Multi-group Agnostic PAC Learnability

Guy N. Rothblum [1] **Gal Yona** [1]

Abstract

An agnostic PAC learning algorithm finds a predictor that is competitive with the best predictor in a benchmark hypothesis class, where competitiveness is measured with respect to a given loss function. However, its predictions might be quite suboptimal for structured subgroups of individuals, such as protected demographic groups. Motivated by such fairness concerns, we study "multi-group agnostic PAC learnability": fixing a measure of loss, a benchmark class $\mathcal{H}$ and a (potentially) rich collection of subgroups $\mathcal{G}$, the objective is to learn a single predictor such that the loss experienced by every group $g \in \mathcal{G}$ is not much larger than the best possible loss for this group within $\mathcal{H}$. Under natural conditions, we provide a characterization of the loss functions for which such a predictor is guaranteed to exist. For any such loss function we construct a learning algorithm whose sample complexity is logarithmic in the size of the collection $\mathcal{G}$. Our results unify and extend previous positive and negative results from the multi-group fairness literature, which applied for specific loss functions.

1. Introduction

Machine learning tools are used to make and inform increasingly consequential decisions about individuals. Examples range from medical risk prediction to hiring decisions and criminal justice. Automated classification and risk prediction come with benefits, but they also raise substantial societal concerns. One prominent concern is that these algorithms might discriminate against protected and/or disadvantaged groups. In particular, a learned predictor might perform differently on a protected subgroup compared to the general population. The growing literature on algorithmic fairness has studied different concerns. Many works aim to ensure parity or balance between demographic groups, e.g. similar rates of positive predictions or similar false positive or false negative rates (Hardt et al., 2016; Kleinberg et al., 2016). Other works consider accuracy guarantees, such as calibration (Dawid, 1982), for protected groups. Protections at the level of a single group might be too weak (Dwork et al., 2012), and recent works have studied extending these notions to the setting of multiple overlapping groups (Hébert-Johnson et al., 2018; Kearns et al., 2018).

In this work, we focus on the setting of (supervised) agnostic learning (Kearns et al., 1994). Given an i.i.d. training set of labeled data, the goal is to learn a predictor h that performs well on the underlying distribution. Performance is measured by a loss function, and with respect to a fixed class $\mathcal{H}$: the loss incurred by the predictor h should be competitive with the best predictor in $\mathcal{H}$. To capture a wide variety of settings, we aim to be quite general in our treatment of different loss functions.

With fairness in mind, the agnostic learning paradigm raises a fundamental concern: since the predictor's loss is measured over the entire underlying distribution, it might not reflect the predictor's performance on sub-populations such as protected demographic groups. Indeed, it has been demonstrated that standard machine learning tools, when applied to standard data sets, produce predictors whose performance on protected demographic groups is quite poor (Buolamwini & Gebru, 2018).

Motivated by these concerns, we study *multi-group* agnostic learning. For a rich collection $\mathcal{G}$ of (potentially) overlapping groups, our goal is to learn a single predictor h, such that the loss experienced by every group $g \in \mathcal{G}$ (when classified by h) is not much larger than the loss of the best predictor *for that group* in the class $\mathcal{H}$. We emphasize that this should hold for all groups in $\mathcal{G}$ simultaneously.

To see how this objective is different from the usual agnostic PAC, consider the simple example in which $\mathcal{H}$ is the class of hyperplanes and we have two subgroups $S, T \subseteq \mathcal{X}$. Suppose that the data is generated such that every group g has a hyperplane h_g that has very low error on it (but that these are different, so e.g. h_T has large loss on S and vice versa). This means that there is no classifier $h \in \mathcal{H}$ that perfectly labels the data. If S is small compared to T, then the agnostic learning objective could be satisfied by h_T, the

*Equal contribution [1]Weizmann Institute of Science, Rehovot, Israel. Correspondence to: Guy N. Rothblum <rothblum@alum.mit.edu>, Gal Yona <gal.yona@weizmann.ac.il>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

optimal classifier for T. For *multi-group* agnostic PAC, the fact that there is some other classifier in $\mathcal{H}$ that perfectly labels S serves to disqualify h_T (more generally, it could be the case that no $h \in \mathcal{H}$ will be multi-PAC). This also highlights that the multi-group objective becomes challenging when the groups in question are intersecting (if the groups are disjoint, we can combine the optimal classifiers for each group (Dwork et al., 2017)).

A recent work by Blum and Lykouris (Blum & Lykouris, 2019) studied this question in an online setting with sequential predictions. Our focus is on the batch setting. They showed that (for every collection of groups and every benchmark hypothesis class) it is possible to achieve competitive loss for all groups, so long as the loss function is *decomposable*: the loss experienced by each group is an average of losses experienced by its members (this result also applies to the batch setting). On the other hand, they showed a loss function (the average of false negative and false positive rates), for which the objective is infeasible even in the batch setting. Since this loss is non-decomposable (the false positive rate of a classifier depends on the negative rate, which is a property of the entire distribution in question), one might conjecture that the multi-group objective is only possible for decomposable losses. However, this conjecture is false: For example, calibration (Dawid, 1982) is a non-decomposable loss that is compatible with the multi-group objective. Indeed, (Hébert-Johnson et al., 2018) propose an algorithm that guarantees predictions are calibrated across groups (in other words, the calibration loss for each $g \in \mathcal{G}$ is close to zero). In particular, the output of this algorithm would satisfy the multi-group agnostic PAC requirement (with respect to the calibration loss, and for every hypothesis class.

1.1. Our contributions

Motivated by these observations, in this work we formalize two main questions:

1. We define a loss function as (multi-group) *compatible* if it is "appropriate" to use with our objective, in the sense that for every hypothesis class $\mathcal{H}$ and collection of groups $\mathcal{G}$, there exists a hypothesis h that is competitive with $\mathcal{H}$ for every group in $\mathcal{G}$. What makes a loss function compatible? Previous works provide several positive and negative results, but there was no clear characterization of compatibility.

