Loss Minimization through the Lens of Outcome Indistinguishability

Parikshit Gopalan
Apple

Lunjia Hu*
Stanford University

Michael P. Kim†
UC Berkeley

Omer Reingold‡
Stanford University

Udi Wieder
VMware

Abstract

We present a new perspective on loss minimization and the recent notion of Omniprediction through the lens of Outcome Indistingusihability. For a collection of losses and hypothesis class, omniprediction requires that a predictor provide a loss-minimization guarantee simultaneously for every loss in the collection compared to the best (loss-specific) hypothesis in the class. We present a generic template to learn predictors satisfying a guarantee we call *Loss Outcome Indistinguishability*. For a set of statistical tests—based on a collection of losses and hypothesis class—a predictor is Loss OI if it is indistinguishable (according to the tests) from Nature's true probabilities over outcomes. By design, Loss OI implies omniprediction in a direct and intuitive manner. We simplify Loss OI further, decomposing it into a calibration condition plus multiaccuracy for a class of functions derived from the loss and hypothesis classes. By careful analysis of this class, we give efficient constructions of omnipredictors for interesting classes of loss functions, including non-convex losses.

This decomposition highlights the utility of a new multi-group fairness notion that we call calibrated multiaccuracy, which lies in between multiaccuracy and multicalibration. We show that calibrated multiaccuracy implies Loss OI for the important set of convex losses arising from Generalized Linear Models, without requiring full multicalibration. For such losses, we show an equivalence between our computational notion of Loss OI and a geometric notion of indistinguishability, formulated as *Pythagorean theorems* in the associated Bregman divergence. We give an efficient algorithm for calibrated multiaccuracy with computational complexity comparable to that of multiaccuracy. In all, calibrated multiaccuracy offers an interesting tradeoff point between efficiency and generality in the omniprediction landscape.

*LH is supported by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness, Omer Reingold's NSF Award IIS-1908774, and Moses Charikar's Simons Investigator award.

†MPK is supported by the Miller Institute for Basic Research in Science and, in part, by the Simons Collaboration on the Theory of Algorithmic Fairness.

‡OR is supported by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness, the Simons Foundation investigators award 689988, and Sloan Foundation Grant 2020-13941.

1 Introduction

Loss minimization is the dominant paradigm in machine learning. Techniques for loss minimization have played a critical role in the development of the theory and practice of supervised learning [KV94, BV04, SS+12, SB14, HR21]. A clean theoretical formulation of the underlying problem is via the notion of agnostic PAC learning [SB14]. We consider real-valued loss functions ℓ that take two arguments, a label $y \in \{0,1\}$ and an action $t \in \mathbb{R}$. Given a loss ℓ, a base class of hypotheses $\mathcal{C}$, and approximation parameter ε, the goal is to find a hypothesis h that achieves near-optimal expected loss (compared to $c \in \mathcal{C}$) over a fixed, but unknown distribution $\mathcal{D}$:[1]

$$\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}}[\ell(\mathbf{y}^*, h(\mathbf{x}))] \leq \min_{c\in\mathcal{C}} \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}}[\ell(\mathbf{y}^*, c(\mathbf{x}))] + \varepsilon.$$

Researchers have devoted significant effort into developing different choices of loss functions [MV08]. Different settings—so the conventional wisdom goes—require the design of different loss functions (e.g., squared, zero-one, logistic) to better encode the objectives of the task at hand (regression, classification, calibration). The choice of loss function dictates the updates during training and hence the resulting loss minimizer. With different loss functions, there are many different optimal hypotheses, and one needs to learn afresh for each loss.

Recent work pushes back against this conventional wisdom. The work of [GKR+22] introduces a solution concept for agnostic PAC learning, which they call *omniprediction*. Intuitively, an omnipredictor $\widetilde{p} : \mathcal{X} \to [0,1]$ is a predictor that can be used to simultaneously minimize loss for many different losses. Formally, an omnipredictor is parameterized by a collection of loss functions $\mathcal{L}$, a class of hypotheses $\mathcal{C}$, and approximation parameter ε. Given any loss $\ell \in \mathcal{L}$, a decision-maker can treat $\widetilde{p}(x)$ as if it were the Bayes optimal predictor $p^*(x) = \mathbf{E}[\mathbf{y}|x]$, selecting an action t that will minimize $\mathbf{E}[\ell(\widetilde{\mathbf{y}}, t)]$ where $\widetilde{\mathbf{y}}$ is drawn according to $\widetilde{p}$. Even though the true labels are drawn according to $p^*(x)$, the resulting decision rule is ε-optimal for ℓ over $c \in \mathcal{C}$. Importantly, the omnipredictor $\widetilde{p}$ is a single prediction function, fixed in advance, but yields optimal decisions for all $\ell \in \mathcal{L}$. The Bayes optimal predictor $p^*(x)$ is easily seen to be an omnipredictor for all losses, the question is whether they can be learnt efficiently. The main result in [GKR+22] is a sweeping feasibility result: they demonstrate that for any efficiently learnable hypothesis class $\mathcal{C}$ and $\varepsilon > 0$, efficient omnipredictors exist for the class $\mathcal{L}_{\text{cvx}}$ of all Lipschitz, convex loss functions. They prove this by showing a connection to *multicalibration*, from the literature on fair prediction [HKRR18].

Multicalibration was developed with the goal of promoting fairness across subpopulations encoded by a class of functions $\mathcal{C}$. In contrast to the loss-minimization paradigm, multicalibration does not frame learning as loss minimization. Rather, the goal of learning is to satisfy a collection of "indistinguishability" constraints. This view on multicalibration was developed in the recent work of [DKR+21], who introduced an alternative paradigm for learning called *outcome indistinguishability* (OI). OI considers two *alternate worlds* on individual-outcome pairs: in the natural world, outcomes $(\mathbf{x}, \mathbf{y}^*)$ are generated by Nature's true joint distribution; in the other simulated world, outcomes $(\mathbf{x}, \widetilde{\mathbf{y}})$ are sampled according to the predictive model $\widetilde{\mathbf{y}} \sim \text{Ber}(\widetilde{p}(\mathbf{x}))$. OI requires the learner to produce a predictor $\widetilde{p}$ in which the two worlds are computationally indistinguishable. More formally, OI is parameterized by a class of distinguisher algorithms $\mathcal{A}$. Each $a \in \mathcal{A}$ receives an individual $x \in \mathcal{X}$, an outcome $y \in \{0,1\}$, and the prediction $\widetilde{p}(x)$ and outputs a value in the

[1] This version where we do not restrict h to belong to $\mathcal{C}$ is sometimes called improper learning.

interval $[0,1]$. For such a collection of algorithms $\mathcal{A}$ and approximation parameter ε, a predictor $\widetilde{p}$ is $(\mathcal{A},\varepsilon)$-outcome indistinguishable[2] if no algorithm $a \in \mathcal{A}$ can distinguish between the two distributions over individual-outcome pairs.

$$\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}}[a(\mathbf{x},\mathbf{y}^*,\widetilde{p}(\mathbf{x}))] \approx_\varepsilon \mathop{\mathbf{E}}_{\substack{\mathbf{x}\sim\mathcal{D}\\ \widetilde{\mathbf{y}}\sim\mathrm{Ber}(\widetilde{p}(\mathbf{x}))}}[a(\mathbf{x},\widetilde{\mathbf{y}},\widetilde{p}(\mathbf{x}))]$$

As multicalibration is a special case of OI, by the results of [GKR+22], one can view omniprediction for convex, Lipschitz losses as a consequence of OI, for an appropriate family of distinguishers. While rigorous, this argument is rather indirect and in our view, it does not provide clear intuition for why there should be a link between loss minimization and indistinguishability. Moreover, the connection to multicalibration established in [GKR+22] is rather constrained in terms of the family of loss functions $\mathcal{L}$. If we want omnipredictors for a more expressive class such as all Lipschitz functions, not just convex ones (where it is known that multicalibration is insufficient [GKR+22, Lemma 6.7]), or simpler omnipredictors for a more restricted class of convex loss functions (such as L_p losses), the results of prior work don't shed much light on how we might proceed.

1.1 Our Contributions

Motivated by ominprediction, we establish a direct and intuitive connection between loss minimization and outcome indistinguishability, through a notion which we call *Loss OI*. Fundamental to our approach is to use loss functions as tools to construct *distinguishers*: given a family $\mathcal{L}$ of loss functions and a family of hypotheses $\mathcal{C}$, we devise a family of distinguishers $\mathcal{U}_{\mathcal{L},\mathcal{C}} = \{u_{\ell,c}\}_{\ell\in\mathcal{L},c\in\mathcal{C}}$ such that if $\widetilde{p}(x)$ is not an omnipredictor, then some distinguisher from this family can tell apart the labels generated by Nature from those generated by the predictor's simulation. We say that any predictor that fools every distinguisher from this family satisfies loss OI. By construction, loss OI implies omniprediction.

We show that loss OI admits a decomposition into two simpler outcome indistinguishability requirements which we call *hypothesis OI* and *decision OI*. Hypothesis OI compares the expected loss of the hypothesis c when labels are generated by Nature versus its simulation by $\widetilde{p}$, for each hypothesis in the class $c \in \mathcal{C}$. Decision OI tests compares the expected loss incurred when we take actions based on the optimal post-processing of the predictions of $\widetilde{p}$ under the two distributions on labels. We give a characterization of these indistinguishability conditions in terms of the *discrete derivative* $\partial\ell : [0,1] \to \mathbb{R}$ of the loss function ℓ, defined as $\partial\ell(t) = \ell(1,t) - \ell(0,t)$. Via this characterization, decision OI amounts to a *weighted* calibration condition derived from $\partial\ell$, which is implied by standard notions of calibration. Hypothesis OI can be expressed as a *multiaccuracy* condition for the class of functions $\partial\mathcal{L}\circ\mathcal{C} = \{\partial\ell \circ c : \ell \in \mathcal{L}, c \in \mathcal{C}\}$. Multiaccuracy [HKRR18,KGZ19] for a given hypothesis family $\mathcal{C}$ is a weaker notion than multicalibration for $\mathcal{C}$. Both notions require access to a weak agnostic learner for $\mathcal{C}$, but multiaccuracy admits simpler and more efficient algorithms in terms of sample complexity and running time.

[2] In fact, [DKR+21] introduce a more general hierarchy of OI notions, whose levels are based on the distinguishers' access to the predictions given by $\widetilde{p}$. The variant where we allow distinguishers access to $\widetilde{p}(\mathbf{x})$ (so-called, *sample-access* OI) is known to be computationally *equivalent* to multicalibration.

Loss OI for specific families. With this decomposition, we turn our attention to specific collections of loss functions $\mathcal{L}$. Since decision OI follows from calibration, to achieve hypothesis OI and loss OI, we analyze the structure of $\partial\mathcal{L} \circ \mathcal{C}$, with the goal of bounding the complexity of such functions.

- **All losses:** We begin with the family $\mathcal{L}_{\text{all}}$ of *all* losses satisfying minimal boundedness conditions. The losses need not be convex or Lipschitz. We show that loss OI is possible for $\mathcal{L}_{\text{all}}$ and any hypothesis class $\mathcal{C}$, provided we can ensure calibration and multiaccuracy over functions on the level sets of $\mathcal{C}$. Specifically, we require multiaccuracy over the collection $\text{level}(\mathcal{C}) = \{f \circ c\}$ for all $c \in \mathcal{C}$ and all maps $f : [-1, 1] \to [-1, 1]$. We can view these as the set of all bounded functions over the level sets of c. This has immediate consequences for Boolean (even discrete) hypothesis classes, since there, the class $\text{level}(\mathcal{C})$ is not much more complex than $\mathcal{C}$ itself: $\mathcal{C}$-multiaccuracy plus calibration implies loss minimization for any loss function.

- **Lipschitz losses:** Under Lipschitzness (but still without convexity), a weaker multiaccuracy condition suffices. We define $\text{Int}(\mathcal{C}, \alpha)$ to be the collection of Boolean functions, which are the indicators of the events that $c(x)$ lies in an interval of width α. We show that for Lipschitz losses, $\partial\mathcal{L} \circ \mathcal{C}$ lies in the linear span of functions in $\text{Int}(\mathcal{C}, \alpha)$. Hence, calibration together with $\text{Int}(\mathcal{C}, \alpha)$-multiaccuracy guarantees loss OI for all Lipschitz loss functions.

- **GLM losses:** GLMs are a popular class of convex loss minimization based models, which include basic learning algorithms such as linear and logistic regression. They can be viewed as minimizing Bregman divergences for predictors which are derived from linear combination of $\mathcal{C}$. For the class of GLM losses $\mathcal{L}_{\text{GLM}}$, we show that $\partial\mathcal{L} \circ \mathcal{C} = \mathcal{C}$. Hence, calibrated multiaccuracy—that is, calibration together with $\mathcal{C}$-mulitaccuracy—guarantees loss OI for all GLM losses. We give an equivalence between of predictors that satisfy Loss OI for $\mathcal{L}_{\text{GLM}}$ and the set of predictors satisfying a certain Pythagorean Theorem in the geometry of the corresponding Bregman divergence.

  Finally, we exhibit a reverse connection by showing that the optimal solution to any L_1-regularized GLM loss minimization problem is multiaccurate. This leads us to fast and practical methods for achieving both multiaccuracy and calibrated multiaccuracy.

Our results for Loss OI are incomparable with the result of [GKR+22] on omnipredictors. On one hand, loss OI is stronger than omniprediction. On the other hand, we require weak agnostic learning for $\partial\mathcal{L} \circ \mathcal{C}$, which might be a much more powerful primitive than weak learning for $\mathcal{C}$ itself (which is sufficient for multicalibration). For the class of convex Lipschitz losses $\mathcal{L}_{\text{cvx}}$ considered in [GKR+22], we show that multicalibration does not imply loss OI, although it implies omniprediction. Our best "upper bound" for $(\mathcal{L}_{\text{cvx}}, \mathcal{C})$-loss OI comes from $\text{Int}(\mathcal{C}, \alpha)$-multiaccuracy, and it applies even when the losses are non-convex. For the subset $\mathcal{L}_{\text{GLM}} \subset \mathcal{L}_{\text{cvx}}$, we show a stronger guarantee (loss OI versus omniprediction) from weaker assumptions (calibrated multiaccuracy versus multicalibration).

Calibrated multiaccuracy. A key takeaway from our results is the surprising power of the notion of calibrated multiaccuracy, where we require predictors to satisfy both multiaccuracy with respect to $\mathcal{C}$ and calibration. It implies loss OI for the class of GLM losses, and for the case

when $\mathcal{C}$ is Boolean. As a group fairness notion, it lies in between the notions of multiaccuracy and multicalibration. We show the running time and sample complexity needed to achieve calibrated multiaccuracy are not much higher than that required for multiaccuracy, by giving a simple algorithm that alternates between ensuring multiaccuracy is achieved (using gradient descent for squared loss), and recalibrating the output. The key insight is that either of these steps reduces the squared loss of the predictor. Hence the number of invocations of the weak learner is not much more in the worst case from that required to achieve multiaccuracy, and significantly smaller than that required for multicalibration.

Perspective. We see the key contribution of our work as conceptual: we bring the OI lens to the problem of loss minimization. Reasoning about the simulated labels $\widetilde{\mathbf{y}}$ turns out to a powerful idea in this context, which has not been explored before, even in prior work on omniprediction. Our framework leverages this to give a *compiler* that translates loss OI for a pair $(\mathcal{L}, \mathcal{C})$ into *low-level* calibration and multiaccuracy conditions. With this setup, the proofs of our results are not technically hard. For instance, our result for GLMs uses the well-known fact that the loss function for any GLM has the form $\ell_g(y,t) = g(t) - yt$. It follows that $\partial\ell(t) = -t$, hence $\mathcal{C}$-multiaccuracy suffices for hypothesis OI (assuming $\mathcal{C}$ is closed under negation).

The loss OI perspective establishes a natural and versatile link between loss minimization and indistinguishability. It broadens our understanding of omniprediction. On one hand, it shows it can be scaled up beyond convex, Lipschitz losses. But it can also scaled down for more limited classes of loss functions to give more efficient constructions. It enables a range of omniprediction guarantees, where the richness of the collection of losses scales with the expressive power of the class for which we require multiaccuracy.

Structure of this manuscript: The remainder of the manuscript is structured as follows. In Section 2, we present a high-level technical overview of our definitions and results. We discuss related work in 2.4. In Section 3, we give preliminaries and formal background. In Section 4, we introduce Loss OI and its relationship to omniprediction and the other notions of indistinguishability. We then show how Loss OI can be formulated in terms of multiaccuracy and calibration. In Section 5, we instantiate present our main result on loss OI for Generalized linear models. We also show an equivalence between our formulation of Loss OI for GLMs and Pythagorean theorems in the geometry of Bregman divergences. In Section 6, we consider other families of loss functions including those that are not necessarily convex or Lipschitz. In Section 7, we present and analyze an efficient algorithm for calibrated multiaccuracy, and establish that it is more efficient than multicalibration. We report on the results from some preliminary experiments that aim to establish the efficiency and effectiveness of calibrated multiaccuracy in Section 8. Proofs are occasionally deferred to Appendix A to streamline the flow.

2 Technical Overview

In this section, we give a more detailed but still high-level explanation of how loss OI gives a indistinguishability viewpoint on loss minimization and omniprediction. The starting point for our investigation is understanding why the Bayes optimal predictor is an omnipredictor for any loss

and concept class. We use $p^* : \mathcal{X} \to [0,1]$ to denote the Bayes optimal predictor, which represents Nature's true probability of positive outcomes.

$$p^*(x) = \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}}[\mathbf{y}^*|\mathbf{x} = x]$$

We consider loss functions $\ell : \{0,1\} \times [0,1] \to \mathbb{R}^+$ that take a label and action as arguments and return a real valued loss. For such a loss ℓ, if the labels are drawn as $\mathbf{y} \sim \mathrm{Ber}(p)$, there exists an optimal action $k_\ell(p) \in [0,1]$ defined as

$$k_\ell(p) = \mathop{\arg\min}_{t\in[0,1]} \mathop{\mathbf{E}}_{\mathbf{y}\sim\mathrm{Ber}(p)}[\ell(\mathbf{y}, t)]$$

We refer to k_ℓ as the optimal post-processing for ℓ. Since the Bayes optimal predictor p^* governs the conditional distribution over outcomes $\mathbf{y}^*$, by averaging over $\mathbf{x} \sim \mathcal{D}$, we conclude that $k_\ell \circ p^*$ satisfies the loss minimization guarantee for any loss, with respect to any hypothesis class $\mathcal{C}$.

$$\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}}[\ell(\mathbf{y}^*, k_\ell(p^*(\mathbf{x})))] \le \min_{c\in\mathcal{C}} \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}}[\ell(\mathbf{y}^*, c(\mathbf{x}))] \tag{1}$$

The challenge of constructing an omnipredictor is, given specific families of losses $\mathcal{L}$ and hypotheses $\mathcal{C}$ respectively, to identify properties of $\widetilde{p}$ that will allow us to replace p^* with $\widetilde{p}$ in the above statement, as long as $\ell \in \mathcal{L}$ and $c \in \mathcal{C}$. Formally, we say that a predictor $\widetilde{p}$ is an $(\mathcal{L}, \mathcal{C}, \varepsilon)$-omnipredictor if for every loss $\ell \in \mathcal{L}$, the post-processed predictor $k_\ell \circ \widetilde{p}$ is an ε-loss minimizer compared to the class $\mathcal{C}$:

$$\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}}[\ell(\mathbf{y}^*, k_\ell(\widetilde{p}(\mathbf{x})))] \le \min_{c\in\mathcal{C}} \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}}[\ell(\mathbf{y}^*, c(\mathbf{x}))] + \varepsilon. \tag{2}$$

2.1 Omniprediction from outcome indistinguishability.

Omniprediction is a statement about Nature's distribution. Equation (2) makes no mention of the simulated predictions $\widetilde{\mathbf{y}}$. It is unclear how considering labels $\widetilde{\mathbf{y}}$ from the predictor's simulation might be useful. Indeed, the simulated labels do not play a role in the [GKR+22] derivation of omniprediction from multicalibration.

