Notation. Throughout the Supplementary Materials, we use $\|x\|^2 = \mathbf{E}_{i\sim\mathcal{D}}[x_i^2]$. We use $\langle x, y\rangle = \mathbf{E}_{i\sim\mathcal{D}}[x_i \cdot y_i]$.

A. Learning Multicalibrated Predictors

Here, we will work with a more technical variant of multicalibration, which implies $(\mathcal{C}, \alpha)$-multicalibration. In particular, this definition will allow us to work with an explicit discretization of the values $v \in [0, 1]$. Throughout, for a predictor f, we refer to the "categories" $S_v(f) = \{i : f_i \in \lambda(v)\} \cap S$ for all $S \in \mathcal{C}$ and $v \in \Lambda[0, 1]$.

Definition ($(\mathcal{C}, \alpha, \lambda)$-multicalibration)**.** *Let $\mathcal{C} \subseteq 2^{\mathcal{X}}$ be a collection of subsets of $\mathcal{X}$. For any $\alpha, \lambda > 0$, a predictor f is* $(\mathcal{C}, \alpha, \lambda)$-multicalibrated *if for all $S \in \mathcal{C}$, $v \in \Lambda[0, 1]$, and all categories $S_v(f)$ such that $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S_v(f)] \geq \alpha\lambda \cdot \mathbf{Pr}_{i\sim\mathcal{D}}[i \in S]$, we have*

$$\mathop{\mathbf{E}}_{i\in S_v(f)}[f_i - p_i^*] \leq \alpha.$$

We claim that if learn a predictor that satisfies $(\mathcal{C}, \alpha, \lambda)$-multicalibration, we can easily transform this predictor into one that satisfies our earlier notion of $(\mathcal{C}, \alpha)$-multicalibration. In particular, let f^λ be the λ-discretization of a predictor f if for all $i \in S_v(f)$, $f_i^\lambda = \mathbf{E}_{i\sim S_v(f)}[f_i]$.

Lemma 1. *For $\alpha, \lambda > 0$, suppose $\mathcal{C} \subseteq 2^{\mathcal{X}}$ is a collection of subsets of $\mathcal{X}$. If f is (α, λ)-multicalibrated on $\mathcal{C}$, then f^λ is $(\alpha + \lambda)$-multicalibrated on $\mathcal{C}$.*

Proof. Consider the categories $S_v(f)$ where $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S_v(f)] < \alpha\lambda \cdot \mathbf{Pr}_{i\sim\mathcal{D}}[i \in S]$. By the λ-discretization, there are at most $1/\lambda$ such categories, so the cardinality of their union is at most $(1/\lambda)\alpha\lambda \cdot \mathbf{Pr}_{i\sim\mathcal{D}}[i \in S] = \alpha \mathbf{Pr}_{i\sim\mathcal{D}}[i \in S]$. Thus, for each $S \in \mathcal{C}$, there is a subset $S' \subseteq S$ with $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S'] \geq (1 - \alpha) \cdot \mathbf{Pr}_{i\sim\mathcal{D}}[i \in S]$ where for all $v \in \Lambda[0, 1]$,

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

Further, λ-discretization will "move" the values of f_i by at most λ, so overall, f^λ will be $(\alpha + \lambda)$-calibrated. □

In this section, we prove Theorem 2 by analyzing Algorithm 1, showing how to implement the algorithm from a small sample. Theorem 1 follows as a corollary of Theorem 2.

A.1. Proof of Theorem 2

Algorithm 1 runs through each possible category S_v and if S_v is large enough, queries the oracle. The algorithm continues searching for uncalibrated categories until f's guesses on all sufficiently large categories receive ✓. By the definition of the guess-and-check oracle, if the query returns ✓ for some category S_v, then $\bar{v}$ is at most $4\omega = \alpha$ far from the true value $\mathbf{E}_{i\sim S_v}[p_i^*]$. Thus, by the stopping condition of the loop, the predictor where all $i \in \lambda(v)$ receive $f_i = \bar{v}$ will be α-calibrated on every large category. Finally, the algorithm updates f to be λ-discretized, so by Lemma 1, f will be $(\mathcal{C}, \alpha + \lambda)$-multicalibrated. Further, the number of updates necessary to terminate is bounded.

Lemma 2. *Suppose $\alpha, \lambda > 0$ and $\mathcal{C} \subseteq 2^{\mathcal{X}}$ where for all $S \in \mathcal{C}$, $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S] \geq \gamma$. Algorithm 1 returns f after receiving at most $O(1/\alpha^3\lambda\gamma)$ guess-and-check responses where $r \in [0, 1]$ and at most $O(|C|/\alpha^4\lambda\gamma)$ responses $r =$ ✓.*

Proof. For some non-✓ response on $S_v = \{i : f_i \in \lambda(v)\} \cap S$, by the properties of the guess-and-check oracle, we can lower bound the progress of each update in terms of the squared error. Let $\delta_v = r - \bar{v}$ be the magnitude of the updates for $i \in S_v$.