2. Is it always the case that for such "compatible" losses, a multi-group predictor can also be found using a finite number of samples? In other words, is there a separation between multi-group *compatibility* and multi-group *learnability*?

Our main technical contributions answer both questions:

1. We prove a partial information-theoretic characterization of the compatible loss functions.

2. For any such loss function that also satisfies a natural uniform convergence property, we show an algorithm that, for any specified finite collection $\mathcal{G}$ and finite hypothesis class $\mathcal{H}$, learns a multi-group agnostic predictor from labeled data. The sample complexity is logarithmic in the sizes of $\mathcal{G}$ and $\mathcal{H}$. Our algorithm is derived by a reduction to *outcome indistinguishability* (OI), a learning objective recently introduced by Dwork *et al.* (Dwork et al., 2020), drawing a new connection between OI and agnostic learning. This shows that (under minimal assumptions on the loss function), multi-group compatibility implies multi-group learnability.

In slightly more detail, we characterize the compatible loss functions assuming an additional *unambiguity* property (we refer to the characterization as "partial" because of this assumption): we assume that once we fix a single individual and specify the distribution of their label, there is a unique prediction that minimizes the loss for that individual. We view this as a very natural assumption on the loss function, see the discussion following Definition 2.2. We show that if a loss function is compatible and unambiguous, then for each individual, the optimal prediction can be obtained by a fixed "local" function f, whose output only depends on the features and on the marginal distribution of that individual's label. The point is that the function f doesn't depend on the global distribution, but the predictions that it specifies still minimize the loss for that distribution. We call loss functions that satisfy this property *f-proper*, and we show that (under the unambiguity assumption) being f-proper is *equivalent* to multi-group agnostic PAC compatibility.

We then construct a universal multi-group agnostic PAC learning algorithm for any f-proper loss function that also satisfies a minimal uniform convergence property (this is necessary for finite sample complexity). The learning algorithm works for any specified finite hypothesis class $\mathcal{H}$ and any finite collection of groups $\mathcal{G}$, and its sample complexity is logarithmic in $|\mathcal{H}|$ and $|\mathcal{G}|$. The algorithm is obtained via a reduction from multi-group agnostic PAC learning to the recently studied task of outcome indistinguishable learning (Dwork et al., 2020). Beyond unifying previous work in *multi-group fairness*, this result can be thought of as a multi-group analogue for the known result that every PAC learnable class is learnable via empirical risk minimization.

1.2. Related work

The literature on algorithmic fairness is broad and growing. Most pertinent to our work is the literature on fairness notions that aim to stake a middle-ground between

the strong, individual-level semantics of *individual fairness* notion (Dwork et al., 2012) and the weaker but useful *group fairness* notions (Hardt et al., 2016; Kleinberg et al., 2016). We broadly refer to such notions as *multi-group* notions. Some of these works are geared towards guaranteeing parity: equalizing some statistic across the (possibly exponentially-many) subgroups in $\mathcal{G}$. For example, (Kearns et al., 2018) study multi-group versions of both Equality of Opportunity and Demographic Parity notions (Hardt et al., 2016). We note however, that as the collection $\mathcal{G}$ becomes richer and richer, it might be the case multi-group parity can only be obtained via trivial predictions or otherwise undesirable behaviour, such as intentionally making worse predictions on some groups to achieve parity (Chen et al., 2018). A different line of works consider guarantees that are not inherently in conflict with accuracy, e.g. because the Bayes optimal predictor always satisfies multi-group fairness (regardless of $\mathcal{G}$). Notable examples include multi-calibration (Hébert-Johnson et al., 2018) and multi-accuracy (Kim et al., 2019), with subsequent works studying extensions in ranking (Dwork et al., 2019), regression (Jung et al., 2020) and online learning (Gupta et al., 2021). As discussed above, Blum and Lykouris (Blum & Lykouris, 2019) study multi-group agnostic PAC learning (which, depending on the loss function, is often similarly aligned with accuracy) in the online setting. In the batch setting, their approach for learning a multi-PAC predictor is via the mixture of experts approach (Jacobs et al., 1991), where the "experts" in question are a set of $|\mathcal{G}|$ predictors, with each h_i being the optimal predictor for group $g_i \in \mathcal{G}$ in the class $\mathcal{H}$. However, this approach fails to produce a multi-PAC predictor when the loss is non-decomposable, such as calibration (this happens even though, by (Hébert-Johnson et al., 2018), such multi-fair predictors exist and can be found in sample complexity that scales with $\log |\mathcal{G}|$).

2. Preliminaries

Setup and notation. We consider binary classification problems, where $\mathcal{X} \subseteq \mathbb{R}^d$ denotes a feature space and $Y = \{0,1\}$ the target labels. As a general convention, we use $\mathcal{D}$ to denote distributions over $\mathcal{X} \times Y$ and $\mathcal{D}_X$ for the marginal distribution of $\mathcal{D}$ over $\mathcal{X}$. The support of a distribution (w.r.t $\mathcal{X}$) is $\text{supp}(\mathcal{D}) = \text{supp}(\mathcal{D}_X)$. We will sometimes be interested in distributions over $\mathcal{X} \times Y$ for which $|\text{supp}(\mathcal{D})| = 1$, i.e. distributions that are supported on a single element $x \in \mathcal{X}$. We refer to these as "singleton distributions". For a distribution $\mathcal{D}$ and an element $x \in \mathcal{X}$, we use $\mathcal{D}_x$ to denote the singleton distribution on $\mathcal{X} \times \mathcal{Y}$ obtained by restricting $\mathcal{D}$ to $X = x$. A predictor is a mapping $h : \mathcal{X} \to [0,1]$, where $h(x)$ is an estimate for the probability that the label of x is 1. We will sometimes consider the special case of classifiers (binary predictors), whose range is $\{0,1\}$. A hypothesis class is a collection of predictors, and is denoted by $\mathcal{H}$. A subgroup is some subset of $\mathcal{X}$. A collection of subgroups is denoted by $\mathcal{G}$. For this work we generally assume $\mathcal{H}$ and $\mathcal{G}$ are finite (but possibly exponentially large in d). For a distribution $\mathcal{D}$ and a group $g \subseteq \mathcal{X}$, we use $\mathcal{D}_g$ to denote the distribution on $\mathcal{X} \times Y$ obtained by restricting $\mathcal{D}$ to $x \in g$.

