Multicalibration: Calibration for the (Computationally-Identifiable) Masses

Úrsula Hébert-Johnson[1] Michael P. Kim[1] Omer Reingold[1] Guy N. Rothblum[2]

Abstract

We develop and study *multicalibration* as a new measure of fairness in machine learning that aims to mitigate inadvertent or malicious discrimination that is introduced at training time (even from ground truth data). Multicalibration guarantees meaningful (calibrated) predictions for *every* subpopulation that can be identified within a specified class of computations. The specified class can be quite rich; in particular, it can contain many overlapping subgroups of a protected group. We demonstrate that in many settings this strong notion of protection from discrimination is provably attainable and aligned with the goal of accurate predictions. Along the way, we present algorithms for learning a multicalibrated predictor, study the computational complexity of this task, and illustrate tight connections to the agnostic learning model.

1. Introduction

Machine-learned predictors are informing decisions that affect all aspects of life; from news article recommendations to criminal sentencing decisions to healthcare diagnostics, increasingly algorithms are used to make predictions about individuals. A potential risk is that these predictors might discriminate against groups of individuals that are protected by law or by ethics. Indeed, examples of such unintended but harmful discrimination have been well-documented across many learning tasks including image classification (Buolamwini & Gebru, 2018) and natural language tasks (Bolukbasi et al., 2016). This work aims to mitigate such risks of algorithmic discrimination in the context of prediction tasks.

The output of a learning algorithm can be discriminatory for a number of reasons. First, the training data may contain biases that should be corrected. Second, the *analysis* of the training data may inadvertently introduce biases that are not borne out in the data. In this work, we focus on the latter concern.

Indeed, even given accurate ground-truth training data, the typical approach to supervised learning – choosing a model that minimizes the expected loss on the training data – runs the risk of choosing a prediction model that is good for the majority population, but overlooks the minority populations. Consider the case where a financial institution trains a model to predict the probability that applicants will default on their loans. If on average, the individuals from S are financially disadvantaged compared to the majority population, the model may assign a fixed, low probability to all $i \in S$, while still achieving good empirical loss by predicting very accurately in the majority population. Such a model discriminates against the *qualified* members of S. Worse yet, this form of discrimination has the potential to amplify S's underrepresentation by refusing to approve members that are capable of repaying the loan.

Focusing on such concerns, we develop a theoretical framework that aims to mitigate such risks of algorithmic discrimination, in the context of prediction tasks. Specifically, we focus on a setting where a learner has access to a small sample of ground truth data D from some domain of individuals $\mathcal{X}$. Each individual $i \in D$ has a boolean label $o_i \in \{0,1\}$ representing the outcome of a certain stochastic event (ad click, loan repayment, cancer diagnosis, etc.) the learner wishes to predict. We suppose that for each $i \in \mathcal{X}$, there is an underlying probability p_i^* which governs the distribution of the resulting outcome o_i. We say a *predictor* $f : \mathcal{X} \to [0,1]$ is a map from individuals $i \in \mathcal{X}$ to an estimate of the true parameters. Next, we discuss desirable properties of predictors that motivate our new perspective on fairness.

Calibration and Multicalibration. If we do not want a predictor f to downplay the fitness of a group $S \subseteq \mathcal{X}$, we can require that it be (approximately) accurate in expectation over S; namely, that $\left|\mathbf{E}_{i\sim S}\left[f_i - p_i^*\right]\right| \leq \alpha$, for some small $\alpha \geq 0$. This means that the expectation of f and p^* over S are almost identical. Calibration strengthens this requirement by essentially asking that for any partic-

[1] Computer Science Department, Stanford University, Stanford, CA [2] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. Correspondence to: Michael P. Kim <mpk@cs.stanford.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

ular value v, if we let $S_v = \{i \in S : f_i = v\}$ be the subset of S of individuals with predicted probability v, then $\left|\mathbf{E}_{i \sim S_v}\left[f_i - p_i^*\right]\right| = |v - \mathbf{E}_{i \sim S_v}[p_i^*]| \leq \alpha$.

While this notion already precludes some forms of discrimination, a principle weakness of calibration as a fairness concept is that the guarantees are too coarse. Indeed, weaknesses of group fairness notions were discussed in (Dwork et al., 2012), as a motivation for introducing an individual fairness notion. A specific way to discriminate while satisfying calibration is to assign every member of S the value $\mathbf{E}_{i \sim S}[p_i^*]$. While being perfectly calibrated over S, the qualified members of S with large values p_i^* will be hurt.

Calibration is typically applied to large, often disjoint, sets of protected groups; that is, the guarantees are only required to hold on average over a population defined by a small number of sensitive attributes, like race or gender. A stronger definition of fairness would ensure that the predictions on *every* subpopulation would be calibrated, including, for instance, the qualified members of S from the example above. The problem with such a notion is that it is information-theoretically unattainable from a small sample of labeled examples, as it essentially requires perfect predictions. As such, we need an intermediary definition that balances the desire to protect important subgroups and the information bottleneck that arises when learning from a small sample.

To motivate our notion, suppose a learning algorithm produces a predictor f. Then, more outcomes are determined, and an auditor finds a subpopulation S whose outcomes outperform the predictions made by f. Perhaps the learning algorithm was lazy and neglected to identify the higher potential in S? Perhaps the individuals of S were simply lucky? How can we tell? To answer these questions, we take the following perspective: on the one hand, we can only expect a learner to produce a predictor that is calibrated on sets that could have been identified *efficiently* from the data at hand; on the other hand, we expect the learner to produce a predictor that is calibrated on *every* efficiently-identifiable subset. This motivates our definition of *multicalibration*, which loosely says: "A predictor f is multicalibrated with respect to a family of subpopulations $\mathcal{C}$ if it is calibrated with respect to every $S \in \mathcal{C}$."

