2, and Theorem 5.2). In doing so, we quantify the often-overlooked uncertainty induced by online SGD-type algorithms that are routinely used to train machine learning models. Our goal is not to design a new fairness algorithm, but to develop and analyze a real-time, practicable method for quantifying uncertainty in the fairness of a fairness-aware classifier trained on streaming data.

(b) The covariance of the asymptotic distribution depends on the unknown global optimum and must therefore be estimated online to construct confidence intervals for DI and DM. We propose an online multiplier bootstrap procedure (Algorithm 2) for constrained optimization, extending the approach of Fang et al. [2018] developed for unconstrained problems, to obtain asymptotically valid confidence intervals for DI and DM.

(c) Our main theoretical contributions are twofold. First, we show that the online multiplier bootstrap estimator converges in distribution to the asymptotic distribution of the model parameter (Theorem 5.1). Second, Theorem 5.2 establishes the asymptotic distribution of the fairness functional itself which is challenging as the function is nonsmooth. While our primary goal is principled UQ for fairness-aware learning, these results also advance online inference for constrained stochastic optimization, a problem of independent interest.

(d) We adopt the synthetic experiments studied in Zafar et al. [2019]. For real data applications, we use the benchmark datasets `Adult` Dua and Graff [2017] and `ACSIncome` Ding et al. [2021] for DI, and `COMPAS` Larson [2016] for DM, as they are widely studied in both offline [del Barrio et al., 2020, Le Quy et al., 2022] and online [Iosifidis et al., 2021, Aguiar et al., 2022, Iosifidis et al., 2019, Sun et al., 2019, Zhang and Zhao, 2020, Zhao et al., 2021] fairness literature.

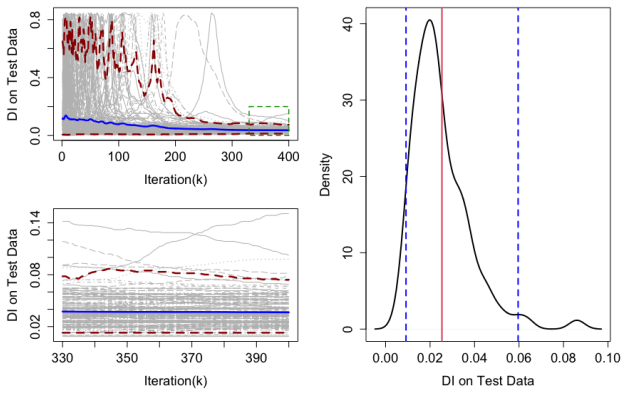


Figure 1: Towards the end of training, test DI varies substantially across independent trials with the same training data distribution, initialization, and test set, even though individual trajectories and the mean remain nearly constant (green box, bottom left). The right plot shows the density of the final DI values: while the mean is 0.025, the 97.5% quantile reaches 0.06, which is 140% larger.

In Section 2 we present related work. In Section 3, we introduce the notations and the problem statement. We state the results on asymptotic distribution of the model parameters in Section 4. We show our main theoretical result on online multiplier bootstrap-based covariance estimation in Section 5. We illustrate our results on real and synthetic datasets in Section 6. Section 7 concludes the paper.

1.1 MOTIVATIONAL TOY EXAMPLE

We illustrate the motivation with a toy example based on the synthetic dataset of Zafar et al. [2019], which is vulnerable to DI. Training a fairness-constrained classifier by an online SGD-type algorithm on an $\mathrm{i.i.d}$ data stream produces a mean test DI ~ 0.025. Suppose 200 practitioners independently repeat the training under the same data distribution. Figure 1 shows that even with same initialization, and test data, the resulting test DI of trained models differ noticeably across trials (green box in top left, magnified below) while individual trajectories and the mean stabilize. The right panel shows that the 97.5% quantile of DI at convergence is 140% larger than the mean, which can be unacceptable in sensitive applications. This variability arises from the stochastic nature of the optimization algorithm, which inherits randomness from the data stream. In practice, a single practitioner observes only one trajectory. Our online bootstrap method, with theoretical guarantees, quantifies this uncertainty in real time using only a single data stream.

2 RELATED WORK

UQ in optimization. Asymptotic properties of stochastic approximation algorithms have been studied in depth Polyak and Juditsky [1992], Kushner and Yin [2003], Benveniste et al. [2012]. [Chen et al., 2016, Zhu et al., 2021, Roy and Balasubramanian, 2023, Jiang et al., 2025, Roy and Ma, 2023], and [Fang et al., 2018, Ramprasad et al., 2023, Samsonov et al., 2025] propose online batch-means estimator and online bootstrap-based approaches respectively for online inference for SGD. However, these methods are not applicable to constrained settings. Moreover, unlike us, these works do not establish the asymptotic distribution of non-smooth functionals of the estimator such as DI or DM.

UQ in fairness. Maity et al. [2021] constructs an asymptotically valid CI for individual unfairness by constructing a gradient-flow-based adversarial test statistic. Our work differs from Maity et al. [2021] in three key aspects: (i) we study group fairness rather than individual fairness; (ii) we quantify uncertainty arising from stochastic training algorithms, rather than from adversarial perturbations; and (iii) their asymptotic normality follows from an $\mathrm{i.i.d}$ sum structure, whereas SGD-type algorithms induce a Markov chain even under $\mathrm{i.i.d}$ data, requiring more involved analysis. In particular, standard sample covariance estimators are not valid in our setting (see Zhu et al. [2021], Fang et al. [2018]). There is another line of work on testing whether a *given fixed classifier* is fair [Si et al., 2021, Taskesen et al., 2021, DiCiccio et al., 2020, Holstein et al., 2019, Cherian and Candès, 2023, Tang et al., 2026]. In contrast, we quantify the real-time undertainty of the fairness metric when the classifier itself exhibits random fluctuations induced by a stochastic training algorithm, with randomness arising from the underlying training data distribution.

Bootstrap methods. There is extensive work on resampling-based bootstrap for offline empirical risk minimizers in i.i.d. settings [Efron, 1979, van der Vaart and Wellner, 1996, Hall, 2013, Forneron and Ng, 2020]. Our setting is fundamentally different. First, classical bootstrap relies on resampling the full dataset, which is infeasible in the online setting where the algorithm operates with one data point at each iteration. Second, existing results typically assume access to the exact minimizer and do not account for optimization-induced randomness. In contrast, we explicitly quantify the randomness resulting from stochastic algorithm and provide inference for the fairness functional of the learned model itself. While bootstrap methods for constrained estimators exist Li [2021], Fang and Santos [2014], they operate in offline frameworks.

3 NOTATIONS

Let $h(\cdot;\theta):\mathbb{R}^d\to\{-1,1\}$ be a classifier parameterized by $\theta\in\mathbb{R}^d$. In a sample (x,y,z), $x\in\mathbb{R}^d$ is the feature, $z\in\{0,1\}$ is the sensitive attribute, and $y\in\{-1,1\}$ is the label. We use $\rho(w)$ to denote the density of a random variable W. Let $L(\cdot,W):\Theta\to\mathbb{R}^+\cup\{0\}$ be the loss function where $W\coloneqq(x,y)$. Let $\delta_\theta(x)$ denote the decision boundary, i.e., $h(x;\theta)=\mathrm{sign}(\delta_\theta(x))$. For some statements we use $DI[DM]$ as subscript to imply that the statement holds for DI and DM both. $W_2(X,Y)$ denotes the Wasserstein-2 distance between random variables X and Y. We present the formal definitions of DI and DM below.

Definition 1 (Disparate Impact) *A binary classifier is fair with respect to DI if the probability of the classifier predicting positive class does not change conditional on the value of the sensitive feature, and*

we measure DI by, $\phi_{DI}(\theta)=|\Delta_{DI}(\theta)|$, *where* $\Delta_{DI}(\theta)=\mathbb{P}(h(x;\theta)=1|z=0)-\mathbb{P}(h(x;\theta)=1|z=1)$.

Definition 2 (Disparate Mistreatment) *A binary classifier is fair with respect to DM if the False Positive Rate of the classifier does not change conditional on the value of the sensitive feature, and*

we measure DM unfairness by $\phi_{DM}(\theta)=|\Delta_{DM}(\theta)|$, *where* $\Delta_{DM}(\theta)=\mathbb{P}(h(x;\theta)=1|y=-1,z=0)-\mathbb{P}(h(x;\theta)=1|y=-1,z=1)$.

DM can be also be defined in terms of False Negative Rate (FNR), False Omission Rate (FOR), and False Discovery Rate (FDR), and our methods work for all the choices.

4 ASYMPTOTICS OF MODEL PARAMETER

In this section, we characterize the asymptotic distribution of the algorithmic estimate of the parameter of a fairness-aware linear binary classifier trained with SGD-type algorithm. First, we state our assumptions below.

Assumption 4.1 *We use the loss function* $L(\cdot;W):\Theta\to\mathbb{R}^+\cup\{0\}$ *to denote both squared loss and* ℓ_2*-regularized Cross-Entropy (CE) loss for all* $W\in\mathbb{R}^{d+1}$. *Denote* $\mathbb{E}\left[L(\theta,W)|\theta\right]\coloneqq l(\theta)$, *and* $\theta^*_{DI[DM]}$ *denotes the minima of* $l(\theta)$ *subject to* $DI[DM]$ *constraint.*

Assumption 4.2 *The data* $\{(W_k,z_k)\}_k$ *are* $\mathrm{i.i.d}$ *and* $\mathbb{E}[\|\varphi(x)\|_2^2]<\infty$. *Denote* $Cov[\nabla L(\theta^*_{DI[DM]},W_k)]\coloneqq S_{DI[DM]}$.

Assumption 4.3 *For any point* $x\in\mathbb{R}^d$, *the decision boundary* $\delta_\theta(x)$ *is a linear function of* θ, *i.e., for some feature map* $\varphi(x)$ *we can write,* $\delta_\theta(x)=\theta^\top\varphi(x)$.

Assumption 4.4 *For each* $z\in\{0,1\}$, *there exist neighborhoods* $\mathcal{U}$ *of* $\{x:\delta_{\theta^*_{DI}}(x)=0\}$ *and* $\mathcal{U}'$ *of* $\{x:\delta_{\theta^*_{DM}}(x)=$

$0\}$ *such that the conditional distributions* $\varphi(x) \mid z$ *and* $\varphi(x) \mid (z, Y = -1)$ *admit Lebesgue densities that are continuous on* $\mathcal{U}$ *and* $\mathcal{U}'$*, respectively.*

Remark 1 (Comments on assumptions) *Our analysis readily extends to any strongly-convex loss but we use CE loss and squared loss in Assumption 4.1 as these are the standard losses for classification Zhu and Liu [2021], Arafa et al. [2021], Li et al. [2016], Hui and Belkin [2020]. Assumption 4.2 allows heavy-tailed noise as it only requires the existence of the covariance of the feature map and the noisy gradient at the optima. Assumption 4.3 holds true for logistic regression, SVM as well as in the Neural Tangent Kernel regime of deep neural network classifier. Assumption 4.4 is mild. It only requires local absolute continuity of the feature distribution in arbitrarily small neighborhoods of the decision boundaries* $\{\delta_{\theta^*_{DI}}(x) = 0\}$ *and* $\{\delta_{\theta^*_{DM}}(x) = 0\}$*. It is satisfied whenever the conditional feature distributions admit continuous densities, e.g., under additive noise with continuous density such as Gaussian noise. Part (3) allows for heavy-tailed feature distributions that requires the existence of only the second moment.*

4.1 DISPARATE IMPACT

We adopt the DI-aware classification framework of Zafar et al. [2019], which, for a given unfairness tolerance level $\epsilon > 0$, is formulated as follows:

$$\theta^*_{DI} = \operatorname{argmin} \, l(\theta) \;\; \text{s.t.} \;\; \left|\mathbb{E}\left[(z - \mathbb{E}\left[z\right])\delta_\theta(x)\right]\right| \leq \epsilon. \quad (1)$$

Our setting differs from Zafar et al. [2019] in the following aspects: (i) We consider an online setting where the algorithm receives one sample per iteration, which is particularly efficient for large datasets where computing full gradients is costly, e.g., `ACSIncome` Ding et al. [2021]. (ii) We assume access to a fixed held-out dataset $\mathcal{D}'$ with $n_c = |\mathcal{D}'|$ i. i. d samples. It is unavoidable, as we need this dataset to impose the fairness constraints. Then, under Assumption 4.3, (1) now becomes,

$$\theta^*_{DI} \coloneqq \operatorname{argmin} \;\; l(\theta) \;\; \text{s.t.} \;\; \theta \in \Theta \coloneqq \left\{\theta \middle| \left|\tilde{x}^\top \theta\right| \leq \epsilon\right\}, \quad (2)$$

where $\tilde{x} = n_c^{-1} \sum_{i \in \mathcal{D}'} (z_i - \bar{z})\varphi(x_i)$, and $\bar{z} = n_c^{-1} \sum_{i \in \mathcal{D}'} z_i$. We use the Stochastic Dual Averaging (SDA) (Algorithm 1) to solve (2). In this algorithm one maintains a sequence $\{u_k\}_k$ where u_k is the weighted sum of the noisy gradients up to iteration k. θ_k is the projection $\Pi_\Theta(-u_{k-1})$ of $-u_{k-1}$ on to the set Θ.