$$\begin{aligned}
&\|p^* - f\|^2 - \|p^* - f'\|^2 \\
&= \mathop{\mathbf{Pr}}_{i\sim\mathcal{D}}[i \in S_v] \cdot \mathop{\mathbf{E}}_{i\sim S_v}[(p_i^* - f_i)^2 - (p_i^* - \pi_{[0,1]}(f_i + \delta_v))^2] \\
&\geq \beta \cdot \mathop{\mathbf{E}}_{i\sim S_v}[(p_i^* - f_i)^2 - (p_i^* - (f_i + \delta_v))^2] \\
&= \beta \cdot \mathop{\mathbf{E}}_{i\sim S_v}[2 \cdot (p_i^* - f_i) \cdot \delta_v - \delta_v^2] \\
&= \beta \cdot \left(2\delta_v \cdot \mathop{\mathbf{E}}_{i\sim S_v}[p_i^* - f_i] - \delta_v^2\right)
\end{aligned}$$

Letting $\nu_v = \mathbf{E}_{i\sim S_v}[p_i^* - f_i]$ and $\delta_v = \nu - \tau$ for some $|\tau| \leq \omega$, we can rearrange as follows.

$$2 \cdot (\nu - \tau) \cdot \nu - (\nu - \tau)^2 = \nu^2 - \tau^2$$

Noting that by the guarantees of the guess-and-check oracle, $\nu \geq 2\omega$ and taking $\omega \leq \alpha/4$, we see that the potential progress is at least $\beta \cdot \Omega(\alpha^2)$ where $\beta = \mathbf{Pr}_{i\sim\mathcal{D}}[i \in S_v] \geq \alpha\lambda\gamma$.

As $\|p^*\|^2 \leq 1$, we make at most $O(1/\alpha^3\lambda\gamma)$ updates upper bounding the number of non-✓ responses. By working with a λ-discretization, there are at most $|C|/\lambda$ categories to consider in every phase, so we receive at most $O(|C|/\alpha^3\lambda^2\gamma)$ ✓ responses. □

Thus, we conclude the following theorem.

Theorem. *For $\alpha, \lambda > 0$ and $\mathcal{C} \subseteq 2^{\mathcal{X}}$ where for all $S \in \mathcal{C}$, $|S| \geq \gamma N$, there is a statistical query algorithm that learns a (α, λ)-multicalibrated predictor with respect to $\mathcal{C}$ in $O(|\mathcal{C}|/\alpha^3\lambda^2\gamma)$ queries.*

Again, note that our output is, in fact, $(\mathcal{C}, \alpha + \lambda)$-multicalibrated, so taking $\lambda = \alpha$, we obtain a $(\mathcal{C}, 2\alpha)$-multicalibrated predictor in $O(|\mathcal{C}|/\alpha^5\gamma)$ queries.

A.2. Answering guess-and-check queries from a random sample

Next, we argue that we can implement a guess-and-check oracle from a set of random samples in a manner that guarantees good generalization. This, in turn, allows us to translate our statistical query algorithm for learning a $(\mathcal{C}, \alpha, \lambda)$-multicalibrated predictor into an algorithm that learns from samples. Naively, we could resample for every update the algorithm makes to the predictor. Suppose that $\mathcal{C}$ is such that for all $S \in \mathcal{C}$, $|S| \geq \gamma N$; let $\beta = \alpha\lambda\gamma$. We could take $n = \tilde{O}(\log(|\mathcal{C}|)/\alpha^2\beta^2)$ samples per update to guarantee generalization, resulting in an overall sample complexity of $\tilde{O}(\log(|\mathcal{C}|)/\alpha^4\beta^4)$. We show how to improve upon this approach further. In particular, we show that there is a differentially private algorithm that can answer the entire sequence of guess-and-check queries accurately. Appealing to known connections between differential privacy and adaptive data analysis, paired with an additional observation that our notion of approximate calibration only requires relative additive error, we can guarantee that our algorithm generalizes given a set of $\tilde{O}(\log(|\mathcal{C}|)/\alpha^{5/2}\beta^{3/2})$ random samples.

Algorithm 1 only interacts with the sample through the guess-and-check oracle. Thus, to give a differentially private implementation of the algorithm, it suffices to give a differentially private implementation of the guess-and-check oracle (Dwork & Roth, 2014).

Consider the sequence of queries that Algorithm 1 makes to the guess-and-check oracle. We say the sequence $\langle (S_1, v_1, \omega_1), \ldots, (S_k, v_k, \omega_k) \rangle$ is a (k, m)-sequence of guess-and-check queries if, over the course of the k queries, the response to at most m of the queries is some $r \in [0, 1]$, and the responses to the remaining queries are all ✓. We will assume that we know a lower bound on the minimum absolute error $\beta = \min_{j \in [k]} \mathbf{Pr}_{i \sim \mathcal{D}}[i \in S_j] \cdot \omega_j$ over all of the queries. We say that some algorithm $\mathcal{A}$ responds to a guess-and-check query (S, v, ω) according to a random sample X if its response satisfies the guess-and-check properties with $\mathbf{E}_{i \sim S}[p_i^*]$ its empirical estimate on X,

$$\hat{p}_S(X) = \frac{|S|}{|S \cap X|} \sum_{i \in S \cap X} o_i.$$

Responding to such a sequence in a differentially private manner can be achieved using techniques from the private multiplicative weights mechanism.

