Low-Degree Multicalibration

Parikshit Gopalan* VMware Research
Michael P. Kim† UC Berkeley
Mihir Singhal‡ MIT
Shengjia Zhao§ Stanford University

March 2, 2022

Abstract

Introduced as a notion of algorithmic fairness, multicalibration has proved to be a powerful and versatile concept with implications far beyond its original intent. This stringent notion—that predictions be well-calibrated across a rich class of intersecting subpopulations—provides its strong guarantees at a cost: the computational and sample complexity of learning multicalibrated predictors are high, and grow exponentially with the number of class labels. In contrast, the relaxed notion of multiaccuracy can be achieved more efficiently, yet many of the most desirable properties of multicalibration cannot be guaranteed assuming multiaccuracy alone. This tension raises a key question: *Can we learn predictors with multicalibration-style guarantees at a cost commensurate with multiaccuracy?*

In this work, we define and initiate the study of *Low-Degree Multicalibration*. Low-Degree Multicalibration defines a hierarchy of increasingly-powerful multi-group fairness notions that spans multiaccuracy and the original formulation of multicalibration at the extremes. Our main technical contribution demonstrates that key properties of multicalibration, related to fairness and accuracy, actually manifest as low-degree properties. Importantly, we show that low-degree multicalibration can be significantly more efficient than full multicalibration. In the multi-class setting, the sample complexity to achieve low-degree multicalibration improves exponentially (in the number of classes) over full multicalibration. Our work presents compelling evidence that low-degree multicalibration represents a sweet spot, pairing computational and sample efficiency with strong fairness and accuracy guarantees.

*pgopalan@vmaware.com
†mpkim@berkeley.edu Supported by the Miller Institute for Basic Research in Science and by the Simons Collaboration on the Theory of Algorithmic Fairness
‡mihirs@mit.edu This work completed during an internship at VMware Research.
§sjzhao@stanford.edu

1 Introduction

Machine learning models are increasingly used to aid decision-making in professional, personal, medical, and legal spheres. This ubiquity has brought increased concern about whether these models make fair predictions, especially on underrepresented *subpopulations*. Typically in supervised learning, models are trained to minimize the expected loss over the *entire population*, and can be less accurate on such subpopulations. A particular fairness concern is that predictive models may commit *algorithmic stereotyping*, where every member of a subpopulation receives similar predictions, despite internal diversity within the subpopulation. Such shortcomings of the standard supervised learning framework have been documented extensively within the research community and in the popular press. In response, a growing area of research investigates *multi-group* fairness notions, that require predictions to perform well not simply overall, but even when restricting attention to structured subgroups [HKRR18, KNRW18, KRR18, KGZ19, SCM20, BL20, DKR+21, JLP+21, RY21, TH21, GJN+21, DKR+22].

Central within the study of multi-group fairness is the notion of *multicalibration*. Calibration is a classic notion from the forecasting literature [Daw82] that was introduced to the literature on fairness in prediction tasks by [KMR17]. For a binary prediction task, calibration requires that amongst the individuals which receive prediction $f(x) = v$, the true expectation is v. Defined by [HKRR18], multicalibration strengthens the classic notion, requiring a predictor to be calibrated simultaneously across a large, possibly-overlapping collection of subpopulations. We model the collection by a hypothesis class $\mathcal{C} \subseteq \{c : \mathcal{X} \to \{0, 1\}\}$ and say that a predictor f is mutlicalibrated over $\mathcal{C}$ if, for all predicted values $v \in [0, 1]$, and for all $c \in \mathcal{C}$

$$\mathbf{E}\left[\, c(\mathbf{x}) \cdot (\mathbf{y} - v) \mid f(\mathbf{x}) = v \,\right] \approx 0.$$

Intuitively, an expressive class $\mathcal{C}$ will contain subpopulations that go beyond "protected groups" (typically defined marginally in terms of a single attribute). In this way, while calibration provides only marginal guarantees, multicalibration requires predictors to capture the variation within subpopulations, to give confident (but not overconfident) predictions, thus providing strong protections against algorithmic stereotyping.

Beyond its origins as a notion of fairness, multicalibration has proved surprisingly versatile and powerful in diverse contexts. The notion of multicalibration makes no mention of loss minimization, yet the work of [GKR+22] shows that multicalibrated predictors implicitly obtain optimal loss, simultaneously for all Lipschitz, convex losses. Specifically, given any (fixed) $\mathcal{C}$-multicalibrated predictor f, for every such loss ℓ, the predictor f guarantees loss competitive with the hypothesis $c_\ell \in \mathcal{C}$ chosen to the minimize the loss over $\mathcal{C}$ (in fact, over linear combinations of hypotheses in $\mathcal{C}$). This property leads to an *omniprediction* guarantee, where one can learn a single predictor f, without knowledge of the choice of loss at the time of learning.

In another direction, [DKR+21] demonstrate that the multicalibration framework is equivalent to a certain pseudorandomness condition, which they call *outcome indistinguishability*. Intuitively, a predictor f is outcome indistinguishable to a family of distinguisher algorithms $\mathcal{A}$ if given a sample $(\mathbf{x}, \mathbf{y})$ no distinguisher $A \in \mathcal{A}$ can tell whether $\mathbf{y}$ was sampled from Nature's true conditional distribution of $\mathbf{y}|\mathbf{x}$, or according to the predicted probability $f(\mathbf{x})$. This indistinguishability perspective has seen application in characterizing the feasibility of multi-group strengthenings of agnostic PAC learnability [RY21]. Multicalibration has also been extended to diverse settings of un-

certainty quantification for real-valued outcomes [JLP+21], importance weights [GRSW21], online prediction [GJN+21], and adaptation under covariate shift [KKG+22].

It is perhaps not a surprise that the power of multicalibration comes at a cost—both in terms of samples and computation. Computationally, [HKRR18] show that a *weak agnostic learner* for the class $\mathcal{C}$ is necessary and sufficient to learn $\mathcal{C}$-multicalibrated predictors. Using the weak learner, they design a boosting-style algorithm that produces a multicalibrated predictor by combining hypotheses $c \in \mathcal{C}$ using nontrivial Boolean logic.[1] The number of calls to the weak learner and the sample complexity are governed by an approximation parameter α which quantifies the *deviation* from perfect multicalibration. The sample complexity depends inverse polynomially in the parameter α, with fairly large exponent. This dependence becomes even worse when we generalize multicalibration to the multi-class setting with $l > 2$ class labels, where a single prediction is a vector of probabilities in l dimensions. Reasoning about expectations conditioned on the predictions leads to complexity that scales as $1/\alpha^{\Omega(\ell)}$, exponential in the number of class labels. Consequently, achieving multicalibration is practically infeasible with more than a few classes.

In the work defining multicalibration, [HKRR18] also introduced a weaker fairness notion known as *multiaccuracy*, which only requires that $f(\mathbf{x})$ and $\mathbf{y}$ have similar expectations over subpopulations in $\mathcal{C}$, without conditioning on the predicted values.

$$\mathbf{E}\left[\ c(\mathbf{x}) \cdot (\mathbf{y} - f(\mathbf{x}))\ \right] \approx 0$$

The simpler notion of multiaccuracy is comparatively easier to obtain, quantitatively and qualitatively. One can view $\mathcal{C}$-multiaccuracy as a first-order optimality condition on $\lambda \in \mathbb{R}^{|\mathcal{C}|}$ for predictors of the form $f(x) = \sum_{c \in \mathcal{C}} \lambda_c \cdot c(x)$. Thus, while multiaccuracy also requires a weak learning oracle, one can learn a multiaccurate predictor simply by minimizing the squared or logistic loss over linear combinations of $c \in \mathcal{C}$ (without specialized boolean logic), using techniques like coordinate ascent or gradient boosting. Further, standard concentration inequalities demonstrate that the sample complexity to obtain multiaccuracy scales as α^{-2}, even as the number of classes l grows.

In exchange for its efficiency, multiaccuracy is known to provide considerably weaker guarantees than multicalibration. Many of the most desirable fairness and accuracy properties that can be derived from multicalibration cannot be derived from multiaccuracy alone, including its guarantees for loss minimization [HKRR18, RY21, GKR+22], the fairness properties of the ranking induced by predictions [DKR+19], and multi-group confidence intervals [JLP+21]. This tension—between guarantees and efficiency—brings us to the motivating question behind our work:

Are there notions that retain important properties of multicalibration,
but are computable much more efficiently (comparable to multiaccuracy)?

2 Overview of Contributions

We introduce a hierarchy of multicalibration notions that enable a tradeoff between the strength of multi-group guarantees and the complexity required to learn the predictor. At the extremes,

[1] In fact, multicalibration has been shown to be tightly connected to the boosting-by-branching-programs framework of [MM02, KMV08]. See [GKR+22] for a discussion.

our hierarchy recovers the existing notions of multiaccuracy and multicalibration, but our interest is in the intermediate notions. We establish guarantees about these notions, and show that many desirable properties of multicalibration kick in at low levels of the hierarchy. In doing so, we gain new insights into the power of (full) multicalibration. We complement these with algorithmic results showing that computing predictors in the low levels of the hierarchy can be significantly more efficient than the original formulation of multicalibration. Our three main contributions can be summarized as follows:

(1) Our primary conceptual contribution is the definition of *low-degree multicalibration* and its associated hierarchy. For $k \in \mathbb{N}$, the kth level of the hierarchy defines a notion of multicalibration that constrains the first k moments of the predictor, conditioned on subpopulations in $\mathcal{C}$. The lowest level of the hierarchy corresponds to multiaccuracy. As we go higher, the multicalibration constraints become more stringent. In the limit, we approach a notion we call *smooth* multicalibration; a relaxation of the original formulation of [HKRR18], which we refer to as *full* multicalibration.

(2) With the hierarchy in place, we study the fairness and accuracy properties obtainable via low-degree multicalibration. Our main contribution is to provide a rich toolbox for reasoning about multicalibrated predictions $f(x)$, by comparing to the moments of the Bayes optimal predictions $f^*(x) = \mathbf{E}[\, \mathbf{y} \mid \mathbf{x} = x \,]$. Our key technical result establishes novel *moment sandwiching bounds* for multicalibrated predictors. For instance, in lieu of k^{th} moment matching (which would give $E[f(\mathbf{x})^k] \approx \mathbf{E}[f^*(\mathbf{x})^k]$, but is impossible to achieve for $k > 1$), we show that degree-k multicalibration implies that $\mathbf{E}[f(\mathbf{x})^k] \leq \mathbf{E}[f^*(\mathbf{x})^k]$, even when conditioned on subpopulations defined by $c \in \mathcal{C}$. Using these tools, we can relate the confusion rates (generalized false error rates) of any degree-2 multicalibrated f to those of the optimal predictor. Our results reveal wide gaps even between multiaccuracy and degree-2 multicalibration; predictors satisfying the latter cannot exhibit overconfidence over subpopulations in $\mathcal{C}$, unlike multiaccurate predictors.

(3) Finally, we show that low-degree multicalibration can provide significant savings over full multicalibration. In particular, we show that for l-class prediction tasks, the sample complexity to obtain low-degree multicalibration is polynomial in l, whereas obtaining the same guarantees using full multicalibration requires sample complexity exponential in l. Even in the case of binary prediction, low-degree multicalibration obtains improvements over full multicalibration, by polynomial factors in the approximation parameter $1/\alpha$. These bounds suggest that the low-degree notions may be practically-realizable, providing strong guarantees in settings where the existing notions of multicalibration cannot be achieved.

In all, we develop a more refined picture of multiaccuracy, multicalibration, and the guarantees that lie in between. Our results establish that—in addition to the calibration class $\mathcal{C}$ and approximation parameter α—the degree of multicalibration is a meaningful "knob" that can be tuned to the needs and constraints of a given setting. Low-degree multicalibration provides a new perspective and set of techniques that we anticipate will be useful to practitioners and theoreticians alike.

Organization of manuscript. The remainder of the manuscript is structured as follows.

First, we continue this section with a high-level overview of our contributions, focusing on the binary prediction setting. In Section 2.1, we begin with an intuitive explanation of how one might

discover low-degree multicalibration. Then, in Section 2.2, we present the definitions of the new variants of multicalibration and their relation to one another; in Section 2.3, we present the novel moment sandwiching bounds for multicalibrated predictors; and in Section 2.4, we present the bounds on the complexity of achieving each variant of multicalibration. We conclude the overview with Section 2.5, where we provide further discussion of low-degree multicalibration and how it relates to other works on multi-group fairness, agnostic learning, and calibration.

The technical portions of the manuscript follow the structure of the overview. In Section 3, we give formal definitions of the notions within the low-degree multicalibration hierarchy, handling the multi-class setting. Then, we establish Proposition 1 and other the relationships and robustness properties of notions in the hierarchy. In Section 4, we establish our main technical result, Theorem 2, along with other key fairness properties of low-degree multicalibration. In Section 5, we describe Algorithm 1 and analyze the sample complexity as in Theorem 3. In Section 6, we highlight a proof-of-concept experimental evaluation of low-degree multicalibration.

2.1 Towards Multi-Group Moment Matching

In this motivating vignette, we focus on the setting of binary prediction: we are given samples from a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, for domain $\mathcal{X}$ and label space $\mathcal{Y} = \{0, 1\}$. We use $f : \mathcal{X} \to [0, 1]$ to denote our hypothesis, and use $f^* : \mathcal{X} \to [0, 1]$ to denote the Bayes optimal predictior, defined as $f^*(x) = \mathbf{E}[\, \mathbf{y} \mid \mathbf{x} = x \,]$, the true expected outcome over $\mathcal{D}$ of $\mathbf{y}$ given $\mathbf{x} = x$.[2] Ideally, our learned hypothesis f should be a close approximation to f^*. Every variant of multicalibration is parameterized by a hypothesis class $\mathcal{C} \subseteq \{c : \mathcal{X} \to \{0, 1\}\}$. We think of $\mathcal{C}$ as an expressive but bounded class, where the functions $c \in \mathcal{C}$ have a simple representation; for instance, we may assume that the VC-dimension of $\mathcal{C}$ is finite.

To begin, we describe the intuition for multiaccuracy and how one might strengthen it, without appealing to full multicalibration. Towards the goal of approximating f^*, multiaccuracy imposes a first-order condition of matching expectations over $c \in \mathcal{C}$.

$$\mathbf{E}[\, c(\mathbf{x}) \cdot (\mathbf{y} - f(\mathbf{x})) \,] \approx 0 \implies \mathbf{E}[\, c(\mathbf{x}) \cdot f^*(x) \,] \approx \mathbf{E}[\, c(\mathbf{x}) \cdot f(\mathbf{x}) \,]$$

Naturally, the next step might be to try and match second moments, and require that $\mathbf{E}[c(\mathbf{x}) \cdot f(\mathbf{x})^2] \approx \mathbf{E}[c(\mathbf{x}) \cdot f^*(\mathbf{x})^2]$. This ask, however, is information-theoretically infeasible. In our setting, we only get to see samples $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$ for discrete $\mathbf{y} \sim \mathrm{Ber}(f^*(\mathbf{x}))$ and do not have access to the values $f^*(\mathbf{x})$ themselves. Initially, in such a setting, it seems impossible to say anything meaningful about higher-order moments of $f^*(\mathbf{x})$.

Short of second-moment matching, degree-2 multicalibration imposes the following constraint for all $c \in \mathcal{C}$.

$$\mathbf{E}[\, c(\mathbf{x}) \cdot f(\mathbf{x})(\mathbf{y} - f(\mathbf{x})) \,] \approx 0 \implies \mathbf{E}[\, c(\mathbf{x}) \cdot f(\mathbf{x}) f^*(\mathbf{x}) \,] \approx \mathbf{E}\left[\, c(\mathbf{x}) \cdot f(\mathbf{x})^2 \,\right]$$

As a sanity check, observe that this is a *valid* constraint, since $f = f^*$ does satisfy it. As in multiaccuracy, the condition can be audited using only access to samples $(\mathbf{x}, \mathbf{y})$ and the predictions $f(\mathbf{x})$, without knowledge of $f^*(\mathbf{x})$. But what do we gain from adding this constraint? As we will see, this

[2] We use boldface for random variables. All expectations are taken over $\mathcal{D}$.

"degree-2" constraint turns out to be surprisingly powerful. Short of second moment matching, we prove that the degree-2 condition (approximately) implies the following second moment inequality:

$$\mathbf{E}\left[\ c(\mathbf{x}) \cdot f(\mathbf{x})^2\ \right] \leq \mathbf{E}\left[\ c(\mathbf{x}) \cdot f^*(\mathbf{x})^2\ \right]$$

It is unclear that this inequality by itself can be audited from samples alone, yet it is implied by the conditions of degree-2 multicalibration, which are indeed auditable. Intuitively, the inequality says that—unlike multiaccuracy—this degree-2 variant of multicalibration prevents overconfident predictions conditioned on $c \in \mathcal{C}$. It gives us an avenue to reason about $f(x)$ by comparing it to $f^*(x)$, using its first two moments.

