Classification with Abstention but Without Disparities

Nicolas Schreuder[1] **Evgenii Chzhen**[2]

[1]CREST, ENSAE Paris, Palaiseau, France
[2]LMO, Université Paris-Saclay, CNRS, Inria, Orsay, France

Abstract

Classification with abstention has gained a lot of attention in recent years as it allows to incorporate human decision-makers in the process. Yet, abstention can potentially amplify disparities and lead to discriminatory predictions. The goal of this work is to build a general purpose classification algorithm, which is able to abstain from prediction, while avoiding disparate impact. We formalize this problem as risk minimization under fairness and abstention constraints for which we derive the form of the optimal classifier. Building on this result, we propose a post-processing classification algorithm, which is able to modify any off-the-shelf score-based classifier using only unlabeled sample. We establish finite sample risk, fairness, and abstention guarantees for the proposed algorithm. In particular, it is shown that fairness and abstention constraints can be achieved independently from the initial classifier as long as sufficiently many unlabeled data is available. The risk guarantee is established in terms of the quality of the initial classifier. Our post-processing scheme reduces to a sparse linear program allowing for an efficient implementation, which we provide. Finally, we validate our method empirically showing that moderate abstention rates allow to bypass the risk-fairness trade-off.

1 INTRODUCTION

In recent years classification with abstention or with reject option has gained a considerable amount of attention from both statistical and machine learning communities. Probably the earliest appearance of classification with reject option can be found in the works of Chow [1957, 1970] in the context of information retrieval and an initial statistical treatment was given in [Györfi et al., 1979]. Much later, Herbei and Wegkamp [2006] provided non-parametric analysis for the problem of binary classification with a fixed rejection cost in the spirit of Audibert et al. [2007]. Several extensions followed later, all working with fixed cost of rejection [Yuan and Wegkamp, 2010, Wegkamp and Yuan, 2011, Bartlett and Wegkamp, 2008].

Following the conformal prediction literature [see, *e.g.,* Vovk et al., 2005], Lei [2014] considers a framework where ones wants to minimize the reject rate under a pre-specified accuracy constraint, meanwhile Denis and Hebiri [2020] target its reversed formulation. Both derive finite sample guarantees for plug-in type classification procedures and instanciate their analysis to standard non-parametric class of distributions. In a similar direction, several practical methods [Grandvalet et al., 2008, Nadeem et al., 2009] have been proposed in the machine learning community to address the problem of classification with abstention. Recently, Bousquet and Zhivotovskiy [2019], Neu and Zhivotovskiy [2020], Puchkin and Zhivotovskiy [2021] show that abstention can significantly improve regret bounds and convergence rates for the problems of online and batch classification.

Crucially, in our work we view abstention as a mechanism to lighten the burden of fairness constraints and bypass the risk-fairness trade-off [Agarwal et al., 2018, Menon and Williamson, 2018, Chzhen and Schreuder, 2020]: one can enjoy the best of both worlds – a simultaneously fair and accurate classifier – at the cost of rejection. A majority of observations are still classified in an automatic manner, while the rejected ones can be handled by, *e.g.,* human experts. Importantly, in our setting, the rejection rate is rigorously controlled by the practitioner depending on the number of available experts. In addition, since it is illusory to assume that a data-dependent classifier can make error-less and trustworthy decisions, it is desirable to put human experts back in the loop for sensitive tasks. The rejection mechanism partially transfers the burden of optimizing those conflicting quantities to human experts, who can eventually have access to more information to make a better informed decision (*e.g.,* a doctor can ask for extra medical examination for its

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:1227–1236.

final diagnosis).

Fairness in binary classification is a very popular topic with various types of algorithmic and statistical contributions [see, *e.g.,* Hardt et al., 2016, Barocas et al., 2019]. However, abstention framework has not yet received a lot of attention in the context of fair learning. Notable exceptions are work of Madras et al. [2018], Jones et al. [2020]. The latter demonstrates that an imprudent use of abstention might amplify potential disparities already present in the data. In particular, they show that in the framework of prediction without disparate treatment [Zafar et al., 2017] the use of the same rejection threshold across sensitive groups might result in a large group-wise risks disparities. As a potential remedy, our work offers a theoretically grounded way to enforce fairness constraints as well as a desired group-dependent reject rates. The idea of relying on a reject mechanism to enforce fairness has only been explored once, in Madras et al. [2018]. The authors introduce "learning to defer" framework – an extension of classification with abstention – where the cost of rejection is allowed to depend on the prediction of an external decision-maker (*e.g.,* a human expert). The authors argue that by making the automated model aware of the potential biases and weaknesses of the external decision-maker, it can globally optimize for accuracy and fairness. The authors enforce Equalized Odds [Hardt et al., 2016] through regularization of the risk and thus cannot control explicitly the reject rate, which might potentially lead to a huge external decision-maker costs. While the authors provide empirical evidences of their claims, theoretical justification of their results remains open. Our work offers a completely theory-driven way to enforce both fairness and rejection constraints while optimizing for accuracy, leading to a computationally efficient post-processing algorithm.

Contributions. Our work combines and extends previous results in abstention framework with recent results on fair binary classification. Namely, similarly to [Denis and Hebiri, 2020], we aim at minimizing misclassification risk under a control over *group-wise* reject rates. As we would like to avoid disparate impact, we explicitly add this as a constraint to our framework. We derive the optimal form of a reject classifier, which minimizes the misclassification risk under the discussed constraints. Our explicit characterization of the optimal reject classifier provides a better understating of the interplay between, on one side, the fairness and rejection constraints and, on the other side, the accuracy. We propose a data-driven post-processing algorithm which enjoys generic plug-and-play finite sample guarantees. An appealing feature of our post-processing algorithm is that it can be used on top of *any* pre-trained classifier, thus avoiding the – potentially high – cost of re-fitting a classifier from scratch. From numerical perspective, the proposed method reduces to a solution of a sparse linear program, allowing us to leverage efficient LP solvers. Numerical experiments validate our theoretical result demonstrating that the proposed method successfully enforces fairness and rejection constraints in practice, while achieving a high level of accuracy.

