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

Michael P. Kim*
Stanford University
mpk@cs.stanford.edu

Amirata Ghorbani
Stanford University
amiratag@stanford.edu

James Zou
Stanford University
jamesz@stanford.edu

Abstract

Machine learning predictors are successfully deployed in applications ranging from disease diagnosis, to predicting credit scores, to image recognition. Even when the overall accuracy is high, the predictions often have systematic biases that harm specific subgroups, especially for subgroups that are minorities in the training data. We develop a rigorous framework of *multiaccuracy* auditing and post-processing to improve predictor accuracy 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. We prove guarantees on the convergence rate of the algorithm and show that it improves overall accuracy at each step. Importantly, if the initial model is accurate on an identifiable subgroup, then the post-processed model will be also. We demonstrate the effectiveness of this approach on diverse applications in image classification, finance, and population health. MULTIACCURACY BOOST can improve subpopulation accuracy (e.g. for "black women") even when the sensitive features (e.g. race, gender) are not known to the algorithm.

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 (Buolamwini and Gebru 2018), 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%!

*Supported by a Google Faculty Research Award and 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 (Hébert-Johnson et al. 2018), 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. Multiaccuracy auditing makes no assumptions about the original classifier; in particular, it can handle inadvertent and malicious forms of discrimination.

Our contributions We develop a framework for auditing and post-processing prediction models for multiaccuracy. We describe a new algorithm, MULTIACCURACY BOOST, where a simple learning algorithm – the auditor – is used to identify subpopulations where f_0 is systematically biased. 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 (Dwork et al. 2012; Hardt, Price, and Srebro 2016; Kearns et al. 2017), and is reasonable in settings such as gender detection where we would like to boost the classifier's accuracy across 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 an experimental case study based on the Gender Shades example (Buolamwini and Gebru 2018).[1] We train an initial prediction model that achieve good overall classification error, but exhibits biases against minority 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, enforcing multiaccuracy 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. We also show that the auditing process may be useful in understanding why prediction models are making mistakes. Specifically, the multiaccuracy auditor can be used to produce examples of inputs where the predictor is erring significantly. These results are reported in Section 4 and the appendix.

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 prediction 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.

[1] We also evaluate performance on a semi-synthetic medical diagnosis task and adult income prediction, reported in the Appendix.

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.

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 (Hébert-Johnson et al. 2018).

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 $c \in \mathcal{C}$:*

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

$(\mathcal{C}, \alpha)$-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 χ_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 α-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 depth-4 decision trees, then multiaccuracy guarantees unbiasedness, not just on the marginal populations defined by race and separately gender, but by the subpopulations defined by the combinations 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.

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 (Kearns et al. 2017; Hébert-Johnson et al. 2018; Kim, Reingold, and Rothblum 2018), albeit for different tasks.

To achieve $(\mathcal{C}, \alpha)$-multiaccuracy we could audit with a naive learning algorithm that iterates over statistical tests $c \in \mathcal{C}$. Given an algorithm $\mathcal{A}$ that efficiently learns the class $\mathcal{C}$, we can speed up the auditing process; for instance, if we take $\mathcal{C}$ to be the class of linear tests, we can use efficient algorithms for linear regression to audit. 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.

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 γ, the density of the protected subpopulation S, grows. For instance, in our case study on gender detection, we train on a large unbalanced data set, but then audit using a data set of balanced diversity collected for the Gender Shades study (Buolamwini and Gebru 2018).

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 $\mathcal{A} : (\mathcal{X} \times [-1, 1])^m \to [-1, 1]^{\mathcal{X}}$ that learns a class $\mathcal{C}$, and for any accuracy parameter $\alpha > 0$, outputs a hypothesis $f : \mathcal{X} \to [0, 1]$ that is $(\mathcal{C}, \alpha)$-multiaccurate. The post-processing algorithm is an iterative procedure similar to boosting (Schapire and Freund 2012), that uses the multiplicative weights framework to improve suboptimal predictions identified by the auditor. This approach is similar to the algorithm given in (Hébert-Johnson et al. 2018) in the context of fairness and (Trevisan, Tulsiani, and Vadhan 2009) 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 $S \in \mathcal{S}$:
  $h_{t,S} \leftarrow \mathcal{A}(D_t; (f_t - y)_S)$
  `// audit ft on X,X0,X1 with fresh data`
- $S^* \leftarrow \mathrm{argmax}_{S\in\mathcal{S}}\, \mathbf{E}_{x\sim D_t}[h_{t,S}(x) \cdot (f_t(x) - y(x))]$
  `// 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`

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.