The key insight is that *in the simulated world of labels $\widetilde{\mathbf{y}}$, $\widetilde{p}$ is the Bayes optimal predictor.* So Equation (2) holds with $\varepsilon = 0$. Indeed, we just apply Equation (1) with $\mathbf{y}^* = \widetilde{\mathbf{y}}$ and $p^* = \widetilde{p}$ to get

$$\mathop{\mathbf{E}}_{\substack{\mathbf{x}\sim\mathcal{D}\\ \widetilde{\mathbf{y}}\sim\mathrm{Ber}(\widetilde{p}(\mathbf{x}))}}[\ell(\widetilde{\mathbf{y}}, k_\ell(\widetilde{p}(\mathbf{x})))] \le \min_{c\in\mathcal{C}} \mathop{\mathbf{E}}_{\substack{\mathbf{x}\sim\mathcal{D}\\ \widetilde{\mathbf{y}}\sim\mathrm{Ber}(\widetilde{p}(\mathbf{x}))}}[\ell(\widetilde{\mathbf{y}}, c(\mathbf{x}))] \tag{3}$$

If $\widetilde{p}$ has the property that the expectations on either side of the Equation don't change much when we replace $\widetilde{\mathbf{y}}$ with $\mathbf{y}^*$, then this will imply our desired omniprediction guarantee (Equation (2)). But this condition is a form of outcome indistinguishability, tailored to distinguishers constructed from $\mathcal{L}$ and $\mathcal{C}$. Loss OI is a crisp formulation of this notion.

Loss OI. Loss OI is parameterized by a loss class $\mathcal{L}$ and a concept class $\mathcal{C}$, which induce the following collection of distinguishers:

$$u_{\ell,c}(y,p,x) = \ell(y,c(x)) - \ell(y,k_\ell(p)) \tag{4}$$
$$\mathcal{U}_{\mathcal{L},\mathcal{C}} = \{u_{\ell,c} : \ell \in \mathcal{L}, c \in \mathcal{C}\}$$

For a given loss ℓ, the distinguisher $u_{\ell,c} : \mathcal{Y}\times[0,1]\times\mathcal{X} \to \mathbb{R}$ measures the excess loss of the prediction $c(x)$ compared to the optimal post-processing k_ℓ applied to the predicted label distribution p. For a fixed $x \in \mathcal{X}$, if we generated labels $\widetilde{\mathbf{y}} \sim \mathrm{Ber}(\widetilde{p}(x)$, then $k_\ell(\widetilde{p}(x))$ is the optimal action, so $u_{\ell,c}(\widetilde{\mathbf{y}}, \widetilde{p}(x), x) \geq 0$. Hence, the expected value over $\mathbf{x} \sim \mathcal{D}$ is also non-negative. For omniprediction to hold, it would suffice if

$$\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}}[u_{\ell,c}(\mathbf{y}^*, \widetilde{p}(\mathbf{x}), \mathbf{x})] = \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}}[\ell(\mathbf{y}^*, c(\mathbf{x}))] - \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}}[\ell(\mathbf{y}^*, k_\ell(\widetilde{p}(\mathbf{x})))] \geq 0.$$

Loss OI imposes the stronger condition that the expectation under Nature's distribution and the simulation are (approximately) equal. For a loss class $\mathcal{L}$, a concept class $\mathcal{C}$, $\varepsilon > 0$, a predictor $\widetilde{p}$ is $(\mathcal{L},\mathcal{C},\varepsilon)$-loss OI if for all $\ell \in \mathcal{L}$ and for all $c \in \mathcal{C}$, the following approximate equality holds.

$$\mathop{\mathbf{E}}_{\substack{\mathbf{x}\sim\mathcal{D}\\ \widetilde{\mathbf{y}}\sim\mathrm{Ber}(\widetilde{p}(\mathbf{x}))}}[u_{\ell,c}(\widetilde{\mathbf{y}}, \widetilde{p}(\mathbf{x}), \mathbf{x})] \approx_\varepsilon \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}}[u_{\ell,c}(\mathbf{y}^*, \widetilde{p}(\mathbf{x}), \mathbf{x})] \tag{5}$$

By design, Loss OI guarantees omniprediction. In fact, it is a strictly stronger notion. In Section 4.1, we show that while $\mathcal{C}$-multicalibration implies omniprediction for $\mathcal{L}_{\mathrm{cvx}}$, it does not imply loss OI even for the ℓ_4 loss.

Proposition 1. *If a predictor $\widetilde{p}$ is $(\mathcal{L},\mathcal{C},\varepsilon)$-loss OI, then $\widetilde{p}$ is an $(\mathcal{L},\mathcal{C},\varepsilon)$-omnipredictor. The converse does not hold.*

The [GKR+22] proof of omniprediction was tailored specifically to multicalibration and the specific class of convex loss functions $\mathcal{L}_{\mathrm{cvx}}$. In contrast, Loss OI is a versatile notion that may be applied to any class of loss functions. By approaching the question of omniprediction via loss OI, we arrive at an easy-to-state set of sufficient conditions to obtain omniprediction for any class of losses $\mathcal{L}$ and hypothesis class $\mathcal{C}$.

Characterizing Loss OI via calibration and multiaccuracy. We define loss OI using distinguisher functions $\{u_{\ell,c}\}$ that depend on both $c(\mathbf{x})$ and $\widetilde{p}(\mathbf{x})$. It is known from the work of [DKR+21] that when distinguishers receive simultaneous access to $c(\mathbf{x})$ and $\widetilde{p}(\mathbf{x})$, outcome indistinguishability can implement (full) multicalibration. However, the distinguishers $u_{\ell,c}$ have very specific structure, which permits a decomposition of loss OI into two modular conditions, involving two different distinguishers that each depend on the label and one out of $c(\mathbf{x})$ and $\widetilde{p}(\mathbf{x})$ *separately*. The first set of distinguishers will simply compare the loss of hypotheses $c \in \mathcal{C}$ for each loss $\ell \in \mathcal{L}$, a condition we call *hypothesis OI*.

$$\mathbf{E}[\ell(\widetilde{\mathbf{y}}, c(\mathbf{x}))] \approx_\varepsilon \mathbf{E}[\ell(\mathbf{y}^*, c(\mathbf{x}))] \tag{6}$$

The second set of distinguishers evaluates the loss achieved by the predictor $\widetilde{p}$ under optimal post-processing for each loss, a condition we call *decision OI*.

$$\mathbf{E}[\ell(\widetilde{\mathbf{y}}, k_\ell(\widetilde{p}(\mathbf{x})))] \approx_\varepsilon \mathbf{E}[\ell(\mathbf{y}^*, k_\ell(\widetilde{p}(\mathbf{x})))] \tag{7}$$

Subtracting (7) from (6), we obtain (5), albeit with a slightly larger error parameter. In other words, if $\widetilde{p}$ satisfies both hypothesis OI and decision OI, then $\widetilde{p}$ satisfies loss OI.

It turns out that decision OI is easy to achieve, we show that it is implied by calibration. Recall that a predictor is α-calibrated if $\mathbf{E}[\mathbf{y}|\widetilde{p}(\mathbf{x}) = v] \approx_\alpha v$. Using a more nuanced notion called weighted calibration from [GKSZ22], we can get an exact characterization of decision OI (see Theorem 4.9).

To present a characterization of hypothesis OI, we need a couple of definitions. For a class of functions $\mathcal{C}$ and approximation $\alpha \geq 0$, a predictor $\widetilde{p}$ is $(\mathcal{C}, \alpha)$-multiaccurate if for every $c \in \mathcal{C}$, the correlation between c and $\mathbf{y}^* - \widetilde{p}(\mathbf{x})$ is at most α. Formally, we require

$$|\mathbf{E}[c(\mathbf{x}) \cdot (\mathbf{y}^* - \widetilde{p}(\mathbf{x}))]| \leq \alpha.$$

For a loss function ℓ, we define the discrete derivative $\partial\ell$ as $\partial\ell(t) = \ell(1,t) - \ell(0,t)$. For a loss class $\mathcal{L}$ and hypothesis class $\mathcal{C}$, we consider the class of functions $\partial\mathcal{L} \circ \mathcal{C} = \{\partial\ell \circ c : \ell \in \mathcal{L}, c \in \mathcal{C}\}$. We can characterize Hypothesis OI in terms of $\partial\mathcal{L} \circ \mathcal{C}$-multiaccuracy.

Proposition 2. *(Decomposition for Loss OI) For loss class $\mathcal{L}$, hypothesis class $\mathcal{C}$, and $\varepsilon \geq 0$, predictor $\widetilde{p}$ is $(\mathcal{L}, \mathcal{C}, \varepsilon)$-hypothesis OI iff it is $(\partial\mathcal{L} \circ \mathcal{C}, \varepsilon)$-multiaccurate. Thus, if $\widetilde{p}$ is ε-calibrated and $(\partial\mathcal{L} \circ \mathcal{C}, \varepsilon)$-multiaccurate, then it is $(\mathcal{L}, \mathcal{C}, O(\varepsilon))$-loss OI, and hence an $(\mathcal{L}, \mathcal{C}, O(\varepsilon))$-omnipredictor.*

Thus we have decomposed loss OI into two constraints on our predictors: calibration, and multiaccuracy for the class $\partial\mathcal{L} \circ \mathcal{C}$. This presents an alternative (and possibly more efficient) route to obtaining omnipredictors than via multicalibration.

Non-convex losses. Using our decomposition theorem we show that, perhaps surprisingly, loss-OI and omniprediction are feasible even for non-convex losses, given a sufficiently powerful learner for functions derived from $\mathcal{C}$. We require the losses to be bounded: $\|\partial\ell\|_\infty \leq 1$. But otherwise, the losses can be arbitrary, we do not assume Lipschitzness or convexity. Define the set

$$\text{level}(\mathcal{C}) = \{f \circ c : f \in \mathcal{F},\ c \in \mathcal{C}\} \quad \text{where } \mathcal{F} = \{f : \text{Im}(\mathcal{C}) \to [-1, 1]\}$$

That is, level($\mathcal{C}$) consists of all possible bounded post-processings of $c \in \mathcal{C}$; in particular the functions $f \in \mathcal{F}$ only get to distinguish between the *level sets* of each $c \in \mathcal{C}$. The importance of level($\mathcal{C}$) stems from the fact that $\partial\ell \circ c$ belongs to this class, hence level($\mathcal{C}$)-multiaccuracy suffices for Hypothesis-OI over all loss functions.

Proposition 3. *For any class of loss functions $\mathcal{L}$, if $\widetilde{p}$ is $(\text{level}(\mathcal{C}), \alpha)$-multiaccurate, then $\widetilde{p}$ is $(\mathcal{L}, \mathcal{C}, \alpha)$-hypothesis OI. Hence if $\widetilde{p}$ is α-calibrated and $(\text{level}(\mathcal{C}), \alpha)$-multiaccurate, then for any loss class $\mathcal{L}$, $\widetilde{p}$ is $(\mathcal{L}, \mathcal{C}, O(\alpha))$-loss OI.*

Thus, omnipredictors for every bounded loss function are computable, with complexity scaling with the complexity of weak agnostic learning for level($\mathcal{C}$). While level($\mathcal{C}$) could in general be far more expressive than $\mathcal{C}$ itself, there are important special cases, including when $\mathcal{C}$ is a family of Boolean functions, where it is not much larger than $\mathcal{C}$. In these settings, we get loss-OI for arbitrary losses from calibration and $\mathcal{C}$-multiaccuracy. This includes natural loss functions such as weighted 0-1 loss which are important for classification.

Lipschitz losses If we are willing to assume that the losses are Lipschitz, then we can obtain hypothesis OI from a weaker multiaccuracy condition. Intuitively, if the loss ℓ is Lipschitz in t, then so is $\partial\ell$, so we only need to consider Lipschitz post-processings. We can achieve this guarantee by enforcing multiaccuracy over the class of functions $\mathrm{Int}(\mathcal{C}, \alpha)$ which are the indicators of the event that $c(x)$ lies in a certain interval $I \subset [-1, 1]$ of width α, over all $c \in \mathcal{C}$ and intervals I.

We show that $\mathrm{Int}(\mathcal{C}, \alpha)$-multiaccuracy suffices to give Hypothesis OI for Lipschitz losses.

Proposition 4. *For any class of 1-Lipschitz loss functions $\mathcal{L}$, if $\widetilde{p}$ is $(\mathrm{Int}(\mathcal{C}, \alpha), \alpha^2)$-multiaccurate then $\widetilde{p}$ is $(\mathcal{L}, \mathcal{C}, O(\alpha))$-hypothesis OI. If $\widetilde{p}$ is also calibrated, then $\widetilde{p}$ is a $(\mathcal{L}, \mathcal{C}, O(\alpha))$-omnipredictor.*

2.2 Loss OI in GLMs.

GLMs are a important class of models from statistics that generalize linear and logistic regression [MN89, Agr15]. On a technical level, we start with an arbitrary strictly convex function f and look for a predictor that minimizes the associated Bregman divergence D_f. We restrict the predictor to have the (canonical) form $g'(h)$ where g is the Legendre dual of f and h comes from a hypothesis class $\mathcal{H}$ which is typically taken to be linear combinations over some base class $\mathcal{C}$. When we take $f = x^2$, this recipe gives linear regression with the squared loss. When $f = -H_2(x)$ is the negative binary entropy, we get logistic regression where our predictor is the sigmoid applied to linear function. The class of losses $\mathcal{L}_{\mathrm{GLM}}$ that arise in this manner are strongly convex. Thus, by the results of [GKR+22], $\mathcal{C}$-multicalibration suffices to obtain omniprediction for $\mathcal{L}_{\mathrm{GLM}}$.

Our first result on GLMs shows that the class $\partial\mathcal{L}_{\mathrm{GLM}} \circ \mathcal{C} = \mathcal{C}$. This holds for the simple reason that every loss $\ell_g \in \mathcal{L}_{\mathrm{GLM}}$ has the form $\ell_g(y, t) = g(t) - yt$, hence $\partial\ell(t) = -t$ is linear in t. This means that $\mathcal{C}$-multiaccuracy—not a derived class—plus calibration suffices for loss OI for GLMs.

Proposition 5. *If $\widetilde{p}$ is $(\mathcal{C}, \alpha)$-multiaccurate and α-calibrated, then it is $(\mathcal{L}_{\mathrm{GLM}}, \mathcal{C}, O(\alpha))$-Loss OI and an $(\mathcal{L}_{\mathrm{GLM}}, \mathcal{C}, O(\alpha))$-omnipredictor.*

These results highlight the power of calibrated multiaccuracy which gives omniprediction for all GLM losses. Before this, we only knew how to achieve this using the stronger notion of multicalibration. Is it really much easier to achieve calibrated multiaccuracy? A key piece of the answer comes from our next result shows a reverse connection between multiaccuracy and GLM optimality with ℓ_1-regularization. We state the result informally here.

Proposition 6 (Informal)**.** *For any GLM loss and $\alpha > 0$, the optimizer of the ℓ_1-regularized GLM optimization over the class $\mathcal{C}$ is $(\mathcal{C}, \alpha)$-multiaccurate.*

This result immediately gives a (number of) efficient avenues for computing a $\mathcal{C}$-multiaccurate predictor: run any ℓ_1-regularized GLM learner, like Lasso [Tib96] for linear regression. It also suggests a template for achieving calibrated multiaccuracy: we can alternate between the GLM learner and a calibration procedure such as isotonic regression until convergence [ZE01]. We will analyze a simple algorithm based on this template and show that its complexity is comparable to that of achieving multiaccuracy, and considerably lower than what is needed to achieve multicalibration.

Finally, we consider the Loss OI conditions for GLM losses. We show that, in this setting, the computational indistinguishability notion of Loss OI is equivalent to a geometric indistinguishability

condition, formalized by Pythagorean theorems in the associated Bregman divergence. We state the result fairly formally, deferring background on GLMs and Bregman divergences to the technical section.

Theorem 7. *Let f be a strictly convex function, and g be its Legendre dual, and let D_f be the corresponding Bregman divergence. A predictor $\widetilde{p}$ is $(\ell_g, \mathcal{H}, \alpha)$-Loss OI if and only if the following approximate Pythagorean theorem holds approximately.*

$$\mathbf{E}[D_f(p^*(\mathbf{x}), g'(h(\mathbf{x})))] \approx_\alpha \mathbf{E}[D_f(p^*(\mathbf{x}), \widetilde{p}(\mathbf{x}))] + \mathbf{E}[D_f(\widetilde{p}(\mathbf{x}), g'(h(\mathbf{x})))]$$

Intuitively, the Pythagorean theorem says that the "distance" between p^* and a predictor derived from the class $\mathcal{H}$ can be broken down into "orthogonal" components: the distance between p^* and $\widetilde{p}$ plus the distance between $\widetilde{p}$ and the predictor from $\mathcal{H}$. In other words, if a predictor $\widetilde{p}$ is $\mathcal{C}$-multiaccurate and calibrated, then it is simultaneously a "projection" of the best GLMs towards the statistically optimal predictor p^*.

2.3 Algorithms for Calibrated Multiaccuracy.

For a given hypothesis clas $\mathcal{C}$, we define the following classes of predictors.

- Let MA(α) denote the set of predictors that are $(\mathcal{C}, \alpha)$-multiaccurate.
- Let calMA(α) denote the set of predictors that are α-calibrated and $(\mathcal{C}, \alpha)$-multiaccurate.
- Let MC(α) denote the set of predictors that are $(\mathcal{C}, \alpha)$-multicalibrated.

Then we have MA(α) $\supseteq$ calMA(α) $\supseteq$ MC(α). We compare the complexity of computing a predictor in each of these classes given access to a (ρ, σ)-weak learner for $\mathcal{C}$ [BLM01, KS05, KMV08]. Such a learner, when given access to a distribution $(\mathbf{x}, \mathbf{z})$ where $\mathbf{x} \sim \mathcal{D}_\mathcal{X}$ and $\mathbf{z}$ are labels in $\{\pm 1\}$, if there exists $c \in \mathcal{C}$ such that $\mathbf{E}[c(\mathbf{x})\mathbf{z}] \geq \rho$, will return c' such that $\mathbf{E}[c'(\mathbf{x})\mathbf{z}] \geq \sigma$. If no such c exists it returns $\perp$. The complexity of learning the predictor in any of the aforementioned classes is governed by the number of oracle calls to the weak learner.

We present Algorithm 7.2 for achieving calibrated multiaccuracy that alternates between ensuring multiaccuracy (using the weak learner), and calibrating the predictor. The key insight that makes it efficient is that either step can be seen to reduce the same potential function, which is the squared distance from the Bayes optimal predictor. This results in a worst-case complexity for calMA that is not too different than just for achieving the weaker guarantee of MA (since that algorithm is also analyzed using the same potential).

We compare the number of oracle calls needed for computing a predictor in each of MA, calMA and MC. We emphasize that this is a comparison between the best known upper bounds. For MA, we use the [HKRR18] algorithm as analyzed in Lemma 7.6. For calMA, we use our analysis of Algorithm 7.2 in Theorem 4.9. For MC, we use the analysis of the algorithm from [GKR+22, Section 9], which is derived from the boosting by branching programs algorithm by [MM02].