The multicalibration hierarchy arises by lifting this intuition to higher degree polynomials: at the kth level, we obtain guarantees on the kth moments. Indeed, we show that a number of meaningful guarantees that hold for low-degree multicalibration, but not for multiaccuracy. Our results are in direct analogy with classic results in pseudorandomness, where k-wise—even pairwise—independence is known to be surprisingly powerful, in contrast to 1-wise independence [LW06]. As with pseudorandomness, we are able to prove new guarantees for multicalbration in its full generality, by showing that they hold even for low-degree multicalibration.

2.2 A Hierarchy of Multicalibration

With the motivation in place, we are ready to define the multicalibration hierarchy. In this overview, we present the notions focusing on binary predictors. In Section 3, we extend these definitions to the multi-class setting where outcomes are categorical random variables with $l \geq 2$ labels. As with previous notions, our variants of multicalibration are parametrized by a hypothesis class $\mathcal{C}$ and an approximation parameter $\alpha > 0$. A new ingredient of our definitions will be a family $\mathcal{W} \subseteq \{w : [0,1] \to [0,1]\}$ of *weight* functions that we will compose with the predictions of our model, which gives rise to a generic weighted version of multicalibration.

Definition. *Given a hypothesis class $\mathcal{C}$ and a weight class $\mathcal{W}$, we say that the predictor $f : \mathcal{X} \to [0,1]$ is $(\mathcal{C}, \mathcal{W}, \alpha)$-multicalibrated if for every $c \in \mathcal{C}$ and $w \in \mathcal{W}$ it holds that*

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

The primary conceptual contribution of this work is to identify choices of weight classes $\mathcal{W}$ that give rise to novel meaningful notions of multicalibration. We consider four notions, which fixing a hypothesis class $\mathcal{C}$, give increasingly stronger guarantees.

(1) **Multiaccuracy.** Taking $\mathcal{W} = \{1\}$ to consist of only the constant function $w(z) = 1$ for all z, we recover *multiaccuracy*.[3] Let $\mathrm{MA}(\alpha)$ denote the set of $(\mathcal{C}, \alpha)$-multiaccurate predictors.

(2) **Low-degree multicalibration.** We define a hierarchy of variants of low-degree multicalibration, by taking $\mathcal{W}$ to be families of low-degree polynomials. Formally, *degree-k* multicalibration uses weight functions defined by sparse degree-$(k-1)$ polynomials. We adopt this convention because using degree-$(k-1)$ polynomials as weight functions allow us to reason about the kth

[3] We adopt the formulation of multiaccuracy, initially defined in [KGZ19].

moments of our predictions. In the case of binary prediction, it suffices to take $\mathcal{W}_k = \{t^j\}_{j=0}^{k-1}$ to be the monomial basis. Let $\mathrm{MC}_k(\alpha)$ denote the set of $(\mathcal{C}, \alpha)$-degree-k multicalibrated predictors.

(3) **Smooth multicalibration.** Beyond the hierarchy, we consider multicalibration using the family all 1-Lipshcitz functions as our weight class. We refer to the resulting notion as $(\mathcal{C}, \alpha)$-*smooth* multicalibration. Smooth multicalibration directly extends the notion of smooth calibration [KF08, FH18], introduced to address issues of robustness in defining calibration. Let $\mathrm{MC}^{\mathrm{s}}(\alpha)$ denote the set of α-smooth multicalibrated predictors.

(4) **Full multicalibration.** Instead of explicitly conditioning on the predicted values, we define full multicalibration using indicator functions on the prediction intervals as our weight class. In binary prediction,[4] we define the δ-interval basis to be $\mathcal{I}_\delta = \{\mathbb{1}_{[(j-1)\delta, j\delta)} : j \in \lceil 1/\delta \rceil\}$ where $\mathbb{1}_{[a,b)}$ indicates membership in the interval $[a, b)$. We refer to $(\mathcal{C}, \mathcal{I}_\delta, \alpha)$-multicalibration as $(\mathcal{C}, \alpha, \delta)$-*full* multicalibration to distinguish it from smooth multicalibration, and use $\mathrm{MC}^{\mathrm{f}}_\delta(\alpha)$ to denote the set of predictors satisfying it.

With the variants in place, our first results establish the relationship between the multicalibration notions for a fixed class $\mathcal{C}$.

Proposition 1. *Fix a hypothesis class $\mathcal{C}$ and $\alpha \geq 0$. By construction, multiaccuracy and degree-1 multicalibration are equivalent. For every $k \geq 1$, increasing the degree leads to a more restrictive notion. In other words, the hierarchy satisfies the following inclusions.*

$$\mathrm{MA}(\alpha) = \mathrm{MC}_1(\alpha) \supseteq \mathrm{MC}_2(\alpha) \cdots \supseteq \mathrm{MC}_k(\alpha)$$

Further, low-degree multicalibration is a relaxation of smooth multicalibration, which is a relaxation of full multicalibration. That is, for any $k \geq 1$ and $\alpha \geq \delta \geq 0$,

$$\mathrm{MC}_k(k\alpha) \supseteq \mathrm{MC}^{\mathrm{s}}(\alpha) \supseteq \mathrm{MC}^{\mathrm{f}}_\delta(\alpha\delta - \delta^2)$$

[KF08, FH18] were motivated to introduce smooth calibration in order to address the lack of robustness in the classical notion of calibration. We show that smooth multicalibration has similar robustness guarantees; for instance, predictors that are close to smoothly multicalibrated functions are also smoothly multicalibrated (with modest degradation in the approximation parameter). We also show robustness to small (in ℓ_∞) perturbations to the weight function. This property is important algorithmically, since it allows us to infer smooth multicalibration (defined over all 1-Lipshcitz weight functions) from a small basis of functions that can uniformly approximate every 1-Lipschitz function in ℓ_∞.

2.3 Fairness from Low-Degree Multicalibration

We now return to our motivating question and ask: *does the hierarchy give strong fairness guarantees at low levels?* Our main technical contribution is a toolbox for reasoning about properties of predictors at low levels of the hierarchy, using their first few moments. Using this, we establish that the gurarantees of multicalibration for important measures of fairness, like the false error rates indeed manifest at low levels of the hierarchy. The hammer in this toolbox is the following *moment sandwiching bound* stated below for binary predictors with $\alpha = 0$.

[4] For multi-class prediction, we use an l-dimensional analog of the interval basis.

Theorem 2 (Informal)**.** *Suppose that $f : \mathcal{X} \to [0,1]$ is a $(\mathcal{C}, 0)$-degree-k multicalibrated predictor. Then, for every degree $d \leq k$ and every $c \in \mathcal{C}$,*

$$\mathbf{E}[c(\mathbf{x})f^*(\mathbf{x})^d] \underset{(a)}{\geq} \mathbf{E}[c(\mathbf{x})f(\mathbf{x})^d] \underset{(b)}{=} \mathbf{E}[c(\mathbf{x})f(\mathbf{x})^{d-1}f^*(\mathbf{x})] \underset{(c)}{\geq} \mathbf{E}[c(\mathbf{x})f(\mathbf{x})^{d-1}] \cdot \mathbf{E}[c(\mathbf{x})f^*(\mathbf{x})]$$

Our bounds are loosely inspired by the sandwiching bounds for importance weights from multicalibrated partitions proved in [GRSW21]. While we prove these bounds assuming only degree-k multicalibration, no analogous statements were known to hold, even for the original (stronger) notion of multicalibration.

The utility of this sandwiching bound can be seen by thinking about each inequality separately. First, note that (b) follows immediately by the definition of degree-k multicalibration. We dig into (a) and (c) separately. The upper bound in (a) says that first k moments of the multicalibrated predictor f are dominated by those of the ground truth predictor f^*, conditioned on any $c \in \mathcal{C}$. Given that the exact kth moment of f^* is inaccessible through samples, this inequality is useful in lieu of exact moment matching.

Specifically, while we can't perfectly match moments for $d \geq 2$, this inequality implies that the predictions of f cannot be overconfident. For example, for a degree-2 multicalibrated f, the inequality implies that conditioned on any $c \in \mathcal{C}$, the variance of f is no more than the variance of f^*, and their expectations match. In other words, any variation in the predictions of f over $c \in \mathcal{C}$ can be attributed to true variation in the distribution of $f^*(\mathbf{x})|c(\mathbf{x}) = 1$. Concretely, if all points with $c(\mathbf{x}) = 1$ were identical under f^*, then f cannot treat them differently! This stands in stark constrast with multiaccuracy, where many of the "failure modes" of multiaccuracy exploit this weakness [DKR+19, DKR+21].

The lower bound in (c) can be interpreted as saying that $f^{k-1}(\mathbf{x})$ is positively correlated with the label $\mathbf{y}$, by expressing the difference in expectations as the covariance between $f^{k-1}(\mathbf{x})$ and $\mathbf{y}$.

$$\mathrm{Cov}[f(\mathbf{x})^{k-1}, \mathbf{y}] = \mathbf{E}[f(\mathbf{x})^{k-1}f^*(\mathbf{x})] - \mathbf{E}[f(\mathbf{x})^{k-1}]\,\mathbf{E}[f^*(\mathbf{x})] \geq 0$$

In this light, the bounds from degree-2 multicalibration can be summarized succinctly: the covariance between $f(\mathbf{x})$ and the ground truth labels $\mathbf{y}$ within any $c \in \mathcal{C}$ is always positive, but never more than the covariance between the optimal $f^*(\mathbf{x})$ and $\mathbf{y}$ in c. Again, neither claim necessarily holds under multiaccuracy.

As a concrete application, we show how sandwiching bounds allow us to reason about the true and false error rates of f, which have been intensively studied in algorithmic fairness [HPS+16, KMR17]. In particular, the bounds extend even when we condition on the value of the true label. For instance, conditioning on $\mathbf{y} = 1$, we can bound the generalized true positive rate of f compared to that of f^*, even conditioned on subpopulations.[5]

$$\mathbf{E}\left[\, c(\mathbf{x})f(\mathbf{x}) \,\right] \leq \mathbf{E}\left[\, c(\mathbf{x})f(\mathbf{x}) \mid \mathbf{y} = 1 \,\right] \leq \mathbf{E}\left[\, c(\mathbf{x})f^*(\mathbf{x}) \mid \mathbf{y} = 1 \,\right]$$

This bound crystallizes the inutition that multicalibrated predictors prevent overconfidence. Even though f has positive correlation with $\mathbf{y}$ over c, the predictions don't assign excessively high probabilities to points which are more likely to be 1. Such overconfidence is a well-known issue in

[5] An analogous statement can be made about the true negative rates.

large neural networks [GPSW17]. Whereas full multicalibration addresses overconfidence at a fine-grained level, conditioning on every value of the prediction, even degree-2 multicalibration gives a qualitatively similar guarantee.

2.4 The Pragmatic Appeal of Low-Degree Multicalibration

We establish the feasibility of low-degree and smooth multicalibration by giving an efficient learning algorithm. More generally, following the boosting-style approach of [HKRR18], we show that a weak agnostic learner for $\mathcal{C}$ [BLM01, KK09] suffices to obtain $(\mathcal{C}, \mathcal{W}, \alpha)$-multicalibration for any finite weight class $\mathcal{W}$. This procedure, described as Algorithm 1, makes a number of calls to the weak learner bounded polynomially in $|\mathcal{W}|$ and $1/\alpha$. The algorithm immediately demonstrates the feasibility of degree-k multicalibration for any fixed $k \in \mathbb{N}$. For smooth multicalibration, where the definition involves an infinite family $\mathcal{W}$, our algorithm uses constructions of finite bases that uniformly approximate the family of 1-Lipschitz functions.

We now delve further into the sample efficiency of each notion. In principle, the relaxed notion of low-degree multicalibration is implied by smooth or full multicalibration. However, there is a significant degradation in the accuracy parameter α, which translates to a blowup in computational and sample complexity. Concretely, suppose we train a predictor f to satisfy $(\mathcal{C}, \alpha_0, \delta)$-full multicalibration; in order to guarantee that f satisfies $(\mathcal{C}, \alpha)$-degree-k multicalibration, we need to take $\alpha_0 \ll \alpha$ much smaller than if we train for degree-k multicalibration directly. For a fair comparison, we compare the sample complexity needed by each variant to obtain the same guarantee: $(\mathcal{C}, \alpha)$ degree-k multicalibration. We make the assumption that $\mathcal{C}$ has a sample-optimal weak agnostic learner in terms of the dependence on α_0.[6] We state the theorem informally, using shorthand $m \sim B$ to denote $m \leq \widetilde{O}(B)$ (see Section 5, Theorem 5.5 for a formal statement).

Theorem 3 (Informal)**.** *For a $\mathcal{C}$ be a hypothesis class, let $\mathsf{VC}(\mathcal{C})$ denote its VC-dimension. For any $\alpha > 0$ and $k \in \mathbb{N}$, the sample complexity to obtain a predictor $f \in \mathrm{MC}_k(\alpha)$ (with constant failure probability) is bounded as m_k using degree-k multicalibration, m_s using smooth multicalibration, and m_i using full multicalibration, for m_k, m_s, m_i bounded as follows.*

$$m_k \sim \frac{l \cdot (\mathsf{VC}(\mathcal{C}) + k)}{\alpha^4} \qquad m_s \sim \frac{k^4 l \cdot \mathsf{VC}(\mathcal{C})}{\alpha^4} + \frac{k^{l+3} l^l}{\alpha^{l+3}} \qquad m_i \sim \frac{(2kl)^{4(l+1)} \cdot \mathsf{VC}(\mathcal{C})}{\alpha^{4(l+1)}}$$

Note the dependence is $\mathrm{poly}(l, 1/\alpha)$ for low-degree multicalibration, but $\Omega(1/\alpha)^{bl}$ for smooth and full multicalibration, with a factor 4 difference in the constant b between them. This suggests that in the multi-class setting, low-degree multicalibration can lead to exponential savings, as compared to the stronger notions. Even under tighter analyses tailored to the binary prediction setting, we obtain polynomial savings in $1/\alpha$ from low-degree multicalibration.

Given that many desirable properties of multicalibration start to take effect at small degrees (even degree-2), these results suggest a pragmatic win for low-degree multicalibration. In the realistic setting where a learner is given a fixed data set of sample, the learner may achieve stronger fairness and accuracy properties by training for low-degree multicalibration directly, as compared to either multiaccuracy or full multicalibration.

[6] Suboptimal dependence of the learner on α_0 will *increase* the gaps in sample complexity between low-degree and the stronger variants of multicalibration.

While these sample complexities give asymptotic upper bounds, in Section 6, we report on a proof-of-concept experiment that explore the findings in a semi-synthetic setup. By using a semi-synthetic setup, we can access the ground-truth f^* values, thereby evaluating quantities like the gap in moments between f^* and the learned predictors. We show that, even in a standard binary prediction setting, the sample efficiency gap is not a merely asymptotic phenomenon, but is realized in a setting with a few thousand samples from the data distribution.

2.5 Discussion and Related Works

Low-degree multicalibration adds to a growing list of works that study multicalibration and related notions. On a conceptual level, our work is closely related to the idea of outcome indistinguishability (OI), introduced by [DKR+21, DKR+22]. OI is not a single notion, but rather an extensible framework for reasoning about the guarantees of predictions in the language of computational indistinguishability. In particular, [DKR+21] also define a hierarchy of notions, based on the way distinguisher algorithms may access the predictions $f(\mathbf{x})$. The first two levels of their hierarchy correspond tightly to multiaccuracy and multicalibration; intuitively, multiaccuracy distinguishers do not get to observe $f(\mathbf{x})$, where as multicalibration distinguishers may depend on $f(\mathbf{x})$ arbitrarily. Low-degree multicalibration refines the idea of access to $f(\mathbf{x})$, where the corresponding distinguishers can functionally depend on $f(\mathbf{x})$ in restricted ways.

On a technical level, our sandwiching bounds, which shed light on how multicalibrated predictors control for uncertainty relative to the uncertainty of the optimal predictions, are actually inspired by work on unsupervised distribution learning of [GRSW21]. Their work establishes analogous moment sandwiching bounds for density estimates given by so-called *multicalibrated partitions*. Closely related—but not to be confused with low-degree multicalibration—is the idea of *moment multicalibration* due to [JLP+21]. Moment multicalibration obtains guarantees of uncertainty quantification in the setting of *real-valued* outcomes, by (full) multicalibrating on higher moments of the outcomes. Importantly, in contrast to our setting, the relevant moments of real-valued outcomes can be estimated to arbitrary precision with enough samples. Using these estimates, [JLP+21] derive Chebyshev-style confidence intervals across subpopulations for their multicalibrated predictions.

Very recently, multicalibration and OI have played a key technical role in obtaining strong guarantees for omniprediction [GKR+22] and multi-group agnostic learning [RY21]. Both results are known to follow from multicalibration, but not from multiaccuracy. In light of our work, it would be interesting to revisit these results to see whether the proofs use the full power of full multicalibration or whether we can recover the guarantees using low-degree techniques. More broadly, we speculate that the low-degree multicalibration hierarchy suggests a certain *proof system*—akin to the convex programming sum-of-squares hierarchy [BS14]—in which properties derived from low-degree moments of f^* could be derived for low-degree multicalibrated predictors for free. Exploring this intuition further and how it may connect with OI is a fascinating direction for future research.