Notation. For each $K \in \mathbb{N}$ we denote by $[K]$ the set of the first K positive integers. The standard Euclidean inner product is denoted by $\langle \cdot, \cdot \rangle$. For a real number $a \in \mathbb{R}$ we write $(b)_+$ (*resp.* $(a)_-$) to denote the positive (*resp.* the negative) part of a. For two real numbers a, b we denote by $a \vee b$ (*resp.* $a \wedge b$) the maximum (*resp.* the minimum) between the two. We denote by $1 \in \mathbb{R}^K$ the vector composed of ones and by $e_s \in \mathbb{R}^K$ the s^{th} basis vector of $\mathbb{R}^K$.

2 PROBLEM PRESENTATION

Consider a triplet $(X, S, Y) \sim \mathbb{P}$, where $X \in \mathbb{R}^d$ is the feature vector, $S \in [K]$ is the sensitive attribute, and $Y \in \{0, 1\}$ is the binary label to be predicted. A classifier is a mapping $g : \mathbb{R}^d \times [K] \to \{0, 1, r\}$. That is, any classifier g is able to provide a prediction in $\{0, 1\}$, or to abstain from prediction by outputting r. With any classifier g, we associate the following quantities:

$$
\begin{aligned}
&R(g) := \mathbb{P}(Y \neq g(X,S) \mid g(X,S) \neq r) \ , \\
&\mathrm{NAb}_s(g) := \mathbb{P}(g(X,S) \neq r \mid S = s) \ , \\
&\mathrm{NAb}(g) := \mathbb{P}(g(X,S) \neq r) \ , \\
&\mathrm{PT}_s(g) := \mathbb{P}(g(X,S) = 1 \mid S = s, g(X,S) \neq r) \ , \\
&\mathrm{PT}(g) := \mathbb{P}(g(X,S) = 1 \mid g(X,S) \neq r)) \ .
\end{aligned} \tag{1}
$$

The first one is the risk of a classifier, which measures the probability of incorrect prediction, given that an actual prediction was issued. The second two quantities measure the group-wise and marginal prediction rates. The last two quantities describe the group-wise and marginal rates of positive predictions given that the prediction was made. Intuitively, a good classifier has low risk R, high NAb_s, and low disparities between $\mathrm{PT}_s(g)$.

Fairness constraint. We formalize fairness through the notion of Demographic Parity [see for instance, Barocas et al., 2019]. A predictor g is said to satisfy Demographic Parity (or, equivalently, to avoid Disparate Impact) if the distribution of its prediction is independent from the sensitive attribute. Formally, in the standard binary classification framework it means that for any $z \in \{0, 1\}$ and for any $s, s' \in [K]$,

$$
\mathbb{P}(g(X,S) = z \mid S = s) = \mathbb{P}(g(X,S) = z \mid S = s') \ .
$$

In the setting of classification with abstention, we naturally want to condition on the fact that the classifier issues a prediction, that is, $g(X,S) \neq r$. Using the quantities introduced in Eq. (1), the latter reduces to

$$
\forall s \in [K], \quad \mathrm{PT}_s(g) = \mathrm{PT}(g) \ .
$$

Penalized version. There are various trade-offs that one can consider between the quantities in Eq. (1). For instance, adapting the approach of Herbei and Wegkamp [2006] to the context of fairness, one can target a prediction which avoids disparate impact and minimizes penalized risk. Formally, it amounts to solving the following problem:

$$\min_{g:\mathbb{R}^d\times[K]\to\{0,1,r\}} R(g) + \sum_{s=1}^{K} \lambda_s \mathrm{NAb}_s(g) \quad , \qquad \textbf{(P-DPWA)}$$
$$\text{s.t. } \forall s \in [K], \quad \mathrm{PT}_s(g) = \mathrm{PT}(g)$$

for some $\lambda_s \geq 0$, $s \in [K]$. This approach also resembles the one employed by Madras et al. [2018], who additionally penalized for fairness violation instead of directly controlling it. The main issue with the formulation (**P-DPWA**) is connected with the choice of the penalization parameters $\lambda_s \geq 0$, $s \in [K]$, which do not have simple and intuitive interpretation. Indeed, it is impossible to know beforehand which $\lambda_s \geq 0$, $s \in [K]$ will result in a usable reject rate, forcing the practitioner to explore the whole space of the hyperparameters $\lambda_s \geq 0$, $s \in [K]$. Instead of the above formulation, we consider the problem in which one is able to *explicitly* control the rejection rate. In particular, such an approach allows us to develop a *parameter-free* post-processing method.

Explicit control of reject. Given $\alpha = (\alpha_1, \ldots, \alpha_K)^\top \in [0,1]^K$, our goal is to find a solution of the following problem

$$\min_{g:\mathbb{R}^d\times[K]\to\{0,1,r\}} R(g)$$
$$\text{s.t. } , \forall s \in [K], \begin{cases} \mathrm{NAb}_s(g) = \alpha_s \\ \mathrm{PT}_s(g) = \mathrm{PT}(g) \end{cases} \quad . \qquad \textbf{(DPWA)}$$

It will be shown later that, under a mild assumption on the distribution of the conditional expectation $\mathbb{E}[Y \mid X, S]$, the above problem admits a global minimizer written in the form of group-wise thresholding.

The first constraint in (**DPWA**) specifies the abstention level accepted for each class while the second constraint, as before, demands the classifier g to avoid disparate impact. Notably, in this formulation, the parameter vector $\alpha \in [0,1]^K$ has a simple and intuitive interpretation – it allows to fix precisely *different* levels of rejects for different groups. This, for instance, can be beneficial, if $g(x,s) = r$ is followed by the intervention of a human decision-maker, who replaces the classifier. One can force a higher rejection rate (*i.e.,* a higher rate of human intervention) for disadvantaged groups by lowering the corresponding $\alpha_s \in [0,1]$. Crucially, we implicitly assume that the practitioner is able to treat unclassified instances in an accurate and fair manner. While this assumption is void for the theoretical contributions of our paper, we *warn* the practitioner that it must not be overlooked once our method is deployed in real world.