Lemma 3 ((Hardt & Rothblum, 2010)). *Suppose $\varepsilon, \delta, \omega, \xi > 0$ and suppose $X \sim (\mathcal{X} \times \{0, 1\})^n$ is a set of n random samples. Then there exists an (ε, δ)-differentially private algorithm $\mathcal{A}$ that responds to any (k, m)-sequence of guess-and-check queries with minimum absolute error β according to X provided*

$$n = \Omega\left(\sqrt{\frac{\log(k/\xi) \cdot m \cdot \log(1/\delta)}{\varepsilon \cdot \beta^2}}\right)$$

with probability at least $1 - \xi$ over the randomness of $\mathcal{A}$.

Using this differentially private algorithm, we can apply generalization bounds based on privacy developed in (Dwork et al., 2015a;b;c; Bassily et al., 2016) to show that, with a modest increase in sample complexity, we can respond to all k guess-and-check queries.

Theorem. *Let $s_k = \langle (S_1, v_1, \omega), \ldots, (S_k, v_k, \omega) \rangle$ be a (k, m)-sequence of guess-and-check queries such that for all $j \in [k]$, $\mathbf{Pr}_{i \sim \mathcal{D}}[i \in S_j] \geq \beta$. Then there is an algorithm $\mathcal{A}$ that, given n random samples $X \sim (\mathcal{X} \times \{0, 1\})^n$, responds to s_k such that for all $j \in [k]$, the response $\mathcal{A}(S_j, v_j, \hat{\omega}_j; X)$ satisfies the guess-and-check properties with window $\omega = \alpha/4$ provided*

$$n = \Omega\left(\frac{\log(|\mathcal{C}|/\alpha\beta\xi)}{\alpha^{5/2} \cdot \beta^{3/2}}\right)$$

with probability at least $1 - \xi$ over the randomness of $\mathcal{A}$ and the draw of X.

This theorem implies that, asymptotically, we can answer the k adaptively chosen guess-and-check queries with only a $\sqrt{1/\alpha\beta}$ factor increase in the sample complexity compared to if we knew the queries in advance. The theorem follows from tailoring the proof of the main "transfer" theorem of (Bassily et al., 2016) (Theorem 3.4) specifically to the requirements of our guess-and-check oracle and applying the differentially private mechanism described in Lemma 3. Combining these theorems and Algorithm 1 and the fact that $\beta = \alpha\lambda\gamma$, we obtain an algorithm for learning α-multicalibrated predictors from random samples.

Theorem. *Suppose $\alpha, \lambda, \gamma, \xi > 0$, and $\mathcal{C} \subseteq 2^{\mathcal{X}}$ where for all $S \in \mathcal{C}$, $\mathbf{Pr}_{i \sim \mathcal{D}}[i \in S] \geq \gamma$. Then there is an algorithm that learns an $(\mathcal{C}, \alpha, \lambda)$-multicalibrated predictor with probability at least $1 - \xi$ from $n = O\left(\frac{\log(|\mathcal{C}|/\alpha\lambda\gamma\xi)}{\alpha^4 \cdot \lambda^{3/2} \cdot \gamma^{3/2}}\right)$ samples.*

A.3. Runtime analysis of Algorithm 1

Here, we present a high-level runtime analysis of Algorithm 1 for learning an $(\mathcal{C}, \alpha, \lambda)$-calibrated predictor on $\mathcal{C}$. In Lemma 2, we claim an upper bound of $O(|\mathcal{C}|/\alpha^3\lambda^2\gamma)$ on the number of guess-and-check queries needed before Algorithm 1 converges. Here, we formally argue that each of these queries can be implemented in the random sample model without much overhead, which upper-bounds the

running time of the algorithm overall. This upper bound is not immediate from our earlier analysis, as the sets and our predictor are represented implicitly as circuits.

Claim. *Algorithm 1 runs in time* $O(|\mathcal{C}| \cdot t \cdot \mathrm{poly}(1/\alpha, 1/\lambda, 1/\gamma))$*, where* t *is an upper bound on the time it takes to evaluate set membership for* $S \in \mathcal{C}$*.*