2.1. General loss functions

A loss function L is some mapping from a distribution $\mathcal{D}$ and a predictor h (technically, its' restriction to $\mathcal{D}$) to $[0,1]$. We are typically interested in how L changes as we keep the first argument (the distribution) fixed, and only change the second argument (the predictor in question). We thus typically use $L_{\mathcal{D}}(h)$ to denote the loss of h w.r.t. a distribution $\mathcal{D}$. For a sample $S = \{(x_i, y_i)\}_{i=1}^m$ we use $L_S(h)$ to denote the empirical loss, calculated as $L_{\hat{\mathcal{D}}}(h)$, where $\hat{\mathcal{D}}$ is the empirical distribution defined by the sample S. Note that this setup is extremely general, and assumes nothing about the loss (except that it is bounded and can't depend on what happens outside $\mathcal{D}$). In machine learning it is common to consider more structured losses, in which $L_{\mathcal{D}}(h)$ is the expected loss of h on a random example drawn according to $\mathcal{D}$. We refer to such structured losses as *decomposable* losses.

Definition 2.1 (Decomposable losses)**.** *A loss function L is* decomposable *if there exists a function $\ell : X \times Y \times [0,1] \to [0,1]$ such that for every distribution $\mathcal{D}$ and classifier h, $L_{\mathcal{D}}(h) = \mathbf{E}_{(x,y)\sim\mathcal{D}}[\ell(x,y,h(x))]$.*

For example, for binary classifiers a standard decomposable loss is the 0-1 loss, in which $\ell(x,y,h(x)) = \mathbf{1}[h(x) \neq y]$. For predictors, an example of a standard decomposable loss is the squared loss, in which $\ell(x,y,h(x)) = (h(x) - y)^2$.

Beyond decomposable losses. While decomposable losses are standard and common, there are many loss functions of interest that don't have this form – especially in the literature on algorithmic fairness. For this reason, we focus on a general notion of loss functions (which does not explicitly assume losses are decomposable) in our exploration of multi-group agnostic PAC learning. Notable examples of such losses, as used in the algorithmic fairness literature, include the following notions:

Calibration (Chouldechova, 2017; Hébert-Johnson et al., 2018; Kleinberg et al., 2016; Shabat et al., 2020): A predictor is *calibrated* if for every value $v \in [0,1]$, conditioned on $p(x) = v$, the true expectation of the label is close to v. Intuitively, this means that the outputs of the predictor can reasonably be thought of as probabilities, and hence is a fundamental requirement in the literature on forecasting (Dawid, 1982; Foster & Vohra, 1998). For example, in weather prediction, calibrated forecasts ensure that, out of all the days on which the forecaster predicted say 0.8, it

really rained on 80% of them. Calibration loss measures the extent to which a predictor is miscalibrated; e.g., the *expected calibration error* (Kumar et al., 2018) measures the deviation from calibration w.r.t a value v drawn from the distribution induced by the prediction in questions. This loss is not decomposable because it is a global function of the predictions, not a property of the prediction for a single $x \in \mathcal{X}$.

One-sided error rates (Blum & Lykouris, 2019; Blum & Stangl, 2019; Chouldechova, 2017; Hardt et al., 2016; Kearns et al., 2018): The *false positive rate* (similarly, false negative rate) measures the probability of a random example being labeled as $h(x) = 1$, conditioned on the true label being $y = 0$. This isn't a decomposable loss because the exact contribution of a single misclassification depends on the frequency of the negative labels, which is a global property.

Individual fairness (Dwork et al., 2012; Rothblum & Yona, 2018): Individual fairness (IF) requires that individuals that are considered similar w.r.t the task at hand are treated similarly by the classifier. Similarity is specified by a metric $d : \mathcal{X} \times \mathcal{X} \to [0,1]$. If $d(x,x') \approx 0$ (x, x' are similar), then it should be the case that $h(x) \approx h(x')$. An IF loss may quantify the expected violation of this requirement, e.g. over a draw of a pair (x,x') i.i.d from $\mathcal{D}$. This objective isn't decomposable because the requirement w.r.t a single $x \in \mathcal{X}$ depends on the extent to which there are other similar x' in $\text{supp}(\mathcal{D})$.

We note that both the latter two losses (false positive rate and Individual fairness) are typically not interesting on their own – e.g. because the constant classifier $h \equiv 0$ minimizes them. Typically we are interested in these losses as either an additional constraint on an existing objective, or paired with an additional loss (e.g. IF + accuracy based loss, or false positive rate + false negative rate)

In the rest of this section we continue to specify two minimal conditions for the losses we consider. We will use the notation $\mathcal{L}$ for the class of losses that satisfy both conditions. The first condition is *unambiguity*. It guarantees that distributions over a single example x have a unique loss minimizer.

Definition 2.2 (Unambiguity)**.** *A loss L is unambiguous if for every singleton distribution $\mathcal{D}_x$ over $\mathcal{X} \times Y$, there is a unique prediction $h(x)$ that minimizes the loss. That is,* $\left|\arg\min_{h(x)\in[0,1]} L_{\mathcal{D}_x}(h)\right| = 1$.

Standard decomposable losses satisfy this condition because the function $\ell(h(x),y)$ typically has a unique minimum w.r.t its first argument. For example when ℓ corresponds to the squared loss the optimal labeling is $h(x) = \mathbf{E}[y|x]$ and when ℓ corresponds to the expected 0-1 loss it is $h(x) = 1[\mathbf{E}[y|x] \geq 0.5]$. With respect to the fairness-motivated losses mentioned above, unambiguity fails w.r.t individual fairness and the one-sided error rates (exactly because they can both be minimized by the constant all-zeroes classifier, regardless of the true distribution). But as mentioned above, these losses are rarely interesting on their own, and combined losses (individual fairness with an unambiguous loss, or an average of false positive rate and false negative rate) are unambiguous. In other words, all losses that are of practical interest to us are unambiguous.