Multicalibration was initially developed as a strengthening of calibration to provide meaningful fairness guarantees, not just on the basis of marginally-defined groups, but intersectionally [HKRR18, KGZ19]. The risk of inequity due to miscalibrated predictions has been well-documented [KMR16, PRW+17, GKR19], especially in the setting of medical risk prediction [OPVM19, BYR+21]. Despite its origins as a complexity-theoretic fairness notion, mutlicalibration has already seen clinical application to address such equity issues [BYR+21] and to develop a COVID-19 risk predictor

in the early days of the pandemic [BRA+20]. While individual-level calibration is generally impossible [BCRT19], recent works have investigated settings where guarantees at an individual-level are possible, through hedging [ZE21] and randomization [ZME20].

Calibration has a rich history in the forecasting literature, as a criterion for uncertainty quantification [Bri50, Daw84, FV98, KF08, FH18]. For multi-class prediction tasks, the strongest definition is known as distribution calibration [KF15, SDKF19], and is known to require exponentially many samples in the number of labels. Consequently, practitioners commonly use very weak notions such as confidence calibration [P+99, GPSW17], class-wise calibration [KPNK+19]. To address this tension, [ZKS+21] recently proposed a notion of *decision calibration*, which ensures that the predictions are calibrated with respect to downstream decisions, avoiding the infeasibility of calibrating to the predictions, but strengthening the marginal guarantees of confidence and class-wise calibration. Naturally, one could extend the definition of decision calibration to decision multicalibration, with similar motivations to low-degree multicalibration but incomparable guarantees.

3 Defining the Multicalibration Hierarchy

Notation. In a generic supervised learning problem, we are given a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the domain and $\mathcal{Y}$ is the set of labels. We consider the multi-class setting where $\mathcal{Y} = [l]$ for $l \geq 2$. We represent each outcome $i \in [l]$ by the "one-hot" encoding $e_i \in \mathbb{R}^l$. We denote sampling from $\mathcal{D}$ by $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$ where $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathbb{R}^l$, so that $\mathbf{y}_\ell$ is the indicator for label equalling $\ell \in [l]$. Let Δ_l denote the space of probability distributions over $[l]$. We associate every distribution $p \in \Delta_l$ with a vector in $p \in \mathbb{R}^l$ where $p_\ell = \Pr_{\mathbf{y} \sim p}[\mathbf{y} = e_\ell]$. An l-class predictor is a function $f : \mathcal{X} \to \Delta_l$ which maps each point to a distribution over labels. Throughout, we use $f^* : \mathcal{X} \to \Delta_l$ to denote the Bayes optimal predictor, defined as

$$f^*(x) = \mathbf{E}[\ \mathbf{y} \mid \mathbf{x} = x\ ]$$

where for each $\ell \in [l]$, $f^*_\ell(x) = \Pr[\ \mathbf{y}_\ell = 1 \mid \mathbf{x} = x\ ]$. In other words, $f^*(x)$ governs the true distribution over classes for a given individual $x \in \mathcal{X}$. Some of our results will be stated for the special case of binary prediction ($l = 2$), in which case, we let $f : \mathcal{X} \to [0, 1]$, where $f(\mathbf{x})$ estimates of $\Pr[\ \mathbf{y} = 1 \mid \mathbf{x}\ ]$.

3.1 Multicalibration with Weight Classes

We present a unified framework for defining variants of multicalibration that captures the original notions, as defined by [HKRR18], but also naturally captures the novel notions of low-degree and smooth multicalibration. We present all of the definitions in terms of l-class predictors, which generalizes their original definitions for binary predictors.

Multicalibration is parameterized by a hypothesis class of functions $\mathcal{C} \subseteq \{c : \mathcal{X} \to [0, 1]\}$. The class $\mathcal{C}$ may be finite or infinite, but importantly, we think of $\mathcal{C}$ as having a simple representation. Natural choice of $\mathcal{C}$ include linear/logistic hypotheses, decision forests of a fixed depth, or neural networks of a fixed size and architecture. Departing from prior works, we additionally parameterize multicalibration by a *weight class* $\mathcal{W} \subseteq \{w : \Delta_l \to [0, 1]^l\}$. The weight functions will be applied on the predictions from f.

The classic calibration constraint requires that predictions be accurate in expectation, even after conditioning on the predicted value. Intuitively, the class of weight functions serves as an analog of "conditioning" on a prediction. Taking different choices of $\mathcal{W}$ will realize the original and novel variants of multicalibration. Generically, we define multicalibration in terms of the calibration class $\mathcal{C}$, the weight class $\mathcal{W}$, and an approximation parameter $\alpha \geq 0$.

Definition 3.1. *Given a hypothesis class $\mathcal{C} \subseteq \{\mathcal{X} \to [0,1]\}$ a weight class $\mathcal{W} \subseteq \{\Delta_l \to [0,1]^l\}$ and $\alpha \geq 0$, a predictor $f : \mathcal{X} \to \Delta_l$ is $(\mathcal{C}, \mathcal{W}, \alpha)$-multicalibrated if for every $c \in \mathcal{C}$ and $w \in \mathcal{W}$*

$$\left| \mathop{\mathbf{E}}_{\mathcal{D}} \left[\; c(\mathbf{x}) \cdot \left\langle w(f(\mathbf{x})), \mathbf{y} - f(\mathbf{x}) \right\rangle \; \right] \right| \leq \alpha. \tag{2}$$

With this general framework in place, we instantiate it with various choices of weight functions to derive four notions of increasing strength.

Multiaccuracy. [HKRR18, KGZ19] The weakest notion, multiaccuracy, requires that predictions be accurate in expectation on each $c \in \mathcal{C}$. Specifically, an l-class predictor $f : \mathcal{X} \to \Delta_l$ is $(\mathcal{C}, \alpha)$-multiaccurate if for each $\ell \in [l]$,

$$\left| \mathop{\mathbf{E}}_{\mathcal{D}} \left[\; c(\mathbf{x}) \cdot (\mathbf{y}_\ell - f(x)_\ell) \; \right] \right| \leq \alpha. \tag{3}$$

To instantiate $(\mathcal{C}, \alpha)$-multiaccuracy from $(\mathcal{C}, \mathcal{W}, \alpha)$-multicalibration, we can take $\mathcal{W} = \{w_\ell\}_{\ell \in [l]}$ where $w_\ell(p) = e_\ell$ for all $p \in \Delta_l$. In words, w_ℓ is constantly the ℓ^{th} standard basis vector. In the case where $f : \mathcal{X} \to [0,1]$ is a binary predictor, we can simply take $\mathcal{W} = \{w_0\}$ to contain the constant function $w_0(p) = 1$ for all $p \in [0,1]$. We use $\mathrm{MA}(\alpha)$ to denote the set of all predictors that satisfy Equation (3).

$$\mathrm{MA}(\alpha) = \{f : \mathcal{X} \to \Delta_l \text{ satisfying (3)}\}$$

Low-degree multicalibration. Intuitively, multiaccuracy enforces a collection of linear (in $f(\mathbf{x})$) constraints based on $\mathcal{C}$. In low-degree multicalibration, we strengthen these constraints by taking the class of weight functions to be the family of low-degree polynomials. For instance, in the binary prediction setting, degree-2 multicalibration enforces the following nonlinear constraint.

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

A degree-k polynomial is a function

$$q(z) = \sum_{S \in [l]^j :\; j \leq k} q_S \cdot \prod_{i \in S} z_i$$

where S goes over all multisets of elements from $[l]$, of size at most k. $\mathcal{P}_k$ is the family of all degree-k polynomials that satisfy the following two conditions

$$q(z) \in [0,1] \quad \forall z \in \Delta_l, \tag{4}$$

$$\sum_{S \in [l]^j :\; j \leq k} |q_S| \leq 1. \tag{5}$$

We refer to (4) as boundedness and (5) as sparsity. For $k \geq 1$, we define the weight class $\mathcal{W}_k$ as all functions where each coordinate belongs to $\mathcal{P}_{k-1}$.

$$\mathcal{W}_k = \left\{ w : \Delta_l \to [0,1]^l \text{ such that } w_i \in \mathcal{P}_{k-1} \quad \forall i \in [l] \right\}$$

Note that we define the degree-k weight class in terms of degree-$(k-1)$ polynomials. We adopt this convention because, as we will see, using degree-$(k-1)$ polynomials allows us to reason about the kth moments of the predictor. With this class in place, we can define degree-k multicalibration.

Definition 3.2. *For $k \geq 1$, a predictor $f : \mathcal{X} \to \Delta_l$ is $(\mathcal{C}, \alpha)$-degree-k multicalibrated if it is $(\mathcal{C}, \mathcal{W}_k, \alpha)$-multicalibrated.*

Let $\mathrm{MC}_k(\alpha)$ be the set of degree-k-multicalibrated predictors with approximation parameter α.

Proposition 3.3. *For any calibration class $\mathcal{C}$ and approximation parameter α, multiaccuracy and degree-1-multicalibration are identical, hence*

$$\mathrm{MA} = \mathrm{MC}_1(\alpha),$$

and for every $k \geq 2$, degree-k-multicalibration implies degree-$(k-1)$-multicalibration, hence

$$\mathrm{MC}_{k-1}(\alpha) \supseteq \mathrm{MC}_k(\alpha).$$

In other words, low-degree multicalibration is a strengthening of multiaccuacy, and becomes a more restrictive notion as we increase the degree k. The proposition follows immediately by the construction of $\mathcal{P}_k$ and the fact that $\mathcal{P}_{k-1} \subseteq \mathcal{P}_k$.

Smooth multicalibration. As we increase k, intuitively, low-degree multicalibration will begin to approximate multicalibration with arbitrary, smooth weight functions. This motivates our definition of smooth multicalibration using Lipschitz functions. We consider weight functions $w : \Delta_l \to [0,1]^l$ that are $\ell_1 \to \ell_\infty$ Lipschitz; that is, for all $z, z' \in \Delta_l$

$$\left\| w(z) - w(z') \right\|_\infty \leq \left\| z - z' \right\|_1 . \tag{6}$$

We consider the class $\mathcal{L}_{1\to\infty}$ of all such Lipschitz functions.

$$\mathcal{L}_{1\to\infty} = \left\{ w : \Delta_l \to [0,1]^l \text{ satisfying (6)} \right\}$$

Definition 3.4. *A predictor $f : \mathcal{X} \to \Delta_l$ is $(\mathcal{C}, \alpha)$-smoothly multicalibrated if it is $(\mathcal{C}, \mathcal{L}_{1\to\infty}, \alpha)$-multicalibrated.*

While we define smooth multicalibration in terms of 1-Lipschitz weight functions, the property naturally extends to any weight function with bounded Lipschitz constant. Denote $r\mathcal{L}_{1\to\infty}$ as the set of weight functions $w : \Delta_l \to [0,1]^l$, where for all $z, z' \in \Delta_l$, $\|w(z) - w(z')\|_\infty \leq r \cdot \|z - z'\|_1$.

Proposition 3.5. *For any calibration class $\mathcal{C}$, approximation parameter α, and constant $r > 1$, if $f : \mathcal{X} \to \Delta_l$ is $(\mathcal{C}, \alpha)$-smoothly multicalibrated, then for any $w \in r\mathcal{L}_{1\to\infty}$,*

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

This proposition is an immediate consequence of the smooth multicalibration guarantee and linearity of expectation.

Full Multicalibration. [HKRR18] The strongest variant of multicalibration corresponds to the classic notion of calibration, where the expectation is taken conditional on the predicted value. We present our generalization to the multi-class setting.

For a measurable set $S \subseteq \Delta_l$, let $\mathbb{1}_S$ be the indicator function of the set; that is, for any $z \in \Delta_l$,

$$\mathbb{1}_S(z) = \begin{cases} 1 & z \in S \\ 0 & z \notin S \end{cases}$$

Let Π be a partition of Δ_l and denote by $\mathcal{I}_\Pi = \{\mathbb{1}_P : P \in \Pi\}$. For binary predictors, we focus on the *interval basis*. For $\delta > 0$, we define a partition Π_δ of the interval $[0, 1]$ to be

$$\Pi_\delta = \left\{ [(j-1)\delta, j\delta) \text{ for } j \in \{1, \ldots, \lceil 1/\delta \rceil\} \right\}.$$

We use $\mathcal{I}_\delta = \mathcal{I}_{\Pi_\delta}$ to denote the basis of indicator functions on each δ-interval. Taking $\mathcal{I}_\delta$ as our weight class in the binary setting, $(\mathcal{C}, \mathcal{I}_\delta, \alpha)$-multicalibration gives us the notion we call $(\mathcal{C}, \alpha, \delta)$-full multicalibration, which recovers the original notion of multicalibration, as defined by [HKRR18].[7] For $l > 2$, we define the partition Π_δ^l of the interval $[0, 1]^l$ to be the l-wise Cartesian product of Π_δ, whose elements are products of δ-width intervals in each dimension. We set $\mathcal{I}_\delta^l = \mathcal{I}_{\Pi_\delta^l}$ and use these as weight functions in full multicalibration. We will use the easy bound $|\Pi_l^\delta| \leq \lceil 1/\delta \rceil^l$, it also holds that $|\Pi_l^\delta| \leq l \lceil 1/\delta \rceil^{l-1}$.

Definition 3.6. *A predictor $f : \mathcal{X} \to \Delta_l$ is $(\mathcal{C}, \alpha, \delta)$-full multicalibrated if it is $(\mathcal{C}, \mathcal{I}_\delta^l, \alpha)$-multicalibrated.*

3.2 Understanding the Hierarchy

Proposition 3.3 demonstrates that there is a hierarchy of notions of low-degree multicalibration. Next, we show how low-degree multicalibration compares to smooth and full multicalibration. We show how for any $k \in \mathbb{N}$, for appropriately chosen approximation parameters, smooth multicalibration can implement degree-k multicalibration. Then, we show how full multicalibration can implement smooth multicalibration.

Theorem 3.7. *Fix a calibration class $\mathcal{C}$ and an approximation parameter $\alpha \geq 0$. For any $k \geq 2$, every $(\mathcal{C}, \alpha)$-smoothly multicalibrated predictor is $(\mathcal{C}, (k-1)\alpha)$-degree-$k$ multicalibrated.*

$$\mathrm{MC}^{\mathrm{s}}(\alpha) \subseteq \mathrm{MC}_k((k-1)\alpha).$$

The theorem follows by bounding the $\ell_1 \to \ell_\infty$ Lipschitz constant of weight functions $w \in \mathcal{W}_k$ by $(k-1)$ and then appealing to Proposition 3.5.

Lemma 3.8. *For $k \geq 2$, the degree-k weight functions are $(k-1)$-Lipschitz; that is,*

$$\mathcal{W}_k \subseteq (k-1)\mathcal{L}_{1\to\infty}.$$

[7] Technically, [HKRR18] work with a version of this notion that requires α error, defined *relative* to the probability mass of $c^{-1}(1)$ and $\Pr[f(\mathbf{x}) \in [(j-1)\delta, j\delta)]$, whereas our notion gives an absolute approximation guarantee. By adjusting the choice of α appropriately, it is easy to implement one notion as the other. We follow other works adopting this absolute-error convention, to avoid many of the hassles of dealing with conditional expectations.

Proof. To prove the claim, we need to show for any $z, z' \in \Delta_l$,

$$\left\|w(z) - w(z')\right\|_\infty \leq (k-1) \cdot \left\|z - z'\right\|_1 .$$

Recall that each coordinate of $w : \Delta_l \to [0,1]^l$ can be expressed as an sparse degree-$(k-1)$ polynomial $w_i \in \mathcal{P}_{k-1}$. To obtain this $\ell_1 \to \ell_\infty$ bound, we start with the max over class predictions $i \in [l]$, and bound this quantity in terms of the gradient of $w_i \in \mathcal{P}_{k-1}$.

$$\begin{aligned}
\left\|w(z) - w(z')\right\|_\infty &\leq \max_{i \in [l]} \left|w(z)_i - w(z')_i\right| \\
&\leq \max_{i \in [l]} \max_{z^* \in \Delta_l} \langle \nabla w_i(z^*), z - z' \rangle \qquad (7) \\
&\leq \max_{i \in [l]} \max_{z^* \in \Delta_l} \left\|\nabla w_i(z^*)\right\|_\infty \cdot \left\|z - z'\right\|_1
\end{aligned}$$

where (7) follows by the smoothness of w_i and the mean value theorem, establishing that for some $\bar{z} \in \Delta_l$, $|w(z)_i - w(z')_i| = \langle \nabla w_i(\bar{z}), z - z' \rangle$. Thus, to bound $\left\|\nabla w_i(z^*)\right\|_\infty$ it suffices to bound the gradient over any $q \in \mathcal{P}_{k-1}$ on $z^* \in \Delta_l$. Using the degree and sparsity bounds, a simple convexity argument demonstrates that for any polynomial $q \in \mathcal{P}_{k-1}$ and $i \in [l]$,

$$\max_{z \in \Delta_l} \frac{\partial q}{\partial z_i}(z) \leq k - 1.$$

In all, we derive the bound

$$\left\|w(z) - w(z')\right\|_\infty \leq \max_{q \in \mathcal{P}_{k-1}} \max_{z^* \in \Delta_l} \left\|\nabla q(z^*)\right\|_\infty \cdot \left\|z - z'\right\|_1 \leq (k-1) \cdot \left\|z - z'\right\|_1 .$$

□

Next we show that full multicalibration with a sufficiently small error parameter leads to smooth multicalibration.

