Multicalibration as Boosting for Regression

Ira Globus-Harris Declan Harrison Michael Kearns Aaron Roth
Jessica Sorrell

February 1, 2023

Abstract

We study the connection between multicalibration and boosting for squared error regression. First we prove a useful characterization of multicalibration in terms of a "swap regret" like condition on squared error. Using this characterization, we give an exceedingly simple algorithm that can be analyzed both as a boosting algorithm for regression and as a multicalibration algorithm for a class $\mathcal{H}$ that makes use only of a standard squared error regression oracle for $\mathcal{H}$. We give a weak learning assumption on $\mathcal{H}$ that ensures convergence to Bayes optimality without the need to make any realizability assumptions — giving us an agnostic boosting algorithm for regression. We then show that our weak learning assumption on $\mathcal{H}$ is both necessary and sufficient for multicalibration with respect to $\mathcal{H}$ to imply Bayes optimality. We also show that if $\mathcal{H}$ satisfies our weak learning condition relative to another class $\mathcal{C}$ then multicalibration with respect to $\mathcal{H}$ implies multicalibration with respect to $\mathcal{C}$. Finally we investigate the empirical performance of our algorithm experimentally using an open source implementation that we make available on GitHub[1].

1 Introduction

We revisit the problem of boosting for regression, and develop a new agnostic regression boosting algorithm via a connection to multicalibration. In doing so, we shed additional light on multicalibration, a recent learning objective that has emerged from the algorithmic fairness literature [Hébert-Johnson et al., 2018]. In particular, we characterize multicalibration in terms of a "swap-regret" like condition, and use it to answer the question "what property must a collection of functions $\mathcal{H}$ have so that multicalibration with respect to $\mathcal{H}$ implies Bayes optimality?", giving a complete answer to problem asked by Burhanpurkar et al. [2021]. Using our swap-regret characterization, we derive an especially simple algorithm for learning a multicalibrated predictor for a class of functions $\mathcal{H}$ by reduction to a standard squared-error regression algorithm for $\mathcal{H}$. The same algorithm can also be analyzed as a boosting algorithm for squared error regression that makes calls to a weak learner for squared error regression on subsets of the original data distribution *without the need to relabel examples* (in contrast to Gradient Boosting as well as existing multicalibration algorithms). This lets us specify a weak learning condition that is sufficient for convergence to the Bayes optimal predictor (even if the Bayes optimal predictor does not have zero error), avoiding the kinds of realizability assumptions that are implicit in analyses of boosting algorithms that converge to zero error. We conclude that ensuring multicalibration with respect to $\mathcal{H}$ corresponds to boosting for squared error regression in which $\mathcal{H}$ forms the set of weak learners. Finally we define a weak learning condition for $\mathcal{H}$ relative to a constrained class of functions $\mathcal{C}$ (rather than with respect to the Bayes optimal predictor). We show that multicalibration with respect to $\mathcal{H}$ implies multicalibration with respect to $\mathcal{C}$ if $\mathcal{H}$ satisfies the weak learning condition with respect to $\mathcal{C}$, which in turn implies accuracy at least that of the best function in $\mathcal{C}$.

Multicalibration Consider a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ defined over a domain $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ of feature vectors $x \in \mathcal{X}$ paired with real valued labels y. Informally, a regression function $f : \mathcal{X} \to \mathbb{R}$ is *calibrated* if for every v in the range of f, $\mathbb{E}_{(x,y)\sim\mathcal{D}}[y|f(x) = v] = v$. In other words, $f(x)$ must be an unbiased estimator of y,

[1]Our code repository can be found at `https://github.com/Declancharrison/Level-Set-Boosting`

even conditional on the value of its own prediction. Calibration on its own is a weak condition, because it only asks for f to be unbiased on *average* over all points x such that $f(x) = v$. For example, the constant predictor that predicts $f(x) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y]$ is calibrated. Thus calibration does not imply accuracy—a calibrated predictor need not make predictions with lower squared error than the best constant predictor. Calibration also does not imply that f is equally representative of the label distribution on different subsets of the feature space $\mathcal{X}$. For example, given a subset of the feature space $G \subseteq \mathcal{X}$, even if f is calibrated, it may be that f is not calibrated on the conditional distribution conditional on $x \in G$—it might be e.g. that $\mathbb{E}[y|f(x) = v, x \in G] \gg v$, and $\mathbb{E}[y|f(x) = v, x \notin G] \ll v$. To correct this last deficiency, Hébert-Johnson et al. [2018] defined *multi-calibration*, which is a condition parameterized by a subset of groups $G \subseteq \mathcal{X}$ each defined by an indicator function $h : \mathcal{X} \to \{0, 1\}$ in some class $\mathcal{H}$. It asks (informally) that for each such $h \in \mathcal{H}$, and for each v in the range of f, that $\mathbb{E}[h(x)(y-v)|f(x) = v] = 0$. Since h is a binary indicator function for some set G, this is equivalent to asking for calibration not just marginally over $\mathcal{D}$, but simultaneously for calibration over $\mathcal{D}$ conditional on $x \in G$. Kim et al. [2019] and Gopalan et al. [2022] generalize multicalibration beyond group indicator functions to arbitrary real valued functions $h : \mathcal{X} \to \mathbb{R}$. Intuitively, as $\mathcal{H}$ becomes a richer and richer set of functions, multicalibration becomes an increasingly stringent condition. But if $\mathcal{H}$ consists of the indicator functions for e.g. even a very large number of *randomly selected* subsets $G \subseteq \mathcal{X}$, then the constant predictor $f(x) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y]$ will still be approximately multicalibrated with respect to $\mathcal{H}$. What property of $\mathcal{H}$ ensures that multicalibration with respect to $\mathcal{H}$ implies that f is a Bayes optimal regression function? This question was recently asked by Burhanpurkar et al. [2021] — and we provide a necessary and sufficient condition.

Boosting for Regression Boosting refers broadly to a collection of learning techniques that reduce the problem of "strong learning" (informally, finding an error optimal model) to a series of "weak learning" tasks (informally, finding a model that has only a small improvement over a trivial model)—See Schapire and Freund [2013] for a textbook treatment. The vast majority of theoretical work on boosting studies the problem of binary classification, in which a weak learner is a learner that obtains classification error bounded below 1/2. Several recent papers Kim et al. [2019], Gopalan et al. [2022] have made connections between algorithms for guaranteeing multicalibration and boosting algorithms for binary classification.

In this paper, we show a direct connection between multicalibration and the much less well-studied problem of boosting for squared error regression [Friedman, 2001, Duffy and Helmbold, 2002]. There is not a single established notion for what constitutes a weak learner in the regression setting (Duffy and Helmbold [2002] introduce several different notions), and unlike boosting algorithms for classification problems which often work by calling a weak learner on a reweighting of the data distribution, existing algorithms for boosting for regression typically resort to calling a learning algorithm on *relabelled* examples. We give a boosting algorithm for regression that only requires calling a squared error regression learning algorithm on subsets of examples from the original distribution (without relabelling), which lets us formulate a weak learning condition that is sufficient to converge to the Bayes optimal predictor, without making the kinds of realizability assumptions implicit in the analysis of boosting algorithms that assume one can drive error to zero.

1.1 Our Results

We focus on classes of real valued functions $\mathcal{H}$ that are closed under affine transformations — i.e. classes such that if $f(x) \in \mathcal{H}$, then for any pair of constants $a, b \in \mathbb{R}$, $(af(x) + b) \in \mathcal{H}$ as well. Many natural classes of models satisfy this condition already (e.g. linear and polynomial functions and regression trees), and any neural network architecture that does not already satisfy this condition can be made to satisfy it by adding two additional parameters (a and b) while maintaining differentiability. Thus we view closure under affine transformations to be a weak assumption that is enforceable if necessary.

First in Section 3 we prove the following characterization for multicalibration over $\mathcal{H}$, for any class $\mathcal{H}$ that is closed under affine transformations. Informally, we show that a model f is multicalibrated with respect

to $\mathcal{H}$ if and only if, for every v in the range of f:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2|f(x)=v] \leqslant \min_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h(x)-y)^2|f(x)=v]$$

(See Theorem 3.2 for the formal statement). This is a "swap regret"-like condition (as in Foster and Vohra [1999] and Blum and Mansour [2005]), that states that f must have lower squared error than any model $h \in \mathcal{H}$, even conditional on its own prediction. Using this characterization, in Section 4 we give an exceedingly simple algorithm for learning a multicalibrated predictor over $\mathcal{H}$ given a squared error regression oracle for $\mathcal{H}$. The algorithm simply repeats the following over t rounds until convergence, maintaining a model $f : \mathcal{X} \to \{0, 1/m, 2/m, \ldots, 1\}$ with a discrete range with support over multiples of $1/m$ for some discretization factor m:

1. For each level set $v \in \{0, 1/m, 2/m, \ldots, 1\}$, run a regression algorithm to find the $h_v^t \in \mathcal{H}$ that minimizes squared error on the distribution $\mathcal{D}|(f_{t-1}(x) = v)$, the distribution conditional on $f_{t-1}(x) = v$.
2. Replace each level set v of $f_{t-1}(x)$ with $h_v^t(x)$ to produce a new model f_t, and round its output to the discrete range $\{0, 1/m, 2/m, \ldots, 1\}$

Each iteration decreases the squared error of f_t, ensuring convergence, and our characterization of multicalibration ensures that we are multicalibrated with respect to $\mathcal{H}$ at convergence. Compared to existing multicalibration algorithms (e.g. the split and merge algorithm of Gopalan et al. [2022]), our algorithm is exceptionally simple and makes use of a standard squared-error regression oracle on subsets of the original distribution, rather than using a classification oracle or requiring example relabelling.

We can also view the same algorithm as a boosting algorithm for squared error regression. Suppose $\mathcal{H}$ (or equivalently our weak learning algorithm) satisfies the following weak learning assumption: informally, that on any restriction of $\mathcal{D}$ on which the Bayes optimal predictor is non-constant, there should be some $h \in \mathcal{H}$ that obtains squared error better than that of the best constant predictor. Then our algorithm converges to the Bayes optimal predictor. In Section A we give uniform convergence bounds which guarantee that the algorithm's accuracy and multicalibration guarantees generalize out of sample, with sample sizes that are linear in the pseudodimension of $\mathcal{H}$.

We then show in Section 5 that in a strong sense this is the "right" weak learning assumption: Multicalibration with respect to $\mathcal{H}$ implies Bayes optimality if and only if $\mathcal{H}$ satisfies this weak learning condition. This gives a complete answer to the question of when multicalibration implies Bayes optimality.

In Section 6, we generalize our weak learning condition to a weak learning condition relative to a constrained class of functions $\mathcal{C}$ (rather than relative to the Bayes optimal predictor), and show that if $\mathcal{H}$ satisfies the weak learning condition relative to $\mathcal{C}$, then multicalibration with respect to $\mathcal{H}$ implies multicalibration with respect to $\mathcal{C}$, and hence error that is competitive with the best model in $\mathcal{C}$.

We give a fast, parallelizable implementation of our algorithm and in Section 7 demonstrate its convergence to Bayes optimality on two-dimensional datasets useful for visualization, as well as evaluate the accuracy and calibration guarantees of our algorithm on real Census derived data using the Folktables package Ding et al. [2021].

1.2 Additional Related Work

Calibration as a statistical objective dates back at least to Dawid [1982]. Foster and Vohra [1999] showed a tight connection between marginal calibration and internal (equivalently swap) regret. We extend this characterization to multicalibration. Multicalibration was introduced by Hébert-Johnson et al. [2018], and variants of the original definition have been studied by a number of works [Kim et al., 2019, Jung et al., 2021, Gopalan et al., 2022, Kim et al., 2022, Roth, 2022]. We use the ℓ_2 variant of multicalibration studied in Roth [2022]—but this definition implies all of the other variants of multicalibration up to a change in parameters. Burhanpurkar et al. [2021] first asked the question "when does multicalibration with respect to $\mathcal{H}$ imply accuracy", and gave a sufficient condition: when $\mathcal{H}$ contains (refinements of) the levelsets of the

Bayes optimal regression function, together with techniques for attempting to find these. This can be viewed as a "strong learning" assumption, in contrast to our weak learning assumption on $\mathcal{H}$.

Boosting for binary classification was introduced by Schapire [1990] and has since become a major topic of both theoretical and empirical study — see Schapire and Freund [2013] for a textbook overview. Both Kim et al. [2019] and Gopalan et al. [2022] have drawn connections between algorithms for multicalibration and boosting for binary classification. In particular, Gopalan et al. [2022] draw direct connections between their split-and-merge multicalibration algorithm and agnostic boosting algorithms of Kalai [2004], Kanade and Kalai [2009], Kalai et al. [2008]. Boosting for squared error regression is much less well studied. Freund and Schapire [1997] give a variant of Adaboost (Adaboost.R) that reduces regression examples to infinite sets of classification examples, and requires a base regressor that optimizes a non-standard loss function. Friedman [2001] introduced the popular gradient boosting method, which for squared error regression corresponds to iteratively fitting the residuals of the current model and then applying an additive update, but did not give a theoretical analysis. Duffy and Helmbold [2002] give a theoretical analysis of several different boosting algorithms for squared error regression under several different weak learning assumptions. Their algorithms require base regression algorithms that can be called (and guaranteed to succeed) on arbitrarily relabelled examples from the training distribution, and given their weak learning assumption, their analysis shows how to drive the error of the final model arbitrarily close to 0. Weak learning assumptions in this style implicitly make very strong realizabilty assumptions (that the Bayes error is close to 0), but because the weak learner is called on relabelled samples, it is difficult to enunciate a weak learning condition that is consistent with obtaining Bayes optimal error, but not better. The boosting algorithm we introduce only requires calling a standard regression algorithm on subsets of the examples from the training distribution, which makes it easy for us to define a weak learning condition that lets us drive error to the Bayes optimal rate without realizability assumptions — thus our results can be viewed as giving an agnostic boosting algorithm for regression.

2 Preliminaries

We study prediction tasks over a domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Here $\mathcal{X}$ represents the *feature* domain and $\mathcal{Y}$ represents the label domain. We focus on the bounded regression setting where $\mathcal{Y} = [0, 1]$ (the scaling to $[0, 1]$ is arbitrary). We write $\mathcal{D} \in \Delta\mathcal{Z}$ to denote a distribution over labelled examples, $\mathcal{D}_{\mathcal{X}}$ to denote the induced marginal distribution over features, and write $D \sim \mathcal{D}^n$ to denote a dataset consisting of n labelled examples sampled i.i.d. from $\mathcal{D}$. We will be interested in the squared error of a model f with respect to distribution $\mathcal{D}$, $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(y - f(x))^2]$. We abuse notation and identify datasets $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with the empirical distribution over the examples they contain, and so we can write the empirical squared error over D: as $\mathbb{E}_{(x,y)\sim D}[(y - f(x))^2] = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$. When taking expectations over a distribution that is clear from context, we will frequently suppress notation indicating the relevant distribution for readability.

We write $R(f)$ to denote the range of a function f, and when $R(f)$ is finite, use m to denote the cardinality of its range: $m = |R(f)|$. We are interested in finding models that are *multicalibrated* with respect to a class of real valued functions $\mathcal{H}$. We use an ℓ_2 notion of multicalibration as used in Roth [2022]:

Definition 2.1 (Multicalibration)**.** *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ and a model $f : \mathcal{X} \to [0, 1]$ that maps onto a countable subset of its range. Let $\mathcal{H}$ be an arbitrary collection of real valued functions $h : \mathcal{X} \to \mathbb{R}$. We say that f is α-approximately multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}$ if for every $h \in \mathcal{H}$:*

$$K_2(f, h, \mathcal{D}) = \sum_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x) = v] \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y - v)|f(x) = v] \right)^2 \leqslant \alpha.$$

We say that f is α-approximately calibrated if:

$$K_2(f, \mathcal{D}) = \sum_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x) = v] \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y - v)|f(x) = v] \right)^2 \leqslant \alpha.$$

If $\alpha = 0$, then we simply say that a model is multicalibrated *or* calibrated. *We will sometimes refer to $K_2(f, \mathcal{D})$ as the mean squared calibration error of a model f.*

Remark 2.2. *When the functions $h(x)$ have binary range, we can view them as indicator functions for some subset of the data domain $S \subseteq \mathcal{X}$, in which case multicalibration corresponds to asking for calibration conditional on membership in these subsets S. Allowing the functions h to have real valued range is only a more general condition. Our notion of approximate multicalibration takes a weighted average over the level sets v of the predictor f, weighted by the probability that $f(x) = v$. This is necessary for any kind of out of sample generalization statement — otherwise we could not even necessarily measure calibration error from a finite sample. Other work on multicalibration use related measures of multicalibration that we think of as ℓ_1 or ℓ_∞ variants, that we can write as $K_1(f, h, \mathcal{D}) = \sum_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x) = v] \left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[h(x)(y - v)|f(x) = v]\right|$ and $K_\infty(f, h, \mathcal{D}) = \max_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x) = v] \left(\mathbb{E}_{(x,y)\sim\mathcal{D}}[h(x)(y - v)|f(x) = v]\right)$. These notions are related to each other: $K_2(f, h, \mathcal{D}) \leqslant K_1(f, h, \mathcal{D}) \leqslant \sqrt{K_2(f, h, \mathcal{D})}$ and $K_\infty(f, h, \mathcal{D}) \leqslant K_1(f, h, \mathcal{D}) \leqslant mK_\infty(f, h, \mathcal{D})$ [Roth, 2022].*

We will characterize the relationship between multicalibration and Bayes optimality.

