# Multiaccuracy: Black-Box Post-Processing for Fairness in Classification

Michael P. Kim*† mpk@cs.stanford.edu
Amirata Ghorbani* amiratag@stanford.edu
James Zou jamesz@stanford.edu

**Abstract**

Prediction systems are successfully deployed in applications ranging from disease diagnosis, to predicting credit worthiness, to image recognition. Even when the overall accuracy is high, these systems may exhibit systematic biases that harm specific subpopulations; such biases may arise inadvertently due to underrepresentation in the data used to train a machine-learning model, or as the result of intentional malicious discrimination. We develop a rigorous framework of *multiaccuracy* auditing and post-processing to ensure accurate predictions across *identifiable subgroups*.

Our algorithm, Multiaccuracy Boost, works in any setting where we have black-box access to a predictor and a relatively small set of labeled data for auditing; importantly, this black-box framework allows for improved fairness and accountability of predictions, even when the predictor is minimally transparent. We prove that Multiaccuracy Boost converges efficiently and show that if the initial model is accurate on an identifiable subgroup, then the post-processed model will be also. We experimentally demonstrate the effectiveness of the approach to improve the accuracy among minority subgroups in diverse applications (image classification, finance, population health). Interestingly, Multiaccuracy Boost can improve subpopulation accuracy (e.g. for "black women") even when the sensitive features (e.g. "race", "gender") are not given to the algorithm explicitly.

## 1 Introduction

Despite the successes of machine learning at complex tasks that involve making predictions about people, there is growing evidence that "state-of-the-art" models can perform significantly less accurately on minority populations than on the majority population. Indeed, a notable study of three commercial face recognition systems known as the "Gender Shades" project [BG18], demonstrated significant performance gaps across different populations at classification tasks. While all systems achieved roughly 90% accuracy at gender detection on a popular benchmark, a closer investigation revealed that the system was significantly less accurate on female subjects compared to males and on dark-skinned individuals compared to light-skinned. Worse yet, this discrepancy in accuracy compounded when comparing dark-skinned females to light-skinned males; classification accuracy differed between these groups by as much as 34%! This study confirmed empirically the intuition that machine-learned classifiers may optimize predictions to perform well on the majority population, inadvertently hurting performance on the minority population in significant ways.

*These authors contributed equally.

†Supported by NSF Grant CCF-1763299.

A first approach to address this serious problem would be to update the training distribution to reflect the distribution of people, making sure historically-underrepresented populations are well-represented in the training data. While this approach may be viewed as an eventual goal, often for historical and social reasons, data from certain minority populations is less available than from the majority population. In particular, we may not immediately have enough data from these underrepresented subpopulations to train a complex model. Additionally, even when adequate representative data is available, this process necessitates retraining the underlying prediction model. In the common setting where the learned model is provided as a service, like a commercial image recognition system, there may not be sufficient incentive (financial, social, etc.) for the service provider to retrain the model. Still, the clients of the model may want to improve the accuracy of the resulting predictions across the population, even when they are not privy to the inner workings of the prediction system.

At a high level, our work focuses on a setting, adapted from [HKRR18], that is common in practice but distinct from much of the other literature on fairness in classification. We are given black-box access to a classifier, $f_0$, and a relatively small "validation set" of labeled samples drawn from some representative distribution $\mathcal{D}$; our goal is to *audit* $f_0$ to determine whether the predictor satisfies a strong notion of subgroup fairness, *multiaccuracy*. Multiaccuracy requires (in a sense that we make formal in Section 2) that predictions be unbiased, not just overall, but on every identifiable subpopulation. If auditing reveals that the predictor does not satisfy multiaccuracy, we aim to *post-process* $f_0$ to produce a new classifier $f$ that is multiaccurate, without adversely affecting the subpopulations where $f_0$ was already accurate.

Even if the initial classifier $f_0$ was trained in good faith, it may still exhibit biases on significant subpopulations when evaluated on samples from $\mathcal{D}$. This setting can arise when minority populations are underrepresented in the distribution used to train $f_0$ compared to the desired distribution $\mathcal{D}$, as in the Gender Shades study [BG18]. In general, we make no assumptions about how $f_0$ was trained. In particular, $f_0$ may be an adversarially-chosen classifier, which explicitly aims to give erroneous predictions within some protected subpopulation while satisfying marginal statistical notions of fairness. Indeed, the influential work on "Fairness Through Awareness" [DHP+12], followed by [KNRW17, HKRR18], demonstrated the weakness of statistical notions of fairness (such as statistical parity, equalized odds, and calibration), showing that prediction systems can exhibit material forms of discrimination against protected populations, even though they satisfy statistical fairness conditions. Left unaddressed, such forms of discrimination may discourage the participation of minority populations, leading to further underrepresentation of these groups. Thus, our goal will be to mitigate systematic biases broadly enough to handle inadvertent and malicious forms of discrimination.

**Our contributions** We investigate a notion of fairness – multiaccuracy – originally proposed in [HKRR18], and develop a framework for auditing and post-processing for multiaccuracy. We develop a new algorithm, Multiaccuracy Boost, where a simple learning algorithm – the auditor – is used to identify subpopulations in $\mathcal{D}$ where $f_0$ is systematically making more mistakes. This information is then used to iteratively post-process $f_0$ until the multiaccuracy condition – unbiased predictions in each identifiable subgroup – is satisfied. Our notion of multiaccuracy differs from parity-based notions of fairness, and is reasonable in settings such as gender detection where we would like to boost the classifier's accuracy across many subgroups. We prove convergence guarantees for Multiaccuracy Boost and show that post-processing for multiaccuracy may actually improve

the *overall* classification accuracy. We describe the post-processing algorithm in Section 3.

Empirically, we validate MULTIACCURACY BOOST in several different case studies: gender detection from images as in Gender Shades [BG18], a semi-synthetic medical diagnosis task, and adult income prediction. In all three cases, we use standard, initial prediction models that achieve good overall classification error but exhibit biases against significant subpopulations. After post-processing, the accuracy improves across these minority groups, even though minority-status is not explicitly given to the post-processing algorithm as a feature. As long as there are features in the audit set correlated with the (unobserved) human categories, then MULTIACCURACY BOOST is effective at improving the classification accuracy across these categories. As suggested by the theory, MULTIACCURACY BOOST actually improves the overall accuracy, by identifying subpopulations where the initial models systematically erred; further, post-processing does not significantly affect performance on groups where accuracy was already high. We show that MULTIACCURACY BOOST, which only accesses $f_0$ as a black-box, performs comparably and sometimes even better than very strong white-box alternatives which has full access to $f_0$. These results are reported in Section 4.

In Section 4.1, we explore the gender detection example further, investigating some of the practical aspects of multiaccuracy auditing and post-processing. In particular, we observe that the representation of images used for auditing (and post-processing) matters; we show that auditing is more effective when using an embedding of the images that was trained using an unsupervised autoencoder compared to using the internal representation of the neural network used for prediction. These findings seem consistent with the guiding philosophy, put forth by [DHP+12], that maintaining "awareness" is paramount to detecting unfairness. We also show that the auditing process, which we use algorithmically as a way to boost the accuracy of the classifier, can also be used to help people understand why their prediction models are making mistakes. Specifically, the output of the multiaccuracy auditor can be used to produce examples of inputs where the predictor is erring significantly; this provides human interpretation for biases of the original classifier.

## 2 Setting and multiaccuracy

**High-level setting.** Let $\mathcal{X}$ denote the input space; we denote by $y : \mathcal{X} \to \{0, 1\}$ the function that maps inputs to their label. Let $\mathcal{D}$ represent the validation data distribution supported on $\mathcal{X}$; the distribution $\mathcal{D}$ can be viewed as the "true" distribution, on which we will evaluate the accuracy of the final model. In particular, we assume that the important subpopulations are sufficiently represented on $\mathcal{D}$ (cf. Remark on data distribution). Our post-processing learner receives as input a small sample of labeled validation data $\{(x, y(x))\}$, where $x \sim \mathcal{D}$, as well as black-box access to an initial regression / classification model $f_0 : \mathcal{X} \to [0, 1]$. The goal is to output a new model (using calls to $f_0$) that satisfies the multiaccuracy fairness conditions (described below).

Importantly, we make no further assumptions about $f_0$. Typically, we will think of $f_0$ as the output of a learning algorithm, trained on some other distribution $\mathcal{D}_0$ (also supported on $\mathcal{X}$); in this scenario, our goal is to mitigate any inadvertently-learned biases. That said, another important setting assumes that $f_0$ is chosen *adversarially* to discriminate against a protected population of individuals, while aiming to appear accurate and fair on the whole; here, we aim to protect subpopulations against malicious misclassification. The formal guarantees of multiaccuracy provide meaningful protections from both of these important forms of discrimination.

**Additional Notation.** For a subset $S \subseteq \mathcal{X}$, we use $x \sim S$ to denote a sample from $\mathcal{D}$ conditioned on membership in $S$. We take the characteristic function of $S$ to be $\chi_S(x) = 1$ if $x \in S$ and 0 otherwise. For a hypothesis $f : \mathcal{X} \to [0,1]$, we denote the classification error of $f$ with respect to a subset $S \subseteq \mathcal{X}$ as $\mathrm{er}_S(f; y) = \mathbf{Pr}_{x \sim S}[\bar{f}(x) \neq y(x)]$, where $\bar{f}(x)$ rounds $f(x)$ to $\{0,1\}$. For a function $z : \mathcal{X} \to [-1,1]$ and a subset $S \subseteq \mathcal{X}$, let $z_S$ be the restriction to $S$ where $z_S(x) = z(x)$ if $x \in S$ and $z_S(x) = 0$ otherwise. We use $\ell_{\mathcal{D}}(f; y) = \mathbf{E}_{x \sim \mathcal{D}}[\ell_x(f; y)]$ to denote the expected cross-entropy loss of $f$ on $x \in \mathcal{X}$ where $\ell_x(f; y) = -y(x) \cdot \log(f(x)) - (1 - y(x)) \cdot \log(1 - f(x))$.

## 2.1 Multiaccuracy

The goal of multiaccuracy is to achieve low classification error, not just on $\mathcal{X}$ overall, but also on subpopulations of $\mathcal{X}$. This goal is formalized in the following definition adapted from [HKRR18].

