Statistical Guarantees for Fairness Aware Plug-In Algorithms

Anonymous Authors[1]

Abstract

A plug-in algorithm to estimate Bayes Optimal Classifiers for fairness-aware binary classification has been proposed in (Menon & Williamson, 2018). However, the statistical efficacy of their approach has not been established. We prove that the plug-in algorithm is statistically consistent. We also derive finite sample guarantees associated with learning the Bayes Optimal Classifiers via the plug-in algorithm. Finally, we propose a protocol that modifies the plug-in approach, so as to simultaneously guarantee fairness and differential privacy with respect to a binary feature deemed sensitive.

1. Introduction and Related Work

Bayes Optimal Classifiers (BOCs) (Devroye et al., 1996) are of significant importance, since they achieve the least average error possible for any classification task. However, BOCs are generally specified in terms of unknown distributional quantities. Constructing sound estimators for BOCs, provided access to only a finite training sample, is thus of utmost practical relevance. One approach to estimating the BOC is through constructing 'plug-in' estimators. The plug-in principle applied to a broad class of problems, including that of binary classification, is well studied in the statistics literature (Audibert et al., 2007; Denis & Hebiri, 2017; Yang, 1999). Indeed, the existence of a plug-in classifier that is optimal in the minimax sense is established in (Audibert et al., 2007; Yang, 1999). In their work, (Menon & Williamson, 2018) propose a plug-in algorithm to estimate the BOCs corresponding to fairness-aware learning (FAL) tasks. However, (Menon & Williamson, 2018) do not provide guarantees on the statistical efficacy of their algorithm. In this paper, we plug this gap by ***proving that the plug-in algorithm of** (Menon & Williamson, 2018), **is indeed statistically consistent***. We also ***characterise the sample complexity associated with the task of learning a low regret classifier via the plug-in algorithm***. Closest to our work is that of (Chzhen et al., 2019), wherein an asymptotic study (for a different fairness aware plug-in classifier) is carried out. The work of (Chzhen et al., 2019) however, focuses on settings wherein perfect fairness constraints are imposed. It is well established that due to inherent fairness-accuracy trade-offs, ensuring perfect fairness without considerable loss in accuracy is generally not possible (Menon & Williamson, 2018; Zhao & Gordon, 2019; Chen et al., 2018). We thus focus on approximate notions of two fairness metrics, Demographic Parity (DPar) and Equality of Opportunity (EO). Further, the approach of (Chzhen et al., 2019) requires access to the sensitive variable (denoted $\overline{Y}$ hereon) at test time which is often not permitted. The plug-in approach of (Menon & Williamson, 2018) however does not necessitate test-time access to $\overline{Y}$. Indeed, real-world settings may impose even more stringent requirements on $\overline{Y}$. For example, we may be required to ensure that our model does not leak information about the sensitive attribute, $\overline{Y}$, corresponding to any individual. In such cases, a possible solution is to protect individuals via Differential Privacy (DP) (Dwork et al., 2006). The literature combining fairness and privacy (Jagielski et al., 2019; Cummings et al., 2019; Mozannar et al., 2020), is emerging and limited. Such settings motivate us to ***propose an easy to deploy, modified version of the plug-in algorithm, referred to as DP Plug-in***. The framework ensures that $\overline{Y}$ is protected via DP. Using publicly available data sets, we demonstrate empirically, that the ***DP Plug-in algorithm achieves strong privacy-fairness-accuracy guarantees***, as it outperforms the private, fair approach of (Jagielski et al., 2019) across 3 out of 4 experimental sets ups considered.

2. Background and Notation

For brevity, we only introduce the main features from (Menon & Williamson, 2018) pertinent to our study in this section. We present other useful definitions and results from (Menon & Williamson, 2018) in section A of the supplement. Additionally, our focus in the main thesis of this paper will be on the approximate EO criterion for the case when $\overline{Y}$ is unavailable during test-time. Analogous (and relatively simpler) analyses for 1) the case when $\overline{Y}$ is available at test time, and for 2) the approximate DPar criterion are presented in sections B and C of the

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

supplement for completeness.