Definition 2.3 (Bayes Optimal Predictor)**.** *Let $f^* : \mathcal{X} \to [0, 1]$. We say that f^* is the Bayes optimal predictor for $\mathcal{D}$ if:*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y - f^*(x))^2] \leqslant \min_{f:\mathcal{X}\to[0,1]} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y - f(x))^2]$$

The Bayes Optimal predictor satisfies: $f^*(x) = \mathbb{E}_{(x',y)\sim\mathcal{D}}[y|x' = x]$. We say that a function $f : \mathcal{X} \to [0, 1]$ is γ-approximately Bayes optimal if

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y - f(x))^2] \leqslant \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y - f^*(x))^2] + \gamma.$$

Throughout this paper, we will denote the Bayes optimal predictor as f^*.

3 A Characterization of Multicalibration

In this section we give a simple "swap-regret" like characterization of multicalibration for any class of functions $\mathcal{H}$ that is closed under affine transformations:

Definition 3.1. *A class of functions $\mathcal{H}$ is closed under affine transformations if for every $a, b \in \mathbb{R}$, if $h(x) \in \mathcal{H}$ then $h'(x) := ah(x) + b \in \mathcal{H}$.*

As already discussed, closure under affine transformation is a mild assumption: it is already satisfied by many classes of functions $\mathcal{H}$ like linear and polynomial functions and decision trees, and can be enforced for neural network architectures when it is not already satisfied by adding two additional parameters a and b without affecting our ability to optimize over the class.

The first direction of our characterization states that if f fails the multicalibration condition for some $h \in \mathcal{H}$, then there is some other $h' \in \mathcal{H}$ that improves over f in terms of squared error, *when restricted to a level set of f*. The second direction states the opposite: if f is calibrated (but not necessarily multicalibrated), and if there is some level set of f on which h improves over f in terms of squared error, then in fact f must fail the multicalibration condition for h.

Theorem 3.2. *Suppose $\mathcal{H}$ is closed under affine transformation. Fix a model $f : \mathcal{X} \to \mathbb{R}$ and a levelset $v \in R(f)$ of f. Then:*

1. *If there exists an $h \in \mathcal{H}$ such that:*

$$\mathbb{E}[h(x)(y - v)|f(x) = v] \geqslant \alpha,$$

for $\alpha > 0$, then there exists an $h' \in \mathcal{H}$ such that:

$$\mathbb{E}[(f(x) - y)^2 - (h'(x) - y)^2|f(x) = v] \geqslant \frac{\alpha^2}{\mathbb{E}[h(x)^2|f(x) = v]},$$

2. If f is calibrated and there exists an $h \in \mathcal{H}$ such that

$$\mathbb{E}[(f(x)-y)^2 - (h(x)-y)^2 | f(x)=v] \geqslant \alpha,$$

then:

$$\mathbb{E}[h(x)(y-v)|f(x)=v] \geqslant \frac{\alpha}{2}.$$

Proof. We prove each direction in turn.

Lemma 3.3. *Fix a model $f : \mathcal{X} \to \mathbb{R}$. Suppose for some $v \in R(f)$ there is an $h \in \mathcal{H}$ such that:*

$$\mathbb{E}[h(x)(y-v)|f(x)=v] \geqslant \alpha$$

Let $h' = v + \eta h(x)$ for $\eta = \frac{\alpha}{\mathbb{E}[h(x)^2|f(x)=v]}$. Then:

$$\mathbb{E}[(f(x)-y)^2 - (h'(x)-y)^2|f(x)=v] \geqslant \frac{\alpha^2}{\mathbb{E}[h(x)^2|f(x)=v]}$$

Proof. We calculate:

$$\begin{aligned}
& \mathbb{E}[(f(x)-y)^2 - (h'(x)-y)^2|f(x)=v] \\
= \; & \mathbb{E}[(v-y)^2 - (v+\eta h(x)-y)^2|f(x)=v] \\
= \; & \mathbb{E}[v^2 - 2vy + y^2 - (v+\eta h(x))^2 + 2y(v+\eta h(x)) - y^2|f(x)=v] \\
= \; & \mathbb{E}[2y\eta h(x) - 2v\eta h(x) - \eta^2 h(x)^2|f(x)=v] \\
= \; & \mathbb{E}[2\eta h(x)(y-v) - \eta^2 h(x)^2|f(x)=v] \\
\geqslant \; & 2\eta\alpha - \eta^2 \mathbb{E}[h(x)^2|f(x)=v] \\
= \; & \frac{\alpha^2}{\mathbb{E}[h(x)^2|f(x)=v]}
\end{aligned}$$

Where the last line follows from the definition of η. □

The first direction of Theorem 3.2 follows from Lemma 3.3, and the observation that since $\mathcal{H}$ is closed under affine transformations, the function h' defined in the statement of Lemma 3.3 is in $\mathcal{H}$. Now for the second direction.

Lemma 3.4. *Fix a model $f : \mathcal{X} \to \mathbb{R}$. Suppose for some $v \in R(f)$ there is an $h \in \mathcal{H}$ such that:*

$$\mathbb{E}[(\bar{y}_v - y)^2 - (h(x)-y)^2|f(x)=v] \geqslant \alpha,$$

where $\bar{y}_v = \mathbb{E}[y \mid f(x)=v]$. Then it must be that:

$$\mathbb{E}[h(x)(y-\bar{y}_v)|f(x)=v] \geqslant \frac{\alpha}{2}$$

Proof. We calculate:

$$
\begin{aligned}
& \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y-\bar{y}_v)|f(x)=v] \\
= & \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)y|f(x)=v] - \bar{y}_v \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)|f(x)=v] \\
= & \frac{1}{2}\left(2\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)y|f(x)=v] - 2\bar{y}_v \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)|f(x)=v]\right) \\
\geqslant & \frac{1}{2}\left(2\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)y|f(x)=v] - 2\bar{y}_v \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)|f(x)=v] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h(x)-\bar{y}_v)^2|f(x)=v]\right) \\
= & \frac{1}{2}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[2h(x)y - h(x)^2 - \bar{y}_v^2|f(x)=v]\right) \\
= & \frac{1}{2}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[2h(x)y - h(x)^2 - 2\bar{y}_v y + \bar{y}_v^2|f(x)=v]\right) \\
= & \frac{1}{2}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\bar{y}_v - y)^2 - (h(x)-y)^2|f(x)=v]\right) \\
\geqslant & \frac{\alpha}{2}
\end{aligned}
$$

where the 3rd to last line follows from adding and subtracting $\bar{y}_v^2$. □

For any calibrated f it follows that $v = \mathbb{E}[y \mid f(x) = v] = \bar{y}_v$, and so for calibrated f we have that if

$$\mathbb{E}[(v-y)^2 - (h(x)-y)^2|f(x)=v] \geqslant \alpha,$$

then:

$$\mathbb{E}[h(x)(y-v)|f(x)=v] \geqslant \frac{\alpha}{2}.$$

□

4 An Algorithm (For Multicalibration And Regression Boosting)

We now give a single algorithm, and then show how to analyze it both as an algorithm for obtaining a multicalibrated predictor f, and as a boosting algorithm for squared error regression.

Let $m \in \mathbb{N}^+$ be a discretization term, and let $[1/m] := \{0, \frac{1}{m}, \ldots, \frac{m-1}{m}, 1\}$ denote the set of points in $[0,1]$ that are multiples of $1/m$. We will learn a model f whose range is $[1/m]$, which we will enforce by *rounding* its outputs to this range as necessary using the following operation:

Definition 4.1 (Round$(f;m)$)**.** *Let $\mathcal{F}$ be the family of all functions $f : \mathcal{X} \to \mathbb{R}$. Let* Round $: \mathcal{F} \times \mathbb{N}^+ \to \mathcal{F}$ *be a function such that* Round$(f;m)$ *outputs $\tilde{h}(x) = \min_{v\in[1/m]} |h(x) - v|$.*

Unlike other algorithms for multicalibration which make use of *agnostic learning* oracles for binary classification, our algorithm makes use of an algorithm for solving squared-error regression problems over $\mathcal{H}$:

Definition 4.2. *$A_{\mathcal{H}}$ is a squared error regression oracle for a class of real valued functions $\mathcal{H}$ if for every $\mathcal{D} \in \Delta\mathcal{Z}$, $A_{\mathcal{H}}(\mathcal{D})$ outputs a function $h \in \mathcal{H}$ such that*

$$h \in \arg\min_{h'\in\mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h'(x)-y)^2].$$

For example, if $\mathcal{H}$ is the set of all linear functions, then $A_{\mathcal{H}}$ simply solves a linear regression problem (which has a closed form solution). Algorithm 1 (LSBoost[2])repeats the following operation until it no longer decreases overall squared error: it runs squared error regression on each of the level-sets of f_t, and then replaces those levelsets with the solutions to the regression problems, and rounds the output to $[1/m]$.

We will now analyze the algorithm first as a multicalibration algorithm, and then as a boosting algorithm. For simplicity, in this section we will analyze the algorithm as if it is given direct access to the distribution $\mathcal{D}$. In practice, the algorithm will be run on the empirical distribution over a dataset $D \sim \mathcal{D}^n$, and the multicalibration guarantees proven in this section will hold for this empirical distribution. In Section A we prove generalization theorems, which allow us to translate our in-sample error and multicalibration guarantees over D to out-of-sample guarantees over $\mathcal{D}$.

Algorithm 1: LSBoost$(f, \alpha, A_{\mathcal{H}}, \mathcal{D}, B)$

Let $m = \frac{2B}{\alpha}$.
Let $f_0 = \text{Round}(f; m)$, $\text{err}_0 = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f_0(x) - y)^2]$,
$\text{err}_{-1} = \infty$ and $t = 0$.
while $(\text{err}_{t-1} - \text{err}_t) \geqslant \frac{\alpha}{2B}$ **do**
  for each $v \in [1/m]$ **do**
    Let $\mathcal{D}_v^{t+1} = \mathcal{D}|(f_t(x) = v)$.
    Let $h_v^{t+1} = A_{\mathcal{H}}(\mathcal{D}_v^{t+1})$.
  Let:

$$\tilde{f}_{t+1}(x) = \sum_{v\in[1/m]} \mathbb{1}[f_t(x) = v] \cdot h_v^{t+1}(x) \quad f_{t+1} = \text{Round}(\tilde{f}_{t+1}, m)$$

  Let $\text{err}_{t+1} = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f_{t+1}(x) - y)^2]$ and $t = t + 1$.
Output f_{t-1}.

4.1 Analysis as a Multicalibration Algorithm

Theorem 4.3. *Fix any distribution $\mathcal{D} \in \Delta\mathcal{Z}$, any model $f : \mathcal{X} \to [0, 1]$, any $\alpha < 1$, any class of real valued functions $\mathcal{H}$ that is closed under affine transformations, and a squared error regression oracle $A_{\mathcal{H}}$ for $\mathcal{H}$. For any bound $B > 0$ let:*

$$\mathcal{H}_B = \{h \in \mathcal{H} : \max_{x\in\mathcal{X}} h(x)^2 \leqslant B\}$$

be the set of functions in h with squared magnitude bounded by B. Then LSBoost$(f, \alpha, A_{\mathcal{H}}, \mathcal{D}, B)$ (Algorithm 1) halts after at most $T \leqslant \frac{2B}{\alpha}$ many iterations and outputs a model f_{T-1} such that f_{T-1} is α-approximately multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}_B$.

Remark 4.4. *Note the form of this theorem — we do not promise multicalibration at approximation parameter α for all of $\mathcal{H}$, but only for $\mathcal{H}_B$ — i.e. those functions in $\mathcal{H}$ satisfying a bound on their squared value. This is necessary, since $\mathcal{H}$ is closed under affine transformations. To see this, note that if $\mathbb{E}[h(x)(y-v)] \geqslant \alpha$, then it must be that $\mathbb{E}[c \cdot h(x)(y - v)] \geqslant c \cdot \alpha$. Since $h'(x) = ch(x)$ is also in $\mathcal{H}$ by assumption, approximate multicalibration bounds must always also be paired with a bound on the norm of the functions for which we promise those bounds.*

Proof. Since f_0 takes values in $[0, 1]$ and $y \in [0, 1]$, we have $\text{err}_0 \leqslant 1$, and by definition $\text{err}_T \geqslant 0$ for all T. By construction, if the algorithm has not halted at round t it must be that $\text{err}_t \leqslant \text{err}_{t-1} - \frac{\alpha}{2B}$, and so the algorithm must halt after at most $T \leqslant \frac{2B}{\alpha}$ many iterations to avoid a contradiction.

It remains to show that when the algorithm halts at round T, the model f_{T-1} that it outputs is α-approximately multi-calibrated with respect to $\mathcal{D}$ and $\mathcal{H}_B$. We will show that if this is not the case, then $\text{err}_{T-1} - \text{err}_T > \frac{\alpha}{2B}$, which will be a contradiction to the halting criterion of the algorithm.

[2]LSBoost can be taken to stand for either "Level Set Boost" or "Least Squares Boost", at the reader's discretion.

Suppose that f_{T-1} is not α-approximately multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}_B$. This means there must be some $h \in \mathcal{H}_B$ such that:

$$\sum_{v \in [1/m]} \Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x) = v] \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y-v)|f_{T-1}(x) = v] \right)^2 > \alpha$$

For each $v \in [1/m]$ define

$$\alpha_v = \Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x) = v] \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y-v)|f_{T-1}(x) = v] \right)^2$$

So we have $\sum_{v\in[1/m]} \alpha_v > \alpha$.

Applying the 1st part of Theorem 3.2 we learn that for each v, there must be some $h_v \in \mathcal{H}$ such that:

$$\begin{aligned} \mathbb{E}[(f_{T-1}(x) - y)^2 - (h_v(x) - y)^2 | f_{T-1}(x) = v] &> \frac{1}{\mathbb{E}[h(x)^2 | f_{T-1}(x) = v]} \cdot \frac{\alpha_v}{\Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x) = v]} \\ &\geqslant \frac{1}{B} \frac{\alpha_v}{\Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x) = v]} \end{aligned}$$

where the last inequality follows from the fact that $h \in \mathcal{H}_B$ Now we can compute:

$$\begin{aligned} & \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (\tilde{f}_T(x) - y)^2] \\ = & \sum_{v\in[1/m]} \Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x) = v] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (\tilde{f}_T(x) - y)^2 | f_{T-1}(x) = v] \\ = & \sum_{v\in[1/m]} \Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x) = v] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (h_v^T(x) - y)^2 | f_{T-1}(x) = v] \\ \geqslant & \sum_{v\in[1/m]} \Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x) = v] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (h_v(x) - y)^2 | f_{T-1}(x) = v] \\ \geqslant & \sum_{v\in[1/m]} \frac{\alpha_v}{B} \\ > & \frac{\alpha}{B} \end{aligned}$$

Here the third line follows from the definition of $\tilde{f}_T$ and the fourth line follows from the fact $h_v \in \mathcal{H}$ and that h_v^T minimizes squared error on $\mathcal{D}_v^T$ amongst all $h \in \mathcal{H}$.

Finally we calculate:

$$\begin{aligned} & \mathrm{err}_{T-1} - \mathrm{err}_T \\ = & \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (f_T(x) - y)^2] \\ = & \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (\tilde{f}_T(x) - y)^2] + \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\tilde{f}_T(x) - y)^2 - (f_T(x) - y)^2] \\ > & \frac{\alpha}{B} + \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\tilde{f}_T(x) - y)^2 - (f_T(x) - y)^2] \\ > & \frac{\alpha}{B} - \frac{1}{m} \\ \geqslant & \frac{\alpha}{2B} \end{aligned}$$

where the last equality follows from the fact that $m \geqslant \frac{2B}{\alpha}$.