Proposition 4.1 *Let Assumption 4.1-4.3 be true and* $\eta_k = k^{-a}$ *where* $a \in (1/2, 1)$*. Then, for the updates of Algorithm 1, we have,* $\bar{\theta}_k \overset{a.s.}{\to} \theta^*_{DI}$*, and*

$$\sqrt{k}(\bar{\theta}_k - \theta^*_{DI}) \overset{d}{\sim} N\left(0, \Sigma_{DI}\right),$$

Algorithm 1 Weighted Stochastic Dual Averaging

Input: η_k, $u_0 \in \mathbb{R}^d$

1: **for** $k = 1, \cdots, T$ **do**
2: **Update** $\theta_k = \operatorname{argmin}_{\theta \in \Theta} \left\{\langle u_{k-1}, \theta\rangle + \frac{1}{2}\left\|\theta\right\|_2^2\right\} = \Pi_\Theta(-u_{k-1})$
3: $u_k = u_{k-1} + \eta_k \nabla L(\theta_k, W_k)$
4: **end for**

where Π_Θ is the projection operator on to the set Θ.
Output: $\bar{\theta}_T = \frac{1}{T}\sum_{k=1}^T \theta_k \in \mathbb{R}^d$.

where $P_A = I - \tilde{x}\tilde{x}^\top / \left\|\tilde{x}\right\|_2^2$*, and* $\Sigma_{DI} = P_A \left(\nabla^2 l(\theta^*_{DI})\right)^\dagger P_A S_{DI} P_A \left(\nabla^2 l(\theta^*_{DI})\right)^\dagger P_A$.

The proof of Proposition 4.1 is provided in Appendix C.1.

4.2 DISPARATE MISTREATMENT

Learning a linear classifier to mitigate DM with respect to FPR requires solving the following optimization problem Wan et al. [2021], Celis et al. [2019].

$$\theta^*_{DM} \coloneqq \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \; l(\theta) + R_2 \gamma(\theta; \mathcal{D}'^-), \quad (3)$$

where $\mathcal{D}'^-$ contains the samples from $\mathcal{D}'$ with label -1, $n_c^- = |\mathcal{D}'^-|$, and $\gamma(\theta; \mathcal{D}'^-) \coloneqq \left((n_c^-)^{-1} \sum_{i \in \mathcal{D}'^-} (z_i - \bar{z}) \min(0, y_i \varphi(x_i)^\top \theta)\right)^2$.

$R_2 > 0$ controls the unfairness tolerance similar to ϵ in (2). The points of discontinuity of $\min(0, y_i \varphi(x_i)^\top \theta)$ are given by $y_i \varphi(x_i)^\top \theta = 0$. Under Assumption 4.4, the Hessian of $\min(0, y_i \varphi(x_i)^\top \theta)$ is 0 almost everywhere. We can bypass this issue by defining the Hessian and the gradient at $y_i \varphi(x_i)^\top \theta = 0$ to be 0 Blanc et al. [2020]. One could also replace the function $\min(0, y_i \varphi(x_i)^\top \theta)$ by a differentiable function $-\log(1 + \exp(-\tau y_i \varphi(x_i)^\top \theta))/\tau$, $\tau > 0$ which can approximate $\min(0, y_i \varphi(x_i)^\top \theta)$ arbitrarily well as $\tau \to \infty$.

As (3) is unconstrained, we use vanilla SGD with Polyak-Ruppert averaging Polyak and Juditsky [1992] to solve (3).

$$\theta_k = \theta_{k-1} - \eta \left(\nabla L(\theta_{k-1}, w_k) + R_2 \nabla \gamma(\theta_{k-1}; \mathcal{D}'^-)\right)$$
$$\bar{\theta}_T = \frac{1}{T}\sum_{k=1}^T \theta_k \quad (4)$$

In this setting, the asymptotic normality of $\bar{\theta}_T$ follows from Polyak and Juditsky [1992]. We state the result here for completion but omit the proof.

Proposition 4.2 (Theorem 3, Polyak and Juditsky [1992]) *Let Assumptions 4.1-4.3 be true. Then, choosing* $\eta_k = k^{-a}$

with $1/2 < a < 1$*, for the updates of SGD* (4) *we have,* $\bar{\theta}_k \overset{a.s.}{\to} \theta^*_{DM}$*, and*

$$\sqrt{k}\left(\bar{\theta}_k - \theta^*_{DM}\right) \sim N(0, \Sigma_{DM}),$$

where,

$$\Sigma_{DM} = \nabla^2 l(\theta^*_{DM})^{-1}\left(S_{DM} + R_2^2 \nabla\gamma(\theta^*_{DM}; \mathcal{D}'^{-})\nabla\gamma(\theta^*_{DM}; \mathcal{D}'^{-})\right) \nabla^2 l(\theta^*_{DM})^{-1}.$$

Constructing CI for $\bar{\theta}_k$ requires estimating $\Sigma_{DI[DM]}$ as they depend on unknown $\theta^*_{DI[DM]}$. We estimate $\Sigma_{DI[DM]}$ in Section 5, and quantify the uncertainty in test fairness functional $\phi_{DI[DM]}(\bar{\theta}_k)$ which is our main theoretical contribution.

5 ONLINE CI ESTIMATION

In this section we present the bootstrap-based online CI estimation algorithm. We split the estimation into two steps:

Step 1. We generate B perturbed samples $\{\bar{\theta}^b_k\}_{b=1}^B$ (line 4-6 for DI, and line 11 for DM in Algorithm 2) at each time k to estimate the distribution of $\bar{\theta}_k$, where $\{V^b_k\}_{k,b}$ is a sequence of i. i. d random variables such that $V^b_k \sim \mathcal{V}$ with $\mathbb{E}\left[V^b_k\right] = 1$, var $\left[V^b_k\right] = 1$, and $\theta^b_0 = \theta_0$ for all $k \geq 1, b = 1, 2, \cdots, B$. In principle, one could obtain B independent copies of $\bar{\theta}_k$ by running the algorithm on B independent data streams. Such an approach, besides being computationally expensive and data-intensive, is infeasible in online fairness-aware classification Sun et al. [2019], Iosifidis et al. [2021], Huang et al. [2022], Wang et al. [2023], where only a single data stream is available. Consistent with the online setting, in Algorithm 2, each perturbed trajectory $\{\bar{\theta}^b_k\}_{b=1}^B$ is updated using only the current data point W_k at iteration k. The empirical distribution of $\bar{\theta}^b_n - \bar{\theta}_n$ is the approximation of the distribution of $\bar{\theta}_n - \theta^*$ where $\bar{\theta}^b_n = n^{-1}\sum_{i=1}^n \theta^b_i$ as we show in Theorem 5.1.

Step 2. To estimate the CI, we use a small held-out dataset $\tilde{\mathcal{D}}$ independent of $\mathcal{D}'$. Use of a held-out dataset for inference is common in statistical literature Barber et al. [2022], Sesia and Romano [2021], Cherian and Candès [2023], Barber et al. [2021]. In our setting, this is unavoidable as group fairness measures cannot be meaningfully computed from a single data point. Let n_0 and n_1 denote the number of samples in $\tilde{\mathcal{D}}$ with $z_i = 0$ and $z_i = 1$ respectively. Let $n_{0,-}$ and $n_{1,-}$ denote the number of samples in $\tilde{\mathcal{D}}$ with $z_i = 0, y_i = -1$ and $z_i = 1, y_i = -1$ respectively. Using Definition 1 and 2, we generate B samples of fairness estimates $\{\hat{\phi}_{DI}(\bar{\theta}^b_k)\}_{b=1}^B$ and $\{\hat{\phi}_{DM}(\bar{\theta}^b_k)\}_{b=1}^B$. For any given θ, $\hat{\phi}_{DI}(\theta)$, and $\hat{\phi}_{DM}(\theta)$ are evaluated on $\tilde{\mathcal{D}}$ as follows.

$$\hat{\phi}_{DI}(\theta) = \left|\frac{\sum_{i\in\tilde{\mathcal{D}}, z_i=0} q_i}{n_0} - \frac{\sum_{i\in\tilde{\mathcal{D}}, z_i=1} q_i}{n_1}\right|,$$

$$\hat{\phi}_{DM}(\theta) = \left|\frac{\sum_{i\in\tilde{\mathcal{D}}, z_i=0, y_i=-1} q_i}{n_{0,-}} - \frac{\sum_{i\in\tilde{\mathcal{D}}, z_i=1, y_i=-1} q_i}{n_{1,-}}\right|, \quad (5)$$

Algorithm 2 Online Bootstrap for $1-\alpha$ CI Estimation

Input: $\{\eta_k\}_k$, $\theta_0, u_0 \in \mathbb{R}^d$, $\mathcal{V}$, fairness criterion, α

1: **for** $k = 1, \cdots, T$ **do**
2: **for** $b = 1, 2, \cdots, B$
3: **if** fairness criterion = DI **then**
4: **Update** $\theta^b_k = \Pi_\Theta(-u^b_{k-1})$
5: **Sample** $V^b_k \sim \mathcal{V}$
6: $u^b_k = u^b_{k-1} + \eta_k V^b_k \nabla L(\theta^b_k, W_k)$
7: **end if**
8: **if** fairness criterion = DM **then**
9: $\theta^b_k = \theta^b_{k-1} - \eta_k V^b_k \nabla L(\theta^b_k, W_k)$
10: **end if**
11: $\bar{\theta}^b_k = k^{-1}(\bar{\theta}^b_{k-1} + \theta^b_k)$
12: **end for**
13: **end for**
14: Compute $\{\hat{\phi}_{DI[DM]}(\bar{\theta}^b_T)\}_{b=1}^B$ using (5)
15: **Output:**
16: $[\hat{\phi}_{DI[DM],\alpha/2}, \hat{\phi}_{DI[DM],1-\alpha/2}]$

where $q_i = \mathbb{1}(\theta^\top\varphi(x_i) > 0)$. For $0 < \alpha < 1$, to estimate CI of significance level α of $\phi_{DI}(\bar{\theta}_k)$, and $\phi_{DM}(\bar{\theta}_k)$ we use the sample CIs $[\hat{\phi}_{DI,\alpha/2}, \hat{\phi}_{DI,1-\alpha/2}]$, and $[\hat{\phi}_{DM,\alpha/2}, \hat{\phi}_{DM,1-\alpha/2}]$ of $\{\hat{\phi}_{DI}(\bar{\theta}^b_k)\}_{b=1}^B$ and $\{\hat{\phi}_{DM}(\bar{\theta}^b_k)\}_{b=1}^B$ respectively.

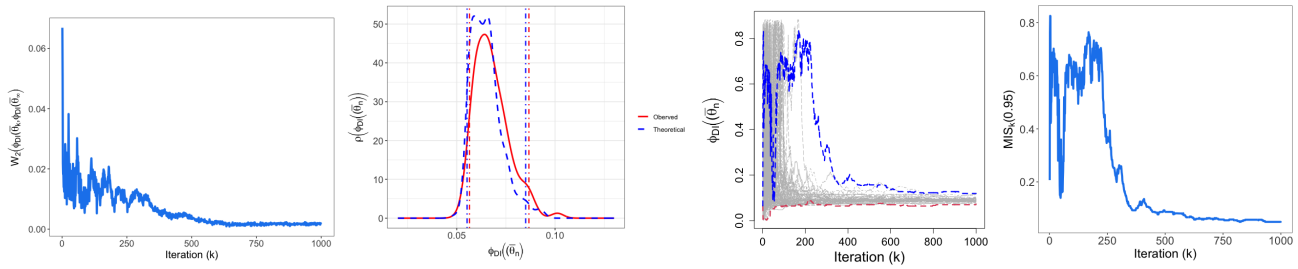
Fang et al. [2018] presents an online bootstrap-based covariance estimation of averaged SGD iterates for strongly convex objectives in the unconstrained setting. While their results apply directly to SGD (4), they do not extend to Algorithm 1, which solves a constrained optimization problem. Our first theoretical contribution is to prove analogous convergence guarantees for Algorithm 2 when applied to Algorithm 1. Beyond fairness-aware classification, our results enable online inference for the constrained stochastic optimization framework of Duchi and Ruan [2016]. To the best of our knowledge, this is the first theoretical result on online bootstrap-based inference for constrained stochastic optimization. We now state our first main theoretical result.