Access to a finite training sample, $S = \{x_i, y_i, \overline{y}_i\}_{i=1}^n$ drawn i.i.d from some unknown distribution $\mathbb{P}$ is assumed in (Menon & Williamson, 2018). $\forall i \in [n]$ the triplet $(x_i, y_i, \overline{y}_i)$ is a realisation of the random variable triplet $(X, Y, \overline{Y})$ comprising of the feature, label and sensitive attribute respectively. Let $\pi = \mathbb{P}(Y = 1), \overline{\pi} = \mathbb{P}(\overline{Y} = 1)$ and $\beta = \mathbb{P}(\overline{Y} = 1|Y = 1)$. Assume that $\pi, \overline{\pi}, \beta > 0$

Let $(X, Y) \sim \mathcal{D}$, $(X, \overline{Y}|Y = 1) \sim \overline{\mathcal{D}}_{EO}$. $f : \chi \to [0, 1]$ denotes a randomised classifier on measurable domain χ. f yields predictions via $(\hat{Y}|X = x) \sim Bernoulli(f(x))$. Regression functions w.r.t. $\mathcal{D}, \overline{\mathcal{D}}_{EO}$ are given by $\eta(x) = \mathbb{P}(Y = 1|X = x)$, $\overline{\eta}_{EO}(x, y) = \mathbb{P}(\overline{Y} = 1|X = x, Y = y)$ respectively.

A central object of interest in (Menon & Williamson, 2018), is the notion of cost-sensitive risks (CSR). Denoting false positive and negative rates of f w.r.t $\mathcal{D}$, by $FPR_{\mathcal{D}}(f)$ and $FNR_{\mathcal{D}}(f)$ respectively, the CSR of a classifier f w.r.t a distribution $\mathcal{D}$, parameterised by $c \in [0, 1]$ is given by:

$$CS(f; \mathcal{D}, c) := c\,(1 - \pi)\,FPR_{\mathcal{D}}(f) + \pi(1 - c)FNR_{\mathcal{D}}(f)$$

Definition 2.2 *A binary classifier f, with corresponding predictor $\hat{Y}$ admits Equality of Opportunity if:* $\mathbb{P}(\hat{Y} = 1|Y = 1, \overline{Y} = 0) = \mathbb{P}(\hat{Y} = 1|Y = 1, \overline{Y} = 1)$

Thus, EO requires parity in the TPRs between groups as explicated in Definition 2.2. Obtaining perfect fairness while retaining non-trivial accuracy is generally not possible, and so (Menon & Williamson, 2018) introduce approximate measures of fairness which require the additive or multiplicative disparity between prediction rates to be small. A key lemma in (Menon & Williamson, 2018) draws an equivalence between the super-level sets of approximate fairness measures and CSRs. This in turn leads to a reduction of the FAL problem to a problem with constraints on cost-sensitive risks (The reader may refer to section A of the supplement for a more thorough presentation of these results). The FAL problem in (Menon & Williamson, 2018) is thus posed in terms of CSRs as follows:

Problem 2.1 **(Cost-sensitive FAL)** *For trade-off parameter $\lambda \in \mathbb{R}$, and cost parameters $c, \overline{c} \in (0, 1)^2$, minimise the fairness-aware cost-sensitive risk:*

$$R_{FA}(f; \mathcal{D}, \overline{\mathcal{D}}_{EO}, c, \overline{c}, \lambda) = CS(f; \mathcal{D}, c) - \lambda CS(f; \overline{\mathcal{D}}_{EO}, \overline{c})$$

Equipped with this soft constrained FAL problem formulated in terms of CSRs, (Menon & Williamson, 2018) derive the BOCs corresponding to such problems. We present the BOC for approx. EO and the plug-in algorithm of (Menon & Williamson, 2018) to estimate this BOC, in Theorem 2.2 and Algorithm 1 respectively. It is this (optimal) classifier and algorithm that will make for the key objects of our theoretical analysis in Section 3.