Proof. As before, let $\beta = \alpha\lambda\gamma$. First, for each $S \in \mathcal{C}$, we need to evaluate $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S_v]$ for $S_v = \{i : f_i \in \lambda(v)\} \cap S$ for each of the $O(1/\lambda)$ values $v \in \Lambda[0, 1]$. We do this by sampling $i \sim \mathcal{D}$ and evaluating whether $i \in S$, and if so, checking the current value of f_i. Each of the membership queries takes at most t time and each evaluation of f_i takes at most $O(t/\alpha^2\beta)$ time by the same argument as our upper bound on the circuit size. After $\tilde{O}(1/\lambda\beta^2)$ samples, we will be able to detect with constant probability which of the S_v have density $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S_v] \geq \beta$. Further, if $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S_v]$ is large, we can estimate $\bar{v}$ by evaluating the current predictor on samples from S_v, by rejection sampling. Similarly, to answer the guess-and-check queries, we will estimate the true empirical estimate of the query based on samples from S_v and respond based on a noisy comparison between the $\bar{v}$ and the estimate of $\sum_{i\in S_v} o_i$. These estimates can all be computed in $\mathrm{poly}(1/\alpha, 1/\beta)$. Then, each update to the predictor can be implemented in time proportional to the bit complexity of the arithmetic computations, which is upper bounded by t. Repeating this process for each $S \in \mathcal{C}$ gives the upper bound of $O(|\mathcal{C}| \cdot t \cdot \mathrm{poly}(1/\alpha, 1/\lambda, 1/\gamma))$. Finally, applying the upper bound on the number of guess-and-check queries from Lemma 2, the claim follows. □

A.4. The circuit complexity of multicalibrated predictors

An interesting corollary of our algorithm is a theorem about the complexity of representing a multicalibrated predictor. Indeed, from the definition of multicalibration alone, it is not immediately clear that there should be succinct descriptions of multicalibrated predictors; after all, $\mathcal{C}$ could contain many sets. We argue that the cardinality of $\mathcal{C}$ is not the operative parameter in determining the circuit complexity of a predictor f that is multicalibrated on $\mathcal{C}$; instead it is the circuit complexity necessary to describe sets $S \in \mathcal{C}$, as well as the cardinality of the subsets in $\mathcal{C}$, and the degree of approximation.

Leveraging Lemma 2, we can see that Algorithm 1 actually gives us a way to build up a circuit that computes the mapping from individuals to the probabilities of our learned multicalibrated predictor f. Suppose that for all sets $S \in \mathcal{C}$, set membership can be determined by a circuit family of bounded complexity; that is, for all $S \in \mathcal{C}$, there is some c_S with size at most s, such that $c_S(i) = 1$ if and only if $i \in S$. Then we can use this family of circuits to build a circuit that implements f. We assume that we maintain real-valued numbers up to $b \geq \log(1/\alpha)$ bits of precision.

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

Proof. We describe how to construct a circuit c_f that, on input i, will output the prediction f_i according to the predictor learned by our algorithm. We initialize c_f to be the constant function $c_f(i) = 1/2$ for all $i \in \mathcal{X}$. Throughout, we will update c_f based on the current outputs of c_f.

Consider an iteration of Algorithm 1 where for some S described by $c_S \in \mathcal{C}$, we update f based on a category $S_v = S \cap \{i : f_i \in \lambda(v)\}$. This occurs when the guess-and-check query returns some $r = \tilde{q}(S_v, \bar{v}, \omega) \in [0, 1]$. Our goal is to implement the update to f (i.e. update c_f), such that for all $i \in S_v$, the new value $f_i = r$ and all other values are unchanged.

We achieve this update by testing membership $i \in S$ and separately testing if the current value $c_f(i) = v$; if both tests pass, then we update the value output by $c_f(i)$ to be r. Specifically, we include a copy of c_S and hard-code v and $\delta_v = r - \bar{v}$ into the circuit; if $c_S(i) = 1$ and the current value of $c_f(i)$ is in $\lambda(v)$, then we update $c_f(i)$ to add the hardcoded δ_v to its current estimate of f_i; if either test fails, then $c_f(i)$ remains unchanged. This logic can be implemented with addition and subtraction circuits to a precision of λ with boolean circuits of size $O(b)$. We string these update circuits together, one for each iteration. Learning an $(\alpha/2, \alpha/2)$-multicalibrated predictor with Algoirthm 1 only requires $O(\alpha^4\gamma)$ updates. By this upper bound, we obtain an $O(\alpha^4\gamma)$ upper bound on the resulting circuit size. □

B. Multicalibration and Weak Agnostic Learning

B.1. Multicalibration from weak agnostic learning

In this section, we show how we can use a weak agnostic learner to solve the search problem that arises at each iteration of Algorithm 1: namely, to find an update that will make progress towards mulitcalibration. Formally, we show the following theorem.

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

That is, if there is an algorithm for learning the concept class $\mathcal{C}$ over the hypothesis class of real-valued functions $\mathcal{H} = \{h : \mathcal{X} \to [-1, 1]\}$ on the distribution of individuals in polynomial time in $\log(|\mathcal{C}|), 1/\rho$, and $1/\tau$, then there is an algorithm for learning an α-multicalibrated predictor on the large sets in $\mathcal{C}$ that runs in time polynomial in $\log(|\mathcal{C}|), 1/\alpha, 1/\lambda, 1/\gamma$. For clarity of presentation in the reduction, we make no attempts to optimize the sample complexity or running time. Indeed, the exact sample complexity and running time will largely depend on how strong the weak learning guarantee is for the specific class $\mathcal{C}$.