**Definition** (Multiaccuracy)**.** *Let $\alpha \geq 0$ and let $\mathcal{C} \subseteq [-1,1]^{\mathcal{X}}$ be a class of functions on $\mathcal{X}$. A hypothesis $f : \mathcal{X} \to [0,1]$ is* $(\mathcal{C}, \alpha)$-multiaccurate *if for all $c \in \mathcal{C}$:*

$$\mathop{\mathbf{E}}_{x \sim \mathcal{D}}[c(x) \cdot (f(x) - y(x))] \leq \alpha. \tag{1}$$

$(\mathcal{C}, \alpha)$-multiaccuracy guarantees that a hypothesis appears unbiased according to a class of statistical tests defined by $\mathcal{C}$. As an example, we could define the class in terms of a collection of subsets $S \subseteq \mathcal{X}$, taking $\mathcal{C}$ to be $\chi_S$ (and its negation) for each subset in the collection; in this case, $(\mathcal{C}, \alpha)$-multiaccuracy guarantees that for each $S$, the predictions of $f$ are at most $\alpha$-biased.

Ideally, we would hope to take $\mathcal{C}$ to be the class of *all* statistical tests. Requiring multiaccuracy with respect to such a $\mathcal{C}$, however, requires learning the function $y(x)$ exactly, which is information-theoretically impossible from a small sample. In practice, if we take $\mathcal{C}$ to be a *learnable* class of functions, then $(\mathcal{C}, \alpha)$-multiaccuracy guarantees accuracy on all *efficiently-identifiable* subpopulations.

For instance, if we took $\mathcal{C}$ to be the class of width-4 conjunctions, then multiaccuracy guarantees unbiasedness, not just on the marginal populations defined by race and separately gender, but by the subpopulations defined by the intersection of race, gender, and two other (possibly "unprotected") features. In particular, the subpopulations that multiaccuracy protects can be overlapping and include groups beyond traditionally-protected populations. This form of computationally-bounded intersectionality provides strong protections against forms of discrimination, like subset targeting, discussed in [DHP+12, HKRR18].

## 2.2 Classification accuracy from multiaccuracy

Multiaccuracy guarantees that the predictions of a classifier appear unbiased on a rich class of subpopulations; ideally though, we would state a guarantee in terms of the classification accuracy, not just the bias. Intuitively, as we take $\mathcal{C}$ to define a richer class of tests, the guarantees of multiaccuracy become stronger. This intuition is formalized in the following proposition.

**Proposition 1.** *Let $\hat{y} : \mathcal{X} \to \{-1, 1\}$ as $\hat{y}(x) = 1 - 2y(x)$. Suppose that for $S \subseteq \mathcal{X}$ with $\mathbf{Pr}_{x \sim \mathcal{D}}[x \in S] \geq \gamma$, there is some $c \in \mathcal{C}$ such that $\mathbf{E}_{x \sim \mathcal{D}}[|c(x) - \hat{y}_S(x)|] \leq \tau$. Then if $f$ is $(\mathcal{C}, \alpha)$-multiaccurate,* $\mathrm{er}_S(f; y) \leq 2 \cdot (\alpha + \tau)/\gamma$.

That is, if there is a function in $\mathcal{C}$ that correlates well with the label function on a significant subpopulation $S$, then multi-accuracy translates into a guarantee on the *classification accuracy* on this subpopulation.

**Remark on data distribution.** Note that in our definition of multiaccuracy, we take an expectation over the distribution $\mathcal{D}$ of validation data. Ideally, $\mathcal{D}$ should reflect the true population distribution or could be aspirational, increasing the representation of populations who have experienced historical discrimination; for instance, the classification error guarantee of Proposition 1 improves as $\gamma$, the density of the protected subpopulation $S$, grows. Still, if we take the multiaccuracy error term $\alpha$ small enough, then we may still hope to improve the accuracy on less-represented subpopulations. Our experiments suggest that applying the multiaccuracy framework with an unbalanced valdiation distribution may still help improve the accuracy on underrepresented groups.

### 2.3 Auditing for multiaccuracy

With the definition of $(\mathcal{C}, \alpha)$-multiaccuracy in place, a natural question to ask is how to test if a hypothesis $f$ satisfies the definition; further, if $f$ does not satisfy $(\mathcal{C}, \alpha)$-multiaccuracy, can we update $f$ efficiently to satisfy the definition, while maintaining the overall accuracy? We will use a learning algorithm $\mathcal{A}$ to audit a classifier $f$ for multiaccuracy. The algorithm $\mathcal{A}$ receives a small sample from $\mathcal{D}$ and aims to learn a function $h$ that correlates with the *residual* function $f - y$. In Section 3, we describe how to use such an auditor to solve the post-processing problem. This connection between subpopulation fairness and learning is also made in [KNRW17,HKRR18,KRR18], albeit for different tasks.

**Definition** (Multiaccuracy auditing)**.** *Let $\alpha > 0, m \in \mathbb{N}$, and let $\mathcal{A} : (\mathcal{X} \times [-1, 1])^m \to [-1, 1]^{\mathcal{X}}$ be a learning algorithm. Suppose $D \sim \mathcal{D}^m$ is a set of independent samples. A hypothesis $f : \mathcal{X} \to [0, 1]$ passes* $(\mathcal{A}, \alpha)$-multiaccuracy auditing *if for $h = \mathcal{A}(D; f - y)$:*

$$\mathop{\mathbf{E}}_{x \sim \mathcal{D}} [h(x) \cdot (f(x) - y(x))] \leq \alpha. \tag{2}$$

A special case of $(\mathcal{A}, \alpha)$-multiaccuracy auditing uses a naive learning algorithm that iterates over statistical tests $c \in \mathcal{C}$. Concretely, in our experiments, we audit with ridge regression and decision tree regression; both auditors are effective at identifying subpopulations on which the model is underperforming. Further, in the image recognition setting, we show that auditing can be used to produce interpretable synopses of the types of mistakes a predictor makes.

## 3 Post-processing for multiaccuracy

Here, we describe an algorithm, Multiaccuracy Boost, for post-processing a pre-trained model to achieve multiaccuracy. The algorithm is given black-box access to an initial hypothesis $f_0 : \mathcal{X} \to [0, 1]$ and a learning algorithm $\mathcal{A} : (\mathcal{X} \times [-1, 1])^m \to [-1, 1]^{\mathcal{X}}$, and for any accuracy parameter $\alpha > 0$, outputs a hypothesis $f : \mathcal{X} \to [0, 1]$ that passes $(\mathcal{A}, \alpha)$-multiaccuracy auditing. The post-processing algorithm is an iterative procedure similar to boosting [FS95, SF12], that uses the multiplicative weights framework to improve suboptimal predictions identified by the auditor. This approach is similar to the algorithm given in [HKRR18] in the context of fairness and [TTV09] in the context of pseudorandomness. Importantly, we adapt these algorithms so that Multiaccuracy Boost exhibits what we call the "do-no-harm" guarantee; informally, if $f_0$ has low classification error on some subpopulation $S \subseteq \mathcal{X}$ identified by $\mathcal{A}$, then the resulting classification error on $S$ cannot increase significantly. In this sense, achieving our notion of fairness need not adversely affect the utility of the classifier.

**Algorithm 1:** MULTIACCURACY BOOST

**Given:**

- initial hypothesis $f_0 : \mathcal{X} \to [0,1]$
- auditing algorithm $\mathcal{A} : (\mathcal{X} \times [-1,1])^m \to [-1,1]^{\mathcal{X}}$
- accuracy parameter $\alpha > 0$
- validation data $D = D_0, \ldots, D_T \sim \mathcal{D}^m$

**Let:**

- $\mathcal{X}_0 \leftarrow \{x \in \mathcal{X} : f_0(x) \leq 1/2\}$
- $\mathcal{X}_1 \leftarrow \{x \in \mathcal{X} : f_0(x) > 1/2\}$ `// partition X according to f0`
- $\mathcal{S} \leftarrow \{\mathcal{X}, \mathcal{X}_0, \mathcal{X}_1\}$

**Repeat:** from $t = 0, 1, \ldots, T$

- For $S \in \mathcal{S}$: `// audit ft on X,X0,X1 with fresh data`
  $h_{t,S} \leftarrow \mathcal{A}(D_t; (f_t - y)_S)$
- $S^* \leftarrow \text{argmax}_{S \in \mathcal{S}} \mathbf{E}_{x \sim D_t}[h_{t,S}(x) \cdot (f_t(x) - y(x))]$ `// take largest residual`
- if $\mathbf{E}_{x \sim D_t}[h_{t,S^*}(x) \cdot (f_t(x) - y(x))] \leq \alpha$:
  **return** $f_t$ `// terminate when at most alpha`
- $f_{t+1}(x) \propto e^{-\eta h_{t,S^*}(x)} \cdot f_t(x) \qquad \forall x \in S^*$ `// multiplicative weights update`

A key algorithmic challenge is to learn a multiaccurate predictor without overfitting to the small sample of validation data. In theory, we prove bounds on the sample complexity necessary to guarantee good generalization as a function of the class $\mathcal{C}$, the error parameter $\alpha$, and the size of subpopulations we wish to protect $\gamma$. In practice, we need to balance the choice of $\mathcal{C}$ (or $\mathcal{A}$) and the number of iterations of our algorithm to make sure that the auditor is discovering true signal, rather than noise in the validation data. Indeed, if the auditor $\mathcal{A}$ learns an expressive enough class of functions, then our algorithm will start to overfit at some point; we show empirically that multiaccuracy post-processing improves the generalization error before overfitting. Next, we give an overview of the algorithm, and state its formal guarantees in Section 3.1.

At a high level, MULTIACCURACY BOOST starts by partitioning the input space $\mathcal{X}$ based on the initial classifier $f_0$ into $\mathcal{X}_0 = \{x \in \mathcal{X} : f_0(x) \leq 1/2\}$ and $\mathcal{X}_1 = \{x \in \mathcal{X} : f_0(x) > 1/2\}$; note that we can partition $\mathcal{X}$ simply by calling $f_0$. Partitioning the search space $\mathcal{X}$ based on the predictions of $f_0$ helps to ensure that the $f$ we output maintains the initial accuracy of $f_0$; in particular, it allows us to search over just the positive-labeled examples (negative, resp.) for a way to improve the classifier. The initial hypothesis may make false positive predictions and false negative predictions for very different reasons, even if in both cases the reason is simple enough to be identified by the auditor.