Theorem 2.2 **(BOC for FAL)** *Pick any costs $c, \overline{c} \in (0, 1)^2$ and trade-off parameter $\lambda \in \mathbb{R}$. Then:*

$$Argmin_{R_{FA}}(f; \mathcal{D}, \overline{\mathcal{D}}_{EO}) = \{H_\alpha \circ s^*(x) \,|\alpha \in [0, 1]\}$$

$$where,\ s^*(x) = \left\{1 - \frac{\lambda}{\overline{\pi}}\left(\overline{\eta}_{EO}(x, 1) - \overline{c}\right)\right\}\eta(x) - c$$

$$and,\ H_\alpha(z) = \mathbb{I}(z > 0) + \alpha\mathbb{I}(z = 0)$$

Algorithm 1 Plugin approach to FAL, EO setting

Input: Sample $S = \{x_i, y_i, \overline{y}_i\}_{i=1}^n$ from distribution $\mathbb{P}$; cost parameters $c, \overline{c}$; trade-off parameter λ

Estimate: π via $\hat{\pi} = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{y_i = 1\}$
Estimate: η:$\chi \to [0, 1]$ using appropriate CPE on $\{x_i, y_i\}_{i=1}^n$
Estimate: $\overline{\eta}_{EO}$: $(\chi, Y) \to [0, 1]$ using appropriate CPE on S
Compute: $\hat{s}(x) = \left\{1 - \frac{\lambda}{\hat{\overline{\pi}}}(\hat{\overline{\eta}}_{EO}(x, 1) - \overline{c})\right\}\hat{\eta}(x) - c$

Return: $\hat{f}(x) = H_\alpha(\hat{s}(x))$ for any $\alpha \in [0, 1]$

3. Theory

In this section, we analyse the asymptotic and non-asymptotic properties of the plug-in algorithm (Algorithm 1). Recall from Problem 2.1, that our goal is to minimise the fairness aware cost-sensitive risk, which for a given choice of cost parameters $c, \overline{c} \in [0, 1]^2$ and trade-off parameter $\lambda \in \mathbb{R}$ is given by $CS(f; \mathcal{D}, c) - \lambda CS(f; \overline{\mathcal{D}}_{EO}, \overline{c})$.

In the language of (Narasimhan et al., 2014), we introduce the notion of performance measure. A performance measure, defined w.r.t a distribution $\mathbb{P}$, and performance metric Ψ, is a mapping from the space of measurable functions $\mathcal{F}$ to the reals, i.e., $\mathfrak{P}_{\mathbb{P}}^{\Psi} : \mathcal{F} \to \mathbb{R}$. In our setting, the performance metric is simply given by the negative of the objective function of Problem 2.1. Unless stated otherwise, we will denote $\overline{\mathcal{D}}_{EO}$ by $\overline{\mathcal{D}}$ and $\overline{\eta}_{EO}$ by $\overline{\eta}$ from hereon. Our performance measure is given by:

$$\mathfrak{P}_{\mathbb{P}}^{\Psi}(f) = \Psi[TPR_{\mathcal{D}}(f), TNR_{\mathcal{D}}(f), \pi, TPR_{\overline{\mathcal{D}}}(f), TNR_{\overline{\mathcal{D}}}(f), \beta] = -\left\{CS\left(f; \mathcal{D}, c\right) - \lambda CS\left(f; \overline{\mathcal{D}}, \overline{c}\right)\right\}$$

Thus, a classifier's performance measure explains its merit with regards to the combined, fairness-utility objective, appropriately balanced by cost and trade-off parameters. Performance metric Ψ, makes explicit that our performance measure is a function of the classifier's TPRs and TNRs with respect to $\mathcal{D}$ and $\overline{\mathcal{D}}_{EO}$, as well as distributional quantities π and β. The CSRs, and thus the performance measure, are linear in TPRs, TNRs and class probabilities implying the performance measure is continuous in the arguments of Ψ. The regret of a classifier f, w.r.t performance measure $\mathfrak{P}_{\mathbb{P}}^{\Psi}$ is defined as: $regret_{\mathbb{P}}^{\Psi}(f) = \mathfrak{P}_{\mathbb{P}}^{\Psi,*} - \mathfrak{P}_{\mathbb{P}}^{\Psi}(f)$ where, $\mathfrak{P}_{\mathbb{P}}^{\Psi,*} = \mathfrak{P}_{\mathbb{P}}^{\Psi}(f^*)$. In our case, f^* is the BOC introduced in Theorem 2.2

3.1. Asymptotic Analysis

In this sub-section, we prove that the plug-in procedure yields an estimator $\hat{f}$ which is Ψ-consistent, implying that $regret_{\mathbb{P}}^{\Psi}(\hat{f}) \xrightarrow{p} 0$, where $\xrightarrow{p}$ denotes convergence in probability. We denote the estimators of $\eta, \overline{\eta}$ by $\hat{\eta}$ and $\hat{\overline{\eta}}$ respectively. In order to proceed we make the following assumptions:

Assumption 1 $\mathbb{P}_{X|Y=1}\left(\gamma(x) \leqslant c\right), \mathbb{P}_{X|Y=-1}\left(\gamma(x) \leqslant c\right),$

$\mathbb{P}_{X|Y=1,\overline{Y}=1}(\gamma(x) \leqslant c)$ and $\mathbb{P}_{X|Y=1,\overline{Y}=-1}(\gamma(x) \leqslant c)$ are continuous at c, where $\gamma(x) = (1 + \frac{\lambda \overline{c}}{\pi})\eta(x) - \frac{\lambda}{\pi}\overline{\eta}(x,1)\eta(x)$, i.e, $\gamma(x)$ is $s^*(x)$ in Theorem 2.2 without the constant term c

Assumption 2 *Class probability estimators (CPEs) $\hat{\eta}, \hat{\overline{\eta}}$ are L-1 consistent, i.e., $\mathbb{E}_X[|\eta(x) - \hat{\eta}(x)|] \xrightarrow{p} 0$; $\mathbb{E}_{X,Y}\left[|\overline{\eta}(x,y) - \hat{\overline{\eta}}(x,y)|\right] \xrightarrow{p} 0$*

Remark: As noted in (Narasimhan et al., 2014; Chzhen et al., 2019), Assumption 2 is not a very strong one, as an appropriately regularized ERM yields an L-1 consistent class probability estimator for proper losses (Menon et al., 2013; Agarwal, 2013).

Assumption 3 *Domain χ is compact and there exist constants $a, B \in \mathbb{R}_+$, such that the PDFs, $f_{X|Y=-1}, f_X, f_{X|Y=1}$ satisfy $\forall x \in \chi$, $0 < a \leqslant f_{X|Y=-1}(x), f_X(x), f_{X|Y=1}(x) \leqslant B$*

Remark: We make this assumption for technical convenience. This is akin to the 'strong density assumption' defined in (Audibert et al., 2007). This assumption is not necessary for the case when $\overline{Y}$ is available at test time, or for either case relating to the approximate Demographic Parity criterion

We now state our key lemma that facilitates the consistency result. Denoting the estimator derived via the plug-in procedure for $\gamma(x)$ by $\hat{\gamma}(x) = (1 + \frac{\lambda\overline{c}}{\pi})\hat{\eta}(x) - \frac{\lambda}{\pi}\hat{\overline{\eta}}(x,1)\hat{\eta}(x)$, we have:

Lemma 3.1 *Provided Assumptions 2 and 3 hold, $\hat{\gamma}$ is L-1 consistent, i.e., $\mathbb{E}_X[|\gamma(x) - \hat{\gamma}(x)|] \xrightarrow{p} 0$*

The validity of Lemma 3.1 allows us to leverage the proof template of (Narasimhan et al., 2014) which in turn proves the plug-in algorithm's consistency.

Theorem 3.2 *Provided Assumptions 1, 2 and 3 hold, the plug-in algorithm is Ψ-consistent, i.e., the algorithm yields $\hat{f} = sign \circ \{\hat{\gamma} - c\}$, s.t., $\mathfrak{P}^{\Psi}_{\mathbb{P}}(\hat{f}) \xrightarrow{p} \mathfrak{P}^{\Psi,*}_{\mathbb{P}}$, i.e., $regret^{\Psi}_{\mathbb{P}}(\hat{f}) \xrightarrow{p} 0$*

Proof sketch: Lemma 3.1 and Assumption 1 allow us to show that the plug-in yields an estimator $\hat{f}$ which is s.t.:

$$TPR_{\mathcal{D}}(\hat{f}) \xrightarrow{p} TPR_{\mathcal{D}}(f^*);\ TNR_{\mathcal{D}}(\hat{f}) \xrightarrow{p} TNR_{\mathcal{D}}(f^*)$$

$$TPR_{\overline{\mathcal{D}}}(\hat{f}) \xrightarrow{p} TPR_{\overline{\mathcal{D}}}(f^*);\ TNR_{\overline{\mathcal{D}}}(\hat{f}) \xrightarrow{p} TNR_{\overline{\mathcal{D}}}(f^*)$$

The result then follows by the Continuous Mapping Theorem (Mann & Wald, 1943), since Ψ is continuous in its arguments. Complete proofs and detailed discussion for the results presented in this section can be found in section B of the supplement.