- For MA(α), the number of calls made by the algorithm of [HKRR18] is bounded by $O(1/\sigma^2)$.

- For calMA(α), the number of calls made by Algorithm 7.2 bounded by $O(1/\sigma^2)$.
- For MC(α), the number of calls made by the algorithm of [GKR+22] is bounded by $O(1/\alpha^2\sigma^4)$. The weak learning assumption required is also somewhat stronger, see Section 7 and Appendix A.5 for a detailed discussion.

The comparison above shows that MA and calMA have similar complexities in terms of the worst-case number of calls to the weak learner. The number of calls required for MC is significantly larger. These results suggest that calibrated multiaccuracy is an interesting multi-group notion in its own right, that lies in between MA and MC. It offers an interesting tradeoff point between efficiency and generality in the omniprediction landscape. It is an interesting open problem to ask if it captures any of the desirable fairness properties of MC, or even of low-degree multicalibration [GKSZ22].

Finally, we show that calibrated multiaccuracy (and hence omniprediction for GLM losses) cannot be achieved by any algorithm that outputs a hypothesis which is a Single Index Model (SIM): these are functions of the form $u(\sum w_c c(x))$. In particular, this implies that known algorithms like the Isotron [KS09, KKKS11] which work in the realizable setting but produce a SIM as hypothesis cannot give an omnipredictor in the non-realizable setting.

We present some preliminary experiments which support the efficiency and omniprediction claims in Section 8. Importantly, the implementation is fewer than 100 lines of python code using standard regression and calibration libraries in sklearn, whereas multicalibration is more complex [GRSW22]. For a collection of common losses (incldluing some non-GLM losses), the calibrated MA predictor always competes with and sometimes outdoes the best linear predictor tailored to the loss.

2.4 Related Work and Discussion

Our work is inspired by and most closely related to the work of [GKR+22] which introduced omnipredictors, and the outcome indistinguishability framework of [DKR+21]. The relation of our results to the former is detailed in depth in Section 1.1. The outcome indistinguishability framework establishes general connections between multi-group fairness notions and appropriate levels in OI hierarchy. Here, we use their framework to focus on more fine-grained notions of OI that are tailored towards loss minimization and omniprediction. The framework of Loss OI is quite versatile, and has already been extended by [KP22] to the "performative" prediction setting, where predictions can influence the distribution over outcomes.

Rothblum and Yona [RY21] employed the notion of outcome indistinguishability in order to obtain loss-minimization over a rich family of sub populations. Their notion of loss functions is more general than ours. But they fix a single loss function in their discussion whereas we seek to address general families of loss functions. A major distinction is that our work studies the complexity of loss OI for broad families of loss functions and relates them to distinguishers that do not depend on the loss function.

The work of [GKSZ22] on low-degree multicalibration was also motivated by the goal of finding intermediate notions of multigroup fairness between MA and MC. They propose the hierarchy $\{\text{MC}_d\}$ of degree-d multicalibrated predictors which interpolates between these two notions. They show that several desirable fairness properties of MC are already achieved at low levels of the

hierarchy, at a computational cost similar to that of MA. Our results on calibrated multiaccuracy are similar in spirit but incomparable, we show how omniprediction for some important convex losses can already be obtained at calMA, at a computational cost comparable to that of MA.

There is a vast body of work on Generalized Linear Models [Agr15, Rig16]. Classically, the focus is on the setting where the function f defining the Bregman divergence and hence the *link function* f' and its inverse g' are known. The resulting program is convex and can be solved using the iteratively reweighted least squares algorithm [Rig16, MN89]. The set of convex losses $\mathcal{L}_{\mathrm{GLM}}$ derived from GLMs are also referred to as matching losses in the literature [AHW95].

The more challenging setting is where the link function is unknown. This is sometimes called the SIM (single index model) problem in the literature. To our knowledge, all work with provable guarantees (prior to the work of [GKR+22]) hold only for the realizable setting: the data are generated so that $\mathbf{E}[\mathbf{y}^*|\mathbf{x}] = g'(h(\mathbf{x}))$ for some $h \in \mathrm{Lin}(\mathcal{C})$, both g' and h are unknown. The first algorithm to give guarantees in this scenario was [Kal04], who finds a hypothesis that is close in squared error to the ground truth $g' \circ h$, and is represented as branching program. The elegant Isotron algorithm for this problem was introduced and analyzed in [KS09, KKKS11], it is a proper learning algorithm where the output is of the form $u \circ \widetilde{h}$, where $\widetilde{h} \in \mathrm{Lin}(\mathcal{C})$ and u is monotone.

Both our work and the work of [GKR+22] depart from these works in that they do not require the realizability assumption. We give a single predictor $\widetilde{p}$, with the guarantee that for any convex f (with Legendre dual g), $D_f(p^*, \widetilde{p})$ is comparable to $D_f(p^*, g' \circ h)$ for any $h \in \mathrm{Lin}(\mathcal{C})$. Under the realizability assumption, for any strongly convex function f bounding D_f implies a squared loss bound [KKKS11,Kan18], thus our results imply bounds for the squared loss. In the agnostic setting, squared loss and bounds on the divergence D_f are incomparable. The works of [SSS11, GKKT17] apply polynomial kernel techniques to the problem of loss minimization when the inverse link function is sigmoid or the ReLU for families of losses including ℓ_1 and the squared loss. In these settings, a polynomial dependence on the accuracy parameter ε is not possible.

Bregman divergences and Pythagorean theorems for them are studied in information geometry [Nie18, CT06], although the term is broadly used for inequalities arising from projections onto convex bodies. That a stronger guarantee than omniprediction holds true for the squared loss was observed in the work of [GKR+22, Lemma 8.4]. This guarantee was subsequently shown to hold even with degree-2 multicalibration [GKSZ22, Proposition A.1]. Our results generalize this to all GLM losses, and only assumes calibrated multiaccuracy, while also showing that for such losses, Pythagorean theorems are equivalent to loss OI.

3 Preliminaries

Let $\mathcal{D}$ be a distribution on labelled examples $(\mathbf{x}, \mathbf{y}^*)$ comprising of points $\mathbf{x}$ from a domain $\mathcal{X}$ and binary outcomes[3] $\mathbf{y}^* \in \{0, 1\}$. We let $\mathcal{D}_{\mathcal{X}}$ denote the marginal distribution over $\mathcal{X}$. We will occasionally refer to the distribution $\mathcal{D}$ as Nature. We assume sample access to Nature. $\mathrm{Ber}(p)$ denotes the Bernoulli distribution on $\{0, 1\}$ with parameter p. For a real valued function $f : \mathcal{T} \to \mathbb{R}$, let $\|f\|_\infty = \max_{\mathcal{T}} |f(x)|$. For a family of such functions $\mathcal{F}$, let $\|\mathcal{F}\|_\infty = \max_{f \in \mathcal{F}} \|f\|_\infty$.

[3] All our results can be extended to multi-class setting where there are finitely many distinct classes, but we work with the binary setting for simplicity.

Predictors: A predictor is a function $\widetilde{p} : \mathcal{X} \to [0,1]$ be a predictor, where $\widetilde{p}(x)$ is interpreted as an estimate of the label being 1, conditioned on x. For a predictor $\widetilde{p} : \mathcal{X} \to [0,1]$, we define the distribution $(\mathbf{x}, \widetilde{\mathbf{y}}) \sim \mathcal{D}(\widetilde{p})$ on $\mathcal{X} \times \{0,1\}$ where $\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}$ is sampled according to Nature's marginal distribution over inputs and conditioned on $\mathbf{x}$, $\widetilde{\mathbf{y}} \sim \mathrm{Ber}(\widetilde{p}(\mathbf{x}))$ so that

$$\widetilde{p}(x) = \mathbf{E}[\widetilde{\mathbf{y}} | \mathbf{x} = x].$$

We use $p^*(x) \in [0,1]$ to denote the Bayes optimal prediction for an individual $x \in \mathcal{X}$.

$$p^*(x) = \mathbf{E}[\mathbf{y}^* | \mathbf{x} = x]$$

In other words, using the optimal predictor $\mathcal{D}(p^*) = \mathcal{D}$ recovers the true distribution, Nature.

Calibration: Intuitively, a predictor is calibrated if, conditioned on the prediction $\widetilde{p}(\mathbf{x}) = v$, the expected outcome is close to v.

$$\mathbf{E}[\mathbf{y}^* | \widetilde{p}(\mathbf{x}) = v] \approx v$$

Formally, we quantify approximate calibration through *expected calibration error*.

Definition 3.1. *(ECE and Approximate calibration) We define the expected calibration error (ECE) of a predictor $\widetilde{p}$ as*

$$\mathsf{ECE}(\widetilde{p}) = \mathop{\mathbf{E}}_{\widetilde{p}(\mathbf{x})} \left| \mathop{\mathbf{E}}_{\mathbf{y}|\widetilde{p}(\mathbf{x})} [\mathbf{y} - \widetilde{p}(\mathbf{x})] \right|.$$

For $\alpha \geq 0$, a predictor $\widetilde{p} : \mathcal{X} \to [0,1]$ is α-calibrated if $\mathsf{ECE}(\widetilde{p}) \leq \alpha$.

A predictor $\widetilde{p}$ is perfectly calibrated if $\alpha = 0$, so that $\mathbf{E}_{\mathcal{D}}[\mathbf{y}^* | \widetilde{p}(\mathbf{x}) = v] = v$. While the notion of approximate calibration is well-defined for all predictors, checking for calibration efficiently requires the predictor to be discretized. When efficiency is a consideration, we will assume that the supported values of the predictor are multiples of some $\delta \in [0,1]$; such assumptions are standard in the calibration literature [FV98, HKRR18]. For such predictors, one can check for α-calibration given black-box access to $\widetilde{p}$ in time $\mathrm{poly}(1/\alpha, 1/\delta)$, using labeled samples.

Following [GKSZ22], we will allow for weighted notions of calibration, parametrized by a family of weight functions $\mathcal{W} = \{w : [0,1] \to \mathbb{R}\}$. Intuitively, we think of a weight function as highlighting predictions belonging to certain regions of $[0,1]$.

Definition 3.2. *Let $\mathcal{W} = \{w : [0,1] \to \mathbb{R}\}$ be a family of weight functions. For a predictor $\widetilde{p} : \mathcal{X} \to [0,1]$ we define*

$$\mathsf{CE}(\mathcal{W}, \widetilde{p}) = \max_{w \in \mathcal{W}} \left| \mathop{\mathbf{E}}_{\mathcal{D}} [w(\widetilde{p}(\mathbf{x}))(\mathbf{y}^* - \widetilde{p}(\mathbf{x}))] \right|.$$

We collect some simple properties of weighted calibration in the next lemma, the proof is in Section A.1. The first is that ECE is captured by considering weight functions bounded in absolute value by 1. The second is that α-calibration implies a bound on $\mathsf{CE}(\mathcal{W}, \widetilde{p})$ for any family of weights $\mathcal{W}$.

Lemma 3.3. *1. Let $\mathcal{W}^f$ denote the space of all functions $w : [0,1] \to [-1,1]$. Then*

$$\mathsf{ECE}(\widetilde{p}) = \mathsf{CE}(\mathcal{W}^f, \alpha).$$

2. If $\widetilde{p}$ is α-calibrated, then for any family $\mathcal{W}$ of weight functions,

$$\mathsf{CE}(\mathcal{W}, \widetilde{p}) \leq \|\mathcal{W}\|_\infty \alpha.$$

We will sometimes use weaker notions of calibration. An important special case is where we take $\mathcal{W}_1$ to be the set of all 1-Lipschitz weight functions bounded in the range $[-1, 1]$. We say that a predictor $\widetilde{p}$ is α-smoothly calibrated if it $\mathsf{CE}(\mathcal{W}_1, \widetilde{p}) \leq \alpha$.

Loss functions and decision functions: A loss function is a function $\ell : \{0,1\} \times \mathbb{R} \to \mathbb{R}$. For instance, we define the squared loss by $\ell_2(y,t) = \|y-t\|_2^2$ and the ℓ_p loss by $\ell_p(y,t) = \|y-t\|_p^p$. We define b_ℓ, the Lipschitz constant of ℓ, to be the smallest constant so that $|\ell(y,t_1) - \ell(y,t_2)| \leq b_\ell |t_1 - t_2|$. We let Lip_b denote the set of all b-Lipschitz functions. We say that a loss ℓ is convex, if for each $y \in \{0,1\}$, $\ell(y,t)$ is a convex function of t. In a generic loss minimization problem, given a loss function ℓ and a class $\mathcal{H}$ of hypotheses, one tries to find the hypothesis $h \in \mathcal{H}$ which minimizes $\mathbf{E}[\ell(\mathbf{y}, h(\mathbf{x}))]$. We extend the definition of ℓ via linearity so that the first argument can take values in $[0,1]$. We define

$$\ell(p,t) = \mathop{\mathbf{E}}_{\mathbf{y} \sim \mathrm{Ber}(p)}[\ell(\mathbf{y}, t)] = p \cdot \ell(1,t) + (1-p) \cdot \ell(0,t).$$

A decision function is a function $k : [0,1] \to \mathbb{R}$. We think of k as taking predictions $p \in [0,1]$ from a predictor and mapping them to actions $k(p) \in \mathbb{R}$. Decision functions are used to select a suitable action for a loss function, given a prediction of the distribution of labels. For a loss ℓ, we define the Bayes-optimal decision function $k_\ell : [0,1] \to \mathbb{R}$ by

$$k_\ell(p) = \arg\min_{t \in \mathbb{R}} \ell(p,t).$$

For proper losses like the squared error $(y-t)^2$, k_ℓ is simply the identity function. For the ℓ_1 loss $|y-t|$, $k_{\ell_1}(p)$ rounds p to the nearest value in $\{0,1\}$.

Hypotheses: A bounded hypothesis class is a family of functions $\mathcal{C} \subseteq \{c : \mathcal{X} \to [-1,1]\}$. We will assume that $\mathcal{C}$ contains the constant function 1 and is closed under negation. Our results will typically assume some learnability properties of the class $\mathcal{C}$, such as having bounded dimension and being weakly learnable. We define the class $\mathrm{Lin}(\mathcal{C}, B)$ to contain all functions of the form

$$h(x) = \sum_{c \in \mathcal{C}} w_c c(x), \quad \sum_{c \in \mathcal{C}} |w_c| \leq B.$$

Note that $|h(x)| \leq B$ for all $h \in \mathrm{Lin}(\mathcal{C}, B)$. We will consider loss minimization problems with the hypothesis class $\mathcal{H} = \mathrm{Lin}(\mathcal{C}, B)$ (e.g linear or logistic regression). Here B can be viewed as a regularization parameter.

Multicalibration: Originally introduced as a form of "multi-group" fairness [HKRR18], *multicalibration* and related notions have seen application beyond fair prediction in recent years. Intuitively, multicalibration requires that the predictions of $\widetilde{p}$ appear calibrated even when we restrict

our attention to structured subpopulations. [HKRR18] formalizes the collection of subpopulations through a concept class $\mathcal{C}$. Importantly, the multicalibration guarantee holds simultaneously for every $c \in \mathcal{C}$.

First, we define a weaker notion called multiaccuracy [HKRR18, KGZ19], which requires that predictions appear accurate in expectation (unbiased) over each $c \in \mathcal{C}$.

Definition 3.4. *Let $\mathcal{C} = \{c : \mathcal{X} \to [-1, 1]\}$ be a family of hypotheses and $\alpha \geq 0$. We say that the predictor $\widetilde{p} : \mathcal{X} \to [0, 1]$ is $(\mathcal{C}, \alpha)$-multiaccurate if for every $c \in \mathcal{C}$ it holds that*

$$\left| \mathop{\mathbf{E}}_{\mathcal{D}}[c(\mathbf{x})(\mathbf{y}^* - \widetilde{p}(\mathbf{x}))] \right| \leq \alpha$$

Multicalibration strengthens both calibration and multiaccuracy, requiring approximate calibration over each $c \in \mathcal{C}$. We adapt the definitions in [HKRR18, GKR+22] to our notion of approximate calibration.

Definition 3.5. *Let $\mathcal{C} = \{c : \mathcal{X} \to [-1, 1]\}$ be a family of hypotheses and $\alpha \geq 0$. We say that the predictor $\widetilde{p} : \mathcal{X} \to [0, 1]$ is $(\mathcal{C}, \alpha)$-multicalibrated if for every $c \in \mathcal{C}$ it holds that*

$$\mathop{\mathbf{E}}_{\widetilde{p}(\mathbf{x})} \left| \mathop{\mathbf{E}}_{\mathbf{y}|\widetilde{p}(\mathbf{x})} [c(\mathbf{x})(\mathbf{y}^* - \widetilde{p}(\mathbf{x}))] \right| \leq \alpha$$

By averaging over the predicted values, we can see that $(\mathcal{C}, \alpha)$-multicalibration implies $(\mathcal{C}, \alpha)$-multiaccuracy. Since we assume $1 \in \mathcal{C}$, $(\mathcal{C}, \alpha)$-multicalibration also implies α-calibration.

In defining multiaccuracy and multicalibration, we assume that the hypotheses are bounded by 1 in absolute value. For general hypotheses families $\mathcal{H}$, we define the multiaccuracy error as

$$\mathsf{MAE}(\mathcal{H}, \widetilde{p}) = \max_{h \in \mathcal{H}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}}[h(\mathbf{x})(\mathbf{y}^* - \widetilde{p}(\mathbf{x}))] \right| \right].$$

We will generally reserve the term $(\mathcal{C}, \alpha)$-multiaccuracy to denote a bounded hypothesis class $\mathcal{C}$ where $\mathsf{MAE}(\mathcal{C}, \widetilde{p}) \leq \alpha$. The hypotheses classes $\mathcal{H}$ most relevant to us are of the form $\mathcal{H} = \mathrm{Lin}(\mathcal{C}, B)$. For these, we can derive bounds on the multiaccuracy error from bounds for the base hypotheses in $\mathcal{C}$, that decay linearly with B. The proof is via linearity of expectation.

Lemma 3.6. *If the predictor $\widetilde{p}$ is $(\mathcal{C}, \alpha)$-multiaccurate, then for $B \geq 1$ and $\mathcal{H} = \mathrm{Lin}(\mathcal{C}, B)$ we have*

$$\mathsf{MAE}(\mathcal{H}, \widetilde{p}) \leq B\alpha.$$

Omnipredictors: The notion of omniprediction introduced by [GKR+22] asks for a single predictor which can do as well as the best hypothesis in a hypothesis class $\mathcal{H}$ for a family $\mathcal{L}$ of loss functions.

Definition 3.7. *We say that the predictor $\widetilde{p} : \mathcal{X} \to [0, 1]$ is an $(\mathcal{L}, \mathcal{H}, \delta)$-omnipredictor if for every loss $\ell \in \mathcal{L}$ and hypothesis $h \in \mathcal{H}$,*

$$\mathbf{E}[\ell(\mathbf{y}^*, k_\ell(\widetilde{p}(\mathbf{x})))] \leq \mathbf{E}[\ell(\mathbf{y}^*, h(\mathbf{x}))] + \delta.$$

Outcome Indistinguishability: Outcome indistinguishability introduced by [DKR+21] provides an elegant framework for reasoning about the quality predictions made by a predictor $\widetilde{p}$, by measuring their ability to fool statistical tests when nature's labels $\mathbf{y}^*$ and replaced by simulated labels $\widetilde{\mathbf{y}}$. The notion is parameterized by a class of algorithms $\mathcal{A} \subseteq \{a : \mathcal{X} \times \{0,1\} \times [0,1] \to [-1,1]\}$, whose goal is to "distinguish" Nature's distribution and the modeled distribution.