After partitioning the input space, the procedure iteratively uses the learning algorithm $\mathcal{A}$ to

search over $\mathcal{X}$ (and within the partitions $\mathcal{X}_0, \mathcal{X}_1$) to find any function which correlates significantly with the current residual in prediction $f - y$. If $\mathcal{A}$ successfully returns some function $h : \mathcal{X} \to [-1, 1]$ that identifies a significant subpopulation where the current hypothesis is inaccurate, the algorithm updates the predictions multiplicatively according to $h$. In order to update the predictions simultaneously for all $x \in \mathcal{X}$, at the $t$th iteration, we build $f_{t+1}$ by incorporating $h_t$ into the previous model $f_t$. This approach of augmenting the model at each iteration is similar to boosting. To guarantee good generalization of $h$, we assume that $\mathcal{A}$ uses a fresh sample $D_t \sim \mathcal{D}^m$ per iteration. In practice, when we have few samples, we can put all of our samples in one batch and use noise-addition techniques to reduce overfitting [DFH+15, RZ16]; this connection to adaptive data analysis was studied formally in [HKRR18].

From the stopping condition, it is clear that when the algorithm terminates, $f_T$ passes $(\mathcal{A}, \alpha)$-multiaccuracy auditing. Thus, it remains to bound the number of iterations $T$ before Multiaccuracy Boost terminates. Additionally, as described, the algorithm evaluates statistics like $\mathbf{E}_{x\sim\mathcal{D}}[h(x) \cdot (f(x) - y(x))]$, which depends on $y(x)$ for all $x \in \mathcal{X}$; we need to bound the number of validation samples we need to guarantee good generalization to the unseen population. Theorem 2 provides formal guarantees on the convergence rate and the sample complexity from $\mathcal{D}$ needed to estimate the expectations sufficiently-accurately.

**Do no harm.** The distinction between our approach and most prior works on fairness (especially [KNRW17]) is made clear from the "do-no-harm" property that Multiaccuracy Boost exhibits, stated formally as Theorem 3. In a nutshell, the property guarantees that on any subpopulation $S \subseteq \mathcal{X}$ that $\mathcal{A}$ audits, the classification error cannot increase significantly from $f_0$ to the post-processed classifier. Further, the bound we prove is very pessimistic. Both in theory and in practice, we do not expect any increase to occur. In particular, the convergence analysis of Multiaccuracy Boost follows by showing that at every update, the average cross-entropy loss on the population we update must drop significantly. Termination is guaranteed because after too many iterations of auditing, the post-processing will have learned $y$ perfectly. Thus, if we use Algorithm 1 to post-process a model that is already achieves high accuracy on the validation distribution the resulting model's accuracy should not deteriorate in significant ways; empirically, we observe that classification accuracy (on held-out test set) tends to improve over $\mathcal{D}$ after multiaccuracy post-processing.

### 3.1 Formal guarantees of Multiaccuracy Boost

For clarity of presentation, we describe the formal guarantees of our algorithm assuming that $\mathcal{A}$ provably agnostic learns a class of tests $\mathcal{C}$, in order to describe the sample complexity appropriately. The guarantees on the rate of convergence do not rely on such an assumption. We show that, indeed, Algorithm 1 must converge in a bounded number of iterations. The proof follows by showing that, for an appropriately chosen $\eta$ (on the order of $\alpha$), each update improves the cross-entropy loss over the updated set $S$, so the bound depends on the initial cross-entropy loss.

To estimate the statistics used in Multiaccuracy Boost, we need to bound the sample complexity required for the auditor to generalize. Informally, we say the *dimension* $d(\mathcal{C})$ of an agnostically learnable class $\mathcal{C}$ is a value such that $m \geq \Omega\left(\frac{d(\mathcal{C})+\log(1/\delta)}{\alpha^2}\right)$ random samples from $\mathcal{D}$ guarantee uniform convergence over $\mathcal{C}$ with accuracy $\alpha$ with failure probability at most $\delta$. Examples of upper bounds on this notion of dimension include $\log(|\mathcal{C}|)$ for a finite class of tests, the VC-

dimension [KV94] for boolean tests, and the metric entropy [BLM13] of real-valued tests. We state the formal guarantees as Theorem 2.

**Theorem 2.** *Let $\alpha, \delta > 0$; suppose $\mathcal{A}$ agnostic learns a class $\mathcal{C} \subseteq [-1,1]^{\mathcal{X}}$ of dimension $d(\mathcal{C})$. Then, using $\eta = O(\alpha)$, Algorithm 1 converges to a $(\mathcal{C}, \alpha)$-multiaccurate hypothesis $f_T$ in $T = O\left(\frac{\ell_{\mathcal{D}}(f_0;y)}{\alpha^2}\right)$ iterations from $m = \tilde{O}\left(T \cdot \frac{d(\mathcal{C})+\log(1/\delta)}{\alpha^2}\right)$ random samples with probability $\geq 1-\delta$.*

Roughly speaking, for constant $\alpha, \delta$, the sample complexity scales with the dimension of the class $\mathcal{C}$. For many relevant classes $\mathcal{C}$ for which we would want to enforce $(\mathcal{C}, \alpha)$-multiaccuracy, this dimension will be significantly smaller than the amount of data required to train an accurate initial $f_0$. Note also that our sample complexity is completely generic and we make no effort to optimize the exact bound. In particular, for structured $\mathcal{C}$ and $\mathcal{A}$, better uniform convergence bounds can be proved. Further, appealing to a recent line of work on adaptive data analysis initiated by [DFH+15, RZ16], we can avoid resampling at each iteration as in [HKRR18].

**Do no harm.** The do-no-harm property guarantees that the classification error on any subpopulation that $\mathcal{A}$ audits cannot increase significantly. As we assume $\mathcal{A}$ can identify a very rich class of overlapping sets, in aggregate, this property gives a strong guarantee on the utility of the resulting predictor. Further, the proof of Theorem 3 reveals that this worst-case bound is very pessimistic and can be improved with stronger assumptions.

**Theorem 3** (Do-no-harm)**.** *Let $\alpha, \beta, \gamma > 0$ and $S \subseteq \mathcal{X}$ be a subpopulation where $\mathbf{Pr}_{x\sim\mathcal{D}}[x \in S] \geq \gamma$. Suppose $\mathcal{A}$ audits the characteristic function $\chi_S(x)$ and its negation. Let $f : \mathcal{X} \to [0,1]$ be the output of Algorithm 1 when given $f_0 : \mathcal{X} \to [0,1]$, $\mathcal{A}$, and $\alpha \leq \beta\gamma$ as input. Then the classification error of $f$ on the subset $S$ is bounded as*

$$\mathrm{er}_S(f; y) \leq 3 \cdot \mathrm{er}_S(f_0; y) + 4\beta. \tag{3}$$

**Derivative learning for faster convergence** Here, we propose auditing with an algorithm $\mathcal{A}_\ell$, described formally in Algorithm 2 in the Appendix, that aims to learn a smoothed version of the partial derivative function of the cross-entropy loss with respect to the *predictions* $\frac{\partial \ell(f;y)}{\partial f(x)} = \frac{1}{1-f(x)-y(x)}$, which grows in magnitude as $|f(x) - y(x)|$ grows. We show that running MULTIACCURACY BOOST with $\mathcal{A}_\ell$ converges in a number of iterations that grows with $\log(1/\alpha)$, instead of polynomially, as we would expect for a smooth, strongly convex objective [SS+12, B+15]. This sort of gradient method does not typically make sense when we don't have information about $y(x)$ for all $x \in \mathcal{X}$; nevertheless, if there is a simple way to improve $f$, we might hope to *learn* the partial derivative as a function of $x \in \mathcal{X}$ in order to update $f$. This application of MULTIACCURACY BOOST is similar in spirit to gradient boosting techniques [MBBF00, Fri01], which interpret boosting algorithms as running gradient descent on an appropriate cost-functional.

In principle, if the magnitude of the residual $|f(x) - y(x)|$ is not too close to 1 for most $x \in \mathcal{X}$, then the learned partial derivative function should correlate well with the true gradient. Empirically, we observe that $\mathcal{A}_\ell$ is effective at finding ways to improve the model quite rapidly. Formally, we show the following linear convergence guarantee.

**Proposition 4.** *Let $\alpha, B, L > 0$ and $\mathcal{C} \subseteq [-B, B]^{\mathcal{X}}$. Suppose we run Algorithm 1 with $\eta = O(1/L)$ on initial model $f_0$ with auditor $\mathcal{A}_\ell$ defined in Algorithm 2. Then, Algorithm 1 converges in $T = O\left(L \cdot \log(\ell_{\mathcal{D}}(f_0; y)/\alpha)\right)$ iterations.*

# 4 Experimental Evaluation

We evaluate the empirical performance of Multiaccuracy Boost in three case studies. The first and most in-depth case study aims to emulate the conditions of the Gender Shades study [BG18], to test the effectiveness of multiaccuracy auditing and post-processing on this important real-world example. In Section 4.1, we show experimental results for auditing using two different validation data sets. In particular, one data set is fairly unbalanced and similar to the data used to train, while the other data set was developed in the Gender Shades study and is very balanced. For each experiment, we report for various subpopulations, the population percentage in $\mathcal{D}$, accuracies of the initial model, our black-box post-processed model, and white-box benchmarks. In Section 4.1.1, we discuss further subtleties of applying the multiaccuracy framework involving the representation of inputs passed to $\mathcal{A}$ for auditing; in Section 4.1.2, we show how auditing can be used beyond post-processing. In particular, the hypotheses that $\mathcal{A}$ learns can be used to highlight subpopulations – in an interpretable way – on which the model is making mistakes.

We evaluate the effectiveness of multiaccuracy post-processing on two other prediction tasks. In both these case studies, we take the training and auditing data distribution $\mathcal{D}$ to be the same, though the number of examples used for auditing will be quite small. Multiaccuracy still improves the performance on significant subpopulations. These results suggest some interesting hypotheses about why machine-learned models may incorporate biases in the first place, which warrant further investigation.

## 4.1 Multiaccuracy improves gender detection

In this case study, we replicate the conditions of the Gender Shades study [BG18] to evaluate the effectiveness of the multiaccuracy framework in a realistic setting. For our initial model, we train an inception-resnet-v1 [SIVA17] gender classification model using the CelebA data set with more than 200,000 face images [LLWT15]. The resulting test accuracy on CelebA for binary gender classification is 98.4%.