In a nutshell, multicalibration guarantees highly-accurate predictions for every subpopulation of individuals identified by a specified collection $\mathcal{C}$ of subpopulations of individuals. While our results can be applied to any set system $\mathcal{C}$, typically, we will think of $\mathcal{C}$ as a collection of subsets where set membership can be determined *efficiently* – for instance, subpopulations defined by the conjunctions of a small number of boolean features or by small decision trees. In this sense, we can take $\mathcal{C}$ to be sets identified by a class of bounded computations. As we increase the expressiveness of $\mathcal{C}$, the fairness guarantee becomes stronger; no subpopulation that can be identified within the class will be overlooked.

In the mortgage repayment example above, if the qualified members of S can be identified by some computation $c \in \mathcal{C}$, then the resulting predictor cannot ignore the variance within S. We emphasize that the class $\mathcal{C}$ can be quite rich and, in particular, can contain many overlapping subgroups of a protected group S. In this sense, multicalibration goes far beyond calibration for a handful of sensitive groups, providing calibration for all *computationally-identifiable* subsets, where the notion of computational-identifiability is parameterized by the expressiveness of $\mathcal{C}$.

1.1. Our Contributions

We investigate the new notion of *multicalibration* from an algorithmic and complexity theoretic perspective. We present a simple, general-purpose algorithm for learning a predictor from a small set of labeled examples that is multicalibrated with respect to *any* given class $\mathcal{C}$. The algorithm is an iterative method, similar to boosting, that can be viewed as a form of functional gradient descent. A number of subtleties arise when learning a multicalibrated predictor due to the fact that the calibration constraints change based on the current set of predictions made by the predictor. To guarantee generalization from a small sample of training examples, we leverage results from a new line of work connecting differential privacy to robust adaptive data analysis (Dwork et al., 2015a;b;c; Bassily et al., 2016).

We place no explicit restrictions on the hypothesis class of the learned predictor; instead, we show that *implicitly* our algorithm learns a model that provably generalizes well to unseen data, which may be of independent interest. We demonstrate this implicit generalization by showing the predictions we learn are *compressible*, in a sense similar to decomposition lemmas from pseudorandomness (Trevisan et al., 2009). In the language of circuit complexity, we show that we can build a circuit, only slightly larger than the circuits from $\mathcal{C}$, that implements the learned predictor. As a corollary, the learned predictor is efficient in both space to represent and time to evaluate.

We also study the *computational complexity* of learning multicalibrated predictors for structured classes $\mathcal{C}$. We show a strong connection between the complexity of learning a multicalibrated predictor and agnostic learning (Haussler, 1992; Kearns et al., 1994). In the positive direction, if there is an efficient (weak) agnostic learner (Kalai et al., 2008; Feldman, 2010) for a class $\mathcal{C}$, then we can achieve similarly efficient multicalibration over $\mathcal{C}$. In the other direction, we show that learning a multicalibrated predictor on all sets defined by $\mathcal{C}$ is as hard as weak agnostic learning $\mathcal{C}$. In this sense, the complexity of learning a multicalibrated predictor with respect to a class $\mathcal{C}$ is *equivalent* to the complexity of weak agnostic learning $\mathcal{C}$.

Finally, we demonstrate that the goal of multicalbration is aligned with the goal of achieving high-utility predictions. In particular, given any predictor h, we can post-process h to obtain a multicalibrated predictor f whose squared error is no worse than that of h. The complexity of evaluating the predictor f is only slightly larger than that of h. In this sense, unlike many fairness notions, multicalibration is not at odds with predictive power and can be paired with any predictive model at essentially no cost to its accuracy.

2. Multicalibration Preliminaries

Let $\mathcal{X}$ denote the domain of (feature vectors of) individuals; we wish to predict whether some event will occur for each individual. For each $i \in \mathcal{X}$, we assume there is some unknown probability $p_i^* \in [0,1]$; we make no assumptions on the structure of $p^* : \mathcal{X} \to [0,1]$. In particular, we assume that there is enough uncertainty in the outcomes that it may be hard to learn p^* directly. Let $\mathcal{D}$ denote the distribution over individuals, supported on $\mathcal{X}$; for $S \subseteq \mathcal{X}$, let $i \sim S$ denote a sample drawn from $\mathcal{D}$ conditioned on membership in S.[1] In our learning setting, the algorithm has access to a small number of labeled individuals $D \subseteq \mathcal{X}$, where for each $i \in D$, the label is the outcome $o_i \sim \mathrm{Ber}(p_i^*)$ of an independent Bernoulli trial. Given these samples, the learner aims to produce a predictor $f : \mathcal{X} \to [0,1]$ that achieves multicalibration, described formally next.

Multicalibration. The most basic property we might hope for from a predictor is unbiasedness, i.e. that the predictions are accurate in expectation.

Definition (Accuracy in expectation). *For any $\alpha > 0$ and $S \subseteq \mathcal{X}$, a predictor f is* α-accurate-in-expectation (AE) *with respect to S if*

$$\left| \mathop{\mathbf{E}}_{i \sim S}[f_i - p_i^*] \right| \leq \alpha. \tag{1}$$

While this condition is necessary to achieve unbiased predictions, it is not sufficient to prevent all forms of discrimination; in particular, a predictor can be unbiased on a set S while introducing variance that is not borne out in the data, artificially treating similar individuals differently. Calibration mitigates this form of discrimination by considering the expected values over categories $S_v = \{i : f_i = v\}$ defined by the predictor f. Specifically, α-calibration with respect to S requires that for all but an α-fraction of a set S, the average of the true probabilities of the individuals receiving prediction v is α-close to v.

Definition (Calibration). *For any $v \in [0,1]$, $S \subseteq \mathcal{X}$, and predictor f, let $S_v = \{i : f_i = v\}$. For $\alpha \in [0,1]$, f is* α-calibrated *with respect to S if there exists some $S' \subseteq S$ with $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S'] \geq (1-\alpha) \cdot \mathbf{Pr}_{i\sim\mathcal{D}}[i \in S]$ such that for all $v \in [0,1]$,*

$$\left| \mathop{\mathbf{E}}_{i \sim S_v \cap S'}[f_i - p_i^*] \right| \leq \alpha. \tag{2}$$

Note that α-calibration with respect to S implies 2α-AE with respect to S. Our definition only requires the notion of calibration to hold on a $(1-\alpha)$-fraction of each S; this is for technical reasons due to learning from a small sample and needing to discretize the range $[0,1]$ of the learned predictor.