The second condition we will require is that the empirical risk $L_S(\cdot)$ can really be used to approximate the true risk $L_{\mathcal{D}}(\cdot)$, for a sufficiently large sample S. To build up to this notion we first recall the standard definition of uniform convergence for hypotheses classes.

Definition 2.3 (Uniform Convergence for hypotheses classes)**.** *We say that a hypothesis class $\mathcal{H}$ has the uniform convergence property (w.r.t. a domain $X \times Y$ and a loss function L) if there exists a function $m^{UC}_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ such that for every $\varepsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over $X \times Y$, if S is a sample of $m \geq m^{UC}_{\mathcal{H}}(\varepsilon,\delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, with probability of at least $1-\delta$, $\forall h \in \mathcal{H}$: $\quad |L_S(h) - L_D(h)| \leq \varepsilon$.*

In our context, we will be interested in uniform convergence as a property of the *loss function*. We will say that a loss L has uniform convergence (w.r.t finite classes) with sample complexity $m^{UC}_L : (0,1)^2 \times \mathbb{N} \to \mathbb{N}$ if every finite class $\mathcal{H}$ has the uniform convergence property w.r.t L with sample complexity $m^{UC}_{\mathcal{H}}(\varepsilon,\delta) \leq m^{UC}_L(\varepsilon,\delta,|\mathcal{H}|)$. Specifically, we will be interested in losses that have the uniform convergence property with sample complexity that depends polynomially on $1/\varepsilon, 1/\delta$ and $\log|\mathcal{H}|$. This gives rise to the following definition:

Definition 2.4 (Uniform convergence for loss functions)**.** *A loss L has the uniform convergence property (w.r.t finite classes) with sample complexity $m^{UC}_L : (0,1)^2 \times \mathbb{N}$ if there exists a polynomial $f : \mathbb{R}^3 \to \mathbb{N}$ such that for every $\varepsilon, \delta \in (0,1)$ and $k \in \mathbb{N}$,*

$$m^{UC}_L(\varepsilon,\delta,k) \triangleq \max_{\mathcal{H}:|\mathcal{H}|=k} m^{UC}_{\mathcal{H}}(\varepsilon,\delta) \leq f(1/\varepsilon, 1/\delta, \log(k))$$

The uniform convergence property is satisfied by any decomposable loss function. This follows by a combination of Heoffding's bound (for a single h) and a union bound to get a simultaneous guarantee for every $h \in \mathcal{H}$. Out of the fairness-motivated losses we discussed above, only the loss $L_{\mathcal{D}}(h) = a \cdot \texttt{FPR}_{\mathcal{D}}(h) + b \cdot \texttt{FNR}_{\mathcal{D}}(h)$ doesn't have uniform convergence, as we prove in Appendix A. For calibration, uniform convergence follows as a special case of the bounds in (Shabat et al., 2020); for individual fairness, the argument is similar to the standard argument for decomposable losses, only this time the concentration argument is w.r.t pairs of samples from $\mathcal{D}$ (Rothblum & Yona, 2018).

loss	notation	in $\mathcal{L}$?
decomposable	$L^{\ell}_{\mathcal{D}}(h)$	✓
calibration	$L^{\mathtt{C}}_{\mathcal{D}}(h)$	✓
IF + decomposable	$a \cdot L^{\mathtt{IF}}_{\mathcal{D}}(h) + b \cdot L^{\ell}_{\mathcal{D}}(h)$	✓
error rates	$a \cdot L^{\mathtt{FPR}}_{\mathcal{D}}(h) + b \cdot L^{\mathtt{FNR}}_{\mathcal{D}}(h)$	✗

Figure 1. Summary: loss functions and the set $\mathcal{L}$ (all losses that are unambiguous and have the uniform convergence property.)

To summarize, we have defined a collection of "reasonable" losses $\mathcal{L}$ as all loss functions that are both unambiguous and have the uniform convergence property. We have argued that of the loss functions we discussed, this only rules out the one-sided error rate loss; see Figure 1 for an overview.

2.2. Learning paradigms

In this work we draw a connection between an extension of the classic notion of agnostic PAC learnability in the presence of multiple groups and the recently introduced complexity-theoretic inspired framework of Outcome Indistinguishability (Dwork et al., 2020), which builds on previous work in algorithmic fairness (Hébert-Johnson et al., 2018). In this section we review both of these learning paradigms, starting with agnostic PAC learnability. For consistency, we give the definition w.r.t finite classes.

Definition 2.5 (Agnostic-PAC learnability)**.** *A hypothesis class $\mathcal{H}$ is agnostic PAC learnable with respect to a set $X \times Y$ and a loss function L if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\varepsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over $\mathcal{X} \times Y$, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns h such that, with probability of at least $1-\delta$ (over the choice of the m training examples),*

$$L_{\mathcal{D}}(h) \leq L_{\mathcal{D}}(\mathcal{H}) + \varepsilon \tag{1}$$

Additionally, the sample complexity $m_{\mathcal{H}}$ should be polynomial in $1/\varepsilon, 1/\delta$ and in $\log(|\mathcal{H}|)$.

Outcome indistinguishability. A predictor $\tilde{p} : \mathcal{X} \to [0,1]$ can be viewed as providing a generative model for outcomes, where for $x \in \mathcal{X}$ a binary outcome is sampled as $y \sim \mathrm{Ber}(\tilde{p}(x))$. Given a distribution $\mathcal{D}$ and a predictor $\tilde{p}$, we define the "modeled distribution" $\tilde{\mathcal{D}} = \mathcal{D}(\tilde{p})$ as the distribution over $\mathcal{X} \times Y$ obtained by first drawing x according to the marginal distribution of $\mathcal{D}$ on $\mathcal{X}$ and then labeling it as $y \sim \mathrm{Ber}(\tilde{p}(x))$. The framework of Outcome Indistinguishability (OI) aims to guarantee that outcomes produced by $\tilde{p}$ are indistinguishable from the true outcomes under $\mathcal{D}$. This guarantee is formulated with respect to a fixed class $\mathcal{A}$ of distinguishers, and the requirement is that every distringuisher $A \in \mathcal{A}$ behaves similarly on samples from the real distribution $\mathcal{D}$ and on samples from the "modeled distribution" $\tilde{\mathcal{D}}$.