Theorem 3.9. *Fix a calibration class $\mathcal{C}$ and an approximation parameter $\alpha \geq 0$. Every $(\mathcal{C}, \beta, \delta)$-full multicalibrated predictor is $(\mathcal{C}, \alpha)$-smooth multicalibrated for*

$$\alpha = \beta \cdot \lceil 1/\delta \rceil^l + l\delta.$$

Hence for $\delta \leq \alpha/l$ and $1/\delta \in \mathbb{N}$, we have $\mathrm{MC}^{\mathrm{f}}_\delta(\alpha\delta^l - l\delta^{l+1}) \subseteq \mathrm{MC}^{\mathrm{s}}(\alpha)$.

To prove Theorem 3.9, we first establish a few key properties of calibration defined by weight functions that will be useful throughout our discussion. We start by showing that $(\mathcal{C}, \mathcal{W}, \alpha)$-multicalibration is robust to small perturbations of the weight functions in $\mathcal{W}$.

Lemma 3.10. *Let $v, w : \Delta^l \to [0,1]^l$ be weight functions such that*

$$\max_{p \in \Delta^l} \left\|v(p) - w(p)\right\|_\infty \leq \eta.$$

Then for any l-class predictor $f : \mathcal{X} \to \Delta^l$ and any $c : \mathcal{X} \to [0,1]$,

$$\left| \mathop{\mathbf{E}}_{\mathcal{D}} [\, c(\mathbf{x}) \langle v(f(\mathbf{x})), \mathbf{y} - f(\mathbf{x}) \rangle \,] - \mathop{\mathbf{E}}_{\mathcal{D}} [\, c(\mathbf{x}) \langle w(f(\mathbf{x})), \mathbf{y} - f(\mathbf{x}) \rangle \,] \right| \leq 2\eta.$$

Proof. Observe that for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$\begin{aligned}|c(x) \cdot (\langle v(f(x)) - w(f(x)), y - f(x)\rangle)| &\le |c(x)| \cdot \|v(f(x)) - w(f(x))\|_\infty \cdot \|y - f(x)\|_1 \\ &\le 2\eta\end{aligned}$$

since $|c(x)| \le 1$ and $\|y - f(x)\|_1 \le 2$. The claim follows by averaging over $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$. □

A key application of the robustness to weight functions is that one can infer smooth multicalibration—a condition defined in terms of an infinite weight class—from multicalibration for a finite basis of functions that uniformly approximates every function in $\mathcal{L}_{1\to\infty}$.

Definition 3.11. *A collection of weight functions $\mathcal{W} = \{w_i\}_{i=1}^k$ is a (η, L) basis for $\mathcal{L}_{1\to\infty}$ if for every $u \in \mathcal{L}_{1\to\infty}$, there exists $v : \Delta_l \to [0,1]^l$ such that $\|u - v\|_\infty \le \eta$ where*

$$v = \sum_{i=1}^k \lambda_i w_i, \ \sum_{i=1}^k |\lambda_i| \le L.$$

Typically, both k and L will be functions of η. Our interest in ℓ_∞ approximations is motivated by the following lemma.

Lemma 3.12. *If $\mathcal{W}$ is (η, L) basis for $\mathcal{L}_{1\to\infty}$ and the predictor f is $(\mathcal{C}, \mathcal{W}, \beta)$-multicalibrated on $\mathcal{C}$, then f is $(\mathcal{C}, \alpha)$-smoothly multicalibrated where $\alpha = \beta L + 2\eta$.*

Proof. For a weight function $u \in \mathcal{L}_1$, let $v = \sum_i \lambda_i w_i$ be the ℓ_∞ approximation guaranteed from the definition of $\mathcal{W}$. Then by linearity of expectation

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathcal{D}}[c(\mathbf{x})\langle v(f(\mathbf{x}), (\mathbf{y} - f(\mathbf{x}))\rangle] &= \mathop{\mathbf{E}}_{\mathcal{D}}\left[c(\mathbf{x}) \sum_{i=1}^n \lambda_i \langle w_i(f(\mathbf{x})), \mathbf{y} - f(\mathbf{x})\rangle\right] \\ &= \sum_{i=1}^n \lambda_i \mathop{\mathbf{E}}_{\mathcal{D}}[c(\mathbf{x})\langle w_i(f(\mathbf{x})), \mathbf{y} - f(\mathbf{x})\rangle]\end{aligned}$$

Taking absolute values and using the assumption that f is $(\beta, \mathcal{W})$-multicalibrated on $\mathcal{C}$ gives

$$\left|\mathop{\mathbf{E}}_{\mathcal{D}}[c(\mathbf{x})v(f(\mathbf{x}))(\mathbf{y} - f(\mathbf{x}))]\right| \le \left|\sum_{i=1}^n \lambda_i \beta\right| \le \beta L$$

To complete the proof, we now use Lemma 3.10 to conclude that f is $\alpha = \beta L + 2\eta$-smoothly multicalibrated on $\mathcal{C}$. □

Proposition 3.13. *For any $l \ge 2$, the set I_δ^l is a $(l\delta/2, \lceil 1/\delta \rceil^l)$ basis for $\mathcal{L}_{1\to\infty}$.*

Proof. Recall that Π_δ^l partitions $[0,1]^l$ in a set of cubes (products of intervals). For each cube $\pi \in \Pi_\delta^l$, we pick the center point $z_\pi \in \pi$, note that it satisfies $\|z - x\|_1 \le \delta l/2$ for every $x \in \pi$. Given a weight function $u \in \mathcal{L}_1$, we approximate it by the function v where

$$v(x) = \sum_{\pi \in \Pi_\delta^l} u(z_\pi) \mathbb{1}_\pi(x)$$

which assigns the value $u(z_\pi)$ to every point $x \in \pi$. Since $u \in \mathcal{L}_{1\to\infty}$, for $x \in \pi$

$$\|v(x) - u(x)\|_\infty = \|u(x) - u(z_\pi)\|_\infty \le \|x - z_\pi\|_1 \le l\delta/2,$$

$$\sum_{\pi \in \Pi_\delta^l} |u(z_\pi)| \le |\Pi_\delta^l| \le \lceil 1/\delta \rceil^l$$

hence the claim follows. □

Combining Proposition 3.13 with Lemma 3.12 completes the proof of Theorem 3.9. In the setting of binary labels, one gets the following strengthening, which improves the dependence on δ, by observing that $|\Pi_\delta| \le \lceil 1/\delta \rceil$.

Proposition 3.14. *For binary classification, let $\delta < \alpha$ and $1/\delta \in \mathbb{N}$. If f is $(\mathcal{C}, \alpha\delta - \delta^2, \delta)$-full multicalibrated then it is $(\mathcal{C}, \alpha)$-smoothly multicalibrated, hence $\mathrm{MC}^{\mathrm{f}}_\delta(\alpha\delta - \delta^2) \subseteq \mathrm{MC}^{\mathrm{s}}(\alpha)$.*

3.3 Robustness of Low-Degree and Smooth Multicalibration

While calibration is an intuitively desirable property of predictors, one frustration in reasoning about calibrated and multicalibrated predictors is that full (multi)calibration is not robust to small perturbations in predictions. For instance, consider a binary prediction setting where $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1$ is equally partitioned such that in one half all $\mathcal{X}_0$ have associated label $y = 0$, and in the other half $\mathcal{X}_1$ all have $y = 1$. Then, the constant predictor $f(x) = 1/2$ is perfectly calibrated, but the ϵ-close predictor $f(x) = 1/2 - \epsilon$ for $x \in \mathcal{X}_0$ and $f(x) = 1/2 + \epsilon$ for $x \in \mathcal{X}_1$, is far from calibrated.

In this section, we show that—in stark contrast to full multicalibration—low-degree and smooth multicalibration are both robust to small perturbations of the predictor.

Theorem 3.15. *Let $f, g : \mathcal{X} \to \Delta^l$ be l-class predictors such that for $\delta > 0$,*

$$\mathbf{E}\left[\ \|f(\mathbf{x}) - g(\mathbf{x})\|_1\ \right] \le \delta.$$

- *If f is $(\mathcal{C}, \alpha)$-degree-k-multicalibrated, then g is $(\mathcal{C}, \alpha + (2k-1)\delta)$-degree-$k$ multicalibrated.*
- *If f is $(\mathcal{C}, \alpha)$-smoothly multicalibrated, then g is $(\mathcal{C}, \alpha + 3\delta)$-smoothly multicalibrated.*

This theorem is a consequence of the following more general lemma which applies to any weight family $\mathcal{W}$ with bounded $\ell_1 \to \ell_\infty$ Lipschitz constant.

Lemma 3.16. *Let $f, g : \mathcal{X} \to \Delta^l$ be l-class predictors such that for $\delta > 0$,*

$$\mathbf{E}\left[\ \|f(\mathbf{x}) - g(\mathbf{x})\|_1\ \right] \le \delta.$$

Let $\mathcal{W} \subseteq r\mathcal{L}_{1\to\infty}$. If f is $(\mathcal{C}, \mathcal{W}, \alpha)$-multicalibrated, then g is $(\mathcal{C}, \mathcal{W}, \alpha + (2r+1)\delta)$-multicalibrated.

Proof. Fix any $w \in r\mathcal{L}_{1\to\infty}$. By the Lipschitz property, we derive the following inequalities:

$$|\langle w(f(x)) - w(g(x)), y\rangle| \le \|w(f(x)) - w(g(x))\|_\infty \le r\,\|f(x) - g(x)\|_1$$

and

$$\begin{aligned}
&|\langle w(f(x)), f(x)\rangle - \langle w(g(x)), g(x)\rangle| \\
&= |\langle w(f(x)), f(x) - g(x)\rangle + \langle w(f(x)) - w(g(x)), g(x)\rangle| \\
&\leq \|w(f(x)\|_\infty \cdot \|f(x) - g(x)\|_1 + \|w(f(x)) - w(g(x))\|_\infty \cdot \|g(x)\|_1 \\
&\leq (r+1)\,\|f(x) - g(x)\|_1 .
\end{aligned}$$

Thus, for every $x \in \mathcal{X}$, by the triangle inequality, we bound the difference of the expressions by

$$|\langle w(f(x)), y - f(x)\rangle - \langle w(g(x)), y - g(x)\rangle| \leq (2r+1)\,\|f(x) - g(x)\|_1$$

Fix $c \in \mathcal{C}$ and consider the difference for f and g in the smooth multicalibration constraint for c.

$$\begin{aligned}
&\left|\mathop{\mathbf{E}}_{\mathcal{D}}[\ c(\mathbf{x})\langle w(f(\mathbf{x})), \mathbf{y} - f(\mathbf{x})\rangle\ ] - \mathop{\mathbf{E}}_{\mathcal{D}}[\ c(\mathbf{x})\langle w(g(\mathbf{x})), \mathbf{y} - g(\mathbf{x})\rangle\ ]\right| \\
&= \left|\mathop{\mathbf{E}}_{\mathcal{D}}[\ c(\mathbf{x})\left(\langle w(f(\mathbf{x})), \mathbf{y} - f(\mathbf{x})\rangle - \langle w(g(\mathbf{x})), \mathbf{y} - g(\mathbf{x})\rangle\right)\ ]\right| \\
&\leq \max_{x\in\mathcal{X}} |c(x)| \cdot \left|\mathop{\mathbf{E}}_{\mathcal{D}}[\ \langle w(f(\mathbf{x})), \mathbf{y} - f(\mathbf{x})\rangle - \langle w(g(\mathbf{x})), \mathbf{y} - g(\mathbf{x})\rangle\ ]\right| \\
&\leq (2r+1)\mathop{\mathbf{E}}_{\mathcal{D}}[\ \|f(\mathbf{x}) - g(\mathbf{x})\|_1\ ] \\
&= (2r+1)\delta
\end{aligned}$$

where we use the fact that $|c(x)| \leq 1$ for all $c \in \mathcal{C}$. Since by assumption f is $(\mathcal{C}, \mathcal{W}, \alpha)$ multicalibrated, we get

$$\begin{aligned}
\left|\mathop{\mathbf{E}}_{\mathcal{D}}[\ c(\mathbf{x})\langle w(g(\mathbf{x})), \mathbf{y} - g(\mathbf{x})\rangle\ ]\right| &\leq \left|\mathop{\mathbf{E}}_{\mathcal{D}}[\ c(\mathbf{x})\langle w(f(\mathbf{x})), \mathbf{y} - f(\mathbf{x})\rangle\ ]\right| + (2r+1)\delta \\
&\leq \alpha + (2r+1)\delta.
\end{aligned}$$

This holds for all $c \in \mathcal{C}, w \in \mathcal{W}$, hence g is $(\mathcal{C}, \mathcal{W}, \alpha + (2r+1)\delta)$-multicalibrated. □

4 Moment Sandwiching from Low-Degree Multicalibration

In this section, we prove the key technical results, establishing moment sandwiching bounds for low-degree multicalibrated predictors. We begin by establishing the general theorem, then show various corollaries, for bounds on the confusion probabilities in multi-class prediction, and the correlation between the predicted values and the true outcomes.

We begin with some useful notational shorthand. Recall that $\mathcal{C} = \{c : \mathcal{X} \to [0,1]\}$. We denote the measure of c under $\mathcal{D}$ by $\mu_c = \mathbf{E}_{\mathcal{D}}[c(\mathbf{x})]$. Define the distribution $\mathcal{D}_c$ over $\mathcal{X} \times \{0,1\}$ obtained by conditioning on c by

$$\mathcal{D}_c(x,y) = \frac{c(x)\mathcal{D}(x,y)}{\mu_c}$$

For ease of notation, we will sometimes use $\alpha_c = \alpha/\mu_c$.

Lemma 4.1. *Let $k \geq 2$ and $f \in \mathrm{MC}_k(\alpha)$. For every degree $d \in [k]$, label $\ell \in [l]$ and $c \in \mathcal{C}$,*

$$\left| \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^{d-1} f^*_\ell(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d] \right| \leq \frac{\alpha}{\mu_c} = \alpha_c. \tag{8}$$

Proof. We consider the function $w : \Delta^l \to [0,1]^l$ where $w_\ell(f) = f_\ell^{d-1}$ and $w_j(f) = 0$ for $j \neq \ell$. It is easy to see that $w \in \mathcal{P}_{k-1}$. For this choice of w, degree k multicalibration implies

$$\left| \mathop{\mathbf{E}}_{\mathcal{D}}[c(\mathbf{x}) f_\ell(\mathbf{x})^{d-1} (\mathbf{y}_\ell - f_\ell(\mathbf{x}))] \right| \leq \alpha.$$

Switching to the conditional distribution $\mathcal{D}_c$, we can rewrite this as

$$\mu_c \left| \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^{d-1} (f^*_\ell(\mathbf{x}) - f_\ell(\mathbf{x}))] \right| \leq \alpha$$

since $\mathbf{E}_{[}\, \mathbf{y}_\ell | x] = f^*(x)$. Dividing both sides by μ_c and rearranging gives the desired bound. □

Note that this guarantee is meaningful only if μ_c is larger than α. This is to be expected, since we cannot hope to get strong conditional guarantees for sets that are very small.

Our goal is to show the following sandwiching bound which we state for the setting $\alpha = 0$ for clarity. For all $d \in [k], \ell \in [l], c \in \mathcal{C}$,:

$$\mathop{\mathbf{E}}_{\mathcal{D}_c}[f^*_\ell(\mathbf{x})^d] \geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d] = \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^{d-1} f^*_\ell(\mathbf{x})] \geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^{d-1}] \mathop{\mathbf{E}}_{\mathcal{D}_c}[f^*_\ell(\mathbf{x})].$$

The middle equality is by Lemma 4.1, which is immediate from the definition of degree-k multicalibration. The key ingredients are the outer inequalities. The lower bound can be interpreted as saying that f_ℓ^{d-1} is positively correlated with the label being ℓ, since

$$\mathrm{Cov}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^{d-1}, \mathbf{y}_\ell] = \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^{d-1} f^*_\ell(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^{d-1}] \mathop{\mathbf{E}}_{\mathcal{D}_c}[f^*_\ell(\mathbf{x})] \geq 0.$$

The upper bound can be seen as an upper bound *in lieu of exact moment matching*, it says that the first k moments of f_ℓ are dominated by those of the ground truth f^*_ℓ for every $\ell \in [l]$.

When $\alpha > 0$, the overall form of the inequalities stays the same, with some slack depending on α_c. But the two terms in the middle are only approximately equal. This makes it easier to state the bounds separately, which we do in Theorem 4.2 and Corollary 4.3.