Theorem 5.1 *Let Assumptions 4.1-4.3 be true. Then, choosing* $\eta_k = k^{-a}$*,* $1/2 < a < 1$ *in Algorithm 2, for any* $b \in 1, 2, \cdots, B$*, we have,* $\bar{\theta}^b_k \overset{a.s.}{\to} \theta^*_{DI[DM]}$*, and*

1. $\sqrt{k}\left(\bar{\theta}^b_k - \bar{\theta}_k\right) \overset{d}{\sim} N\left(0, \Sigma_{DI}\right)$,
2. $\sup_{q\in\mathbb{R}^d}\left|\mathbb{P}(\sqrt{k}(\bar{\theta}^b_k - \bar{\theta}_k) \leq q|\mathcal{F}_k) - \mathbb{P}(\sqrt{k}(\bar{\theta}_k - \theta^*_{DI}) \leq q)\right| \overset{P}{\to} 0$,

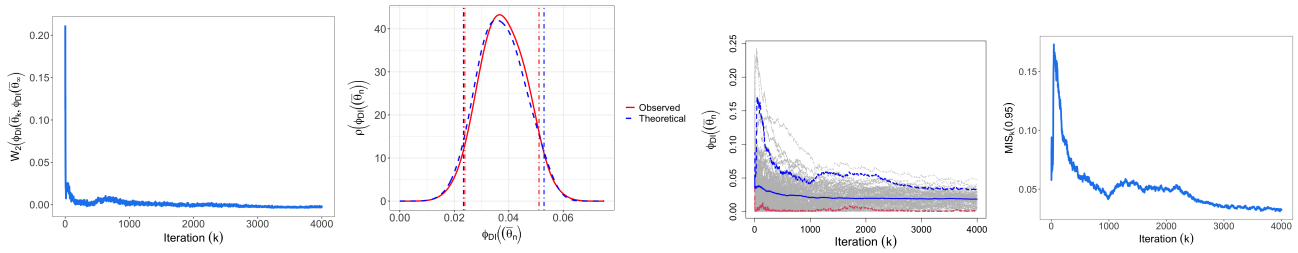
Statements 1, and 2 hold when DI *is replaced with* DM*.*

Remark 2 *Statement 1 in Theorem 5.1 says that* $\bar{\theta}^b_k$ *is asymptotically normal with* $O(k^{-1/2})$ *rate with mean* $\bar{\theta}_k$



(a) (b) (c) (d)

Figure 2: (a) Convergence of $W_2\left(\phi_{DI}(\bar{\theta}_k), \phi_{DI}(\bar{\theta}_\infty)\right)$ (b) comparison of theoretical asymptotic density and observed density of $\phi_{DI}(\bar{\theta}_k)$ (c) Online Bootstrap 95% CI (d) $MIS(\hat{\phi}_{DI,0.025}, \hat{\phi}_{DI,0.975}; 0.95)$ under CE loss for synthetic data.



(a) (b) (c) (d)

Figure 3: (a) Convergence of $W_2\left(\phi_{DI}(\bar{\theta}_k), \phi_{DI}(\bar{\theta}_\infty)\right)$ (b) comparison of theoretical asymptotic density and observed density of $\phi_{DI}(\bar{\theta}_k)$ (c) Online Bootstrap 95% CI (d) $MIS(\hat{\phi}_{DI,0.025}, \hat{\phi}_{DI,0.975}; 0.95)$ under CE loss for ACSIncome data.

and covariance $\Sigma_{DI[DM]}$, the asymptotic covariance of $\bar{\theta}_k$. This implies that the fluctuations of $\bar{\theta}_k^b$ around $\bar{\theta}_k$ can be used to quantify the unobservable fluctuations of $\bar{\theta}_k$ around $\theta^*_{DI[DM]}$. Statement 2 strengthens Statement 1 saying that the probability of $\bar{\theta}_k^b$ lying in a region around $\bar{\theta}_k$ converges to the probability of $\bar{\theta}_k$ lying in the same region around $\theta^*_{DI[DM]}$ with $O(k^{-1/2})$ rate.

Remark 3 (On computation and interpretability) *The bootstrap procedure at most changes the gradient computation by $O(1)$ since B does not grow with iteration k. Moreover, these computations can be parallelized easily. Theorem 5.1 shows that one can construct CI for $\bar{\theta}_k$ based on $\{\bar{\theta}_k^b\}_{b=1}^B$. This could shed light on the features which are significant for fairness-aware classification leading to better model interpretability.*

As discussed above, the result for $\phi_{DM}(\bar{\theta}k^b)$ follows directly from Theorem 2 of Fang et al. [2018]. We therefore focus on proving Proposition 5.1 for $\phi_{DI}(\bar{\theta}_k^b)$, which constitutes our first main theoretical contribution. A proof outline is given below, with full details deferred to Appendix D.1.

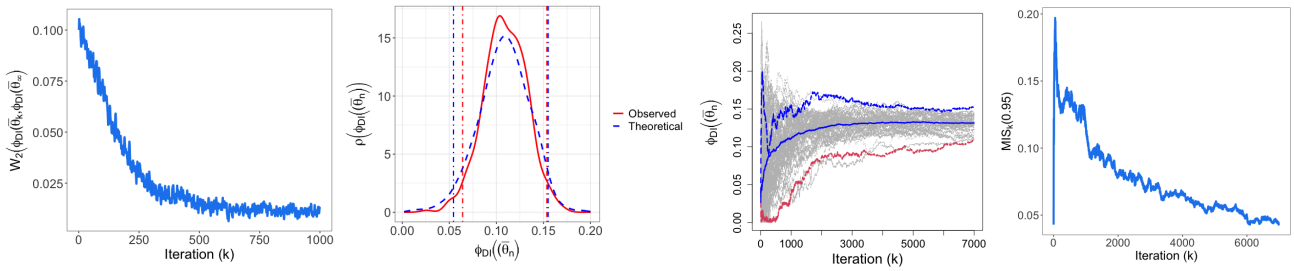
Proof outline. First we show since we are multiplying by i. i. d perturbation variable V_k^b, where $\mathbb{E}[V_k^b] = 1$, the unbiasedness of the noisy gradient is preserved. Then we show the following regularity condition on the perturbed noisy gradient, $\mathbb{E}\left[{V_k^b}^2 \left\|\nabla L(\theta, W) - \nabla L(\theta^*, W)\right\|_2^2\right] \leq 2L_E \left\|\theta - \theta^*\right\|_2^2$. Then using Theorem 3 of Duchi and Ruan [2016] we show that the bootstrapped iterates identify the active constraints of the form $A^b\theta_k^b = \epsilon$ correctly after finite number of steps K. This implies that for $k \geq K$, $P_A\theta_k = \theta_k$ where, $P_A = I - A^\top(AA^\top)^\dagger A = I - \tilde{x}\tilde{x}^\top / \left\|\tilde{x}\right\|_2^2$. Then we show that the projected bootstrapped iterates $P_A\theta_k^b$ satisfies $\sqrt{k}(\bar{\theta}_k^b - \theta^*_{DI}) \xrightarrow{d} N(0, \Sigma_{DI})$, and

$$\sqrt{k}\left(\bar{\theta}_k^b - \bar{\theta}_k\right) = P_H^\dagger \frac{1}{\sqrt{k}} \sum_{i=1}^{k} (V_i^b - 1) P_A \nabla L(\theta^*_{DI}, W_i) + o_{\mathbb{P}}(1),$$

where $P_H = P_A \nabla^2 l(\theta^*_{DI}) P_A$. Then the rest of the proof mainly follows by Martingale CLT Hall and Heyde [2014].

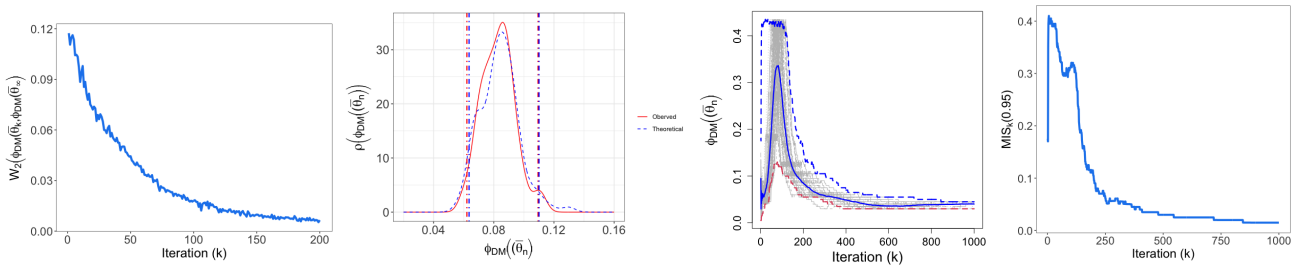
5.1 ASYMPTOTIC NORMALITY OF FAIRNESS

In the previous section, we have established the asymptotic normality of $\bar{\theta}k$ and the validity of bootstrap-based CIs for the model parameter. We now characterize the asymptotic distribution and bootstrap guarantees for the test fairness functional $\phi_{DI[DM]}$ in Theorem 5.2, which is our second main theoretical contribution.



(a) (b) (c) (d)

Figure 4: (a) Convergence of $W_2\left(\phi_{DI}(\bar{\theta}_k), \phi_{DI}(\bar{\theta}_\infty)\right)$ (b) comparison of theoretical asymptotic density and observed density of $\phi_{DI}(\bar{\theta}_k)$ (c) Online Bootstrap 95% CI (d) $MIS(\hat{\phi}_{DI,0.025}, \hat{\phi}_{DI,0.975}; 0.95)$ under CE loss for Adult data.



(a) (b) (c) (d)

Figure 5: (a) Convergence of $W_2\left(\phi_{DM}(\bar{\theta}_k), \phi_{DM}(\bar{\theta}_\infty)\right)$ (b) comparison of theoretical asymptotic density and observed density of $\phi_{DM}(\bar{\theta}_k)$ (c) Online Bootstrap 95% CI (d) $MIS(\hat{\phi}_{DM,0.025}, \hat{\phi}_{DM,0.975}; 0.95)$ under CE loss for synthetic data.

Theorem 5.2 (Asymptotic distribution of $\phi_{DI[DM]}$) *Let Assumptions 4.1 - 4.4 hold. Let $g := \nabla\Delta(\theta^*)$ and set $\sigma_\Delta^2 := g^\top \Sigma g$. Then, we have,*

1. If $\phi(\theta^*) \neq 0$, then

$$\sqrt{k}\big(\phi(\bar{\theta}_k) - \phi(\theta^*)\big) \Rightarrow N(0, \sigma_\Delta^2).$$

If $\phi(\theta^*) = 0$, then

$$\sqrt{k} * \phi(\bar{\theta}_k) \Rightarrow |N(0, \sigma_\Delta^2)|.$$

2. If $\Delta(\theta^*) \neq 0$, then

$$\sup_{c\in\mathbb{R}} \Big| \mathbb{P}\left(\sqrt{k}\big(\phi(\bar{\theta}_k^b) - \phi(\bar{\theta}_k)\big) \leq c | \mathcal{F}_k\right) - \mathbb{P}\left(\sqrt{k}\big(\phi(\bar{\theta}_k) - \phi(\theta^*)\big) \leq c\right) \Big| \xrightarrow{\mathbb{P}} 0.$$

If $\Delta(\theta^*) = 0$, then

$$\sup_{c\geq 0} \Big| \mathbb{P}\left(\sqrt{k}\phi(\bar{\theta}_k^b) \leq c \Big| \mathcal{F}_k\right) - \mathbb{P}\left(\sqrt{k}\phi(\bar{\theta}_k) \leq c\right) \Big| \xrightarrow{\mathbb{P}} 0,$$

and both limits coincide with the CDF of $|N(0, \sigma_\Delta^2)|$.