Definition 3.8 (Outcome Indistinguishability)**.** *A predictor $\widetilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{A}, \varepsilon)$-outcome indistinguishable if for every $a \in \mathcal{A}$,*

$$\left| \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}}[a(\mathbf{x}, \mathbf{y}^*, \widetilde{p}(\mathbf{x}))] - \mathop{\mathbf{E}}_{(\mathbf{x},\widetilde{\mathbf{y}})\sim\mathcal{D}(\widetilde{p})}[a(\mathbf{x}, \widetilde{\mathbf{y}}, \widetilde{p}(\mathbf{x}))] \right| \leq \varepsilon.$$

In fact, [DKR+21] consider various levels of OI which are defined by the degree of access to the predictions made available to the tests. In their language, Definition 3.8 corresponds to "sample-access OI" where the distinguisher receives access to $\mathbf{x}$, $\widetilde{p}(\mathbf{x})$, and outcomes sampled either from $\mathbf{y}^* \sim \mathrm{Ber}(p^*(\mathbf{x}))$ or $\widetilde{\mathbf{y}} \sim \mathrm{Ber}(\widetilde{p}(\mathbf{x}))$.

Also of relevance to us are special cases of this model. The first, so-called "no-access OI" corresponds to a restriction where the distinguishers do not receive $\widetilde{p}(\mathbf{x})$, and simply has access to either $(\mathbf{x}^*, \mathbf{y}^*) \sim \mathcal{D}$ or $(\mathbf{x}, \widetilde{\mathbf{y}}) \sim \mathcal{D}(\widetilde{p})$. Sample-access OI and No-access OI are in tight correspondence with multicalibration and multiaccuracy, respectively [DKR+21]. Another interesting special case of sample-access OI is when we are given access to $\widetilde{p}(\mathbf{x})$ but not to the point $\mathbf{x}$. Here, the goal is to distinguish between $(\mathbf{y}^*, \widetilde{p}(\mathbf{x})) \sim \mathcal{D}$ and $\widetilde{\mathbf{y}}, \widetilde{p}(\mathbf{x})) \sim D(\widetilde{p})$. OI for this model is tightly connected to calibration: for boolean outcomes, it follows that perfect calibration implies that these distributions are identical.

4 Outcome Indistinguishability for loss functions

We define notions of outcome indistinguishability for a predictor $\widetilde{p}$ with regard to distinguishers that are derived from a loss function ℓ. We allow distinguishers that take on real values, such a function distinguishes two distributions if its expected values differ significantly between them.

We define the notion of Loss OI formally. Here we compare the difference (between Nature and the predictor's model) in the expected loss suffered when using the hypothesis $c \in \mathcal{C}$ compared to when using the Bayes-optimal decision function k_ℓ based on the predictor $\widetilde{p}$.

Definition 4.1. (Loss OI) *Let $\mathcal{L}$ be a family of loss functions, $\mathcal{C}$ be a family of hypotheses, and $\varepsilon > 0$. For each $\ell \in \mathcal{L}, c \in \mathcal{C}$, define the distinguisher $u_{\ell,c} : \{0,1\} \times [0,1] \times \mathcal{X} \to \mathbb{R}$ by*

$$u_{\ell,c}(y, \widetilde{p}(x), x) = \ell(y, c(x)) - \ell(y, k_\ell(\widetilde{p}(x))). \tag{8}$$

We say that the predictor $\widetilde{p}$ is $(\mathcal{L}, \mathcal{C}, \varepsilon)$-loss-OI if for every loss $\ell \in \mathcal{L}$ and hypothesisc $\in \mathcal{C}$,

$$\left| \mathop{\mathbf{E}}_{\mathcal{D}}[u_{\ell,c}(\mathbf{y}^*, \widetilde{p}(\mathbf{x}), \mathbf{x})] - \mathop{\mathbf{E}}_{\mathcal{D}(\widetilde{p})}[u_{\ell,c}(\widetilde{\mathbf{y}}, \widetilde{p}(\mathbf{x}), \mathbf{x})] \right| \leq \varepsilon.$$

We define two additional, simpler notions. First is that of decision OI, which informally states that applying the Bayes optimal decision functions to the predictions of $\widetilde{p}$ and computing the expected loss cannot distinguish between $\mathbf{y}^*$ and $\widetilde{\mathbf{y}}$.

Definition 4.2. (Decision OI) *Let $\mathcal{L}$ be a family of loss functions, and $\varepsilon > 0$. We say that predictor $\widetilde{p}$ is $(\mathcal{L}, \varepsilon)$-decision-OI if for every $\ell \in \mathcal{L}$ it holds that*

$$\left| \mathop{\mathbf{E}}_{\mathcal{D}}[\ell(\mathbf{y}^*, k_\ell(\widetilde{p}(\mathbf{x}))] - \mathop{\mathbf{E}}_{\mathcal{D}(\widetilde{p})}[\ell(\widetilde{\mathbf{y}}, k_\ell(\widetilde{p}(\mathbf{x}))] \right| \leq \varepsilon.$$

Our next notion is hypothesis OI, which stipulates that no hypothesis from $\mathcal{C}$ results in significantly different expected loss whether the labels come from nature or the simulation.

Definition 4.3. (Hypothesis OI) *Let $\mathcal{L}$ be a family of loss functions, $\mathcal{C}$ a family of hypotheses and $\varepsilon > 0$. We say that the predictor $\widetilde{p}$ is $(\mathcal{L}, \mathcal{C}, \varepsilon)$-hypothesis-OI for $\varepsilon \geq 0$ if for loss $\ell \in \mathcal{L}$ and every hypothesis $c \in \mathcal{C}$ it holds that*

$$|\mathbf{E}[\ell(\mathbf{y}^*, c(\mathbf{x}))] - \mathbf{E}[\ell(\widetilde{\mathbf{y}}, c(\mathbf{x}))]| \leq \varepsilon.$$

We show that Loss OI is implied by having both Decision OI and Hypothesis OI simultaneously.

Lemma 4.4. *(***Decomposition lemma)** *If the predictor $\widetilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{L}, \varepsilon_1)$-decision-OI and $(\mathcal{L}, \mathcal{C}, \varepsilon_2)$-hypothesis-OI, then it is $(\mathcal{L}, \mathcal{C}, \varepsilon_1 + \varepsilon_2)$-loss-OI.*

Proof. For each $\ell \in \mathcal{L}$ and $c \in \mathcal{C}$ we can write

$$\begin{aligned}
&\mathbf{E}[u_{\ell,c}(\mathbf{y}^*, \widetilde{p}(\mathbf{x}), \mathbf{x})] - \mathbf{E}[u_{\ell,c}(\widetilde{\mathbf{y}}, \widetilde{p}(\mathbf{x}), \mathbf{x})] \\
&= \mathbf{E}[\ell(\mathbf{y}^*, c(\mathbf{x})) - \ell(\mathbf{y}^*, k_\ell(\widetilde{p}(\mathbf{x})))] - \mathbf{E}[\ell(\widetilde{\mathbf{y}}, c(\mathbf{x})) - \ell(\widetilde{\mathbf{y}}, k_\ell(\widetilde{p}(\mathbf{x})))] \\
&= \mathbf{E}[(\ell(\mathbf{y}^*, c(\mathbf{x}))] - \mathbf{E}[\ell(\widetilde{\mathbf{y}}, c(\mathbf{x}))] + \mathbf{E}[\ell(\widetilde{\mathbf{y}}, k(\widetilde{p}(\mathbf{x})))] - \mathbf{E}[\ell(\mathbf{y}^*, k_\ell(\widetilde{p}(\mathbf{x})))]. \qquad (9)
\end{aligned}$$

Hence by the triangle inequality,

$$\begin{aligned}
|\mathbf{E}[u_{\ell,c}(\mathbf{y}^*, \widetilde{p}(\mathbf{x}), \mathbf{x})] - \mathbf{E}[u_{\ell,c}(\widetilde{\mathbf{y}}, \widetilde{p}(\mathbf{x}), \mathbf{x})]| &\leq |\mathbf{E}[(\ell(\mathbf{y}^*, c(\mathbf{x}))] - \mathbf{E}[\ell(\widetilde{\mathbf{y}}, c(\mathbf{x}))]| \\
&\quad + |\mathbf{E}[\ell(\widetilde{\mathbf{y}}, k(\widetilde{p}(\mathbf{x})))] - \mathbf{E}[\ell(\mathbf{y}^*, k_\ell(\widetilde{p}(\mathbf{x})))]| \\
&\leq \varepsilon_1 + \varepsilon_2.
\end{aligned}$$

where the first term is bounded by hypothesis-OI and the second is bounded by decision-OI. □

4.1 Loss-OI implies Omniprediction

Our interest in the notion of loss-OI stems from the fact that it implies omniprediction.

Proposition 4.5 (Formal Restatement of Proposition 1)**.** *If the predictor $\widetilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{L}, \mathcal{C}, \varepsilon)$-loss-OI, then it is an $(\mathcal{L}, \mathcal{C}, \varepsilon)$-omnipredictor.*

Proof. A consequence of loss-OI is that for every $\ell \in \mathcal{L}$ and $c \in \mathcal{C}$, we have

$$\mathbf{E}[u_{\ell,c}(\mathbf{y}^*, \widetilde{p}(\mathbf{x}), \mathbf{x})] \geq \mathbf{E}[u_{\ell,c}(\widetilde{\mathbf{y}}, \widetilde{p}(\mathbf{x}), \mathbf{x})] - \varepsilon. \qquad (10)$$

But for every $x \in \mathcal{X}$, by the definition of the Bayes-optimal decision function k_ℓ we have

$$\mathbf{E}[u_{\ell,c}(\widetilde{\mathbf{y}}, \widetilde{p}(\mathbf{x}), \mathbf{x})|\mathbf{x} = x] = \mathbf{E}[\ell(\widetilde{\mathbf{y}}, c(\mathbf{x})) - \ell(\widetilde{\mathbf{y}}, k_\ell(\widetilde{p}(\mathbf{x}))|\mathbf{x} = x] \geq 0$$

since $k_\ell(\widetilde{p}(x))$ is defined to be action that minimizes expected loss for $\widetilde{\mathbf{y}} \sim \mathrm{Ber}(\widetilde{p}(x))$. Averaging over all $\mathbf{x} \sim \mathcal{D}$ gives

$$\mathbf{E}[u_{\ell,c}(\widetilde{\mathbf{y}}, \widetilde{p}(\mathbf{x}), \mathbf{x})] = \mathbf{E}[\ell(\widetilde{\mathbf{y}}, c(\mathbf{x}))] - \mathbf{E}[\ell(\widetilde{\mathbf{y}}, k_\ell(\widetilde{p}(\mathbf{x})))] \geq 0.$$

Plugging this into Equation (10) gives

$$\mathbf{E}[u_{\ell,c}(\mathbf{y}^*, \widetilde{p}(\mathbf{x}), \mathbf{x})] = \mathbf{E}[\ell(\mathbf{y}^*, c(\mathbf{x})) - \ell(\mathbf{y}^*, k_\ell(\widetilde{p}(\mathbf{x})))] \geq -\varepsilon.$$

Rearranging, we get that for every $\ell \in \mathcal{L}, c \in \mathcal{C}$,

$$\mathbf{E}[\ell(\mathbf{y}^*, k_\ell(\widetilde{p}(\mathbf{x})))] \leq \mathbf{E}[\ell(\mathbf{y}^*, c(\mathbf{x}))] + \varepsilon.$$

hence $\widetilde{p}$ is an $(\mathcal{L}, \mathcal{C}, \varepsilon)$-omnipredictor. □

The converse of this statement is not true. We show that omniprediction does not imply Loss-OI for any class $\mathcal{L}$ than includes the ℓ_4 loss. We prove an even stronger statement, that multicalibration does not imply loss-OI. This statement is stronger because of the result of [GKR+22] that multicalibration implies omniprediction for a broad class of convex loss functions. We define the ℓ_p loss for all $p \geq 1$ as

$$\ell_p(y, z) = \frac{1}{p}|y - z|^p$$

where the normalization by p makes it 1-Lipschitz. Let $L_p = \{\ell_p\}_{p \geq 1}$. We prove the following result which separates multicalibration from loss OI.

Theorem 4.6. *There exist a distribution $\mathcal{D}$, a class $\mathcal{C}$ and a predictor $\widetilde{p}$ such that*

- *$\widetilde{p}$ is $(\mathcal{C}, 0)$-multicalibrated, so it is an $(L_p, \mathcal{C}, 0)$-omnipredictor.*
- *$\widetilde{p}$ is not $(\{\ell_4\}, \mathcal{C}, \varepsilon)$-loss OI for any $\varepsilon < 4/9$.*

The proof which is given in Section A.2 uses Fourier analysis on the Boolean cube.

4.2 Loss OI from Calibration and Multiaccuracy

In order to analyze the notions of OI, we need to compare the expected loss under different distributions on labels for a certain action. The notion of *discrete derivative* of a loss function will aid these comparisons.

Definition 4.7. *Given a loss $\ell : \{0,1\} \times \mathbb{R} \to \mathbb{R}$, define the function $\partial\ell : \mathbb{R} \to \mathbb{R}$ as*

$$\partial\ell(t) = \ell(1, t) - \ell(0, t).$$

The following lemma justifies the analogy to partial derivatives.

Lemma 4.8. *For random variables $\mathbf{y}, \mathbf{y}' \in \{0,1\}$, and $t \in \mathbb{R}$ we have*

$$\mathbf{E}[\ell(\mathbf{y}, t)] - \mathbf{E}[\ell(\mathbf{y}', t)] = \mathbf{E}[(\mathbf{y} - \mathbf{y}')\partial\ell(t)]. \tag{11}$$

Proof. By definition

$$\mathbf{E}[\ell(\mathbf{y},t)] = \mathbf{E}[\mathbf{y}\ell(1,t) + (1-\mathbf{y})\ell(0,t)] = \mathbf{E}[\mathbf{y}\partial\ell(t)] + \ell(0,t)$$

We write a similar expression for $\mathbf{y}'$ and subtract. □

We now present characterizations of decision-OI and hypothesis-OI in terms of weighted calibration and multiaccuracy errors for suitably defined classes of functions. Combined with Lemma 4.4, this gives a decomposition of loss OI as a calibration condition and a multiaccuracy condition.

Theorem 4.9. *Let $\mathcal{L}$ be a family of loss functions and $\mathcal{C}$ be a hypothesis class.*

1. *Define the family of hypotheses $\partial\mathcal{L} \circ \mathcal{C} = \{\partial\ell \circ c\}_{\ell\in\mathcal{L}, c\in\mathcal{C}}$. The predictor $\widetilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{L}, \mathcal{C}, \varepsilon_1)$-hypothesis-OI where $\varepsilon_1 = \mathsf{MAE}(\mathcal{C}', \widetilde{p})$.*
2. *Define the family of weight functions $\mathcal{W}' = \{\partial\ell \circ k_\ell\}_{\ell\in\mathcal{L}}$. The predictor $\widetilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{L}, \varepsilon_2)$-decision-OI where $\varepsilon_2 = \mathsf{CE}(\mathcal{W}', \widetilde{p})$.*

Proof. We first prove Part (1). Conditioned on $\mathbf{x} = x$, by Equation (11) with $t = c(x)$ we can write

$$\mathbf{E}[\ell(\widetilde{\mathbf{y}}, c(\mathbf{x}))|\mathbf{x} = x] - \mathbf{E}[\ell(\mathbf{y}^*, c(\mathbf{x}))|\mathbf{x} = x] = \mathbf{E}[(\widetilde{p}(\mathbf{x}) - \mathbf{y}^*)\partial\ell(c(\mathbf{x}))|\mathbf{x} = x].$$

Hence taking expectations over $\mathbf{x}$ and absolute values,

$$|\mathbf{E}[\ell(\widetilde{\mathbf{y}}, c(\mathbf{x}))] - \mathbf{E}[\ell(\mathbf{y}^*, c(\mathbf{x}))]| \le \max_{c\in\mathcal{C}} |\mathbf{E}[(\widetilde{p}(\mathbf{x}) - \mathbf{y}^*)\partial\ell(c(\mathbf{x}))]|\,.$$

The LHS corresponds to hypothesis OI, while the RHS to $\mathcal{C}'$ multiaccuracy error for $\mathcal{C}' = \{\partial\ell \circ c\}$.

We now consider Part (2). Conditioned on $\mathbf{x} = x$, by Equation (11) with $t = k_\ell(\widetilde{p}(x))$,

$$\mathbf{E}[\ell(\widetilde{\mathbf{y}}, k_\ell(\widetilde{p}(x)))|\mathbf{x} = x] - \mathbf{E}[\ell(\mathbf{y}^*, k_\ell(\widetilde{p}(x)))|\mathbf{x} = x] = \mathbf{E}[(\widetilde{p}(x) - \mathbf{y}^*)\partial\ell(k_\ell(\widetilde{p}(x)))|\mathbf{x} = x].$$

We now take expectations over $\mathbf{x}$, followed by absolute values to get

$$\mathbf{E}[\ell(\widetilde{\mathbf{y}}, k_\ell(\widetilde{p}(x)))] - \mathbf{E}[\ell(\mathbf{y}^*, k_\ell(\widetilde{p}(x)))] = \mathbf{E}[(\widetilde{p}(x) - \mathbf{y}^*)\partial\ell(k_\ell(\widetilde{p}(x)))|\mathbf{x} = x].$$

The LHS corresponds to loss-OI while the RHS measures the weighted calibration error for $\mathcal{W}' = \{\partial\ell \circ k_\ell\}_{\ell\in\mathcal{L}}$. □

It is easy to see that the characterizations above are tight. For instance if $\mathsf{MAE}(\mathcal{C}', \widetilde{p})$ is larger than ε', then there exist a c, ℓ pair that distinguishes between $\mathbf{y}^*$ and $\widetilde{\mathbf{y}}$ with advantage ε'.

5 Loss-OI for Generalized Linear Models

In this section we study Loss OI and omniprediction in the context of Generalized Linear Models (GLMs), which are well-studied in machine learning and statistics [Kal04, KS09, KKKS11, Agr15, Rig16, Kan18, AHW95]. We give a self-contained description of GLMs in Section 5.1, where we introduce the family $\mathcal{L}_{\mathrm{GLM}}$ of convex losses that arise from GLMs. Our main results about GLMs are the following:

1. We show an information-geometric characterization of loss-OI for losses in $\mathcal{L}_{\mathrm{GLM}}$, showing an equivalence to a Pythagorean theorem for the associated Bregman divergence (Theorem 5.3).
2. We show that calibrated multiaccuracy implies loss OI for $\mathcal{L}_{\mathrm{GLM}}$ (Theorem 5.5).
3. As a partial converse, we show that the solution to the ℓ_1 regularized GLM loss minimization problem over $\mathrm{Lin}(\mathcal{C})$ is multiaccurate for $\mathcal{C}$ (Theorem 5.6).

5.1 Preliminaries about GLMs

Bregman divergences: Let $f : [0,1] \to \mathbb{R}$ be a strictly convex, twice differentiable function and let $f' : [0,1] \to \mathbb{R}$ denote its derivative. The Bregman divergence $D_f : [0,1] \times [0,1] \to \mathbb{R}^+$ corresponding to f is defined as

$$D_f(v^*, v) = f(v^*) - f(v) - (v^* - v)f'(v)$$

We say that f is λ-strictly convex if $f''(v) \geq \lambda$ for $v \in [0,1]$. For such f we have the inequality

$$f(v^*) \geq f(v) + (v^* - v)f'(v) + \frac{\lambda}{2}(v^* - v)^2.$$

Hence $D_f(v^*, v) \geq \lambda(v^* - v)^2/2$, and it vanishes iff $v^* = v$.