The 2nd inequality follows from the fact that for every pair (x, y):

$$(\tilde{f}_T(x) - y)^2 - (f_T(x) - y)^2 \geqslant -\frac{1}{m}$$

To see this we consider two cases. Since $y \in [0, 1]$, if $\tilde{f}_T(x) > 1$ or $\tilde{f}_T(x) < 0$ then the Round operation decreases squared error and we have $(\tilde{f}_T(x) - y)^2 - (f_T(x) - y)^2 \geqslant 0$. In the remaining case we have $f_T(x) \in [0, 1]$ and $\Delta = \tilde{f}_T(x) - f_T(x)$ is such that $|\Delta| \leqslant \frac{1}{2m}$. In this case we can compute:

$$\begin{aligned}(\tilde{f}_T(x) - y)^2 - (f_T(x) - y)^2 &= (f_T(x) + \Delta - y)^2 - (f_T(x) - y)^2 \\ &= 2\Delta(f(x) - y) + \Delta^2 \\ &\geqslant -2|\Delta| + \Delta^2 \\ &\geqslant -\frac{1}{m}\end{aligned}$$

□

4.2 Analysis as a Boosting Algorithm

We now analyze the same algorithm (Algorithm 1) as a boosting algorithm designed to boost a "weak learning" algorithm $A_{\mathcal{H}}$ to a strong learning algorithm. Often in the boosting literature, a "strong learning" algorithm is one that can obtain accuracy arbitrarily close to perfect, which is only possible under strong realizability assumptions. In this paper, by "strong learning", we mean that Algorithm 1 should output a model that is close to Bayes optimal, which is a goal we can enunciate for any distribution $\mathcal{D}$ without needing to make realizability assumptions. (Observe that if the Bayes optimal predictor has zero error, then our meaning of strong learning corresponds to the standard meaning, so our analysis is only more general).

We now turn to our definition of weak learning. Intuitively, a weak learning algorithm should return a hypothesis that makes predictions that are slightly better than trivial whenever doing so is possible. We take "trivial" predictions to be those of the best *constant* predictor as measured by squared error — i.e. the squared error obtained by simply returning the label mean. A "weak learning" algorithm for us can be run on any restriction of the data distribution $\mathcal{D}$ to a subset $S \subseteq \mathcal{X}$, and must return a hypothesis with squared error slightly better than the squared error of the best constant prediction, whenever the Bayes optimal predictor f^* has squared error slightly better than a constant predictor; on restrictions for which the Bayes optimal predictor also does not improve over constant prediction, our weak learning algorithm is not required to do better either.

Traditionally, "weak learning" assumptions do not distinguish between the optimization ability of the algorithm and the representation ability of the hypothesis class it optimizes over. Since we have defined a squared error regression oracle $A_{\mathcal{H}}$ as exactly optimizing the squared error over some class $\mathcal{H}$, we will state our weak learning assumption as an assumption on the representation ability of $\mathcal{H}$—but this is not important for our analysis here. To prove Theorem 4.6 we could equally well assume that $A_{\mathcal{H}}$ returns a hypothesis h that improves over a constant predictor whenever one exists, without assuming that h optimizes squared error over all of $\mathcal{H}$.

Definition 4.5 (Weak Learning Assumption)**.** *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ and a class of functions $\mathcal{H}$. Let $f^*(x) = \mathbb{E}_{y \sim \mathcal{D}(x)}[y]$ denote the true conditional label expectation conditional on x. We say that $\mathcal{H}$ satisfies the* γ-weak learning *condition relative to $\mathcal{D}$ if for every $S \subseteq \mathcal{X}$ with $\Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[x \in S] > 0$, if:*

$$\mathbb{E}[(f^*(x) - y)^2 | x \in S] < \min_{c \in \mathbb{R}} \mathbb{E}[(c - y)^2 | x \in S] - \gamma$$

then there exists an $h \in \mathcal{H}$ such that:

$$\mathbb{E}[(h(x) - y)^2 | x \in S] < \min_{c \in \mathbb{R}} \mathbb{E}[(c - y)^2 | x \in S] - \gamma$$

When $\gamma = 0$ we simply say that $\mathcal{H}$ satisfies the weak learning condition relative to $\mathcal{D}$.

Observe why our weak learning assumption is "weak": the Bayes optimal predictor f^* may improve arbitrarily over the best constant predictor on some set S in terms of squared error, but in this case we only require of $\mathcal{H}$ that it include a hypothesis that improves by some γ which might be very small.

Since the γ-weak learning condition does not make any requirements on $\mathcal{H}$ on sets for which $f^*(x)$ improves over a constant predictor by less than γ, the best we can hope to prove under this assumption is γ-approximate Bayes optimality, which is what we do next.

Theorem 4.6. *Fix any distribution $\mathcal{D} \in \Delta\mathcal{Z}$, any model $f : \mathcal{X} \to [0,1]$, any $\gamma > 0$, any class of real valued functions $\mathcal{H}$ that satisfies the γ-weak learning condition relative to $\mathcal{D}$, and a squared error regression oracle $A_{\mathcal{H}}$ for $\mathcal{H}$. Let $\alpha = \gamma$ and $B = 1/\gamma$ (or any pair such that $\alpha/B = \gamma^2$). Then LSBoost$(f, \alpha, A_{\mathcal{H}}, \mathcal{D}, B)$ halts after at most $T \leqslant \frac{2}{\gamma^2}$ many iterations and outputs a model f_{T-1} such that f_{T-1} is 2γ-approximately Bayes optimal over $\mathcal{D}$:*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2] \leqslant \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x) - y)^2] + 2\gamma$$

where $f^*(x) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y]$ is the function that minimizes squared error over $\mathcal{D}$.

Proof. At each round t before the algorithm halts, we have by construction that $\mathrm{err}_t \leqslant \mathrm{err}_{t-1} - \frac{\alpha}{2B}$, and since the squared error of f_0 is at most 1, and squared error is non-negative, we must have $T \leqslant \frac{2B}{\alpha} = \frac{2}{\gamma^2}$.

Now suppose the algorithm halts at round T and outputs f_{T-1}. It must be that $\mathrm{err}_T > \mathrm{err}_{T-1} - \frac{\gamma^2}{2}$. Suppose also that f_{T-1} is not 2γ-approximately Bayes optimal:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (f^*(x) - y)^2] > 2\gamma$$

We can write this condition as:

$$\sum_{v\in[1/m]} \Pr[f_{T-1}(x) = v] \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (f^*(x) - y)^2 | f_{T-1}(x) = v] > 2\gamma$$

Define the set:

$$S = \{v \in [1/m] : \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (f^*(x) - y)^2 | f_{T-1}(x) = v] \geqslant \gamma\}$$

to denote the set of values v in the range of f_{T-1} such that conditional on $f_{T-1}(x) = v$, f_{T-1} is at least γ-sub-optimal. Since we have both $y \in [0,1]$ and $f_{T-1}(x) \in [0,1]$, for every v we must have that $\mathbb{E}[(f_{T-1}(x) - y)^2 - (f^*(x) - y)^2 | f_{T-1}(x) = v] \leqslant 1$. Therefore we can bound:

$$\begin{aligned} 2\gamma &< \sum_{v\in[1/m]} \Pr[f_{T-1}(x) = v] \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (f^*(x) - y)^2 | f_{T-1}(x) = v] \\ &\leqslant \Pr_{(x,y)\sim\mathcal{D}}[x \in S] + (1 - \Pr_{(x,y)\sim\mathcal{D}}[x \in S])\gamma \end{aligned}$$

Solving we learn that:

$$\Pr_{(x,y)\sim\mathcal{D}}[x \in S] \geqslant \frac{2\gamma - \gamma}{(1-\gamma)} \geqslant 2\gamma - \gamma = \gamma$$

Now observe that by the fact that $\mathcal{H}$ is assumed to satisfy the γ-weak learning assumption with respect to $\mathcal{D}$, at the final round T of the algorithm, for every $v \in S$ we have that h_v^T satisfies:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (h_v^T(x) - y)^2 | f_{T-1}(x) = v] \geqslant \gamma$$

Let $\tilde{\mathrm{err}}_T = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\tilde{f}_T(x) - y)^2]$ Therefore we have:

$$\begin{aligned} \mathrm{err}_{T-1} - \tilde{\mathrm{err}}_T &= \sum_{v\in[1/m]} \Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x) = v] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (h_v^T(x) - y)^2 | f_{T-1}(x) = v] \\ &\geqslant \Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x) \in S]\gamma \\ &\geqslant \gamma^2 \end{aligned}$$

We recall that $|\tilde{\text{err}}_T - \text{err}_T| \leqslant 1/m = \frac{\gamma^2}{2}$ and so we can conclude that

$$\text{err}_{T-1} - \text{err}_T \geqslant \frac{\gamma^2}{2}$$

which contradicts the fact that the algorithm halted at round T, completing the proof. □

5 When Multicalibration Implies Accuracy

We analyzed the same algorithm (Algorithm 1) as both an algorithm for obtaining multicalibration with respect to $\mathcal{H}$, and, when $\mathcal{H}$ satisfied the weak learning condition given in Definition 4.5, as a boosting algorithm that converges to the Bayes optimal model. In this section we show that this is no coincidence: multicalibration with respect to $\mathcal{H}$ implies Bayes optimality if and only if $\mathcal{H}$ satisfies the weak learning condition from Definition 4.5,

First we define what we mean when we say that multicalibration with respect to $\mathcal{H}$ implies Bayes optimality. Note that the Bayes optimal model $f^*(x)$ is multicalibrated with respect to any set of functions, so it is not enough to require that there *exist* Bayes optimal functions f that are multicalibrated with respect to $\mathcal{H}$. Instead, we have to require that *every* function that is multicalibrated with respect to $\mathcal{H}$ is Bayes optimal:

Definition 5.1. *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$. We say that multicalibration with respect to $\mathcal{H}$ implies Bayes optimality over $\mathcal{D}$ if for every $f : \mathcal{X} \to \mathbb{R}$ that is multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}$, we have:*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2] = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x) - y)^2]$$

Where $f^*(x) = \mathbb{E}_{y\sim\mathcal{D}(x)}[y]$ is the function that has minimum squared error over the set of all functions.

Recall that when the weak learning parameter γ in Definition 4.5 is set to 0, we simply call it the "weak learning condition" relative to $\mathcal{D}$. We first state and prove our characterization for the exact case when $\gamma = 0$, because it leads to an exceptionally simple statement. We subsequently extend this characterization to relate approximate Bayes optimality and approximate multicalibration under quantitative weakenings of the weak learning condition.

Theorem 5.2. *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$. Let $\mathcal{H}$ be a class of functions that is closed under affine transformation. Multicalibration with respect to $\mathcal{H}$ implies Bayes optimality over $\mathcal{D}$ if and only if $\mathcal{H}$ satisfies the weak learning condition relative to $\mathcal{D}$.*

Proof. To avoid measurability issues we assume that models f have a countable range (which is true in particular whenever $\mathcal{X}$ is countable).

First we show that if $\mathcal{H}$ satisfies the weak learning condition relative to $\mathcal{D}$, then multicalibration with respect to $\mathcal{H}$ implies Bayes optimality over $\mathcal{D}$. Suppose not. Then there exists a function f that is multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}$, but is such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2] > \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x) - y)^2]$$

By linearity of expectation we have:

$$\sum_{v\in R(f)} \Pr[f(x) = v] \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2 - (f^*(x) - y)^2 | f(x) = v] > 0$$

In particular there must be some $v \in R(f)$ with $\Pr_{x\sim\mathcal{D}_\mathcal{X}}[f(x) = v] > 0$ such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2 | f(x) = v] > \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x) - y)^2 | f(x) = v]$$

Let $S = \{x : f(x) = v\}$. Observe that if $\mathcal{H}$ is closed under affine transformation, the constant function $h(x) = 1$ is in $\mathcal{H}$, and hence multicalibration with respect to $\mathcal{H}$ implies calibration. Since f is calibrated, we know that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(v-y)^2|x \in S] = \min_{c\in\mathbb{R}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(c-y)^2|x \in S]$$

Thus by the weak learning assumption there must exist some $h \in \mathcal{H}$ such that:

$$\mathbb{E}[(v-y)^2 - (h(x)-y)^2|x \in S] = \mathbb{E}[(f(x)-y)^2 - (h(x)-y)^2|f(x) = v] > 0$$

By Theorem 3.2, there must therefore exist some $h' \in \mathcal{H}$ such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h'(x)(y-v)|f(x) = v] > 0$$

implying that f is *not* multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}$, a contradiction.

In the reverse direction, we show that for any $\mathcal{H}$ that does *not* satisfy the weak learning condition with respect to $\mathcal{D}$, then multicalibration with respect to $\mathcal{H}$ and $\mathcal{D}$ does not imply Bayes optimality over $\mathcal{D}$. In particular, we exhibit a function f such that f is multicalibrated with respect to $\mathcal{H}$ and $\mathcal{D}$, but such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2] > \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2]$$

Since $\mathcal{H}$ does not satisfy the weak learning assumption over $\mathcal{D}$, there must exist some set $S \subseteq \mathcal{X}$ with $\Pr[x \in S] > 0$ such that

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2|x \in S] < \min_{c\in\mathbb{R}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(c-y)^2|x \in S]$$

but for every $h \in \mathcal{H}$:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h(x)-y)^2|x \in S] \geqslant \min_{c\in\mathbb{R}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(c-y)^2|x \in S]$$

.

Let $c(S) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y|x \in S]$. We define $f(x)$ as follows:

$$f(x) = \begin{cases} f^*(x) & x \notin S \\ c(S) & x \in S \end{cases}$$

We can calculate that:

$$\begin{aligned}
& \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2] \\
= & \Pr_{(x,y)\sim\mathcal{D}}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(c(S)-y)^2|x \in S] + \Pr_{(x,y)\sim\mathcal{D}}[x \notin S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2|x \notin S] \\
> & \Pr_{(x,y)\sim\mathcal{D}}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2|x \in S] + \Pr_{(x,y)\sim\mathcal{D}}[x \notin S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2|x \notin S] \\
= & \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2]
\end{aligned}$$

In other words, f is not Bayes optimal. So if we can demonstrate that f is multicalibrated with respect to $\mathcal{H}$ and $\mathcal{D}$ we are done. Suppose otherwise. Then there exists some $h \in \mathcal{H}$ and some $v \in R(f)$ such that

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y-v)|f(x) = v] > 0$$

By Theorem 3.2, there exists some $h' \in \mathcal{H}$ such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h'(x)-y)^2|f(x) = v] < \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2|f(x) = v]$$

We first observe that it must be that $v = c(S)$. If this were not the case, by definition of f we would have that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h'(x) - y)^2 | f(x) = v] < \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x) - y)^2 | f(x) = v]$$

which would contradict the Bayes optimality of f^*. Having established that $v = c(S)$ we can calculate:

$$\begin{aligned}
& \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h'(x) - y)^2 | f(x) = c(S)] \\
= & \Pr_{(x,y)\sim\mathcal{D}}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h'(x) - y)^2 | x \in S] + \\
& \Pr_{(x,y)\sim\mathcal{D}}[x \notin S, f(x) = c(S)] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h'(x) - y)^2 | x \notin S, f(x) = c(S)] \\
\geqslant & \Pr_{(x,y)\sim\mathcal{D}}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h'(x) - y)^2 | x \in S] + \\
& \Pr_{(x,y)\sim\mathcal{D}}[x \notin S, f(x) = c(S)] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2 | x \notin S, f(x) = c(S)]
\end{aligned}$$

where in the last inequality we have used the fact that by definition, $f(x) = f^*(x)$ for all $x \notin S$, and so is pointwise Bayes optimal for all $x \notin S$.

Hence the only way we can have $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(h'(x) - y)^2 | f(x) = c(S)] < \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2 | f(x) = c(S)]$ is if:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h'(x) - y)^2 | x \in S] < \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(c(S) - y)^2 | x \in S]$$

But this contradicts our assumption that $\mathcal{H}$ violates the weak learning condition on S, which completes the proof.

□

We now turn our attention to deriving a relationship between approximate multicalibration and approximate Bayes optimality. To do so, we'll introduce an even weaker weak learning condition that has one additional parameter ρ, lower bounding the mass of sets S that we can condition on while still requiring the weak learning condition to hold. We remark that Algorithm 1 can be analyzed as a boosting algorithm under this weaker weak learning assumption as well, with only minor modifications in the analysis.

Definition 5.3 ((γ, ρ)-weak learning condition)**.** *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ and let $\mathcal{H}$ be a class of arbitrary real-valued functions. We say that $\mathcal{H}$ satisfies the* (γ, ρ)-weak learning condition *for $\mathcal{D}$ if the following holds. For every set $S \subseteq \mathcal{X}$ such that $\Pr_{x\sim\mathcal{D}_\mathcal{X}}[x \in S] > \rho$, if*

$$\mathop{\mathbb{E}}_{(x,y)\sim D}[(f^* - y)^2 \mid x \in S] < \mathop{\mathbb{E}}_{(x,y)\sim D}[(\bar{y}_S - y)^2 \mid x \in S] - \gamma,$$

where $\bar{y}_S = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y \mid x \in S]$, then there exists $h \in \mathcal{H}$ such that

$$\mathop{\mathbb{E}}_{(x,y)\sim D}[(h(x) - y)^2 \mid x \in S] < \mathop{\mathbb{E}}_{(x,y)\sim D}[(\bar{y}_S - y)^2 \mid x \in S] - \gamma.$$

We may now prove our theorem showing that approximate multicalibration with respect to a class $\mathcal{H}$ implies approximate Bayes optimality if and only if $\mathcal{H}$ satisfies the (γ, ρ)-weak learning condition. We recall Remark 4.4, which notes that we must restrict approximate multicalibration to a bounded subset of $\mathcal{H}$, as we will assume that $\mathcal{H}$ is closed under affine transformation.