Both the statements hold for DI and DM. To avoid messy notations, we omit the $DI[DM]$ subscripts here.

The main theoretical challenge towards proving Theorem 5.2 is the nonsmoothness of $\phi_{DI[DM]}$. We defer the proof to Appendix D.1 due to lack of space.

Remark 4 (Operational use with $\hat{\phi}$) *Theorem 5.2 characterizes the sampling fluctuations of the population fairness functional $\phi(\bar{\theta}_k)$. In practice, we report $\hat{\phi}(\bar{\theta}_k)$ computed on an independent held-out dataset $\tilde{\mathcal{D}}$. Since the indicator class $\{x \mapsto \mathbb{1}(\theta^\top \varphi(x) > 0) : \theta \in \Theta\}$ has finite VC dimension, standard uniform laws of large numbers imply $\sup_{\theta\in\Theta} |\hat{\phi}(\theta) - \phi(\theta)| \xrightarrow{\mathbb{P}} 0$ as $|\tilde{\mathcal{D}}| \to \infty$. Consequently, $|\hat{\phi}(\bar{\theta}_k) - \phi(\bar{\theta}_k)| \xrightarrow{\mathbb{P}} 0$, and thus $\hat{\phi}(\bar{\theta}_k)$ consistently estimates $\phi(\bar{\theta}_k)$. The result is formally stated in Proposition D.2 along with proof in Appendix D.1.*

This implies that the uncertainty in the test fairness of a classifier trained via a stochastic algorithm can be consistently quantified using the bootstrap confidence interval constructed from $\{\hat{\phi}(\bar{\theta}_k^b)\}_{b=1}^B$.

We use Mean Interval Score (MIS), a standard scoring function which favors narrower well-calibrated interval, to evaluate the quality of the estimated CI Wu et al. [2021].

Given a sample of size N of a random variable s, and an

estimate $[l, u]$ of CI of level α, $MIS_N(l, u; \alpha)$ is defined as

$$MIS_N(l,u;\alpha) = u - l + \frac{2}{N\alpha}\sum_{i=1}^{N}(s_i - u)\mathbb{1}(s_i > u) + \frac{2}{N\alpha}\sum_{i=1}^{N}(l - s_i)\mathbb{1}(s_i < l). \quad (6)$$

6 EXPERIMENTS

In this section, we present our results on synthetic and real datasets. For each dataset, we show that the observed density converges to the theoretical asymptotic density in Wasserstein-2 (W_2) distance, and MIS for the CI generated by Algorithm 2 decreases over time. We choose ϵ for which the mean accuracy under CE loss matches with the benchmarks. We provide the accuracy plots in Appendix for completion. We choose $\mathcal{V} = \text{Unif}[1-\sqrt{3}, 1+\sqrt{3}]$. For data application, we choose the benchmark datasets `Adult` Dua and Graff [2017], and `ACSIncome` Ding et al. [2021] for DI and `COMPAS` dataset Larson [2016] for DM.

6.1 DISPARATE IMPACT

6.1.1 Synthetic Dataset

Similar to Zafar et al. [2019], we choose $\rho(x|y = 1) = N\left(\left[1.5; 1.5\right], \left[0.4, 0.2; 0.2, 0.3\right]\right)$, $\rho(x|y = 1) = N\left(\left[-1.5; -1.5\right], \left[0.6, 0.1; 0.1, 0.4\right]\right)$, and the sensitive attribute z as a Bernoulli random variable with $\mathbb{P}(z = 1) = \rho(x'|y = 1)/(\rho(x'|y = 1) + \rho(x'|y = -1))$, where $x' = \left[\cos(\pi/3), -\sin(\pi/3); \sin(\pi/3), \cos(\pi/3)\right] x$. Figure 2(a) shows that $W_2(\phi_{DI}(\bar{\theta}_k), \phi_{DI}(\bar{\theta}_\infty))$, where $\bar{\theta}_\infty \sim N(\theta^*_{DI}, \Sigma_{DI}/k)$, decreases over iterations. Figure 2(b) shows that after 1000 (4000) iterations, the observed density of $\phi_{DI}(\bar{\theta}_{1000})$ ($\phi_{DI}(\bar{\theta}_{4000})$), and the theoretical density given by $\phi_{DI}(\bar{\theta}_\infty)$ (dashed line) almost overlap. The vertical dashed lines mark the observed (red) and theoretical (blue) 95% CI. Figure 2(c)-(d) show that DI values across repetitions are well contained in the computed CI and $MIS(\hat{\phi}_{DI,0.025}, \hat{\phi}_{DI,0.975}; 0.95)$ becomes small respectively. In (c), the grey lines show multiple repetitions.

6.1.2 Adult Dataset

Here we use the `Adult` dataset Dua and Graff [2017] which contains income data of adults. Two classes indicate whether income is $\geq 50K$ or $< 50K$. We use 13 features for classification. `Sex` is the sensitive feature. More details on the dataset is provided in the Appendix. Similar to Figure 2, in Figure 4: (a)-(b) shows the convergence to the asymptotic distribution in terms of W_2 distance, and CI. Online CI estimation results are in (c)-(d).

6.1.3 ACSIncome Dataset

The `ACSIncome` dataset Ding et al. [2021] is derived from the U.S. Census Bureau's American Community Survey (ACS) and is designed as a modern replacement for the `Adult` dataset. The task is binary classification, where the response indicates whether an individual's annual income exceeds $50,000. We use the standard 10 demographic and employment-related features provided Ding et al. [2021]. `Sex` is treated as the sensitive feature. Compared to `Adult`, `ACSIncome` is much larger with 1,664,500 samples making offline setting computationally expensive and the online setting the more natural choice. Similar to Figure 2, in Figure 3: (a)-(b) show convergence to the asymptotic distribution in terms of W_2 distance and confidence intervals, and online CI estimation results are presented in (c)-(d).

6.2 DISPARATE MISTREATMENT

6.2.1 Synthetic Dataset

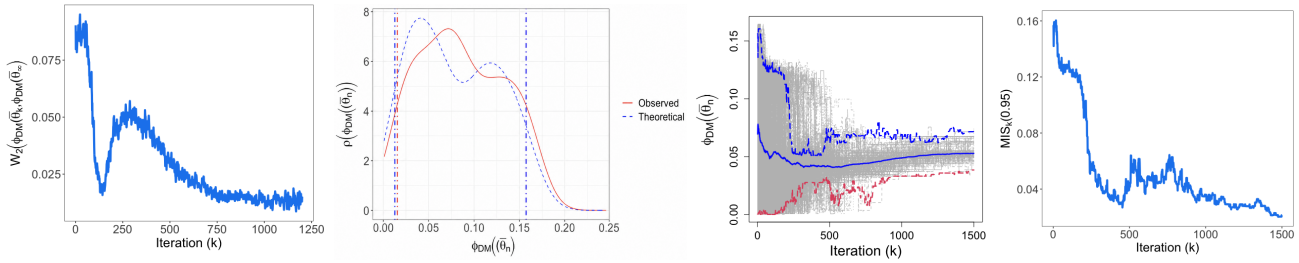
We use the synthetic dataset introduced in Zafar et al. [2019]. Let $\rho(x|z = 0, y = 1) = \rho(x|z = 1, y = 1) = N\left(\left[2; 2\right], \Sigma_1\right)$, $\rho(x|z = 0, y = -1) = N\left(\left[1; 1\right], \Sigma_1\right)$, and $\rho(x|z = 1, y = -1) = N\left(\left[-2; -2\right], \Sigma_1\right)$ where $\Sigma_1 = \left[3, 1; 1, 3\right]$. In Figure 5: (a) shows that $W_2(\phi_{DM}(\bar{\theta}_k), \phi_{DM}(\bar{\theta}_\infty))$, where $\bar{\theta}_\infty \sim N(\theta^*_{DM}, \Sigma_{DM}/k)$, becomes small over iterations; (b) shows that the observed density of $\phi_{DM}(\bar{\theta}_{1000})$, and the theoretical density of $\phi_{DI}(\bar{\theta}_\infty)$ almost overlap. Similar to Section 6.1.1, results on CI estimation are in (c)-(d).

6.2.2 COMPAS Dataset

We use the ProPublica COMPAS dataset Larson [2016], which contains data on criminal defendants, to classify whether a subject will recidivate within two years (positive class) or not (negative class). After preprocessing as in Zafar et al. (2019), the dataset contains 5,278 subjects with 5 features. We use `Race` as the sensitive feature. Figure 6(a)-(b) show convergence to the asymptotic distribution in terms of W_2 distance and CI. Figure 6(c)-(d) show that the estimated CI tightly contains the trajectories over multiple repetitions, and $MIS(\hat{\phi}_{DM,0.025}, \hat{\phi}_{DM,0.975}; 0.95)$ captures the CI width correctly.

7 DISCUSSION AND FUTURE WORK

In this work, we establish asymptotic normality of the model parameter for a linear binary classifier trained with SDA and SGD under DI and DM constraints, respectively. To the best of our knowledge, this is the first study of real-time uncertainty quantification for DI and DM in fairness-aware



(a) (b) (c) (d)

Figure 6: (a) Convergence of $W_2\left(\phi_{DM}(\bar{\theta}_k), \phi_{DM}(\bar{\theta}_\infty)\right)$ (b) comparison of theoretical asymptotic density and observed density of $\phi_{DM}(\bar{\theta}_k)$ (c) Online Bootstrap 95% CI (d) $MIS(\hat{\phi}_{DM,0.025}, \hat{\phi}_{DM,0.975}; 0.95)$ under CE loss for COMPAS data.

online classification. To estimate the asymptotic covariance $\Sigma_{DI[DM]}$, which depends on the unknown minimizer $\theta^*_{DI[DM]}$, we propose an online multiplicative bootstrap procedure (Algorithm 2). Our theoretical contributions are twofold: (i) we prove distributional convergence of the bootstrap estimator and validity of the resulting CIs for the model parameter; (ii) we derive the asymptotic distribution of the test fairness functional $\phi_{DI[DM]}$, addressing its nonsmoothness, and show that the proposed CI consistently estimates the asymptotically valid CI. We illustrate our results on synthetic and real benchmark datasets.

Extending these results to neural networks is an important direction. Another promising avenue is fairness inference with multiple sensitive attributes and classes.

References

Gabriel Aguiar, Bartosz Krawczyk, and Alberto Cano. A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework. *arXiv preprint arXiv:2204.03719*, 2022.

Ahmed Arafa, Marwa Radad, Mohammed Badawy, and Nawal El-Fishawy. Regularized logistic regression model for cancer classification. In *2021 38th National Radio Science Conference (NRSC)*, volume 1, pages 251–261. IEEE, 2021.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*, 2022.

Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR, 2020.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019.

Richard J Chen, Tiffany Y Chen, Jana Lipkova, Judy J Wang, Drew FK Williamson, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. Algorithm fairness in ai for medicine and healthcare. *arXiv preprint arXiv:2110.00603*, 2021.

Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *arXiv preprint arXiv:1610.08637*, 2016.

John J Cherian and Emmanuel J Candès. Statistical inference for fairness auditing. *arXiv preprint arXiv:2305.03712*, 2023.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

Eustasio del Barrio, Paula Gordaliza, and Jean-Michel Loubes. Review of mathematical frameworks for fairness in machine learning. *arXiv preprint arXiv:2005.13755*, 2020.

Ahmet Demirkaya, Jiasi Chen, and Samet Oymak. Exploring the role of loss functions in multiclass classification. In *2020 54th annual conference on information sciences and systems (ciss)*, pages 1–5. IEEE, 2020.

Cyrus DiCiccio, Sriram Vasudevan, Kinjal Basu, Krishnaram Kenthapadi, and Deepak Agarwal. Evaluating fairness using permutation tests. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1467–1477, 2020.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

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

John Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *arXiv preprint arXiv:1612.05612*, 2016.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979. doi: 10.1214/aos/1176344552. URL `https://doi.org/10.1214/aos/1176344552`.

Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *The Journal of Machine Learning Research*, 19(1):3053–3073, 2018.

Zheng Fang and Andres Santos. Inference on directionally differentiable functions. *arXiv preprint arXiv:1404.3763*, 2014.