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 tth 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.

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 α, and the size of subpopulations we wish to protect γ. To guarantee good generalization, 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 (Dwork et al. 2015; Russo and Zou 2016). In practice, we need to balance the choice of $\mathcal{C}$ 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.

From the stopping condition, it is clear that when the algorithm terminates, f_T will be $(\mathcal{C}, \alpha)$-multiaccurate. 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 we can estimate accurately and efficiently from a small sample. We provide formal guarantees on the convergence rate and the sample complexity from $\mathcal{D}$ in the Appendix.

Do no harm. The distinction between our approach and most prior works on fairness (especially (Kearns et al. 2017)) is made clear from the "do-no-harm" property that MULTIACCURACY BOOST exhibits, stated formally as Theorem 2. 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. 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 2 reveals that this worst-case bound is very pessimistic and can be improved with stronger assumptions. 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.

Theorem 2 (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*

$$\mathrm{er}_S(f; y) \leq 3 \cdot \mathrm{er}_S(f_0; y) + 4\beta. \tag{2}$$

4 Case Study: Gender Detection

We aim to replicate the conditions of the Gender Shades study (Buolamwini and Gebru 2018), to test the effectiveness of multiaccuracy auditing and post-processing on this important real-world example.[2] For our initial model, we train an inception-resnet-v1 (Szegedy et al. 2017) gender classification model using the CelebA data set with more than 200,000 face images (Liu et al. 2015). The resulting test accuracy on CelebA for binary gender classification is 98.4%. Even though the overall accuracy of this f_0 is high, the error rate is much worse for females compared to males and worse for blacks compared to non-blacks; these results are qualitatively very similar to those observed by the commercial gender detection systems.

We applied MULTIACCURACY BOOST using the PPB data set (developed by Buolamwini and Gebru) which has balanced representation across gender and race. Specifically, we audit using ridge regression. Instead of auditing over the raw input pixels, we use a representation derived from a variational autoencoder (VAE) trained on CelebA dataset using Facenet (Schroff, Kalenichenko, and Philbin 2015) library. The PPB data set is very small; thus, this experiment can be viewed as a stress test to evaluate 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**).

In addition to evaluating the effectiveness of the multiaccuracy approach, we compare our post-processing results 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, whereas the auditor is "blind" – it is not explicitly given the race or gender and knows nothing about the inner workings of f_0.

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

Table 1: **Results for the PPB 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 $\mathcal{D}$.

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

[2]See the appendix for two other case studies.

5 iterations and substantially reduced error across subpopulations. We report the overall classification accuracy as well as accuracy on different subpopulations We report the population percentage in $\mathcal{D}$, accuracy of the initial model, our black-box post-processed model, and white-box benchmarks in Table 1 for each subpopulations – e.g. **DF** indicates dark-skinned female. In particular, we highlight the subpopulations of **DF** and **LM**; the classification error improves significantly on **DF** but does not hurt the accuracy on **LM** significantly.

Multiaccuracy auditing as diagnostic As was shown in (Buolamwini and Gebru 2018), 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. We test this idea on the PPB data set, evaluating the test images' representations with the hypotheses the auditor returns.

In Figure 1, 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 the auditor. In the first round of auditing, the three highest-scoring images (top row) are all women, both black and white. The least active images (second 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 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.

5 Discussion

In this work, we propose multiaccuracy as a framework 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.