As discussed and studied in (Dwork et al., 2020), there are several different ways to instantiate this framework based on the power of the distinguishers. In particular, two axes of variation are (i) the input to the distinguisher (single sample vs multi-sample) and (ii) the access level it receives to the predictor $\tilde{p}$ in question. In this work, we focus on multi-sample Sample-Access OI distinguishers, where the inputs to each distinguisher $A \in \mathcal{A}^k$ are $k-$tuples of the form $\{(x_i, y_i, \tilde{p}_i)\}_{i=1}^k$, where for every $i \in [k]$, (x_i, y_i) is sampled from either $\mathcal{D}$ or $\tilde{\mathcal{D}}$:

Definition 2.6 (Outcome Indistinguishability (Dwork et al., 2020))**.** *Fix a distribution $\mathcal{D}$, a collection of distinguishers $\mathcal{A}^k$ and $\tau > 0$. A predictor $\tilde{p} : \mathcal{X} \to [0,1]$ satisfies $(\mathcal{A}^k, \tau)$-OI if for every $A \in \mathcal{A}^k$,*

$$\Big| \Pr_{\{(x_i,y_i)\}_{i=1}^k \sim \mathcal{D}^k} [A(\{(x_i, y_i, \tilde{p}(x_i)\}_{i=1}^k) = 1] - \Pr_{\{(x_i,y_i)\}_{i=1}^k \sim \tilde{\mathcal{D}}^k} [A(\{(x_i, y_i, \tilde{p}(x_i)\}_{i=1}^k) = 1] \Big| \leq \tau$$

Much like the definition of (regular) PAC learning, we say that an algorithm *learns* multi-sample OI if on receiving sufficiently many i.i.d samples from the distribution, it is guaranteed to return a predictor that is "probably approximately" OI.

Definition 2.7 (OI learnability)**.** *A family of k-sample distinguishers $\mathcal{A}^k$ is multi-sample OI learnable if with respect to a set $\mathcal{X} \times Y$ if there exists a function $m_{\mathcal{A}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\tau, \eta \in (0,1)$ and for every distribution $\mathcal{D}$, when running the learning algorithm on $m \geq m_{\mathcal{A}}(\tau, \eta)$ i.i.d example generated by $\mathcal{D}$, the algorithm returns h such that with probability at least $1-\eta$ (over the choice of the m training examples), h satisfies $(\tau, \mathcal{A}^k)$-OI.*

Additionally, the sample complexity $m_{\mathcal{A}}$ should be polynomial in $1/\tau, 1/\eta$ and in $\log(\mathcal{A}^k)$.

Dwork *et al.* (Dwork et al., 2020) showed that every distinguisher class $\mathcal{A}^k$ is OI-learnable. In our analysis we use the following theorem, which bounds the sample complexity of this algorithm and follows from (Dwork et al., 2020).

Theorem 2.8 (from (Dwork et al., 2020))**.** *Fix a class of $k-$sample distinguishers $\mathcal{A}^k$. There exists an algorithm $\mathtt{OI}_{\mathcal{A}^k}$ that satisfies the requirement of Definition (2.7), and whose sample complexity is $O\left(\frac{k \cdot \log|A^k|/\eta}{\tau^4}\right)$.*

3. Agnostic PAC with multiple groups

The objective of agnostic PAC learning is to output a predictor h that satisfies $L_{\mathcal{D}}(h) \lesssim L_{\mathcal{D}}(\mathcal{H})$ (Equation 1). Putting aside questions of computational complexity and sample complexity, this objective is itself always feasible. In other words, regardless of the loss in question, it can always be obtained using a learner who has full knowledge of the distribution $\mathcal{D}$ and is not constrained in runtime. From an information-theoretic perspective, this makes the interesting question the question of what can be done in finite samples – hence the focus on agnostic PAC *learnability*.

A multi-group extension of agnostic PAC asks for a predictor that satisfies the above, but simultaneously for every group g in a collection $\mathcal{G}$: $L_{\mathcal{D}_g}(h) \lesssim L_{\mathcal{D}_g}(\mathcal{H})$, where $\mathcal{D}_g$ denotes the restriction of $\mathcal{D}$ to samples from g.

When $\mathcal{G}$ consists of intersecting groups, however, it is not immediately clear that this objective is always feasible: it might not be satisfied by *any* predictor $h : \mathcal{X} \to [0,1]$. For a simple (but contrived) example, let h^0, h^1 denote the all-zeroes and all-ones predictors, and consider a loss L that specifies that $L_{\mathcal{D}_S}(h^0) = 0$ and $L_{\mathcal{D}_T}(h^1) = 0$ (and for any other classifier h, the loss of every distribution is always 1). Then the multi-group objective w.r.t $\mathcal{G} = \{S, T\}$ requires that we label the intersection $S \cap T$ as both 1 and 0, which is impossible. The following lemma proves that this can be the case even for seemingly natural losses, that are in $\mathcal{L}$.

Proposition 3.1. *There exists $L \in \mathcal{L}$ for which some problem instances (corresponding to a distribution $\mathcal{D}$, class $\mathcal{H}$, subgroups $\mathcal{G}$ and $\varepsilon > 0$) don't admit a multi-group agnostic PAC solution. In other words, there is no classifier h for which $L_{\mathcal{D}_g}(h) \lesssim L_{\mathcal{D}_g}(\mathcal{H})$ for every group $g \in \mathcal{G}$.*

The loss in question is a weighted combination of an individual fairness loss with an accuracy loss. We construct two intersecting groups S and T and a similarity metric d that specifies $d(x, x') = 0$ if and only if $x \in S - T$ and $x' \in S \cap T$. However, the true labels of these individuals are different: for $x \in S - T$ the label is $y = 0$ but for $x \in S \cap T$ the label is $y = 1$. This creates a situation in which group T is optimizing strictly for accuracy (and so it wants the intersection to be labeled as $\hat{y} = 1$) whereas for the group S the dominating term is IF (and so it wants everyone, and in particular the intersection, to be labeled as $\hat{y} = 0$). Thus, any classifier that is simultaneously multi-PAC w.r.t both T and S must label the intersection as both 0 and 1, which is impossible. See Appendix B for the full proof.