This formulation allows to bypass the usual trade-off between fairness and accuracy at the price of rejection. Indeed, note that a classifier that solves (**DPWA**) is fair for any parameters $(\alpha_s)_{s\in[K]}$. At the same time, setting $\alpha_1 = \ldots = \alpha_K = \tilde{\alpha}$ for some $\tilde{\alpha} \in (0,1]$, one can observe that by varying $\tilde{\alpha}$ we can recover the accuracy of a classifier without constraints while still satisfying Demographic Parity. This will be later empirically confirmed in Section 7. We again emphasize that the accuracy gain comes at a price of a possible reject region, which, depending on the application at hand might or might not constitute a reasonable price.

3 OPTIMAL CLASSIFIER

Our first theoretical contribution is the derivation of a classification strategy g^*, which is a solution of (**DPWA**). We define the conditional expectation of the label Y knowing (X,S) as

$$\eta(X,S) = \mathbb{E}[Y \mid X, S] \quad .$$

It is known that the Bayes optimal rule for the problem of binary classification with misclassification risk is given by the point-wise thresholding of $\eta(X,S)$ on the level $1/2$ [Devroye et al., 2013]. In our case the classifier does not correspond to the Bayes decision. Instead, it is a solution of a constrained optimization problem with constraints that depend on the unknown data distribution $\mathbb{P}$. In several frameworks, which are also formulated as risk minimization under distribution dependent constraints, it is possible to obtain a closed form expression of a minimizer under fairly mild assumptions. In particular, it is the case for the classification with reject option [Chow, 1970, Lei, 2014, Denis and Hebiri, 2020] as well as classification under various fairness constraints [Hardt et al., 2016, Chzhen et al., 2019, del Barrio et al., 2020]. Similarly to the above contributions, we will make a mild assumption on the behaviour of $\eta(X,S)$, which is, for instance, naturally satisfied whenever $\eta(X,S)$ admits a density *w.r.t.* the Lebesgue measure.

Assumption 3.1. *The random variables* $(\eta(X,S) \mid S = s)$ *are non-atomic for all* $s \in [K]$.

One can actually get rid of this assumption, as explained in Lei [2014], by switching from deterministic classification strategies, which are valued in $\{0,1,r\}$, to randomized classifiers, which output a distribution over $\{0,1,r\}$.

To present the main result of this section, we introduce the notations $p_s := \mathbb{P}(S = s)$, $\bar{\alpha} := \sum_{s\in[K]} p_s \alpha_s$ and we define the following function

$$G(x,s,\lambda,\gamma) = \left| \frac{p_s}{2\bar{\alpha}}(1 - 2\eta(x,s) - \langle \gamma, 1\rangle) + \frac{\langle \gamma, e_s\rangle}{2\alpha_s} \right| - \frac{p_s}{2\bar{\alpha}}(1 - \langle \gamma, 1\rangle) - \langle \lambda, e_s\rangle - \frac{\langle \gamma, e_s\rangle}{2\alpha_s} \quad ,$$

which plays a key role in the derivation of an optimal classifier for the problem (**DPWA**). We now state the first result of this work, which provides a form of g^* – solution for (**DPWA**).

Theorem 3.2. *Under Assumption 3.1, an optimal classifier for the problem* (**DPWA**) *is given for all* $(x,s) \in \mathbb{R}^d \times [K]$ *by*

$$g^*(x,s) = \begin{cases} r & \text{if } G(x,s,\lambda^*,\gamma^*) \leq 0 \\ \mathbb{1}\left(\eta(x,s) \geq \frac{1}{2} + c_{\gamma^*,s}\right) & \text{otherwise} \end{cases},$$

where (λ^*,γ^*) *are solutions of*

$$\min_{(\lambda,\gamma)} \left\{ \langle \lambda, \alpha \rangle + \sum_{s=1}^{K} \mathbb{E}_{X|S=s}[(G(X,S,\lambda,\gamma))_+] \right\} ,$$

and $c_{\gamma^*,s} := \frac{1}{2}\left(\frac{\bar{\alpha}\gamma_s^*}{\alpha_s p_s} - \langle 1, \gamma^* \rangle\right)$.

Let us mention that unlike other similar results described above, the main difficulty in the proof of Theorem 3.2 lies in the fact the misclassification risk in our case involves conditioning on the event which itself depends on the classifier that we want to find. Theorem 3.2 is instructive and allows to develop an intuition which is similar to that of the original rule derived by Chow [1957, 1970]. To be more precise, denoting by

$$t_{\gamma^*,s} := (1 - \langle \gamma, 1 \rangle) + \frac{\bar{\alpha}\,\langle \gamma, e_s \rangle}{p_s \alpha_s} ,$$

the reject region is expressed as a strip around $t_{\gamma,s}$:

$$|\eta(x,s) - t_{\gamma,s}| \leq t_{\gamma,s} + \frac{\bar{\alpha}\lambda_s}{p_s} .$$

We highlight that the center as well as the size of this strip is group-dependent. Interestingly, the position of the strip only depends on the Lagrange multiplier controlling for the fairness constraint, while its width is determined by both constraints.

4 EMPIRICAL METHOD

The form of the optimal classifier suggests to develop a post-processing algorithm, which receives an estimator $\hat{\eta}(x,s)$ of $\eta(x,s)$ and an additional *unlabeled* set of samples to estimate (γ^*,λ^*). Indeed, observe that the optimal classifier g^* is know up to the quantities $\eta(x,s), \gamma^*, \lambda^*$.

Remark 4.1. *For simplicity of exposition we assume that the marginal distribution of S is known, that is, we have access to* $p_s := \mathbb{P}(S=s)$. *Note that S follows multinomial distribution, and, in practice, we can estimate these probabilities by their empirical counterparts, which is the direction that we take in our experimental section. Our proofs generalize straightforwardly for the case of unknown* p_s, *but such modification results in additional, unnecessary, complications.*

We denote by $\hat{\eta}(X,S)$ any off-the-shelf estimator of $\eta(X,S)$. For instance, one can take k-NN [Stone, 1977, Devroye et al., 2013], locally polynomial estimator [Korostelev and Tsybakov, 2012], logistic regression [Bühlmann and Van de Geer, 2011], random forest [Breiman, 2001, Biau and Scornet, 2016, Mourtada et al., 2020] to name a few. Our theoretical guarantees on the misclassification risk will explicitly depend on the quality of this off-the-shelf estimator, hence it is advisable to use those methods which are supported by statistical guarantees. Yet, our algorithm remains valid even for inconsistent estimators $\hat{\eta}$ in the sense that the resulting classifier after post-processing will (nearly) satisfy the prescribed constraints independently from $\hat{\eta}$.