We applied Multiaccuracy Boost to this $f_0$ using two different auditing distributions. In the first case, we audit using data from the LFW+a[1] set [WHT11, HRBLM07], which has similar demographic breakdowns as CelebA (i.e. $\mathcal{D} \approx \mathcal{D}_0$). In the second case, we audit using the PPB data set (developed in [BG18]) which has balanced representation across gender and race (i.e. $\mathcal{D} \neq \mathcal{D}_0$). These experiments allows us to track the effectiveness of Multiaccuracy Boost as the representation of minority subpopulations changes. In both cases, the auditor is "blind" – it is not explicitly given the race or gender of any individual – and knows nothing about the inner workings of the classifier. Specifically, we take the auditor to be a variant of $\mathcal{A}_\ell$ (Algorithm 2) that performs ridge regression to fit $\frac{\partial \ell_x(f;y)}{\partial f(x)} = \frac{1}{1-f(x)-y(x)}$.[2] Instead of training the auditor on raw input pixels, we use the low dimensional representation of the input images derived by a variational autoencoder (VAE) trained on CelebA dataset using Facenet [SKP15] library. (For more discussion of the representation used during auditing, cf. Section 4.1.1.)

To test the initial performance of $f_0$, we evaluated on a random subset of the LFW+a data containing 6,880 face images, each of which is labeled with both gender and race – black (**B**) and non-black (**N**). For gender classification on LFW+a, $f_0$ achieves 94.4% accuracy. Even though the

[1] We fixed the original data set's label noise for sex and race.

[2] To help avoid outliers, we smooth the loss and use a quadratic approximation for $\left|\frac{\partial \ell_x(f;y)}{\partial f(x)}\right| > 10$.

overall accuracy is high, the error rate is much worse for females (23.1%) compared to males (0.7%) and worse for blacks (10.2%) compared to non-blacks (5.1 %); these results are qualitatively very similar to those observed by the commercial gender detection systems studied in [BG18]. We applied Multiaccuracy Boost, which converged in 7 iterations. The resulting classifier's classification error in minority subpopulations was substantially reduced, even though the auditing distribution was similar as the training distribution.

We compare Multiaccuracy Boost against a strong white-box baseline. Here, we retrain the network of $f_0$ using the audit set. Specifically, we retrain the last two layers of the network, which gives the best results amongst retraining methods. We emphasize that this baseline requires white-box access to $f_0$, which is often infeasible in practice. Multiaccuracy Boost accesses $f_0$ only as a black-box without any additional demographic information, and still achieves comparable, if not improved, error rates compared to retraining. We report the overall classification accuracy as well as accuracy on different subpopulations – e.g. **BF** indicates black female – in Table 1.

| | **All** | **F** | **M** | **B** | **N** | **BF** | **BM** | **NF** | **NM** |
|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{D}$ | 100 | 21.0 | 79.0 | 4.9 | 95.1 | 2.1 | 18.8 | 2.7 | 76.3 |
| $f_0$ | 5.4 | 23.1 | 0.7 | 10.2 | 5.1 | 20.4 | 2.1 | 23.4 | 0.6 |
| MA | 4.1 | 11.3 | 3.2 | 6.0 | 4.9 | 8.2 | 4.3 | 11.7 | 3.2 |
| RT | 3.8 | 11.2 | 1.9 | 7.5 | 3.7 | 11.6 | 4.3 | 11.1 | 1.8 |

Table 1: **Results for LFW+a gender classification.** $\mathcal{D}$ denotes the percentages of each population in the data distribution; $f_0$ denotes the classification error (%) of the initial predictor; MA denotes the classification error (%) of the model after post-processing with Multiaccuracy Boost; RT denotes the classification error (%) of the model after retraining on $\mathcal{D}$.

The second face dataset, PPB, in addition to being more balanced, is much smaller; thus, this experiment can be viewed as a stress test, evaluating the data efficiency of our post-processing technique. The test set has 415 individuals and the audit set has size 855. PPB annotates each face as dark (**D**) or light-skinned (**L**). As with LFW+a, we evaluated the test accuracy of the original $f_0$, the multiaccurate post-processed classifier, and retrained classifier on each subgroup. Multiaccuracy Boost converged in 5 iterations and again, substantially reduced error despite a small audit set and the lack of annotation about race or skin color (Table 2).

| | **All** | **F** | **M** | **D** | **L** | **DF** | **DM** | **LF** | **LM** |
|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{D}$ | 100 | 44.6 | 55.4 | 46.4 | 53.6 | 21.4 | 25.0 | 23.2 | 30.4 |
| $f_0$ | 9.9 | 21.6 | 0.4 | 18.8 | 2.2 | 39.8 | 1.0 | 5.2 | 0.0 |
| MA | 3.9 | 6.5 | 1.8 | 7.3 | 0.9 | 12.5 | 2.9 | 1.0 | 0.8 |
| RT | 2.2 | 3.8 | 0.9 | 4.2 | 0.4 | 6.8 | 1.9 | 1.0 | 0.0 |

Table 2: **Results for the PPB gender classification data set.** $\mathcal{D}$ denotes the percentages of each population in the data distribution; $f_0$ denotes the classification error (%) of the initial predictor; MA denotes the classification error (%) of the model after post-processing with Multiaccuracy Boost; RT denotes the classification error (%) of the model after retraining on $\mathcal{D}$.

To further test the data efficiency of Multiaccuracy Boost, we evaluate the effect of audit set size on the resulting accuracy of each method. In Fig. 1, we report the performance of Multiaccuracy Boost versus the white-box retraining method for different sizes of audit set. The plot

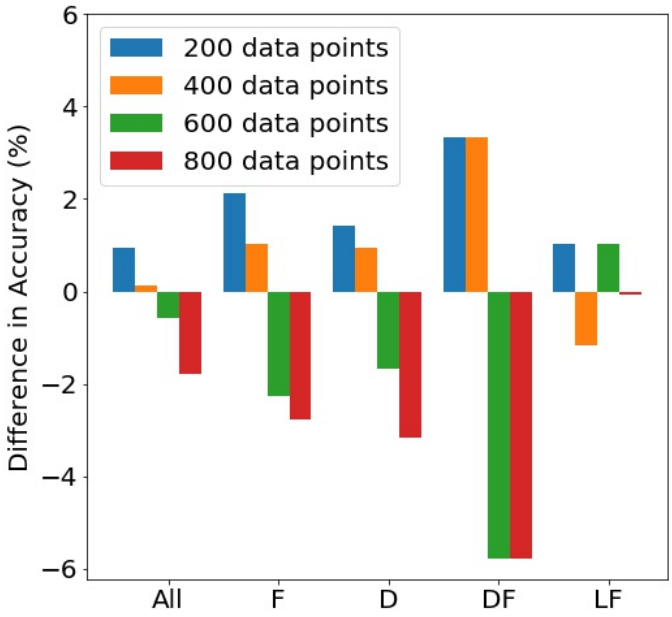


Figure 1: **Multiaccuracy vs. Retraining**: Difference in classification accuracy (i.e. % accuracy after Multiaccuracy Boost − % accuracy after retraining) is plotted on the vertical axis; this difference represents the accuracy advantage of Multiaccuracy Boost compared to retraining. As the size of the audit set shrinks, Multiaccuracy Boost has better performance both in overall accuracy and accuracy of the subgroups with the most initial bias because it is more data efficient.

displays the difference in accuracy for the overall population along with the subgroups that suffered the most initial bias. It shows that the performance of Multiaccuracy Boost may actually be better than the white-box retraining baseline when validation data is especially scarce.

#### 4.1.1 Representation matters

As discussed earlier, in the reported gender detection experiments, we audit for multiaccuracy using ridge regression over an encoding of images produced by a variational autoencoder. Using the representation of images produced by this encoding intuitively makes sense, as the autoencoder's reconstruction objective aims to preserve as much information about the image as possible while reducing the dimension. Still, we may wonder whether multiaccuracy auditing over a different representation of the images would perform better. In particular, since we are interested in improving the accuracy on the gender detection task, it seems plausible that a representation of the images based on the internal layers of the initial prediction network might preserve the information salient to gender detection more effectively.

We investigate the importance of the representation used to audit empirically. In particular, we also evaluate the performance of Multiaccuracy Boost using the same auditor $\mathcal{A}_\ell$ run over two

other sets of features, given by the last-layer and the second-to-last layer of the initial prediction residual network $f_0$. In Table 3, we show that using the unsupervised VAE representation yields the best results. Still, the representations from the last and second-to-last layers are competitive with that of the VAE, and in some subpopulations are better.

Collectively, these findings bolster the argument for "fairness through awareness", which advocates that in order to make fair predictions, sensitive information (like race or gender) should be given to the (trustworthy) classifier. While none of these representations explicitly encode sensitive group information, the VAE representation does preserve information about the original input, for instance skin color, that seems useful in understanding the group status. The prediction network is trained to have the best prediction accuracy (on an unbalanced training data set), and thus, the representations from the network reasonably may contain less information about these sensitive features. These results suggest that the effectiveness of multiaccuracy does depend on the representation of inputs used for auditing, but so long as the representation is sufficiently expressive, Multiaccuracy Boost may be robust to the exact encoding of the features.

| | **All** | **F** | **M** | **D** | **L** | **DF** | **DM** | **LF** | **LM** |
|---|---|---|---|---|---|---|---|---|---|
| **LFW+a:** | | | | | | | | | |
| VAE | 4.1 | 11.3 | 3.2 | 6.0 | 4.9 | 8.2 | 4.3 | 11.7 | 3.2 |
| $R_{1,f_0}$ | 4.9 | 13.6 | 2.6 | 6.3 | 4.9 | 8.8 | 4.3 | 14.1 | 2.6 |
| $R_{2,f_0}$ | 4.5 | 12.6 | 2.4 | 6.3 | 4.4 | 8.8 | 4.3 | 13.1 | 2.3 |
| **PPB:** | | | | | | | | | |
| VAE | 3.9 | 6.5 | 1.8 | 7.3 | 0.9 | 12.5 | 2.9 | 1.0 | 0.8 |
| $R_{1,f_0}$ | 4.3 | 7.6 | 1.7 | 7.8 | 1.3 | 13.6 | 2.9 | 2.1 | 0.8 |
| $R_{2,f_0}$ | 5.1 | 9.7 | 1.3 | 9.4 | 1.3 | 17.0 | 2.9 | 3.1 | 0.0 |

Table 3: **Effect of representation on the Multiaccuracy Boost performance** VAE denotes the denotes the classification error (%) using the VAE representation; $R_{1,f_0}$ denotes the classification error (%) using the classifier's last layer representation, $R_{2,f_0}$ denotes the classification error (%) using the classifier's second to last layer representation

#### 4.1.2 Multiaccuracy auditing as diagnostic

As was shown in [BG18], we've demonstrated that models trained in good faith on unbalanced data may exhibit significant biases on the minority populations. For instance, the initial classification error on black females is significant, whereas on white males, it is near 0. Importantly, the only way we were able to report these accuracy disparities was by having access to a rich data set where gender and race were labeled. Often, this demographic information will not be available; indeed, the CelebA images are not labeled with race information, and as such, we were unable to evaluate the subpopulation classification accuracy on this set. Thus, practitioners may be faced with a problem: even if they know their model is making undesirable mistakes, it may not be clear if these mistakes are concentrated on specific subpopulations. Absent any identification of the subpopulations on which the model is underperforming, collecting additional training data may not actually improve performance across the board.