Jean-Jacques Forneron and Serena Ng. Inference by stochastic optimization: A free-lunch bootstrap. *arXiv preprint arXiv:2004.09627*, 2020.

Peter Hall. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.

Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

Lingxiao Huang and Nisheeth Vishnoi. Stable and fair classification. In *International Conference on Machine Learning*, pages 2879–2890. PMLR, 2019.

Wen Huang, Kevin Labille, Xintao Wu, Dongwon Lee, and Neil Heffernan. Achieving user-side fairness in contextual bandits. *Human-Centric Intelligent Systems*, pages 1–14, 2022.

Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*, 2020.

Vasileios Iosifidis, Thi Ngoc Han Tran, and Eirini Ntoutsi. Fairness-enhancing interventions in stream classification. In *Database and Expert Systems Applications: 30th International Conference, DEXA 2019, Linz, Austria, August 26–29, 2019, Proceedings, Part I 30*, pages 261–276. Springer, 2019.

Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. Online fairness-aware learning with imbalanced data streams. *arXiv preprint arXiv:2108.06231*, 2021.

Liwei Jiang, Abhishek Roy, Krishnakumar Balasubramanian, Damek Davis, Dmitriy Drusvyatskiy, and Sen Na. Online covariance estimation in nonsmooth stochastic approximation. In *The Thirty Eighth Annual Conference on Learning Theory*, pages 3079–3123. PMLR, 2025.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *Uncertainty in artificial intelligence*, pages 862–872. PMLR, 2020.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Je Larson. Surya ma u, lauren kirchner, and julia angwin. 2016. *How We Analyzed the COMPAS Recidivism Algorithm. ProPublica (5 2016)*, 2016.

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452, 2022.

Jessie Li. The proximal bootstrap for finite-dimensional regularized estimators. In *AEA Papers and Proceedings*, volume 111, pages 616–620. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2021.

Wenfa Li, Hongzhe Liu, Peng Yang, and Wei Xie. Supporting regularized logistic regression privately and efficiently. *PloS one*, 11(6):e0156479, 2016.

Subha Maity, Songkai Xue, Mikhail Yurochkin, and Yuekai Sun. Statistical inference for individual fairness. *arXiv preprint arXiv:2103.16714*, 2021.

Barbara Martinez Neda, Yue Zeng, and Sergio Gago-Masague. Using machine learning in admissions: Reducing human and algorithmic bias in the selection process. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, pages 1323–1323, 2021.

Clara Moreau and Julia Weber. Quantifying algorithmic fairness: A novel perspective through uncertainty estimation. *European Journal of Emerging Artificial Intelligence*, 1 (01):96–112, 2024.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Apostolos F Psaros, Xuhui Meng, Zongren Zou, Ling Guo, and George Em Karniadakis. Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *Journal of Computational Physics*, page 111902, 2023.

Sandro Radovanović, Andrija Petrović, Boris Delibašić, and Milija Suknović. Enforcing fairness in logistic regression algorithm. In *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–7. IEEE, 2020.

Pratik Ramprasad, Yuantong Li, Zhuoran Yang, Zhaoran Wang, Will Wei Sun, and Guang Cheng. Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, 118(544): 2901–2914, 2023.

Abhishek Roy and Krishnakumar Balasubramanian. Online covariance estimation for stochastic gradient descent under markovian sampling. *arXiv preprint arXiv:2308.01481*, 2023.

Abhishek Roy and Yi-An Ma. A central limit theorem for algorithmic estimator of saddle point. *arXiv preprint arXiv:2306.06305*, 2023.

Sergey Samsonov, Marina Sheshukova, Eric Moulines, and Alexey Naumov. Statistical inference for linear stochastic approximation with markovian noise. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Elizabeth D Schifano, Jing Wu, Chun Wang, Jun Yan, and Ming-Hui Chen. Online updating of statistical inference in the big data setting. *Technometrics*, 58(3):393–403, 2016.

Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34:6304–6315, 2021.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Nian Si, Karthyek Murthy, Jose Blanchet, and Viet Anh Nguyen. Testing group fairness via optimal transport projections. In *International Conference on Machine Learning*, pages 9649–9659. PMLR, 2021.

Yi Sun, Ivan Ramirez, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Learning fair classifiers in online stochastic settings. *CoRR, abs/1908.07009*, 2019.

Jie Tang, Chuanlong Xie, Xianli Zeng, and Lixing Zhu. Empirical likelihood-based fairness auditing: Distribution-free certification and flagging. *arXiv preprint arXiv:2601.20269*, 2026.

Bahar Taskesen, Jose Blanchet, Daniel Kuhn, and Viet Anh Nguyen. A statistical test for probabilistic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 648–665, 2021.

Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. *Mobile Health: Sensors, Analytic Methods, and Applications*, pages 495–517, 2017.

Aad Van Der Vaart and Jon A Wellner. A note on bounds for vc dimensions. *Institute of Mathematical Statistics collections*, 5:103, 2009.

Aad W van der Vaart and Jon A Wellner. The bootstrap. In *Weak Convergence and Empirical Processes: With Applications to Statistics*, pages 345–359. Springer, 1996.

Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. Modeling techniques for machine learning fairness: A survey. *arXiv preprint arXiv:2111.03015*, 2021.

Zichong Wang, Nripsuta Saxena, Tongjia Yu, Sneha Karki, Tyler Zetty, Israat Haque, Shan Zhou, Dukka Kc, Ian Stockwell, Albert Bifet, et al. Preventing discriminatory decision-making in evolving data streams. *arXiv preprint arXiv:2302.08017*, 2023.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, 2017.

Dongxia Wu, Liyao Gao, Xinyue Xiong, Matteo Chinazzi, Alessandro Vespignani, Yi-An Ma, and Rose Yu. Quantifying uncertainty in deep spatiotemporal forecasting. *arXiv preprint arXiv:2105.11982*, 2021.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1):2737–2778, 2019.

Wenbin Zhang and Liang Zhao. Online decision trees with fairness. *arXiv preprint arXiv:2010.08146*, 2020.

Chen Zhao, Feng Chen, and Bhavani Thuraisingham. Fairness-aware online meta-learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2294–2304, 2021.

Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, pages 1–12, 2021.

Yunzhang Zhu and Renxiong Liu. An algorithmic view of l2 regularization and some path-following algorithms. *Journal of Machine Learning Research*, 22(138):1–62, 2021.

A APPENDIX

B DETAILS ON THE LOSS FUNCTIONS

B.1 CROSS-ENTROPY LOSS

In case of CE loss $l(\theta)$ is of the form

$$l(\theta) = \mathbb{E}\left[\log\left(1 + \exp\left(-y\varphi(x)^\top\theta\right)\right)\right] + R(\theta), \tag{7}$$

where $R(\theta)$ is a strongly convex regularizer, e.g., $R(\theta) = \varkappa\|\theta\|_2^2$. It is easy to see that canonical CE loss is strictly convex and hence we need $\varkappa > 0$ but $\varkappa$ is allowed to be arbitrarily small.

B.2 SQUARED LOSS

We use the one-hot encoding of the class for the squared loss. One-hot encoding of a class c is given by the vector y_v where $y_v[i] = \mathbb{1}(i = c),\ i = 1, 2$. In this case, the decision boundary is given by,

$$\delta_\theta(x) = \left[1, -1\right]\Theta\varphi(x),$$

where $\varphi(x) \in \mathbb{R}^d$, and $\Theta \in \mathbb{R}^{2\times d}$. Then the optimization problem we solve here is given by

$$\begin{aligned} \min \quad & \mathbb{E}\left[\|y_v - \Theta\varphi(x)\|_2^2\right] + R(\theta) \\ \text{subject to} \quad & \frac{1}{n_c}\left|\sum_{i=1}^{n_c}(z_i - \bar{z})\left[1, -1\right]\Theta\varphi(x_i)\right| \le \epsilon. \end{aligned} \tag{8}$$

Let $\Theta_i, i = 1, 2$ denote the i-th row of Θ. Let $\Theta_v = \left[\Theta_1^\top; \Theta_2^\top\right]$ be the vectorized version of Θ, and let $x_{v,i} = (\varphi(x_i)^\top, -\varphi(x_i)^\top)^\top$. Then the optimization problem (8) can also written as the following.

$$\begin{aligned} \min \quad & \mathbb{E}\left[(y_v[1] - \Theta_1\varphi(x))^2 + (y_v[2] - \Theta_2\varphi(x))^2\right] + R(\theta) \\ \text{subject to} \quad & \frac{1}{n_c}\left|\sum_{i=1}^{n_c}(z_i - \bar{z})x_{v,i}^\top\Theta_v\right| \le \epsilon. \end{aligned} \tag{9}$$

where $R(\theta)$ is a strongly convex regularizer, e.g., $R(\theta) = \varkappa\|\theta\|_2^2$. Note that for squared loss, if the features are linearly independent, then the covariance matrix $\mathbb{E}\left[\varphi(x)\varphi(x)^\top\right]$ is positive definite and $\varkappa$ can be set to 0.

C PROOFS OF SECTION 4

Before we state the proofs, note that for both, CE loss (7) and squared loss (8), the following conditions hold. Let θ^* denote the global optimal of the optimization problem. With a little abuse of notation, let $l(\theta)$ denote both squared and CE loss.

Fact 1. $l(\theta)$ has Lipschitz continuous gradient, i.e., for all $\theta_1, \theta_2 \in \Theta \subseteq \mathbb{R}^d$ we have,

$$\|\nabla l(\theta_1) - \nabla l(\theta_2)\|_2 \le L_G\|\theta_1 - \theta_2\|_2.$$

where $L_G > 0$ is a constant. There exists a $L_H, \delta > 0$ such that, for $\theta \in \Theta \cap \{\theta : \|\theta - \theta^*\|_2 \le \delta\}$, we have

$$\left\|\nabla l(\theta) - \nabla l(\theta^*) - \nabla^2 l(\theta)(\theta - \theta^*)\right\|_2 \le L_H\|\theta - \theta^*\|_2^2.$$

Fact 2. The following bound holds on the growth of $\|\nabla l(\theta)\|_2$,

$$\|\nabla l(\theta)\|_2^2 \le C(1 + \|\theta\|_2^2),$$

for some constant $C > 0$.

Fact 3. There is a constant L_E, such that, for all $\theta \in \Theta^*$, and $W \sim \mathcal{D}$,

$$\mathbb{E}\left[\|\nabla L(\theta, W) - \nabla L(\theta^*, W)\|_2^2\right] \le L_E\|\theta - \theta^*\|_2^2. \tag{10}$$

C.1 PROOF PROPOSITION 4.1

Proof. [Proof of Proposition 4.1] Here the constraint can be split into two constraints $\tilde{x}^\top \theta \leq \epsilon$ and $(-\tilde{x})^\top \theta \leq \epsilon$. Clearly only one of these constraints can be active at a time. At the optima $\theta = \theta^*_{DI}$, let's denote the active constraint by $A\theta^*_{DI} = \epsilon$. Then the KKT condition is given by,

$$\nabla l(\theta^*_{DI}) + \lambda^* A = 0,$$

where $\lambda^* > 0$. Since our constraint is linear, using Assumption 4.1, for any $s \in \mathbb{R}^d$, we have

$$s^\top \left[\nabla^2 l(\theta^*_{DI}) + \nabla^2 (A\theta^*_{DI})\right] s \geq \mu \left\|s\right\|_2^2. \tag{11}$$

This implies Assumption C of Duchi and Ruan [2016]. We first need the following result from 4.1.