Theorem 4.2 (Formal restatement of Theorem 2)**.** *Fix $k \geq 2$ and a predictor $f : \mathcal{X} \to \Delta_l$ where $f \in \mathrm{MC}_k(\alpha)$. For every degree $d \in [k]$, label $\ell \in [l]$ and $c \in \mathcal{C}$, the following sandwiching bound holds:*

$$d\frac{\alpha}{\mu_c} + \mathop{\mathbf{E}}_{\mathcal{D}_c}[f^*_\ell(\mathbf{x})^d] \geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d] \geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f^*_\ell(\mathbf{x})] \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^{d-1}] - \frac{\alpha}{\mu_c}. \tag{9}$$

Proof. We first prove Equation (9), starting from the upper bound on $\mathbf{E}[f_\ell(\mathbf{x})^d]$. We claim that

$$\mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d]^{\frac{d-1}{d}} \mathop{\mathbf{E}}_{\mathcal{D}_c}[f^*_\ell(\mathbf{x})^d]^{\frac{1}{d}} \geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^{d-1} f^*_\ell(\mathbf{x})] \geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d] - \alpha_c$$

The the upper bound is by Holder's inequality whereas the lower bound is by Equation (8). Dropping the term in the middle,

$$\mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d]^{\frac{d-1}{d}} \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})^d]^{\frac{1}{d}} \geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d] - \alpha_c$$

We may assume f is not identically 0, since otherwise the upper bound is trivial, and hence that its moments are positive. Divide both sides by $\mathbf{E}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d]$ to get

$$\left(\frac{\mathbf{E}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})^d]}{\mathbf{E}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d]}\right)^{\frac{1}{d}} \geq \left(1 - \frac{\alpha_c}{\mathbf{E}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d]}\right)$$

Raising both sides to the power d, and using $(1-\eta)^d \geq 1 - d\eta$ for $k \in \mathbb{Z}^+$ and $\eta > 0$

$$\frac{\mathbf{E}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})^d]}{\mathbf{E}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d]} \geq \left(1 - \frac{\alpha_c}{\mathbf{E}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d]}\right)^d \geq \left(1 - d\frac{\alpha_c}{\mathbf{E}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d]}\right)$$

Multiplying throughout by $\mathbf{E}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d]$

$$\mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})^d] \geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d] - d\alpha_c \tag{10}$$

which gives the desired upper bound for $d \in \{1, \ldots, k\}$.

We now prove the lower bound on $\mathbf{E}[f^d]$ in Equation (9). By multiaccuracy and convexity,

$$\mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})] \leq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})] + \alpha_c \leq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d]^{\frac{1}{d}} + \alpha_c$$
$$\mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^{d-1}] \leq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d]^{\frac{d-1}{d}}$$

Multiplying these together gives

$$\mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^{d-1}]\,\mathbf{E}[f_\ell^*(\mathbf{x})] \leq (\mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d]^{\frac{1}{d}} + \alpha_c)\mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d]^{\frac{d-1}{d}} \leq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d] + \alpha_c$$

where the last inequality uses the fact that $f_\ell \in [0,1]$. This completes the proof of Equation (9). □

Corollary 4.3. *Under the conditions of Theorem 4.2, the following sandwiching bound holds:*

$$(d+1)\frac{\alpha}{\mu_c} + \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})^d] \geq \quad \mathbf{E}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^{d-1} f_\ell^*(\mathbf{x})] \quad \geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})]\mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^{d-1}] - 2\frac{\alpha}{\mu_c}. \tag{11}$$

Proof. We start from the bound on $\mathbf{E}[f^d]$ in Equation (9) and use Equation (8) which implies that the bounds hold for $\mathbf{E}[f^{d-1}f^*]$ with an additional slack of α_c. □

4.1 Bounding the True Positive Rates

Imagine a binary prediction problem on a population $P = A \cup B$ where one half A has a 70% chance of having outcome $\mathbf{y} = 1$, and the other half B has only a 30% chance. When we consider this population in isolation, multiaccuracy is a very weak condition. In particular, the predictor $g : \mathcal{X} \to [0, 1]$ that predicts $g(\mathbf{x}) = 1$ on A, and $g(\mathbf{x}) = 0$ on B satisfies accuracy-in-expectation over P. This predictor is over-confident in its predictions; it does not give an accurate sense of the uncertainty in its predictions. A key property of calibration is that it disallows such overconfidence, by explicitly requiring that when $f(\mathbf{x}) \approx v$, then $\mathbf{E}[\mathbf{y}|f(\mathbf{x})] \approx v$.

A different way to quantify confidence is to look at $\mathbf{E}[f(\mathbf{x})|\mathbf{y} = 1]$ and $\mathbf{E}[f(\mathbf{x})|\mathbf{y} = 0]$. The latter has been termed the *generalized false positive rate*, so we refer to the former as the *generalized true positive rate*. Common sense suggests that we want a low false positive rate, and a high true positive rate. But how low a false positive rate is desirable? A quick calculation shows that the predictor g above has a generalized false positive rate of 0.3, whereas even the Bayes optimal predictor f^* has a higher false positive rate of 0.58! In general, only seeking to minimize the false positive rate and false negative rate risks preferring predictors that overstate their confidence, since it does not incentivize the predictor to convey the level of uncertainty in its predictions.

We give bounds on the true positive rates of any predictor in MC_2.

Definition 4.4. *For a predictor $f : \mathcal{X} \to \Delta_l$, label $\ell \in [l]$ and $c \in \mathcal{C}$, define the true positive rate for f on ℓ conditioned on $c \in \mathcal{C}$ to be $\tau_c(f, \ell) = \mathbf{E}_{\mathcal{D}_c}[f_\ell(\mathbf{x})|\mathbf{y}_\ell = 1]$.*

Lemma 4.5. *Every predictor $f \in \mathrm{MC}_2$ satisfies*

$$\frac{3\alpha}{\mathbf{E}_{\mathcal{D}_c}[f^*_\ell(\mathbf{x})]\mu_c} + \tau_c(f^*, \ell) \geq \tau_c(f, \ell) \geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})] - \frac{2\alpha}{\mathbf{E}_{\mathcal{D}_c}[f^*_\ell(\mathbf{x})]\mu_c}$$

The lower bound says the true positive rate for label ℓ is at least as much as the overall positive rate for that label in the population $\mathcal{C}$. This is equivalent to positive correlation between $f_\ell(\mathbf{x})$ and $\mathbf{y}_\ell$ conditioned on c. The upper bound asserts that the correlation is not exaggerated, being upper bounded by the true positive rate of the Bayes optimal predictor. Our proof relies on the following characterization of true positive rates:

Lemma 4.6. *We have*

$$\tau_c(f, \ell) = \frac{\mathbf{E}_{\mathcal{D}_c}[f_\ell(\mathbf{x})f^*_\ell(x)]}{\mathbf{E}_{\mathcal{D}_c}[f^*_\ell(\mathbf{x})]}$$

Proof. By the definition of τ_c,

$$\begin{aligned}
\tau_c(f, \ell) &= \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})|\mathbf{y}_\ell = 1] \\
&= \frac{\sum_x \mathcal{D}_c(\mathbf{x} = x \wedge \mathbf{y}_\ell = 1)f_\ell(x)}{\sum_x \mathcal{D}_c(\mathbf{x} = x \wedge \mathbf{y}_\ell = 1)} \\
&= \frac{\sum_x \mathcal{D}_c(\mathbf{x} = x)f^*_\ell(x)f_\ell(x)}{\sum_x \mathcal{D}_c(\mathbf{x} = x)f^*_\ell(x)} \\
&= \frac{\mathbf{E}_{\mathcal{D}_c}[f_\ell(\mathbf{x})f^*_\ell(\mathbf{x})]}{\mathbf{E}_{\mathcal{D}_c}[f^*_\ell(\mathbf{x})]}.
\end{aligned}$$

□

Proof of Lemma 4.5. Equation (11) for $d = 2$ gives

$$3\frac{\alpha}{\mu_c} + \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})^2] \geq \quad \mathbf{E}_{\mathcal{D}_c}[f_\ell(\mathbf{x})f_\ell^*(\mathbf{x})] \quad \geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})] \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})] - 2\frac{\alpha}{\mu_c}. \tag{12}$$

The claimed bounds follow by dividing throughout by $\mathbf{E}[f_\ell^*]$ and observing that by Lemma 4.6,

$$\tau_c(f, \ell) = \frac{\mathbf{E}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})f_\ell(\mathbf{x})]}{\mathbf{E}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})]}$$
$$\tau_c(f^*, \ell) = \frac{\mathbf{E}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})^2]}{\mathbf{E}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})]}.$$

□

We can define higher moment analogues of the true positive rate by considering the quantity $\mathbf{E}[f_\ell(\mathbf{x})^d|\mathbf{y}_\ell = 1]$ for $d \geq 2$; for which we prove the following bound:

Lemma 4.7. *For every $d \in [k-1]$ and $c \in \mathcal{C}$, every $f \in \mathrm{MC}_k$ satisfies the following bounds :*

$$\frac{(d+1)\alpha}{\mathbf{E}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})]\mu_c} + \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})^d|\mathbf{y}_\ell = 1] \geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d|\mathbf{y}_\ell = 1] \geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^d] - \frac{2\alpha}{\mathbf{E}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})]\mu_c}$$

4.2 Bounds for the Confusion Matrix

We generalize the discussion of error rates to the multi-class setting. Here, we are concerned with the confusion matrix, whose ijth entry corresponds to the probability of predicting j when the true label is i. In this setting, we think about sampling a predicted label $\mathbf{z}$ according to the prediction $f(\mathbf{x})$; this bears similarity to the model of outcome indistinguishability. With some manipulation, it is not hard to see that the collision and confusion rates can be audited directly within the OI framework [DKR+21, DKR+22] to test the closeness of the matrices as in Lemma 4.8.

Given a distribution $\mathcal{D}$ on $\mathcal{X} \times [l]$ and a predictor $f : \mathcal{X} \to \Delta_l$, let us assume that the predictor f generates a label $\mathbf{z} \in [l]$ where $\Pr[\mathbf{z} = j|x] = f_j(x)$. The $l \times l$ confusion matrix has entries $b_{ij} = \Pr_{\mathcal{D}}[\mathbf{y} = i \wedge \mathbf{z} = j]$ for $i, j \in [l]$. It is easy to see that

$$\Pr_{\mathcal{D}}[\mathbf{y} = i \wedge \mathbf{z} = j] = \mathop{\mathbf{E}}_{\mathcal{D}}[f_i^*(\mathbf{x})f_j(\mathbf{x})].$$

Associating points in Δ_l with column vectors in $\mathbb{R}^l$,and using f^T to denote the transpose of f, we can define the confusion matrix as

$$B_{\mathcal{D}}(f^*, f) = \mathop{\mathbf{E}}_{\mathcal{D}}[f^*(\mathbf{x})f(\mathbf{x})^T].$$

Define the max norm of a matrix $A = \{a_{ij}\}_{i,j\in[l]}$ by $\|A\|_{\max} = \max_{i,j\in[l]} |a_{ij}|$.

Lemma 4.8. *Let $f \in \mathrm{MC}_2(\alpha)$ and $c \in \mathcal{C}$. Then we have*

$$\|B_{\mathcal{D}_c}(f^*, f) - B_{\mathcal{D}_c}(f, f)\|_{\max} \leq \frac{\alpha}{\mu_c}.$$

Proof. For $i, j \in [l]$, define the function $w : \Delta \to [0,1]^l$ by $w_i(f) = f_j$ and $w_{i'}(f) = 0$ for $i' \neq i$. Clearly $w \in \mathcal{P}_1$, hence by the definition of degree-2 multicalibration applied to w,

$$\mu_c \left| \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_j(\mathbf{x})(\mathbf{y}_i - f_i(\mathbf{x}))] \right| \leq \alpha.$$

We can rewrite this condition as

$$\left| \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_i^*(\mathbf{x}) f_j(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_i(\mathbf{x}) f_j(\mathbf{x})] \right| \leq \alpha_c$$

which implies the claim. □

Next we show that the confusion matrix for f is dominated by that for f^* in the PSD order.

Lemma 4.9. *For any $f \in \mathrm{MC}_2(\alpha)$ and $c \in \mathcal{C}$, we have*

$$B(f, f) \preccurlyeq B(f^*, f^*) + 2l\alpha \cdot I_{l \times l}.$$

Proof. Fix a unit vector $u \in \mathbb{R}^\ell$. By the definition of the PSD order, it suffices to show that

$$\mathop{\mathbf{E}}_{\mathcal{D}_c}[(u^T f(\mathbf{x}))^2] \leq \mathop{\mathbf{E}}_{\mathcal{D}_c}[(u^T f^*(\mathbf{x}))^2] + 2l\alpha.$$

We have

$$\begin{aligned}
\mathbf{E}[(u^T f(\mathbf{x}))^2] &= \sum_{ij} u_i u_j \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_i(\mathbf{x}) f_j(\mathbf{x})] \\
&\leq \sum_{ij} u_i u_j \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_i^*(\mathbf{x}) f_j(\mathbf{x})] + \alpha \sum_{ij} u_i u_j \\
&\leq \mathop{\mathbf{E}}_{\mathcal{D}_c}[(u^t f^*(\mathbf{x}))(u^t f(\mathbf{x}))] + l\alpha \\
&\leq \mathop{\mathbf{E}}_{\mathcal{D}_c}[(u^t f^*(\mathbf{x}))^2]^{1/2} \mathop{\mathbf{E}}_{\mathcal{D}_c}[(u^t f(\mathbf{x}))^2]^{1/2} + l\alpha
\end{aligned}$$

Dividing both sides by $\mathbf{E}_{\mathcal{D}_c}[(u^T f(\mathbf{x}))^2]$ we get

$$\frac{\mathbf{E}_{\mathcal{D}_c}[(u^T f^*(\mathbf{x}))^2]^{1/2}}{\mathbf{E}_{\mathcal{D}_c}[(u^T f(\mathbf{x}))^2]^{1/2}} \geq 1 - \frac{l\alpha}{\mathbf{E}_{\mathcal{D}_c}[(u^T f(\mathbf{x}))^2]}$$

Squaring and using $(1-\eta)^2 \geq 1 - 2\eta$ gives

$$\mathop{\mathbf{E}}_{\mathcal{D}_c}[(u^T f^*(\mathbf{x}))^2] \geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[(u^T f(\mathbf{x}))^2] - 2l\alpha,$$

which implies the desired bound. □

As an application, we can generalize our bounds on true positive rates to allow combinations of labels. As motivation, consider a model trained to assign images from a large label set $[l]$. We might want to know how well the model does on the set of cat images, where cats correspond to a subset $L \subseteq [l]$. In analogy with the true positive rate on label, we could measure the probability that the image belongs to L, and the predicted label is also in L, without distinguishing between labels within L. This motivates our next definition. We use the notation $f_L(x) = \sum_{\ell \in L} f_\ell(x)$.

Definition 4.10. *For a predictor $f : \mathcal{X} \to \Delta_l$, set $L \subseteq [l]$ and $c \in \mathcal{C}$, define the true positive rate for f on L conditioned on $c \in \mathcal{C}$ to be*

$$\tau_c(f, L) = \mathop{\mathbf{E}}_{\mathcal{D}_c}\Big[f_L(\mathbf{x}) | \sum_{l \in L} \mathbf{y}_\ell = 1\Big].$$

We present the following generalization of Lemma 4.5:

Lemma 4.11. *For a subset $L \subseteq L$ of labels, every predictor $f \in \mathrm{MC}_2$ satisfies*

$$\frac{2l|L|^2\alpha}{\mathbf{E}_{\mathcal{D}_c}[f_L(\mathbf{x})]\mu_c} + \tau_c(f^*, L) \geq \tau_c(f, L) \geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_L(\mathbf{x})] - \frac{(|L|^2 + |L|)\alpha}{\mathbf{E}_{\mathcal{D}_c}[f^*_L(\mathbf{x})]\mu_c}.$$

Proof. In analogy with Lemma 4.6, we can show that

$$\tau_c(f, L) = \frac{\mathbf{E}_{\mathcal{D}_c}[f_L(\mathbf{x}) f^*_L(x)]}{\mathbf{E}_{\mathcal{D}_c}[f^*_L(\mathbf{x})]} = \frac{\mathbb{1}_L^T B_{\mathcal{D}_c}(f^*, f)\mathbb{1}_L}{\mathbf{E}_{\mathcal{D}_c}[f^*_L(\mathbf{x})]}.$$

Using Lemma 4.8, we have

$$\begin{aligned}
\mathbb{1}_L^T B_{\mathcal{D}_c}(f^*, f)\mathbb{1}_L &\geq \mathbb{1}_L^T B_{\mathcal{D}_c}(f, f)\mathbb{1}_L - |L|^2\alpha_c \\
&= \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_L(\mathbf{x})^2] - |L|^2\alpha_c \\
&\geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_L(\mathbf{x})]^2 - |L|^2\alpha_c \\
&\geq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_L(\mathbf{x})] \mathop{\mathbf{E}}_{\mathcal{D}_c}[f^*_L(\mathbf{x})] - (|L|^2 + |L|)\alpha_c
\end{aligned}$$

where the last line used the following consequence of multiaccuracy

$$\left|\mathop{\mathbf{E}}_{\mathcal{D}_c}[f_L(\mathbf{x}) - f^*_L(\mathbf{x})]\right| \leq |L|\alpha_c$$

Lemma 4.9 implies an upper bound of

$$\begin{aligned}
\mathbb{1}_L^T B_{\mathcal{D}_c}(f^*, f)\mathbb{1}_L &\leq \mathbb{1}_L^T B_{\mathcal{D}_c}(f^*, f^*)\mathbb{1}_L + 2l|L|^2\alpha_c \\
&\leq \mathop{\mathbf{E}}_{\mathcal{D}_c}[f^*_L(\mathbf{x})^2] + 2l|L|^2\alpha + c.
\end{aligned}$$

The claim now follows by dividing throughout by $\mathbf{E}_{\mathcal{D}_c}[f^*_L(\mathbf{x})]$ □

4.3 Covariance Guarantees for Degree-2 Multicalibration

We present a detailed analysis of the covariance between the predictions $f_\ell(\mathbf{x})$ and label being ℓ conditioned on c. As one might expect, the ground truth predictor is indeed positively correlated with the labels. Multiaccuracy guarantees that the expectations of $f_\ell(\mathbf{x})$ and $\mathbf{y}_\ell$ are equal conditioned on c, but this need not imply that they are positively correlated. Our main result of this section is MC_2 guarantees positive correlation, and that the correlation is conservative compared with the ground truth predictor. We also provide an example showing that the correlation can indeed be negative for $f \in \mathrm{MC}_1$.