For a collection of subsets $\mathcal{C}$, we say that a predictor is $(\mathcal{C},\alpha)$-multicalibrated if it is α-calibrated simultaneously on all $S \in \mathcal{C}$.

Definition (Multicalibration). *Let $\mathcal{C} \subseteq 2^{\mathcal{X}}$ be a collection of subsets of $\mathcal{X}$ and $\alpha \in [0,1]$. A predictor f is $(\mathcal{C},\alpha)$-*multicalibrated *if for all $S \in \mathcal{C}$, f is α-calibrated with respect to S.*

Discretization. Even though α-calibration is a meaningful definition if we allow for arbitrary predictions $f_i \in [0,1]$, computationally, we need to maintain some discretization on the values $v \in [0,1]$. Formally, we will use the following technical definition.

Definition (λ-discretization). *Let $\lambda > 0$. The* λ-discretization *of $[0,1]$, denoted by $\Lambda[0,1] = \left\{\frac{\lambda}{2}, \frac{3\lambda}{2}, \ldots, 1 - \frac{\lambda}{2}\right\}$, is the set of $1/\lambda$ evenly spaced real values over $[0,1]$. For $v \in \Lambda[0,1]$, let*

$$\lambda(v) = [v - \lambda/2, v + \lambda/2)$$

be the λ-interval centered around v (except for the final interval, which will be $[1-\lambda, 1]$).

If we take $\lambda = \alpha$, then the λ-discretization of a $(\mathcal{C},\alpha)$-multicalibrated predictor will be $(\mathcal{C}, 2\alpha)$-multicalibrated.

In what follows, we give an overview of our results and a flavor of the proof techniques. We defer complete coverage of the results and formal proofs to the archival version (Hébert-Johnson et al., 2017).

3. Learning Multicalibrated Predictors

The first question to address is whether multicalibration is feasible. For instance, it could be the case that the requirements of multicalibration are so strong that they would require learning and representing an arbitrarily complex function p^* very precisely, which can be infeasible in our setting. Our first result characterizes the complexity of representing a multicalbrated predictor. We demonstrate that

[1] We remark that in order to guarantee a meaningful notion of fairness, we assume that the subpopulations we wish to protect are sufficiently represented in the distribution $\mathcal{D}$, in order to see these populations in a random sample. Understanding how much representation is necessary in practice remains an interesting question for future empirical investigations.

multicalibration, indeed, can be achieved efficiently: for any p^* and any collection of large subsets $\mathcal{C}$, there exists a predictor that is α-multicalibrated on $\mathcal{C}$, whose complexity is only slightly larger than the complexity required to describe the sets of $\mathcal{C}$. For concreteness, we use circuit size as our measure of complexity in the following theorem.

Theorem 1. *Suppose $\mathcal{C} \subseteq 2^{\mathcal{X}}$ is collection of sets where for $S \in \mathcal{C}$, there is a circuit of size s that computes membership in S and $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S] \geq \gamma$. For any $p^* : \mathcal{X} \to [0,1]$, there is a predictor that is $(\mathcal{C},\alpha)$-multicalibrated implemented by a circuit of size $O(s/\alpha^4\gamma)$.*

3.1. The Algorithm

In fact, we prove Theorem 1 algorithmically by learning $(\mathcal{C},\alpha)$-multicalibrated predictors from labeled samples. Our algorithm is an iterative procedure. At a high level, the algorithm maintains a candidate predictor f, and at each iteration, corrects the candidate values of some subset that violates calibration until the candidate predictor is α-calibrated on every $S \in \mathcal{C}$. We show that even if $\mathcal{C}$ is very large (e.g. exponential in the other relevant parameters), the number of updates we make and thus, the complexity of the learned model is bounded (polynomially in $1/\alpha$, $1/\gamma$).

Recall that calibration over a set S requires that on the subsets $S_v = \{i \in S : f_i = v\}$ (which we will refer to throughout as *categories*), the expected value of the true probabilities $\mathbf{E}_{i\sim S_v}[p_i^*]$ on this set is close to v. As such, the algorithm is easiest to describe in the statistical query model, where we query for estimates of the true statistics on subsets of the population and update the predictor based on these estimates. In particular, given a statistical query oracle that guarantees tolerance $\omega = O(\alpha\gamma)$, the estimates will be accurate enough to guarantee α-calibration on sets S with such that $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S] \geq \gamma$.

Adaptive Generalization. When we turn to adapting the algorithm to learn from random samples, the algorithm answers these statistical queries using the empirical estimates on some random sample from the population. Standard uniform convergence arguments (Kearns & Vazirani, 1994) show that if the set of queries we might ask is fixed in advance, then we could bound the sample complexity needed to answer these *non-adaptive* queries as $\tilde{O}(\log|\mathcal{C}|/\omega^2)$. Note, however, that the categories S_v whose expectations we query are selected *adaptively* (i.e. with dependence on the results of prior queries). In particular, the definition of the categories S_v depends on the current values of the predictor f; thus, when we update f based on the result of a statistical query, the set of categories on which we might ask a statistical query changes. In this case, we cannot simply apply concentration inequalities and take a union bound to guarantee good generalization without resampling every time we update the predictor.

To avoid this blow-up in sample complexity, we appeal to recently-uncovered connections between differential privacy and adaptive data analysis developed in (Dwork et al., 2015a;b;c; Bassily et al., 2016). To answer the statistical queries, our algorithm deliberately interacts with the data through a so-called guess-and-check oracle. In particular, each time the algorithm needs to know the value of a statistical query on a set S, rather than asking the query directly, we require that the algorithm submit its current guess $f_S = \mathbf{E}_{i\sim S}[f_i]$ to the oracle, as well as an acceptable relative error bound $\alpha \in [0,1]$. Intuitively, if the algorithm's guess is far from the window centered around the true expectation, then the oracle will respond with the answer to a statistical query with tolerance $\alpha \cdot \mathbf{Pr}_{i\sim\mathcal{D}}[i \in S]$. If, however, the guess is sufficiently close to the true value, then the oracle responds with ✓ to indicate that the current guess is close to the expectation, without revealing another answer.