Lemma C.1 (Theorem 3 Duchi and Ruan [2016]) *Let Assumption 4.1- Assumption 4.3 hold. Then, with probability one, there exists some $K < \infty$, such that for $k \geq K$,*

$$A\theta_k = \epsilon.$$

This implies that after a finite number of steps, Algorithm 1 almost surely identifies the active set at the optima A correctly. Let P_A denote the projection matrix onto the null space of the active constrained. Then, almost surely,

$$P_A = I - A^\top (AA^\top)^\dagger A = I - \tilde{x}(\tilde{x}^\top \tilde{x})^\dagger \tilde{x}^\top = I - \frac{\tilde{x}\tilde{x}^\top}{\left\|\tilde{x}\right\|_2^2}.$$

Then Proposition 4.1 follows from Theorem 4 of Duchi and Ruan [2016]. ∎

D PROOFS OF SECTION 5

D.1 PROOF OF THEOREM 5.1

We first show that $\bar{\theta}^b_k$ converges almost surely to θ^* for any $b = 1, 2, \cdots, B$. In Algorithm 2, the stochastic gradient term $\nabla L(\theta^b_k, W_k)$ is replaced by the perturbed stochastic gradient $V^b_k \nabla L(\theta^b_k, W_k)$. Note that by (10), and the fact that V^b_k is independent of θ^b_k, and W_k, we have the following inequality for the perturbed noisy gradient,

$$\mathbb{E}\left[{V^b_k}^2 \left\|\nabla L(\theta, W) - \nabla L(\theta^*, W)\right\|_2^2\right] \leq 2L_E \left\|\theta - \theta^*\right\|_2^2. \tag{12}$$

Then using Theorem 2 of Duchi and Ruan [2016] we have,

Proposition D.1 (Theorem 2, Duchi and Ruan [2016]) *Under Assumption 4.1-4.3, for any $b \in \{1, 2, \cdots, B\}$, we have,*

$$\bar{\theta}^b_k \overset{a.s.}{\to} \theta^*.$$

We re-state Theorem 5.1 here for convenience.

Theorem D.1 *Let Assumptions 4.1-4.3 be true. Then, choosing $\eta_k = k^{-a}$, $1/2 < a < 1$ in Algorithm 2, for any $b \in 1, 2, \cdots, B$, we have, $\bar{\theta}_k \overset{a.s.}{\to} \theta^*_{DI[DM]}$, and*

1. $\sqrt{n}\left(\bar{\theta}^b_k - \bar{\theta}_k\right) \overset{d}{\sim} N\left(0, \Sigma_{DI[DM]}\right)$,

2.

$$\sup_{q \in \mathbb{R}^d} \left|\mathbb{P}(\sqrt{k}(\bar{\theta}^b_k - \bar{\theta}_k) \leq q) - \mathbb{P}(\sqrt{k}(\bar{\theta}_k - \theta^*_{DI[DM]}) \leq q)\right| \overset{P}{\to} 0,$$

3.

$$\sup_{c \in [0,1]} \left|P\left(\sqrt{k}\left(\hat{\phi}_{DI[DM]}(\bar{\theta}^b_k) - \phi_{DI[DM]}(\bar{\theta}_k)\right) \leq c\right) - P\left(\sqrt{k}\left(\phi_{DI[DM]}(\bar{\theta}_k) - \phi_{DI[DM]}(\theta^*_{DI[DM]})\right) \leq c\right)\right| \overset{P}{\to} 0.$$

Proof. [Proof of Theorem 5.1]

1. Note that with (12), all the assumptions required for Lemma C.1 hold. Then, from Lemma C.1 we have that, after $k \geq K$, Algorithm 2 almost surely identifies the active constraint at optima. For a perturbed trajectory let A^b denote this active constraint, i.e., $A^b\theta_k^b = \epsilon$ for $k \geq K$. Let P_{A^b} be the orthogonal projector to the null space of A^b. Note that $P_{A^b} = P_A$ since P_A is an even function of A. Then, for $k \geq K$, $P_A(\theta_k^b - \theta_{DI}^*) = \theta_k^b - \theta_{DI}^*$. So proving the asymptotic normality of the projected sequence $P_A(\theta_k^b - \theta_{DI}^*)$ is sufficient in this case. From Algorithm 2, we have,

$$u_k^b - u_{k-1}^b = \eta_k V_k^b \nabla L(\theta_k^b, W_k). \tag{13}$$

Here the constraint can be split into two constraints $\tilde{x}^\top\theta \leq \epsilon$ and $(-\tilde{x})^\top\theta \leq \epsilon$. Clearly only one of these constraints can be active at a time. Let's denote the active constraint by $A^b\theta_k^b = \epsilon$. Then by KKT conditions for the projection step (line 4 of Algorithm 2, we have, for some $\lambda_k^b \geq 0$, and $\mu_k^b \geq 0$,

$$\theta_{k+1}^b + u_k^b + \lambda_k^b A^b - \mu_k^b A^b = 0. \tag{14}$$

Then combining (13), and (14), we get

$$\theta_{k+1}^b = \theta_k^b - \eta_k V_k^b \nabla L(\theta_k^b, W_k) - A^b(\lambda_k^b - \lambda_{k-1}^b) + A^b(\mu_k^b - \mu_{k-1}^b).$$

Multiplying both sides by the projection matrix $P_A = I - {A^b}^\top(A^b{A^b}^\top)^\dagger A^b = I - \tilde{x}\tilde{x}^\top / \|\tilde{x}\|_2^2$, and using $P_A A^b = 0$, we have,

$$\begin{aligned}
&P_A(\theta_{k+1}^b - \theta_{DI}^*) \\
=&P_A(\theta_k^b - \theta_{DI}^*) - \eta_k V_k^b P_A \nabla L(\theta_k^b, W_k) - P_A A^b(\lambda_k^b - \lambda_{k-1}^b) + P_A A^b(\mu_k^b - \mu_{k-1}^b) \\
=&(I - \eta_k P_A \nabla^2 l(\theta_{DI}^*) P_A) P_A(\theta_k^b - \theta_{DI}^*) + \eta_k P_A \nabla^2 l(\theta_{DI}^*) P_A(\theta_k^b - \theta_{DI}^*) - \eta_k V_k^b P_A \nabla L(\theta_k^b, W_k) \\
=&(I - \eta_k P_A \nabla^2 l(\theta_{DI}^*) P_A) P_A(\theta_k^b - \theta_{DI}^*) - \eta_k (I - P_A)\nabla^2 l(\theta_{DI}^*) P_A(\theta_k^b - \theta_{DI}^*) + \eta_k \nabla^2 l(\theta_{DI}^*) P_A(\theta_k^b - \theta_{DI}^*) \\
&+ \eta_k P_A \nabla l(\theta_{DI}^*) - \eta_k P_A \nabla l(\theta_k^b) + \eta_k P_A \nabla l(\theta_k^b) - \eta_k V_k^b P_A \nabla L(\theta_k^b, W_k) \\
=&(I - \eta_k P_A \nabla^2 l(\theta_{DI}^*) P_A) P_A(\theta_k^b - \theta_{DI}^*) - \eta_k P_A \zeta_k^b + \eta_k P_A \xi_k^b + \eta_k P_A D_k^b + \epsilon_k^b,
\end{aligned}$$

where

$$\begin{aligned}
\xi_k^b &\coloneqq \nabla l(\theta_k^b) - V_k^b \nabla L(\theta_k^b, W_k), \\
\zeta_k^b &\coloneqq \nabla L(\theta_k^b) - \nabla l(\theta_{DI}^*) - \nabla^2 l(\theta_{DI}^*)(\theta_k - \theta_{DI}^*), \text{and} \\
\epsilon_k^b &\coloneqq -\eta_k P_A \nabla^2 l(\theta_{DI}^*)(I - P_A)(\theta_k^b - \theta_{DI}^*).
\end{aligned}$$

The last inequality follows from the fact that we have $P_A \nabla l(\theta_{DI}^*) = 0$ by optimality properties. The following decomposition of $P_A\xi_k^b$ takes place.

$$\begin{aligned}
P_A\xi_k^b &= P_A\nabla l(\theta_k^b) - P_A\nabla l(\theta_{DI}^*) + V_k^b P_A \nabla L(\theta_{DI}^*, W_k) - V_k^b P_A \nabla L(\theta_k^b, W_k) - V_k^b P_A \nabla L(\theta_{DI}^*, W_k) \\
&= \xi_k^b(0) + \xi_k^b(\theta_k^b),
\end{aligned}$$

where,

$$\xi_k^b(0) \coloneqq -V_k^b P_A \nabla L(\theta_{DI}^*, W_k) \quad \xi_k^b(\theta_k^b) \coloneqq P_A\left(\nabla l(\theta_k^b) - \nabla l(\theta_{DI}^*)\right) + V_k^b P_A\left(\nabla L(\theta_{DI}^*, W_k) - \nabla L(\theta_k^b, W_k)\right).$$

Then, using Young's lemma, and 10, we have,

$$\mathbb{E}\left[\left\|\xi_k^b(\theta_k^b)\right\|_2^2\right] \leq C\left\|\theta_k^b - \theta_{DI}^*\right\|_2^2,$$

for some constant $C > 0$. Then using Proposition 2, specifically equation (65) of Duchi and Ruan [2016], we have,

$$\frac{1}{\sqrt{n}}\sum_{k=1}^n\left(\theta_k^b - \theta_{DI}^*\right) = (P_A\nabla^2 l(\theta_{DI}^*)P_A)^\dagger \frac{1}{\sqrt{n}}\sum_{k=1}^n V_k^b P_A \nabla L(\theta_{DI}^*, W_k) + o_{\mathbb{P}}(1).$$

Similarly, from Algorithm 1, we have,

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(\theta_k-\theta^*_{DI})=(P_A\nabla^2 l(\theta^*_{DI})P_A)^\dagger\frac{1}{\sqrt{n}}\sum_{k=1}^{n}P_A\nabla L(\theta^*_{DI},W_k)+o_{\mathbb{P}}(1).$$

Combining the above two equations we get,

$$\sqrt{n}\left(\bar{\theta}^b_n-\bar{\theta}_n\right)=(P_A\nabla^2 l(\theta^*_{DI})P_A)^\dagger\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(V^b_k-1)P_A\nabla L(\theta^*_{DI},W_k)+o_{\mathbb{P}}(1).$$

Then by Martingale central limit theorem, we have that the asymptotic distribution of $\sqrt{n}\left(\bar{\theta}^b_n-\bar{\theta}_n\right)$ is Gaussian with mean 0 and asymptotic covariance,

$$\begin{aligned}\Sigma_{DI}&=\lim_{n\to\infty}\mathbb{E}\left[(P_A\nabla^2 l(\theta^*_{DI})P_A)^\dagger\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n}(V^b_k-1)P_A\nabla L(\theta^*_{DI},W_k)(V^b_i-1)\left(P_A\nabla L(\theta^*_{DI},W_i)\right)^\top\left(P_A\nabla^2 l(\theta^*_{DI})P_A\right)^{\dagger\top}\right]\\
&\overset{(1)}{=}\lim_{n\to\infty}(P_A\nabla^2 l(\theta^*_{DI})P_A)^\dagger\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\left[(V^b_k-1)^2\right]P_A\mathbb{E}\left[\nabla L(\theta^*_{DI},W_k)\nabla L(\theta^*_{DI},W_k)^\top\right]P_A^\top(P_A\nabla^2 l(\theta^*_{DI})P_A)^{\dagger\top}\\
&\overset{(2)}{=}(P_A\nabla^2 l(\theta^*_{DI})P_A)^\dagger P_A S_{DI}P_A^\top(P_A\nabla^2 l(\theta^*_{DI})P_A)^{\dagger\top}\\
&\overset{(3)}{=}P_A\nabla^2 l(\theta^*_{DI})^\dagger P_A S_{DI}P_A\nabla^2 l(\theta^*_{DI})^\dagger P_A.\end{aligned}$$

Equality (1) follows from the martingale difference property of $\{(V^b_k-1)\nabla L(\theta^*_{DI},W_k)\}_k$. Equality (2) follows from the facts that V^b_k is independent of $\nabla L(\theta^*_{DI},W_k)$, $\mathbb{E}\left[(V^b_k-1)^2\right]=1$, and Assumption 4.2. Equation (3) follows from the fact $P_A^2=P_A$ since P_A is a projection matrix. This proves first statement of Theorem 5.1

2. Let ϑ denote the Gaussian distribution $N(0,\Sigma_{DI})$. Then, from the previous part we can write,

$$\sup_{q\in\mathbb{R}^d}\left|\mathbb{P}(\sqrt{n}\left(\bar{\theta}^b_n-\bar{\theta}_n\right)\le q)-\mathbb{P}(\vartheta\le q)\right|\xrightarrow{P}0. \tag{15}$$

Similarly, from Proposition 4.1, we have

$$\sup_{q\in\mathbb{R}^d}\left|\mathbb{P}(\sqrt{n}\left(\bar{\theta}_n-\theta^*_{DI}\right)\le q)-\mathbb{P}(\vartheta\le q)\right|\xrightarrow{P}0. \tag{16}$$

Combining (15), and (16), we get the second part of the Theorem D.1,

$$\sup_{q\in\mathbb{R}^d}\left|\mathbb{P}(\sqrt{n}\left(\bar{\theta}^b_n-\bar{\theta}_n\right)\le q)-\mathbb{P}(\sqrt{n}\left(\bar{\theta}_n-\theta^*_{DI}\right)\le q)\right|\xrightarrow{P}0. \tag{17}$$

■

Proof. [Proof of Theorem 5.2] Since the result holds for both DI and DM, in this proof, we will remove the DI and DM subscript from θ^* and Δ unless there is any ambiguity. Assumption 4.4 implies that the map $\theta\mapsto\mathbb{P}(\varphi(X)^\top\theta>0\mid\cdot)$ is Fréchet differentiable at θ^*, hence so is $\Delta(\theta)$. Therefore, by first order Taylor expansion we have,

$$\Delta(\bar{\theta}_k)=\Delta(\theta^*)+\nabla\Delta(\theta^*)^\top(\bar{\theta}_k-\theta^*)+O(\|\bar{\theta}_k-\theta^*\|_2^2).$$

Now, since $\bar{\theta}_k\xrightarrow{a.s.}\theta^*$, by part (1) of Theorem 5.1, we have,

$$\begin{aligned}\sqrt{k}\big(\Delta(\bar{\theta}_k)-\Delta(\theta^*)\big)&=\nabla\Delta(\theta^*)^\top\sqrt{k}(\bar{\theta}_k-\theta^*)+o_{\mathbb{P}}(1)\\
&\Rightarrow N(0,\sigma^2_\Delta),\end{aligned}$$

with $\sigma^2_\Delta=\nabla\Delta(\theta^*)^\top\Sigma\nabla\Delta(\theta^*)$.