Legendre transform: Given f as above, the Legendre transform $g : \mathbb{R} \to \mathbb{R}$ of f is defined as

$$g(t) = \max_{v \in [0,1]} t \cdot v - f(v). \tag{12}$$

Let g' denote its derivative. We collect some properties of f and g, we refer the reader to [Nie10] for proofs of these facts.

- Let $I = [f'(0), f'(1)]$ so that $f' : [0,1] \to I$ is an injection. Then we have $g' : \mathbb{R} \to [0,1]$. Restricted to the interval I, we have $g' = f'^{-1}$, so that $g'(f'(v)) = v$ for $v \in [0,1]$.
- For $t \in I$, we have the identity

$$g(t) = tg'(t) - f(g'(t)). \tag{13}$$

GLMs: Consider the following Bregman divergence minimization problem:

$$\min_{\widetilde{p}:\mathcal{X}\to[0,1]} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[D_f(p^*(\mathbf{x}), \widetilde{p}(\mathbf{x}))] \tag{14}$$

Without restrictions on the structure of $\widetilde{p}$, the unique minimizer is given by p^*. Generalized linear models parameterize $\widetilde{p}$ using the so-called canonical link in a way that renders the resulting program convex In a GLM, we consider predictors $\widetilde{p}$ belonging to the class of *generalized linear models* $\{g' \circ h\}_{h \in \mathrm{Lin}(\mathcal{C},B)}$ and solve the minimization problem

$$\min_{\substack{\widetilde{p}=g'\circ h \\ h\in\mathrm{Lin}(\mathcal{C},B)}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[D_f(p^*(\mathbf{x}), \widetilde{p}(\mathbf{x}))] \tag{15}$$

Since $g' : \mathbb{R} \to [0,1]$, $g' \circ h$ is indeed a predictor for any $h : \mathcal{X} \to \mathbb{R}$. The key advantage of this choice of (inverse) link function is that it results in a convex optimization problem, for an appropriately defined loss function.

Definition 5.1. *Define the loss function $\ell_g : \{0,1\} \times \mathbb{R} \to \mathbb{R}$ by*

$$\ell_g(y,t) = g(t) - yt = \int_0^t (g'(s) - y)ds. \tag{16}$$

We define its extension (in the first argument) to the interval $[0,1]$ by $\ell(p,t) = g(t) - pt$.

The following lemma encapsulates the relation between the loss ℓ_g and the Bregman divergence D_f. This lemma can be derived from known results in the literature on generalized linear models and surrogate loss minimization [Nie10, AHW95, Rig16], we present the proof in Appendix for completeness.

Lemma 5.2. *For the loss function ℓ_g defined above,*

1. *We have $\ell_g(p^*, t) + f(p^*) = D_f(p^*, g'(t))$.*
2. *Program (15) is equivalent to the following convex program:*

$$\min_{h \in \mathrm{Lin}(\mathcal{C},B)} \mathbf{E}[\ell_g(\mathbf{y}, h(\mathbf{x}))] \tag{17}$$

For λ-strongly convex functions f, it holds that for any predictors p^* and $\widetilde{p}$,

$$\mathop{\mathbf{E}}_{\mathcal{D}}[D_f(p^*(\mathbf{x}), \widetilde{p}(\mathbf{x}))] \geq \lambda \, \mathbf{E}[(p^*(\mathbf{x}) - \widetilde{p}(\mathbf{x}))^2]/2.$$

5.2 Loss-OI for GLMs and Pythagorean theorems

We relate loss-OI for losses of the form ℓ_g to Pythagorean theorems for the Bregman divergence D_f. Let $p^*, \widetilde{p}, p : \mathcal{X} \to [0,1]$ be predictors. An exact Pythagorean bound for $(p^*, \widetilde{p}, p)$ is the statement

$$\mathbf{E}[D_f(p^*(\mathbf{x}), p(\mathbf{x}))] = \mathbf{E}[D_f(p^*(\mathbf{x}), \widetilde{p}(\mathbf{x}))] + \mathbf{E}[D_f(\widetilde{p}(\mathbf{x}), p(\mathbf{x}))].$$

In an approximate bound, the absolute value of the difference of the LHS and RHS is bounded. In our setting p^* will be the Bayes optimal predictor, $\widetilde{p}$ will be calibrated and $\mathcal{C}$-multiaccurate, while $p = g' \circ h$ belongs to the class of GLMs. The Pythagorean theorem says that minimizing the divergence to p^* for models p, is equivalent to minimizing the divergence to $\widetilde{p}$, which is clearly in the spirit of outcome indistinguishability. When we take $f = x^2/2$, D_f is just the squared Euclidean distance and the Pythagorean theorem has the familiar form of

$$\mathbf{E}[(p^*(\mathbf{x}) - p(\mathbf{x}))^2] = \mathbf{E}[(p^*(\mathbf{x}) - \widetilde{p}(\mathbf{x}))^2] + \mathbf{E}[(\widetilde{p}(\mathbf{x}) - p(\mathbf{x}))^2].$$

Here the statement implies that the error $p^* - \widetilde{p}$ is *orthogonal* to the space spanned by $\widetilde{p} - p$ over all generalized linear models p.

Theorem 5.3. *Let $f : [0,1] \to \mathbb{R}$ be a convex function and g be its Legendre dual. The predictor $\widetilde{p}$ is $(\ell_g, \mathcal{H}, \alpha)$-loss OI iff the following approximate Pythagorean bound holds for every model $h \in \mathcal{H}$:*

$$\left|\mathbf{E}[D_f(p^*(\mathbf{x}), \widetilde{p}(\mathbf{x}))] + \mathbf{E}[D_f(\widetilde{p}(\mathbf{x}), g'(h(\mathbf{x})))] - \mathbf{E}[D_f(p^*(\mathbf{x}), g'(h(\mathbf{x})))]\right| \leq \alpha$$

Proof. Recall that loss-OI corresponds to fooling the distinguishers

$$d_h(y, \widetilde{p}(x), x) = \ell_g(y, h(x)) - \ell_g(y, f'(\widetilde{p}(x)).$$

For the distribution $\mathcal{D}$ we have

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathcal{D}}[\ell_g(\mathbf{y}^*, h(\mathbf{x})) - \ell_g(\mathbf{y}^*, f'(\widetilde{p}(\mathbf{x}))] &= \mathbf{E}[D_f(p^*(\mathbf{x}), g'(h(\mathbf{x}))) - f(p^*(\mathbf{x})) - (D_f(p^*(\mathbf{x}), \widetilde{p}(\mathbf{x})) - f(p^*(\mathbf{x})))] \\ &= \mathbf{E}[D_f(p^*(\mathbf{x}), g'(h(\mathbf{x}))) - D_f(p^*(\mathbf{x}), \widetilde{p}(\mathbf{x}))]. \end{aligned} \tag{18}$$

For the distribution $\mathcal{D}(\widetilde{p})$ we have

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathcal{D}(\widetilde{p})}[\ell_g(\widetilde{\mathbf{y}}, h(\mathbf{x})) - \ell_g(\widetilde{\mathbf{y}}, f'(\widetilde{p}(\mathbf{x}))] &= \mathbf{E}[D_f(\widetilde{p}(\mathbf{x}), g'(h(\mathbf{x}))) - D_f(\widetilde{p}(\mathbf{x}), \widetilde{p}(\mathbf{x}))] \\ &= \mathbf{E}[D_f(\widetilde{p}(\mathbf{x}), g'(h(\mathbf{x})))]. \end{aligned} \tag{19}$$

Loss-OI, which asserts that the LHS of Equations (18) and (19) are within α, is equivalent to

$$\left|\mathbf{E}[D_f(p^*(\mathbf{x}), g'(h(\mathbf{x})))] - \mathbf{E}[D_f(p^*(\mathbf{x}), \widetilde{p}(\mathbf{x}))] - \mathbf{E}[D_f(\widetilde{p}(\mathbf{x}), g'(h(\mathbf{x})))]\right| \leq \alpha$$

□

5.3 Loss OI for GLMs from calibrated multiaccuracy

We first define the class of losses we will consider.

Definition 5.4. *For $D \geq 1$, let $\mathcal{F}^D$ be the set of all twice-differentiable strictly convex functions $f : [0,1] \to \mathbb{R}$, such that $|f''(x)| \leq D$. Let $\mathcal{G}^D$ denote the set of Legendre duals of functions in $\mathcal{F}$. Let $\mathcal{L}^D_{\text{GLM}} = \{\ell_g : g \in \mathcal{G}^D\}$ denote the associated loss functions.*

Let $\mathcal{W}^D$ be the set of all D-Lipschitz weight functions bounded in the range $[-1, 1]$. We say that the predictor $\widetilde{p}$ is α-smoothly calibrated if $\mathsf{CE}(\mathcal{W}^1, \widetilde{p}) \leq \alpha$. For any $D \geq 1$, this implies that $\mathsf{CE}(\mathcal{W}^D, \widetilde{p}) \leq D\alpha$.

Theorem 5.5. *Let $\mathcal{C}$ be a bounded hypothesis class and let $\mathcal{H} = \text{Lin}(\mathcal{C}, B)$ for $B \geq 0$. If the predictor $\widetilde{p}$ is α_1-smoothly calibrated, and $(\mathcal{C}, \alpha_2)$-multiaccurate, then for any $D \geq 1$, it is $(\mathcal{L}^D_{\text{GLM}}, \mathcal{H}, \alpha)$-loss OI for $\alpha = D\alpha_1 + B\alpha_2$.*

Proof. For any $g \in \mathcal{G}_B$, we have

$$\partial\ell_g(t) = \ell(1, t) - \ell(0, t) = (g(t) - t) - g(t) = -t.$$

Next we show that $k_{\ell_g}(p) = f'(p)$. For $p \in [0,1]$, we wish to find $t = k_\ell(p)$ which minimizes $k_{\ell_g}(p, t)$. By Lemma 5.2, this is the same as minimizing $D_f(p, g'(t))$. If we take $t = f'(p)$ then

$g'(t) = g'(f'(p)) = p$, which gives $D_f(p, g'(t)) = D_f(p, p) = 0$. Hence $k_\ell = f'$ is D-Lipshcitz, since $|f''(t)| \leq D$, and so is $\partial\ell \circ k_{\ell_g} = -f'(t)$. Hence if $\widetilde{p}$ is α-smoothly calibrated, then it is $(\partial\ell_g \circ f', D\alpha_1)$-calibrated.

Further, we have $\partial\ell_g \circ c = -c$. Hence being $(\mathcal{C}, \alpha)$-multiaccurate implies $(\mathcal{L}_D, \mathrm{Lin}(\mathcal{C}, B), B\alpha_2)$-hypothesis OI. We now apply the Decomposition lemma (Lemma 4.4). □

Finally, we show that multiaccuracy and GLMs are intimately connected. Indeed the infinity norm of the gradient vector corresponds to the multiaccuracy error of the predictor.

Theorem 5.6. *Let h^* be the optimal solution to the ℓ_1-regularized GLM loss minimization problem:*

$$\min_{h \in \mathrm{Lin}(\mathcal{C})} [\ell_g(\mathbf{y}, h(\mathbf{x})] + \alpha \sum_c |w_c| \text{ where } h(x) = \sum_c w_c c(x) \tag{20}$$

The predictor $g' \circ h^*$ is $(\mathcal{C}, \alpha)$-multiaccurate.

Proof. For $h(x) = \sum_c w_c c(x)$ define $L(w) = \mathbf{E}[\ell_g(\mathbf{y}, g'(h(\mathbf{x}))]$ so that L is a convex function of w. We can use the chain rule together with Equation (31), derived in the Appendix, to write

$$\frac{\partial L}{\partial w_c} = \mathbf{E}\left[\frac{d\ell(\mathbf{y}, t)}{dt}\Big|_{t=h(\mathbf{x})} \frac{\partial h(\mathbf{x})}{\partial w_c}\right] = \mathbf{E}[(g'(h(\mathbf{x})) - \mathbf{y})c(\mathbf{x})].$$

Let $\mathrm{sign}(t)$ be the sub-gradient of $|t|$, so that when $t = 0$, it can take any value in $[-1, 1]$. Note that $|\mathrm{sign}(t)| \leq 1$ for all t. If w^* is the parameter vector of h^*, then the (sub)-gradient of the loss vanishing is equivalent to the following equality holding for every $c \in \mathcal{C}$:

$$\mathbf{E}[(g'(h(\mathbf{x})) - \mathbf{y})c(\mathbf{x})] + \alpha \mathrm{sign}(w_c) = 0$$

Rearranging and taking absolute values,

$$\left|\mathbf{E}[(g'(h(\mathbf{x})) - \mathbf{y})c(\mathbf{x})]\right| \leq \alpha \left|\mathrm{sign}(w_c)\right| \leq \alpha.$$

□

This tells us that multiaccuracy is computationally easy to achieve, assuming access to a weak agnostic learner for the class $\mathcal{C}$. In particular, one could use least-squares with L_1-regularization which gives the familiar Lasso algorithm, or logistic regression with L_1 regularization. The weak-agnostic learner lets us identify a coordinate c along which the gradient is large, we use it to update h.

6 Loss OI for general families of losses

In this section, we instantiate the loss-OI framework to derive omniprediction guarantees for more general classes of losses than the convex, Lipschitz losses considered in the work of [GKR+22]. In particular, we explore the effect of relaxing each of those requirements. Our approach is to fix a loss class $\mathcal{L}$, then analyze for any class of hypotheses $\mathcal{C}$, the structure of the class $\{\partial\ell \circ c\}$. Doing so lets us derive loss OI guarantees where the complexity of the weak learning primitive we need grows with the expressiveness of $\mathcal{L}$. We also present results for the ℓ_p losses.

6.1 Arbitrary losses

Define the class $\mathcal{L}_{\text{all}}$ to consist of all loss functions ℓ such that $\|\partial\ell\|_\infty \le 1$. We can work with any constant in place of 1 by rescaling. [4] Let $\mathcal{C} = \{c : \mathcal{X} \to \mathbb{R}\}$ be a possibly unbounded hypothesis class. Define the class level($\mathcal{C}$) to be all functions on the level sets of $\mathcal{C}$ with range $[-1, 1]$. Formally, we define level($\mathcal{C}$) $= \{f \circ c\}$ where $f : \text{Im}(\mathcal{C}) \to [-1, 1]$ and $c \in \mathcal{C}$.

Theorem 6.1. *For a hypothesis class $\mathcal{C} = \{c : \mathcal{X} \to \mathbb{R}\}$, if $\widetilde{p}$ is α_1-calibrated and (level($\mathcal{C}$), α_2)-multiaccurate, then it is ($\mathcal{L}_{\text{all}}, \mathcal{C}, \alpha_1 + \alpha_2$)-loss OI.*

Proof. By the definition of $\mathcal{L}_{\text{all}}$, for the weight family $\mathcal{W}' = \{\partial\ell \circ k_\ell\}$ we have $\|\mathcal{W}'\|_\infty \le 1$. Hence by Lemma 3.3, if $\widetilde{p}$ is α-calibrated, then $\mathsf{CE}(\mathcal{W}', \widetilde{p}) \le \alpha$. By Theorem 4.9, if $\widetilde{p}$ is α_1-calibrated, then it satisfies ($\mathcal{L}_{\text{all}}, \alpha_1$)-decision OI.

If $\ell \in \mathcal{L}$, then $\partial\ell \circ c \in$ level($\mathcal{C}$), since we assume that $\|\partial\ell\|_\infty \le 1$. Hence by Theorem 4.9 being (level($\mathcal{C}$), α_2)-multiaccurate implies ($\mathcal{L}, \mathcal{C}, \alpha_2$)-hypothesis OI.

By the decomposition lemma, these conditions together imply ($\mathcal{L}, \mathcal{C}, \alpha_1 + \alpha_2$)-loss OI. □

6.1.1 On the complexity of level($\mathcal{C}$)

In general, the class level($\mathcal{C}$) might be much more expressive than $\mathcal{C}$ itself. Since $\mathcal{C}$-multiaccuracy is known to be equivalent to weak agnostic learning for the class $\mathcal{C}$, achieving multiaccuracy for level($\mathcal{C}$) might be computationally more complex than achieving it for $\mathcal{C}$. For instance, if $\mathcal{C}$ contains linear combinations of features x_i, then level($\mathcal{C}$) contains all halfspaces. However in the case where $\mathcal{C}$ has small range, they might not be too different. For Boolean functions, we can show the following:

Lemma 6.2. *If $\mathcal{C} = \{c : \mathcal{X} \to \{0, 1\}\}$ consists of Boolean functions, then ($\mathcal{C}, \alpha$)-multiaccuracy implies (level($\mathcal{C}$), 3α)-multiaccuracy.*

Proof. Take any $f : \{0, 1\} \to [-1, 1]$. Since $\mathcal{C}$ is Boolean, for $t \in \{0, 1\}$ we can write $f(t) = at + b$ where $|a| + |b| \le 3$. Hence $f \circ c \in \text{Lin}(\mathcal{C}, 3)$. We now apply Lemma 3.6. □

Combining Theorem 6.1 and Lemma 6.2, we get the following corollary:

Corollary 6.3. *Let $\mathcal{C} = \{c : \mathcal{X} \to \{0, 1\}\}$ be a class of Boolean functions. If $\widetilde{p}$ is α-calibrated and ($\mathcal{C}, \alpha$)-multiaccurate, then it is ($\mathcal{L}_{\text{all}}, \mathcal{C}, 4\alpha$)-loss OI.*

For Boolean hypotheses, the class $\mathcal{L}_{\text{all}}$ includes the 0-1 loss and its weighted variants. In this case the omniprediction guarantee is equivalent to agnostic learning. Thus for Boolean functions, calibrated multiaccuracy (which is multiaccuracy and calibration) suffices for agnostic learning.

Another important class where level($\mathcal{C}$) is not more complex than $\mathcal{C}$ is decision trees. When $\mathcal{C} = \{c : \mathcal{X} \to [-1, 1]\}$ is the class of decision trees (of bounded size/depth), then level($\mathcal{C}$) $= \mathcal{C}$, since given a decision tree c, the decision tree where we replace each leaf label v by $f(v)$ computes $f \circ c$ with the same size and depth.

[4] Strictly speaking, we don't require boundedness of $\partial\ell$ over its entire domain, it suffices if $\partial\ell$ is bounded for $\text{Im}(\mathcal{C}) \cup \text{Im}(k_\ell)$.

6.2 Lipschitz losses

Under suitable assumptions of Lipschitzness, we can replace level($\mathcal{C}$) with a simpler class functions. We first define the class of losses we consider.

Definition 6.4. *Define* Lip *to be the set of loss functions ℓ where*

- $\mathrm{Im}(k_\ell) \subseteq [-1, 1]$.
- *On the interval $[-1, 1]$, $\partial\ell$ is 1-Lipschitz and $|\partial\ell(t)| \leq 1$.*

Clearly Lip $\subseteq \mathcal{L}_{\text{all}}$. Let $\mathcal{C}$ be a bounded hypothesis class. For a given $\delta \geq 0$, partition the interval $[-1, 1]$ into intervals $\{I_j^\delta\}_{j=1}^{2/\delta}$ of width δ where $m \leq 2/\delta$. Define the family of functions

$$\mathrm{Int}(\mathcal{C}, \delta) = \{\mathbb{1}[c(x) \in I_j^\delta]\}_{j\in[m], c\in\mathcal{C}}.$$

Lemma 6.5. *If $\widetilde{p}$ is $(\mathrm{Int}(\mathcal{C}, \alpha), \alpha^2)$-multiaccurate, then it is $(\mathrm{Lip}, \mathcal{C}, 3\alpha)$-hypothesis OI.*

Proof. We first show that every function $\partial\ell \circ c$ can be uniformly approximated by a linear combination of functions from $\mathrm{Int}(\mathcal{C}, \delta)$. More precisely, for every $\ell \in$ Lip, there exist constants $l_1, \ldots, l_m \in [-1, 1]$ such that

$$\max_{t\in[-1,1]} \left| \partial\ell(c(x)) - \sum_{j=1}^{m} l_j \mathbb{1}[c(x) \in I_j] \right| \leq \delta/2. \tag{21}$$

For each interval I_j, let t_j be its midpoint so that $|t - t_j| \leq \delta/2$ for $t \in I_j$. Let $l_j = \partial\ell(t_j) \in [-1, 1]$. Since $\partial\ell$ is 1-Lipschitz on $[-1, 1]$, for every $t \in I_j$, $|\partial\ell(t) - l_j| \leq \delta/2$. Equation (21) follows by setting $t = c(x)$.