Definition (Guess-and-check oracle)**.** *Let $\tilde{q} : 2^{\mathcal{X}} \times [0,1] \times [0,1] \to [0,1] \cup \{✓\}$. $\tilde{q}$ is a* guess-and-check oracle *if for $S \subseteq \mathcal{X}$ with $p_S = \mathbf{E}_{i\sim S}[p_i^*]$, $v \in [0,1]$, and any $\alpha > 0$, the response to $\tilde{q}(S,v,\alpha)$ satisfies the following conditions:*

- *if $|p_S - v| < 2\alpha$, then $\tilde{q}(S,v,\alpha) =$ ✓*
- *if $|p_S - v| > 4\alpha$, then $\tilde{q}(S,v,\alpha) \in [0,1]$*
- *if $\tilde{q}(S,v,\alpha) \neq$ ✓, then*

$$p_S - \alpha \leq \tilde{q}(S,v,\alpha) \leq p_S + \alpha.$$

Note that if the guess is such that $|p_S - v| \in [2\alpha, 4\alpha]$, the the oracle may respond with some α-accurate $r \in [0,1]$ or with ✓. If we have a lower bound $\omega = \mathbf{Pr}_{i\sim\mathcal{D}}[i \in S]\cdot\alpha$ on a sequence of guess-and-check queries, we can implement the queries using a statistical query oracle with tolerance $\tau \leq \omega$; the advantage of using this guess-and-check framework is that it can be implemented using tools developed for differential privacy (Hardt & Rothblum, 2010). This will in turn allow us to give an algorithm for learning $(\mathcal{C},\alpha)$-multicalibrated predictors from a small number of samples that generalizes well.

With the definition of this mechanism in place, we give a description of the procedure in Algorithm 1.

Implicit Representation of $\mathcal{C}$. While this procedure will work for any collection $\mathcal{C}$ for efficiency's sake (in the algorithm and the learned predictor), it is important that we have some implicit representation of $S \in \mathcal{C}$ – i.e. membership tests can be evaluated by a simple model like a decision tree, neural network, etc. In particular, even though the algorithm updates the predictions for all $i \in \mathcal{X}$, this update can be done implicitly by stringing together a "circuit" that tests membership, followed by the appropriate addition if the individual passes the test.

Algorithm 1 – Learning a $(\mathcal{C}, \alpha)$-multicalibrated predictor

```
Let $\alpha, \lambda > 0$ and let $\mathcal{C} \subseteq 2^{\mathcal{X}}$.
Let $\tilde{q}(\cdot,\cdot,\cdot)$ be a guess-and-check oracle.

• Initialize:   $f = (1/2, \ldots, 1/2) \in [0,1]^{\mathcal{X}}$

• Repeat:
    ○ For each $S \in \mathcal{C}$ and $v \in \Lambda[0,1]$:
        – Let $S_v = S \cap \{i : f_i = \lambda(v)\}$
        – if $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S_v] < \alpha\lambda \cdot \mathbf{Pr}_{i\sim\mathcal{D}}[i \in S]$:
          continue
        – Let $\bar{v} = \mathbf{E}_{i\sim S_v}[f_i]$
        – Let $r = \tilde{q}(S_v, \bar{v}, \alpha/4)$
        – If $r \neq \checkmark$:
          update $f_i \leftarrow f_i + (r - \bar{v})$ for all $i \in S_v$
          (project onto $[0,1]$ if necessary)
    ○ If no $S_v$ updated:  exit

• For $v \in \Lambda[0,1]$:
    ○ Let $\bar{v} = \mathbf{E}_{i\sim\lambda(v)}[f_i]$
    ○ For $i \in \lambda(v)$: $f_i \leftarrow \bar{v}$

• Output $f$
```

Formally, we prove the following theorem.

Theorem 2. *Suppose $\mathcal{C} \subseteq 2^{\mathcal{X}}$ is collection of sets such that for all $S \in \mathcal{C}$, $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S] \geq \gamma$, and suppose set membership can be evaluated in time t. Then Algorithm 1 run with $\lambda = \alpha$ learns a predictor of $f : \mathcal{X} \to [0,1]$ that is $(\mathcal{C}, 2\alpha)$-multicalibrated for p^* from $O(\log(|\mathcal{C}|)/\alpha^{11/2}\gamma^{3/2})$ samples in time $O(|\mathcal{C}| \cdot t \cdot \text{poly}(1/\alpha, 1/\gamma))$.*

4. Multicalibration and Weak Agnostic Learning

Observing the linear dependence in the running time on $|\mathcal{C}|$, it is natural to try to develop a learning procedure with subpolynomial, or even polylogarithmic, dependence on $|\mathcal{C}|$. Our next results aim to characterize when this optimistic goal is possible – and when it is not. We emphasize that the algorithm of Theorem 2 learns a multicalibrated predictor for *arbitrary* $p^* : \mathcal{X} \to [0,1]$ and $\mathcal{C}$. In the setting where we cannot exploit structure in p^* to learn efficiently, we might hope to exploit structure, if it exists, in the collection of subsets $\mathcal{C}$. Indeed, we demonstrate a connection between our goal of learning a multicalibrated predictor and weak agnostic learning, introduced in the literature on agnostic boosting (Ben-David et al., 2001; Kalai et al., 2008; Kanade & Kalai, 2009; Feldman, 2010). More formally, we require a (ρ, τ)-weak agnostic learner in the sense first introduced by (Kalai et al., 2008) and generalized by (Feldman, 2010). We describe the distribution-specific learner of (Feldman, 2010), where the samples and inner product in the definition are taken over the fixed data distribution $\mathcal{D}$.