Theorem 4.12. *For $f \in \mathrm{MC}_2(\alpha)$ and $c \in \mathcal{C}$, we have*

$$\mathrm{Var}_{\mathcal{D}_c}[f_\ell(\mathbf{x})] - 2\alpha_c \le \mathrm{Cov}_{\mathcal{D}_c}[f_\ell(\mathbf{x}), \mathbf{y}_\ell] \le \mathrm{Cov}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x}), \mathbf{y}_\ell] + 6\alpha_c. \tag{13}$$

Our key technical lemma uses these to show that conditioned on any $c \in \mathcal{C}$, every predictor in MC_2 has variance not much larger than the Bayes optimal predictor. Its proof will use the following properties every $f \in \mathrm{MC}_2(\alpha)$ satisfies by Equation (8).

$$\left|\mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathcal{D}_c}[\mathbf{y}]\right| \le \alpha_c, \tag{14}$$

$$\left|\mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^2] - \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x}) f_\ell^*(\mathbf{x})\right| \le \alpha_c. \tag{15}$$

Lemma 4.13. *For $f \in \mathrm{MC}_2$ and $c \in \mathcal{C}_1$,*

$$\mathrm{Var}_{\mathcal{D}_c}[f_\ell(\mathbf{x})] \le \mathrm{Var}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})] + 4\alpha_c. \tag{16}$$

$$\mathrm{Cov}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})\mathbf{y}] = \mathrm{Var}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})]. \tag{17}$$

$$|\mathrm{Cov}_{\mathcal{D}_c}[f_\ell(\mathbf{x})\mathbf{y}] - \mathrm{Var}_{\mathcal{D}_c}[f_\ell(\mathbf{x})]| \le 2\alpha_c. \tag{18}$$

Proof. We observe that by Equation (10) with $d = 2$,

$$\mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})^2] \ge \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})^2] - 2\alpha_c. \tag{19}$$

By Equation (14) we have

$$\left|\mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})]^2 - \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})^2]\right| \le \left|\mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})] + \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(x)]\right| \left|\mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(x)]\right| \le 2\alpha_c.$$

Applying this together with Equation (19) gives

$$\begin{aligned}\mathrm{Var}_{\mathcal{D}_C}[f_\ell^*(\mathbf{x})] - \mathrm{Var}_{\mathcal{D}_c}[f_\ell(\mathbf{x})] &\ge \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})^2] - \mathop{\mathbf{E}}_{\mathcal{D}_C}[f_\ell(\mathbf{x})^2] + \mathop{\mathbf{E}}_{\mathcal{D}_C}[f_\ell(\mathbf{x})]^2 - \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})]^2 \\ &\ge -2\alpha_c - 2\alpha_c = -4\alpha_c\end{aligned}$$

which implies the desired bound.

By definition of f^*, $\mathbf{E}[\mathbf{y}|\mathbf{x}] = f_\ell^*(\mathbf{x})$. Hence

$$\mathrm{Cov}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})\mathbf{y}] = \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})\mathbf{y}] - \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})]\mathop{\mathbf{E}}_{\mathcal{D}}[\mathbf{y}] = \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})^2] - \mathop{\mathbf{E}}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})]^2 = \mathrm{Var}_{\mathcal{D}_c}[f_\ell^*(\mathbf{x})]. \tag{20}$$

By Equation (15),

$$\left|\mathbf{E}[f_\ell(\mathbf{x})\mathbf{y}] - \mathbf{E}[f_\ell(\mathbf{x})^2]\right| \le \alpha_c \tag{21}$$

Using Equation (14) we have

$$\left|\mathbf{E}[f_\ell(\mathbf{x})]\,\mathbf{E}[\mathbf{y}] - \mathbf{E}[f_\ell(\mathbf{x})]^2\right| \le \mathbf{E}[f_\ell(\mathbf{x})\,|\mathbf{E}[f_\ell(\mathbf{x}) - \mathbf{y}]| \le \alpha_c\,\mathbf{E}[f_\ell(\mathbf{x})] \le \alpha_c. \tag{22}$$

Finally, using the definitions of variance and covariance, we have

$$|\mathrm{Cov}[f_\ell(\mathbf{x}), \mathbf{y}] - \mathrm{Var}[f_\ell(\mathbf{x})]| \le \left|\mathbf{E}[f_\ell(\mathbf{x})\mathbf{y}] - \mathbf{E}[f_\ell(\mathbf{x})^2]\right| + \left|\mathbf{E}[f_\ell(\mathbf{x})]\,\mathbf{E}[\mathbf{y}] - \mathbf{E}[f_\ell(\mathbf{x})]^2\right| \le 2\alpha_c$$

where we use Equations (21) and (22). □

Proof of Theorem 4.2. The lower bound is an immediate consequence of Equation (18). The upper bound follows since

$$\mathrm{Cov}_{\mathcal{D}_c}[f_\ell(\mathbf{x}), y] \leq \mathrm{Var}_{\mathcal{D}_c}[f_\ell(\mathbf{x})] + 2\alpha_c \leq \mathrm{Var}_{\mathcal{D}_c}[f^*_\ell(x)] + 6\alpha_c = \mathrm{Cov}_{\mathcal{D}_c}[f^*_\ell(x), y] + 6\alpha_c$$

where we use Equations (18), (16) and (17) respectively. □

We complement this by an example showing that the correlation between $f_\ell(\mathbf{x})$ and $\mathbf{y}_\ell$ can be negative for $f \in \mathrm{MA} = \mathrm{MC}_1$. In particular, we show that this is true even when $\mathbf{y}_\ell$ is obtained by standard methods such as least-squres or logistic regression (on linear combinations from $\mathcal{C}$).

Lemma 4.14. *There exist a distribution $\mathcal{D}$ on $\{0,1\}^2 \times \{0,1\}$, a constraint set $\mathcal{C}$, a constraint c, and a label $\ell \in \{0,1\}$, such that if $f \in \mathrm{MA}$ is obtained by least-squares or logistic regression, then*

$$\mathrm{Cov}_{D_c}[f_\ell(\mathbf{x}), \mathbf{y}_\ell] = -1/12 < 0.$$

The example we provide is in the binary classification setting, with $\mathcal{D} = \{0,1\}^2$ and the sets $\mathcal{C}$ all being edges (that is, the sets $x_1 = 0$, $x_1 = 1$, $x_2 = 0$, and $x_2 = 1$). The covariance we achieve is $-1/12$, though we note that it can be made arbitrarily close to $-1/4$, the minimum possible covariance of two $[0,1]$ random variables. The full example is provided in Appendix A.

5 The Complexity of Low-Degree Multicalibration

Here, we establish upper bounds on the time and sample complexity for obtaining low-degree multicalibration. We begin by describing a completely generic multicalibration algorithm that works for any class of weight functions $\mathcal{W}$ and for any number of class labels $k \in \mathbb{N}$. Importantly, following [HKRR18, KGZ19], the algorithm reduces the task of learning a weighted multicalibrated predictor to the task of *weak agnostic learning* the class $\mathcal{C}$. We analyze the algorithm in terms of its oracle-efficiency, assuming access to a weak agnostic learner.

With upper bounds on the complexity of learning a multicalibrated predictor using a generic weight class $\mathcal{W}$, we instantiate the bounds for the low-degree, smooth, and indicator variants of multicalibration. We show that, for meaningful settings of the parameters, low-degree multicalibration is considerably more sample efficient than the original formulation of multicalibration. This effect is particularly pronounced as the number of class labels l grows.

5.1 Learning Weighted Multicalibrated Predictors

In Algorithm 1, we describe a procedure for learning multicalibrated predictors. The algorithm assumes oracle-access to a weak agnostic learner [KMV08, Fel09].

Definition 5.1 (Weak Agnostic Learning)**.** *For a data distribution $\mathcal{D}$ supported on $\mathcal{X} \times [-1,1]$, a weak agnostic learner for a hypothesis class $\mathcal{C} \subseteq \{c : \mathcal{X} \to [0,1]\}$ is a learning procedure that takes labeled data $D = \{(x_i, z_i)\}_{i=1}^m$, where each sample $(\mathbf{x}, \mathbf{z}) \sim \mathcal{D}$. For $\alpha > 0$, the learning procedure returns an element of $\mathcal{C} \cup \{\perp\}$*

$$c \leftarrow \mathsf{WAL}_{\mathcal{C},\alpha}(D)$$

satisfying the following properties.

(1) if there exists $c' \in \mathcal{C}$ such that $\mathbf{E}_{\mathcal{D}}[\, c'(\mathbf{x}) \cdot \mathbf{z} \,] \notin [-\alpha, \alpha]$, then $c \neq \perp$ and $\mathbf{E}_{\mathcal{D}}[\, c(\mathbf{x}) \cdot \mathbf{z} \,] \geq \alpha/2$.

(2) if $c = \perp$, then for all $c' \in \mathcal{C}$, $\mathbf{E}_{\mathcal{D}}[\, c'(\mathbf{x}) \cdot \mathbf{z} \,] \in [-\alpha, \alpha]$.

We say the sample complexity of the weak agnostic learner, $m = m(\mathcal{C}, \alpha, \beta)$, is the number of samples from $\mathcal{D}$ necessary to guarantee properties (1) and (2) with probability at least $1 - \beta$.

Intuitively, a weak agnostic learner searches for some $c \in \mathcal{C}$ that correlates nontrivially with the labels given by z.[8] With this definition in place, we can describe the Weighted Multicalibration algorithm and state its guarantees.

Algorithm 1 Weighted Multicalibration
Input: training data $\{(x_i, y_i)\}_{i=1}^m$
Concept class $\mathcal{C} \subseteq \{c : \mathcal{X} \to [0,1]\}$,
Weight class $\mathcal{W} \subseteq \left\{w : \Delta_l \to [0,1]^l\right\}$,
approximation $\alpha > 0$,
step size η
Output: $(\mathcal{C}, \mathcal{W}, \alpha)$-multicalibrated predictor $f : \mathcal{X} \to [0,1]^l$

$f_0(\cdot) \leftarrow (1/2, \ldots, 1/2) \in [0,1]^l$
$mc \leftarrow$ false
$t \leftarrow 0$
while $\neg mc$ **do**
  $mc \leftarrow$ true
  for $w \in \mathcal{W}$ **do**
    $c_{t+1} \leftarrow \mathsf{WAL}_{\mathcal{C},\alpha}\left(\left\{\left(x_i, \langle w(f_t(x_i)), y_i - f_t(x_i)\rangle\right)\right\}_{i=1}^m\right)$
    if $c_{t+1} = \perp$ **then**
      continue
    else
      $\Delta_{t+1}(\cdot) \leftarrow w(f_t(\cdot)) \cdot c_{t+1}(\cdot)$
      $f_{t+1}(\cdot) \leftarrow \pi_{[0,1]^l}\left(f_t(\cdot) + \eta \cdot \Delta_{t+1}(\cdot)\right)$ *// where $\pi_{[0,1]^l}$ projects onto $[0,1]^l$*
      $mc \leftarrow$ false
      $t \leftarrow t + 1$
      break
    end if
  end for
end while
return f_t

The algorithm is an iterative boosting-style procedure. We initialize the hypothesis to be the constant function $f_0(\mathbf{x}) = (1/2, \ldots, 1/2) \in [0,1]^l$. Then, in the tth iteration, for each $w \in \mathcal{W}$ we reduce the problem of searching for some $c \in \mathcal{C}$ where f_t is miscalibrated to the problem of weak agnostic learning. If we find some $c \in \mathcal{C}$ such that

$$\mathbf{E}[\, c(\mathbf{x})\langle w(f_t(\mathbf{x})), \mathbf{y} - f_t(\mathbf{x})\rangle \,] > \alpha$$

[8] Our analysis does not assume that WAL is a proper learner. All of the results hold equally for improper weak agnostic learners.

then we can use $c(\mathbf{x}) \cdot w(f_t(\mathbf{x}))$ to update the predictor to be better calibrated in this direction. If for all $w \in \mathcal{W}$ we fail to find any $c \in \mathcal{C}$ that correlates with the residual, then we return the current hypothesis. We describe the procedure in Algorithm 1.

Analysis of the algorithm. The exact running time of the algorithm depends intimately on the model of computation and the time complexity of weak agnostic learning, which for most classes $\mathcal{C}$ will dominate the time complexity. With this in mind, we bound the iteration complexity T of the algorithm, noting that each iteration makes at most $|\mathcal{W}|$ calls to $\mathsf{WAL}_{\mathcal{C},\alpha}$, which results in a time complexity bounded by $T \cdot |\mathcal{W}|$ times the complexity of weak agnostic learning $\mathcal{C}$.

First, we argue correctness—that if the algorithm terminates, then the returned hypothesis satisfies multicalibration.

Lemma 5.2. *If Algorithm 1 returns a hypothesis $f : \mathcal{X} \to [0,1]^l$, then f is $(\mathcal{C}, \mathcal{W}, \alpha)$-multicalibrated.*

Proof. Observe that Algorithm 1 only returns a hypothesis f_t if, in the tth iteration, for every $w \in \mathcal{W}$, the call to the weak agnostic learner $\mathsf{WAL}_{\mathcal{C},\alpha}$ returns $\perp$. By the weak agnostic learning property (2), returning $\perp$ in every call indicates that for all $w \in \mathcal{W}$ and for all $c \in \mathcal{C}$, the correlation between c and the weighted residual is bounded by α in magnitude.

$$\mathbf{E}\left[\ c(\mathbf{x}) \cdot \langle w(f_t(\mathbf{x})), \mathbf{y} - f_t(\mathbf{x}) \rangle\ \right] \in [-\alpha, \alpha]$$

By definition, this means that f_t is $(\mathcal{C}, \mathcal{W}, \alpha)$-multicalibrated. □

Next, we argue that the number of iterations that the algorithm ever runs for is bounded polynomially in l and $1/\alpha$.

Lemma 5.3. *Algorithm 1 returns f_T after $T \leq l/\alpha^2$ iterations.*

Proof. The iteration complexity follows by a potential argument. Using the expected squared error as a potential function, we lower bound the progress at each iteration. Specifically, we use the following potential function,

$$\phi(f) = \mathbf{E}\left[\ \|f^*(\mathbf{x}) - f(\mathbf{x})\|^2\ \right].$$

By the assumption that $f^* : \mathcal{X} \to [0,1]^l$ and our choice of $f_0(\mathbf{x})_i = 1/2$ for all $i \in [l]$, the initial potential value is at most $\phi(f_0) \leq l/4$.

Consider the change in potential after the tth update.

$$\begin{aligned}
&\mathbf{E}\left[\ \|f^*(\mathbf{x}) - f_t(\mathbf{x})\|^2\ \right] - \mathbf{E}\left[\ \|f^*(\mathbf{x}) - f_{t+1}(\mathbf{x})\|^2\ \right] \\
&= \mathbf{E}\left[\ \|f^*(\mathbf{x}) - f_t(\mathbf{x})\|^2\ \right] - \mathbf{E}\left[\ \|f^*(\mathbf{x}) - f_t(\mathbf{x}) - \eta \cdot \Delta_{t+1}(\mathbf{x})\|^2\ \right] \\
&= 2\eta \cdot \mathbf{E}\left[\ \langle f^*(\mathbf{x}) - f_t(\mathbf{x}), \Delta_{t+1}(\mathbf{x}) \rangle\ \right] - \eta^2 \cdot \mathbf{E}\left[\ \|\Delta_{t+1}(\mathbf{x})\|^2\ \right] \\
&\geq 2\eta \cdot \mathbf{E}\left[\ \langle f^*(\mathbf{x}) - f_t(\mathbf{x}), w(f_t(\mathbf{x})) \cdot c_{t+1}(\mathbf{x}) \rangle\ \right] - \eta^2
\end{aligned}$$

By the weak agnostic learning property (1), we know that if c_{t+1} is the output from $\mathsf{WAL}_{\mathcal{C},\alpha}$ with the given training data, c_{t+1} has nontrivial correlation with the labels.

$$\mathbf{E}\left[\ \langle f^*(\mathbf{x}) - f_t(\mathbf{x}), w(f_t(\mathbf{x})) \cdot c_{t+1}(\mathbf{x})\rangle\ \right] = \mathbf{E}\left[\ c_{t+1}(\mathbf{x}) \cdot \langle w(f_t(\mathbf{x})), \mathbf{y} - f_t(\mathbf{x})\rangle\ \right] \geq \alpha/2$$

Plugging this bound into the last inequality and taking $\eta = \alpha/2$, the progress in ϕ in each iteration is at least $\alpha^2/4$. By the initial bound on $\phi(f_0)$, the total number of iterations is upper bounded by $T \leq l/\alpha^2$. □

We upper bound the sample complexity necessary to run Algorithm 1 in terms of the number of iterations, the cardinality of the weight class $\mathcal{W}$, and the sample complexity of the weak agnostic learner for $\mathcal{C}$. Note that in the tth iteration, for each $w \in \mathcal{W}$, we assign x_i a label that depends on f_t. This dependence on prior hypotheses (and thus prior access to the data), results in an adaptive data analysis problem. Naively, we can handle this by resampling at each iteration. We obtain the following generic bound.