Hence, under $(\mathrm{Int}(\mathcal{C}, \delta), \alpha^2)$-multiaccuracy, it follows that

$$|\mathbf{E}[(\mathbf{y} - \widetilde{p}(\mathbf{x}))\partial\ell(c(\mathbf{x}))]| \leq \frac{\delta}{2} + \sum_{j=1}^{m} |l_j| \left|\mathbf{E}[(\mathbf{y} - \widetilde{p}(\mathbf{x}))\mathrm{Int}_{I_j}(c(\mathbf{x}))]\right| \leq \frac{\delta}{2} + \frac{2\alpha^2}{\delta} \leq 3\alpha$$

by choosing $\delta = \alpha$. □

Since $|k_\ell(p)| \leq 1$, we have $|\partial\ell \circ k_\ell(p)| \leq 1$. Hence the family of weight functions $\mathcal{W} = \{\partial\ell \circ k_\ell\}_{\ell\in\mathcal{L}}$ is bounded by 1. So we can bound $\mathcal{C}(\mathcal{W}, \widetilde{p}) \leq \alpha$ if $\widetilde{p}$ is α-calibrated. Hence the decomposition lemma together with Lemma 6.5 gives the following claim.

Theorem 6.6. *Let $\mathcal{C}$ be a bounded hypothesis class. If $\widetilde{p}$ is α-calibrated and $(\mathrm{Int}(\mathcal{C}, \alpha), \alpha^2)$-multiaccurate, then it is $(\mathrm{Lip}, \mathcal{C}, 4\alpha)$-loss OI.*

6.3 Low degree losses

Definition 6.7. *Define $\mathcal{L}_{d,B}$ to be the set of all loss functions where*

$$\partial\ell(t) = \sum_{i=0}^{d} b_j t^j, \text{ where } \sum_j |b_j| \leq B.$$

For a hypothesis class $\mathcal{C} : \mathcal{X} \to \mathbb{R}$, let $\mathcal{C}^d$ denote the hypotheses class $\{c(x)^j\}_{c\in\mathcal{C}, j\in[d]}$.

We refer to such losses as low-degree losses. We have already seen that $\mathcal{L}_{\text{GLM}} \in \mathcal{L}_{1,1}$, since $\partial\ell_g(t) = -t$. The following lemma is an immediate consequence of our definitions and the decomposition lemma.

Lemma 6.8. *Let $d \in \mathbb{Z}^+$ and $B \in \mathbb{R}^+$ If $\widetilde{p}$ is α-calibrated and $(\mathcal{C}^d, \alpha/B)$-multiaccurate, then it is $(\mathcal{L}_{d,B}, \mathcal{C}, 2\alpha)$-loss OI.*

The ℓ_p losses naturally yield low-degree losses.

- For unbounded $\mathcal{C}$ and p even, for $\ell(y,t) = (y-t)^p/p$, we have
$$\partial\ell_p(t) = \frac{1}{p}((1-t)^p - t^p).$$
Since p is even, the t^p term cancels and we have $\ell_p \in \mathcal{L}_{p-1,2^p}$.
- The same expression for $\partial\ell_p$ also holds for odd p if t is bounded to lie in the range $[0,1]$. In this setting, $\partial\ell_p$ is a degree p polynomial for $t \in [0,1]$. In particular, when $p = 1$, $\partial\ell_1 = 1-2t$ for $t \in [0,1]$. If we restrict $\mathcal{C}$ to be bounded in the range $[0,1]$, then we only need a bound for $t \in [0,1]$.

Thus by Lemma 6.8, in these settings, we get loss OI for the ℓ_p losses from calibration and multiaccuracy for $\mathcal{C}^p$.

7 Efficient algorithms for calMA

Let $\mathcal{C} = \{c : \mathcal{X} \to [-1,1]\}$ which contains the constant 1 function and is closed under negation. We recall the following classes

- Let $\text{MA}(\alpha)$ denote the set of predictors that are $(\mathcal{C}, \alpha)$-multiaccurate.
- Let $\text{calMA}(\alpha)$ denote the set of predictors that are α-calibrated and $(\mathcal{C}, \alpha)$-multiaccurate.
- Let $\text{MC}(\alpha)$ denote the set of predictors that are $(\mathcal{C}, \alpha)$-multicalibrated.

Then we have $\text{MA}(\alpha) \supseteq \text{calMA}(\alpha) \supseteq \text{MC}(\alpha)$. In this section we will give an efficient algorithm to compute a predictor in $\text{calMA}(\alpha)$. Like with algorithms for MA and MC, we will assume oracle access to a weak agnostic learner for $\mathcal{C}$. The complexity of the algorithm hinges on the number of calls made to the weak agnostic learner. The main takeaway from this section is the worst-case number of calls needed for calMA is similar to that for MA and lower than what is needed for MC.

Weak agnostic learning. We first define the notions of a weak agnostic learner and a discrete predictor which will be needed for our algorithm.

Definition 7.1. *Let $\mathcal{D}_{\mathcal{X}}$ be a distribution over $\mathcal{X}$. A (ρ, σ)-weak learner* WL *for $\mathcal{C}$ under $\mathcal{D}_{\mathcal{X}}$ is an algorithm* WL*, whose input is specified by a function $f : \mathcal{X} \to [-1, 1]$.*

- *The algorithm is given sample access to f via samples $(\mathbf{x}, \mathbf{z}) \in \mathcal{X} \times \{\pm 1\}$ where $\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}$, and $\mathbf{E}[\mathbf{z}|\mathbf{x} = x] = f(x)$.*
- *If there exists $c \in \mathcal{C}$ such that $\mathbf{E}_{\mathcal{D}}[c(\mathbf{x})f(\mathbf{x})] \geq \rho$, then $\mathrm{WL}(f, \rho) = c' \in \mathcal{C}$ such that $\mathbf{E}_{\mathcal{D}}[c'(\mathbf{x})f(\mathbf{x}))] \geq \sigma$.*
- *If no such c exists, the weak learner returns $\mathrm{WL}(f) = \perp$.*

Some observations about our definition:

- A non-proper learner is allowed to return hypothesis from a class $\mathcal{C}'$ which is different from $\mathcal{C}$. Our analysis goes through unchanged in this setting, we set $\mathcal{C}' = \mathcal{C}$ for simplicity. Typically, the weak learner will only succeed with probability $1 - \delta$. However, standard amplification allows us to make δ small at an added cost of $\log(1/\delta)$, so we ignore this failure probability for simplicity. Assuming that $\mathcal{C}$ is closed under negation is also a notational convenience, it lets us suppress the sign of the correlation in the updates.
- In our algorithms, $f(x)$ will take the form $p^*(x) - p_t(x)$ where p^* is the Bayes optimal predictor and $p_t : \mathcal{X} \to [0, 1]$ is our current hypothesis predictor. The weak agnostic learner requires sample access to f, which can be simulated via a standard trick in the literature on distribution-specific agnostic boosting [KK09, Fel09]. Note that $p^*(x) - p_t(x) \in [-1, 1]$. In order to simulate sample access to f, we draw a sample $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$. Then we generate $\mathbf{z} \in \{\pm 1\}$ so that $\mathbf{E}[\mathbf{z}] = \mathbf{y} - p_t(\mathbf{x})$. Since $\mathbf{y} - p_t(\mathbf{x}) \in [-1, 1]$, this uniquely specifies the distribution of $\mathbf{z}$. Moreover

  $$\mathbf{E}[\mathbf{z}|\mathbf{x} = x] = \mathbf{E}[\mathbf{y}|\mathbf{x} = x] - p_t(x) = p^*(x) - p_t(x) = f(x).$$

  Alternatively, some weak learners may accept real-valued labels; in this case, we can use $z = y - p_t(x)$ to label $x \sim \mathcal{D}_{\mathcal{X}}$.

We define the following norms over the space of predictors $p : \mathcal{X} \to [0, 1]$:

$$\begin{aligned} l_1(p_1, p_2) &= \mathbf{E}[|p_1(\mathbf{x}) - p_2(\mathbf{x})|] \\ l_2(p_1, p_2) &= \mathbf{E}[(p_1(\mathbf{x}) - p_2(\mathbf{x}))^2]^{1/2} \\ l_\infty(p_1, p_2) &= \max_{x \in \mathcal{X}} |p_1(\mathbf{x}) - p_2(\mathbf{x})| \end{aligned}$$

and observe that $l_1(p_1, p_2) \leq l_2(p_1, p_2) \leq l_\infty(p_1, p_2)$. Our algorithms will use the potential function

$$l_2(p^*, p)^2 = \mathbf{E}[(p^*(\mathbf{x}) - p_2(\mathbf{x}))^2].$$

We record the following technical lemma showing that multiaccuracy error and squared loss are robust under perturbations of the predictor. The proof is in Section A.4.

Lemma 7.2. *For any predictors p_1, p_2 such that $l_1(p_1, p_2) \le \delta$,*

$$\left|l_2(p^*, p_1)^2 - l_2(p^*, p_2)^2\right| \le 2\delta. \tag{22}$$

Further, if p_1 is $(\mathcal{C}, \alpha)$-multiaccurate, then p_2 is $(\mathcal{C}, \alpha + \delta)$-multiaccurate.

7.1 Discrete predictors and Calibration

We say that a predictor $p : \mathcal{X} \to [0, 1]$ is δ-discrete if its predictions are integer multiples of δ. For every predictor p, we associate it with a δ-discrete predictor p^δ as follows. We partition the interval $[0, 1]$ into $m = \lceil 1/2\delta \rceil$ intervals $\{I_1, \ldots, I_m\}$ of width 2δ each, where $I_j = [(2j-2)\delta, 2j\delta)$. For $x \in I_j$, we define

$$p^\delta(x) = (2j-1)\delta, \quad \bar{p}(x) = \mathbf{E}[\mathbf{y} | p(\mathbf{x}) \in I_j]$$

Some observation about these predictors:

1. The predictor p^δ is δ-discrete and $l_\infty(p, p^\delta) \le \delta$. Hence its squared loss and its multiaccuracy error are not much greater than that of p (by Lemma 7.2). But it need not be calibrated.
2. We can view $\bar{p}$ as the result of recalibrating p^δ, so $\mathsf{ECE}(\bar{p}) = 0$. Since $\bar{p}$ is obtained by calibrating p^δ, its squared loss is less than that of p^δ (since the mean is the constant value that minimizes the squared error), and even less than that of p for suitable parameter settings.
3. The predictor $\bar{p}$ is not efficiently computable since it is defined in terms of the expectation of $\mathbf{y}$ under $\mathcal{D}$. In Lemma 7.4, we give an efficient approximation $\widehat{p}$ to it (via random sampling), at the cost of a small increase in ECE.

We formalize observation (2) below, relating the reduction in squared loss to $\mathsf{ECE}(p^\delta)$.

Lemma 7.3. *For the predictors $p^\delta, \bar{p}$ defined above,*

$$l_2(p^*, p^\delta)^2 - l_2(p^*, \bar{p})^2 \ge \mathsf{ECE}(p^\delta)^2. \tag{23}$$

Proof. We write the LHS of Equation (23) as

$$\mathbf{E}[(p^*(\mathbf{x}) - p^\delta(\mathbf{x})^2] - \mathbf{E}[(p^*(\mathbf{x}) - \bar{p}(\mathbf{x})^2] = \mathbf{E}[(\bar{p}(\mathbf{x}) - p^\delta(\mathbf{x}))(2p^*(\mathbf{x}) - p^\delta(\mathbf{x}) - \bar{p}(x))]$$

We consider the distribution on intervals induced by choosing $\mathbf{x} \sim \mathcal{D}$ and $\mathbf{I_j} \ni p(\mathbf{x})$. Since p^δ and $\bar{p}$ are constant for each interval I_j, we can write $p^\delta(I_j)$ and $\bar{p}(I_j)$ for their values in this interval without ambiguity. Hence by first taking expectations over $\mathbf{I_j}$ and then $p(\mathbf{x}) \in \mathbf{I_j}$

$$\begin{aligned}
\mathbf{E}[(\bar{p}(\mathbf{x}) - p^\delta(\mathbf{x}))(2p^*(\mathbf{x}) - p^\delta(\mathbf{x}) - \bar{p}(x))] &= \mathop{\mathbf{E}}_{\mathbf{I_j}}\left[(\bar{p}(\mathbf{I_j}) - p^\delta(\mathbf{I_j})) \mathop{\mathbf{E}}_{\mathbf{x}|p(\mathbf{x})\in I_j}[2p^*(\mathbf{x}) - p^\delta(\mathbf{I_j}) - \bar{p}(\mathbf{I_j})]\right] \\
&= \mathop{\mathbf{E}}_{\mathbf{I_j}}\left[(\bar{p}(\mathbf{I_j}) - p^\delta(\mathbf{I_j}))^2\right] \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}\left[(\bar{p}(\mathbf{x}) - p^\delta(\mathbf{x}))^2\right].
\end{aligned}$$

where the penultimate line uses $\mathbf{E}[p^*(\mathbf{x})|\mathbf{x} \in I_j] = \bar{p}(I_j)$.

Since $p^\delta, \bar{p}$ are both constant one each interval I_j, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}\left[(\bar{p}(\mathbf{x}) - p^\delta(\mathbf{x}))^2\right] &= \mathop{\mathbf{E}}_{\mathbf{I_j}}\left[\mathbf{E}[\bar{p}(\mathbf{x}) - p^\delta(\mathbf{x}))|p(\mathbf{x}) \in I_j]^2\right] \\
&= \mathop{\mathbf{E}}_{\mathbf{I_j}}\left[\mathbf{E}[\mathbf{y} - p^\delta(\mathbf{x}))|p(\mathbf{x}) \in I_j]^2\right] \\
&\geq \mathop{\mathbf{E}}_{\mathbf{I_j}}\left[\left|\mathbf{E}[\mathbf{y} - p^\delta(\mathbf{x}))|p(\mathbf{x}) \in I_j]\right|\right]^2 \\
&= \mathsf{ECE}(p^\delta)^2
\end{aligned}$$

where the first inequality uses the convexity of x^2. □

Lemma 7.4. *Let $\mu, \delta \in [0,1]$. Given acess to a predictor p and random samples from $\mathcal{D}$,*

- *There exists an algorithm $\mathsf{estECE}(p,\mu)$ which returns an estimate of $\mathsf{ECE}(p^\delta)$ within additive error μ. The algorithm runs in time and sample complexity $\widetilde{O}(1/(\delta\mu^3))$.*
- *There exists an algorithm $\mathrm{reCAL}(p,\delta)$ which returns a predictor $\widehat{p}$ which has $\mathsf{ECE}(\widehat{p}) \leq \delta$ and $l_1(\bar{p},\widehat{p}) \leq \delta$. The algorithm has time and sample complexity $\widetilde{O}(1/\delta^4)$.*

For both algorithms, the stated guarantees hold with a failure probability can be made arbitrarily small by standard amplification. For simplicity, we have omitted this from the statement. The proof of this claim is through standard sampling arguments and use of Chernoff bounds. We record the following corollary, which follows from Lemmas 7.2, 7.3 and 7.4. The proofs are in Appendix A.4.

Corollary 7.5. *For the predictors $p, p^\delta, \widehat{p}$ defined above,*

$$l_2(p^*,p)^2 - l_2(p^*,\widehat{p})^2 \geq \mathsf{ECE}(p^\delta)^2 - 4\delta. \tag{24}$$

7.2 Multiaccuracy and Calibrated Multiaccuracy

We now state and analyze the algorithm for achieving calibrated multiaccuracy. To begin, we recall the algorithm of [HKRR18] for learning multiaccurate predictors. We present a formulation that will be useful for our main algorithm (Algorithm 7.2). The algorithm MA is given a predictor p_0 as input. It returns as output a predictor $\widetilde{p} \in \mathrm{MC}(\alpha)$, such that the squared distance to p^* only decreases. We use $\Pi(h)$ to denote the clip operator which takes a possibly real valued function $h : \mathcal{X} \to \mathbb{R}$ and truncates any values that lie outside $[0,1]$ to the closest value in $[0,1]$.

The counter t tracks the number of non-trivial updates made to the predictor. If we return at $t = T$, then the total number of calls to WL is $T+1$, where the last call does not yield a non-trivial update. We present the proof in Appendix A.4 for completeness.

Lemma 7.6. *[HKRR18] Assume that algorithm $\mathrm{MA}(p_0,\alpha)$ returns the predictor p_T where $T \geq 0$. Then $p_T \in \mathrm{MA}(\alpha)$ and*

$$l_2(p^*,p_0)^2 - l_2(p^*,p_T)^2 \geq T\sigma^2.$$

Algorithm 1 MA

Input: Predictor $p_0 : \mathcal{X} \to [0,1]$
Error parameter $\alpha \in [0,1]$.
Oracle access to a (ρ, σ) Weak learner WL for $\mathcal{C}$ under $\mathcal{D}_{\mathcal{X}}$ where $\alpha \geq \rho$.
Output: Predictor p_T.

```
t ← 0
ma ← false
while ¬ma do
  c_{t+1} ← WL(p* − p_t).
  if c_{t+1} = ⊥ then
    ma ← true
  else
    h_{t+1} ← p_t + σc_{t+1}.
    p_{t+1} ← Π(h_{t+1}).
    t ← t + 1.
  end if
end while
return p_t.
```

We now present our algorithm for finding a predictor in calMA(α). We will set the discretization δ to be small compared to α (α^2/C for some constant C). The algorithm may be viewed as starting with an arbitrary predictor p_0 and then running the following steps:

1. We set $p_{t+1} = \mathsf{MA}(p_t)$ to get a predictor that is mulitaccurate.
2. We estimate the calibration error $\mathsf{ECE}(p^{\delta}_{t+1})$ using estECE.
   (a) If the calibration error is large, we recalibrate it to $\widehat{p}_{t+1}$ using reCAL so that the calibration error drops to δ and repeat the loop.
   (b) Else, we return the predictor p^{δ}_{t+1}.

When we terminate, both multiaccuracy and calibration are achieved. Both steps reduce the potential function $l_2(p^*, p_t)^2$, (for suitable choices of parameters) which allow us to bound the overall number of iterations.

Some observations about the execution of Algorithm 7.2: the counter t tracks the number of executions of the while loop. Assume that we return at $t = T$. The updates p_t to q_t are made by reCAL, except the last update where $q_T = p^{\delta}_T$.

Theorem 7.7. *For $\alpha > 0$, let ρ, σ, δ be as given in Algorithm 7.2, and let q_T be the predictor returned.*

1. *$q_T \in$ calMA(α) and it is δ-discrete.*
2. *The number of iterations of the while loop is bounded by $O(1/\alpha^2)$.*
3. *The total number of calls to* WL *is bounded by $O(1/\sigma^2)$.*

Algorithm 2 calMA

Input: Predictor $p_0 : \mathcal{X} \to [0,1]$
Error parameter $\alpha \in [0,1]$.
Oracle access to a (ρ, σ)-Weak learner WL for $\mathcal{C}$ under $\mathcal{D}_{\mathcal{X}}$ where $\alpha - \alpha^2/32 \geq \rho$.
Output: Predictor q_T.