By mean value theorem,

$$\Delta(\bar{\theta}^b_k)-\Delta(\bar{\theta}_k)=\nabla\Delta(\tilde{\theta}_k)^\top(\bar{\theta}^b_k-\bar{\theta}_k),$$

for some intermediate point $\tilde{\theta}_k$ between $\bar{\theta}_k^b$ and $\bar{\theta}_k$. By part (1) of Theorem 5.1, $\bar{\theta}_k \xrightarrow{a.s.} \theta^*$. By part (2) of Theorem 5.1,

$$\sqrt{k}(\bar{\theta}_k^b - \bar{\theta}_k) = O_{\mathbb{P}}(1),$$

and therefore

$$\bar{\theta}_k^b - \bar{\theta}_k = O_{\mathbb{P}}(k^{-1/2}) \xrightarrow{\mathbb{P}} 0.$$

Since

$$\tilde{\theta}_k = \bar{\theta}_k + t_k(\bar{\theta}_k^b - \bar{\theta}_k),$$

for some $t_k \in (0,1)$, we have

$$\|\tilde{\theta}_k - \theta^*\| \le \|\bar{\theta}_k - \theta^*\| + \|\bar{\theta}_k^b - \bar{\theta}_k\|.$$

Therefore,

$$\tilde{\theta}_k \xrightarrow{\mathbb{P}} \theta^*.$$

Since $\nabla\Delta(\cdot)$ is continuous at θ^*,

$$\nabla\Delta(\tilde{\theta}_k) = \nabla\Delta(\theta^*) + o_{\mathbb{P}}(1).$$

Therefore,

$$\sqrt{k}\big(\Delta(\bar{\theta}_k^b) - \Delta(\bar{\theta}_k)\big) = \nabla\Delta(\theta^*)^\top \sqrt{k}(\bar{\theta}_k^b - \bar{\theta}_k) + o_{\mathbb{P}}(1),$$

Now using Slutsky's theorem, and part (2) of Theorem 5.1, we have,

$$\sqrt{k}\big(\Delta(\bar{\theta}_k^b) - \Delta(\bar{\theta}_k)\big) \xrightarrow{d} N(0, \sigma_\Delta^2).$$

Now consider the following two cases.
Case 1: $\Delta(\theta^*) \neq 0$. Since $f(u) = |u|$ is continuously differentiable at $u = \Delta(\theta^*)$ with derivative $f'(\Delta(\theta^*)) = \operatorname{sign}(\Delta(\theta^*))$, we have,

$$\sqrt{k}\Big(|\Delta(\bar{\theta}_k)| - |\Delta(\theta^*)|\Big) = \operatorname{sign}(\Delta(\theta^*))\sqrt{k}(\Delta(\bar{\theta}_k) - \Delta(\theta^*)) + \sqrt{k}R_k, \tag{18}$$

where $R_k = 0$ when $\Delta(\bar{\theta}_k)$ and $\Delta(\theta^*)$ are of the same sign, and $0 \le R_k \le 2$ otherwise. Observe that,

$$R_k \neq 0 \implies |\Delta(\bar{\theta}_k) - \Delta(\theta^*)| > |\Delta(\theta^*)|/2.$$

Then, for any $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}(\sqrt{k}R_k > \epsilon) \le& \mathbb{P}(R_k \neq 0) \\
\le& P\left(|\Delta(\bar{\theta}_k) - \Delta(\theta^*)| > |\Delta(\theta^*)|/2\right) \\
=& P\left(\sqrt{k}|\Delta(\bar{\theta}_k) - \Delta(\theta^*)| > \sqrt{k}|\Delta(\theta^*)|/2\right).
\end{aligned} \tag{19}$$

But we know, $\sqrt{k}(\Delta(\bar{\theta}_k) - \Delta(\theta^*)) = O_{\mathbb{P}}(1)$ whereas the RHS of (19) is diverging with k. So,

$$\mathbb{P}(\sqrt{k}R_k > \epsilon) \to 0.$$

Then, from (18), we have,

$$\sqrt{k}\Big(\phi(\bar{\theta}_k) - \phi(\theta^*)\Big) = \operatorname{sign}(\Delta(\theta^*))\sqrt{k}(\Delta(\bar{\theta}_k) - \Delta(\theta^*)) + o_{\mathbb{P}}(1). \tag{20}$$

By the CLT for $\Delta(\bar{\theta}_k)$,

$$\sqrt{k}\big(\phi(\bar{\theta}_k) - \phi(\theta^*)\big) \xrightarrow{d} N(0, \sigma_\Delta^2).$$

The bootstrap statement follows by the same argument applied to $\phi(\bar{\theta}_k^b) - \phi(\bar{\theta}_k)$, combined with the conditional CLT for $\sqrt{k}(\Delta(\bar{\theta}_k^b) - \Delta(\bar{\theta}_k))$ and Slutsky's theorem.

Case 2: $\Delta(\theta^*) = 0$. In this case, since $\sqrt{k}\Delta(\bar{\theta}_k) \Rightarrow N(0, \sigma_\Delta^2)$, by the continuous mapping theorem, we have,

$$\sqrt{k}\phi(\bar{\theta}_k) = \left|\sqrt{k}\Delta(\bar{\theta}_k)\right| \Rightarrow |N(0, \sigma_\Delta^2)|,$$

which is half-normal. The same argument applied conditionally on data to $\bar{\theta}_k^b$ gives $\sqrt{k}\phi(\bar{\theta}_k^b) \Rightarrow |N(0, \sigma_\Delta^2)|$. ∎

Proposition D.2 (Consistency of $\hat{\phi}_{DI[DM]}$) *Let Assumptions 4.1–4.4 hold. Then, as $|\tilde{\mathcal{D}}| \to \infty$,*

$$|\hat{\phi}_{DI}(\bar{\theta}_k) - \phi_{DI}(\bar{\theta}_k)| \xrightarrow{\mathbb{P}} 0.$$

The same conclusion holds for $\hat{\phi}_{DM}$.

Proof. [Proof of Proposition D.2] Let $m := |\tilde{\mathcal{D}}|$ and define for $z \in \{0, 1\}$

$$\hat{p}_{z,m}(\theta) = \frac{\sum_{i \in \tilde{\mathcal{D}}} \mathbb{1}(z_i = z) q_i}{n_{z,m}}, \qquad p_z(\theta) = \mathbb{P}(\theta^\top \varphi(x) > 0 \mid z),$$

where $n_{z,m} = \sum_{i \in \tilde{\mathcal{D}}} \mathbb{1}(z_i = z)$.

Consider the class

$$\mathcal{F} = \big\{(x, z) \mapsto \mathbb{1}(z = 0)\mathbb{1}(\theta^\top \varphi(x) > 0),\ \mathbb{1}(z = 1)\mathbb{1}(\theta^\top \varphi(x) > 0) : \theta \in \Theta\big\}.$$

Since halfspaces in $\mathbb{R}^d$ have finite VC dimension, $\mathcal{F}$ has finite VC dimension Van Der Vaart and Wellner [2009], Shalev-Shwartz and Ben-David [2014]. Hence, by the uniform law of large numbers for VC classes,

$$\sup_{\theta \in \Theta} \left| \frac{1}{m} \sum_{i \in \tilde{\mathcal{D}}} \mathbb{1}(Z_i = z)\mathbb{1}(\theta^\top \varphi(X_i) > 0) - \mathbb{E}[\mathbb{1}(Z = z)\mathbb{1}(\theta^\top \varphi(X) > 0)] \right| \xrightarrow{\mathbb{P}} 0.$$

By the law of large numbers, $n_{z,m}/m \to \mathbb{P}(z) > 0$ in probability. Combining numerator and denominator convergence yields

$$\sup_{\theta \in \Theta} |\hat{p}_{z,m}(\theta) - p_z(\theta)| \xrightarrow{\mathbb{P}} 0.$$

Since $\phi_{DI}(\theta) = |p_0(\theta) - p_1(\theta)|$ and $|\cdot|$ is Lipschitz,

$$\sup_{\theta \in \Theta} |\hat{\phi}_{DI}(\theta) - \phi_{DI}(\theta)| \le \sup_{\theta \in \Theta} |\hat{p}_{0,m}(\theta) - p_0(\theta)| + \sup_{\theta \in \Theta} |\hat{p}_{1,m}(\theta) - p_1(\theta)| \xrightarrow{\mathbb{P}} 0.$$

Since $\bar{\theta}_k$ is independent of $\tilde{\mathcal{D}}$,

$$|\hat{\phi}_{DI}(\bar{\theta}_k) - \phi_{DI}(\bar{\theta}_k)| \le \sup_{\theta \in \Theta} |\hat{\phi}_{DI}(\theta) - \phi_{DI}(\theta)|,$$

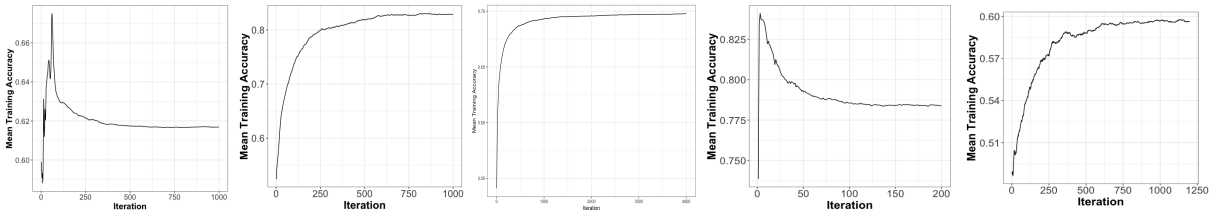
which converges to zero in probability. ∎

E FURTHER DETAILS ON EXPERIMENTS

Dataset	ϵ	B	$\lvert\tilde{\mathcal{D}}\rvert$	$\lvert\mathcal{D}'\rvert$	R_2	#Repetitions
Synthetic (DI)	0.002	100	100	200	–	100
`Adult`	0.00001	200	1000	1000	–	200
`ACSIncome`	0.01	100	1000	1000	–	200
Synthetic (DM)	–	100	100	250	500	100
`COMPAS`	–	200	400	250	300	200

Table 1: The choice of parameters across experiments.

Across all experiments, to find the global optima we initially run the algorithm once until $\|\theta_k - \theta_{k-1}\|_2 < 10^{-7}$. CE loss is strictly convex and squared loss is almost-surely strongly convex for i. i. d data. We add a regularizer $R(\theta) = \varkappa \|\theta\|_2^2 / 2$ to the canonical expected losses to ensure strong convexity. Note that $\varkappa$ can be tuned to ensure sparsity in high-dimensional cases. We choose small $\varkappa > 0$ since our goal is not sparsity but just to ensure strong convexity. For all the experiments we choose the step size $\eta_k = 0.01k^{-0.501}$, and $\varkappa = 0.0001$. We continue the training until the accuracy and fairness have reached stability. The detailed choice of the experimental setup is provided in Table 1.