3.2. Non-Asymptotic Analysis

In this section, our objective is to characterise the sample complexity requirements associated with learning a classifier that yields small regret, via the plug-in algorithm of (Menon & Williamson, 2018). In our problem formulation, the performance measure of a classifier, is a linear function of its true positive and true negative rates. This implies that the performance measure is non-decomposable, since it cannot be expressed as a summation/ expectation over individual instances. This is contrary to the case associated with most standard loss functions that feature in the ML literature, and thus the finite-sample analysis for our performance measure is non-standard. We provide a strategy that allows us to precisely relate the sample complexity of this task to the sample complexity associated with learning the regression functions, η and $\overline{\eta}$, as well as other distributional quantities. We defer the detailed derivation of this strategy to section C of the supplement. We assume in this section that, $\pi = \mathbb{P}(Y = 1)$ is known. While we can remove this assumption and modify our analysis to obtain equivalent results, we found doing so makes the underlying algebra/ geometry much more convoluted, without adding significant insight. Thus, for simplicity, we proceed by assuming π is known.

Recall, by Assumption 2, that we are working in a setting wherein the class probability estimators (CPEs), $\hat{\eta}, \hat{\overline{\eta}}$ are L-1 consistent. Convergence in the L-1 norm implies convergence in probability, so we can meaningfully define the sample complexity associated with learning the regression function η via the CPE $\hat{\eta}$:

Definition 3.4 *The sample complexity of learning η, is a mapping $m_\eta : (0,1)^3 \to \mathbb{N}$, where $m_\eta((\epsilon, \delta^{'}), \delta)$ is the minimal (integer) number of training samples required to ensure that, with probability $\geqslant (1-\delta)$: $\mathbb{P}_X(|\eta(x) - \hat{\eta}(x)| \geqslant \epsilon) \leqslant \delta^{'}$*

Note, we show in Lemma B.1 of the supplement that $\hat{\overline{\eta}}(\cdot, 1)$ is also L-1 consistent. The sample complexity of learning $\overline{\eta}(\cdot, 1)$ is thus analogously defined, and we denote this by $m_{\overline{\eta}}$. Our non-asymptotic result is derived via a geometric argument based in the plane of regression functions, i.e., the $(\overline{\eta}(\cdot,1), \eta)$-plane. We define some key objects pertaining to our derivation. Consider in the $(\overline{\eta}(\cdot,1), \eta)$-plane, the hyperbola $H(\lambda, \pi, c, \overline{c}) := \{(1 + \frac{\lambda\overline{c}}{\pi})\eta - \frac{\lambda}{\pi}\overline{\eta}(\cdot,1)\eta - c = 0\}$. Also, for $\epsilon \in (0, \frac{1}{2})$, let $X_M := \{x \in \chi :$ *the square of length* 2ϵ *centred at* $(\overline{\eta}(x,1), \eta(x))$ *intersects the hyperbola* $H(\lambda, \pi, c, \overline{c})$ *in the* $(\overline{\eta}(\cdot,1), \eta)$*-plane*$\}$. Having defined $H(\lambda, \pi, c, \overline{c})$ and X_M, we now state our non-asymptotic result:

Theorem 3.5 *Let $\delta, \delta^{'}, \epsilon \in (0, \frac{1}{2})$. Pick any $t > Q = 4G\{max\{c(1-\pi), (1-c)\pi, |\lambda|\overline{c}(1-\beta), |\lambda|(1-\overline{c})\beta\}\}$, where $G = max\{\frac{B}{\pi}, \frac{B}{1-\pi}, \frac{B}{\pi\beta}, \frac{B}{\pi(1-\beta)}\}$, and $B = \delta^{'} + \mathbb{P}_X(X_M)$. Provided access to $n \geqslant max\{m_\eta((\epsilon, \frac{\delta^{'}}{2}), \frac{\delta}{8}), m_{\overline{\eta}}((\epsilon, \frac{\delta^{'}}{2}), \frac{\delta}{8})\}$ training samples drawn i.i.d. from $\mathbb{P}$, the plug-in algorithm yields an estimator $\hat{f}$, such that, with probability at least $(1-\delta)$: $regret^{\psi}_{\mathbb{P}}(\hat{f}) \leqslant t$*