We prove the theorem by using the weak learner for $\mathcal{C}$ to learn a (α, λ)-multicalibrated predictor. Recall Algorithm 1: we maintain a predictor f and iteratively look for a set $S \in \mathcal{C}$ where f violates the calibration constraint on $S_v = \{i : f_i \in \lambda(v)\} \cap S$ for some value v. In fact, the proof of Lemma 2 reveals that we are not restricted to updates on S_v for $S \in \mathcal{C}$. As long as there is some uncalibrated category S_v, we can find an update that makes nontrivial progress in ℓ_2^2 distance from p^* – even if this update is not on any $S \in \mathcal{C}$ – then we can bound the number of iterations it will take before there are no more uncalibrated categories. We show that a weak agnostic learner allows us to find such an update.

Proof. Throughout the proof, let $\beta = \alpha\lambda\gamma$, $\rho = \alpha\beta/2$, and $\tau = \rho^d$ for some constant $d \geq 1$. Let f be a predictor initialized to be the constant function $f_i = 1/2$ for all $i \in \mathcal{X}$.

Consider the search problem that arises during Algorithm 1 immediately after updating the predictor f. Let $\mathcal{X}_v = \{i : f_i \in \lambda(v)\}$ be the set of individuals in the λ-interval surrounding v. Our goal is to determine if there is some $v \in \Lambda[0, 1]$ and $S \in \mathcal{C}$ such that $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S_v] \geq \beta$, where

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

We reduce this search problem to the problem of weak agnostic learning over $\mathcal{C}$ on the distribution $\mathcal{D}$. For any $v \in \Lambda[0, 1]$, if $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in X_v] < \beta$, then clearly there is no uncalibrated category S_v with $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S_v] \geq \beta$; for each $v \in \Lambda[0, 1]$, we will test if $\mathcal{X}_v$ is large enough by taking $O(\log(1/\beta\xi)/\beta)$ random draws from $\mathcal{X}$.

We assume that f is overall $\tau/4$-calibrated on $\mathcal{X}$; if f were not, we can update f_i for all $i \in \mathcal{X}_v$ for the violated values to make $\Omega(\tau^2)$ progress as the analysis of Algorithm 1.

For each $v \in \Lambda[0, 1]$, we consider the following learning problem. For $i \in \mathcal{X}_v$, let $\Delta_i = \frac{f_i - o_i}{2}$. For $i \in \mathcal{X} \setminus \mathcal{X}_v$, let $\Delta_i = 0$. We claim that if there is some S_v satisfying (1), then for $i \sim \mathcal{D}_{\mathcal{X}}$, the labeled samples of either (i, Δ_i) or $(i, -\Delta_i)$ satisfy the weak learning promise for $\rho = \alpha\beta/4$. Note that we assume the learner takes enough samples to guarantee that the empirical estimates using outcomes concentrate around their underlying expectations; for the sake of clarity of presentation, we make no attempt to optimize the sample complexity in this section.

Claim. *Let $c_S : \mathcal{X} \to \{-1, 1\}$ be the boolean function associated with some $S \in \mathcal{C}$. For $v \in \Lambda[0, 1]$, if $S_v = \{i : f_i \in \lambda(v)\} \cap S$ satisfies $\mathbf{E}_{i\sim S_v}[f_i - p_i^*] \geq \alpha$, then*

$$\langle c_S, \Delta \rangle \geq \rho.$$

Note that the supposition of the claim is satisfied when (1) holds without the absolute value. In the case where (1) holds in the other direction, the claim will hold for $-\Delta$. The argument will be identical.

$$\begin{aligned}
\langle c_S, \Delta \rangle &= \frac{1}{2} \mathop{\mathbf{E}}_{i\sim\mathcal{D}}[(f_i - o_i) \cdot c_S(i)] \\
&= \frac{1}{2} \mathop{\mathbf{Pr}}_{i\sim\mathcal{D}}[i \in \mathcal{X}_v] \cdot \mathop{\mathbf{E}}_{i\sim\mathcal{X}_v}[(f_i - o_i) \cdot c_S(i)] \\
&= \frac{1}{2} \mathop{\mathbf{Pr}}_{i\sim\mathcal{D}}[i \in S_v] \cdot \mathop{\mathbf{E}}_{i\sim S_v}[(f_i - o_i)] \\
&\quad - \frac{1}{2} \mathop{\mathbf{Pr}}_{i\sim\mathcal{D}}[i \in \mathcal{X}_v \setminus S_v] \cdot \mathop{\mathbf{E}}_{i\sim\mathcal{X}_v\setminus S_v}[(f_i - o_i)] \\
&\geq \mathop{\mathbf{Pr}}_{i\sim\mathcal{D}}[i \in S_v] \cdot \mathop{\mathbf{E}}_{i\sim S_v}[(f_i - o_i)] - \tau/4 \qquad (2) \\
&\geq \beta\alpha/2 - \tau/4 \qquad (3) \\
&\geq \rho
\end{aligned}$$

where the inequality (2) follows from the assumption that f is $\tau/4$-calibrated on $\mathcal{X}$, (3) follows from the assumption that $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S_v] \geq \beta$ and our assumption on $\mathbf{E}_{i\sim S_v}[f_i - p_i^*]$. Noting that $\tau \leq \rho$ gives the claim.