Theorem 5.4. *Fix any distribution $\mathcal{D} \in \Delta\mathcal{Z}$, any model $f : \mathcal{X} \to [0, 1]$, and any class of real valued functions $\mathcal{H}$ that is closed under affine transformation. Let:*

$$\mathcal{H}_1 = \{h \in \mathcal{H} : \max_{x\in\mathcal{X}} h(x)^2 \leqslant 1\}$$

be the set of functions in $\mathcal{H}$ upper-bounded by 1 on $\mathcal{X}$. Let $m = |R(f)|$, $\gamma > 0$, and $\alpha \leqslant \frac{\gamma^3}{16m}$. Then if $\mathcal{H}$ satisfies the $(\gamma, \gamma/m)$-weak learning condition and f is α-approximately multicalibrated with respect to $\mathcal{H}_1$ on $\mathcal{D}$, then f has squared error

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2] \leqslant \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*-y)^2] + 3\gamma.$$

Conversely, if $\mathcal{H}$ does not satisfy the $(\gamma, \gamma/m)$-weak learning condition, there exists a model $f : \mathcal{X} \to [0,1]$ that is α-approximately multicalibrated with respect to $\mathcal{H}_1$ on $\mathcal{D}$, for $\alpha = \gamma$, and is perfectly calibrated on $\mathcal{D}$, but f has squared error

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2] \geqslant \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*-y)^2] + \gamma^2/m.$$

Proof. We begin by arguing that α-approximate multicalibration with respect to $\mathcal{H}_1$ on $\mathcal{D}$ implies approximate Bayes optimality when $\mathcal{H}$ satisfies the $(\gamma, \gamma/m)$-weak learning condition. Suppose not, and there exists a function f that is α-multicalibrated with respect to $\mathcal{H}_1$, but

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*-y)^2] < \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2] - 3\gamma.$$

Then there must exist some $v \in R(f)$ such that $\Pr_{(x,y)\sim\mathcal{D}}[f(x) = v] > \gamma/m$ and

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*-y)^2 \mid f(x)=v] < \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2 \mid f(x)=v] - 2\gamma.$$

We observe that since $\mathcal{H}$ is closed under affine transformation, the constant function $h(x) = 1$ is in $\mathcal{H}$, and so α-approximate multicalibration with respect to $\mathcal{H}_1$ implies α-approximate calibration as well. Thus by definition,

$$\Pr[f(x)=v] \cdot \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[v-y \mid f(x)=v]\right)^2 \leqslant \alpha.$$

Letting $\bar{y}_v = \mathbb{E}[y \mid f(x) = v]$, our lower-bound that $\Pr[f(x) = v] > \gamma/m$ gives us that $(v-\bar{y}_v)^2 < \alpha m/\gamma \leqslant \left(\frac{\gamma}{4}\right)^2$. We now use this upper-bound on calibration error in conjuction with our lower-bound on distance from Bayes optimality to show that the squared error of the constant predictor $\bar{y}_v$ must also be far from Bayes optimal.

$$\begin{aligned}
\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2 \mid f(x)=v] &< \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2 \mid f(x)=v] - 2\gamma \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(v-\bar{y}_v+\bar{y}_v-y)^2 \mid f(x)=v] - 2\gamma \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\bar{y}_v-y)^2 \mid f(x)=v] + (v-\bar{y}_v)^2 - 2\gamma \\
&< \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\bar{y}_v-y)^2 \mid f(x)=v] - \gamma.
\end{aligned}$$

The $(\gamma, \gamma/m)$-weak learning condition then guarantees that there exists some $h \in \mathcal{H}$ such that

$$\mathop{\mathbb{E}}_{(x,y)\sim D}[(h-y)^2 \mid f(x)=v] < \mathop{\mathbb{E}}_{(x,y)\sim D}[(\bar{y}_v-y)^2 \mid f(x)=v] - \gamma.$$

By Lemma 3.4, the fact that h improves on the squared loss of $\bar{y}_v$ by an additive factor γ, on the set of x such that $f(x) = v$, implies that $\mathbb{E}[h(x)(y-\bar{y}_v) \mid f(x)=v] > \gamma/2$. Because f is α-approximately calibrated on $\mathcal{D}$, we can use the existence of such an h to witness a failure of multicalibration:

$$\begin{aligned}
&\mathbb{E}[h(y-v) \mid f(x)=v] \\
&\quad= \mathbb{E}[h(x)(y-\bar{y}_v+\bar{y}_v-v) \mid f(x)=v] \\
&\quad= \mathbb{E}[h(x)(y-\bar{y}_v) \mid f(x)=v] + \mathbb{E}[h(x)(\bar{y}_v-v) \mid f(x)=v] \\
&\quad> \gamma/2 - |\bar{y}_v - v| \\
&\quad> \gamma/4.
\end{aligned}$$

Then

$$\Pr[f(x) = v] \cdot \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y - v) \mid f(x) = v] \right)^2 > \frac{\gamma^3}{16m},$$

contradicting our assumption that f is α-approximately multicalibrated with respect to $\mathcal{H}_1$ for $\alpha < \frac{\gamma^3}{16m}$. Therefore approximate multicalibration with respect to $\mathcal{H}_1$ must imply that f is approximately Bayes optimal.

It remains to show the other direction, that α-approximate multicalibration with respect to a class $\mathcal{H}_1$ implies approximate Bayes optimality only if $\mathcal{H}$ satisfies the $(\gamma, \gamma/m)$-weak learning condition. If this claim were not true for the stated parameters, then there must exist a class $\mathcal{H}$ such that every predictor f that:

- is α-approximately multicalibrated with respect to $\mathcal{H}_1$
- is perfectly calibrated on $\mathcal{D}$
- has range with cardinality $|R(f)| = m$

also has squared error within γ^2/m of Bayes optimal, but $\mathcal{H}$ does not satisfy the weak learning condition. We will show that no such class exists by defining, for any class $\mathcal{H}$ not satisfying the weak learning condition, a predictor f that is α-approximately multicalibrated with respect to that class, but has squared error that is not within γ^2/m of Bayes optimal.

Recall that if a class $\mathcal{H}$ does not satisfy the $(\gamma, \gamma/m)$-weak learning condition, then there must be some set $S_\mathcal{H}$ such that $\Pr[x \in S_\mathcal{H}] > \gamma/m$, there does not exist an $h \in \mathcal{H}$ such that

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h - y)^2 \mid x \in S_\mathcal{H}] < \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\bar{y}_{S_\mathcal{H}} - y)^2 \mid x \in S_\mathcal{H}] - \gamma,$$

but for the Bayes optimal predictor, it holds that its squared loss satisfies

$$\mathop{\mathbb{E}}_{(x,y)\sim D}[(f^* - y)^2 \mid x \in S_\mathcal{H}] < \mathop{\mathbb{E}}_{(x,y)\sim D}[(\bar{y}_{S_\mathcal{H}} - y)^2 \mid x \in S_\mathcal{H}] - \gamma,$$

where $\bar{y}_{S_\mathcal{H}} = \mathbb{E}[y \mid x \in S_\mathcal{H}]$. For some hypothesis class $\mathcal{H}$ not satisfying the weak learning condition, and associated set $S_\mathcal{H}$, let $f_\mathcal{H}$ be defined as follows:

$$f_\mathcal{H}(x) = \begin{cases} f^*(x), & x \notin S_\mathcal{H} \\ \bar{y}_{S_\mathcal{H}}, & x \in S_\mathcal{H}. \end{cases}.$$

Note that, because $f_\mathcal{H}$ is constant on $S_\mathcal{H}$, there must be some $v \in R(f)$ such that the level set $S_v = \{x \in \mathcal{X} : f(x) = v\}$ contains $S_\mathcal{H}$. To see that $f_\mathcal{H}$ is α-approximately multicalibrated with respect to $\mathcal{H}_1$, we first consider the contribution to multicalibration error from the level sets not containing $S_\mathcal{H}$. For all $h \in \mathcal{H}$ and $v \in R(f)$ such that $v \neq \bar{y}_{S_\mathcal{H}}$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y - f_\mathcal{H}(x)) \mid f_\mathcal{H}(x) = v] &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y - f^*(x)) \mid f_\mathcal{H}(x) = v] \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x} \mathop{\mathbb{E}}_{y\sim\mathcal{D}_y(x)}[h(x)y \mid f_\mathcal{H}(x) = v] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}[h(x)f^*(x) \mid f_\mathcal{H}(x) = v] \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x} \mathop{\mathbb{E}}_{y\sim\mathcal{D}_y(x)}[h(x)y \mid f_\mathcal{H}(x) = v] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x} \mathop{\mathbb{E}}_{y\sim\mathcal{D}_y(x)}[h(x)y \mid f_\mathcal{H}(x) = v] \\
&= 0.
\end{aligned}$$

For the level set S_v for which $S_\mathcal{H} \subseteq S_v$, we know from the argument above that the elements $x \in S_v \backslash S_\mathcal{H}$ contribute nothing to the multicalibration error, as $f(x) = f^*(x)$ on these elements. So,

$$\begin{aligned}
\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y - f_\mathcal{H}(x)) \mid f(x) = v] &= \Pr_{x\sim\mathcal{D}_\mathcal{X}}[x \in S_\mathcal{H}] \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y - \bar{y}_{S_\mathcal{H}}) \mid x \in S_\mathcal{H}] \\
&\quad + \Pr_{x\sim\mathcal{D}_\mathcal{X}}[x \notin S_\mathcal{H}] \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y - f^*(x)) \mid x \in S_v \backslash S_\mathcal{H}] \\
&= \Pr_{x\sim\mathcal{D}_\mathcal{X}}[x \in S_\mathcal{H}] \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y - \bar{y}_{S_\mathcal{H}}) \mid x \in S_\mathcal{H}]
\end{aligned}$$

Therefore if $f_{\mathcal{H}}$ is not α-approximately multicalibrated with respect to $\mathcal{H}_1$ on $\mathcal{D}$, it must be the case that there exists some $h \in \mathcal{H}_1$ such that $\mathbb{E}[h(x)(y - \bar{y}_{S_{\mathcal{H}}}) \mid x \in S_{\mathcal{H}}] > \sqrt{\alpha}$. Then by Theorem 3.2, there must exist a $h' \in \mathcal{H}$ such that

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\bar{y}_{S_{\mathcal{H}}} - y)^2 - (h'(x) - y)^2 \mid x \in S_{\mathcal{H}}] > \alpha = \gamma.$$

But $S_{\mathcal{H}}$ was defined to be a subset of $\mathcal{X}$ for which no such h' exists and for which $\Pr[x \in S_{\mathcal{H}}] > \gamma/m$. This would contradict our assumption that $\mathcal{H}$ does not satisfy the $(\gamma, \gamma/m)$-weak learning condition on $\mathcal{D}$, and therefore $f_{\mathcal{H}}$ is α-approximately multicalibrated with respect to $\mathcal{H}_1$ on $\mathcal{D}$.

It remains to prove that $f_{\mathcal{H}}$ is far from Bayes optimal.

$$\begin{aligned}
\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{\mathcal{H}}(x) - y)^2] &= \Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in S_{\mathcal{H}}] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\bar{y}_{S_{\mathcal{H}}} - y)^2 \mid x \in S_{\mathcal{H}}] + \Pr[x \notin S_{\mathcal{H}}] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x) - y)^2 \mid x \notin S_{\mathcal{H}}] \\
&\geqslant \Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in S_{\mathcal{H}}] \left(\mathop{\mathbb{E}}_{(x,y)\mathcal{D}}[(f^* - y)^2 \mid x \in S_{\mathcal{H}}] + \gamma \right) + \Pr[x \notin S_{\mathcal{H}}] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x) - y)^2 \mid x \notin S_{\mathcal{H}}] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^* - y)^2] + \gamma \Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in S_{\mathcal{H}}] \\
&\geqslant \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^* - y)^2] + \gamma^2/m.
\end{aligned}$$

□

6 Weak Learners With Respect to Constrained Classes

Thus far we have studied function classes $\mathcal{H}$ that satisfy a weak learning condition with respect to the Bayes optimal predictor f^*. But we can also study function classes $\mathcal{H}$ that satisfy a weak learning condition defined with respect to another constrained class of real valued functions.

Definition 6.1 (Weak Learning Assumption Relative to $\mathcal{C}$)**.** *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ and two classes of functions $\mathcal{H}$ and $\mathcal{C}$. We say that $\mathcal{H}$ satisfies the* γ-weak learner *condition relative to $\mathcal{C}$ and $\mathcal{D}$ if for every $S \subseteq \mathcal{X}$ with $\Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in S] > 0$, if:*

$$\min_{c\in\mathcal{C}} \mathop{\mathbb{E}}_{(x,y)\sim D}[(c(x) - y)^2 \mid x \in S] < \mathop{\mathbb{E}}_{(x,y)\sim D}[(\bar{y}_S - y)^2 \mid x \in S] - \gamma,$$

where $\bar{y}_S = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y \mid x \in S]$, then there exists $h \in \mathcal{H}$ such that

$$\mathop{\mathbb{E}}_{(x,y)\sim D}[(h(x) - y)^2 \mid x \in S] < \mathop{\mathbb{E}}_{(x,y)\sim D}[(\bar{y}_S - y)^2 \mid x \in S] - \gamma.$$

When $\gamma = 0$ we simply say that $\mathcal{H}$ satisfies the weak learning condition relative to $\mathcal{C}$ and $\mathcal{D}$.

We will show that if a predictor f is multicalibrated with respect to $\mathcal{H}$, and $\mathcal{H}$ satisfies the weak learning assumption with respect to $\mathcal{C}$, then in fact:

1. f is multicalibrated with respect to $\mathcal{C}$, and
2. f has squared error at most that of the minimum error predictor in $\mathcal{C}$.

In fact, Gopalan et al. [2022] show that if f is multicalibrated with respect to $\mathcal{C}$, then it is an *omnipredictor* for $\mathcal{C}$, which implies that f has loss no more than the best function $c(x) \in \mathcal{C}$, where loss can be measured with respect to any Lipschitz convex loss function (not just squared error). Thus our results imply that to obtain an omnipredictor for $\mathcal{C}$, it is sufficient to be multicalibrated with respect to a class $\mathcal{H}$ that satisfies our weak learning assumption with respect to $\mathcal{C}$.

Theorem 6.2. *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ and two classes of functions $\mathcal{H}$ and $\mathcal{C}$ that are closed under affine transformations. Then if $f : \mathcal{X} \to [0, 1]$ is multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}$, and if $\mathcal{H}$ satisfies the weak learning condition relative to $\mathcal{C}$ and $\mathcal{D}$, then in fact f is multicalibrated with respect to $\mathcal{D}$ and $\mathcal{C}$ as well.*

Proof. We assume for simplicity that f has a countable range (which is without loss of generality e.g. whenever $\mathcal{X}$ is countable). Suppose for contradiction that f is not multicalibrated with respect to $\mathcal{C}$ and $\mathcal{D}$. In this case there must be some $c \in \mathcal{C}$ such that:

$$\sum_{v \in R(f)} \Pr[f(x) = v] \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [c(x)(y-v) | f(x) = v] \right)^2 > 0$$

Since $\mathcal{C}$ is closed under affine transformations (and so both c and $-c$ are in $\mathcal{C}$), there must be some $c' \in \mathcal{C}$ and some $v \in R(f)$ with $\Pr[f(x) = v] > 0$ such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [c'(x)(y-v) | f(x) = v] > 0$$

Therefore, by the first part of Theorem 3.2, there must be some $c'' \in \mathcal{C}$ such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [(c''(x) - y)^2 | f(x) = v] < \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [(v - y)^2 | f(x) = v]$$

Since $\mathcal{H}$ is closed under affine transformations, the function $h(x) = 1$ is in $\mathcal{H}$ and so multicalibration with respect to $\mathcal{H}$ implies calibration. Thus $v = \bar{y}_{S_v}$ for $S_v = \{x : f(x) = v\}$. Therefore, the fact that $\mathcal{H}$ satisfies the weak learning condition relative to $\mathcal{C}$ and $\mathcal{D}$ implies that there must be some $h \in \mathcal{H}$ such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [(h(x) - y)^2 | f(x) = v] < \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [(v - y)^2 | f(x) = v]$$

Finally, the second part of Theorem 3.2 implies that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [h(x)(y-v) | f(x) = v] > 0$$

which is a violation of our assumption that f is multicalibrated with respect to $\mathcal{H}$ and $\mathcal{D}$, a contradiction. □

Theorem 6.3. *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ and two classes of functions $\mathcal{H}$ and $\mathcal{C}$. Then if $f : \mathcal{X} \to [0,1]$ is calibrated and multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}$, and if $\mathcal{H}$ satisfies the weak learning condition relative to $\mathcal{C}$ and $\mathcal{D}$, then:*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [(f(x) - y)^2] \leqslant \min_{c \in \mathcal{C}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [(c(x) - y)^2]$$

Proof. We assume for simplicity that f has a countable range (which is without loss of generality e.g. whenever $\mathcal{X}$ is countable). Suppose for contradiction that there is some $c \in \mathcal{C}$ such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [(c(x) - y)^2] < \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [(f(x) - y)^2]$$

Then there must be some $v \in R(f)$ with $\Pr[f(x) = v] > 0$ and:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [(c(x) - y)^2 | f(x) = v] < \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [(v - y)^2 | f(x) = v]$$

Since f is calibrated, $v = \bar{y}_{S_v}$ for $S_v = \{x : f(x) = v\}$. Therefore, the fact that $\mathcal{H}$ satisfies the weak learning condition relative to $\mathcal{C}$ and $\mathcal{D}$ implies that there must be some $h \in \mathcal{H}$ such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [(h(x) - y)^2 | f(x) = v] < \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [(v - y)^2 | f(x) = v]$$

Finally, the second part of Theorem 3.2 implies that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [h(x)(y-v) | f(x) = v] > 0$$

which is a violation of our assumption that f is multicalibrated with respect to $\mathcal{H}$ and $\mathcal{D}$, a contradiction. □

We now turn to approximate versions of these statements. To do so, we need a refined version of one direction of Theorem 3.2 that shows us that if f witnesses a failure of multicalibration with respect to some $h \in \mathcal{H}$, then there is another function $h' \in \mathcal{H}$ that can be used to improve on f's squared error, *while controlling the norm* of h'.