Definition (Weak agnostic learner)**.** *Let $\rho \geq \tau > 0$, $\mathcal{C} \subseteq 2^{\mathcal{X}}$, and $\mathcal{H} \subseteq [-1,1]^{\mathcal{X}}$. A (ρ, τ)-weak agnostic learner $\mathcal{L}$ for a concept class $\mathcal{C}$ with hypothesis class $\mathcal{H}$ solves the following promise problem: given a collection of labeled samples $\{(i, y_i)\}$ where $i \sim \mathcal{D}$ and $y_i \in [-1,1]$, if there is some $c \in \mathcal{C}$ such that $\langle c, y \rangle > \rho$, then $\mathcal{L}$ returns some $h \in \mathcal{H}$ such that $\langle h, y \rangle > \tau$.*

Intuitively, if there is a concept $c \in \mathcal{C}$ that correlates nontrivially with the observed labels, then the weak agnostic learner returns a hypothesis h (not necessarily from $\mathcal{C}$), that is also nontrivially correlated with the observed labels. In particular, ρ and τ are typically taken to be $\rho = 1/p(d)$ and $\tau = 1/q(d)$ for polynomials $p(d) \leq q(d)$, where $d = \log(|\mathcal{C}|)$.

Efficient Multicalibration from Agnostic Learning. Our next result shows that efficient weak agnostic learning over $\mathcal{C}$ implies efficient learning of α-multicalibrated predictors on $\mathcal{C}$.

Theorem 3. *Let $\rho, \tau > 0$ and $\mathcal{C} \subseteq 2^{\mathcal{X}}$ be some concept class. If $\mathcal{C}$ admits a (ρ, τ)-weak agnostic learner that runs in time $T(|\mathcal{C}|, \rho, \tau)$, then there is an algorithm that learns a predictor that is $(\mathcal{C}, \alpha)$-multicalibrated on $\mathcal{C}' = \{S \in \mathcal{C} : \mathbf{Pr}_{i\sim\mathcal{D}}[i \in S] \geq \gamma\}$ in time $O(T(|\mathcal{C}|, \rho, \tau) \cdot \text{poly}(1/\alpha, 1/\lambda, 1/\gamma))$ as long as $\rho \leq \alpha^2\lambda\gamma/2$ and $\tau = \text{poly}(\alpha, \lambda, \gamma)$.*

Recall, in our algorithm for learning multicalibrated predictors, we maintain a candidate predictor f, and iteratively search for some set $S \in \mathcal{C}$ on which f is not calibrated. To solve this search problem more quickly, we frame the search as weak agnostic learning over a concept class derived from $\mathcal{C}$ and over the hypothesis class of $\mathcal{H} = \{h : \mathcal{X} \to [-1,1]\}$.

Specifically, consider the concept class defined by the collection of subsets $\mathcal{C}$, where for each $S \in \mathcal{C}$, we include the concept $c_S : \mathcal{X} \to \{-1,1\}$ where $c_S(i) = 1$ if and only if $i \in S$. We show how to design a "labeling" $\ell : \mathcal{X} \to [-1,1]$ for individuals such that if f violates the calibration constraint on any $S \in \mathcal{C}$, then the concept c_S correlates nontrivially with the labels over the distribution of individuals, i.e. $\langle c_S, \ell \rangle \geq \rho$ for some $\rho > 0$. Specifically, we will consider for each $v \in \Lambda[0,1]$, the following learning problem. For $i \in \mathcal{X}_v$, let $\ell_i = \frac{f_i - o_i}{2}$. For $i \in \mathcal{X} \setminus \mathcal{X}_v$, let $\ell_i = 0$. We claim that if there is some S_v currently in violation of multicalibration, then for $i \sim \mathcal{D}$, the labeled samples of either (i, ℓ_i) or $(i, -\ell_i)$ satisfy the weak learning promise for $\rho = \alpha\beta/2$.

Thus, if f is not yet multicalibrated on $\mathcal{C}$, then we are promised that there is some concept c_S with nontrivial correlation with the labels; we observe that this promise is exactly the requirement for a weak agnostic learner, as defined in (Kalai et al., 2008; Feldman, 2010). In particular, given labeled samples $(i, \ell(i))$ sampled according to $\mathcal{D}$, if there is a concept c_S with correlation at least ρ with ℓ, then the weak agnostic learner returns a hypothesis h that is τ correlated with ℓ for some $\tau < \rho$. The catch is that this hypothesis may not be in our concept class $\mathcal{C}$, so we cannot directly "correct" any $S \in \mathcal{C}$. Nevertheless, the labeling on individuals ℓ is designed such that given the hypothesis h, we can still extract an update to f that will make global progress towards the goal of attaining calibration. As long as τ is nontrvially lower bounded, we can upper bound the number of calls we need to make to the weak learner.

Efficient Agnostic Learning from Multicalibration. Our results so far show that under the right structural assumptions on p^* or on $\mathcal{C}$, a multicalibrated predictor may be learned more efficiently than our upper bound for the general case. Returning to the general case, we may wonder if these structural assumptions are necessary; we answer this question in the positive. We show that for worst-case p^* learning a multicalibrated predictor on $\mathcal{C}$ is as hard as weak agnostic learning for the class $\mathcal{C}$.

Theorem 4. *Let $\alpha, \gamma > 0$ and suppose $\mathcal{C} \subseteq 2^{\mathcal{X}}$ is a concept class. If there is an algorithm for learning a $(\mathcal{C}', \alpha)$-multicalibrated predictor on $\mathcal{C}' = \{S \in \mathcal{C} : \mathbf{Pr}_{i\sim\mathcal{D}}[i \in S] \geq \gamma\}$ in time $T(|C|, \alpha, \gamma)$ then we can implement a (ρ, τ)-weak agnostic learner for $\mathcal{C}$ in time $O(T(|C|, \alpha, \gamma) \cdot \mathrm{poly}(1/\tau))$ for any $\rho, \tau > 0$ such that $\tau \leq \min\{\rho - 2\gamma, \rho/4 - 4\alpha\}$.*

Specifically, we show how to implement a weak agnostic learner for $\mathcal{C}$, given an algorithm to learn an α-multicalibrated predictor f with respect to $\mathcal{C}$ (in fact, we only need the predictor to be multicalibrated on $\mathcal{C}' = \{S \in \mathcal{C} : \mathbf{Pr}_{i\sim\mathcal{D}}[i \in S] \geq \gamma\}$). The key lemma for this reduction says that if there is some $c \in \mathcal{C}$ that is nontrivially correlated with the labels, then f is also nontrivially correlated with c. In general, agnostic learning is considered a notoriously hard computational problem. In particular, under cryptographic assumptions (Valiant, 1984; Goldreich et al., 1984; Bogdanov & Rosen, 2017), this result implies that there is some constant $t > 0$, such that any algorithm that learns a $(\mathcal{C}, \alpha)$-multicalibrated predictor requires $\Omega(|\mathcal{C}|^t)$ time for arbitrary $\mathcal{C}$.