We demonstrate that multiaccuracy auditing may serve as an effective diagnostic and interpretation tool to help developers identify systematic biases in their models. The idea is simple:

the auditor returns a hypothesis $h$ that essentially "scores" individual inputs $x$ by how wrong the prediction $f_0(x)$ is. If we consider the magnitude of their scores $|h(x)|$, then we may understand better the biases that the encoder is discovering.

As an example, we test this idea on the PPB data set, evaluating the test images' representations with the hypotheses the auditor returns. In Figure 2, we display the images in the test set that get the highest and lowest effect ($|h(x)|$ large and $|h(x)| \approx 0$, respectively) according to the first and second hypothesis returned by $\mathcal{A}_\ell$. In the first round of auditing, the three highest-scoring images (top-left row) are all women, both black and white. Interestingly, all of the least active images (bottom-left row) are men in suits, suggesting that suits may be a highly predictive feature of being a man according to the original classifier, $f_0$. Overall the first round of audit seems to primarily identify gender as the axis of bias in $f_0$. In the second round, after the classifier has been improved by one step of MULTIACCURACY BOOST, the auditor seems to hone in on the "dark-skinned women" subpopulation as the region of bias, as the highest activating images (top-right row) are all dark-skinned women.


Figure 2: **Interpreting Auditors** Here, we depict the PPB test images with the highest and lowest activation of the first and second trained auditor. The images with the highest auditor effects corresponds to images where the auditor detects the largest bias in the classifier. In the first round of auditing, the most biased images are women, both black and white. In the second round of auditing, after the base classifier has been augmented by one step of MULTIACCURACY BOOST, the most biased images are more specifically dark-skinned women.

### 4.2 Additional case studies

Multiaccuracy auditing and post-processing is applicable broadly in supervised learning tasks, not just in image classification applications. We demonstrate the effectiveness of MULTIACCURACY BOOST in two other settings: the adult income prediction task and a semi-synthetic disease prediction task.

**Adult Income Prediction** For the first case study, we utilize the adult income prediction data set [Koh96] with 45,222 samples and 14 attributes (after removing subjects with unknown attributes) for the task of binary prediction of income more than $50k for the two major groups of Black and White. We remove the sensitive features of gender – female (**F**) and male (**M**) and race (for the two major groups) – black (**B**) and white (**W**) – from the data, to simulate settings where sensitive

features are not available to the algorithm training. We trained a base algorithm, $f_0$, which is a neural network with two hidden layers on 27,145 randomly selected individuals. The test set consists of an independent set of 15,060 persons.

We audit using a decision tree regression model (max depth 5) $\mathcal{A}_{\text{dt}}$ to fit the residual $f(x)-y(x)$. $\mathcal{A}_{\text{dt}}$ receives samples of validation data drawn from the same distribution as training; that is $\mathcal{D} = \mathcal{D}_0$. In particular, we post-process with 3,017 individuals sampled from the same adult income dataset (disjoint from the training set of $f_0$). The auditor is given the same features as the original prediction model, and thus, is not given the gender or race of any individual. We evaluate the post-processed classifier on the same independent test set. MULTIACCURACY BOOST converges in 50 iterations with $\eta = 1$.

As a baseline, we trained four separate neural networks with the same architecture as before (two hidden layers) for each of the four subgroups using the audit data. As shown in Table 4, multiaccuracy post-processing achieves better accuracy both in aggregate and for each of the subgroups. Importantly, the subgroup-specific models requires explicit access to the sensitive features of gender and race. Training a classifier for each subgroup, or explicitly adding subgroup accuracy into the training objective, assumes that the subgroup is already identified in the data. This is not feasible in the many applications where, say, race or more granular categories are not given. Even when the subgroups are identified, we often do not have enough samples to train accurate classifiers on each subgroup separately. This example illustrates that multiaccuracy can help to boost the overall accuracy of a black-box predictor in a data efficient manner.

| | **All** | **F** | **M** | **B** | **W** | **BF** | **BM** | **WF** | **WM** |
|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{D}$ | 100.0 | 32.3 | 67.7 | 90.3 | 9.7 | 4.8 | 4.9 | 27.4 | 62.9 |
| $f_0$ | 19.3 | 9.3 | 24.2 | 10.5 | 20.3 | 4.8 | 15.8 | 9.8 | 24.9 |
| MA | 14.7 | 7.2 | 18.3 | 9.4 | 15.0 | 4.5 | 13.9 | 7.3 | 18.3 |
| SS | 19.7 | 9.5 | 24.6 | 10.5 | 19.9 | 5.5 | 15.3 | 10.2 | 25.3 |

Table 4: **Results for Adult Income Data Set** $\mathcal{D}$ denotes the percentages of each population in the data distribution; $f_0$ denotes the classification error (%) of the initial predictor; MA denotes the classification error (%) of the model after post-processing with MULTIACCURACY BOOST; SS denotes the classification error (%) of the subgroup-specific models trained separately for each population.

#### 4.2.1 Semi-Synthetic Disease Prediction

We design a disease prediction task based on real individuals, where the phenotype to disease relation is designed to be different for different subgroups, in order to simulate a challenging setting. We used 40,000 individuals sampled from the UK Biobank [SGA+15]. Each individual contains 60 phenotype features. To generate a synthetic disease outcome for each subgroup, we divided the data set into four groups based on gender – male (**M**) and female (**F**) – and age – young (**Y**) and old (**O**). For each subgroup, we create synthetic binary labels using a different polynomial function of the input features with different levels of difficulty. The polynomial function orders are 1, 4, 2, and 6 for OF, OM, YF, and YM subgroups respectively.

For $f_0$, we trained a neural network with two hidden layers on 32,000 individuals, without using the gender and age features. Hyperparameter search was done for the best weight-decay and drop-out parameters. The $f_0$ we discover performs moderately well on every subpopulation, with the

exception of old females (**OF**) where the classification error is significantly higher. Note that this subpopulation had the least representation in $\mathcal{D}_0$. Again, we audit using $\mathcal{A}_{\text{dt}}$ to run decision tree regression with validation data samples drawn from $\mathcal{D} = \mathcal{D}_0$. Specifically, the auditor receives a sample of 4,000 individuals without the gender or age features. As a baseline, we trained a separate classifier for each of the subgroups using the same audit data. As Table 5 shows, MULTIACCURACY BOOST significantly lowers the classification error in the old female population.

| | **All** | **F** | **M** | **O** | **Y** | **OF** | **OM** | **YF** | **YM** |
|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{D}$ | 100 | 39.6 | 60.4 | 34.6 | 65.4 | 15.0 | 19.7 | 24.6 | 40.7 |
| $f_0$ | 18.9 | 29.4 | 12.2 | 21.9 | 17.3 | 36.8 | 10.9 | 24.9 | 12.8 |
| MA | 16.0 | 24.1 | 10.7 | 16.4 | 15.7 | 26.5 | 9.0 | 22.7 | 11.6 |
| SS | 19.5 | 32.4 | 11.0 | 22.1 | 18.1 | 37.6 | 10.3 | 29.3 | 11.3 |

Table 5: **Results for UK Biobank semi-synthetic data set.** $\mathcal{D}$ denotes the percentages of each population in the data distribution; $f_0$ denotes the classification error (%) of the initial predictor; MA denotes the classification error (%) of the model after post-processing with MULTIACCURACY BOOST; SS denotes the classification error (%) of the subgroup-specific models trained separately for each population.

# 5 Discussion

In this work, we propose multiaccuracy auditing and post-processing as a method for improving the fairness and accountability of black-box prediction systems. Here, we discuss how our work compares to prior works, specifically, how it fits into the growing literature on fairness for learning systems. We conclude with further discussion of our results and speculation about future investigations.

## 5.1 Related works

Many different notions of fairness have been proposed in literature on learning and classification [DHP$^+$12, HPS16, ZWS$^+$13, DIKL17, HKRR18, KNRW17, HSNL18, KRR18, RY18]. Many of these works encode some notion of parity, e.g. different subgroups should have similar false positive rates, as an explicit objective/constraint in the training of the original classifier. The fairness properties are viewed as constraints on the classifier that ultimately *limit the model's utility*. A common belief is that in order to achieve equitable treatment for protected subpopulations, the performance on other subpopulations necessarily must degrade.

A notable exception to this pattern is the work of Hébert-Johnson *et al.* [HKRR18], which introduced a framework for achieving fairness notions that aim to provide accurate predictions for many important subpopulations. [HKRR18] introduced the notion of *multiaccuracy*[3] and a stronger notion, dubbed *multicalibration*, in the context of regression tasks. Multicalibration guarantees (approximately) calibrated predictions, not just overall, but on a rich class of structured "identifiable" subpopulations. [HKRR18] provides theoretical algorithms for achieving multiaccuracy and multicalibration, and shows how to post-process a model to achieve multicalibration in a way that *improves* the regression objective across all subpopulations (in terms of squared-error). Our work directly extends the approach of [HKRR18], adapting their work to the binary classification setting. Our

[3] [HKRR18] refers to this notion as "multi-accuracy-in-expectation".

post-processing algorithm, MULTIACCURACY BOOST, builds on the algorithm given in [HKRR18], providing the additional "do-no-harm" property. This property guarantees that if the initial predictor $f_0$ has small classification error on some identifiable group, then the resulting post-processed model will also have small classification error on this group.

Independent work of Kearns *et al.* [KNRW17] also investigated how to achieve statistical fairness guarantees, not just for traditionally-protected groups, but on rich families of subpopulations. [KNRW17] proposed a framework for *auditing* and learning models to achieve fairness notions like statistical parity and equal false positive rates. Both works [HKRR18, KNRW17] connect the task of learning a model that satisfies the notion of fairness to the task of (weak) agnostic learning [Kea98, KSS94, KMV08, Fel10]. [KNRW17] also reduces the problem of learning a classifier satisfying parity-based notions of fairness across subgroups to the problem of auditing; it would be interesting if their notion of auditing can be used by humans as a way to diagnose systematic discrimination.