Lemma 6.4. *Suppose $\mathcal{H}$ is closed under affine transformation. Fix a model $f : \mathcal{X} \to [0,1]$, a levelset $v \in R(f)$, and a bound $B > 0$. Then if there exists an $h \in \mathcal{H}$ such that $\max_{x \in \mathcal{X}} h(x)^2 \leqslant B$ and*

$$\mathbb{E}[h(x)(y - v) | f(x) = v] \geqslant \alpha,$$

for $\alpha \geqslant 0$, then there exists an $h' \in \mathcal{H}$ such that $\max_{x \in \mathcal{X}} h'(x)^2 \leqslant (1 + \frac{\sqrt{B}}{\alpha})^2$ and:

$$\mathbb{E}[(f(x) - y)^2 - (h'(x) - y)^2 | f(x) = v] \geqslant \frac{\alpha^2}{B}.$$

Proof. Let $h'(x) = v + \eta h(x)$ where $\eta = \frac{\alpha}{\mathbb{E}[h(x)^2 | f(x) = v]}$, as in Theorem 3.2. Because $h(x)^2$ is uniformly bounded by B on $\mathcal{X}$, it follows that $\mathbb{E}[h(x)^2] \leqslant B$, and we have already shown in the proof of Theorem 3.2 that this implies

$$\mathbb{E}[(f(x) - y)^2 - (h'(x) - y)^2 | f(x) = v] \geqslant \frac{\alpha^2}{B}.$$

It only remains to bound $\max_{x \in \mathcal{X}} h'(x)^2$. We begin by lower-bounding $\mathbb{E}[h(x)^2 \mid f(x) = v]$ in terms of α.

$$\begin{aligned}
\mathbb{E}[h(x)^2 \mid f(x) = v] &\geqslant \mathbb{E}[h(x) \mid f(x) = v]^2 \\
&\geqslant \mathbb{E}[h(x)(y - v) \mid f(x) = v]^2 \\
&\geqslant \alpha^2.
\end{aligned}$$

It follows that $\eta \leqslant 1/\alpha$, and so

$$\begin{aligned}
\max_{x \in \mathcal{X}} h'(x)^2 &= \max_{x \in \mathcal{X}} (v + \eta h(x))^2 \\
&\leqslant (1 + \eta\sqrt{B})^2 \\
&\leqslant \left(1 + \frac{\sqrt{B}}{\alpha}\right)^2.
\end{aligned}$$

□

We will also need a parameterized version of our weak learning condition. Recalling Remark 4.4, for approximate multicalibration to be meaningful with respect to a class that is closed under affine transformation, we must specify a bounded subset of that class with respect to which a predictor is approximately multicalibrated. Then to show that approximate multicalibration with respect to one potentially unbounded class implies approximate multicalibration with respect to another, we will need to specify the subsets of each class with respect to which a predictor is claimed to be approximately multicalibrated. This motivates a parameterization of our previous weak learning condition relative to a class $\mathcal{C}$. We will need to assume that whenever there is a B-bounded function in $\mathcal{C}$ that improves over the best constant predictor on a restriction of $\mathcal{D}$, there also exists a B-bounded function in $\mathcal{H}$ that improves on the restriction as well.

Definition 6.5 (B-Bounded Weak Learning Assumption Relative to $\mathcal{C}$)**.** *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ and two classes of functions $\mathcal{H}$ and $\mathcal{C}$. Fix a bound $B > 0$ and let $\mathcal{H}_B$ and $\mathcal{C}_B$ denote the sets*

$$\mathcal{H}_B = \{h \in \mathcal{H} : \max_{x \in \mathcal{X}} h(x)^2 \leqslant B\}$$

and

$$\mathcal{C}_B = \{c \in \mathcal{C} : \max_{x \in \mathcal{X}} c(x)^2 \leqslant B\}$$

respectively. We say that $\mathcal{H}$ satisfies the B-bounded γ-bounded weak learning *condition relative to $\mathcal{C}$ and $\mathcal{D}$ if for every $S \subseteq \mathcal{X}$ with $\Pr_{x\sim\mathcal{D}_\mathcal{X}}[x \in S] > 0$, if:*

$$\min_{c\in\mathcal{C}_B} \mathop{\mathbb{E}}_{(x,y)\sim D}[(c(x)-y)^2 \mid x \in S] < \mathop{\mathbb{E}}_{(x,y)\sim D}[(\bar{y}_S - y)^2 \mid x \in S] - \gamma,$$

where $\bar{y}_S = \mathbb{E}[y \mid x \in S]$, then there exists $h \in \mathcal{H}_B$ such that

$$\mathop{\mathbb{E}}_{(x,y)\sim D}[(h(x)-y)^2 \mid x \in S] < \mathop{\mathbb{E}}_{(x,y)\sim D}[(\bar{y}_S - y)^2 \mid x \in S] - \gamma.$$

Theorem 6.6. *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ and two classes of functions $\mathcal{H}$ and $\mathcal{C}$ that are closed under affine transformations. Fix $\alpha_\mathcal{C}, B > 0$. Let $B' = (1+\sqrt{\frac{2B}{\alpha_\mathcal{C}}})^2$ and $\gamma = \frac{\alpha_\mathcal{C}}{4B}$. Fix a function $f : \mathcal{X} \to [0,1]$ that maps into a countable subset of its range, and let $m = |R(f)|$, $\alpha_\mathcal{H} < \frac{\alpha_\mathcal{C}^3}{2^9 m B'^2}$, and $\alpha < \frac{\alpha_\mathcal{C}\gamma^2}{32mB'^2}$. Then if*

- *$\mathcal{H}$ satisfies the B'-bounded γ-weak learning condition relative to $\mathcal{C}$ and $\mathcal{D}$*
- *f is $\alpha_\mathcal{H}$-approximately multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}_{B'}$*
- *f is α-approximately calibrated on $\mathcal{D}$,*

then f is $\alpha_\mathcal{C}$-approximately multicalibrated with respect to $\mathcal{D}$ and $\mathcal{C}_B$.

Proof. Suppose not and there exists some $c \in \mathcal{C}_B$ such that

$$\sum_{v\in R(f)} \Pr_{x\sim\mathcal{D}_x}[f(x)=v] \cdot \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[c(x)(y-v) \mid f(x)=v]\right)^2 > \alpha_\mathcal{C}.$$

Then there must exist some $v \in R(f)$ such that $\Pr[f(x)=v] > \frac{\alpha_\mathcal{C}}{2m}$ and

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[c(x)(y-v) \mid f(x)=v]^2 > \alpha_\mathcal{C}/2.$$

Because $\mathcal{C}$ is closed under affine transformations, $\mathcal{C}_B$ is closed under negation, so there must also exist some $c' \in \mathcal{C}_B$ such that

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[c'(x)(y-v) \mid f(x)=v] > \sqrt{\alpha_\mathcal{C}/2}.$$

Then Lemma 3.3 shows that there is a $c'' \in \mathcal{C}_{(1+\sqrt{\frac{2B}{\alpha_\mathcal{C}}})^2} = \mathcal{C}_{B'}$ such that

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y-f(x))^2 - (y-c''(x))^2 \mid f(x)=v] \geqslant \frac{\alpha_\mathcal{C}}{2B} = 2\gamma.$$

Because f is α-calibrated on $\mathcal{D}$, by definition we have

$$\Pr_{x\sim\mathcal{D}_x}[f(x)=v] \cdot \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[v-y \mid f(x)=v]\right)^2 < \alpha.$$

Letting $\bar{y}_v = \mathbb{E}[y \mid f(x)=v]$, our lower-bound that $\Pr[f(x)=v] > \frac{\alpha_\mathcal{C}}{2m}$ gives us that $(v-\bar{y}_v)^2 < \frac{2\alpha m}{\alpha_\mathcal{C}} \leqslant \frac{\gamma^2}{16B'^2} < \gamma$. So, because v is close to $\bar{y}_v$, we can show the squared error of f must be close to the squared error of $\bar{y}_v$ on this level set.

$$\begin{aligned}
\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y-f(x))^2 \mid f(x)=v] &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y-\bar{y}_v+\bar{y}_v-f(x))^2 \mid f(x)=v] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y-\bar{y}_v)^2 + 2(y-\bar{y}_v)(\bar{y}_v-v) \mid f(x)=v] + (\bar{y}_v-v)^2 \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y-\bar{y}_v)^2 \mid f(x)=v] + (\bar{y}_v-v)^2 \\
&< \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y-\bar{y}_v)^2 \mid f(x)=v] + \gamma.
\end{aligned}$$

Then, because the squared error of c'' on this level set is much less than the squared error of f, we find that c'' must also have squared error less than that of $\bar{y}_v$:

$$\begin{aligned}\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y-\bar{y}_v)^2-(y-c''(x))^2 \mid f(x)=v] &> \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y-f(x))^2-\gamma-(y-c''(x))^2 \mid f(x)=v]\\ &\geqslant 2\gamma-\gamma\\ &=\gamma\end{aligned}$$

We assumed $\mathcal{H}$ satisfies the B'-bounded γ-weak learning condition relative to $\mathcal{C}$, so this gives us a function $h\in\mathcal{H}_{B'}$ such that

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y-\bar{y}_v)^2-(y-h(x))^2 \mid f(x)=v]>\gamma.$$

Then Lemma 3.3 shows that

$$\mathbb{E}[h(x)(y-\bar{y}_v) \mid f(x)=v]>\gamma/2.$$

So h witnesses a failure of multicalibration of f, since it follows that

$$\begin{aligned}\mathbb{E}[h(x)(y-v) \mid f(x)=v] &= \mathbb{E}[h(x)(y-\bar{y}_v) \mid f(x)=v]+\mathbb{E}[h(x)(\bar{y}_v-v) \mid f(x)=v]\\ &> \gamma/2-B'\left|\bar{y}_v-v\right|\\ &\geqslant \gamma/2-\frac{B'\gamma}{4B'}\\ &=\gamma/4\end{aligned}$$

and so

$$\Pr_{x\sim\mathcal{D}_x}[f(x)=v]\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y-v) \mid f(x)=v]\right)^2>\frac{\alpha_{\mathcal{C}}\gamma^2}{32m}>\alpha_{\mathcal{H}},$$

contradicting $\alpha_{\mathcal{H}}$-approximate multicalibration of f on $\mathcal{H}_{B'}$ and $\mathcal{D}$. □

In Gopalan et al. [2022], Gopalan, Kalai, Reingold, Sharan, and Wieder show that any predictor that is approximately multicalibrated for a class $\mathcal{H}$ and distribution $\mathcal{D}$ can be efficiently post-processed to approximately minimize any convex, Lipschitz loss function relative to the class $\mathcal{H}$. The theorem we have just proved can now be used to extend their result to approximate loss minimization over any other class $\mathcal{C}$, so long as $\mathcal{H}$ satisfies the B-bounded γ-weak learning assumption relative to $\mathcal{C}$. Intuitively, this follows from the fact that if f is approximately multicalibrated with respect to $\mathcal{H}$ on $\mathcal{D}$, it is also approximately multicalibrated with respect to $\mathcal{C}$. However, the notion of approximate multicalibration adopted in Gopalan et al. [2022] differs from the one in this work. So, to formalize our intuition above, we will first state the covariance-based definition of approximate multicalibration appearing in Gopalan et al. [2022] and prove a lemma relating it to our own. We note that, going forward, we will restrict ourselves to distributions $\mathcal{D}$ over $\mathcal{X}\times\{0,1\}$, as in this case the two definitions of approximate multicalibration are straightforwardly connected.

Definition 6.7 (Approximate Covariance Multicalibration Gopalan et al. [2022])**.** *Fix a distribution $\mathcal{D}$ over $\mathcal{X}\times\{0,1\}$ and a function $f:\mathcal{X}\to[0,1]$ that maps onto a countable subset of its range, denoted $R(f)$. Let $\mathcal{H}$ be an arbitrary collection of real valued functions $h:\mathcal{X}\to\mathbb{R}$. Then f is* α-approximately covariance multicalibrated *with respect to $\mathcal{H}$ on $\mathcal{D}$ if*

$$\sum_{v\in R(f)}\Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[f(x)=v]\cdot\left|\mathbb{E}[(h(x)-\bar{h}_v)(y-\bar{y}_v) \mid f(x)=v]\right|\leqslant\alpha,$$

where $\bar{h}_v=\mathbb{E}[h(x) \mid f(x)=v]$ and $\bar{y}_v=\mathbb{E}[y \mid f(x)=v]$.

Lemma 6.8. *Fix a distribution $\mathcal{D}$ over $\mathcal{X}\times\{0,1\}$ and a class of functions on $\mathcal{X}$, $\mathcal{H}$. Let $\mathcal{H}_B$ denote the subset*

$$\mathcal{H}_B=\{h\in\mathcal{H}:\max_{x\in\mathcal{X}}h(x)^2\leqslant B\}.$$

Fix a function $f : \mathcal{X} \to [0,1]$ that maps onto a countable subset of its range, denoted $R(f)$. Then if f is α-approximately multicalibrated with respect to $\mathcal{H}_B$ on $\mathcal{D}$, then f is $(\sqrt{\alpha}(1+\sqrt{B}))$-approximately covariance multicalibrated. That is, for all $h \in \mathcal{H}_B$, f satisfies

$$\sum_{v \in R(f)} \Pr[f(x) = v] \cdot \left|\mathbb{E}[(h(x) - \bar{h}_v)(y - \bar{y}_v) \mid f(x) = v]\right| \leqslant \sqrt{\alpha}(1+\sqrt{B}).$$

Proof.

$$\begin{aligned}
&\sum_{v \in R(f)} \Pr[f(x) = v] \cdot \left|\mathbb{E}[(h(x) - \bar{h}_v)(y - \bar{y}_v) \mid f(x) = v]\right| \\
&\qquad = \sum_{v \in R(f)} \Pr[f(x) = v] \cdot \left|\mathbb{E}[h(x)y \mid f(x) = v] - \bar{y}_v \bar{h}_v\right| \\
&\qquad = \sum_{v \in R(f)} \Pr[f(x) = v] \cdot \left|\mathbb{E}[h(x)y \mid f(x) = v] - v\bar{h}_v + v\bar{h}_v - \bar{y}_v \bar{h}_v\right| \\
&\qquad = \sum_{v \in R(f)} \Pr[f(x) = v] \cdot \left|\mathbb{E}[h(x)(y - v) \mid f(x) = v] + \bar{h}_v(v - \bar{y}_v)\right| \\
&\qquad \leqslant \sum_{v \in R(f)} \Pr[f(x) = v] \cdot \left(\left|\mathbb{E}[h(x)(y - v) \mid f(x) = v]\right| + \left|\bar{h}_v(v - \bar{y}_v)\right|\right) \\
&\qquad \leqslant \sqrt{\alpha} + \sqrt{B} \sum_{v \in R(f)} \Pr[f(x) = v] \cdot |v - \bar{y}_v| \\
&\qquad \leqslant \sqrt{\alpha}(1+\sqrt{B}).
\end{aligned}$$

where the second inequality follows from the fact that $\mathbb{E}[x] \leqslant \sqrt{\mathbb{E}[x^2]}$ and the bound $\max_{x \in \mathcal{X}} h(x)^2 \leqslant B$. □

We now recall a theorem of Gopalan et al. [2022], showing that approximate covariance multicalibration with respect to a class $\mathcal{H}$ implies approximate loss minimization relative to $\mathcal{H}$, for convex, Lipschitz losses.