Proposition 5.4. *For a hypothesis class $\mathcal{C}$ and approximation parameter $\alpha_0 > 0$, the sample complexity m to run Algorithm 1 with success probability at least $1 - \beta$ is upper bounded by*

$$m \leq O\left(\frac{l \cdot m(\mathcal{C}, \alpha_0, \beta_0)}{\alpha_0^2}\right)$$

where $m(\mathcal{C}, \alpha_0, \beta_0)$ is the sample complexity of running $\mathsf{WAL}_{\mathcal{C},\alpha_0}$ with failure probability $\beta_0 \leq \frac{\alpha_0^2 \beta}{l \cdot |\mathcal{W}|}$.

Proof. The upper bound follows by using a fresh sample for each iteration. We leverage the upper bound on the number of iterations necessary from Lemma 5.3, $T \leq l/\alpha_0^2$. Then, to obtain an overall failure probability of β, we take β_0 small enough that we can union bound the failure probability of $\mathsf{WAL}_{\mathcal{C},\alpha_0}$ over $T \cdot |\mathcal{W}|$ calls. Again, leveraging the bound on T, we bound $\beta_0 \leq \frac{\alpha_0^2 \beta}{l \cdot |\mathcal{W}|}$. □

Using Proposition 5.4, we obtain a concrete upper bound on the sample complexity based on specifying a weak agnostic learner and a class of weight functions $\mathcal{W}$. For instance, if $\mathcal{C}$ is a finite class, the weak agnostic learner that iterates over $\mathcal{C}$ and evaluates the correlation with labels as a statistical query obtains sample complexity $\log(|\mathcal{C}|/\beta_0)/\alpha_0^2$, for an overall sample complexity of

$$m \leq O\left(\frac{l \cdot \log(l\,|\mathcal{C}|\,|\mathcal{W}|/\alpha_0\beta)}{\alpha_0^4}\right).$$

For VC classes $\mathcal{C}$ that have a weak learner that obtains optimal sample complexity, $m(\mathcal{C}, \alpha_0, \beta_0) \leq (\mathsf{VC}(\mathcal{C}) + \log(1/\beta_0))/\alpha_0^2$, the overall bound goes as

$$m \leq O\left(\frac{l \cdot (\mathrm{VC}(\mathcal{C}) + \log(l\,|\mathcal{W}|/\alpha_0\beta))}{\alpha_0^4}\right). \tag{23}$$

Better analyses. Improved sample complexity analyses are possible for specialized implementations of the weak agnostic learner. Following [HKRR18, Kim20], we can avoid some of the cost of resampling by appealing to generalization guarantees for differentially-private learning algorithms [DMNS06]. Note that Algorithm 1 only touches the data through the weak agnostic learner,

in order to search for a violated constraint for some $c \in \mathcal{C}$. By implementing this search step under differential privacy, we can appeal to the results of [DFH+15, BNS+16, JLN+19], demonstrating that such algorithms guarantee statistical generalization, even under adaptive access to the data. For instance, using a bound from Corollary 6.4 of [BNS+16], in the case where $\mathcal{C}$ is a finite class, we can actually bound total the sample complexity as follows.

$$m \leq O\left(\frac{l^{1/2} \cdot \log(|\mathcal{C}|\,|\mathcal{W}|\,/\alpha) \cdot \log(1/\alpha\beta)^{3/2}}{\alpha^3}\right)$$

This bound follows by viewing each iteration as an optimization over the simultaneous choice of $w \in \mathcal{W}$ and $c \in \mathcal{C}$ to maximize the multicalibration violation. While this approach improves the sample complexity, computationally it requires exhaustive search over the choice of $c \in \mathcal{C}$ and $w \in \mathcal{W}$ to execute the exponential mechanism [MT07].

5.2 Comparing the Sample Complexity Across Notions

As a final comparison, we instantiate the general bound from Proposition 5.4 using the different weight classes $\mathcal{W}$, corresponding to low-degree, smooth, and full multicalibration. To make a fair comparison, we instantiate each bound using the approximation parameter α_0 that is required to guarantee the returned hypothesis satisfies α-degree-$(k+1)$ multicalibration for some $k \in \mathbb{N}$. In other words, for smooth and full multicalibration, we upper bound the sample complexity using the best choice of α_0 known to guarantee the multicalibrated predictor f is degree-$(k+1)$ mutlicalibrated; that is, in every case $f \in \mathrm{MC}_{k+1}(\alpha)$.

While many comparisons of this form could be made based on the properties of $\mathcal{C}$, we assume that $\mathcal{C}$ is a VC class and that $\mathsf{WAL}_{\mathcal{C},\alpha}$ obtains optimal sample complexity. Thus, we instantiate the concrete bound given in (23). Note that, by the relevant choices of α_0, any increase in the sample complexity's dependence on α_0 will *increase* the gap in sample complexities. In this sense, assuming a sample-optimal weak agnostic learner gives a conservative estimate on the gap.

Theorem 5.5 (Formal restatement of Theorem 3)**.** *Suppose $\mathcal{C}$ has a weak agnostic learner with sample complexity*

$$m(\mathcal{C}, \alpha_0, \beta_0) \leq \frac{\mathsf{VC}(\mathcal{C}) + \log(1/\beta_0)}{\alpha_0^2}$$

to obtain desired accuracy α_0 over $\mathcal{C}$ with all but probability β_0. Then, for any $k \in \mathbb{N}$ and any failure probability $\beta > 0$, there exists an implementation of Algorithm 1 to obtain the variants of multicalibration, obtaining sample complexity as follows.

- (Low-Degree).

$$m_k \leq O\left(\frac{l \cdot (\mathsf{VC}(\mathcal{C}) + k \cdot \log(l/\alpha\beta))}{\alpha^4}\right)$$

  to obtain $(\mathcal{C}, \alpha)$-degree-$(k+1)$ multicalibration.

- (Smooth).

$$m_s \le O\left(k^4 l \cdot \left(\frac{\mathsf{VC}(\mathcal{C}) + \log(kl/\alpha\beta)}{\alpha^4} + \frac{(kl)^{l-1}\,\mathrm{poly}(l, \log(k/\alpha))}{\alpha^{l+3}}\right)\right)$$

to obtain $(\mathcal{C}, \alpha/k)$*-smooth multicalibration.*

- (Full).

$$m_i \le O\left(\frac{(2kl)^{4(l+1)} \cdot (\mathsf{VC}(\mathcal{C}) + \log(kl/\alpha\beta))}{\alpha^{4(l+1)}}\right)$$

to obtain $(\mathcal{C}, \alpha_0)$*-full multicalibration for* $\alpha_0 \le (\alpha/2kl)^{l+1}$.

In the case of binary prediction, the full multicalibration bound can be improved to

$$m_{i,\mathrm{bin}} \le O\left(\frac{k^4 \cdot (\mathsf{VC}(\mathcal{C}) + \log(k/\alpha\beta))}{\alpha^8}\right).$$

Proof. The proof instantiates the bound in (23) with an appropriate weight class to achieve the desired notions.

(Low-Degree). To guarantee that the calibration constraint is satisfied for all $w \in \mathcal{W}_{k+1}$, we use a discrete set of functions $\mathcal{M}_k$ defined by monomials of degree $\le k$. In particular, we know that each coordinate of a given w is implemented by some $q \in \mathcal{P}_k$. We consider a finite class of functions, where for each $i \in [l]$ and each monomial $s(z) = \prod_{i \in S} z_i$ of degree $\le k$ (where S is a multiset of elements from $[l]$), we include a 1-sparse function equal to $s(z)$ in the ith coordinate and 0 elsewhere.

$$\mathcal{M}_k = \left\{s^{(i)} : i \in [l], S \in [l]^n, n \le k\right\}$$
$$\text{where } s^{(i)}(z)_j = \begin{cases} \prod_{i \in S} z_i & j = i \\ 0 & \text{o.w.} \end{cases}$$

For $z \in \Delta_l$, these functions satisfy boundedness and sparsity. Further by convexity, obtaining $(\mathcal{C}, \mathcal{M}_k, \alpha)$-multicalibration implies $(\mathcal{C}, \alpha)$-degree-$(k+1)$ multicalibration. The cardinality of this set $|\mathcal{M}_k|$ grows as $O(l^k)$. We plug this bound on the number of weight functions into the generic sample complexity bound.

$$m_{k+1} \le O\left(\frac{l \cdot (\mathsf{VC}(\mathcal{C}) + k\log(l/\alpha\beta))}{\alpha^4}\right)$$

(Smooth). By Theorem 3.7, we can take $\alpha_0 \le \alpha/k$ to guarantee a $(\mathcal{C}, \alpha_0)$-smooth multicalibrated predictor f is also $(\mathcal{C}, \alpha)$-degree-$(k+1)$ multicalibrated. Then, for our choice of weight class $\mathcal{W}$ to guarantee smooth multicalibration, we appeal to Lemma 5.6, which upper bounds the cardinality $|\mathcal{W}| \le \exp\left(\widetilde{O}((l/\alpha_0)^{l-1})\right)$, proved below in Section 5.3. With the choice of α_0, we bound the log of this cardinality as follows.

$$\log|\mathcal{W}| \le \widetilde{O}\left(\left(\frac{l}{\alpha_0}\right)^{l-1}\right) \le \left(\frac{kl}{\alpha}\right)^{l-1} \cdot \mathrm{poly}(l, \log(k/\alpha))$$

Combining these bounds, we can bound the sample complexity for smooth multicalibration as follows.

$$\begin{aligned} m_s &\leq O\left(\frac{k^4 l \cdot (\mathsf{VC}(\mathcal{C}) + \log(kl/\alpha\beta) + \log|\mathcal{W}|)}{\alpha^4}\right) \\ &\leq O\left(k^4 l \cdot \left(\frac{\mathsf{VC}(\mathcal{C}) + \log(kl/\alpha\beta)}{\alpha^4} + \frac{(kl)^{l-1}\,\mathrm{poly}(l, \log(k/\alpha))}{\alpha^{l+3}}\right)\right) \end{aligned}$$

(Full). For $\delta > 0$, by Theorem 3.9, $\mathrm{MC}^{\mathrm{f}}_{\delta}(\alpha\delta^l/k - l\delta^{l+1}) \subseteq \mathrm{MC}^{\mathrm{s}}(\alpha/k) \subseteq \mathrm{MC}_{k+1}(\alpha)$. Balancing terms, we take $\delta = \frac{\alpha}{2kl}$, which results in $\alpha_0 \leq l^{-l}\left(\frac{\alpha}{2k}\right)^{l+1}$ to guarantee that a $(\mathcal{C}, \alpha_0, \delta)$-full multicalibrated predictor is also $(\mathcal{C}, \alpha)$-degree-$(k+1)$ multicalibrated. The interval basis $\mathcal{I}_\delta$ has $1/\delta^l$ functions, so we can bound $\log|\mathcal{I}_\delta|$ as follows.

$$\log|\mathcal{I}_\delta| = l \cdot \log(2kl/\alpha)$$

With the choice of δ and α_0, we bound the sample complexity as follows.

$$\begin{aligned} m_i &\leq O\left(\frac{l \cdot (\mathsf{VC}(\mathcal{C}) + (l+1)\cdot\log(2kl/\alpha\beta) + l\cdot\log(2kl/\alpha))}{l^4(\alpha/2kl)^{4(l+1)}}\right) \\ &\leq O\left((2kl)^{4(l+1)} \cdot \frac{\mathsf{VC}(\mathcal{C}) + \log(kl/\alpha\beta)}{l^2\alpha^{4(l+1)}}\right) \\ &\leq O\left(\frac{(2kl)^{4(l+1)} \cdot (\mathsf{VC}(\mathcal{C}) + \log(kl/\alpha\beta))}{\alpha^{4(l+1)}}\right) \end{aligned}$$

Finally, using the containment of full multicalibration within smooth multicalibration specialized to binary prediction, we can tighten the analysis for $l = 2$. In this case, we can take the binary interval basis $\mathcal{I}_\delta$ of size $1/\delta$ functions for $\delta = \Theta(\alpha_0^{1/2})$, and applying Proposition 3.14, we can take $\alpha_0 \leq O(\alpha^2/k)$, to ensure $(\mathcal{C}, \alpha/k)$-smooth multicalibration. In all, we can bound the sample complexity

$$m_{i,\mathrm{bin}} \leq O\left(\frac{k^4 \cdot (\mathsf{VC}(\mathcal{C}) + \log(k/\alpha\beta))}{\alpha^8}\right),$$

establishing Theorem 5.5. □

Consistent with the prior results on the relationship between the notions of multicalibration, we see that focusing on low-degree multicalibration can lead to significant sample complexity savings. In particular, in the multi-class setting, the low-degree complexity provides exponential savings compared to the smooth and full complexity. For the binary prediction case, the savings from low-degree multicalibration are only polynomial factors in α, but still practically-relevant. Even for very modest values of α, say 0.25, low-degree multicalibration obtains more than a 200-fold decrease in sample complexity.

While this analysis doesn't establish lower bounds on the sample complexity, the point is that the savings are coming from the difference in the necessary choice of α_0. Thus, it seems that any sample complexity upper bound should apply equally well for all notions (in terms of α_0), will result in an improved complexity for low-degree multicalibration. Of particular note, a recent work of [GJN+21]

establishes an (inefficient) algorithm with optimal dependence of α_0^{-2} for full multicalibration in the binary prediction case. Specifically, in our notation, for classes where each group $c \in \mathcal{C}$ has constant measure in $\mathcal{D}$, they achieve $(\mathcal{C}, \alpha_0, \delta)$-full mutlicalibration with probability at least $1 - \beta$ in sample complexity that grows as

$$O\left(\frac{\log(|\mathcal{C}|/\delta\beta)}{\alpha_0^2}\right).$$

Setting α_0 and δ to achieve even $(\mathcal{C}, \alpha)$-multiaccuracy, gives a dependence of α^{-4}. It would be interesting to extend their game-theoretic analysis to low-degree multicalibration, towards obtaining α^{-2} dependence.

5.3 Better bases for smooth multicalibration in high dimensions

In l dimensions, we can construct a sparse basis at the cost of a much larger sized family of weight functions.

Lemma 5.6. *For any $\eta \in (0, 1)$, there exists a $(\eta, 1)$-basis $\mathcal{B}_\eta$ for $\mathcal{L}_{1\to\infty}$ of size $\exp(O(l\eta^{-1})^{l-1} \log \frac{1}{\eta})$.*

Proof. We assume that $1/\eta \in \mathbb{N}$ is an integer. It will be enough to show such a basis for $\mathcal{L}_1$, since we can handle each coordinate of the output separately, at the cost of a multiplicative factor of l on the size of the family.

Break up $[0, 1]^{l-1}$ into $(3l/\eta)^{l-1}$ cubes of side length $\eta/3l$ each. Then, let $\mathcal{B}_\eta$ be the all functions on Δ_l which take one of the constant values $0, \eta/3, 2\eta/3, \ldots, 1$ on each of these cubes (where we have projected away the last coordinate x_l in Δ_l). We have

$$|\mathcal{B}_\eta| \leq (3/\eta + 1)^{(3l/\eta)^{l-1}} = \exp\left(O(l\eta^{-1})^{l-1} \log \frac{1}{\eta}\right)$$

Now, we claim that every $u \in \mathcal{L}_1$ can be approximated to within η by a single function in $\mathcal{B}_\eta$. Indeed, for each of the cubes, round u down to the nearest multiple $\eta/3$ on one of the corners of the cube. Construct the function $v \in \mathcal{B}_\eta$ by letting v take on this rounded value on that cube. Then, v will be within $\eta/3$ of u on that corner of the cube.

We claim that the whole cube (projected up to Δ_l) is within distance $2\eta/3$ in L_1 from the corner. To see this, note that by construction, the distance in L_1 in the first $l - 1$ coordinates is at most $(l - 1) \cdot \eta/3l < \eta/3$. Also, in Δ_l, the distance in the last coordinate is at most the L_1 distance in the rest of the coordinates, so this is at most $\eta/3$ as well.