$\delta \leftarrow \alpha^2/32$.
$\mu \leftarrow \alpha/4$
$q_0 \leftarrow p_0$
$ma \leftarrow$ false
$t \leftarrow 0$
while $\neg ma$ **do**
  $t \leftarrow t+1$
  $p_t \leftarrow \mathsf{MA}(q_{t-1}, \alpha - \delta)$
  if $\mathsf{estECE}(p_t^\delta, \mu) > 3\alpha/4$ **then**
    $q_t \leftarrow \mathrm{reCAL}(p_t, \delta)$
  else
    $q_t \leftarrow p_t^\delta$
    $ma \leftarrow$ true
  end if
end while
return q_t

Proof. We have $q_T = p_T^\delta$, hence it is δ-discrete. The predictor p_T is $(\mathcal{C}, \alpha - \delta)$-multiaccurate, since it is returned by a call to MA. By Lemma 7.2, q_T is $(\mathcal{C}, \alpha)$-multiaccurate. This proves claim (1).

Assume that when we set $p_t = \mathsf{MA}(q_{t-1}, \alpha - \delta)$, this results in $S_t \geq 0$ calls to the weak learner WL. The by Lemma 7.6,

$$l_2(p^*, q_{t-1})^2 - l_2(p^*, p_t)^2 \geq S_t \sigma^2. \tag{25}$$

In every iteration of the while loop except the last, we have $\mathsf{estECE}(p_t^\delta, \mu) \geq 3\alpha/4$. By Lemma 7.4, this means that

$$\mathsf{ECE}(p_t^\delta) \geq \frac{3\alpha}{4} - \mu = \frac{\alpha}{2}.$$

Since $q_t = \widehat{p}_t$, applying Corollary 7.5, we have

$$l_2(p^*, p_t)^2 - l_2(p^*, q_t)^2 \geq \mathsf{ECE}(p_t^\delta)^2 - 4\delta \geq \frac{\alpha^2}{8}. \tag{26}$$

Adding Equations (25) and (26), for $t \in \{1, \ldots, T-1\}$,

$$l_2(p^*, q_{t-1})^2 - l_2(p^*, q_t)^2 \geq \frac{\alpha^2}{8}.$$

Summing this over all t,

$$l_2(p^*, q_0)^2 - l_2(p^*, q_{T-1})^2 \geq (T-1)\frac{\alpha^2}{8}.$$

Since $q_0 = p_0$ and $l_2(p^*, q_{T-1})^2 \geq 0$, we have

$$T \leq 1 + \frac{8}{\alpha^2} l_2(p^*, p_0)^2 = O(1/\alpha^2).$$

To bound the number of calls to the weak learner, we sum Equation (25) over all $t \in [T]$, and Equation (26) over all $t \leq T-1$ to get

$$l_2(p^*, p_T)^2 - l_2(p^*, p_0)^2 \geq \sum_t S_t \sigma^2 + (T-1)\frac{\alpha^2}{8}.$$

This implies that

$$\sum_t S_t \leq 1/\sigma^2$$

Since the number of calls to the weak learner in loop t is bounded by $S_t + 1$, we bound the number of calls by

$$\sum_t (S_t + 1) \leq \frac{1}{\sigma^2} + T = O(1/\sigma^2)$$

where we bound $T = O(1/\sigma^2)$, since $T = O(1/\alpha^2)$ and $\alpha \geq \rho \geq \sigma$. □

A quick remark about sample complexity: in each iteration of the while loop, we use fresh samples in order to ensure that the data and the current hypotheses are independent. This results in an $O(1/\alpha^2)$ sample overhead. It might be possible to improve the sample complexity using adaptive data analysis techniques as in [HKRR18], we leave this open. A similar issue arose in the original analysis of the Isotron algorithm [KS09], this was later remedied in the work of [KKKS11].

7.3 Complexity Comparison

In this section we compare and contrast the complexity of algorithms for computing a predictor in MA, calMA and MC. For an expressive hypothesis class $\mathcal{C}$, the running time is likely to be dominated by the calls to the (ρ, σ) weak learner WL. We compare the number of oracle calls needed for computing a predictor in each of these classes. We emphasize that this is a comparison between the best known upper bounds. For multiaccuracy, we use the [HKRR18] algorithm as analyzed in Lemma 7.6. For multicalibration, we use the analysis of the algorithm from [GKR+22, Section 9], which is derived from the boosting by branching programs algorithm by [MM02].

- For MA(α), the number of calls made by the algorithm of [HKRR18] is bounded by $O(1/\sigma^2)$. We require $\alpha \geq \rho$.
- For calMA(α), the number of calls made by Algorithm 7.2 bounded by $O(1/\sigma^2)$. We require $\alpha - \alpha^2/32 \geq \rho$.
- For MC(α), the number of calls made by the algorithm of [GKR+22] is bounded by $O(1/\alpha^2\sigma^4)$. The weak learning assumption required is also somewhat stronger, see Appendix A.5 for a detailed discussion. For simplicity, one could say that they require a (ρ, σ)-weak learner where $\alpha/2 \geq \rho$ under marginal distributions on $\mathcal{X}$ that are different from $\mathcal{D}_{\mathcal{X}}$.[5]

[5] They are obtained by conditioning $\mathcal{D}_{\mathcal{X}}$ on states with probability as small as $O(\alpha^2\sigma^2)$.

The comparison above shows that MA and calMA have similar complexities in terms of the number of calls to the weak learner. The number of calls required for MC is significantly larger.

Perhaps more importantly, the algorithm for calMA is easy to implement in Python using standard packages for regression and calibration, with some simple additional logic. In contrast, the logic to implement boosting by branching programs is non-trivial, and is not implemented by any standard python libraries to our knowledge [GRSW22]. In practice, the additional complexity of full multicalibration manifests primarily in terms of the samples needed to prevent overfitting. The work of [GRSW22] found that the sample complexity often rendered to algorithm impractical on medium sized real-world datasets. Even if we treat parameters like α and σ as reasonably large constants, the data requirement of calMA compared to MC could easily reduce by 100-fold.

7.4 Calibrated multiaccuracy requires non-linear models

Finally, we discuss the hypothesis class we use to fit a calibrated multiaccurate predictor. Because of the nature of the additive updates, the [HKRR18] algorithm for multiaccuracy returns a linear model $p \in \text{Lin}(\mathcal{C}, 1/\alpha)$. Note, however, that the calMA algorithm does not return such a model. In particular, the recalibration step introduces a nonlinearity, where we must condition on the value of the prediction p_t. The benefits of linearity arise in the simplicity of working with linear models, but also in the sample complexity. The easiest way to bound the sample complexity of achieving calibrated multiaccuracy is to take a fresh sample for each iteration of Algorithm 7.2, which loses a polynomial factor in $1/\alpha$. The sample complexity of multiaccuracy, however, can be bounded more tightly using straightforward uniform convergence arguments.

We may wonder if moving outside the class of linear models, or a slight generalization, is really necessary. Instead, we might hope that algorithms like the Isotron [KS09, KKKS11] may achieve calibrated multiaccuracy. The Isotron returns a so-called single index model (SIM) of the form

$$p(x) = u\left(\sum_{c \in \mathcal{C}} \lambda_c \cdot c(x)\right)$$

where $u : \mathbb{R} \to [0, 1]$ is any monotonic nondecreasing function. The Isotron algorithm is similar to Algorithm 7.2, switching between updating $\{\lambda_c\}$ based on the residuals given the current predictions and choosing u to recalibrate predictions. However, a crucial difference is that there the recalibration uses isotonic regression. This guarantees a calibrated predictor, but is not guaranteed to reduce the squared error. Given the update rule—as in our algorithm—if the procedure terminates, then multiaccuracy and calibration are guaranteed. [KS09], however, only establish convergence in the well-specified setting; in the agnostic setting that we study, it is not clear whether the Isotron is guaranteed to terminate.

Here, we show that the Isotron algorithm might not converge in the agnostic setting. Concretely, we construct a distribution and class $\mathcal{C}$, such that no SIM over the class $\mathcal{C}$ can achieve calibrated multiaccuracy. This simple construction shows the necessity of moving outside the class $\text{Lin}(\mathcal{C}, B)$, as in Algorithm 7.2.

Lemma 7.8 (Informal)**.** *No agnostic learning algorithm that returns a SIM* $p(x) = u\left(\sum_{\mathcal{C}} \lambda_c \cdot c(x)\right)$ *can guarantee* $p \in \text{calMA}(\alpha)$ *for any* $\alpha < 1/10$.

Proof. We exhibit a distribution over the 2-dimensional boolean cube and a collection of functions $\mathcal{C}$ such that calibrated multiaccuracy cannot be achieved by any SIM model. For simplicity, we show the violation for $\alpha = 0.1$ This example can be generalized to any dimension and approximate calibrated multiaccuracy.

Let $\mathcal{X} = \{00, 01, 10, 11\}$, and take the class $\mathcal{C} = \{x_i = 1, x_i = 0 : i \in [2]\}$ to be the four subcubes. Suppose that Bayes optimal probabilities on each $x \in \mathcal{X}$ are given as

$$p^*_{00} = 0 \qquad p^*_{01} = 1/2 \qquad p^*_{10} = 1 \qquad p^*_{11} = 0$$

and take $\mathcal{D}_{\mathcal{X}}$ to be uniform over $\mathcal{X}$.

Consider any predictor p that satisfies multiaccuracy and calibration. The multiaccuracy constraints can be written as:

$$\begin{aligned} x_0 = 0 : &\quad 1/2 \cdot (p_{00} + p_{01}) = 1/4 \\ x_0 = 1 : &\quad 1/2 \cdot (p_{10} + p_{11}) = 1/2 \\ x_1 = 0 : &\quad 1/2 \cdot (p_{00} + p_{10}) = 1/2 \\ x_1 = 1 : &\quad 1/2 \cdot (p_{01} + p_{11}) = 1/4 \end{aligned}$$

By the first and final equations, we see that multiaccuracy implies $p_{00} = p_{11}$. Further, the equations imply that $p_{01} = 1/2 - p_{00}$ and $p_{10} = 1 - p_{00}$.

Next, we consider the calibration constraints. To do this, we must determine the level sets of p. We argue that $p_{00} = p_{11}$ but are not equal to either p_{01} or p_{10}. In particular, if we set p_{00} such that $p_{01} = 1/2 - p_{00} = p_{00} = 1/4$, or $p_{10} = 1 - p_{00} = p_{00} = 1/2$, then we violate calibration. The expectations over the level set in each of these cases is $1/6$ and $1/3$, respectively, violating the calibration constraint by a constant.

In the alternative case, the level sets are $\{p_{00} = p_{11}\}$,$\{p_{01}\}$, and $\{p_{10}\}$. Under calibration, $p_{00} = p_{11}$ must equal 0, by the fact that the true expectation over these sets is 0. In fact, this implies that p must be within ℓ_1 distance α from p^*.

Finally, we note that any SIM computes a unate function, that is monotone according to some orientation of each $\{x_i\}$. The true function p^*, however, is not close to unate. Thus, no SIM can be close to p^* and, by the above analysis, not SIM can achieve calibrated multiaccuracy. □

8 Experiments

As a proof of concept we implemented a naive version of algorithm calMA (7.2), where the weak learner is instantiated by running linear regression with square loss, and the calibration is instantiated by running isotonic regression. Specifically, given a set of features $x_1, \ldots, x_n$, and a label y, the implementation uses as base classifiers the set of linear functions over the features. Multiaccuracy is obtained since linear regression minimizes square loss. The calibration phase of the algorithm is instantiated by running isotonic regression with a freshly sampled calibration set. Both linear regression and isotonic regression are part of the sklearn Python library, thus the algorithm is remarkably simple to implement and consists of less than 100 lines of Python.

Data: The distribution for the 0 label is a mixture of $s \in \{2, 4\}$ well separated Gaussian distributions over $d \in \{2, 4, 10\}$ dimensions. The distribution for the 1 label is the same as the 0 label, but shifted by a unit vector. As the dimensionality d increases, the classification task becomes easier for a linear predictor, and thus we test the algorithm across a range of loss values. While the distributions are simple, they still pose a challenge for simple predictors and demonstrate well the strength of our approach. See Figure 1 for an example in 2 dimensions, where white dots represent points labeled 0 and green dots are points labeled 1. As the data are synthetic, sample complexity is not an issue. We generated 3000 points for the regression and 1000 points for the calibration.

Metrics: We measured the loss calMA suffers and compared it to linear regression with ℓ_1, ℓ_2, exponential loss and log loss, using the correct link function for each loss function. Linear regression with various loss functions is implemented using Python's scipy.optimize.minimize package.

Results: Results are summarized in the tables below. It is clear that calMA competes remarkably well with the optimal linear predictor across all tested loss functions, occasionally performing even better than the optimal linear predictor for the loss function. As a sanity check in the first table we tested the "omniprediction" of the simple ℓ_2 predictor and indeed found it failed to produce the correct result of the ℓ_1 case and log loss. An interesting example in 2 dimensions is visualized in the figures below. In Figure 1 we see the ground truth classification of labels. Linear regression on its own cannot create multiple clusters as can be seen in Figure 3. However, calMA uses multiple linear classifiers and manages to identify the structure of the clusters, as seen in Figure 2.

Algorithm	ℓ_2	ℓ_1	Exp	Log-loss
Optimal	0.21	0.35	1.54	0.61
calMA	0.20	0.32	1.65	0.635
Linear Regression	0.21	0.43	1.61	1.22

Table 1: $s = 2, d = 2$. Number of iterations is 4

Algorithm	ℓ_2	ℓ_1	Exp	Log-loss
Optimal	0.18	0.28	1.51	0.57
calMA	0.18	0.28	1.53	0.58

Table 2: $s = 4, d = 4$. Number of iterations is 5

Algorithm	ℓ_2	ℓ_1	Exp	Log-loss
Optimal	0.08	0.07	1.55	0.57
calMA	0.06	0.08	1.13	0.22

Table 3: $s = 4, d = 10$. Number of iterations is 3

An example in 2 dimensions, with predictions rounded to $\{0, 1\}$

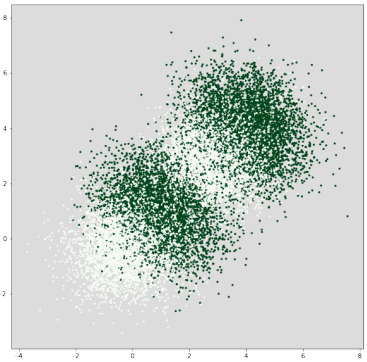

Figure 1: Ground truth

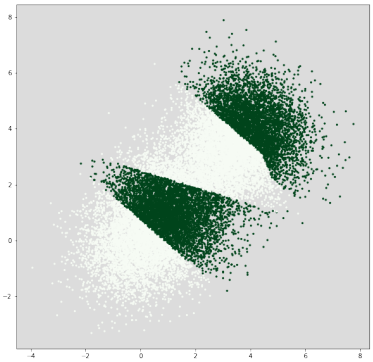

Figure 2: calMA

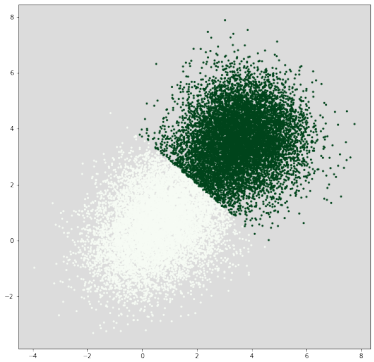

Figure 3: Linear regression

Acknowledgements

PG and **MPK** would like to thank Mihir Singhal and Shengjia Zhao for several discussions while working on [GKSZ22] which inspired some of this work. **PG** would like to thank Adam Klivans, Aravind Gollakota, and Konstantinos Stavropoulos for helpful discussions and comments on earlier versions of this paper and Raghu Meka and Varun Kanade for pointers to the literature.

References

[Agr15] Alan Agresti. *Foundations of Linear and Generalized Linear Models*. Wiley, 2015.

[AHW95] Peter Auer, Mark Herbster, and Manfred K. Warmuth. Exponentially many local minima for single neurons. In *Advances in Neural Information Processing Systems 8, NIPS, Denver, CO, USA, November 27-30, 1995*, pages 316–322. MIT Press, 1995.

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

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006.

[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.

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

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

[GKKT17] Surbhi Goel, Varun Kanade, Adam R. Klivans, and Justin Thaler. Reliably learning the relu in polynomial time. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 1004–1042. PMLR, 2017.

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

[GKSZ22] Parikshit Gopalan, Michael P. Kim, Mihir Singhal, and Shengjia Zhao. Low-degree multicalibration. In *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 3193–3234. PMLR, 2022.

[GRSW22] Parikshit Gopalan, Omer Reingold, Vatsal Sharan, and Udi Wieder. Multicalibrated partitions for importance weights. In *International Conference on Algorithmic Learning Theory, 29-1 April 2022, Paris, France*, volume 167 of *Proceedings of Machine Learning Research*, pages 408–435. PMLR, 2022.

[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.

[HR21] Moritz Hardt and Benjamin Recht. Patterns, predictions, and actions: A story about machine learning. *arXiv preprint arXiv:2102.05242*, 2021.

[Kal04] Adam Kalai. Learning monotonic linear functions. In *Learning Theory, 17th Annual Conference on Learning Theory, COLT 2004*, volume 3120 of *Lecture Notes in Computer Science*, pages 487–501. Springer, 2004.

[Kan18] Varun Kanade. Computational learning theory. learning real-valued functions, Michaelmas Term 2018.

[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.

[KK09] Adam Kalai and Varun Kanade. Potential-based agnostic boosting. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[KKKS11] Sham M. Kakade, Adam Kalai, Varun Kanade, and Ohad Shamir. Efficient learning of generalized linear and single index models with isotonic regression. In *25th Annual Conference on Neural Information Processing Systems 2011.*, pages 927–935, 2011.

[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.

[KP22] Michael P. Kim and Juan C. Perdomo. Making decisions under outcome performativity. *arXiv preprint arXiv:2210.01745*, 2022.

[KS05] Adam Tauman Kalai and Rocco A Servedio. Boosting in the presence of noise. *Journal of Computer and System Sciences*, 71(3):266–290, 2005.

[KS09] Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

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

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

[MN89] P. McCullagh and J. A. Nelder. *Generalized Linear Models (2nd ed.)*. Chapman and Hall, 1989.

[MV08] Hamed Masnadi-Shirazi and Nuno Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. *Advances in neural information processing systems*, 21, 2008.

[Nie10] Frank Nielsen. Legendre transformation and information geometry. Technical Report CIG-MEMO2, September 2010. http://www.informationgeometry.org.

[Nie18] Frank Nielsen. An elementary introduction to information geometry. *CoRR*, abs/1808.08271, 2018.

[Rig16] Philippe Rigollet. Statistics for applications, lecture notes. lecture 10: Generelized linear models., Fall 2016.

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

[SB14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

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

[SSS11] Shai Shalev-Shwartz, Ohad Shamir, and Karthik Sridharan. Learning kernel-based halfspaces with the 0-1 loss. *SIAM J. Comput.*, 40(6):1623–1646, 2011.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[ZE01] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 609–616. Morgan Kaufmann, 2001.