(a) (b) (c) (d) (e)

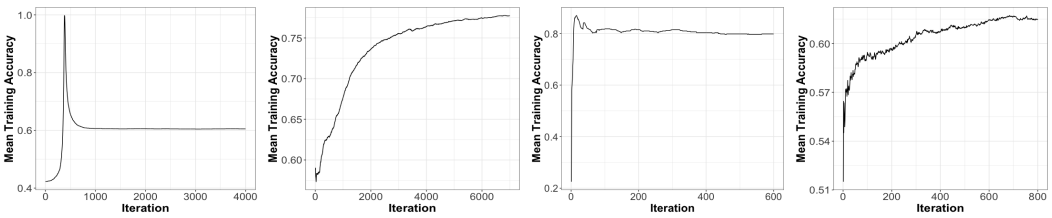
Figure 7: Mean training accuracy of Cross-Entropy loss: (a) DI with synthetic data, (b) DI with Adult data, (c) DI with ACSIncome data, (d) DM with synthetic data, (e) DM with COMPAS data

.

F MEAN TRAINING ACCURACY FOR CROSS-ENTROPY LOSS

Even though our methods and results are not conditional on a particular accuracy level of the classifier, we train the models such that we obtain mean accuracy similar to what has been achieved under these experimental setups and datasets in other works (for example, Zafar et al. [2019]). Figure 7 shows the mean training accuracy for CE loss.

G PLOTS CORRESPONDING TO SQUARED LOSS

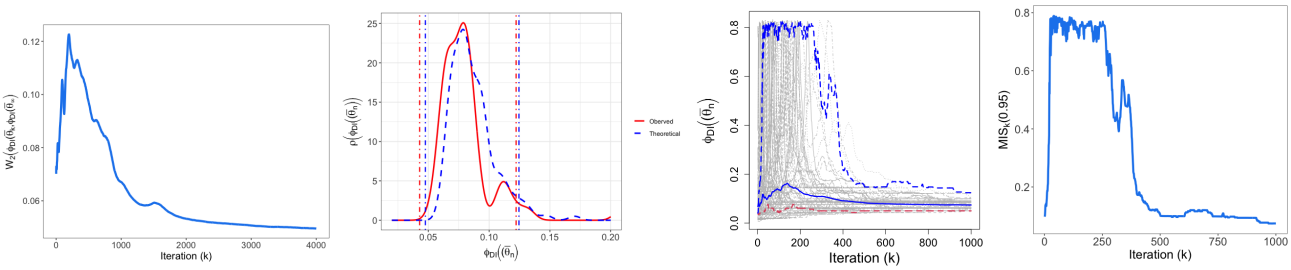


(a) (b) (c) (d)

Figure 8: Mean training accuracy of squared loss: (a) DI with synthetic data, (b) DI with Adult data, (c) DM with synthetic data, (d) DM with COMPAS data

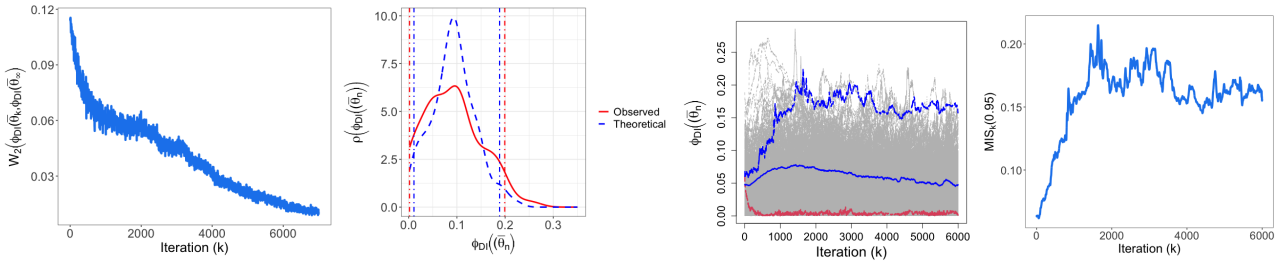
.

For completion, we include the results on squared loss here. Figure 8 shows the mean training accuracy for squared loss. The plots related to uncertainty quantification of fairness on test data are shown in Figure 9 - Figure 12. Similar to CE loss, our estimated online CI matches the observed CI for squared loss as well. The widths of the confidence intervals obtained in all the experiments suggest that the classifier trained under squared loss has higher variability. But a proper comparison of squared loss and CE loss in terms of fairness uncertainty needs a more nuanced analysis which we leave for future work.



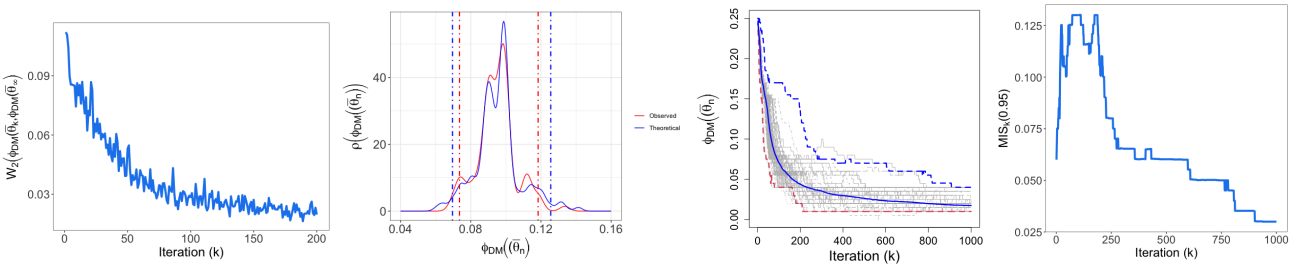
(a) (b) (c) (d)

Figure 9: Top row: Convergence of $W_2\left(\phi_{DI}(\bar{\theta}_k), \phi_{DI}(\bar{\theta}_\infty)\right)$, and comparison of theoretical asymptotic density and observed density of $\phi_{DI}(\bar{\theta}_k)$; Bottom row: Online Bootstrap 95% CI and $MIS(\hat{\phi}_{DI,0.025}, \hat{\phi}_{DI,0.975}; 0.95)$ under squared loss for synthetic data at risk of DI.



(a) (b) (c) (d)

Figure 10: Top row: Convergence of $W_2\left(\phi_{DI}(\bar{\theta}_k), \phi_{DI}(\bar{\theta}_\infty)\right)$, and comparison of theoretical asymptotic density and observed density of $\phi_{DI}(\bar{\theta}_k)$; Bottom row: Online Bootstrap 95% CI and $MIS(\hat{\phi}_{DI,0.025}, \hat{\phi}_{DI,0.975}; 0.95)$ under squared loss for Adult data at risk of DI.



(a) (b) (c) (d)

Figure 11: Top row: Convergence of $W_2\left(\phi_{DM}(\bar{\theta}_k), \phi_{DM}(\bar{\theta}_\infty)\right)$, and comparison of theoretical asymptotic density and observed density of $\phi_{DM}(\bar{\theta}_k)$; Bottom row: Online Bootstrap 95% CI and $MIS(\hat{\phi}_{DM,0.025}, \hat{\phi}_{DM,0.975}; 0.95)$ under squared loss for synthetic data vulnerable to DM.

G.1 NEURAL NETWORK CLASSIFIER: BEYOND NTK REGIME

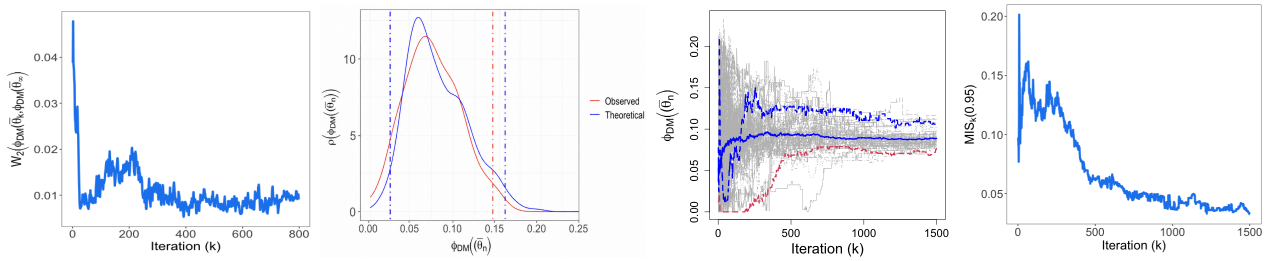
In this section we experimentally demonstrate the potential of applicability of our method to general neural network classifier beyond NTK regime. We leave the extension of our theoretical results to this case to future work. As a proof-of-concept experiment, consider the neural network

$$W_2^\top \tanh(W_1^\top x + b_1) + b_2,$$

where $W_1 \in \mathbb{R}^{d\times h}$, $W_2 \in \mathbb{R}^h$, $b_1 \in \mathbb{R}^h$, and $b_2 \in \mathbb{R}$. Then the decision boundary is given by

$$\delta_\theta(x) = W_2^\top \tanh(W_1^\top x + b_1) + b_2.$$

We illustrate our method on COMPAS dataset with DM as the fairness measure of interest. Here $d = 5$, and we chose a small value of $h = 15$ to avoid operating in NTK regime. Entries of W_1, and W_2 are initialized i.i.d from $N(0, 1/\sqrt{d})$, and $N(0, 1/\sqrt{h})$ respectively. b_1, b_2 are initialized at 0. We use $B = 100$, and $|\tilde{D}| = 400$. Figure 13 shows that our method achieves almost correct coverage probability (slightly overconfident) along with decreasing MIS indicating it identifies narrowest confidence interval with correct coverage. This toy example shows potential of wider applicability and serves as a motivation further research.



(a) (b) (c) (d)

Figure 12: Top row: Convergence of $W_2\left(\phi_{DM}(\bar{\theta}_k), \phi_{DM}(\bar{\theta}_\infty)\right)$, and comparison of theoretical asymptotic density and observed density of $\phi_{DM}(\bar{\theta}_k)$; Bottom row: Online Bootstrap 95% CI and $MIS(\hat{\phi}_{DM,0.025}, \hat{\phi}_{DM,0.975}; 0.95)$ under squared loss for COMPAS data.

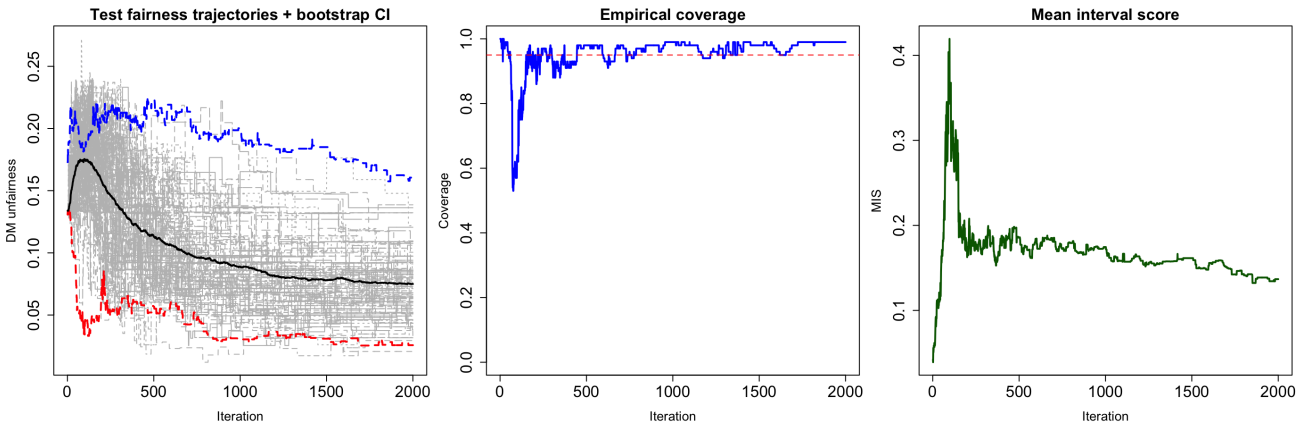


Figure 13: Proof-of-concept experiment with neural network classifier for DM aware classifier for COMPAS dataset. (Left) The random trajectories are contained within the bootstrap-based confidence interval. (Middle) Correct coverage probability. (Right) Decreasing MIS. This experiment shows that our approach is promising even for nonlinear nonconvex models like neural network classifiers.