Thus the whole cube is within $2\eta/3$ in L_1 from the corner, so since $u \in \mathcal{L}_1$, this means that the value of u on the whole cube is within $2\eta/3$ of its value of the corner. Therefore, the value of u on the whole cube is within $\eta/3 + 2\eta/3 = \eta$ of the value of v on the cube. Thus, $\|u - v\|_\infty \leq \eta$. □

6 Experiments

We use a numerical experiment to compare boosting for multiaccuracy (MA), degree-2 multicalibration (MC2) and full multicalibration (MC-full). The goal of this experiment is two-fold. First,

the theoretical sample complexity results are asymptotic; here, we show that qualitatively similar results hold empirically in the finite sample regime. Second, the sandwiching bounds show that MC2 can reduce overconfidence, as compared to multiaccurate predictors; we supplement the theory by showing a setting in which multiaccuracy post-processing does not correct for initial overconfidence, but degree-2 multicalibration does. In combination, these preliminary experiments suggest that the strongest notion of multicalibration is not always better. Given a fixed data set size, the realized fairness guarantees may actually improve by choosing a lower degree of multicalibration.

Metrics. We measure the performance of predictors across the first two moments across subpopulations $c \in \mathcal{C}$. Specifically, we measure the multiaccuracy error as

$$\textbf{multiaccuracy error}: \quad \max_{c \in \mathcal{C}} \mathbf{E}[c(\mathbf{x})(f(\mathbf{x}) - f^*(\mathbf{x}))] - \mathbf{E}[(1 - c(\mathbf{x}))(f(\mathbf{x}) - f^*(\mathbf{x}))] \tag{24}$$

Intuitively, this is the multi-accuracy error on the subpopulation and its complement. Second, we measure the excess variance as

$$\textbf{excess variance}: \quad \max_{c \in \mathcal{C}} \left(\mathrm{Var}[f(\mathbf{x}) \mid c(\mathbf{x}) = 1] - \mathrm{Var}[f^*(\mathbf{x}) \mid c(\mathbf{x}) = 1]\right) \cdot \Pr[c(\mathbf{x}) = 1] \tag{25}$$

which intuitively is how much the variance of the predicted probability exceeds the variance of the optimal predictions over all subpopulations in $\mathcal{C}$.

Setup. To estimate the excess variance we need access to the true probability f^* which is unavailable for real datasets. Therefore, we use a semi-synthetic dataset by fitting a neural network to the real UCI-adult dataset and use the neural network's predicted probability as the "true" Bayes optimal probability. We also fit a generative model (a variational autoencoder) to model the distribution on the features $\mathbf{x}$. Combined we create a synthetic dataset where we can sample $\mathbf{x}$ from the generative model, compute the "true" probability $f^*(\mathbf{x})$, and draw samples $\mathbf{y}$ from the "true" probability. Note that all learning algorithm only have access to the samples $(\mathbf{x}, \mathbf{y})$ and not the "true" probability; we use the "true" probability exclusively for computing the excess variance.

We generate three datasets: a pre-training set, a training set, and a test set. We first pretrain a three-layer neural network on the pre-training set, then use our boosting algorithms to adjust the predictions of the pretrained neural network to achieve multi-accuracy or calibration, and finally use the test set to assess performance. For the calibration class $\mathcal{C}$, we use linear functions with sigmoid activation.

Results. Our results are summarized in Figure 1. We make the following observations.

- The boosting-style algorithm for multi-accuracy (MA), degree 2 multi-calibration (MC2) and full multicalibration (MC-full) all improve the multi-accuracy error on the training set. This is consistent with our result that MC2 and MC-full imply MA, hence by achieving MC2 and MC-full we can also achieve multiaccuracy (MA). However, on the test set, we observe that multicalibration is much more prone to overfitting and the multi-accuracy error increases rapidly without carefully regularization (e.g. by early stopping).

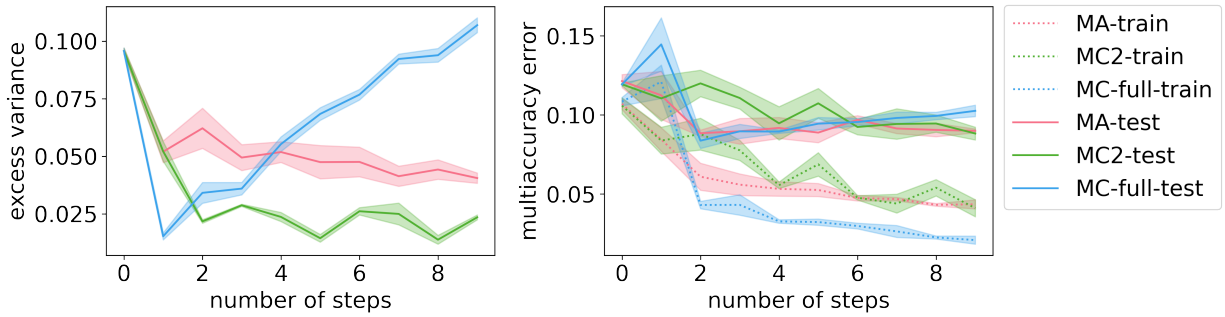

Figure 1: Comparing the excess variance and multiaccuracy error for three methods: boosting for multiaccuracy (MA), degree-2 multicalibration (MC2) and full multicalibration (MC-full). Error bars are 1 standard deviation of the results based on 5 randomly drawn datasets. Both MA and MC2 achieve low multiaccuracy error, but MC2 has much lower excess variance, consistent with our theoretical results. MC-full is very prone to overfitting.

- Boosting algorithms for degree-2 multicalibration (MC2) and full multicalibration (MC-full) can significantly decrease the excess variance. However, degree-2 multi-calibration is much less prone to overfitting and can consistently keep the excess variance low, while multicalibration rapidly overfits.

Overall we observe that running Algorithm 1 for degree-2 multicalibration (MC2) can reduce the excess variance without harming the multiaccuracy error, and generally maintains the generalization performance as compared to multiaccuracy (MA) only. On the other hand, boosting for multicalibration (MC-full) is significantly more prone to overfitting.

Acknowledgments. The authors thank Gal Yona for useful feedback on an earlier draft of this manuscript and Omer Reingold for helpful discussions.

References

[BCRT19] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *arXiv preprint arXiv:1903.04684*, 2019.

[BL20] Avrim Blum and Thodoris Lykouris. Advancing subgroup fairness via sleeping experts. In *Innovations in Theoretical Computer Science Conference (ITCS)*, volume 11, 2020.

[BLM01] Shai Ben-David, Philip M. Long, and Yishay Mansour. Agnostic boosting. In *14th Annual Conference on Computational Learning Theory, COLT*, 2001.

[BNS+16] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1046–1059. ACM, 2016.

[BRA+20] Noam Barda, Dan Riesel, Amichay Akriv, Joseph Levy, Uriah Finkel, Gal Yona, Daniel Greenfeld, Shimon Sheiba, Jonathan Somer, Eitan Bachmat, Guy N. Rothblum, Uri Shalit, Doron Netzer, Ran Balicer, and Noa Dagan. Developing a COVID-19 mortality risk prediction model when individual-level data are not available. *Nat Commun*, 11, 2020.

[Bri50] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[BS14] Boaz Barak and David Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. *arXiv preprint arXiv:1404.5236*, 2014.

[BYR+21] Noam Barda, Gal Yona, Guy N Rothblum, Philip Greenland, Morton Leibowitz, Ran Balicer, Eitan Bachmat, and Noa Dagan. Addressing bias in prediction models by improving subpopulation calibration. *Journal of the American Medical Informatics Association*, 28(3):549–558, 2021.

[Daw82] A. P. Dawid. Objective probability forecasts. *University College London, Dept. of Statistical Science. Research Report 14*, 1982.

[Daw84] A Philip Dawid. Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290, 1984.

[DFH+15] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126, 2015.

[DKR+19] Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Learning from outcomes: Evidence-based rankings. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 106–125. IEEE, 2019.

[DKR+21] Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Outcome indistinguishability. In *ACM Symposium on Theory of Computing (STOC'21)*, 2021.

[DKR+22] Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Beyond bernoulli: Generating random outcomes that cannot be distinguished from nature. In *The 33rd International Conference on Algorithmic Learning Theory*, 2022.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[Fel09] Vitaly Feldman. Distribution-specific agnostic boosting. *arXiv preprint arXiv:0909.2927*, 2009.

[FH18] Dean P. Foster and Sergiu Hart. Smooth calibration, leaky forecasts, finite recall, and nash dynamics. *Games Econ. Behav.*, 109:271–293, 2018.

[FV98] Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

[GJN+21] Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. *arXiv preprint arXiv:2101.01739*, 2021.

[GKR19] Sumegha Garg, Michael P. Kim, and Omer Reingold. Tracking and improving information in the service of fairness. In *Proceedings of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019*, pages 809–824, 2019.

[GKR+22] Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *Innovations in Theoretical Computer Science (ITCS'2022)*, 2022.

[GPSW17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[GRSW21] Parikshit Gopalan, Omer Reingold, Vatsal Sharan, and Udi Wieder. Multicalibrated partitions for importance weights. *arXiv preprint arXiv:2103.05853*, 2021.

[HKRR18] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, 2018.

[HPS+16] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Neural Information Processing Systems*, pages 3315–3323, 2016.

[JLN+19] Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees. *arXiv preprint arXiv:1909.03577*, 2019.

[JLP+21] 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.

[KF08] Sham Kakade and Dean Foster. Deterministic calibration and nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008.

[KF15] Meelis Kull and Peter Flach. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 68–85. Springer, 2015.

[KGZ19] 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.

[Kim20] Michael P. Kim. A complexity-theoretic perspective on fairness. *PhD thesis, Stanford University*, 2020.

[KK09] Adam Kalai and Varun Kanade. Potential-based agnostic boosting. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[KKG+22] 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), 2022.

[KMR16] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[KMR17] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS*, 2017.

[KMV08] Adam Tauman Kalai, Yishay Mansour, and Elad Verbin. On agnostic boosting and parity learning. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 629–638. ACM, 2008.

[KNRW18] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572, 2018.

[KPNK+19] Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Hao Song, Peter Flach, et al. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. *arXiv preprint arXiv:1910.12656*, 2019.

[KRR18] Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Fairness through computationally-bounded awareness. *Advances in Neural Information Processing Systems*, 2018.

[LW06] Michael Luby and Avi Wigderson. Pairwise independence and derandomization. *Foundations and Trends® in Theoretical Computer Science*, 1(4):237–301, 2006.

[MM02] Yishay Mansour and David McAllester. Boosting using branching programs. *Journal of Computer and System Sciences*, 64(1):103–112, 2002.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[OPVM19] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[P+99] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[PRW+17] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

[RY21] Guy N Rothblum and Gal Yona. Multi-group agnostic pac learnability. In *International Conference on Machine Learning*, pages 9107–9115. PMLR, 2021.

[SCM20] Eliran Shabat, Lee Cohen, and Yishay Mansour. Sample complexity of uniform convergence for multicalibration. *Advances in Neural Information Processing Systems*, 33:13331–13340, 2020.

[SDKF19] Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In *International Conference on Machine Learning*, pages 5897–5906. PMLR, 2019.

[TH21] Christopher Tosh and Daniel Hsu. Simple and near-optimal algorithms for hidden stratification and multi-group learning. *arXiv preprint arXiv:2112.12181*, 2021.

[ZE21] Shengjia Zhao and Stefano Ermon. Right decisions from wrong predictions: A mechanism design alternative to individual calibration. In *International Conference on Artificial Intelligence and Statistics*, pages 2683–2691. PMLR, 2021.

[ZKS+21] Shengjia Zhao, Michael P. Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[ZME20] Shengjia Zhao, Tengyu Ma, and Stefano Ermon. Individual calibration with randomized forecasting. *arXiv preprint arXiv:2006.10288*, 2020.

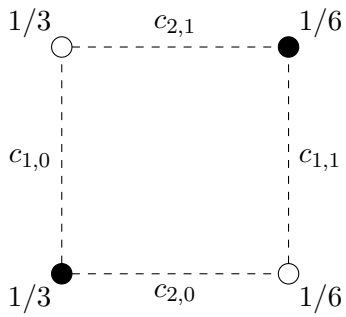

Figure 2: Diagram illustrating Example A.1. Black points indicate points $\mathbf{x}$ such that $\mathbf{y} = 1$ always, and white points indicate where $\mathbf{y} = 0$ always. The points are labeled with their probability under $\mathcal{D}$. The constraints each contain two points, and are drawn as edges.

A Multiaccuracy does not imply positive correlation

Note that many common regression methods do not in fact guarantee a positive or near-positive correlation between f and y, conditioned on a constraint c. We provide an example in the case of binary classification illustrating this.

Example A.1. *Let $\mathcal{X} = \{0,1\}^2$; we will refer to the two coordinates as x_1 and x_2. Let the four constraints $c_{i,b}$, for $i \in \{1,2\}$ and $b \in \{0,1\}$ be 1 exactly when $x_i = b$, and 0 elsewhere. Let the distribution $\mathcal{D}$ have weight $1/3$ on $(0,0)$ and $(0,1)$, and weight $1/6$ on $(1,0)$ and $(1,1)$. Finally, let the value $\mathbf{y}$ be always equal to the parity of $\mathbf{x}$; that is, $\mathbf{y} = x_1 \oplus x_2$. Fig. 2 illustrates this example.*

We will show that, in Example A.1, both L_2 and logistic regression obtain predictions f which have negative correlation with y conditioned on $c_{1,1}$.

First, L_2 *regression* finds the predictor f of the form

$$f(\mathbf{x}) = \sum_{i,b} \lambda_{i,b} c_{i,b}(\mathbf{x}),$$

such that the objective

$$\mathop{\mathbf{E}}_{\mathcal{D}}[(f(\mathbf{x}) - \mathbf{y})^2]$$

is minimized. By first order optimality, f is actually a multiaccurate predictor. We can see that if we apply L_2 regression, the obtained coefficients and predictor f are:

$$\lambda_{1,0} = \lambda_{1,1} = 1/2, \lambda_{2,1} = 1/6, \lambda_{2,0} = -1/6,$$

$$f((0,0)) = f((1,0)) = 2/3, f((0,1)) = f((1,1)) = 1/3. \tag{26}$$

This can be seen by running L_2 regression, or as follows. First, note that the constraints $c_{i,b}$ are linearly dependent (satisfying $c_{1,0} + c_{1,1} = c_{2,1} + c_{2,0}$), so we may assume that $\lambda_{1,0} = 1/2$. The remaining constraints are now linearly independent, so since the objective is strictly convex, there is a unique optimal choice of $\lambda_{i,b}$. Now note that the example exhibits symmetry by exchanging

$x_2 = 0$ and $x_2 = 1$, and flipping each of the $\mathbf{y}$ values. Thus, the value of the objective is conserved under the substitutions

$$\lambda_{1,1} \leftarrow 1 - \lambda_{1,1},\ \lambda_{2,0} \leftarrow -\lambda_{2,1},\ \lambda_{2,1} \leftarrow -\lambda_{2,0}.$$

But since the optimum is unique, this implies that $\lambda_{1,1} = 1/2$ and $\lambda_{2,0} = -\lambda_{2,1}$. Finally, we can obtain the actual value of $\lambda_{2,0}$ by using multiaccuracy, or by directly optimizing the objective. Finally, with the predicted values (26) we obtain the covariance conditioned on $c = c_{1,1}$ equal to

$$\mathrm{Cov}_{\mathcal{D}_c}[f(\mathbf{x}), \mathbf{y}] = -1/12,$$

showing that f is negatively correlated with y conditioned on $c_{1,1}$.

Next, *logistic regression* finds the predictor

$$h(\mathbf{x}) = \frac{1}{1 + \exp\left(-\sum_{i,b} \theta_{i,b} c_{i,b}(\mathbf{x})\right)},$$

maximizing the objective

$$\mathop{\mathbf{E}}_{\mathcal{D}}[\mathbf{y} \log h(\mathbf{x}) + (1 - \mathbf{y}) \log(1 - h(\mathbf{x}))].$$

Again, by first-order optimality h is also a multiaccurate predictor. This time, the obtained coefficients and predictor are:

$$\theta_{1,0} = \theta_{1,1} = 0, \theta_{2,1} = \log 2, \theta_{2,0} = -\log 2,$$

$$h((0,0)) = h((1,0)) = 2/3, h((0,1)) = h((1,1)) = 1/3. \tag{27}$$

This can again be seen by essentially an identical argument as for L_2 regression. Note that this is also the exact same predictor as that of L_2, so we obtain a similar negative correlation.

Note that these examples can also be modified by changing the probabilities under the distribution $\mathcal{D}$ from $1/3$ and $1/6$ to $1/2 - \epsilon$ and ϵ, respectively, as ϵ gets arbitrarily small. The same argument shows that $f((1,0))$ gets arbitrarily close to 1 while $f((1,1))$ gets arbitrarily close to 0. This achieves covariance arbitrarily close to $-1/4$, which is the lowest possible covariance between $[0,1]$ random variables. This is at the cost of $\mathcal{D}(c_{1,1}) = 2\epsilon$ getting arbitrarily small, so the constraint that we condition on gets arbitrarily low in probability, making statements about the constraint less meaningful.