Remark 4.2. *In what follows we assume that the estimator* $\hat{\eta}(X,S)$ *is independent from the* unlabeled *sample (introduced below) and is valued in* $[0,1]$. *In other words, we require a new* unseen *unlabeled sample for the post-processing. As it will be seen from our bound, the assumption that* $\hat{\eta}(X,S)$ *is valued in* $[0,1]$ *is not restrictive, since we can always perform clipping without damaging statistical properties. On a more technical note, we require that* $\mathbb{P}(\hat{\eta}(X,S) = c \mid \hat{\eta}) = 0$ *almost surely for any* $c \in [0,1]$. *Again, this assumption is not restrictive, since we can always randomize the output of* $\hat{\eta}(X,S)$ *by adding a negligible noise coming from a continuous distribution. In Algorithm 1 we use uniformly distributed noise supported on* $[0,\sigma]$, *with* σ *being a small parameter. One can take this parameter* σ *arbitrarily small, preserving the statistical properties of* $\hat{\eta}$.

As mentioned before, to build the post-processing scheme, we will use only *unlabeled* sample. We also do not restrict ourselves to sampling from $\mathbb{P}_{(X,S)}$. Instead, we assume that for all $s \in [K]$ we observe $\{X_i\}_{i \in \mathscr{I}_s}$ sampled i.i.d. from $\mathbb{P}_{X|S=s}$. In the above notation, $\mathscr{I}_s$ have cardinality n_s and they form a partition of $[n]$. That is, we have that $n_1 + \ldots + n_K = n$. The described sampling scheme is potentially appealing in situations when it is possible to gather a lot of data about the minority group without the need of labeling them. In particular, this sampling scheme allows to set $n_1 = \ldots = n_K$, which, since we do not require labeling, is more realistic. The conditional expectation $\mathbb{E}_{X|S=s}$ is estimated based on the following empirical measure

$$\hat{\mathbb{P}}_{X|S=s} = \frac{1}{n_s} \sum_{i \in \mathscr{I}_s} \delta_{X_i} .$$

Before providing the proposed post-processing method, we define the empirical counterpart to the function G as

$$\hat{G}(x,s,\lambda,\gamma) = \left| \frac{p_s}{2\bar{\alpha}}(1 - 2\hat{\eta}(x,s) - \langle \gamma, 1 \rangle) + \frac{\langle \gamma, e_s \rangle}{2\alpha_s} \right| - \frac{p_s}{2\bar{\alpha}}(1 - \langle \gamma, 1 \rangle) - \langle \lambda, e_s \rangle - \frac{\langle \gamma, e_s \rangle}{2\alpha_s}$$

The post-processing classifier with abstention is given by

$$\hat{g}(x,s) = \begin{cases} r & \text{if } \hat{G}(x,s,\hat{\lambda},\hat{\gamma}) \leq 0 \\ \mathbb{1}\left(\hat{\eta}(x,s) > \frac{1}{2} + c_{\hat{\gamma},s}\right) & \text{otherwise} \end{cases}, \quad (2)$$

Algorithm 1: `Post-processing`

Input: base estimator $\hat{\eta}$, unlabeled data $\{X_i\}_{i\in\mathscr{I}_s}$ for $s\in[K]$, noise magnitude σ
`Randomize`:
for $i\in\mathscr{I}_s, s\in[K]$ **do**
  Sample independently $\zeta_i\sim\mathscr{U}([0,\sigma])$
  Set $\hat{\eta}(X_i,s)\leftarrow\hat{\eta}(X_i,s)+\zeta_i$
`Solve`: Eq. (3) based on LP formulation to get $(\hat{\lambda},\hat{\gamma})$
Output: $(\hat{\lambda},\hat{\gamma})$

where $c_{\hat{\gamma},s}:=\frac{1}{2}\left(\frac{\tilde{\alpha}\hat{\gamma}_s}{\alpha_s p_s}-\langle 1,\hat{\gamma}\rangle\right)$ and $(\hat{\lambda},\hat{\gamma})$ is a solution of

$$\min_{(\lambda,\gamma)}\left\{\langle\lambda,\alpha\rangle+\sum_{s=1}^{K}\hat{\mathbb{E}}_{X|S=s}(\hat{G}(X,s,\lambda,\gamma))_+\right\}\ . \tag{3}$$

We summarize the proposed procedure in Algorithm 1 incorporating the randomization step. Note that there is a clear analogy between the result of Theorem 3.2 and the constructed algorithm. Indeed, the latter is an empirical version of the former built via the plug-in approach.

Lemma 4.3. *The minimization problem in Eq. (3) is convex and it admits a global minimizer.*

In Section 6 we will actually prove a stronger statement. Namely, it will be shown that the minimization problem in Eq. (3) is equivalent to a linear program with sparse constraints, which will allow us to provide an efficient implementation of the proposed procedure.

5 FINITE SAMPLE GUARANTEES

In this section we provide finite sample guarantees on the behavior of the post-processing classifier with abstention regarding its performance, its reject rate and its fairness. In order to lighten the presentation of our results, let us define now the sequence

$$u_n^{\delta,K}:=\sqrt{\frac{2\log(4K/\delta)}{2n}}+\frac{2}{n}\ ,\quad\forall n\geq 1\ .$$

The sequence $u_n^{\delta,K}$ behaves as $O(\sqrt{\log(K/\delta)/n})$, that is, it depends logarithmically on the number of sensitive attributes K, on the confidence parameter δ and goes to zero as $n^{-1/2}$ with the growth of n. Our goal in this section is to derive constraint and risk guarantees. Namely, we would like to show that when $n_s\to\infty$ we have for all $s\in[K]$ that

$$\begin{aligned}&|\mathrm{NAb}_s(\hat{g})-\alpha_s|\to 0\\&|\mathrm{PT}_s(\hat{g})-\mathrm{PT}(\hat{g})|\to 0\end{aligned}\quad\text{and}\quad\mathscr{E}(\hat{g}):=R(\hat{g})-R(g^*)\to 0\ .$$

The first part ensures satisfaction of reject and fairness constraints, while the second part shows that the risk of the proposed method is similar to that of g^*. Importantly, both guarantees will be derived in the finite-sample regime and with high probability.