Theorem 6.9. *Fix a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ and a class of real-valued functions on $\mathcal{X}$, $\mathcal{H}$. Fix a function $f : \mathcal{X} \to [0,1]$ that maps onto a countable subset of its range, denoted $R(f)$. Let $\mathcal{L}$ be a class of functions on $\{0,1\} \times \mathbb{R}$ that are convex and L-Lipschitz in their second argument. If f is α-approximately covariance multicalibrated with respect to $\mathcal{H}_B$ on $\mathcal{D}$, then for every $\ell \in \mathcal{L}$ there exists an efficient post-processing function k_ℓ such that*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, k_\ell(f(x)))] \leqslant \min_{h \in \mathcal{H}_B} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, h(x))] + 2\alpha L.$$

Corollary 6.10. *Fix a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ and two classes of real-valued functions on $\mathcal{X}$ that are closed under affine transformation, $\mathcal{H}$ and $\mathcal{C}$. Fix a function $f : \mathcal{X} \to [0,1]$ that maps onto a countable subset of its range, denoted $R(f)$. Let $\mathcal{L}$ be a class of functions on $\{0,1\} \times \mathbb{R}$ that are convex and L-Lipschitz in their second argument. Fix $\alpha_{\mathcal{C}}, B > 0$. Let $B' = (1 + \sqrt{\frac{2B}{\alpha_{\mathcal{C}}}})^2$ and $\gamma = \frac{\alpha_{\mathcal{C}}}{4B}$. Let $\alpha_{\mathcal{H}} < \frac{\alpha_{\mathcal{C}}^3}{2^9 m B'^2}$, and $\alpha < \frac{\alpha_{\mathcal{C}} \gamma^2}{32 m B'^2}$. Then if*

- *$\mathcal{H}$ satisfies the B'-bounded γ-weak learning condition relative to $\mathcal{C}$ and $\mathcal{D}$*
- *f is $\alpha_{\mathcal{H}}$-approximately multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}_{B'}$*
- *f is α-approximately calibrated on $\mathcal{D}$,*

then for every $\ell \in \mathcal{L}$ there exists an efficient post-processing function k_ℓ such that

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, k_\ell(f(x)))] \leqslant \min_{c \in \mathcal{C}_B} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, c(x))] + 2L\sqrt{\alpha_{\mathcal{C}}}(1+\sqrt{B}).$$

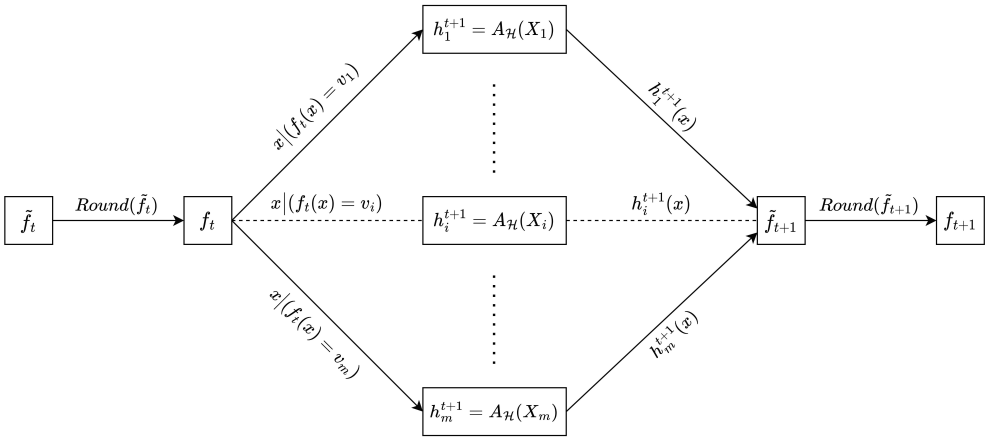

Figure 1: The update process at round t with m level sets during training.

Proof. We have from Theorem 6.6 that given the assumed conditions, f will be $\alpha_{\mathcal{C}}$-approximately multicalibrated with respect to $\mathcal{C}_B$ on $\mathcal{D}$. It follows from Lemma 6.8 that f is $\sqrt{\alpha_{\mathcal{C}}}(1+\sqrt{B})$-approximately covariance multicalibrated with respect to $\mathcal{C}_B$ on $\mathcal{D}$. The result of Gopalan et al. [2022] then gives us that for all $\ell \in \mathcal{L}$, there exists an efficient post-processing function k_ℓ such that

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, k_\ell(f(x)))] \leqslant \min_{c\in\mathcal{C}_B} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, c(x))] + 2L\sqrt{\alpha_{\mathcal{C}}}(1+\sqrt{B}).$$

□

7 Empirical Evaluation

In this section, we study Algorithm 1 empirically via an efficient, open-source Python implementation of our algorithm on both synthetic and real regression problems. Our code is available here: `https://github.com/Declancharrison/Level-Set-Boosting`. An important feature of Algorithm 1 which distinguishes it from traditional boosting algorithms is the ability to parallelize not only during inference, but also during training. Let f_t be the model maintained by Algorithm 1 at round t with m level sets. Given a data set X, f_t creates a partition of X defined by $X_i^{t+1} = \{x | f_t(x) = v_i\}$. Since the X_i are disjoint, each call $h_i^{t+1} = A_{\mathcal{H}}(X_i^{t+1})$ can be made on a separate worker followed by a combine and round operation to obtain $\tilde{f}_{t+1}$ and f_{t+1} respectively, as shown in Figure 1. A parallel inference pass at round t works nearly identically, but uses the historical weak learners h_i^{t+1} obtained from training and applies them to each set X_i^{t+1}.

7.1 Prediction on Synthetic Data

From Theorem 5.2, we know that multicalibration with respect to a hypothesis class $\mathcal{H}$ satisfying our weak learning condition implies Bayes optimality. To visualize the fast convergence of our algorithm to Bayes optimality, we create two synthetic datasets; each dataset contains one million samples with two features. We label these points using two functions, C_0 and C_1, defined below and pictured in Figure 2). We attempt to learn the underlying function with Algorithm 1.

$$C_0(x) = \begin{cases} (x+1)^2 + (y-1)^2, & \text{if } x \leqslant 0, y \geqslant 0 \\ (x-1)^2 + (y-1)^2, & \text{if } x > 0, y \geqslant 0 \\ (x+1)^2 + (y+1)^2, & \text{if } x \leqslant 0, y < 0 \\ (x-1)^2 + (y+1)^2, & \text{if } x > 0, y < 0 \end{cases} \tag{C_0}$$

$$C_1(x) = \begin{cases} x + 20xy^2\cos(-8x)\sin(8y)\left(\frac{(1.5x+4)(x+1)^2}{y+3} + (y-1)^2\right), & \text{if } x \leqslant 0, y \geqslant 0 \\ x + 20xy^2\cos(8x)\sin(8y)\left(\frac{(1.5x+4)(x-1)^2}{y+3} + (y-1)^2\right), & \text{if } x > 0, y \geqslant 0 \\ x + 20xy^2\cos(-8x)\sin(8y)\left(\frac{(1.5x+4)(x+1)^2}{y+3} + (y+1)^2\right), & \text{if } x \leqslant 0, y < 0 \\ x + 20xy^2\cos(8x)\sin(8y)\left(\frac{(1.5x+4)(x-1)^2}{y+3} + (y+1)^2\right), & \text{if } x > 0, y < 0 \end{cases} \tag{C_1}$$

In Figure 3, we show an example of Algorithm 1 learning C_0 using a discretization of five-hundred level sets and a weak learner hypothesis class of depth one decision trees. Each image in figure 3 corresponds to the map produced by Algorithm 1 at the round listed in the top of the image. As the round count increases, the number of non-empty level sets increases until each level set is filled, at which point the updates become more granular. The termination round titled 'final round' occurs at $T = 199$ and paints an approximate map of C_0. The image titled 'out of sample' is the map produced on a set of one million points randomly drawn outside of the training sample, and shows that Algorithm 1 is in fact an approximation of the Bayes Optimal C_0.

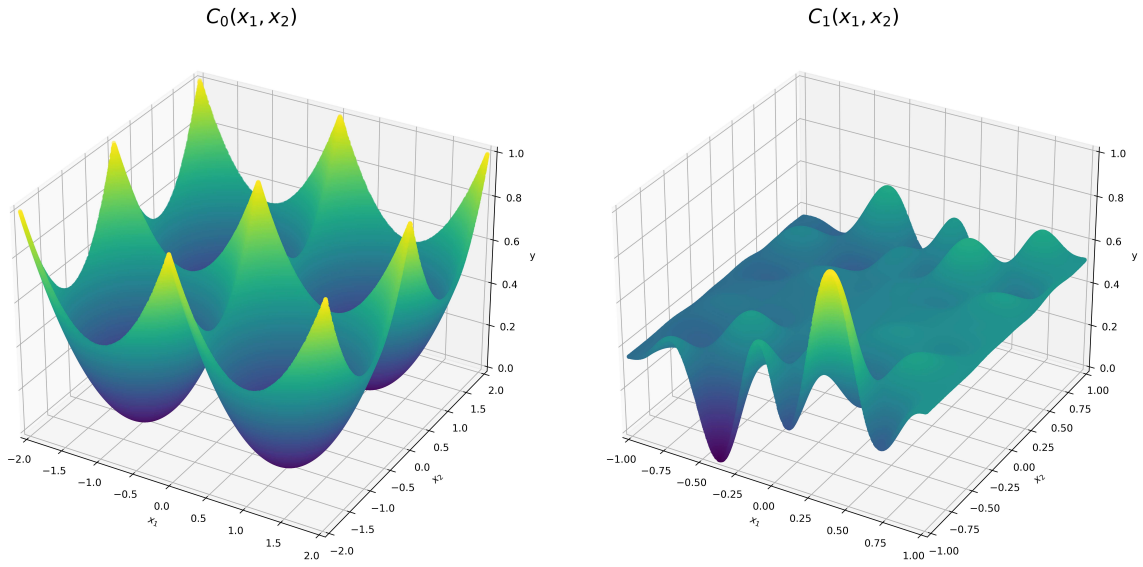

Figure 2: C_0 maps $x_1, x_2 \in [-2, 2]$ to four cylindrical cones symmetric about the origin. C_1 maps $x_1, x_2 \in [-1, 1]$ to a hilly terrain from a more complex function.

Figure 4 plots the same kind of progression as Figure 3, but with a more complicated underlying function C_1 using a variety of weak learner classes. We are able to learn this more complex surface out of sample with all base classes except for linear regression, which results in a noisy out-of-sample plot.

7.2 Prediction on Census Data

We evaluate the empirical performance of Algorithm 1 on US Census data compiled using the Python folktables package Ding et al. [2021]. In this dataset, the feature space consists of demographic information about individuals (see Table 1), and the labels correspond to the individual's annual income.

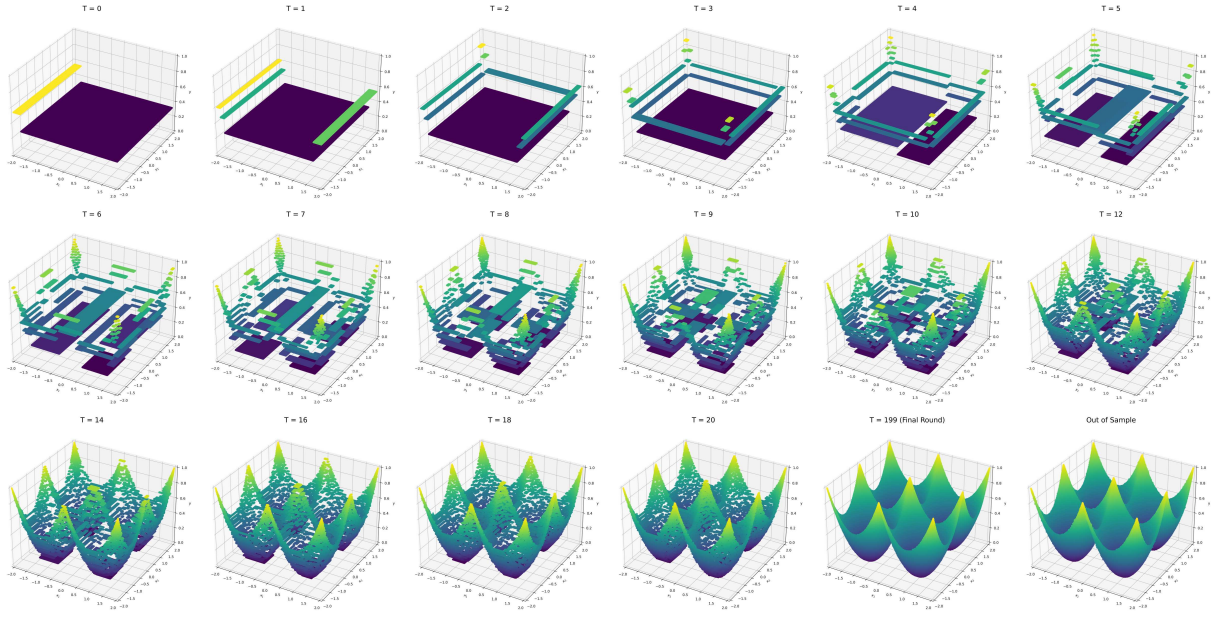

Figure 3: Evolution of Algorithm 1 learning C_0.

feature	description	feature	description
AGEP	age	POBP	place of birth
COW	class of worker	RELP	relationship
SCHL	education level	WKHP	work hours per week
MAR	marital status	SEX	binary sex
OCCP	occupation	RAC1P	race

Table 1: Features included in income prediction task.

We cap income at \$100,000 and then rescale all labels into $[0, 1]$. On an 80/20% train-test split with 500,000 total samples, we compare the performance of Algorithm 1 with Gradient Boosting with two performance metrics: mean squared error (MSE), and mean squared calibration error (MSCE). For less expressive weak learner classes (such as DT(1), see Figure 5), Algorithm 1 has superior MSE out of sample compared to Gradient Boosting through one hundred rounds while maintaining significantly lower MSCE, and converges quicker. However, as the weak learning class becomes more expressive (e.g. increasing decision tree depths), Algorithm 1 is more prone to overfitting than gradient boosting (see Figure 6).

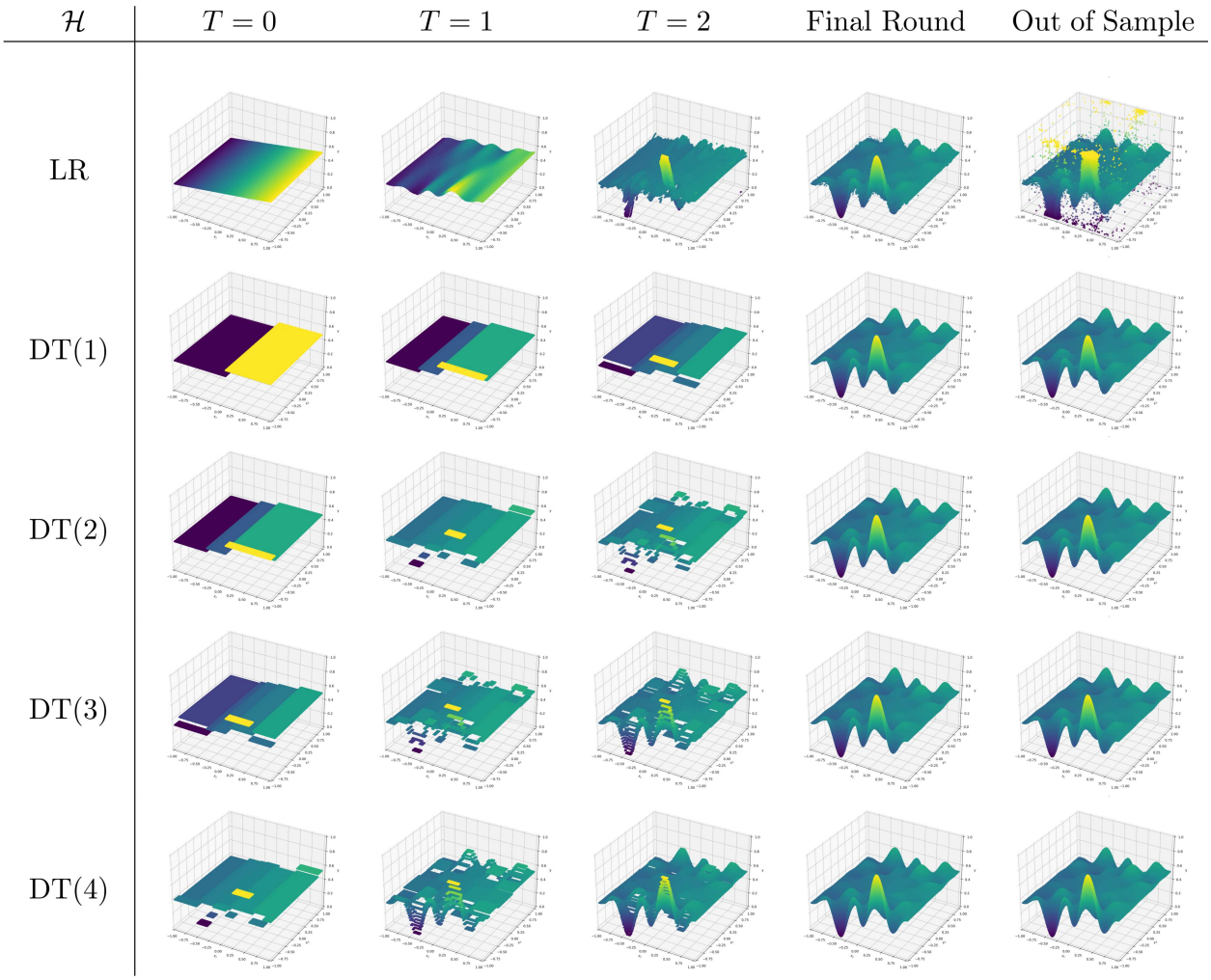

Figure 4: Stages of Algorithm 1 learning C_1 with linear regression (LR) and varying depth d decision trees (DT(d)). In the out of sample plot for linear regression, points are not mapped to their proper position, implying C_1 cannot be learned by boosting linear functions. All other hypothesis classes eventually converge to C_1.