A Additional proofs

A.1 Proofs from Section 3

Proof of Lemma 3.3. Part (1). For $v \in [0,1]$, define $w^*(v) = \text{sign}(\mathbf{E}[\mathbf{y} - v | p(\mathbf{x}) = v])$. For $v \notin \text{Im}(p)$, $w^*(v) \in [-1,1]$ can be arbitrary. Then for any $w \in W_f$,

$$\mathbf{E}[w(p(\mathbf{x}))(\mathbf{y} - p(\mathbf{x}))] \le \mathbf{E}[w^*(p(\mathbf{x}))(\mathbf{y} - p(\mathbf{x}))] = \mathop{\mathbf{E}}_{p(\mathbf{x})}\left[\left|\mathop{\mathbf{E}}_{\mathbf{y}|p(\mathbf{x})}[\,\mathbf{y} - p(\mathbf{x})]\right|\right]$$

which proves the claim.

We prove Part (2) by applying Part (1) to the family $\mathcal{W}' = \mathcal{W}/\left\|\mathcal{W}\right\|_\infty$ which consists of functions bounded in the range $[-1,1]$, and multiplying by $\left\|\mathcal{W}\right\|_\infty$ on either side. □

A.2 Proofs from Section 4

We will use the following result of [GKR+22]. While their result applies to a broader collection of loss functions, we only state it for the class of L_p losses.

Theorem A.1. *[GKR+22] If the predictor $\widetilde{p}$ is $(\mathcal{C}, \alpha)$-multicalibrated, then it is an $(L_p, \mathcal{C}, 2\alpha)$-omnipredictor.*

Proof of Theorem 4.6. The proof uses For this proof alone, it is convenient to use the labels ± 1 rather than $\{0,1\}$, and have predictors $\widetilde{p} : \mathcal{X} \to [-1,1]$ where $\widetilde{p}(\mathbf{x}) = \mathbf{E}[\widetilde{\mathbf{y}} | \mathbf{x} = x]$.

Define the distribution $(\mathbf{x}, \mathbf{y}^*) \sim \mathcal{D}$ where $\mathcal{D}_\mathcal{X}$ is uniform over $\{\pm 1\}^3$ and $\mathbf{y}^* = \chi(\mathbf{x}) = \prod_{i \in [3]} x_i$ is the parity function on 3 bits. Take $\mathcal{C} = \{\sum \alpha_i x_i | \sum_i |\alpha_i| = 1\}$ to be all convex combinations of $\pm x_i$. Consider the predictor $\widetilde{p}(x) = 0$ for all $x \in \{\pm 1\}^3$ so that $\widetilde{\mathbf{y}} | \mathbf{x} = x$ is a uniformly random bit in $\{\pm 1\}$.

We show that $\widetilde{p}$ is $(\mathcal{C}, 0)$-multicalibrated. Since p is constant on the domain, multiaccuracy and multicalibration are equivalent. For reach $i \in [3]$, we have

$$\mathop{\mathbf{E}}_{\mathcal{D}}[x_i(\chi(\mathbf{x}) - \widetilde{\mathbf{y}})] = \mathop{\mathbf{E}}_{\mathcal{D}}[x_i \chi(x)] - \mathop{\mathbf{E}}_{\mathcal{D}}[\chi(\mathbf{x})\mathbf{y}] = 0$$

where the first expectation is 0 because of the orthogonality of characters, and the second because $\mathbf{y}$ and $\mathbf{x}_i$ are independent unbiased random bits. Multiaccuracy for $\mathcal{C}$ now follows from linearity (Lemma 3.6). Now applying Theorem A.1 implies that $\widetilde{p}$ is an $(L_p, \mathcal{C}, 0)$-omnipredictor. Since calibration implies decision-OI, it also implies that $\widetilde{p}$ is decision-OI with 0 error.

Given this, it follows that hypothesis-OI and loss-OI are equivalent, and we show that neither holds. Let $c(x) = (\sum_i x_i)/3$. Then we have

$$\mathbf{E}[\ell_4(\widetilde{\mathbf{y}}, c(x))] = \mathbf{E}[(\widetilde{\mathbf{y}} - c(x))^4] = (1 - 4\,\mathbf{E}[\widetilde{\mathbf{y}} c(x)] + 6\,\mathbf{E}[c(\mathbf{x})^2] - 4\,\mathbf{E}[\widetilde{\mathbf{y}} c(\mathbf{x})^3] + \mathbf{E}[c(\mathbf{x})^4])/4$$
$$\mathbf{E}[\ell_4(\mathbf{y}^*, c(x))] = \mathbf{E}[(\mathbf{y}^* - c(x))^4] = (1 - 4\,\mathbf{E}[\mathbf{y}^* c(x)] + 6\,\mathbf{E}[c(\mathbf{x})^2] - 4\,\mathbf{E}[\mathbf{y}^* c(\mathbf{x})^3] + \mathbf{E}[c(\mathbf{x})^4])/4$$

Hence

$$\mathbf{E}[\ell_4(\widetilde{\mathbf{y}}, c(\mathbf{x}))] - \mathbf{E}[\ell_4(\mathbf{y}^*, c(\mathbf{x}))] = \mathbf{E}[(\mathbf{y}^* - \widetilde{\mathbf{y}}) c(\mathbf{x})] + \mathbf{E}[(\mathbf{y}^* - \widetilde{\mathbf{y}}) c(x)^3] \tag{27}$$

Since $\widetilde{\mathbf{y}}$ is uniform, independent of $\mathbf{x}$, we have $\mathbf{E}[\widetilde{\mathbf{y}}c(\mathbf{x})] = \mathbf{E}[\widetilde{\mathbf{y}}c(\mathbf{x})^3] = 0$. Since $\mathbf{y}^* = \chi(\mathbf{x})$ whereas $c(\mathbf{x}) = (\sum_i \mathbf{x}_i)/3$, by elementary Fourier analysis (see Lemma A.2), we have

$$\mathbf{E}[\mathbf{y}^* c(\mathbf{x})] = 0, \mathbf{E}[\mathbf{y}^* c(\mathbf{x})^3] = 2/9$$

Plugging these bounds back into Equation (27) gives

$$\mathbf{E}[\ell_4(\widetilde{\mathbf{y}}, c(\mathbf{x}))] - \mathbf{E}[\ell_4(\mathbf{y}^*, c(\mathbf{x}))] = 4/9.$$

This shows that hypothesis-OI and hence loss-OI do not hold. □

Lemma A.2. *In the setting of Theorem 4.6, we have*

$$\mathbf{E}[\mathbf{y}^* c(\mathbf{x})] = 0, \tag{28}$$
$$\mathbf{E}[\mathbf{y}^* c(\mathbf{x})^3] = 2/9 \tag{29}$$

Proof. Since $y = \chi(x) = \prod_i x_i$ and $c(x) = (\sum x_i)/3$ Equation (28) follows by orthogonality of characters. For Equation (29),

$$\begin{aligned} c(\mathbf{x})^3 &= \frac{1}{27}\left(\sum_i \mathbf{x}_i\right)^3 \\ &= \frac{1}{27}\left(\sum_i \mathbf{x}_i^3 + \sum_{i\neq j} 3\mathbf{x}_i^2\mathbf{x}_j + 6\mathbf{x}_1\mathbf{x}_2\mathbf{x}_3\right) \\ &= \frac{1}{27}\left(4\sum_{i=1}^3 \mathbf{x}_i + 6\prod_{i=1}^3 \mathbf{x}_i\right). \end{aligned}$$

Hence by the orthogonality of characters,

$$\mathbf{E}[\chi(\mathbf{x})c(\mathbf{x})^3] = \frac{1}{27}\,\mathbf{E}\left[\chi(x)\left(4\sum_i \mathbf{x}_i + 6\prod_{i=1}^3 \mathbf{x}_i\right)\right] = \frac{6}{27}. \tag{30}$$

□

A.3 Proofs from Section 5

Proof of Lemma 5.2. Part (1). Using Equation (13) we have

$$\begin{aligned} \ell_g(p^*, t) &= g(t) - p^* t \\ &= tg'(t) - f(g'(t)) - p^* t \\ &= -f(g'(t) - t(p^* - g'(t)) \\ &= -f(g'(t) - f'(g'(t))(p^* - g'(t)) \\ &= D_f(p^*, g'(t)) - f(p^*) \end{aligned}$$

where the second-last line uses $f'(g'(t)) = t$. We note that this property is a restatement of the Fenchel-Young (in)equality [Nie10, Equation 7], $g(t) + f(p^*) \geq p^* t$ since

$$g(t) + f(p^*) - p^* t = \ell_g(p^*, t) + f(p^*) = D_f(p^*, g'(t)) \geq 0.$$

Part (2). Using Part (1), we can rewrite the objective function in Program (15) as

$$\mathbf{E}[D_f(p^*(\mathbf{x}), g'(h(\mathbf{x}))] = \mathbf{E}[\ell_g(p^*(\mathbf{x}), h(\mathbf{x})] + \mathbf{E}[f(p^*(\mathbf{x}))]$$

Since the $\mathbf{E}[f(p^*(\mathbf{x}))]$ term is independent of h, we can drop it from the objective. Since $p^*(\mathbf{x}) = \mathbf{E}[\mathbf{y}|\mathbf{x}]$, and the definition of $\ell(p, t)$, we have

$$\ell_g(p^*(\mathbf{x}), h(\mathbf{x})) = \mathbf{E}[\ell_g(\mathbf{y}, h(\mathbf{x})|\mathbf{x}]$$

which shows the equivalence to Program (17).

We need to show that this is a convex program. We show that for each $\mathbf{y} \in \{0, 1\}$, $\ell_g(y, t)$ is a convex function of t. Differentiating w.r.t t (or using the fundamental theorem of calculus) we get

$$\frac{d\ell_g(y, t)}{dt} = g'(t) - y. \tag{31}$$

Now we differentiate again and use the fact that $g''(t) \geq 0$ since g is convex. Since we are minimizing a convex loss over the convex set $\mathrm{Lin}(\mathcal{C}, B)$ the resulting program is convex. □

A.4 Proofs from Section 7

Proof of Lemma 7.2. To prove Equation (22), we write

$$\begin{aligned}\left|\mathbf{E}[(p(\mathbf{x}) - p_1(\mathbf{x})^2] - \mathbf{E}[(p(\mathbf{x}) - p_2(\mathbf{x})^2]\right| &= |\mathbf{E}[(p_2(\mathbf{x}) - p_1(\mathbf{x}))(2p(\mathbf{x}) - p_1(\mathbf{x}) - p_2(x)]| \\ &\leq 2\,\mathbf{E}[|p_2(\mathbf{x}) - p_1(\mathbf{x})|] \\ &\leq 2\delta.\end{aligned}$$

The bound on multiaccuracy follows by observing that for any $c : \mathcal{X} \to [-1, 1]$,

$$\begin{aligned}\mathbf{E}[c(\mathbf{x})(\mathbf{y} - p_1(\mathbf{x}))] - \mathbf{E}[c(x)(\mathbf{y} - p_2(\mathbf{x}))] &= \mathbf{E}[c(\mathbf{x})(p_2(\mathbf{x}) - p_1(\mathbf{x}))] \\ &\leq \mathbf{E}[|p_2(\mathbf{x}) - p_1(\mathbf{x})|] \leq \delta.\end{aligned}$$

□

Proof of Lemma 7.4. We take a set of $m = O(\log^2(1/\delta)/(\delta\mu^3))$ samples (x, y) and compute $p^\delta(x)$ for each. For each $j \leq 1/\delta$, let S_j denote the set of samples where $p^\delta(x) = j\delta$ and $m_j = |S_j|$. Define the values

$$\begin{aligned}\bar{y}_j &= \frac{1}{m_j}\sum_i y_i \\ \mathrm{err}_j &= |\bar{y}_j - j\delta| \\ \mathsf{estECE} &= \sum_{j=0}^{1/\delta} \frac{m_j}{m}\mathrm{err}_j\end{aligned} \tag{32}$$

The algorithm returns the value estECE.

We ignore any small values of j such that $\Pr[p^\delta(\mathbf{x}) = j\delta] \le \mu\delta/4$, since except with probability 0.1, such values only contribute $\mu/4$ to $|\mathsf{estECE} - \mathsf{ECE}(p^\delta)|$. Call the other values of j large. For every large j, we have by Chernoff bounds, we have

$$\Pr[m_j \le C(\log(1/\delta)/\mu^2)] \le \frac{\delta}{30}$$

Assuming this event holds, we have

$$\Pr\left[\left|\Pr[p(\mathbf{x}) \in I_j] - \frac{m_j}{m}\right| \ge \frac{\mu}{4}\right] \le \frac{\delta}{30}$$
$$\Pr\left[|\bar{y_j} - \mathbf{E}[\mathbf{y}|p^\delta(\mathbf{x}) = j\delta]| \ge \frac{\mu}{4}\right] \le \frac{\delta}{30}.$$

We take a union bound over all $1/\delta$ large values. Except with error probability 0.2, none of the bad events considered above occur, and we have $|\mathsf{estECE} - \mathsf{ECE}(p^\delta)| \le \mu|$. We can reduce the failure probability by repeating the estimator and taking the median. For simplicity, we ignore the failure probability.

To define the predictor $\widehat{p}$, we repeat the analysis above with $\mu = \delta$. We define $\widehat{p}(x) = \bar{y}_j$ for all $x \in I_j$. We show that it is close to $\bar{p}$ in ℓ_1. The contribution of small values of j to $\mathbf{E}[|\bar{p}(\mathbf{x}) - \widehat{p}(\mathbf{x}))|]$ is no more than $\mu/4$. For large buckets, we have

$$|\bar{y} - \bar{p}(x)| \le \left|\bar{y}_j - \mathbf{E}[\mathbf{y}|p^\delta(\mathbf{x}) = j\delta]\right| \le \delta/2 + \mu/4.$$

Thus overall, the distance is bounded by $(\delta/2 + \mu/4) \le \delta$ by our choice of μ.

Lastly, we bound the calibration error, using the fact that $\bar{p}$ is perfectly calibrated, and $\widehat{p}$ is close to it $\bar{p}$. Note that both $\bar{p}$ and $\widehat{p}$ are constant on all $x \in p^{-1}(I_j)$. Hence

$$\begin{aligned}
\mathsf{ECE}(\widehat{p}) &= \mathop{\mathbf{E}}_{\mathbf{I_j}}\left|\mathop{\mathbf{E}}_{p(\mathbf{x})\in\mathbf{I_j}}[\mathbf{y} - \widehat{p}(\mathbf{x})]\right| \\
&\le \mathop{\mathbf{E}}_{\mathbf{I_j}}\left|\mathop{\mathbf{E}}_{p(\mathbf{x})\in\mathbf{I_j}}[\mathbf{y} - \bar{p}(\mathbf{x})]\right| + \mathop{\mathbf{E}}_{\mathbf{I_j}}\left|\mathop{\mathbf{E}}_{p(\mathbf{x})\in\mathbf{I_j}}[\bar{p}(\mathbf{x}) - \widehat{p}(\mathbf{x})]\right| \\
&= \mathbf{E}[|\bar{p}(\mathbf{x}) - \widehat{p}(\mathbf{x})|] \\
&\le \delta.
\end{aligned}$$

□

Proof of Corollary 7.5. By Lemma 7.3, we know that

$$l_2(p^*, p^\delta)^2 - l_2(p^*, \bar{p})^2 \ge \mathsf{ECE}(p^\delta)^2.$$

By Lemma 7.4, $l_1(\bar{p}, \widehat{p}) \le \delta$. Hence Lemma 7.2 implies that

$$\left|l_2(p^*, \widehat{p})^2 - l_2(p^*, \bar{p})^2\right| \le 2\delta.$$

Similarly, since $l_\infty(p, p^\delta) \le \delta$,

$$\left|l_2(p^*, p)^2 - l_2(p^*, p^\delta)^2\right| \le 2\delta.$$

The claim follows by combining these three equations. □

Proof of Lemma 7.6. Since $\mathrm{WL}(p^* - p_T) = \perp$, by the definition of the weak agnostic learner, for every $c \in \mathcal{C}$,

$$\mathbf{E}[c(\mathbf{x})(p^*(\mathbf{x}) - p_T(\mathbf{x}))] = \mathbf{E}[c(\mathbf{x})(\mathbf{y}^* - p_T(\mathbf{x}))] \leq \sigma.$$

Since $\mathcal{C}$ is closed under negation, the bound also holds in absolute value, hence $p_T \in \mathrm{MA}(\rho) \subseteq \mathrm{MA}(\alpha)$ since $\alpha \geq \rho$.

Assume that $T \geq 1$ and let $t \in \{0, \ldots, T-1\}$. We consider the change in the expected squared error of p_t for every iteration. Note the

$$l_2(p^*, p_t)^2 = \mathbf{E}[(p^*(\mathbf{x}) - p_t(\mathbf{x}))^2] \leq \mathbf{E}[(p^*(\mathbf{x}) - h_{t+1}(\mathbf{x}))^2]$$

since $p_t = \Pi(h_t)$ and projection can only reduce squared error.

$$\begin{aligned}
l_2(p^*, p_t)^2 - l_2(p^*, p_{t+1})^2 &= \mathbf{E}[(p^*(\mathbf{x}) - p_t(\mathbf{x}))^2] - \mathbf{E}[(p^*(\mathbf{x}) - p_{t+1}(\mathbf{x}))^2] \\
&\geq \mathbf{E}[(p^*(\mathbf{x}) - p_t(\mathbf{x}))^2] - \mathbf{E}[(p^*(\mathbf{x}) - h_{t+1}(\mathbf{x}))^2] \\
&= \mathbf{E}[(h_{t+1}(\mathbf{x}) - p_t(\mathbf{x}))(2p^*(\mathbf{x}) - p_t(\mathbf{x}) - h_{t+1}(\mathbf{x})] \\
&= \mathbf{E}[\sigma c'_t(\mathbf{x})(2p^*(\mathbf{x}) - 2p_t(\mathbf{x}) - \sigma c'_t(\mathbf{x})) \\
&= 2\sigma\, \mathbf{E}[c'_t(\mathbf{x})(p^*(\mathbf{x}) - p_t(\mathbf{x}))] - \sigma^2\, \mathbf{E}[c'_t(\mathbf{x})^2] \\
&\geq 2\sigma^2 - \sigma^2 = \sigma^2.
\end{aligned}$$

We sum this over $t \in \{0, \ldots, T-1\}$ to get

$$\mathbf{E}[(p^*(\mathbf{x}) - p_0(\mathbf{x}))^2] - \mathbf{E}[(p^*(\mathbf{x}) - p_T(\mathbf{x}))^2] \geq T\sigma^2. \qquad \square$$

A.5 Complexity analysis for MC

The algorithm from [GKR+22] sets a parameter $\delta \approx \alpha\sigma^2$. It maintains a set of $m = O(1/\delta)$ states, where state j represents a prediction of $j\delta$. In each epoch (where multiple Split operations and a single Merge operation occur), the squared error drops by δ at the cost of $O(m)$ calls to WL. This implies a total of $O(m/\delta) = O(1/\delta^2)$ calls to WL till termination.

One can view the weak learning problem as having a (ρ, σ)-weak learner for arbitrary marginal distributions on $\mathcal{X}$. Alternately, we can stick to the same marginal distribution $\mathcal{D}_{\mathcal{X}}$, but we need to make a stronger assumption on the weak learner, which is for every $\rho \geq 0$, there exists $\sigma(\rho)$ such that if there exists $c \in \mathcal{C}$ such that $\mathbf{E}_{\mathcal{D}}[c(\mathbf{x})f(\mathbf{x})] \geq \rho$, then $\mathrm{WL}(f, \rho) = c' \in \mathcal{C}$ such that $\mathbf{E}_{\mathcal{D}}[c'(\mathbf{x})f(\mathbf{x}))] \geq \sigma(\rho)$. The stronger form of the weak learner is required since the algorithm might present the weak learner with distributions on labels where the correlation with c is rather small; roughly $\Omega(\alpha^3\sigma^3)$, and require it to find a non-trivially correlated hypothesis.