Thus, because the (ρ, τ)-weak agnostic learning promise is satisfied, the learner will return to us some $h : \mathcal{X} \to [-1, 1]$ that is nontrivially correlated with $f - p^*$ on $\mathcal{X}_v$. In particular, if we use this h as an update step, updating $f_i \to v - \eta h_i$ (projecting onto $[0, 1]$ if necessary) for $\eta = \Omega(\tau)$, then we can guarantee that each such update will achieve $\tau^2\beta$ progress in $\|f - p^*\|^2$. The analysis follows in the same way as the analysis of Algorithm 1. □

B.2. Weak agnostic learning from multicalibration

In this section, we show the converse reduction. In particular, we will show that for a concept class $\mathcal{C}$, an efficient algorithm for obtaining an α-multicalibrated predictor with respect to $\mathcal{C}' = \{S \in \mathcal{C} : |S| \geq \gamma N\}$, gives an efficient algorithm for responding to weak agnostic learning queries on $\mathcal{C}$.

Theorem. *Let $\alpha, \gamma > 0$ and suppose $\mathcal{C} \subseteq 2^{\mathcal{X}}$ is a concept class. If there is an algorithm for learning an α-multicalibrated predictor on $\mathcal{C}' = \{S \in \mathcal{C} : \mathbf{Pr}_{i\sim\mathcal{D}}[i \in S] \geq \gamma\}$ in time $T(|C|, \alpha, \gamma)$ then we*

can implement a (ρ, τ)-weak agnostic learner for $\mathcal{C}$ in time $O(T(|C|, \alpha, \gamma) \cdot \mathrm{poly}(1/\tau))$ for any $\rho, \tau > 0$ such that $\tau \leq \min\{\rho - 2\gamma, \rho/4 - 4\alpha\}$.

Proof. Suppose we want to weak agnostic learn over $\mathcal{C}$ on sampled observations from $y \in [-1,1]^N$. We assume there is some $c_S \in \mathcal{C}$ such that $\langle c_S, y\rangle > \rho$.

There are two cases to handle. First, suppose the support of c_S is small; that is, for the corresponding $S \in \mathcal{C}$, $|S| < \gamma$. Then, the correlation between y and the the constant hypothesis $h(i) = -1$ for all $i \in \mathcal{X}$ will be at least $\rho - 2\gamma$. Thus, for $\tau < \rho - 2\gamma$, in the case when the support of c_S is small, then we can return the hypothesis -1. We can test if the constant hypothesis is sufficiently correlated with y in $\mathrm{poly}(1/\tau)\log(1/\xi)$ time by random sampling to succeed with probability at least $1 - \xi$.

Next, we will proceed assuming $|\mathbf{E}_{i\sim\mathcal{X}}[y_i]| < 2\omega$. By the same argument as above, this means $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S] \cdot \mathbf{E}_{i\sim S}[y_i] > \frac{\rho}{2} - \omega$. Suppose we learn an f that is α-multicalibrated with respect to $\mathcal{C}' = \mathcal{C} \cup \{\mathcal{X}\}$ on the labels y. This implies that there is some $\mathcal{X}' \subseteq \mathcal{X}$ such that $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in \mathcal{X}'] \geq 1 - \alpha$ and for all $v \in [-1,1]$, we have $v - \alpha \leq \mathbf{E}_{i\sim X'_v}[y_i] \leq v + \alpha$. In turn, this implies the following inequality.

$$\mathrm{sgn}(v) \cdot \mathop{\mathbf{E}}_{i\sim\mathcal{X}'_v}[y_i] \geq |v| - \alpha \tag{4}$$

Then, let $h^{(f)}$ be the hypothesis defined as $h_i^{(f)} = \mathrm{sgn}(f_i)$. Consider the inner product with y.

$$\begin{aligned}
\langle h^{(f)}, y\rangle &= \mathop{\mathbf{E}}_{i\in\mathcal{X}}[h_i^{(f)} \cdot y_i] && (5)\\
&= \sum_{v\in[-1,1]} \mathop{\mathbf{Pr}}_{i\sim\mathcal{D}}[i \in \mathcal{X}_v] \cdot \mathop{\mathbf{E}}_{i\in\mathcal{X}_v}[h_i^{(f)} \cdot y_i] && (6)\\
&\geq \sum_{v\in[-1,1]} \mathop{\mathbf{Pr}}_{i\sim\mathcal{D}}[i \in \mathcal{X}_v] \cdot \mathrm{sgn}(v) \cdot \mathop{\mathbf{E}}_{i\in\mathcal{X}'_v}[y_i] - \alpha && (7)\\
&\geq \sum_{v\in[-1,1]} \mathop{\mathbf{Pr}}_{i\sim\mathcal{D}}[i \in \mathcal{X}_v] \cdot |v| - 2\alpha && (8)\\
&\geq \sum_{v\in[-1,1]} \mathop{\mathbf{Pr}}_{i\sim\mathcal{D}}[i \in \mathcal{X}'_v] \cdot \left|\mathop{\mathbf{E}}_{i\sim S'_v}[y_i]\right| - 3\alpha && (9)\\
&\geq \mathop{\mathbf{Pr}}_{i\sim\mathcal{D}}[i \in S] \cdot \mathop{\mathbf{E}}_{i\sim S}[y_i] - 4\alpha && (10)\\
&\geq \frac{\rho}{2} - \omega - 4\alpha && (11)
\end{aligned}$$