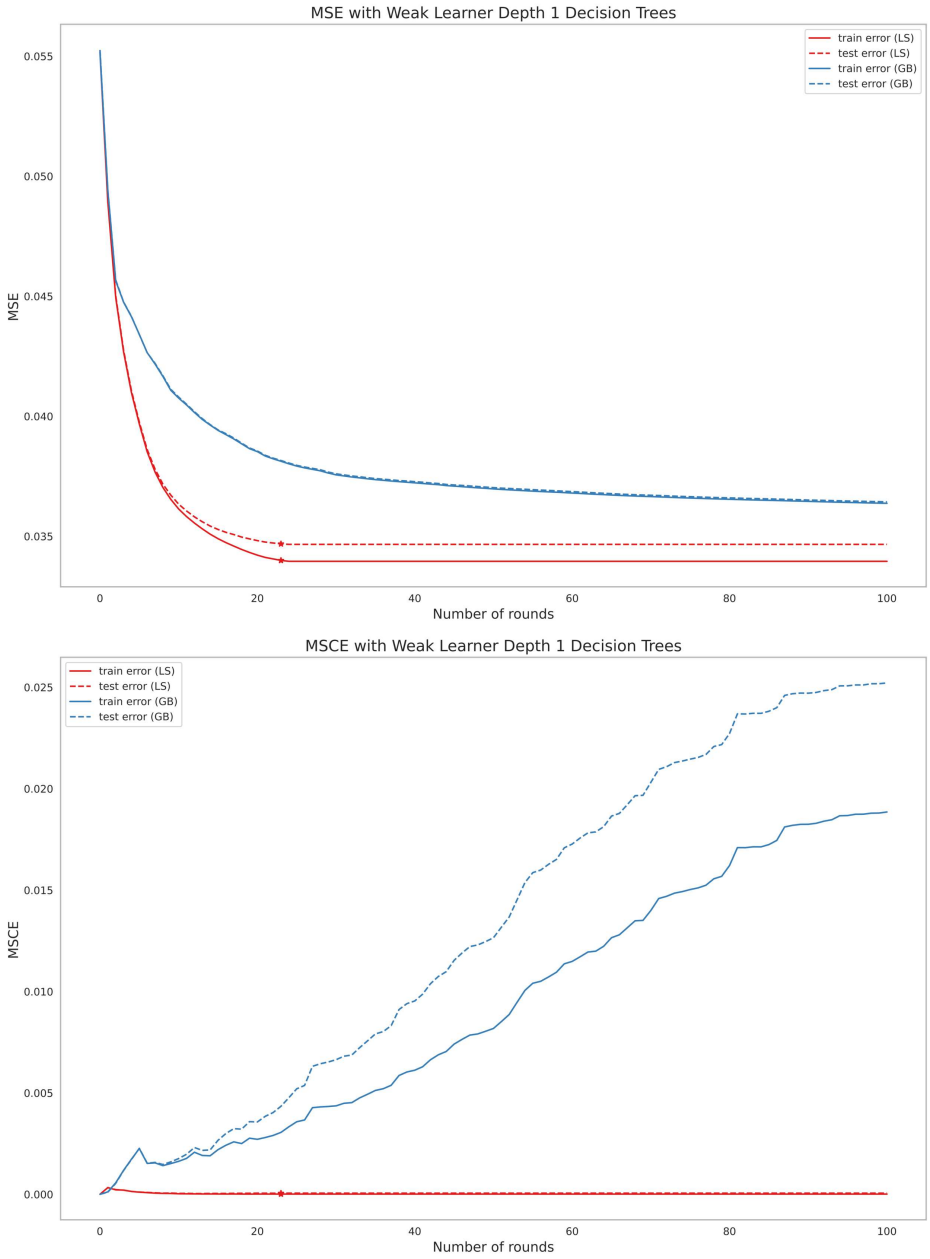

Figure 5: Comparison of Algorithm 1 (LS) and Gradient Boosting (GB), both using depth 1 regression trees. * indicates termination round of Algorithm 1.

In Table 2, we compare the time taken to train n weak learners with Algorithm 1 and with scikit-learn's version of Gradient Boosting on our census data. Recall that our algorithm trains multiple weak learners per round of boosting, and so comparing the two algorithms for a fixed number of calls to the weak learner is distinct from comparing them for a fixed number of rounds. Because models output by Algorithm 1 may be more complex than those produced by Gradient Boosting run for the same number of rounds, we use number of weak learners trained as a proxy for model complexity, and compare the two algorithms holding this measure fixed. We see the trend for Gradient Boosting is linear with respect to number of weak learners, whereas Algorithm 1 does not follow the same linear pattern upfront. This is due to not being able to fully

leverage parallelization of training weak learners in early stages of boosting. At each round, Algorithm 1 calls the weak learner on every large enough level set of the current model, and it is these independent calls that can be easily parallelized. However, in the early rounds of boosting the model may be relatively simple, and so many level sets may be sparsely populated. As the model becomes more expressive over subsequent rounds, the weak learner will be invoked on more sets per round, allowing us to fully utilize parallelizability.

# Weak Learners	DT(1)			DT(2)			DT(3)		
	LS	GB	Faster?	LS	GB	Faster?	LS	GB	Faster?
50 level sets									
100	9.11	11.97	✓	5.86	23.01	✓	6.88	32.92	✓
300	18.70	35.81	✓	14.90	69.17	✓	15.64	102.14	✓
500	27.00	58.19	✓	21.74	115.65	✓	24.77	169.90	✓
1000	46.73	116.49	✓	42.92	231.74	✓	46.38	336.89	✓
100 level sets									
100	7.18	11.97	✓	5.29	23.01	✓	5.06	32.92	✓
300	13.08	35.81	✓	13.55	69.17	✓	14.72	102.14	✓
500	21.20	58.19	✓	19.57	115.65	✓	21.79	169.90	✓
1000	41.99	116.49	✓	36.26	231.74	✓	40.92	336.89	✓
300 level sets									
100	5.87	11.97	✓	9.18	23.01	✓	6.54	32.92	✓
300	13.21	35.81	✓	17.46	69.17	✓	11.13	102.14	✓
500	19.05	58.19	✓	22.20	115.65	✓	19.64	169.90	✓
1000	32.80	116.49	✓	36.61	231.74	✓	27.12	336.89	✓

Table 2: Time (in seconds) comparison of Algorithm 1 (LS) with fifty level sets and Gradient Boosting to train certain numbers of estimators for various weak learner classes.

In Figure 6, we measure MSE and MSCE for Algorithm 1 and Gradient Boosting over rounds of training on our census data. Again, we note that one round of Algorithm 1 is not equivalent to one round of Gradient Boosting, but intend to demonstrate error comparisons and rates of convergence. For the linear regression plots, Gradient Boosting does not reduce either error since combinations of linear models are also linear. As the complexity of the underlying model class increases, Gradient Boosting surpasses Algorithm 1 in terms of MSE, though it does not minimize calibration error.

We notice that Algorithm 1, like most machine learning algorithms, is prone to overfitting when allowed. Future performance hueristics we intend to investigate include validating updates, complexity penalties, and weighted mixtures of updates.

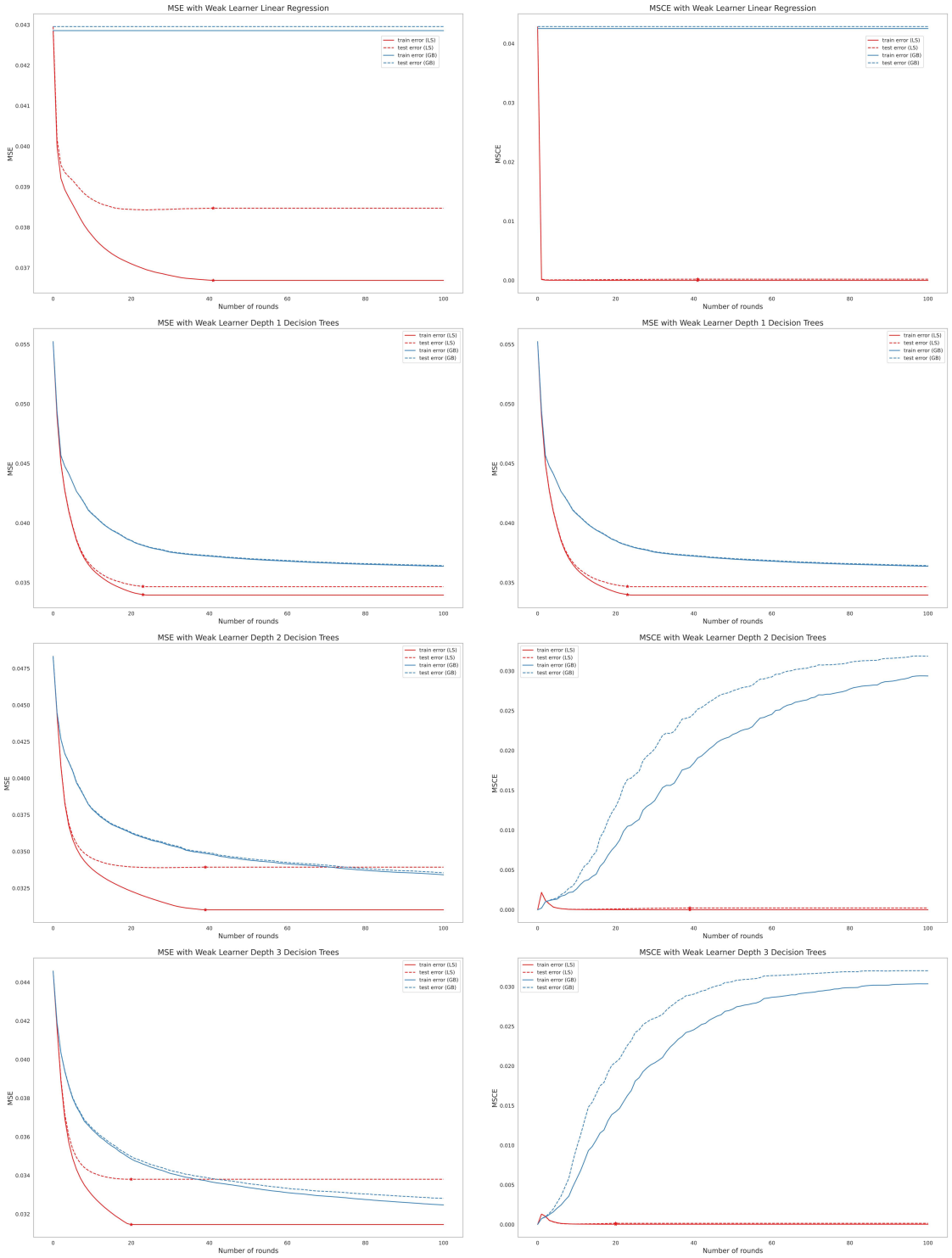

Figure 6: MSE and MSCE comparison of Algorithm 1 (LS) and Gradient Boosting (GB) on linear regression and decision trees of varying depths. * indicates termination round of LS and occurs, from top to bottom, at $T = 41, 23, 39, 20$.

References

Avrim Blum and Yishay Mansour. From external to internal regret. In *International Conference on Computational Learning Theory*, pages 621–636. Springer, 2005.

Maya Burhanpurkar, Zhun Deng, Cynthia Dwork, and Linjun Zhang. Scaffolding sets. *arXiv preprint arXiv:2111.03135*, 2021.

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

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Nigel Duffy and David Helmbold. Boosting methods for regression. *Machine Learning*, 47(2):153–200, 2002.

Dean P Foster and Rakesh Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29(1-2):7–35, 1999.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

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

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *ITCS*, 2022.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory*, pages 2634–2678. PMLR, 2021.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. *arXiv preprint arXiv:2209.15145*, 2022.

Adam Kalai. Learning monotonic linear functions. In *International Conference on Computational Learning Theory*, pages 487–501. Springer, 2004.

Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

Varun Kanade and Adam Kalai. Potential-based agnostic boosting. *Advances in neural information processing systems*, 22, 2009.

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

Michael P Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, and Omer Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4):e2108097119, 2022.

Balas K Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

David Pollard. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

Aaron Roth. Uncertain: Modern topics in uncertainty estimation. https://www.cis.upenn.edu/ aaroth/uncertainty-notes.pdf, 2022.

Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

Robert E Schapire and Yoav Freund. Boosting: Foundations and algorithms. *Kybernetes*, 2013.

Eliran Shabat, Lee Cohen, and Yishay Mansour. Sample complexity of uniform convergence for multicalibration. *Advances in Neural Information Processing Systems*, 33:13331–13340, 2020.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

V.N. Vapnik and A. YA. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities, 1971.

A Generalization Bounds

Our analysis of Algorithm 1 assumed direct access to the data distribution $\mathcal{D}$. In practice, we will run the algorithm on the empirical distribution over a sample of n points $D \sim \mathcal{D}^n$. In this section, we show that when we do this, so long as n is sufficiently large, both our squared error and our multicalibration guarantees carry over from the empirical distribution over D to the distribution $\mathcal{D}$ from which D was sampled. Most generalization bounds for multicalibration algorithms (e.g. Hébert-Johnson et al. [2018], Jung et al. [2021, 2022], Shabat et al. [2020]) are either stated and proven for finite classes $\mathcal{H}$, or are proven for algorithms that do not operate as empirical risk minimization algorithms, but instead gain access to a fresh sample of data from the distribution at each iteration, or are proven for hypotheses classes that are fixed independently of the algorithm. We have a different challenge: Like Hébert-Johnson et al. [2018], Jung et al. [2021] we study an iterative algorithm whose final hypothesis class is not fixed up front, but implicitly defined as a function of $\mathcal{H}$. But we wish to study the algorithms as they are used—as empirical risk minimization algorithms—so we do not want our analysis to depend on using a fresh sample of data at each iteration. And unlike the analysis in Jung et al. [2022], for us $\mathcal{H}$ is continuously large (since it is closed under affine transformations), so we cannot rely on bounds that depend on $\log|\mathcal{H}|$. Instead we give a uniform convergence analysis that depends on the pseudo-dimension of our class of weak learners $\mathcal{H}$:

Definition A.1. *Pseudodimension[Pollard [2012]] Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. We say that a set $S = (x_1, \ldots, x_m, y_1, \ldots, y_m) \in \mathcal{X}^m \times \mathbb{R}^m$ is pseudo-shattered by $\mathcal{H}$ if for any $(b_1, \ldots, b_m) \in \{0,1\}^m$ there exists $h \in \mathcal{H}$ such that $\forall i, h(x_i) > y \iff b_i = 1$ The pseudodimension of $\mathcal{H}$, denoted $\mathrm{Pdim}(\mathcal{H})$ is the largest integer m for which $\mathcal{H}$ pseudo-shatters some set S of cardinality m.*

Although hypotheses in $\mathcal{H}$ are continuously valued, Algorithm 1 outputs functions that have finite range $[1/m]$, and so we can view them as multi-class classification functions. Our analysis will proceed by studying the generalization properties of these multiclass functions, which we will characterize using Natarajan dimension:

Definition A.2 (Shattering for multiclass functions)**.** *Natarajan [1989], Shalev-Shwartz and Ben-David [2014] A set $C \subseteq \mathcal{X}$ is shattered by $\mathcal{H}$ if there exists two functions $f_0, f_1 : C \to [k]$ such that*

1. *For every $x \in C, f_0(x) \neq f_1(x)$.*
2. *For every $B \subseteq C$ there exists a function $h \in \mathcal{H}$ such that*

$$\forall x \in B, h(x) = f_0(x) \text{ and } \forall x \in C \setminus B, h(x) = f_1(x).$$

Definition A.3 (Natarajan dimension)**.** *Natarajan [1989], Shalev-Shwartz and Ben-David [2014] The Natarajan dimension of $\mathcal{H}$, denoted $\mathrm{Ndim}(\mathcal{H})$, is the maximal size of a shattered set $C \subseteq \mathcal{X}$.*

We can then rely the following standard uniform convergence bound for multiclass classification. This statement is slightly modified from the result in Shalev-Schwartz and Ben-David to account for our use of squared error. The result still holds on account of the fact that the Cherhoff bound only relies on the loss function being bounded, and ours is indeed bounded between 0 and 1.