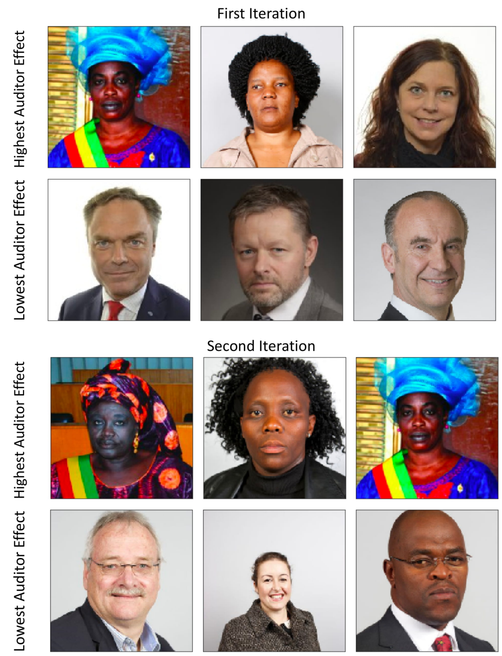

Figure 1: **Interpreting Auditors**
PPB test images of largest and smallest bias detected by the auditor for the first (rows 1-2) and second (rows 3-4) rounds of auditing.

Related works

Many different notions of fairness have been proposed in literature on learning and classification (Dwork et al. 2012; Hardt, Price, and Srebro 2016; Zemel et al. 2013; Dwork et al. 2017; Hébert-Johnson et al. 2018; Kearns et al. 2017; Hashimoto et al. 2018; Kim, Reingold, and Rothblum 2018; Rothblum and Yona 2018). Many of these works encode some notion of parity, e.g. 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, the classifier's performance must degrade.

A notable exception to this pattern is the work of Hébert-Johnson *et al.* (Hébert-Johnson et al. 2018), which first introduced a variant of *multiaccuracy*and *multicalibration* in the context of regression tasks. (Hébert-Johnson et al. 2018) 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 (Hébert-Johnson et al. 2018), adapting their work to the binary classification setting. Our post-processing algorithm, MULTIACCURACY BOOST, builds on the algorithm given in

(Hébert-Johnson et al. 2018), 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.* (Kearns et al. 2017) also investigated how to achieve statistical fairness guarantees, not just for traditionally-protected groups, but on rich families of subpopulations. (Kearns et al. 2017) proposed a framework for *auditing* and learning models to achieve fairness notions like statistical parity and equal false positive rates. Both works (Hébert-Johnson et al. 2018; Kearns et al. 2017) connect the task of learning a model that satisfies the notion of fairness to the task of (weak) agnostic learning (Kearns 1998; Kearns, Schapire, and Sellie 1994; Kalai, Mansour, and Verbin 2008; Feldman 2010). (Kearns et al. 2017) 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 differs from (Kearns et al. 2017) in important ways. In their framework, 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, 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. Additionally, by applying mutliaccuracy over real-valued functions (as opposed to boolean functions), we can get efficient algorithms that provably satisfy $(\mathcal{C}, \alpha)$-multiaccuracy for nontrivial classes $\mathcal{C}$ (specifically, linear hypotheses).

At a technical level, our post-processing algorithm is similar to gradient boosting (Mason et al. 2000; Friedman 2001). Still, our perspective is quite different from the typical boosting setting. Rather than using an expressive class of predictors as the base classifiers to learn the function directly, we focus 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.

A different approach to subgroup fairness is studied by Dwork *et al.* (Dwork et al. 2017). 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 the attributes of interest in the data set and that the groups we wish to protect are partitioned by these attributes. Even if this information is available, *a priori*, it may not always be obvious which subpopulations require special attention. In contrast, multiaccuracy 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 (Dwork et al. 2017).

The present work, along with (Hébert-Johnson et al. 2018; Kearns et al. 2017; Kim, Reingold, and Rothblum 2018), 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 (Dwork et al. 2012; Zemel et al. 2013; Buolamwini and Gebru 2018; Hardt, Price, and Srebro 2016; Dwork et al. 2017; Kim, Reingold, and Rothblum 2018; Rothblum and Yona 2018), discussions of the utility-fairness tradeoffs in fair classification (Angwin et al. 2016; Kleinberg, Mullainathan, and Raghavan 2017; Chouldechova 2017; Chouldechova and G'Sell 2016; Corbett-Davies et al. 2017; Pleiss et al. 2017). 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 (Lipton, Wang, and Smola 2018), but when the learner gets a small number of labeled samples from the new distribution.

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 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 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 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 from a small amount of audit data. Thus, multiaccuracy auditing is limited: if the mistakes appear overly-complicated to the bounded auditor, then the auditor will not be able to identify these mistakes. Our empirical results suggest, that the subpopulations on which a classifier errs may be 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 PPB test sets suggests a lightweight black-box alternative to more sophisticated transfer learning techniques, which may warrant further investigation.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*.