Our approach to post-processing, which uses a learning algorithm as a fairness auditor, is similar in spirit to the approach to learning of [KNRW17], but differs technically in important ways. In particular, in the framework of [KNRW17], the auditor is used during (white-box) training to *constrain* the model selected from a pre-specified hypothesis class; ultimately, this constrains the accuracy of the predictions. In our setting (as in [HKRR18]), we do not restrict ourselves to an explicitly-defined hypothesis class, so we can augment the current model using the auditor; these augmentations *improve* the accuracy of the model.

Indeed, at a technical level, our post-processing algorithm is most similar to work on boosting [FS95, SF12], specifically, gradient boosting [MBBF00, Fri01]. Still, our perspective is quite different from the standard boosting setting. Rather than using an expressive class of predictors as the base classifiers to be able to learn the function directly, our setting focuses on the regime where data is limited and we must restrict our attention to simple classes. Thus, it becomes important that we leverage the expressiveness (and initial accuracy) of $f_0$ if we are to obtain strong performance using the multiaccuracy approach. Further, the termination of MULTIACCURACY BOOST certifies that the final model satisfies $(\mathcal{A}, \alpha)$-multiaccuracy; in general, standard boosting algorithms will not provide such a certificate.

Motivated by unfairness that arises as the result of feedback loops in classification settings, another recent work of Hashimoto *et al.* [HSNL18] aims to improve fairness at a subpopulation level. Specifically, their notion of fairness similarly aims to give accurate (i.e. bounded loss) predictions not just overall, but on *all* significant subpopulations. In the multiaccuracy setting, we argued that this goal was information-theoretically infeasible; [HSNL18] sidesteps this impossibility by optimizing over a fixed hypothesis class, and formulating the problem as a min-max optimization. They give show how to relax the problem of minimizing the worst-case subpopulation loss and reduce the relaxation to a certain robust optimization problem. While their approach does not guarantee optimality, it gives a strong certificate upper-bounding the maximum loss over all subpopulations.

A different approach to subgroup fairness is studied by Dwork *et al.* [DIKL17]. This work investigates the question of how to learn a "decoupled" classifier, where separate classifiers are learned for each subgroup and then combined to achieve a desired notion of fairness. While applicable in some settings, at times, this approach may be untenable. First, decoupling the classification problem requires that we have race, age, and other attributes of interest in the dataset and that the groups we wish to protect are partitioned by these attributes; this information is often not available. Even if this information is available, *a priori*, it may not always be obvious which subpopulations

require special attention. In contrast, the multiaccuracy approach allows us to protect a rich class of overlapping subpopulations without explicit knowledge of the vulnerable populations. An interesting direction for future investigation could try to pair multiaccuracy auditing (to identify subpopulations in need of protection) with the decoupled classification techniques of [DIKL17].

The present work, along with [HKRR18,KNRW17,KRR18], can be viewed as studying information-fairness tradeoffs in prediction tasks (i.e. strengthening the notion of fairness that can be guaranteed using a small sample). These works fit into the larger literature on fairness in learning and prediction tasks [DHP+12, ZWS+13, BG18, HPS16, DIKL17, KRR18, RY18], discussions of the utility-fairness tradeoffs in fair classification [ALMK16, KMR17, Cho17, CG16, CDPF+17, PRW+17]. While fairness and accountability serve as the main motivations for developing the multiaccuracy framework, our results may have broader interest. In particular, multiaccuracy post-processing may be applicable in domain adaptation settings, particularly under label distribution shift as studied recently in [LWS18], but when the learner gets a small number of labeled samples from the new distribution.

### 5.2 Conclusion

The multiaccuracy framework can be applied very broadly; importantly, we can post-process any initial model $f_0$ given only black-box access to $f_0$ and a small set of labeled validation data. We show that in a wide range of realistic settings, post-processing for multiaccuracy helps to mitigate systematic biases in predictors across sensitive subpopulations, even when the identifiers for these subpopulations are not given to the auditor explicitly. In our experiments, we observe that standard supervised learning optimizes for overall performance, leading to settings where certain subpopulations incur substantially worse error rates. Multiaccuracy provides a framework for fairness in classification by improving the accuracy in identifiable subgroups, in a way that suffers no tradeoff between accuracy and utility. We demonstrate – both theoretically and empirically – that post-processing with Multiaccuracy Boost serves as an effective tool for improving the accuracy across important subpopulations, and does not harm the populations that are already classified well.

Multiaccuracy works to the extent that the auditor can effectively identify specific subgroups where the original classifier $f_0$ tends to make mistakes. The power of multiaccuracy lies in the fact that in many settings, we can identify issues with $f_0$ using a relatively small amount of audit data. Thus, multiaccuracy auditing is limited: if the mistakes appear overly-complicated to the bounded auditor (for information- or complexity-theoretic reasons), then the auditor will not be able to identify these mistakes. Our empirical results suggest, however, that in many realistic settings, the subpopulations on which a classifier errs are efficiently-identifiable. This observation may be of interest beyond the context of fairness. In particular, our experiments improving the accuracy of a model trained on CelebA on the LFW+a and PPB test sets suggests a lightweight black-box alternative to more sophisticated transfer learning techniques, which may warrant further investigation.

Our empirical investigations reveal some additional interesting aspects of the multiaccuracy framework. In particular, we've shown that multiaccuracy auditing can identify underrepresented groups receiving suboptimal predictions even when the sensitive attributes defining these groups are not explicitly given to the auditor, which proves useful for diagnosing where models make mistakes. We feel that it may be of further interest within the study of model interpretability. Finally, it is striking that Multiaccuracy Boost tends to improve, not just subgroup accuracy, but also the overall accuracy, even when the minority groups remain underrepresented in the valida-

tion data. While some of these findings may be due to suboptimal training of our initial models, we believe this is not the only factor at play. In particular, we hypothesize that understanding why models incorporate biases during training – and further, why simple interventions like multi-accuracy post-processing can significantly improve generalization error – requires investigating the dynamics of overfitting during training, not just on the population as a whole, but across significant subpopulations.

**Acknowledgements.** The authors thank Omer Reingold and Guy N. Rothblum for their advice and helpful discussions throughout the development of this work; we thank Weihao Kong, Aditi Raghunathan, and Vatsal Sharan for feedback on early drafts of this work.

## References

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

[B+15] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[BG18] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[CDPF+17] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. *KDD*, 2017.

[CG16] Alexandra Chouldechova and Max G'Sell. Fairer and more accurate, but for whom? *FATML*, 2016.

[Cho17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 2017.

[DFH+15] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

[DHP+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, pages 214–226, 2012.

[DIKL17] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for fair and efficient machine learning. *arXiv preprint arXiv:1707.06613*, 2017.

[Fel10] Vitaly Feldman. Distribution-specific agnostic boosting. In *Proceedings of the First Symposium on Innovations in Computer Science'10*, 2010.

[Fri01] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[FS95] Yoav Freund and Robert E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[HKRR18] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Calibration for the (computationally-identifiable) masses. *ICML*, 2018.

[HPS16] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[HRBLM07] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[HSNL18] Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. *ICML*, 2018.

[Kea98] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

[KMR17] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *ITCS*, 2017.

[KMV08] Adam Tauman Kalai, Yishay Mansour, and Elad Verbin. On agnostic boosting and parity learning. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 629–638. ACM, 2008.

[KNRW17] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144v3*, 2017.

[Koh96] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *KDD*, volume 96, pages 202–207. Citeseer, 1996.

[KRR18] Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Fairness through computationally-bounded awareness. *arXiv Preprint*, 1803.03239, 2018.

[KSS94] Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

[KV94] Michael J. Kearns and Umesh Virkumar Vazirani. *An introduction to computational learning theory*. MIT press, 1994.

[LLWT15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[LWS18] Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *ICML*, 2018.

[MBBF00] Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus R Frean. Boosting algorithms as gradient descent. In *Advances in neural information processing systems*, pages 512–518, 2000.

[PRW+17] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. *NIPS*, 2017.

[RY18] Guy N. Rothblum and Gal Yona. Probably approximately metric-fair learning. *ICML*, 2018.

[RZ16] Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *AISTATS*, 2016.

[SF12] Robert E Schapire and Yoav Freund. *Boosting: Foundations and algorithms*. MIT press, 2012.

[SGA+15] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[SIVA17] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.

[SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.

[SS+12] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[TTV09] Luca Trevisan, Madhur Tulsiani, and Salil Vadhan. Regularity, boosting, and efficiently simulating every high-entropy distribution. In *Computational Complexity, 2009. CCC'09. 24th Annual IEEE Conference on*, pages 126–136. IEEE, 2009.

[WHT11] Lior Wolf, Tal Hassner, and Yaniv Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *IEEE transactions on pattern analysis and machine intelligence*, 33(10):1978–1990, 2011.

[ZWS+13] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 325–333, 2013.

**Appendix notation.** We use the inner product $\langle h, g\rangle = \mathbf{E}_{x\sim\mathcal{D}}[h(x)\cdot g(x)]$ and the $p$-norms $\|h\|_p = (\mathbf{E}_{x\sim\mathcal{D}}[|h(x)|^p])^{1/p}$.

# A Multiaccuracy and classification error

Here, we prove Proposition 1.

**Proposition** (Restatement of Propostion 1)**.** *Let $\hat{y} : \mathcal{X} \to \{-1, 1\}$ as $\hat{y}(x) = 1 - 2y(x)$. Suppose that for $S \subseteq \mathcal{X}$ with $\mathbf{Pr}_{x\sim\mathcal{D}}[x \in S] \geq \gamma$, there is some $c \in \mathcal{C}$ such that $\|c - \hat{y}_S\|_1 \leq \tau$. Then if $f$ is $(\mathcal{C}, \alpha)$-multiaccurate,* $\mathrm{er}_S(f; y) \leq 2 \cdot (\alpha + \tau)/\gamma$.

*Proof.* For $i, j \in \{0, 1\}$, let $S_{ij} = \left\{x \in S : y(x) = i \wedge \bar{f}(x) = j\right\}$. Further denote $\beta_{ij} = \mathbf{Pr}_{x\sim\mathcal{D}}[x \in S_{ij}]$. Note that the classification error on a set $S$ is $\mathrm{er}_S(f; y) \leq (\beta_{01} + \beta_{10})/\gamma$.