The next proposition provides a quantitative control on the violation of the reject and demographic parity constraints in the finite sample regime.

Proposition 5.1. *Let $\delta\in(0,1)$. The violation of the constraints by the post-processing classifier with abstention $\hat{g}$ defined in Eq. (2) can be controlled, with probability at least $1-\delta$, for any $s\in[K]$, as*

$$\begin{aligned}&|\mathrm{NAb}_s(\hat{g})-\alpha_s|\leq u_{n_s}^{\delta/2,K},\\&|\mathrm{PT}_s(\hat{g})-\mathrm{PT}(\hat{g})|\leq\frac{6}{\alpha_s}u_{n_s}^{\delta,K}+\frac{6}{\tilde{\alpha}}\sum_{s=1}^{K}p_s u_{n_s}^{\delta,K}\ .\end{aligned}$$

The proofs for the control of the reject rate for the control of the demographic parity constraint are postponed to the supplementary material.

Remarkably Proposition 5.1 is assumption-free. In particular it does not depend on the conditional expectation η as well as it does not depend on the initial estimator $\hat{\eta}$. If one has enough *unlabeled* data than one can get arbitrarily close to exact satisfaction of the constraints. Intuitively, this is the case because the fairness and reject constraints only depend on the conditional distribution of the feature vector X given the sensitive attribute S, not on the relation between the features and the label Y.

We also remark that both bounds of Proposition 5.1 depend on the amount of observation available for each group $s\in[K]$ – it is easier to satisfy constraints for well-represented groups. In particular, it is advisable to collect an unlabeled sample which is balanced in terms of the sensitive attributes. Note that it is explicitly allowed in our framework, since we require samples from $\mathbb{P}_{X|S=s}$ and not from $\mathbb{P}_{(X,S)}$.

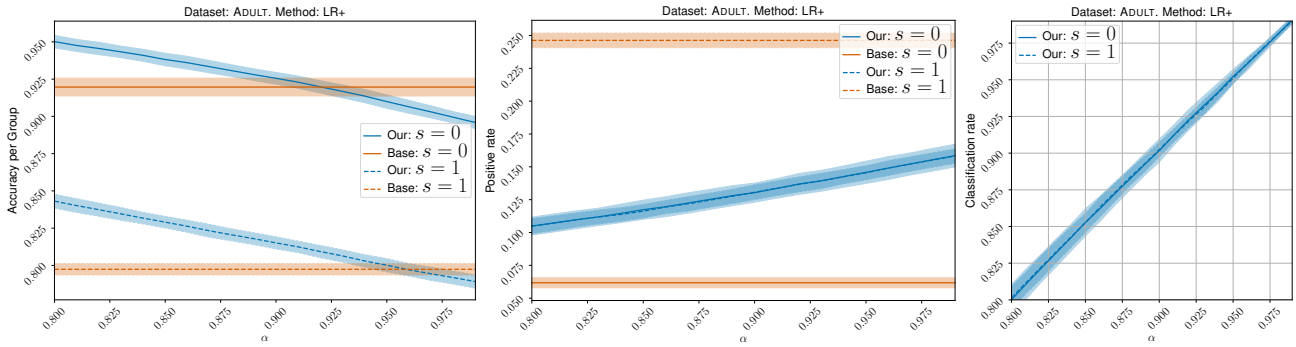
The next result establishes excess risk guarantees for the proposed method.

Proposition 5.2. *Assume that $2u_{n_s}^{\delta,K}<\alpha_s<1-2/n_s$ for any $s\in[K]$ and that Assumption 3.1 holds. Then, for any $\delta\in(0,1)$, the excess risk of the post-processing classifier with abstention $\hat{g}$ defined in Eq. (2) satisfies with probability at least $1-\delta$,*

$$\mathscr{E}(\hat{g})\leq\frac{3}{\tilde{\alpha}}\|\eta-\hat{\eta}\|_1+6\sum_{s=1}^{K}\left(\frac{p_s}{\tilde{\alpha}}+\frac{1}{\alpha_s}\right)u_{n_s}^{\delta,K}\ . \tag{4}$$

For convenience and clarity of exposition we stated separately the control on the constraint and on the excess risk. However, we remark that both Proposition 5.1 and Proposition 5.2 hold on the same high-probability event.

We naturally conclude from Proposition 5.2 that if one has access to a consistent estimator $\hat{\eta}$ of η, *i.e.,* such that



(a) Accuracy per group (b) Positive rate (c) Classification rate

Figure 1: Results on ADULT dataset with Logistic Regression (LR) as the base estimator. Blue lines correspond to our post-processing method; Orange lines correspond to the base classifier. Dashed line correspond to $s = 1$ and solid line to $s = 0$. Shaded areas correspond to the variance of the result over 20 repetitions.

$\|\eta - \hat{\eta}\|_1$ goes to 0 as the sample sizes $(n_s)_{s=1}^K$ go to infinity, then the excess risk can be made arbitrarily small by getting more labeled and unlabeled data.

The only assumption, constraining the reject rates $(\alpha_s)_{s=1}^K$, is quite benign. Recall that α_s is the rate at which the classifier is asked to give a prediction thus, in practice, it is expected to be at least greater than a half. Furthermore, note that it only depends on the size of the unlabeled dataset thus, if one has enough samples, this assumption essentially holds for free. If the sample size is small, than one has to allow the classifier to reject more often in order to satisfy the constraints. Similar constraints are present in other contributions [see *e.g.,* Agarwal et al., 2018, 2019].

Our theoretical analysis is inspired by that of Chzhen et al. [2020]. However, their results hold only in expectation while ours hold with high-probability. Moreover, due to the interplay of the reject and demographic parity constraints, their proof technique requires a non-trivial adaptation to our context.

6 LP REDUCTION

We recall that the proposed post-processing scheme involves solving convex non-smooth minimization problem in Eq. (3). While for low values of K (few sensitive attributes) this problem can be solved via simple grid-search, which would be faster than sub-gradient methods, large values of K can pose significant computational difficulties.