Boucheron, S.; Lugosi, G.; and Massart, P. 2013. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

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

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

Chouldechova, A., and G'Sell, M. 2016. Fairer and more accurate, but for whom? *FATML*.

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

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

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

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

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

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

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232.

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

Hashimoto, T. B.; Srivastava, M.; Namkoong, H.; and Liang, P. 2018. Fairness without demographics in repeated loss minimization. *ICML*.

Hébert-Johnson, Ú.; Kim, M. P.; Reingold, O.; and Rothblum, G. N. 2018. Calibration for the (computationally-identifiable) masses. *ICML*.

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

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144v3*.

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

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

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

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

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

Lipton, Z. C.; Wang, Y.; and Smola, A. J. 2018. Detecting and correcting for label shift with black box predictors. *ICML*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Mason, L.; Baxter, J.; Bartlett, P. L.; and Frean, M. R. 2000. Boosting algorithms as gradient descent. In *Advances in neural information processing systems*, 512–518.

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

Rothblum, G. N., and Yona, G. 2018. Probably approximately metric-fair learning. *ICML*.

Russo, D., and Zou, J. 2016. How much does your data exploration overfit? controlling bias via information usage. *AISTATS*.

Schapire, R. E., and Freund, Y. 2012. *Boosting: Foundations and algorithms*. MIT press.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. *CoRR* abs/1503.03832.

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

Sudlow, C.; Gallacher, J.; Allen, N.; Beral, V.; Burton, P.; Danesh, J.; Downey, P.; Elliott, P.; Green, J.; Landray, M.; et al. 2015. 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.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, 12.

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

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 325–333.

A Additional Experiments

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 (Kohavi 1996) 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 2, 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.

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 (Sudlow et al. 2015). 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 3 shows, MULTIACCURACY BOOST significantly lowers the classification error in the old female population.

	All	**F**	**M**	**B**	**W**	**BF**	**WM**
$\mathcal{D}$	100	32.3	67.7	9.7	90.3	4.8	62.9
f_0	19.3	9.3	24.2	10.5	20.3	4.8	24.9
MA	14.7	7.2	18.3	9.4	15.0	4.5	18.3
SS	19.7	9.5	24.6	10.5	19.9	5.5	25.3

Table 2: **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.

	All	**F**	**M**	**O**	**Y**	**OF**	**YM**
$\mathcal{D}$	100	39.6	60.4	34.6	65.4	15.0	40.7
f_0	18.9	29.4	12.2	21.9	17.3	36.8	12.8
MA	16.0	24.1	10.7	16.4	15.7	26.5	11.6
SS	19.5	32.4	11.0	22.1	18.1	37.6	11.3

Table 3: **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.

B Formal Guarantees of Multiaccuracy

Additional notation. 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 $\ell_x(f;y) = -y(x)\cdot\log(f(x)) - (1-y(x))\cdot\log(1-f(x))$. We use the inner product $\langle h,g\rangle = \mathbf{E}_{x\sim\mathcal{D}}[h(x)\cdot g(x)]$ and the p-norms $\|h\|_p = \left(\mathbf{E}_{x\sim\mathcal{D}}[|h(x)|^p]\right)^{1/p}$.

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.

$$\begin{aligned}
&\mathop{\mathbf{E}}_{x\sim\mathcal{D}}[c(x)\cdot(f(x)-y(x))] && (3)\\
&= \mathop{\mathbf{E}}_{x\sim\mathcal{D}}[\hat{y}(x)_S\cdot(f(x)-y(x))] + \mathop{\mathbf{E}}_{x\sim\mathcal{D}}[z(x)\cdot(f(x)-y(x))] && (4)\\
&\geq \beta_{01}\cdot\mathop{\mathbf{E}}_{x\sim S_{01}}[f(x)-y(x)] + \beta_{10}\cdot\mathop{\mathbf{E}}_{x\sim S_{10}}[y(x)-f(x)] - \tau && (5)
\end{aligned}$$

where (5) 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

$$\mathrm{er}_S(f;y) = (\beta_{01}+\beta_{10})/\gamma \leq 2\cdot(\alpha+\tau)/\gamma. \tag{6}$$

□

Theorem 2 follows by a similar argument.