Proof sketch: Our proof entails showing that, for appropriate t, $\mathbb{P}_{S\sim\mathbb{P}^n}\left[regret_{\mathbb{P}}^{\Psi} > t\right] \leqslant \delta$ holds, so long as we can meaningfully upper bound $\mathbb{P}_X[sign \circ \{f^*(x)\} \neq sign \circ \{\hat{f}(x)\}]$ with probability $\geqslant (1 - \frac{\delta}{4})$. Denoting the upper bound by B, we characterise its form via a geometric argument. Roughly, $B \propto \mathbb{P}_X(X_M)$, where X_M is a specific region enclosing the hyperbola, $H(\lambda, \pi, c, \bar{c})$ in the $(\overline{\eta}(\cdot, 1), \eta)$-plane. Then setting t as described, the result follows. A visual simulation of the underlying geometry can be found in *Figure 1*.

Theorem 3.5 tells us that the regret can be made to decrease arbitrarily, provided a sufficient increase in the number of training samples. The precise rate of decay, depends on 1) the sample complexities associated with learning the regression functions and 2) the rate at which the probability measure , i.e., $\mathbb{P}_X$, decays around the hyperbola $H(\lambda, \pi, c, \bar{c})$ (in the $(\overline{\eta}(\cdot, 1), \eta)$-plane) upon shrinking the region of consideration around it (i.e., the region akin to the 'Projected X_M' region in *Figure 1*). Refer to section C of supplement for a detailed proof and discussion.

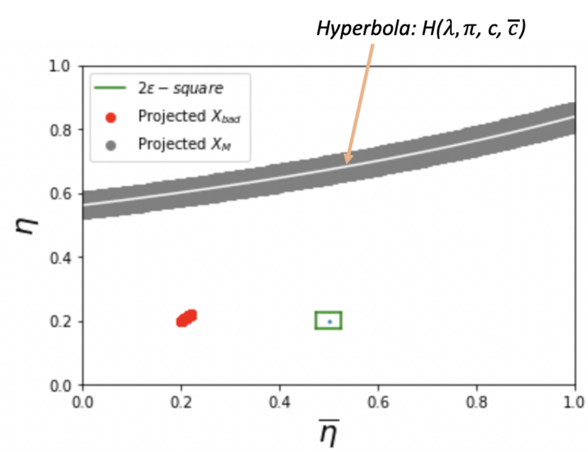

Figure 1. For any point $x \in \chi$: corresponding projected coordinates in the $(\overline{\eta}(\cdot, 1), \eta)$-plane, i.e., $(\overline{\eta}(x, 1), \eta(x))$ lie outside of {projected $X_M \bigcup$ projected X_{bad}}; we can be certain that $sign \circ \{f^*(x)\} = sign \circ \{\hat{f}(x)\}$ - the point centred within the green square of length 2ϵ is one such point.

4. Fairness under Differential Privacy

In this section, we work in a setting wherein there is an additional requirement for our modelling pipeline to mitigate information leakage about the sensitive attribute, $\overline{Y}$. To meet such a requirement we make use of a notion of privacy known as differential privacy (Dwork et al., 2006), which roughly speaking, ensures that an algorithm's output does not differ significantly on data sets that differ in only a single instance.

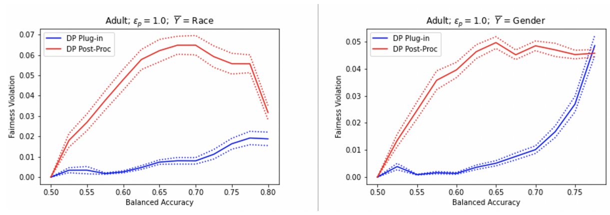

Figure 2. The solid blue curve represents model performance in the 'balanced accuracy-fairness violation plane' corresponding to our method, i.e., the DP Plug-in approach. Whereas the solid red curve, corresponds to model performance for the DP Post-Proc approach. The dotted curves represent the ± 0.2 standard-deviations in fairness violations corresponding to each segment of balanced accuracy considered. Privacy parameter $\epsilon_p = 1.0$