We note that (Blum & Lykouris, 2019) already demonstrated that this impossibility occurs in the batch setting. This motivated them to focus on the 0-1 loss. However, their counter example is for the error-rate type loss we discussed earlier, which is a fixed combination of the false positive rate and false negative rate of a classifier. As noted in Section 2.1, this loss doesn't have the uniform convergence property and is therefore not in $\mathcal{L}$. Proposition 3.1 clarifies that uniform convergence is not the issue, and there are natural (and otherwise reasonable) losses that are not "appropriate" to use with the multi-group objective (in the sense that even an approximate solution is not guaranteed to exist).

In light of this discussion, we proceed to explicitly separate the question of *feasibility* (whether it's always the case that some multi-group solution is guaranteed to exist) from *learnability* (whether we can find it with a "reasonable" number of samples). Formally, we define two notions – compatibility and learnability – that formalize these concepts. Importantly, both are properties of the loss function in question, taking a universal quantifier over the class $\mathcal{H}$ and the groups $\mathcal{G}$ (and the other aspects of the problem instance).

Definition 3.2 (Multi-PAC compatibility)**.** *We say that a loss L is* multi-PAC compatible *if for every distribution $\mathcal{D}$, class $\mathcal{H}$, subgroups $\mathcal{G}$ and $\varepsilon > 0$, there exists $h : \mathcal{X} \to [0,1]$ such that for every group $g \in \mathcal{G}$, $L_{\mathcal{D}_g}(h) \leq L_{\mathcal{D}_g}(\mathcal{H}) + \varepsilon$.*

Multi-PAC learnability strengthens the above requirement by asking that such a solution can also be found by a learning algorithm whose sample complexity is constrained to depend inverse-polynomially on the parameters in question and logarithmically on the sizes of $\mathcal{H}$ and $\mathcal{G}$.

Definition 3.3 (Multi-PAC learnability)**.** *We say that a loss L is* multi-PAC learnable *with sample complexity $m_L^{gPAC} : (0,1)^3 \times \mathbb{N}^2 \to \mathbb{N}$ if there exists a learning algorithm with the following property: For every $\varepsilon, \delta, \gamma \in (0,1)$, for every finite hypothesis class $\mathcal{H}$, for every finite collection of subgroups $G \subseteq 2^{\mathcal{X}}$ and for every distribution $\mathcal{D}$ over $\mathcal{X} \times Y$, when running the learning algorithm on $m \geq m_L^{gPAC}(\varepsilon, \delta, \gamma, |\mathcal{H}|, |\mathcal{G}|)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns h such that, with probability at least $1 - \delta$ (over the choice of the m training examples and the coins of the learning algorithm) $g \in \mathcal{G}_\gamma$, $L_{\mathcal{D}_g}(h) \leq L_{\mathcal{D}_g}(\mathcal{H}) + \varepsilon$, where $\mathcal{G}_\gamma \subseteq \mathcal{G}$ is the subset of groups whose mass under $\mathcal{D}$ is at least γ: $\mathcal{G}_\gamma = \{g \in \mathcal{G} : \mathbf{Pr}_{\mathcal{D}}[x \in g] \geq \gamma\}$.*

Additionally, the sample complexity m_L^{gPAC} should be polynomial in $1/\varepsilon, 1/\delta, 1/\gamma$ and $\log(|\mathcal{H}|), \log(|\mathcal{G}|)$.

4. Compatibility $\iff$ Learnability via OI

In this section we prove our main result: that for loss functions in $\mathcal{L}$, multi-PAC compatibility implies also multi-PAC learnability.

Theorem 4.1. *If a loss $L \in \mathcal{L}$ is multi-group compatible (Definition 3.2), then it is also multi-group learnable (Definition 3.3).*

Towards proving Theorem 4.1, we introduce an additional property of loss functions, which we define below.

Definition 4.2 (f-proper)**.** *For a function $f : \mathcal{X} \times [0,1] \to [0,1]$, we say that a loss L is f-proper if for every distribution $\mathcal{D}$ on $\mathcal{X} \times Y$, the classifier $h_{\mathcal{D}}$ given by $h_{\mathcal{D}}(x) = f(x, \mathbf{E}_{\mathcal{D}}[y|x])$ minimizes the loss w.r.t $\mathcal{D}$: $h_{\mathcal{D}} \in \arg\min_h L_{\mathcal{D}}(h)$.*

Recall that proper losses (or proper scoring functions) are losses that are minimized by the conditional expectation predictor $x \mapsto \mathbf{E}_{\mathcal{D}}[y|x]$ (Buja et al., 2005). Definition 4.2 is a weaker requirement that says that for every distribution a minimizer can be obtained as some *local* transformation of this predictor (i.e. that does not depend on the rest of the distribution).

Recalling that we defined $\mathcal{L}$ as precisely all loss functions that satisfy both uniform convergence and unambiguity, we see that Theorem 4.1 follows as a direct corollary of the following two lemmas.

Lemma 4.3. *If L is unambiguous (Definition 2.2) and multi-group compatible (Definition 3.2), then L is f-proper (Definition 4.2).*

Lemma 4.4. *If L is f-proper (Definition 4.2) and has the uniform convergence property (Definition 2.4), then L is multi-group learnable (Definition 3.3).*

We note that the other direction of Lemma 4.3 is also true (if a loss is f-proper, it is also multi-group compatible): this follows immediately, since labeling everyone in $\text{supp}(\mathcal{D})$ according to f is optimal – and in particular satisfies the multi-group requirement for every $\mathcal{H}$. This means that the notion of $f-$proper provides a partial characterization of compatibility (up to unambiguity).