It turns out that the minimization problem in Eq. (3) is equivalent to Linear Programming (LP) [Matousek and Gärtner, 2007] with sparse constraint matrix. For any matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ we denote by $\mathrm{nnz}(\mathbf{A})$ the number of non-zero elements of $\mathbf{A}$.

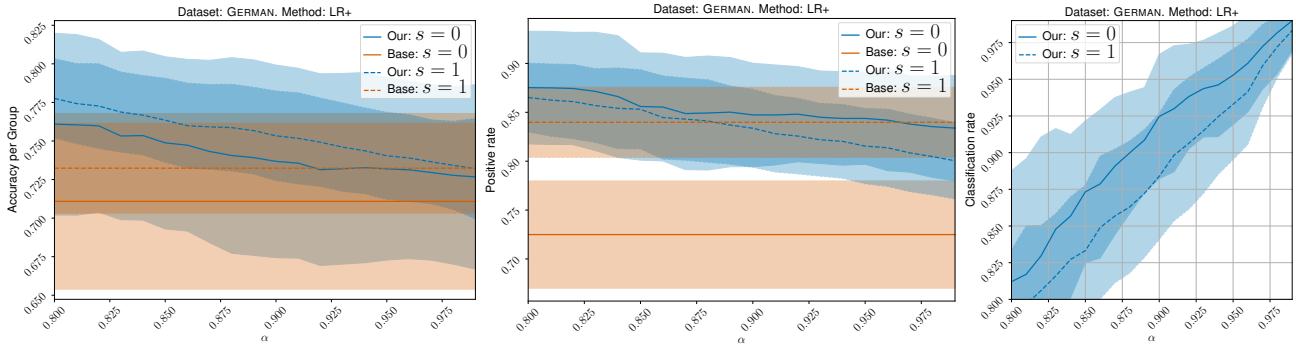
Proposition 6.1. *There exist* $c \in \mathbb{R}^{n+2K}$, $b \in \mathbb{R}^{2n}$, $\mathbf{A} \in \mathbb{R}^{2n\times(n+2K)}$ *with* $\mathrm{nnz}(\mathbf{A}) \leq 4n + nK$, *such that the minimization problem in Eq. (3), is equivalent to*

$$\min_{y\in\mathbb{R}^{n+2K}} \langle c, y\rangle \quad \text{s.t.} \quad \begin{cases} \mathbf{A}y \leq b \\ y_i \geq 0 \quad i \in [n] \end{cases} . \tag{LP}$$

Due to the space considerations, the previous result is stated in existential form, however, all the parameters of the LP are explicit and are provided in the supplementary material. Seminal works of [Khachiyan, 1979, Karmarkar, 1984] confirmed that LP with rational coefficients can be solved in weakly polynomial time. Since then, extremely efficient solvers were developed based on the interior-point and simplex methods. The fact that the post-processing reduces to an LP problem allows us to use these fast solvers. In particular, most of the computational burden lies on the training of the base estimator $\hat{\eta}$ while the post-processing can be performed almost instantly. From theoretical perspective, one can leverage the sparse structure of the problem using, for instance, the result of [Lee and Sidford, 2015] who provide an efficient solver to find an ε solution of an LP in $\tilde{O}((\mathrm{nnz}(\mathbf{A}) + n^2)\sqrt{n}\log(\varepsilon^{-1}))$ time. In particular, the previous guarantee scales only linearly with the number of sensitive attributes and logarithmically with the precision ε. However, in our practical implementation of the proposed method, we use interior point method available as a part of `scipy.optimize.linprog` [Virtanen et al., 2020].

7 EXPERIMENTS

We provide an implementation of the proposed post-processing procedure described in Algorithm 1 using `scipy.optimize.linprog` [Virtanen et al., 2020], which implements interior point method for solving problem (**LP**). The source code is available at `https://github.com/evgchz/dpabst`.



(a) Accuracy per group (b) Positive rate (c) Classification rate

Figure 2: Results on GERMAN dataset with Logistic Regression (LR) as the base estimator. Blue lines correspond to our post-processing method; Orange lines correspond to the base classifier. Dashed line correspond to $s = 1$ and solid line to $s = 0$. Shaded areas correspond to the variance of the result over 20 repetitions.

We consider ADULT [Kohavi, 1996] and GERMAN [Dua and Graff, 2017] datasets, which are standard benchmark datasets in the fairness literature.

ADULT dataset is fetched via `fairlearn.datasets` [Bird et al., 2020]. This dataset contains 14 features and around $48,000$ observations. We dropped those observations that contain missing values. This dataset consists of 1994 US Census entries. Each entry of this dataset corresponds to an individual who is described by 14 characteristics, the binary target variable is equal to 1 if the individual earns more than $\$50K$ per year and it is set to 0 otherwise. In our experiments we take sex as a sensitive attribute.

GERMAN dataset is hosted on the UCI Machine Learning Repository [Dua and Graff, 2017]. Each of the $1,000$ entries represents a person who takes a credit by a bank. The binary target variable is equal to one if the individual is considered as good credit risks based on 20 categorical/symbolic attributes and is set to 0 otherwise. We use ordinal-encoding for ordinal variables and one-hot-encoding for other categorical variables which yields 46 features in total. In our experiments we take sex as sensitive attribute.

We consider the following off-the-shelf methods: Random Forest (RF) and Logistic Regression (LR). We used the `sklearn` [Pedregosa et al., 2011] implementation of the aforementioned methods.

Each dataset of size N we partition in three parts. The first labeled part (60% of N) is used to train the base classifier, the second unlabeled part (20% of N) is used to apply the proposed post-processing, and the third part (20% of N) is used for evaluation of various statistics, which describe performance of the algorithm.