Let $\hat{y}(x) = 1 - 2y(x)$ and suppose $c(x) = \hat{y}(x)_S + z(x)$ where $\|\delta\|_1 \leq \tau$. Then, we derive the following inequality.

$$\begin{aligned}
&\mathop{\mathbf{E}}_{x\sim\mathcal{D}}[c(x)\cdot(f(x) - y(x))] && (4)\\
&= \mathop{\mathbf{E}}_{x\sim\mathcal{D}}[\hat{y}(x)_S \cdot (f(x) - y(x))] + \mathop{\mathbf{E}}_{x\sim\mathcal{D}}[z(x)\cdot(f(x) - y(x))] && (5)\\
&\geq \beta_{01} \cdot \mathop{\mathbf{E}}_{x\sim S_{01}}[f(x) - y(x)] + \beta_{10} \cdot \mathop{\mathbf{E}}_{x\sim S_{10}}[y(x) - f(x)] - \tau && (6)
\end{aligned}$$

where (6) follows by Hölder's inequality, from the fact that the contribution to the expectation of $(1 - 2y(x)) \cdot (f(x) - y(x))$ from $S_{00}$ and $S_{11}$ is lower bounded by 0, and by the definition $\hat{y}_S(x) = 0$ for $x \notin S$. Further, because we know any $x \in S_{01} \cup S_{10}$ is misclassified, we can lower bound the contribution by $1/2$. Thus, if $\mathbf{E}_{x\sim\mathcal{D}}[c(x) \cdot (f(x) - y(x))] \leq \alpha$, then by rearranging we conclude

$$\mathrm{er}_S(f; y) = (\beta_{01} + \beta_{10})/\gamma \leq 2 \cdot (\alpha + \tau)/\gamma. \tag{7}$$

□

Theorem 3 follows by a similar argument.

**Theorem** (Restatement of Theorem 3)**.** *Let $\alpha, \beta, \gamma > 0$ and $S \subseteq \mathcal{X}$ be a subpopulation where $\mathbf{Pr}_{x\sim\mathcal{D}}[x \in S] \geq \gamma$. Suppose for $\mathcal{A}$ audits the characteristic function $\chi_S(x)$ and its negation. Let $f : \mathcal{X} \to [0, 1]$ be the output of Algorithm 1 when given $f_0 : \mathcal{X} \to [0, 1]$, $\mathcal{A}$, and $0 < \alpha \leq \beta\gamma$ as input. Then the classification error of $f$ on the subset $S$ is bounded as*

$$\mathrm{er}_S(f; y) \leq 3 \cdot \mathrm{er}_S(f_0; y) + 4\beta. \tag{8}$$

*Proof.* Suppose that $\mathrm{er}_S(f_0; y) \leq \tau$. Consider $S_1 = \{x \in S : f_0(x) > 1/2\}$; suppose $\mathrm{er}_{S_1}(f_0; y) = \tau_1$.

By assumption, $-\chi_S(x)$ is audited on $\mathcal{X}_1$. Consider $\mathbf{E}_{x\sim S_1}[-\chi_S(x)\cdot(f(x)-y(x))]$.

$$\begin{aligned}
&\mathop{\mathbf{E}}_{x\sim S_1}[-\chi_S(x)\cdot(f(x)-y(x))] && (9)\\
&= \mathop{\mathbf{E}}_{x\sim S_1}[y(x)-f(x)] && (10)\\
&= \mathop{\mathbf{Pr}}_{x\sim S_1}[y(x)=1]\cdot \mathop{\mathbf{E}}_{\substack{x\sim S_1\\ y(x)=1}}[1-f(x)] - \mathop{\mathbf{Pr}}_{x\sim S_1}[y(x)=0]\cdot \mathop{\mathbf{E}}_{\substack{x\sim S_1\\ y(x)=0}}[f(x)] && (11)\\
&\geq \mathop{\mathbf{Pr}}_{x\sim S_1}[y(x)=1\wedge \bar{f}(x)=0]\cdot \mathop{\mathbf{E}}_{\substack{x\sim S_1\\ y(x)=1\\ \bar{f}(x)=0}}[1-f(x)] - \tau_1 && (12)\\
&\geq \frac{1}{2}\mathop{\mathbf{Pr}}_{x\sim S_1}[y(x)=1\wedge \bar{f}(x)=0] - \tau_1 && (13)
\end{aligned}$$

where (12) follows from applying Hölder's inequality and the assumption that $\mathrm{er}_{S_1}(f_0;y)=\tau_1$; and (13) follows from lower bounding the contribution to the expectation based on the true label and the predicted label. Note that $\mathbf{Pr}_{x\sim S}[x\in S_1]\cdot\mathbf{E}_{x\sim S_1}[y(x)-f(x)]\leq \alpha/\gamma=\beta$ by the fact that $f$ passes multiaccuracy auditing by $\mathcal{A}$ and the assumption that $\mathbf{Pr}_{x\sim\mathcal{D}}[x\in S]\geq\gamma$. Rearranging gives the following inequality

$$\mathrm{er}_{S_1}(f;y)\leq \frac{2\beta}{\mathbf{Pr}_{x\sim S}[x\in S_1]}+3\tau_1 \tag{14}$$

where the additional $\tau_1$ comes from accounting for the false positives.

A similar argument holds for $S_0$ with $\mathrm{er}_{S_0}(f_0;y)=\tau_0$, using $\chi_S(x)$. We can expand $\mathrm{er}_S(f;y)$ as a convex combination of the classification error over $S_0$ and $S_1$.

$$\begin{aligned}
&\mathrm{er}_S(f;y) && (15)\\
&= \mathop{\mathbf{Pr}}_{x\sim S}[x\in S_0]\cdot\mathrm{er}_{S_0}(f;y) + \mathop{\mathbf{Pr}}_{x\sim S}[x\in S_1]\cdot\mathrm{er}_{S_1}(f;y) && (16)\\
&\leq \mathop{\mathbf{Pr}}_{x\sim S}[x\in S_0]\cdot \mathop{\mathbf{Pr}}_{x\sim S_0}[y(x)\neq\bar{f}(x)] + \mathop{\mathbf{Pr}}_{x\sim S}[x\in S_1]\cdot \mathop{\mathbf{Pr}}_{x\sim S_1}[y(x)\neq\bar{f}(x)] && (17)\\
&\leq \mathop{\mathbf{Pr}}_{x\sim S}[x\in S_0]\cdot\left(3\tau_0+\frac{2\beta}{\mathbf{Pr}_{x\sim S}[x\in S_0]}\right)\\
&\qquad\qquad + \mathop{\mathbf{Pr}}_{x\sim S}[x\in S_1]\cdot\left(3\tau_1+\frac{2\beta}{\mathbf{Pr}_{x\sim S}[x\in S_1]}\right) && (18)\\
&= 3\cdot\left(\mathop{\mathbf{Pr}}_{x\sim S}[x\in S_0]\cdot\tau_0+\mathop{\mathbf{Pr}}_{x\sim S}[x\in S_1]\cdot\tau_1\right)+4\beta && (19)\\
&\leq 3\tau+4\beta && (20)
\end{aligned}$$

by the fact that $S$ is partitioned into $S_0$ and $S_1$ and $\tau$ is a corresponding convex combination of $\tau_0$ and $\tau_1$. □

# B Analysis of Algorithm 1

Here, we analyze the sample complexity and running time of Algorithm 1.

**Theorem** (Restatement of Theorem 2)**.** *Let $\alpha, \delta > 0$ and suppose $\mathcal{A}$ agnostic learns a class $\mathcal{C} \subseteq [-1,1]^{\mathcal{X}}$ of dimension $d(\mathcal{C})$. Then, using $\eta = O(\alpha)$, Algorithm 1 converges to a $(\mathcal{C}, \alpha)$-multiaccurate hypothesis $f_T$ in $T = O\left(\frac{\ell_{\mathcal{D}}(f_0;y)}{\alpha^2}\right)$ iterations from $m = \tilde{O}\left(T \cdot \frac{d(\mathcal{C})+\log(1/\delta)}{\alpha^2}\right)$ samples with probability at least $1-\delta$ over the random samples.*

## B.1 Sample complexity

We essentially assume the sample complexity issues away by working with the notion of dimension. We give an example proof outline of a standard uniform convergence argument using metric entropy as in [BLM13].

**Lemma 5.** *Suppose $\mathcal{C} \subseteq [-1,1]^{\mathcal{X}}$ has $\varepsilon$-covering number $N_\varepsilon = \mathcal{N}(\varepsilon, \mathcal{C}, \|\cdot\|_1)$. Then, with probability at least $1-\delta$,*

$$\left| \frac{1}{m} \sum_{i=1}^{m} (c(x_i)y(x_i)) - \mathop{\mathbf{E}}_{x \sim \mathcal{D}}[c(x)y(x)] \right| \leq O(\alpha) \tag{21}$$

*provided $m \geq \tilde{\Omega}\left(\frac{\log(N_{\Theta(\alpha)}/\delta)}{\alpha^2}\right)$.*

*Proof.* The lemma follows from a standard uniform convergence argument. First, observe that because every $c : \mathcal{X} \to [-1,1]$ and $y \in \{0,1\}$ that the empirical estimate using $m$ samples has sensitivity $1/m$. Thus, we can apply McDiarmid's inequality to show concentration of the following statistic.

$$\sup_{c \in \mathcal{C}} \left| \frac{1}{m} \sum_{i=1}^{m} (c(x_i)y(x_i)) - \mathop{\mathbf{E}}_{x \sim \mathcal{X}}[c(x)y(x)] \right| \tag{22}$$

Then, using a standard covering argument, for $N = \mathcal{N}(\varepsilon, \mathcal{C}, \|\cdot\|_1)$ the $\varepsilon$-covering number, we can bound the deviation with high probability. Specifically, taking $O\left(\frac{\log(N/\delta)}{\alpha^2}\right)$ samples guarantees that the empirical estimate for each $c \in \mathcal{C}$ will be within $O(\alpha)$ with probability at least $1-\delta$. Taking $\delta$ small enough to union bound against every iteration and adjusting constants shows gives the lemma. □

Note that this analysis is completely generic, and more sophisticated arguments may improve the resulting bounds that leverage structure in the specific $\mathcal{C}$ of interest.

## B.2 Convergence analysis

We will track progress of Algorithm 1 by tracking the expected cross-entropy loss. We show that every update makes the expected cross-entropy loss decrease significantly. As the loss is bounded below by 0, then positive progress at each iteration combined with an upper bound on the initial loss gives the convergence result.

Note that when we estimate the statistical queries from data, we only have access to approximate answers. Thus, per the sample complexity argument above, we assume that each statistical query is $\alpha/4$-accurate. Further, we will update $f_t$ if we find an update $c_t$ where $\langle c_t, f - y \rangle \geq 3\alpha/4$. Thus, at convergence, it should be clear that the resulting hypothesis will be $(\mathcal{C}, \alpha)$-multiaccurate. The goal is to show that this way, Multiaccuracy Boost converges quickly.