The full proofs of Lemmas 4.3 and 4.4 can be found in Appendices C and D, respectively. In the rest of this section we give an overview of both proofs, and end with a discussion of the implications of the equivalence.

4.1. Overview of Lemma 4.3

Let $L \in \mathcal{L}$ be a loss function that is multi-group compatible and unambiguous (we do not use the uniform convergence property here). We show that L is also f-proper, where the function f maps a singleton distribution (specified by an input x and the conditional probability that the label is 1) to whatever prediction minimizes the loss on that distribution. In more detail: for an input $x \in \mathcal{X}$ and a value $z \in [0,1]$, let $\mathcal{D}_{x,z}$ be the singleton distribution over $\{x\} \times \{0,1\}$, where the label is drawn by the Bernoulli distribution $\text{Ber}(z)$. Let $h_{x,z}$ be the predictor that minimizes the loss on this distribution, i.e. $h_{x,z} = \arg\min_h L_{\mathcal{D}_{x,z}(h)}$, and recall that by unambiguity, the prediction that minimizes the loss is unique, and so the value $h_{x,z}(x)$ is well defined. We take $f(x,z) = h_{x,z}(x)$.

It remains to prove that L is an f-proper loss function. Suppose for contradiction that it is not: i.e., there exists a distribution $\mathcal{D}$ for which $h_{\mathcal{D}}(x) = f(x, \mathbf{E}_{\mathcal{D}}[y|x])$ does not minimize the loss, i.e. there exists some predictor h' s.t. $L_{\mathcal{D}}(h') < L_{\mathcal{D}}(h_{\mathcal{D}})$. We show this contradicts the multi-group compatibility of L. To see this, define a collection of sets that includes all the singletons in the support of $\mathcal{D}$, as well as the global set comprised of the entire support of $\mathcal{D}$. By unambiguity, the singletons all "want" to be labeled by $h_{\mathcal{D}}$. On the other hand, the global set wants to be labeled by h'. Whatever predictor we pick, the loss on either the global set or of one of the singletons will not be optimal. Thus, for the above collection of groups $\mathcal{G}$ (comprised of all singletons plus the global set), and for the hypothesis class $\mathcal{H}$ that includes $h_{\mathcal{D}}$ and h', it is impossible to obtain predictions that are competitive with $\mathcal{H}$ for all groups in $\mathcal{G}$ simultaneously.

4.2. Overview of Lemma 4.4

Given a loss function L that is f-proper and has the uniform convergence property, we want to construct a multi-group agnostic PAC learning algorithm that works for any (finite) hypothesis class $\mathcal{H}$ and (finite) collection of groups $\mathcal{G}$. The algorithm will work by a reduction to the task of OI learning (see above): namely, we construct a collection $\mathcal{A}$ of distinguishers, and show that any predictor $\tilde{p}$ that is OI w.r.t this collection can be used to derive a multi-group agnostic predictor h. In particular, we show that if $\tilde{p}$ is OI (w.r.t $\mathcal{A}$), then $\tilde{h}(x) = f(x, \tilde{p}(x))$ is a multi-group agnostic predictor (recall that f is the local transformation for the f-proper loss function L). The collection of distinguishers depends on the loss function L, on the hypothesis class $\mathcal{H}$ and on the collection of groups $\mathcal{G}$. This reduction, together with the OI learning algorithm of Theorem 2.8 (from (Dwork et al., 2020)), gives a "universal" multi-group agnostic learning algorithm for any f-proper loss function. The algorithm is described in Figure 1.

It remains to construct the family of distinguishers $\mathcal{A}$, and to prove the reduction. Towards this, fix a group $g \in \mathcal{G}$ and fix a hypothesis $h \in \mathcal{H}$. We want to guarantee that the loss of the hypothesis $\tilde{h}(x) = f(x, \tilde{p}(x))$ is competitive with the loss of h, where both losses are measured on the distribution $\mathcal{D}_g$ over members of the group g. We begin by observing that this is true when the labels are drawn by $\tilde{p}(x)$ (as in the distribution $\tilde{\mathcal{D}}$). We will use OI (with an appropriately constructed distinguisher) to ensure that it is also true for the "real" distribution $\mathcal{D}_g$.

In more detail, since L is an f-proper loss function, we have:

$$L_{\tilde{\mathcal{D}}_g}(\tilde{h}) \leq L_{\tilde{\mathcal{D}}_g}(h),$$

because in $\tilde{\mathcal{D}}$ the labels are indeed generated by $\tilde{p}$, i.e.

$\tilde{p}(x) = E_{\tilde{\mathcal{D}}}[y|x]$. By uniform convergence, this will remain true even if we consider the empirical loss over a (sufficiently large) i.i.d. sample from $\tilde{\mathcal{D}}_g$. With this in mind, we define a distinguisher $A^k_{g,h,\alpha}$, which takes as input k samples $\{(x_i, y_i, p_i)\}$ and checks whether, for the samples where $x_i \in g$, it is true that the loss obtained by predicting $f(x_i, p_i)$ for each x_i is competitive with the loss obtained by h on those samples (up to an additive factor of α). By the above discussion, when the labels y_i values are drawn by $\mathrm{Ber}(\tilde{p}(x_i))$, and assuming that there are sufficiently many samples in g to guarantee uniform convergence for the loss L, the distinguisher will accept with high probability. See Figure 2 for a full description of the distinguisher.

Now, if $\tilde{p}$ is OI w.r.t. a class that includes the distinguisher $A^k_{g,h,\alpha}$, then the distinguisher should accept with similar probabilities when the labeled examples are drawn by $\tilde{\mathcal{D}}$ or by $\mathcal{D}$ (where in both cases $p_i = \tilde{p}(x_i)$). I.e., $A^k_{g,h,\alpha}$ should also accept w.h.p. when the labeled examples are drawn by $\mathcal{D}$. This can only happen if the predictor $\tilde{h}$ is competitive with the hypothesis h w.r.t. the distribution $\mathcal{D}_g$, which is exactly the guarantee we wanted from $\tilde{h}$!