Theorem (Restatement of Theorem 2)**.** *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*

$$\mathrm{er}_S(f;y) \leq 3\cdot\mathrm{er}_S(f_0;y) + 4\beta. \tag{7}$$

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 $\mathbf{E}_{x\sim S_1}[-\chi_S(x)\cdot(f(x)-y(x))]$.

$$\begin{aligned}
&\mathop{\mathbf{E}}_{x\sim S_1}[-\chi_S(x)\cdot(f(x)-y(x))] && (8)\\
&= \mathop{\mathbf{E}}_{x\sim S_1}[y(x)-f(x)] && (9)\\
&= \mathop{\mathbf{Pr}}_{x\sim S_1}[y(x)=1]\cdot\mathop{\mathbf{E}}_{\substack{x\sim S_1\\ y(x)=1}}[1-f(x)] \\
&\qquad - \mathop{\mathbf{Pr}}_{x\sim S_1}[y(x)=0]\cdot\mathop{\mathbf{E}}_{\substack{x\sim S_1\\ y(x)=0}}[f(x)] && (10)\\
&\geq \mathop{\mathbf{Pr}}_{x\sim S_1}[y(x)=1\wedge\bar{f}(x)=0]\cdot\mathop{\mathbf{E}}_{\substack{x\sim S_1\\ y(x)=1\\ \bar{f}(x)=0}}[1-f(x)] - \tau_1 && (11)\\
&\geq \frac{1}{2}\mathop{\mathbf{Pr}}_{x\sim S_1}[y(x)=1\wedge\bar{f}(x)=0] - \tau_1 && (12)
\end{aligned}$$

where (11) follows from applying Hölder's inequality and the assumption that $\mathrm{er}_{S_1}(f_0;y) = \tau_1$; and (12) 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

$$\mathrm{er}_{S_1}(f;y) \leq \frac{2\beta}{\mathbf{Pr}_{x\sim S}[x \in S_1]} + 3\tau_1 \tag{13}$$

where the additional τ_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.

$$\begin{aligned}
&\mathrm{er}_S(f;y) && (14)\\
&= \mathop{\mathbf{Pr}}_{x\sim S}[x \in S_0]\cdot\mathrm{er}_{S_0}(f;y) + \mathop{\mathbf{Pr}}_{x\sim S}[x \in S_1]\cdot\mathrm{er}_{S_1}(f;y) && (15)\\
&\leq \mathop{\mathbf{Pr}}_{x\sim S}[x \in S_0]\cdot\mathop{\mathbf{Pr}}_{x\sim S_0}[y(x)\neq\bar{f}(x)] \\
&\qquad + \mathop{\mathbf{Pr}}_{x\sim S}[x \in S_1]\cdot\mathop{\mathbf{Pr}}_{x\sim S_1}[y(x)\neq\bar{f}(x)] && (16)\\
&\leq \mathop{\mathbf{Pr}}_{x\sim S}[x \in S_0]\cdot\left(3\tau_0 + \frac{2\beta}{\mathbf{Pr}_{x\sim S}[x \in S_0]}\right) \\
&\qquad + \mathop{\mathbf{Pr}}_{x\sim S}[x \in S_1]\cdot\left(3\tau_1 + \frac{2\beta}{\mathbf{Pr}_{x\sim S}[x \in S_1]}\right) && (17)\\
&= 3\cdot\left(\mathop{\mathbf{Pr}}_{x\sim S}[x \in S_0]\cdot\tau_0 + \mathop{\mathbf{Pr}}_{x\sim S}[x \in S_1]\cdot\tau_1\right) + 4\beta && (18)\\
&\leq 3\tau + 4\beta && (19)
\end{aligned}$$

by the fact that S is partitioned into S_0 and S_1 and τ is a corresponding convex combination of τ_0 and τ_1. □

Analysis of Algorithm 1

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

Theorem 3. *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.

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 (Boucheron, Lugosi, and Massart 2013).