In combination, these results show that the complexity of learning a multicalibrated predictor with respect to a class $\mathcal{C}$ is equivalent to the complexity of weak agnostic learning $\mathcal{C}$.

5. Best-in-class Predictions

Finally, we return our attention to investigating the utility of multicalibrated predictors. Above, we have argued that multicalibration provides a strong protection of groups against discrimination. We show that this protection comes at (next to) no cost in the utility of the predictor. This result adds to the growing literature on fairness-accuracy trade-offs (Fish et al., 2016; Berk et al., 2017; Chouldechova & G'Sell, 2017).

Theorem 5. *Suppose $\mathcal{C} \subseteq 2^{\mathcal{X}}$ is a collection of subsets of $\mathcal{X}$ and $\mathcal{H}$ is a set of predictors. There is a predictor f that is α-multicalibrated on $\mathcal{C}$ such that*

$$\mathop{\mathbf{E}}_{i\sim\mathcal{X}}[(f_i - p_i^*)^2] - \mathop{\mathbf{E}}_{i\sim\mathcal{X}}[(h_i^* - p_i^*)^2] < 6\alpha,$$

where $h^* = \mathrm{argmin}_{h\in\mathcal{H}} \mathbf{E}_{i\sim\mathcal{X}}[(h - p^*)^2]$. Further, suppose that for all $S \in \mathcal{C}$, $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S] \geq \gamma$, and suppose that set membership for $S \in \mathcal{C}$ and $h \in \mathcal{H}$ are computable by circuits of size at most s; then f is computable by a circuit of size at most $O(s/\alpha^4\gamma)$.

We can interpret Theorem 5 in different ways based on the choice of $\mathcal{H}$. Suppose there is some sophisticated learning algorithm that produces some predictor h that obtains exceptional performance, but may violate calibration arbitrarily. If we take $\mathcal{H} = \{h\}$, then this result says: enforcing calibration on h after learning does not hurt the accuracy by much.

Taking a different perspective, we can also think of $\mathcal{H}$ as a set of predictors that, say, are implemented by a circuit class of bounded complexity (e.g. conjunctions of k variables, halfspaces, circuits of size s). Leveraging Theorem 1 and Theorem 2, this theorem shows that for any such class of predictors $\mathcal{H}$ of bounded complexity, there exists a multicalibrated predictor with similar complexity that performs as well as the best $h^* \in \mathcal{H}$. In this sense, with just a slight overhead in complexity, multicalibrated predictors can achieve "best-in-class" predictions.

In contrast to many other notions of fairness, multicalibration does not limit the utility of a predictor. Further, to prove that multicalibration does not negatively impact the utility, we in fact, show a much stronger statement: if applying multicalibration to some $h \in \mathcal{H}$ changes the predictions of h significantly (i.e. if $\mathbf{E}_{i\sim\mathcal{D}}[(f_i - h_i)^2]$ is large), then this change represents an improvement in squared error. In this sense, requiring multicalibration is aligned with the goals of learning a high-utility predictor.

We give a flavor of our approach to proving Theorem 5. Consider some $h \in \mathcal{H}$ and consider the partition of $\mathcal{X}$ into sets according to the predictions of h – in particular, we will first apply a λ-discretization to the range of each h to partition $\mathcal{X}$ into categories. That is, let $S_v(h) = \{i : h_i \in \lambda(v)\}$, and note that $S_v(h)$ is disjoint from $S_{v'}(h)$ for $v \neq v'$, and $\bigcup_{v\in\Lambda[0,1]} S_v(h) = \mathcal{X}$. In addition to calibrating with respect to $S \in \mathcal{C}$, we can also ask for calibration on $S_v(h)$ for all $h \in \mathcal{H}$ and $v \in \Lambda[0, 1]$. Specifically, let

$\mathcal{S}(\mathcal{H}) = \{S_v(h)\}_{h\in\mathcal{H}, v\in\Lambda[0,1]}$; we consider imposing calibration on $\mathcal{C} \cup \mathcal{S}(\mathcal{H})$. Calibrating in this manner protects the groups defined by $\mathcal{C}$ but additionally gives a strong utility guarantee, captured by the following lemma.

Lemma. *Suppose g is a λ-discretized predictor and let $\mathcal{S}(g) = \{S_v(g)\}_{v\in\Lambda[0,1]}$. Suppose f is an arbitrary $(\mathcal{S}(g), \alpha)$-multicalibrated predictor. Then for $v \in \Lambda[0,1]$,*

$$\mathop{\mathbf{E}}_{i\sim S_v(g)}\left[(g_i - f_i)^2\right] - (4\alpha + \lambda) \\ \leq \mathop{\mathbf{E}}_{i\sim S_v(g)}\left[(g_i - p_i^*)^2\right] - \mathop{\mathbf{E}}_{i\sim S_v(g)}\left[(f_i - p_i^*)^2\right].$$

This lemma shows that calibrating on the categories of a predictor not only prevents the squared prediction error from degrading beyond a small additive approximation, but it also guarantees that if calibrating changes the predictor significantly on any category, this change represents significant progress towards the true underlying probabilities on this category. Assuming Lemma 5, Theorem 5 follows.

Note that Lemma 5 shows that this best-in-class property holds not just over the entire domain $\mathcal{X}$, but on every sufficiently large category $S_v(h)$ identified by some $h \in \mathcal{H}$. That is, if f is calibrated on $\mathcal{S}(\mathcal{H})$, then for every category $S_v(h)$, the average squared prediction error $\mathbf{E}_{i\sim S_v(h)}\left[(f_i - p_i^*)^2\right]$ will be at most 6α worse than prediction given by h on this set. If we view $\mathcal{H}$ as defining a set $\mathcal{S}(\mathcal{H})$ of "computationally-identifiable" categories, then we can view any predictor that is calibrated on $\mathcal{S}(\mathcal{H})$ as at least as fair and at least as accurate on this set of computationally-identifiable categories as the predictor that identified the group (up to some small additive approximation).