where the first equalities follow by the definition of $h^{(f)}$; (7) follows by the choice of $\mathcal{X}'$ and α-multicalibration; (8) follows by applying (4) for each $v \in [-1,1]$; (9) follows by substituting v for the empirical average of y over S'_v invoking α-multicalibration for the appropriate choice of $S' \subseteq S$; (10) follows by the triangle inequality; and (11) follows from the assumed inequality on $\mathbf{E}_{i\sim S}[y_i]$.

Thus, $h^{(f)}$ satisfies the (ρ, τ)-weak agnostic learning guarantee for any $\tau \leq \rho/4 - 4\alpha$ by our choice of $\omega = \rho/4$.

□

C. Best-in-class Predictions

Theorem (Best-in-class prediction)**.** *Suppose $\mathcal{C} \subseteq 2^{\mathcal{X}}$ is a collection of subsets of $\mathcal{X}$ and $\mathcal{H}$ is a set of predictors. Then there is a predictor f that is α-multicalibrated on $\mathcal{C}$ such that*

$$\|f - p^*\|^2 - \|h^* - p^*\|^2 < 6\alpha,$$

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

The proof of the theorem actually reveals something stronger: if f is calibrated on the set $\mathcal{S}(\mathcal{H})$, then for every category $S_v(h) \in \mathcal{S}(\mathcal{H})$, if f is significantly different from h on this category – that is, if $\mathbf{E}_{i\in S_v(h)}[(h_i - f_i)^2]$ is large – then f actually achieves significantly improved prediction error on this category compared to h. This is stated formally in Lemma 4.

Lemma 4. *Suppose g is an arbitrary predictor and let $\mathcal{S}(g) = \{S_v(g)\}_{v\in\Lambda[0,1]}$. Suppose f is an arbitrary $(\mathcal{S}(g), \alpha)$-multicalibrated predictor. Then for $v \in \Lambda[0,1]$,*

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

Consequently,

$$\|g - p^*\|^2 - \|f - p^*\|^2 \geq \|g - f\|^2 - (4\alpha + \lambda).$$

This lemma shows that calibrating on the categories of a predictor not only prevents the squared prediction error from degrading beyond a small additive approximation, but it also guarantees that if calibrating changes the predictor significantly on any category, this change represents significant progress towards the true underlying probabilities on this category. Assuming Lemma 4, the theorem follows.

Proof. Note that if f is α-multicalibrated on $\mathcal{C}$, then f is α-multicalibrated on any $\mathcal{C}' \subseteq \mathcal{C}$. Consider enforcing calibration on the collection $\mathcal{C} \cup \mathcal{S}(\mathcal{H})$ as defined above. If f is $(\mathcal{C} \cup \mathcal{S}(\mathcal{H}), \alpha, \lambda)$-calibrated, then it is $(\{S_v(h)\}_{v\in\Lambda[0,1]}, \alpha, \lambda)$-multicalibrated for all $h \in \mathcal{H}$ and specifically for h^*. By Lemma 4, and the fact that the

squared difference is nonnegative, we obtain the following inequality:

$$\begin{aligned}\|h^* - p^*\|^2 - \|f - p^*\|^2 &\geq \|f - h^*\|^2 - (4\alpha + \lambda) \\ &\geq -(4\alpha + \lambda)\end{aligned}$$

This inequality suffices to prove the accuracy guarantee; however, to also guarantee the predictor f can be implemented by a small circuit, we have to be a bit more careful. In particular, when calibrating, we will ignore any $S_v(h)$ such that $|S_v(h)| < \lambda\alpha N$. Note that because we have λ-discretized, there are at most $1/\lambda$ categories; thus, excluding the sets $S_v(h)$ where $|S_v(h)| < \alpha\lambda N$ introduces at most an additional αN error. Taking $\lambda = \alpha$, in turn, this implies that the difference in squared prediction error can be bounded as $\|f - p^*\|^2 - \|h^* - p^*\|^2 \leq 6\alpha N$. Finally, because the sets we want to calibrate on are at least $\alpha^2\gamma N$ in cardinality, the circuit complexity bound follows by applying Lemma 2. □

We turn to proving Lemma 4. The lemma follows by expanding the difference in squared prediction errors and invoking the definition of α-calibration.