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

$$\left|\frac{1}{m}\sum_{i=1}^{m}\left(c(x_i)y(x_i)\right) - \mathop{\mathbf{E}}_{x\sim\mathcal{D}}[c(x)y(x)]\right| \leq O(\alpha) \qquad (20)$$

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.

$$\sup_{c\in\mathcal{C}}\left|\frac{1}{m}\sum_{i=1}^{m}\left(c(x_i)y(x_i)\right) - \mathop{\mathbf{E}}_{x\sim\mathcal{X}}[c(x)y(x)]\right| \qquad (21)$$

Then, using a standard covering argument, for $N = \mathcal{N}(\varepsilon, \mathcal{C}, \|\cdot\|_1)$ the ε-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 δ small enough to union bound against every iteration and adjusting constants shows gives the lemma. $\square$

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.

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. *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 $c_t(x)$.

$$\begin{aligned}&\ell_x(f_t;y) - \ell_x(f_{t+1};y)\\ &= y(x)\cdot\log\left(\frac{f_{t+1}(x)}{f_t(x)}\right) + (1-y(x))\cdot\log\left(\frac{1-f_{t+1}(x)}{1-f_t(x)}\right)\end{aligned} \qquad (22)$$

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 (22) as follows.

$$(22) = y(x)\cdot\log\left(\frac{q_{t+1}(x)}{q_t(x)}\right) + (1-y(x))\cdot\log\left(\frac{1}{1}\right) - \log\left(\frac{1+q_{t+1}(x)}{1+q_t(x)}\right) \qquad (23)$$

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

where (24) 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.

$$-\log\left(\frac{1+q_{t+1}(x)}{1+q_t(x)}\right) \qquad (25)$$

$$= -\log\left(\frac{1+q_t(x)e^{-\eta c_t(x)}}{1+q_t(x)}\right) \qquad (26)$$

$$\geq -\log\left(\frac{1+q_t(x)(1-\eta c_t(x)+\eta^2 c_t(x)^2)}{1+q_t(x)}\right) \qquad (27)$$

$$\geq -\log\left(1 - \frac{q_t(x)}{1+q_t(x)}(\eta c_t(x) - \eta^2 c_t(x)^2)\right) \qquad (28)$$

$$\geq \eta c_t(x) f_t(x) - \eta^2 c_t(x)^2 \qquad (29)$$

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.

$$(24) \geq -\eta c_t(x)y(x) + \eta c_t(x)f_t(x) - \eta^2 c_t(x)^2 \qquad (30)$$

$$= \eta c_t(x)\cdot(f_t(x) - y(x)) - \eta^2 c_t(x)^2 \qquad (31)$$

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.

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

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. □

Linear convergence from gradient learning

Here, we propose auditing with an algorithm $\mathcal{A}_\ell$ 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 (Shalev-Shwartz and others 2012; Bubeck and others 2015). This application of MULTIACCURACY BOOST is similar in spirit to gradient boosting techniques (Mason et al. 2000; Friedman 2001), 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. Consider the following auditor $\mathcal{A}_\ell$. We assume the norms and inner products are estimated accurately using $D \sim \mathcal{D}^m$.

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

Proposition 4. *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 α. For notational convenience, we denote $\nabla_f \ell_{\mathcal{D}}(f; y)$ as $\nabla_f \ell$.

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

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 from angle between grad and f-y`
- $\mathcal{H} \leftarrow \left\{ h \in \mathcal{C} : \|h\|^2 \leq L \cdot \ell(f; y) \right\}$
  `// audit over l2-bounded version of C`
- $h_f \leftarrow \mathrm{argmin}_{h \in \mathcal{H}} \|h - \nabla_f \ell(f; y)\|^2$

if $\ell(f; y) \leq \alpha$ 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

$$\|h_f - \nabla_f \ell\|^2 \leq \frac{1}{2} \cdot \frac{\langle \nabla_f \ell, f - y\rangle^2}{\|f - y\|^2} \tag{36}$$
$$\leq \frac{1}{2} \cdot \frac{\langle \nabla_f \ell, f - y\rangle^2}{\|f - y\|^2} + \frac{1}{16}\|\nabla_f \ell\|^2 \tag{37}$$
$$= \left\| \frac{\langle \nabla_f \ell, f - y\rangle}{\|f - y\|^2}(f - y) - \frac{\nabla_f \ell}{4} \right\|^2 \tag{38}$$

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 earlier, 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 α. $\square$