6. Related Works and Discussion

Calibration. Calibration is a well-studied concept in the literature on statistics and econometrics, particularly forecasting. For a background on calibration in this context, see (Sandroni et al., 2003; Foster & Hart, 2015) and the references therein. Calibration has also been studied in the context of structured predictions where the supported set of predictions is large (Kuleshov & Liang, 2015). Our algorithmic result for multicalibration bears similarity to works from the online learning literature (Blum & Mansour, 2007; Khot & Ponnuswami, 2008; Trevisan et al., 2009). While these works are similar in spirit, none of the algorithmic results apply directly to our setting of multicalibration. We are unaware of prior works drawing connections between calibration and differential privacy / adaptive data analysis.

Parity and Balance. Other works on fairness in classification tend to look at parity-based notions of fairness. Specifically, the notion of statistical parity (Dwork et al., 2012) and balanced error rates (Hardt et al., 2016) aim to enforce some notion of equal treatment across groups of individuals defined by sensitive features, like race, gender, etc. In (Hardt et al., 2016) it is shown how to obtain equalized odds, a definition related to error-rate balance, as a post-processing step of "correcting" any predictor.

While both calibration and balance (as well as other related variants) intuitively seem like good properties to expect in a fair predictor (even if they are a bit weak), it is impossible to satisfy both notions simultaneously (in non-degenerate cases) (Kleinberg et al., 2017; Chouldechova, 2017; Pleiss et al., 2017), and there is much debate about how to proceed given this incompatibility (Corbett-Davies et al., 2017). The inherent conflict between balance and calibration, combined with our observation that calibration is always aligned with the goal of accurate high-utility predictions, implies that at times, balance must be at odds with obtaining predictive utility. In this work, we strengthen the protections implied by calibration, rather than enforcing error-rate balance. While there are certainly contexts in which "equalizing the odds" across groups is a good idea, there are also contexts where calibration is a more appropriate notion of fairness.

One particular critique of balanced error rates as a fairness notion is that given two populations $S, T \subseteq \mathcal{X}$ with different base rates (i.e. $p_i^* > p_j^*$ for $i \in S, j \in T$), the Bayes Optimal predictor p^* will not be balanced. That is, even given access to perfect information about the underlying probabilities, the stochasticity in the outcomes will lead to different false positive and false negative rates. In this sense, balance can be viewed as an *a posteriori* notion of fairness (fairness with respect to outcomes), while our notion of multicalibration is an *a priori* notion of fairness (fairness with respect to given data). In a prediction setting where, given the data, there is still significant uncertainty in the outcome, we feel that multicalibration should be considered as an alternative to balanced error rates. That said, a serious form of discrimination could arise if the uncertainty in outcomes is very different across different subpopulations; this would be a form of *information-theoretic* discrimination that multicalibration could help to identify, but could not remedy directly.

Between Populations and Individuals. Most fairness notions are statistical in nature; roughly, these definitions – including statistical parity (Dwork et al., 2012), balanced error-rates (Hardt et al., 2016), and calibration – say that treatment across groups should be equitable *on-average* (for different notions equitable). In a notable work, (Dwork et al., 2012) critique these broad-strokes statistical definitions and propose an individual notion of fairness, which aims to "treat similar individuals similarly". A key challenge to this approach is that it assumes access to a task-specific metric for *every* pair of individuals. In the practical setting, where we want to learn from a small sample, we cannot hope to achieve such an information-theoretic notion of fairness. One can view multicalibration as a meaningful compromise between group fairness (satisfying calibration) and individual-calibration (closely matching p_i^*). The multicalibration framework presented in this work inspired subsequent work investigating how to interpolate between statistical and individual notions of "metric fairness" for general similarity metrics (Kim et al., 2018b), as well as further theoretical and empirical investigations of multi-accuracy-in-expectation in the context of binary classification (Kim et al., 2018a).

Contemporary independent work of (Kearns et al., 2017) also investigates strengthening the guarantees of notions of group fairness by requiring that these properties hold for a much richer collection of sets. Unlike our work, their definitions require balance or statistical parity on these collection of sets. Despite similar motivations, the two approaches to subgroup fairness differ in substantial ways. As a concrete example, multicalibration is aligned with the incentives of achieving high-utility predictors; this is not necessarily the case with balance-based notions of fairness. Indeed, in the setting considered in this work, one of the motivations for multicalibration is the earlier critique of balance that may only be heightened when considering "multi-balance".

Consider the example from (Dwork et al., 2012) where we wish to predict future success in school. In a population S, the strongest students apply to Engineering whereas in the general population T, they apply to Business. Enforcing balance between the Business applicants and Engineering applicants within both groups would be unfair to qualified applicants in both groups (i.e. the Engineering students of S and the Business students of T). Essentially, carving up the space of individuals into subgroups exaggerates the differences in the base rates, which leads to mistreatment. Preventing discrimination by algorithms is subtle, and different scenarios will call for different notions of protection. Still, these works collectively validate the need to investigate attainable approaches to mitigating discrimination beyond large protected groups.

Corrective Discrimination Multicalibration represents a powerful tool to address a certain form of discrimination, but it is not universally-applicable. Consider the mortgage example again: perhaps the number of members of S that received loans in the past is small (and thus there are too few examples for fine-grained learning within S); perhaps the attributes are too limited to identify the qualified members of S (taking this point to the extreme, perhaps the only available attribute is membership in S). In these cases, the data may be insufficient for multicalibration to provide meaningful guarantees. Further, even if the algorithm was given access to unlimited rich data such that refined values of p^* could be recovered, there are situations where preferential treatment may be in order: after all, the salaries of members of S may be lower due to historical discrimination. For these reasons, the concern that balance is inconsistent with p^* could be answered with: "yes, and purposely so!" Indeed, (Hardt et al., 2016) promotes enforcing a equalized odds as a form of "corrective discrimination." While this type of advocacy is important in many settings, multicalibration represents a different addition to the quiver of anti-discrimination measures, which we also believe is natural and desirable in many settings.