Algorithm 2 DP Plugin approach to FAL, EO setting

Input: Sample $S = \{x_i, y_i, \overline{y}_i\}_{i=1}^n$ from distribution $\mathbb{P}$; cost parameters $c, \bar{c}$; trade-off parameter λ; privacy parameter ϵ_p

Estimate: π via $\hat{\pi} = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{y_i = 1\}$
Estimate: η:$\chi \to [0, 1]$ using appropriate CPE on $\{x_i, y_i\}_{i=1}^n$
Estimate: $\overline{\eta}_{EO}$: $(\chi, Y) \to [0, 1]$ using appropriate CPE on S
Privatise: $\hat{\overline{\eta}}_{EO}$ via appropriate privacy preserving protocol yielding, ϵ_p-DP protected $\hat{\overline{\eta}}_{EO}^{priv}$
Compute: $\hat{s}^{priv}(x) = \left\{1 - \frac{\lambda}{\hat{\pi}}(\hat{\overline{\eta}}_{EO}^{priv}(x, 1) - \bar{c})\right\}\hat{\eta}(x) - c$

Return: $\hat{f}^{priv}(x) = H_\alpha\left(\hat{s}^{priv}(x)\right)$ for any $\alpha \in [0, 1]$

We detail the DP Plugin protocol in Algorithm 2. We compare DP Plug-in's performance against that of the private, fair, post-processing approach (DP Post-Proc) of (Jagielski et al., 2019). We found that our method outperforms the DP Post-Proc approach in three, out of four experimental set ups considered. We present our evaluations on the Adult data set (Dheeru & Taniskidou, 2017) in Figure 2. Complete details for the DP Plug-in algorithm, and for our experimental set up and methodology, can be found in section D of the supplement.

5. Conclusion and Future Work

Our main contributions in this paper included (1) proving the plug-in algorithm of (Menon & Williamson, 2018) is consistent, (2) characterising the sample complexity of learning fairness-aware BOCs via the plug-in algorithm, and (3) proposing an easy to deploy, privacy-preserving protocol for the plug-in algorithm. As future directions, we believe it would be valuable to extend our analysis to the case where the sensitive attribute in non-binary; the case where multiple attributes are deemed sensitive. It would also be useful to study the statistical properties of learning algorithms across other settings, such as those demanding individual fairness, model explainability, or intersections between such areas of ethical and practical importance.

References

Agarwal, S. Surrogate regret bounds for the area under the roc curve via strongly proper losses. In *Conference on Learning Theory*, pp. 338–353. PMLR, 2013.

Audibert, J.-Y., Tsybakov, A. B., et al. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2): 608–633, 2007.

Chen, I., Johansson, F. D., and Sontag, D. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002*, 2018.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Leveraging labeled and unlabeled data for consistent fair binary classification. *arXiv preprint arXiv:1906.05082*, 2019.

Cummings, R., Gupta, V., Kimpara, D., and Morgenstern, J. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pp. 309–315, 2019.

Denis, C. and Hebiri, M. Confidence sets with expected sizes for multiclass classification. *The Journal of Machine Learning Research*, 18(1):3571–3598, 2017.

Devroye, L., Gyorfi, L., and Lugosi, G. A probabilistic theory of pattern recognition. Technical report, 1996.

Dheeru, D. and Taniskidou, E. K. Uci machine learning repository. 2017.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., and Ullman, J. Differentially private fair learning. In *International Conference on Machine Learning*, pp. 3000–3008. PMLR, 2019.

Mann, H. B. and Wald, A. On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, 14 (3):217–226, 1943.

Menon, A., Narasimhan, H., Agarwal, S., and Chawla, S. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, pp. 603–611. PMLR, 2013.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pp. 107–118. PMLR, 2018.

Mozannar, H., Ohannessian, M., and Srebro, N. Fair learning with private demographic data. In *International Conference on Machine Learning*, pp. 7066–7075. PMLR, 2020.

Narasimhan, H., Vaish, R., and Agarwal, S. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In *NIPS*, volume 27, pp. 1493–1501, 2014.

Yang, Y. Minimax nonparametric classification. i. rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, 1999.

Zhao, H. and Gordon, G. J. Inherent tradeoffs in learning fair representations. *arXiv preprint arXiv:1906.08386*, 2019.