Theorem A.4 (Multiclass uniform convergence)**.** *Shalev-Shwartz and Ben-David [2014] Let $\epsilon, \delta > 0$ and let $\mathcal{H}$ be a class of functions $h : \mathcal{X} \to [1/k]$ such that the Natarajan dimension of $\mathcal{H}$ is d. Let $\mathcal{D} \in \Delta(\mathcal{X} \times [0,1])$ be an arbitrary distribution and let $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}_{(x_i,y_i)\sim\mathcal{D}}$ be a sample of n points from $\mathcal{D}$. Then for*

$$n = O\left(\frac{d\log(k) + \log(1/\delta)}{\varepsilon^2}\right),$$

$$\Pr\left[\max_{h\in\mathcal{H}}\left|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y - h(x))^2] - \mathop{\mathbb{E}}_{(x,y)\sim D}[(y - h(x))^2]\right| \geqslant \epsilon\right] \leqslant \delta.$$

Our strategy will be to bound the Natarajan dimension of the class of models that can be output by Algorithm 1 in terms of the pseudodimension of the underlying weak learner, then apply the above uniform convergence result. To do so, we will first use the following lemma, which bounds the Natarajan dimension of functions that can be described as post-processings of binary valued-functions from a class of bounded VC-dimension.

Lemma A.5. *Shalev-Shwartz and Ben-David [2014] Suppose we have ℓ binary classifiers from binary class $\mathcal{H}_{\text{bin}}$ and a rule $r : \{0,1\}^\ell \to [k]$ that determines a multiclass label according to the predictions of the ℓ binary classifiers. Define the hypothesis class corresponding to this rule as*

$$\mathcal{H} = \{r(h_1(\cdot), \ldots, h_\ell(\cdot)) : (h_1, \ldots, h_\ell) \in (\mathcal{H}_{bin})^\ell\}.$$

Then, if $d = \text{VCdim}(\mathcal{H}_{\text{bin}})$,

$$\text{Ndim}(\mathcal{H}) \leqslant 3\ell d\log(\ell d).$$

Recall that the VC-dimension of a binary classifier is defined as follows:

Definition A.6 (VC-dimension)**.** *Vapnik and Chervonenkis [1971] Let $\mathcal{H}$ be a class of binary classifiers $h : \mathcal{X} \to \{0,1\}$. Let $S = \{x_1, \ldots, x_m\}$ and let $\Pi_{\mathcal{H}}(S) = \{(h(x_1), \ldots, h(x_m)) : h \in \mathcal{H}\} \subseteq \{0,1\}^m$. We say that S is shattered by $\mathcal{H}$ if $\Pi_{\mathcal{H}}(S) = \{0,1\}^m$. The Vapnik-Chervonenkis (VC) dimension of $\mathcal{H}$, denoted VCdim($\mathcal{H}$), is the cardinality of the largest set S shattered by $\mathcal{H}$.*

Lemma A.7. *Let $\mathcal{H}_{\text{boost}}$ be the class of models output by* RegressionMulticalibrate$(f, \alpha, A_{\mathcal{H}}, \cdot, B)$ *(Algorithm 1) for any input distribution $\mathcal{D}$ and let d be the pseudodimension of its input weak learner class $\mathcal{H}$.*

$$\text{Ndim}\left(\mathcal{H}_{\text{boost}}\right) \leqslant 24(B/\alpha)^3 d\log\left((2B/\alpha)^3 d\right).$$

Proof. Let m be defined (as in *RegressionMulticalibrate*$(f, \alpha, A_{\mathcal{H}}, \mathcal{D}, B)$) to be $2B/\alpha$. Because our models are always rounded to the nearest value in $[1/m]$, we can think of the model f_t generated in every round of the algorithm multiclass classification problems over m classes. We will show that our final model can be written as a decision rule that maps the outputs of some ℓ Boolean classifiers to $[1/m]$, and that these Boolean classifiers have VC dimension that is bounded by the pseudodimension of the weak learner class. Then, we will apply Lemma A.5 to get an upper bound on the Natarajan dimension of the class of models in terms of α, B, and the pseudodimension of the input weak learner class $\mathcal{H}$.

Consider the initial round of the algorithm. We can convert our (rounded) initial regressor f_0 to a series of m Boolean thresholdings g_v which return 1 when $f_0(x) \geqslant v$:

$$g_v^0 = \begin{cases} 1 & \text{if } f_0(x) \geqslant v, \\ 0 & \text{otherwise.} \end{cases}.$$

These m Boolean thresholdings can then be mapped back to the original prediction over $[1/m]$ using a decision rule $r : \{0,1\}^m \to [1/m]$ which picks the largest of the thresholds that evaluates to 1, and assigns that index to the prediction:

$$r_0(\{g_v^0\}_{v\in[1/m]})(x) = \arg\max_{i\in[1/m]} i\mathbb{1}[g_v(x) = 1].$$

Note that since our initial predictor f_0 was already rounded to take values in $[1/m]$, the largest v such that $f_0(x) \geqslant v$ will be exactly $f_0(x)$, so r_0 is exactly equivalent to f_0. Similarly, at round $t+1$ of RegressionMulticalibrate$(f, \alpha, A_{\mathcal{H}}, \mathcal{D}, B)$, we will show that the model f_{t+1} can be written as a decision rule r_{t+1} over $m + (t+1)m^2$ binary classifiers g, where

$$g_{v,i}^t = \begin{cases} 1 & \text{if } h_v^t(x) \geqslant i - 1/(2m), \\ 0 & \text{otherwise.} \end{cases},$$

Here, the thresholds measure halfway between each level set, as $h_v^t(x)$ has yet to be rounded to the nearest level set. We can write a decision rule that maps these thresholds to classifications over $[1/m]$:

$$r_{t+1}\left(r_t, \{g_{v,i}^{t+1}\}_{i,v\in[1/m]}\right)(x) = \sum_{v\in[1/m]} \mathbb{1}[r_t(x) = v] \arg\max_{i\in[1/m]} \left(i \cdot \mathbb{1}[g_{v,i}^{t+1}(x) = 1]\right),$$

Now, we need to show that this decision rule evaluated at round t is equivalent to f_t. We proceed inductively. For our base case, we have already argued that our initial decision rule r_0 is equivalent to the classifier f_0. Now, say that we have decision rule r_t over binary classifiers g that is equivalent to model f_t. Then, we can write

$$\begin{aligned}
r_{t+1}\left(r_t, \{g_{v,i}^{t+1}\}_{i,v\in[1/m]}\right)(x) &= \sum_{v\in[1/m]} \mathbb{1}[r_t(x) = v] \arg\max_{i\in[1/m]} \left(i \cdot \mathbb{1}[g_{v,i}^{t+1}(x) = 1]\right), \\
&= \sum_{v\in[1/m]} \mathbb{1}[f_t(x) = v] \arg\max_{i\in[1/m]} \left(i \cdot \mathbb{1}[g_{v,i}^{t+1}(x) = 1]\right) \\
&= \sum_{v\in[1/m]} \mathbb{1}[f_t(x) = v] \arg\max_{i\in[1/m]} \left(i \cdot \mathbb{1}[h_v^{t+1}(x) \geqslant i - 1/(2m)]\right) \\
&= \sum_{v\in[1/m]} \mathbb{1}[f_t(x) = v]\text{Round}(h_v^{t+1}(x)) \\
&= f_{t+1}(x),
\end{aligned}$$

where the second line comes from the inductive hypothesis and the second to last line's equality comes from the fact that the largest i such that $h_v^{t+1}(x) - 1/(2m) \geqslant i$ will be the exact rounded prediction of $h_v^{t+1}(x)$.

Now, we need to show that at round $t+1$, the decision rule is a decision rule over $m + (t+1)m^2$ binary classifiers. Note that our initial decision rule r_0 has $m = m + 0 \cdot m^2$ binary classifiers. Say that at round t we have a decision rule r_t over $m + tm^2$ classifiers. In the following round, we build m^2 new Boolean classifiers g_v, i for $v, i \in [1/m]$. So, at round $t+1$ we have $m + tm^2 + m^2 = m + (t+1)m^2$ classifiers total.

From Theorem 4.3, we know that Algorithm 1 halts after at most $T \leqslant 2B/\alpha$ rounds, at which point it outputs model f_{T-1}. So, we can rewrite f_{T-1} as a decision rule r_{T-1} composed of at most $m + (T-1)m^2 < Tm^2$ Boolean models. Plugging in our bound for T and definition of m, this gives us a decision rule r_{T-1} composed of at most $\left(\frac{2B}{\alpha}\right)^3$ Boolean classifiers.

Let $\mathcal{G}$ be the class of Boolean threshold functions over $\mathcal{H}$, i.e. functions $g : \mathcal{X} \to \{0,1\}$ such that

$$g(x) = \begin{cases} 1 & h(x) \geqslant i \\ 0 & h(x) < i, \end{cases}$$

for some $h \in \mathcal{H}$ and $i \in \mathbb{R}$. Say that the VC-dimension of $\mathcal{G}$ is d'. Then, applying lemma A.5, it follows that

$$\begin{aligned}\mathrm{Ndim}(\mathcal{H}_{\mathrm{boost}}) &\leqslant 3\left(\frac{2B}{\alpha}\right)^3 d' \log\left(\left(\frac{2B}{\alpha}\right)^3 d'\right), \\ &= 24\left(\frac{B}{\alpha}\right) d' \log\left(\left(\frac{2B}{\alpha}\right)^3 d'\right).\end{aligned}$$

Now, it remains to show that we can bound the VC-dimension of these thresholding functions by the pseudodimension of the weak learner class $\mathcal{H}$. Note that $\mathcal{G}$ as we have defined it above is a richer hypothesis class than the actual class of thresholding functions used in the above analysis, because it can threshold at any value in $\mathbb{R}$ rather than being restricted to $[1/m]$. Thus, its VC dimension can only be greater than the VC dimension of the class of threshold functions over $\mathcal{H}$ restricted to $[1/m]$, and hence an upper bound on the VC dimension of $\mathcal{G}$ in terms of the pseudodimension of $\mathcal{H}$ will also be an upper bound on the VC dimension of the restricted class of threshold functions.

Let d be the pseudodimension of $\mathcal{H}$, and say that $d < d'$. By the definition of VC-dimension, $\{0,1\}^{d+1}$ must be shattered by $\mathcal{G}$. I.e., for any set of $d+1$ points $x_1, \ldots, x_{d+1} \in \mathcal{X}$ with arbitrary labels $b_1, \ldots, b_{d+1}$, there is some hypothesis $g \in \mathcal{G}$ that realizes those labels on $(x_1, \ldots, x_{d+1})$. Consider the function g that, given the $d+1$ points in $\mathcal{X}$, realizes the labels $b_1, \ldots, b_{d+1}$. By the construction of $\mathcal{G}$, g is a thresholding of some function $h \in \mathcal{H}$ at some point i. So, there is be some $i \in \mathbb{R}$ such that $h(x_i) > i \Rightarrow b_i = 1$ and such that $b_i = 1 \Rightarrow h(x_i) > i$. But this means that $\{0,1\}^{d+1}$ is pseudo-shattered by $\mathcal{H}$, and thus the pseudodimension of $\mathcal{H}$ is not d. Thus, it cannot be the case that $d < d'$, and hence $d' \leqslant d$, i.e. the VC dimension of $\mathcal{G}$ is bounded above by the pseudodimension of $\mathcal{H}$. Plugging this bound into the above bound on the Natarajan dimension gives us that

$$\begin{aligned}\mathrm{Ndim}(\mathcal{H}_{\mathrm{boost}}) &\leqslant 24\left(\frac{B}{\alpha}\right) d' \log\left(\left(\frac{2B}{\alpha}\right)^3 d'\right), \\ &\leqslant 24\left(\frac{B}{\alpha}\right) d \log\left(\left(\frac{2B}{\alpha}\right)^3 d\right).\end{aligned}$$

□

Now, we can state the following uniform convergence theorem for our final model.

Theorem A.8 (Squared Error Generalization for Algorithm 1.)**.** *Let $\epsilon, \delta, \alpha, B > 0$. Let $\mathcal{H}_{\mathrm{boost}}$ be the class of models that can be output by* RegressionMulticalibrate$(f, \alpha, A_{\mathcal{H}}, \cdot, B)$ *(Algorithm 1) for any input distribution $\mathcal{D}$ and let d be the pseudodimension of its input weak learner class $\mathcal{H}$. Let $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}_{(x_i,y_i)\sim\mathcal{D}}$ be a sample of n points drawn i.i.d. from $\mathcal{D}$. Then if*

$$n = O\left(\frac{dB^3 \log^2(dB/\alpha)}{\alpha^3\epsilon^2} + \frac{\log(1/\delta)}{\epsilon^2}\right)$$

$$\Pr\left[\max_{h\in\mathcal{H}_{\mathrm{boost}}} \left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y-h(x))^2] - \mathop{\mathbb{E}}_{(x,y)\sim D}[(y-h(x))^2] \right| \geqslant \epsilon\right] \leqslant \delta.$$

Proof. This follows directly from Theorem A.4 and the bound on the Natarajan dimension in Lemma A.7. □

We also would like to know that our multicalibration guarantees are generalizable. Rather than doing a bespoke analysis here, we can rely on the connection that we have established between failure of multicalibration and ability to improve squared error and argue that if the final hypothesis output by the algorithm was not multicalibrated with high probability then it would be possible to improve its squared error out-of-sample. Thus, by our previous generalization result for squared error, it would be possible to improve the squared error in-sample as well, giving us a contradiction.

Theorem A.9 (Multicalibration generalization guarantee)**.** *Let $\epsilon, \delta, \alpha, B > 0$ and consider the model f_{T-1} output by* RegressionMulticalibrate$(f, \alpha, A_{\mathcal{H}}, D, B)$ *for some sample D of n points drawn i.i.d. from distribution $\mathcal{D}$ such that*

$$n = O\left(\frac{dB^3 \log^2(dB/\alpha)}{\alpha^3 \epsilon^2} + \frac{\log(1/\delta)}{\epsilon^2}\right)$$

Then if $\epsilon \leqslant \frac{\alpha}{4B}$, with probability greater than or equal to $1 - 2\delta$ it follows that f_{T-1} is 2α-approximately multicalibrated with respect to the distribution $\mathcal{D}$.

Proof. Let $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}_{(x_i,y_i)\sim\mathcal{D}}$. Consider the model f_{T-1} output by RegressionMulticalibrate$(f, \alpha, A_{\mathcal{H}}, D, B)$, and recall that within the run of the algorithm there was also a model f_T defined in the final round. Say that model f_{T-1} is not 2α-approximately multicalibrated with respect to $\mathcal{H}_B$ and the true distribution $\mathcal{D}$.

Since the algorithm running on the sample halted, it must have been that the model in the final round improved in squared error by less than $\alpha/(2B)$ when measured with respect to the sample D:

$$\mathop{\mathbb{E}}_{(x,y)\sim D}[(f_{T-1} - y)^2] - \mathop{\mathbb{E}}_{(x,y)\sim D}[(f_T - y)^2] \leqslant (\alpha/2B).$$

Consider what happens if we run the algorithm again, but with f_{T-1} as its initial model and now with the underlying distribution as input rather than the sample of n points. Let f'_T be the model found in the first round of running this process RegressionMulticalibrate$(f_{T-1}, \alpha, A_{\mathcal{H}}, \mathcal{D}, B)$. Since f_{T-1} is not $2\alpha-$approximately multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}_B$, then by an identical argument as in the proof of Theorem 4.3, it it must be that a single round of the algorithm improves the squared error on $\mathcal{D}$ by at least α/B. Thus, $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(f_{T-1} - y)^2] - \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f'_T - y)^2] > \alpha/B$.

We know from our previous convergence bound, Theorem A.8, that with probability $1-\delta$, $|\,\mathbb{E}_{(x,y)\sim D}[(f'_T - y)^2] - \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f_T - y)^2]| < \epsilon$. So, f'_T must with high probability also improve on the sample D:

$$\begin{aligned}
\frac{\alpha}{B} &< \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1} - y)^2] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f'_T - y)^2] \\
&< \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1} - y)^2] - \mathop{\mathbb{E}}_{(x,y)\sim D}[(f'_T - y)^2] + \epsilon \qquad \text{(with probability } \geqslant 1 - \delta) \\
&< \mathop{\mathbb{E}}_{(x,y)\sim D}[(f_{T-1} - y)^2] - \mathop{\mathbb{E}}_{(x,y)\sim D}[(f'_T - y)^2] + 2\epsilon \qquad \text{(with probability } \geqslant 1 - 2\delta) \\
&< \frac{\alpha}{2B} + 2\epsilon,
\end{aligned}$$

where the last line comes from the fact that the error of f'_T on D cannot be less than the error of f_T on D, or else the regression oracle would have found it. Now we have a contradiction: since we have set $\epsilon \leqslant \frac{\alpha}{4B}$,

$$\begin{aligned}
\frac{\alpha}{B} &< \frac{\alpha}{2B} + 2\frac{\alpha}{4B} \\
&= \frac{\alpha}{B}.
\end{aligned}$$

So, it must follow that f_{T-1} is, with probability $1 - 2\delta$, $2\alpha-$approximately multicalibrated. □