The hyperparameters of each base algorithm are tuned via 5-fold cross validation with accuracy as the performance measure. The regularization parameter of LR is searched among 30 values, equally spaced in logarithmic scale between 10^{-4} and 10^{4}. For RF the number of trees has been set to 1000 and the size of the subset of features optimized at each node has been searched in $\{d, \lceil d^{15/16} \rceil, \lceil d^{7/8} \rceil, \lceil d^{3/4} \rceil, \lceil d^{1/2} \rceil, \lceil d^{1/4} \rceil, \lceil d^{1/8} \rceil, \lceil d^{1/16} \rceil, 1\}$ where d is the number of features in the dataset. Recall that our post-processing algorithm is parameter-free, thus, the second step is performed without any tuning. Our setup allows to set different reject rates for different groups. However, the exact values heavily depend on the domain specific knowledge and on the problem itself. Because of that, in our experiments, we set $\alpha_1 = \ldots = \alpha_K = \alpha$ for 20 values of α taking values in the uniform grid over $[.8, .99]$, which correspond to reject rate ranging from 20% to 1%.

Given a classifier with reject option g and a test data $\mathcal{T} = \{(x_i, s_i, y_i)\}_{i=1}^{n_{\text{test}}}$, we evaluate the following statistics

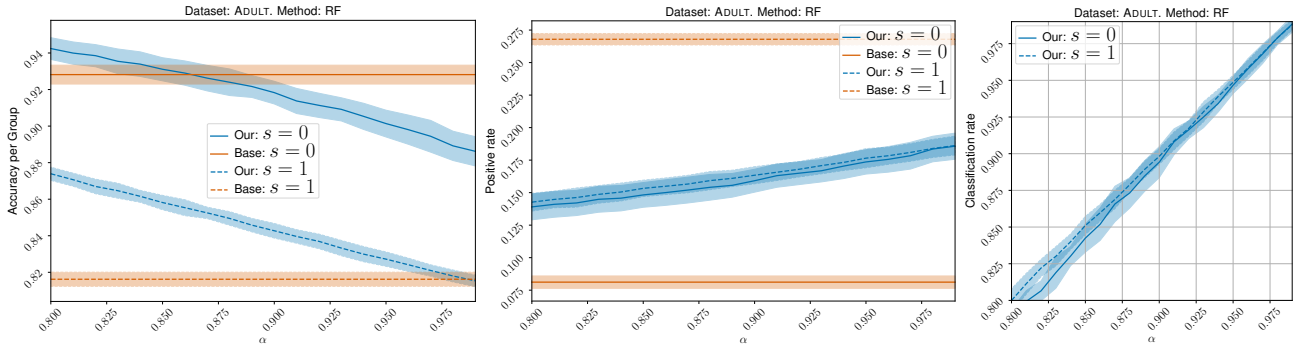
$$\widehat{\text{acc}}_s(g) = \frac{\sum_{i=1}^{n_{\text{test}}} \mathbb{1}\left(g(x_i, s_i) = y_i\right) \mathbb{1}\left(s_i = s\right)}{\sum_{i=1}^{n_{\text{test}}} \mathbb{1}\left(g(x_i, s_i) \neq r\right) \mathbb{1}\left(s_i = s\right)} \quad , \quad s = 1, \ldots, K \ ,$$

$$\widehat{\text{clf}}_s(g) = \frac{\sum_{i=1}^{n_{\text{test}}} \mathbb{1}\left(g(x_i, s_i) \neq r\right) \mathbb{1}\left(s_i = s\right)}{\sum_{i=1}^{n_{\text{test}}} \mathbb{1}\left(s_i = s\right)} \quad , \quad s = 1, \ldots, K \ ,$$

$$\widehat{\text{pos}}_s(g) = \frac{\sum_{i=1}^{n_{\text{test}}} \mathbb{1}\left(g(x_i, s_i) = 1\right) \mathbb{1}\left(s_i = s\right)}{\sum_{i=1}^{n_{\text{test}}} \mathbb{1}\left(g(x_i, s_i) \neq r\right) \mathbb{1}\left(s_i = s\right)} \quad , \quad s = 1, \ldots, K \ .$$

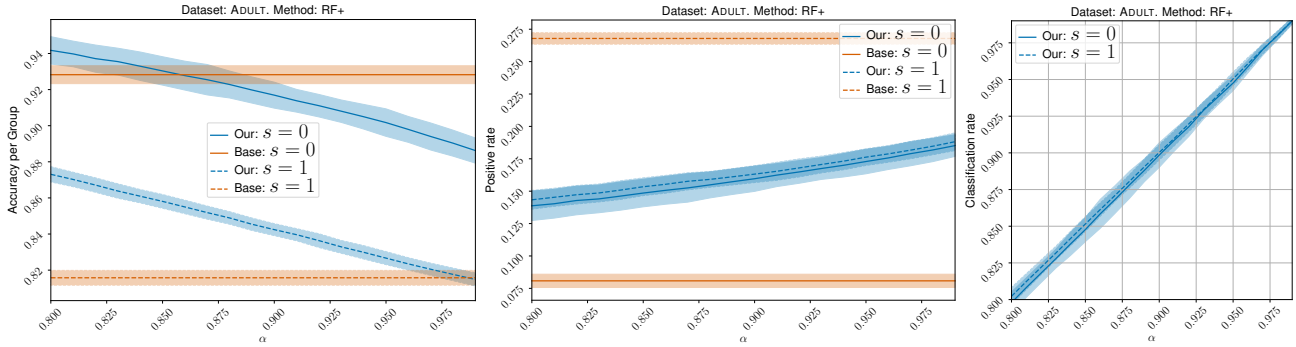
The first statistic measures the accuracy of g, the second the group-wise classification rate of g, and the third one measures the group-wise predicted positive rate of g. It is important to keep in mind that a classifier g which never rejects achieves $\text{clf}_s(g) = 1$ on any dataset.

Figure 1 presents results on ADULT dataset. First of all we observe that the proposed post-processing is effective in imposing reject and fairness constraints as illustrated on Figures 1b-1c. Looking at Figure 1a, we observe that for already moderately low values of rejection our classification algorithm equalizes and even exceeds the accuracy per groups and overall of the base classifier. Figure 2 presents result on GERMAN dataset. Overall conclusions remain the same as for the ADULT dataset. The main difference is an



(a) Accuracy per group (b) Positive rate (c) Classification rate

Figure 3: Results on ADULT dataset with Random Forest (RF) **without** additional randomization as the base estimator. Blue lines correspond to our post-processing method; Orange lines correspond to the base classifier. Dashed line correspond to $s = 1$ and solid line to $s = 0$. Shaded areas correspond to the variance of the result over 20 repetitions.