**Lemma 6.** *Let $\alpha > 0$ and suppose $\mathcal{C} \subseteq [-1,1]^{\mathcal{X}}$. Given access to statistical queries that are $\alpha/4$-accurate, Algorithm 1 converges to a $(\mathcal{C}, \alpha)$-multiaccurate hypothesis in $T = O\left(\frac{\ell_{\mathcal{D}}(f_0;y)}{\alpha^2}\right)$ iterations.*

We state this lemma in terms of a class $\mathcal{C}$ but the proof reveals that any nontrivial update that $\mathcal{A}$ returns suffices to make progress.

*Proof.* We begin by considering the effect of the multiplicative weights update as a univariate update rule. Suppose we use the multiplicative weights update rule to compute $f_{t+1}(x)$ to be proportional to $f_t(x) \cdot e^{-\eta c_t(x)}$ for some $c_t(x)$. We can track how $\ell_x(f;y)$ changes based on the choice of $c_t(x)$.

$$\ell_x(f_t;y) - \ell_x(f_{t+1};y)$$
$$= y(x) \cdot \log\left(\frac{f_{t+1}(x)}{f_t(x)}\right) + (1-y(x)) \cdot \log\left(\frac{1-f_{t+1}(x)}{1-f_t(x)}\right) \quad (23)$$

Recall $f_t(x) = \frac{q_t(x)}{1+q_t(x)}$, so $1 - f_t(x) = \frac{1}{1+q_t(x)}$. Thus, we can rewrite (23) as follows.

$$y(x) \cdot \log\left(\frac{q_{t+1}(x)}{q_t(x)}\right) + (1-y(x)) \cdot \log\left(\frac{1}{1}\right) - \log\left(\frac{1+q_{t+1}(x)}{1+q_t(x)}\right) \quad (24)$$
$$= -\eta c_t(x) y(x) + 0 - \log\left(\frac{1+q_{t+1}(x)}{1+q_t(x)}\right) \quad (25)$$

where (25) follows by the multiplicative weights update rule implies $q_{t+1}(x) = e^{-\eta c_t(x)} q_t(x)$ for $x \in S_t$. Next, we expand the final logarithmic term.

$$-\log\left(\frac{1+q_{t+1}(x)}{1+q_t(x)}\right) = -\log\left(\frac{1+q_t(x)e^{-\eta c_t(x)}}{1+q_t(x)}\right) \quad (26)$$
$$\geq -\log\left(\frac{1+q_t(x)(1-\eta c_t(x)+\eta^2 c_t(x)^2)}{1+q_t(x)}\right) \quad (27)$$
$$\geq -\log\left(1 - \frac{q_t(x)}{1+q_t(x)}(\eta c_t(x) - \eta^2 c_t(x)^2)\right) \quad (28)$$
$$\geq \eta c_t(x) f_t(x) - \eta^2 c_t(x)^2 \quad (29)$$

where (27) follows by upper bounding the Taylor series approximation for $e^z$ for $z \geq -1$; and (29) follows by the fact that $f_t(x) \in [0,1]$. Combining the expressions, we can simplify as follows.

$$(25) \geq -\eta c_t(x) y(x) + \eta c_t(x) f_t(x) - \eta^2 c_t(x)^2 \quad (30)$$
$$= \eta c_t(x) \cdot (f_t(x) - y(x)) - \eta^2 c_t(x)^2 \quad (31)$$

Thus, we can express the change in $\ell_x(f_t;y) - \ell_x(f_{t+1};y)$ after an update based on $c_t(x)$ in terms of the inner product between $c_t$ and $f - y$. In this sense, we can express the local progress during the update at time $t$ in terms of some global progress in the objective.

When we update $x \in \mathcal{X}$ simultaneously according to $c$, we can express the change in expected cross-entropy as follows.

$$\ell_{\mathcal{D}}(f_t;y) - \ell_{\mathcal{D}}(f_{t+1};y) \quad (32)$$
$$\geq \eta \cdot \mathop{\mathbf{E}}_{x\sim\mathcal{X}}[c_t(x) \cdot (f_t(x) - y(x))] - \eta^2 \cdot \mathop{\mathbf{E}}_{x\sim\mathcal{X}}[c_t(x)^2] \quad (33)$$
$$\geq \eta \langle c_t, f_t - y \rangle - \eta^2 \quad (34)$$
$$\geq \eta(\alpha/2 - \eta) \quad (35)$$

where (35) follows from the fact that we assumed that our estimates of the statistical queries were $\alpha/4$-accurate and that we update based on $c_t$ if $\langle c_t, f - y\rangle$ is at least $3\alpha/4$ according to our estimates. Thus, taking $\eta = \alpha/4$, then we see the change in expected cross-entropy over $\mathcal{X}$ is at least $\alpha^2/16$, which shows the lemma. □

# C Linear convergence from gradient learning

Here we show that given an auditing algorithm $\mathcal{A}$ that learns the cross-entropy gradients accurately, Algorithm 1 converges linearly. Consider the following auditor $\mathcal{A}_\ell$. We assume the norms and inner products are estimated accurately using $D \sim \mathcal{D}^m$.

**Algorithm 2:** $\mathcal{A}_\ell$ – smooth cross-entropy auditor

**Given:**

- hypothesis $f : \mathcal{X} \to [0, 1]$;
- class of functions $\mathcal{C} \subseteq [-B, B]^{\mathcal{X}}$; accuracy parameter $\alpha > 0$;
- smoothing parameter $L$;
- validation data $D \sim \mathcal{D}^m$;

**Let:**

- $\varepsilon \leftarrow \frac{\langle \nabla_f \ell, f-y\rangle^2}{\|\nabla \ell\|^2 \|f-y\|^2}$ `// approx factor based on angle between grad and f-y`
- $\mathcal{H} \leftarrow \left\{h \in \mathcal{C} : \|h\|^2 \le L \cdot \ell(f; y)\right\}$ `// audit over l2-bounded version of C`
- $h_f \leftarrow \operatorname{argmin}_{h \in \mathcal{H}} \|h - \nabla_f \ell(f; y)\|^2$

if $\ell(f; y) \le \alpha$ or $\|h_f - \nabla_f \ell(f; y)\|^2 > \frac{\varepsilon}{2} \cdot \|\nabla_f \ell(f; y)\|^2$:
    **return** $h(x) = 0$ `// cross-entropy small or hf bad approx to deriv`
else:
    **return** $h_f$

We claim that this auditor learns the partial derivative function in a way that guarantees linear convergence.

**Proposition** (Restatement of Proposition4)**.** *Let $\alpha, B, L > 0$ and $\mathcal{C} \subseteq [-B, B]^{\mathcal{X}}$. Suppose we run Algorithm 1 on initial model $f_0$ with auditor $\mathcal{A}_\ell$ defined in Algorithm 2. Then, Algorithm 1 converges in $T = O\left(L \cdot \log(\ell_{\mathcal{D}}(f_0; y)/\alpha)\right)$ iterations.*

*Proof.* Note that when $\mathcal{A}_\ell$ returns $h(x) = 0$, then Algorithm 1 terminates. Thus, we will bound the number of iterations until $\ell_{\mathcal{D}}(f; y)$ at most than $\alpha$. For notational convenience, we denote $\nabla_f \ell_{\mathcal{D}}(f; y)$ as $\nabla_f \ell$.

By the definition of $\varepsilon$ and the termination condition, we know that if $\mathcal{A}_\ell$ returns $h_f(x) \neq 0$ then

$h_f$ satisfies the following inequality.

$$\|h_f - \nabla_f \ell\|^2 \leq \frac{1}{2} \cdot \frac{\langle \nabla_f \ell, f - y \rangle^2}{\|f - y\|^2} \tag{36}$$

$$\leq \frac{1}{2} \cdot \frac{\langle \nabla_f \ell, f - y \rangle^2}{\|f - y\|^2} + \frac{1}{16} \|\nabla_f \ell\|^2 \tag{37}$$

$$= \left\| \frac{\langle \nabla_f \ell, f - y \rangle}{\|f - y\|^2} (f - y) - \frac{\nabla_f \ell}{4} \right\|^2 \tag{38}$$

Using this inequality, we can bound the inner product between $h_f$ and $f - y$.

$$\langle h_f, f - y \rangle \tag{39}$$

$$= \langle \nabla_f \ell, f - y \rangle + \langle h_f - \nabla_f \ell, f - y \rangle \tag{40}$$

$$\geq \langle \nabla_f \ell, f - y \rangle - \left\| \frac{\langle \nabla_f \ell, f - y \rangle}{\|f - y\|^2} (f - y) - \frac{\nabla_f \ell}{4} \right\| \cdot \|f - y\| \tag{41}$$

$$\geq \langle \nabla_f \ell, f - y \rangle - \langle \nabla_f \ell, f - y \rangle \cdot \frac{\|f - y\|^2}{\|f - y\|^2} + \frac{1}{4} \cdot \langle \nabla_f \ell, f - y \rangle \tag{42}$$

$$\geq \frac{1}{4} \cdot \ell_{\mathcal{D}}(f; y) \tag{43}$$

where (42) follows from the fact that $\nabla_f \ell$ and $f - y$ are positively correlated; and (43) follows by convexity of $\ell_{\mathcal{D}}$.

Thus, using the analysis of the multiplicative weights update from Section B, we can see that the progress in cross-entropy can be bounded as

$$\ell_{\mathcal{D}}(f_t; y) - \ell_{\mathcal{D}}(f_{t+1}; y) \geq \frac{\eta}{4} \cdot \ell_{\mathcal{D}}(f_t; y) - \eta^2 \cdot \|h_{f_t}(x)\|^2 \tag{44}$$

$$\geq (\frac{\eta}{4} - \eta^2 L) \cdot \ell_{\mathcal{D}}(f_t; y) \tag{45}$$

where (45) follows from the fact that $h_f$ is drawn from a class with Euclidean norm bounded as $\|h_f\|^2 \leq L \cdot \ell_{\mathcal{D}}(f; y)$.

Rearranging and taking $\eta = \frac{1}{8L}$, we arrive at the following inequality that implies linear convergence.

$$\ell_{\mathcal{D}}(f_{t+1}; y) \leq (1 - \frac{\eta}{4} + \eta^2 L) \ell_{\mathcal{D}}(f_t; y) \tag{46}$$

$$\leq e^{-1/64L} \ell_{\mathcal{D}}(f_t; y) \tag{47}$$

Thus, after $O\left(L \cdot \log(\ell_{\mathcal{D}}(f_0; y)/\alpha)\right)$, then the cross-entropy will drop below $\alpha$. □