The class $\mathcal{A}$ of distinguishers includes such a distinguisher for each $g \in \mathcal{G}$ and $h \in \mathcal{H}$, and thus if $\tilde{p}$ is OI w.r.t. $\mathcal{A}$, we conclude that the loss of $\tilde{h}$ is competitive with every $h \in \mathcal{H}$ for every group in $\mathcal{G}$ simultaneously. Note that the distinguishers in $\mathcal{A}$ use multiple samples, and the number of samples must be sufficiently large so that (for any sufficiently-large group g) w.h.p. enough of them fall in g to guarantee uniform convergence.

The sample complexity of the learning algorithm is governed by the sample complexity of OI learning, which is logarithmic in the number of distinguishers. Since the class $\mathcal{A}$ includes $|\mathcal{G}| \cdot |\mathcal{H}|$ distinguishers, the resulting learning algorithm has sample complexity that is logarithmic in $|\mathcal{G}|$ and in $|\mathcal{H}|$. We note that we need $\mathcal{G}$ and $\mathcal{H}$ to be finite because the known OI learning algorithm works for finite collections of distinguishers.

Acknowledgements. We thank Cynthia Dwork, Michael P. Kim and Omer Reingold for fruitful discussions throughout this work.

Algorithm 1 $\texttt{MultiGroup}_{\mathtt{L,f}}(\epsilon, \delta, \gamma, \mathcal{H}, \mathcal{G})$

Parameters: loss function L, function $f : \mathcal{X} \times [0,1] \to [0,1]$
Input: accuracy parameter $\epsilon \in (0,1)$, failure probability $\delta \in (0,1)$, minimal subgroups size parameter $\gamma \in (0,1)$, hypothesis class $\mathcal{H}$, collection of subgroups $\mathcal{G}$.
Output: A classifier h satisfying the (ε, δ)-multi-group guarantee w.r.t $\mathcal{H}$ and $\mathcal{G}$
Set $\varepsilon' = \alpha = \varepsilon/4$ and $\delta' = \eta = \tau = \delta/4$.
Set $k_{\mathcal{G}} = m_L^{UC}(\varepsilon', \delta', |\mathcal{H}| + 1)$.
Set $k = 10 \cdot \frac{1}{\gamma} \cdot \log \frac{1}{\delta'} \cdot k_{\mathcal{G}}$.
Let $\mathcal{A} = \left\{ A^{L,f,k}_{g,h,\alpha} \mid g \in \mathcal{G}, h \in \mathcal{H} \right\}$ be a collection of distinguishers, as defined in Algorithm 2.
Invoke OI as a sub-routine to learn $\tilde{p} \leftarrow \mathtt{OI}(\tau, \eta, \mathcal{A})$.
return $f(\tilde{p})$

Algorithm 2 $A^{L,f,k}_{g,h,\alpha}$ (multi-sample Sample-Access OI distinguisher)

Parameters: number of samples $k \in \mathbb{N}$, group $g \subseteq \mathcal{X}$, classifier $h : \mathcal{X} \to [0,1]$, loss function L, function $f : \mathcal{X} \times [0,1] \to [0,1]$
Input: $\{(x_i, y_i, p_i)\}_{i=1}^k$, where $x_i \in \mathcal{X}$, $y_i \in \{0,1\}$ and $p_i \in [0,1]$
Output: A binary output denoting Accept/Reject
$I_g = \{i : x_i \in g\}$
$S_g = \{(x_i, y_i)\}_{i \in I_g}$
Define a predictor f_g as

$$f_g(x) = \begin{cases} f(x_i, p_i) & \exists i \in [k] \text{ such that } x = x_i \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

if $L_{S_g}(f_g) < L_{S_g}(h) + 2\alpha$ **then**
  return 1
end if
return 0

References

Blum, A. and Lykouris, T. Advancing subgroup fairness via sleeping experts. *arXiv preprint arXiv:1909.08375*, 2019.

Blum, A. and Stangl, K. Recovering from biased data: Can fairness constraints improve accuracy? *arXiv preprint arXiv:1912.01094*, 2019.

Buja, A., Stuetzle, W., and Shen, Y. Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November*, 3, 2005.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C. (eds.), *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 2018. URL http://proceedings.mlr.press/v81/buolamwini18a.html.

Chen, I., Johansson, F. D., and Sontag, D. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, pp. 3539–3550, 2018.

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

Dawid, A. P. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

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

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. Decoupled classifiers for fair and efficient machine learning. *arXiv preprint arXiv:1707.06613*, 2017.

Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., and Yona, G. Learning from outcomes: Evidence-based rankings. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 106–125. IEEE, 2019.

Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., and Yona, G. Outcome indistinguishability. *arXiv preprint arXiv:2011.13426*, 2020.

Foster, D. P. and Vohra, R. V. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

Gupta, V., Jung, C., Noarov, G., Pai, M. M., and Roth, A. Online multivalid learning: Means, moments, and prediction intervals. *arXiv preprint arXiv:2101.01739*, 2021.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Hébert-Johnson, Ú., Kim, M. P., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948, 2018.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Jung, C., Lee, C., Pai, M. M., Roth, A., and Vohra, R. Moment multicalibration for uncertainty estimation. *arXiv preprint arXiv:2008.08037*, 2020.

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

Kearns, M. J., Schapire, R. E., and Sellie, L. M. Toward efficient agnostic learning. *Machine Learning*, 17(2-3): 115–141, 1994.

Kim, M. P., Ghorbani, A., and Zou, J. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Kumar, A., Sarawagi, S., and Jain, U. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pp. 2805–2814. PMLR, 2018.

Rothblum, G. and Yona, G. Probably approximately metric-fair learning. In *International Conference on Machine Learning*, pp. 5680–5688. PMLR, 2018.

Shabat, E., Cohen, L., and Mansour, Y. Sample complexity of uniform convergence for multicalibration. *arXiv preprint arXiv:2005.01757*, 2020.