Proof of Lemma 4. Let S_{vu} represent the set of individuals i where $g_i \in \lambda(v)$ and $f_i = u$. By the assumption that f is α-calibrated on $\mathcal{S}(g)$, we know for every $S_v(g) \in \mathcal{S}(g)$, there is some subset $S'_v(g) \subseteq S_v(g)$ such that $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S'_v(g)] \geq (1-\alpha) \cdot \mathbf{Pr}_{i\sim\mathcal{D}}[i \in S_v(y)]$ for which the predictions of f are approximately correct. In particular, let $S'_{vu} = S'_v(g) \cap S_u(f)$; if f is α-calibrated with respect to $S_v(g)$, this guarantees that for all values $u \in [0,1]$, we have

$$\left|\mathop{\mathbf{E}}_{i\sim S'_{vu}}[p^*_i - u]\right| \leq \alpha. \tag{12}$$

Using this fact, and the fact that the remaining α-fraction of $S_v(g)$ can contribute at most α to the squared error over $S_v(g)$, we can express the difference in the squared errors of g and f on $S_v(g)$:

$$\begin{aligned}&\mathop{\mathbf{E}}_{i\sim S_v(g)}[(g_i - p^*_i)^2] - \mathop{\mathbf{E}}_{i\sim S_v(g)}[(f_i - p^*_i)^2] \\ &= \mathop{\mathbf{E}}_{i\sim S_v(g)}[(v - p^*_i + (g_i - v))^2] - \mathop{\mathbf{E}}_{i\sim S_v(g)}[(f_i - p^*_i)^2] \\ &\geq \mathop{\mathbf{E}}_{i\sim S_v(g)}[(v - p^*_i)^2] - \mathop{\mathbf{E}}_{i\sim S_v(g)}[(f_i - p^*_i)^2] \\ &\quad + 2\mathop{\mathbf{E}}_{i\sim S_v(g)}[(v - p^*_i)(g_i - v)] \\ &\geq \mathop{\mathbf{E}}_{i\sim S_v(g)}[(2(p^*_i - v)(f_i - v) - (f_i - v)^2)] - \lambda. \end{aligned} \tag{13}$$

where (13) follows by the observation that $(g_i - v)^2 \geq 0$ and if $g_i \in \lambda(v)$, then $|g_i - v| \leq \lambda/2$ and $|v - p^*_i|$ is trivially bounded by 1. We bound the first term as follows.

$$\begin{aligned}&\mathop{\mathbf{E}}_{i\sim S_v(g)}[(p^*_i - v)(f_i - v)] \\ &= \sum_{u\in[0,1]} \mathop{\mathbf{Pr}}_{i\sim\mathcal{S}_v}[i \in S_{vu}] \cdot \mathop{\mathbf{E}}_{i\sim S_{vu}}[(p^*_i - v)(u - v)]\end{aligned}$$

Then, for each $u \in [0,1]$,

$$\begin{aligned}&\mathop{\mathbf{E}}_{i\sim S_{vu}}[(p^*_i - v)(u - v)] \\ &= (u - v)\mathop{\mathbf{E}}_{i\sim S_{vu}}[p^*_i - v] \\ &= (u - v)\mathop{\mathbf{E}}_{i\sim S_{vu}}[u - v + p^*_i - u] \\ &= (u - v)^2 + (u - v)\mathop{\mathbf{E}}_{i\sim S_{vu}}[p^*_i - u].\end{aligned}$$

At this point, we note that $|u - v| \leq 1$. Thus, we can bound the contribution of the expectation over S_{vu} by its negative absolute value:

$$\begin{aligned}&\geq (u - v)^2 - |u - v| \cdot \left|\mathop{\mathbf{E}}_{i\sim S_{vu}}[p^*_i - u]\right| \\ &\geq (u - v)^2 - (1 - \alpha) \cdot \left|\mathop{\mathbf{E}}_{i\sim S'_{vu}}[p^*_i - u]\right| - \alpha \\ &\geq (u - v)^2 - 2\alpha\end{aligned}$$

Summing over $u \in [0,1]$,

$$\begin{aligned}&\sum_{u\in[0,1]} \mathop{\mathbf{Pr}}_{i\sim\mathcal{D}}[i \in S_{vu}] \cdot \big((u - v)^2 - 2\alpha\big) \\ &= \mathop{\mathbf{E}}_{i\sim S_v(g)}[(v - f_i)^2] - 2\alpha,\end{aligned}$$

where we bound the sums over S_{vu} by invoking α-calibration and applying (12). Plugging this bound into (13), we see that

$$\begin{aligned}&\mathop{\mathbf{E}}_{i\sim S_v(g)}\left[(g_i - p^*_i)^2 - (f_i - p^*_i)^2\right] \\ &\geq 2\left(\mathop{\mathbf{E}}_{i\sim S_v(g)}[(v - f^*_i)^2] - 2\alpha\right) - \lambda - \mathop{\mathbf{E}}_{i\sim S_v(g)}[(v - f_i)^2] \\ &= \mathop{\mathbf{E}}_{i\in S_v(g)}[(v - f^*_i)^2] - (4\alpha - \lambda).\end{aligned}$$

Summing over $v \in [0,1]$, we can conclude

$$\|g - p^*\|^2 - \|f - p^*\|^2 \geq \|f - g\|^2 - (4\alpha - \lambda)$$

showing the lemma. □

References

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

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

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

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

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

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