(a) Accuracy per group (b) Positive rate (c) Classification rate

Figure 4: Results on ADULT dataset with Random Forest (RF) **with** additional randomization as the base estimator. Blue lines correspond to our post-processing method; Orange lines correspond to the base classifier. Dashed line correspond to $s = 1$ and solid line to $s = 0$. Shaded areas correspond to the variance of the result over 20 repetitions.

increase in variance of the result. This effect should not be attributed to the method itself but rather to the size of the two datasets. Indeed, ADULT contains around $40,000$ observation, while GERMAN contains only $1,000$ observations. Hence, it is simply a more difficult task to learn stable classification algorithms on the GERMAN dataset. Remarkably, already 1% of reject rate allows to maintain the accuracy of the base classifier while significantly improving its fairness as illustrated on Figure 2a.

We would also like to highlight the importance of the additive noise perturbation present in Algorithm 1. To this end, we consider RF classifier, which naturally does not lead to continuous estimator $\hat{\eta}(X,S)$ due to its partitioning nature. On Figure 3 we display the performance of our algorithm without any additional randomization and on Figure 4 follow Algorithm 1 with $\sigma = 10^{-3}$. One can see that on Figure 3c the behaviour of our procedure fails to satisfy rejection rate constraints for lower values of α, even, considering the fact, that we have a rather large dataset. In contrast, this phenomenon disappears once the noise is added (see Figure 4c), confirming our theoretical findings. It is important to emphasize that this additional randomization has only a little impact on the group-wise accuracy, which suggest that the randomization step is always advisable in practice.

8 CONCLUSION

We proposed a classification with abstention algorithm which is able to satisfy Demographic Parity and whose reject rate is controlled explicitly. Our procedure is based on a post-processing scheme of any base estimator and can be computed efficiently using LP solvers. We derived distribution-free finite-sample guarantees demonstrating that the proposed method is able to achieve the prescribed constraints with high probability. Under additional mild assumption, we showed the risk of the proposed procedure nearly matches that of the theoretical minimum, provided the initial estimator is consistent. Our experimental results support the developed theory and suggest that by allowing small reject rate it is possible to avoid the accuracy-fairness trade-off.

Acknowledgements

This work was supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129. PMLR, 2019.

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

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

P. Bartlett and M. Wegkamp. Classification with a reject option using a hinge loss. *J. Mach. Learn. Res.*, 9:1823–1840, 2008.

Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. Technical Report MSR-TR-2020-32, Microsoft, May 2020.

Olivier Bousquet and Nikita Zhivotovskiy. Fast classification rates without standard margin assumptions. *arXiv preprint arXiv:1910.12756*, 2019.

Leo Breiman. Random forests. *Machine learning*, 45(1): 5–32, 2001.

Peter Bühlmann and Sara Van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

C. Chow. On optimum error and reject trade-off. *IEEE Trans. Inform. Theory*, 16:41–46, 1970.

C. K. Chow. An optimum character recognition system using decision functions. *IRE Trans. Electron. Comput.*, 6(4):247–254, 1957. doi: 10.1109/TEC.1957.5222035.

Evgenii Chzhen and Nicolas Schreuder. A minimax framework for quantifying risk-fairness trade-off in regression. *arXiv preprint arXiv:2007.14265*, 2020.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. In *NeurIPS 2019-33th Annual Conference on Neural Information Processing Systems*, 2019.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression via plug-in estimator and recalibration with statistical guarantees. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

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

Christophe Denis and Mohamed Hebiri. Consistency of plug-in confidence sets for classification in semi-supervised learning. *Journal of Nonparametric Statistics*, 32(1):42–72, 2020.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

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

Yves Grandvalet, Alain Rakotomamonjy, Joseph Keshet, and Stéphane Canu. Support vector machines with a reject option. *Advances in neural information processing systems*, 21:537–544, 2008.

L. Györfi, Z. Györfi, and I. Vajda. Bayesian decision with rejection. *Problems of Control and Information Theory*, 8, 01 1979.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3323–3331, 2016.

R. Herbei and M. Wegkamp. Classification with reject option. *Canad. J. Statist.*, 34(4):709–721, 2006.

Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. Selective classification can magnify disparities across groups, 2020.

Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311, 1984.

Leonid Genrikhovich Khachiyan. A polynomial algorithm in linear programming. In *Doklady Akademii Nauk*, volume 5, pages 1093–1096. Russian Academy of Sciences, 1979.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

Aleksandr Petrovich Korostelev and Alexandre B Tsybakov. *Minimax theory of image reconstruction*, volume 82. Springer Science & Business Media, 2012.

Yin Tat Lee and Aaron Sidford. Efficient inverse maintenance and faster algorithms for linear programming. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 230–249. IEEE, 2015.

J. Lei. Classification with confidence. *Biometrika*, 101(4): 755–769, 2014.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Jiri Matousek and Bernd Gärtner. *Understanding and using linear programming*. Springer Science & Business Media, 2007.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118, New York, NY, USA, 23–24 Feb 2018. PMLR.

Jaouad Mourtada, Stéphane Gaïffas, Erwan Scornet, et al. Minimax optimal rates for mondrian trees and forests. *Annals of Statistics*, 48(4):2253–2276, 2020.

Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In *Machine Learning in Systems Biology*, pages 65–81, 2009.

Gergely Neu and Nikita Zhivotovskiy. Fast rates for online prediction with abstention. In *Conference on Learning Theory*, pages 3030–3048. PMLR, 2020.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Nikita Puchkin and Nikita Zhivotovskiy. Exponential savings in agnostic active learning through abstention. *arXiv preprint arXiv:2102.00451*, 2021.

Charles J Stone. Consistent nonparametric regression. *The annals of statistics*, pages 595–620, 1977.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stefan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, lhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antonio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 2020.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

M. Wegkamp and M. Yuan. Support vector machines with a reject option. *Bernoulli*, 17(4):1368–1385, 2011.

M. Yuan and M. Wegkamp. Classification methods with reject option based on convex risk minimization. *J. Mach. Learn. Res.*, 11:111–130, 2010.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.