Consider another example where multicalibration is appropriate, but equalizing error rates might not be: suppose a genomics company offers individuals a prediction of their likelihood of developing certain genetic disorders. These disorders have different rates across different populations; e.g., Tay-Sachs disease is rare in the general population, but occurs much more frequently in the Ashkenazi population. We certainly do not want to enforce balance on the Ashkenazi population by down-weighting the prediction that individuals would have Tay-Sachs (as they are endogenously more likely to have the disease). However, we also don't want the company to base its prediction solely on the Ashkenazi feature. Instead, enforcing multicalibration would require that the learning algorithm investigate both the Ashkenazi and non-Ashkenazi population to predict accurately in each group (even if this means a higher false positive rate in the Ashkenazi population). In this case, relying on p^* seems to be well-aligned with promoting fairness.

Conclusion. Multicalibration addresses a specific form of discrimination that can occur in prediction systems learned from data. In particular, multicalibration requires that the learned predictor accurately reflects the "computationally-identifiable" variance present in the data, without introducing spurious variance. Multicalibration is most appropriate in settings where perfect predictions at an individual level are considered the fairest predictions, but where we do not have rich enough training data to make perfect predictions. Importantly, in this context, *there is no fairness-utility tradeoff!* Enforcing multicalibration only improves the predictive power of the resulting model. Instead, this work identifies and aims to address a "fairness-information" tradeoff; while we cannot achieve the information-theoretic ideal predictions from a small sample of training data, we show that attaining a meaningful complexity-theoretic relaxation of this goal is feasible through multicalibration. Finally, we consider the interplay between multicalibration and "corrective discrimination," such as the transformation of (Hardt et al., 2016), to be an important direction for further research.

Acknowledgments

The authors thank Cynthia Dwork, Roy Frostig, Parikshit Gopalan, Moritz Hardt, Aditi Raghunathan, Jacob Steinhardt, and Greg Valiant for helpful discussions related to this work. We thank the anonymous reviewers for their detailed feedback. MPK was supported by NSF grant CNS-122864. OR was supported by NSF grant CCF-1749750. GNR was supported by ISF grant No. 5219/17.

References

Bassily, R., Nissim, K., Smith, A., Steinke, T., Stemmer, U., and Ullman, J. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1046–1059. ACM, 2016.

Ben-David, S., Long, P., and Mansour, Y. Agnostic boosting. In *Computational Learning Theory*, pp. 507–516. Springer, 2001.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. A convex framework for fair regression. *FATML*, 2017.

Blum, A. and Mansour, Y. From external to internal regret. *Journal of Machine Learning Research*, 8(Jun): 1307–1324, 2007.

Bogdanov, A. and Rosen, A. Pseudorandom functions: Three decades later. In *Tutorials on the Foundations of Cryptography*, pp. 79–158. Springer, 2017.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pp. 4349–4357, 2016.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 2017.

Chouldechova, A. and G'Sell, M. Fairer and more accurate, but for whom? *FATML*, 2017.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. *KDD*, 2017.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, pp. 214–226, 2012.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pp. 2350–2358, 2015a.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015b.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. L. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 117–126. ACM, 2015c.

Feldman, V. Distribution-specific agnostic boosting. In *Proceedings of the First Symposium on Innovations in Computer Science10*, 2010.

Fish, B., Kun, J., and Lelkes, Á. D. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 144–152. SIAM, 2016.

Foster, D. P. and Hart, S. Smooth calibration, leaky forecasts, finite recall, and nash dynamics. Technical report, Working paper, 2015.

Goldreich, O., Goldwasser, S., and Micali, S. How to construct random functions. In *Foundations of Computer Science, 1984. 25th Annual Symposium on*, pp. 464–479. IEEE, 1984.

Hardt, M. and Rothblum, G. N. A multiplicative weights mechanism for privacy-preserving data analysis. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pp. 61–70. IEEE, 2010.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

Haussler, D. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

Hébert-Johnson, Ú., Kim, M. P., Reingold, O., and Rothblum, G. N. Calibration for the (computationally-identifiable) masses. *arXiv Preprint*, 1711.08513, 2017.

Kalai, A. T., Mansour, Y., and Verbin, E. On agnostic boosting and parity learning. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pp. 629–638. ACM, 2008.

Kanade, V. and Kalai, A. Potential-based agnostic boosting. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A.

(eds.), *Advances in Neural Information Processing Systems 22*, pp. 880–888. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3676-potential-based-agnostic-boosting.pdf.

Kearns, M., Neel, S., Roth, A., and Wu, Z. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017.

Kearns, M. J. and Vazirani, U. V. *An introduction to computational learning theory*. MIT press, 1994.

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

Khot, S. and Ponnuswami, A. K. Minimizing wide range regret with time selection functions. In *COLT*, pp. 81–86, 2008.

Kim, M. P., Ghorbani, A., and Zou, J. Multiaccuracy: Blackbox post-processing for fairness in classification. *arXiv preprint arXiv:1805.12317*, 2018a.

Kim, M. P., Reingold, O., and Rothblum, G. N. Fairness through computationally-bounded awareness. *arXiv preprint arXiv:1803.03239*, 2018b.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *ITCS*, 2017.

Kuleshov, V. and Liang, P. S. Calibrated structured prediction. In *Advances in Neural Information Processing Systems*, pp. 3474–3482, 2015.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. *NIPS*, 2017.

Sandroni, A., Smorodinsky, R., and Vohra, R. V. Calibration with many checking rules. *Mathematics of operations Research*, 28(1):141–153, 2003.

Trevisan, L., Tulsiani, M., and Vadhan, S. Regularity, boosting, and efficiently simulating every high-entropy distribution. In *Computational Complexity, 2009. CCC'09. 24th Annual IEEE Conference on*, pp. 126–136. IEEE, 2009.

Valiant, L. G. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.