

# Loss Minimization Through the Lens Of Outcome Indistinguishability

**Parikshit Gopalan** ✉

Apple, Cupertino, CA, USA

**Lunjia Hu** ✉

Stanford University, CA, USA

**Michael P. Kim** ✉

Miller Institute, UC Berkeley, CA, USA

**Omer Reingold** ✉

Stanford University, CA, USA

**Udi Wieder** ✉

VMware Research, Palo Alto, CA, USA

---

## Abstract

---

We present a new perspective on loss minimization and the recent notion of Omniprediction through the lens of Outcome Indistinguishability. 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.

**2012 ACM Subject Classification** Theory of computation → Theory and algorithms for application domains

**Keywords and phrases** Loss Minimization, Indistinguishability

**Digital Object Identifier** 10.4230/LIPIcs.ITCS.2023.60

**Related Version** *Full Version:* <https://arxiv.org/abs/2210.08649> [9]

**Funding** *Lunjia Hu:* 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. *Michael P. Kim:* supported by the Miller Institute for Basic Research in Science and, in part, by the Simons Collaboration on the Theory of Algorithmic Fairness.

*Omer Reingold:* supported by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness, the Simons Foundation investigators award 689988, and Sloan Foundation Grant 2020-13941.



© Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder; licensed under Creative Commons License CC-BY 4.0

14th Innovations in Theoretical Computer Science Conference (ITCS 2023).

Editor: Yael Tauman Kalai; Article No. 60; pp. 60:1–60:20

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

**Acknowledgements** We would like to thank Konstantinos Stavropoulos for finding a bug in an earlier version of this paper and suggesting a fix. **PG** and **MPK** would like to thank Mihir Singhal and Shengjia Zhao for several discussions while working on [11] 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.

## 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 [21, 4, 31, 30, 13]. A clean theoretical formulation of the underlying problem is via the notion of agnostic PAC learning [30]. We consider real-valued loss functions  $\ell$  that take two arguments, a label  $y \in \{0, 1\}$  and an action  $t \in \mathbb{R}$ . Given a loss  $\ell$ , a base class of hypotheses  $\mathcal{C}$ , and approximation parameter  $\varepsilon$ , 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  $\mathcal{D}$ :<sup>1</sup>

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

Researchers have devoted significant effort into developing different choices of loss functions [25]. 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 [10] introduces a solution concept for agnostic PAC learning, which they call *omniprediction*. Intuitively, an omnipredictor  $\tilde{p}: \mathcal{X} \rightarrow [0, 1]$  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  $\varepsilon$ . Given any loss  $\ell \in \mathcal{L}$ , a decision-maker can treat  $\tilde{p}(x)$  as if it were the Bayes optimal predictor  $p^*(x) = \mathbf{E}[\mathbf{y}|x]$ , selecting an action  $t$  that will minimize  $\mathbf{E}[\ell(\tilde{\mathbf{y}}, t)]$  where  $\tilde{\mathbf{y}}$  is drawn according to  $\tilde{p}$ . Even though the true labels are drawn according to  $p^*(x)$ , the resulting decision rule is  $\varepsilon$ -optimal for  $\ell$  over  $c \in \mathcal{C}$ . Importantly, the omnipredictor  $\tilde{p}$  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 [10] 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 [14].

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 [6], 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

<sup>1</sup> This version where we do not restrict  $h$  to belong to  $\mathcal{C}$  is sometimes called improper learning.

distribution; in the other simulated world, outcomes  $(\mathbf{x}, \tilde{\mathbf{y}})$  are sampled according to the predictive model  $\tilde{\mathbf{y}} \sim \text{Ber}(\tilde{p}(\mathbf{x}))$ . OI requires the learner to produce a predictor  $\tilde{p}$  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  $\tilde{p}(x)$  and outputs a value in the interval  $[0, 1]$ . For such a collection of algorithms  $\mathcal{A}$  and approximation parameter  $\varepsilon$ , a predictor  $\tilde{p}$  is  $(\mathcal{A}, \varepsilon)$ -outcome indistinguishable<sup>2</sup> if no algorithm  $a \in \mathcal{A}$  can distinguish between the two distributions over individual-outcome pairs.

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

As multicalibration is a special case of OI, by the results of [10], 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 [10] 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 [10, 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 omniprediction, 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  $\tilde{p}(x)$  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  $\tilde{p}$ , 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  $\tilde{p}$  under the two distributions on labels. We give a characterization of these indistinguishability conditions in terms of the *discrete derivative*  $\partial \ell : [0, 1] \rightarrow \mathbb{R}$  of the loss function  $\ell$ , 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 [14, 22] 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.

<sup>2</sup> In fact, [6] introduce a more general hierarchy of OI notions, whose levels are based on the distinguishers' access to the predictions given by  $\tilde{p}$ . The variant where we allow distinguishers access to  $\tilde{p}(\mathbf{x})$  (so-called, *sample-access OI*) is known to be computationally *equivalent* to multicalibration.

### 1.1.1 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  $f : [-1, 1] \rightarrow [-1, 1]$ . 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  $\alpha$ . 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  $\mathcal{C}$ -multiaccuracy – guarantees loss OI for all GLM losses. We give an equivalence between 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 [10] 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 [10], 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).

### 1.1.2 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.

### 1.1.3 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  $\tilde{y}$  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 be 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.

### 1.1.4 Structure of this manuscript

This manuscript represents an Extended Abstract of the full paper, which can be found on the arXiv [9]. The full 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  $p^* : \mathcal{X} \rightarrow [0, 1]$  to denote the Bayes optimal predictor, which represents Nature’s true probability of positive outcomes.

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

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

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

We refer to  $k_\ell$  as the optimal post-processing for  $\ell$ . 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  $\mathcal{C}$ .

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

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

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

## 2.1 Omniprediction from outcome indistinguishability

Omniprediction is a statement about Nature's distribution. Equation (2) makes no mention of the simulated predictions  $\tilde{\mathbf{y}}$ . It is unclear how considering labels  $\tilde{\mathbf{y}}$  from the predictor's simulation might be useful. Indeed, the simulated labels do not play a role in the [10] derivation of omniprediction from multicalibration.

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

$$\mathbf{E}_{\substack{\mathbf{x} \sim \mathcal{D} \\ \tilde{\mathbf{y}} \sim \text{Ber}(\tilde{p}(\mathbf{x}))}}[\ell(\tilde{\mathbf{y}}, k_\ell(\tilde{p}(\mathbf{x})))] \leq \min_{c \in \mathcal{C}} \mathbf{E}_{\substack{\mathbf{x} \sim \mathcal{D} \\ \tilde{\mathbf{y}} \sim \text{Ber}(\tilde{p}(\mathbf{x}))}}[\ell(\tilde{\mathbf{y}}, c(\mathbf{x}))] \quad (3)$$

If  $\tilde{p}$  has the property that the expectations on either side of the Equation don't change much when we replace  $\tilde{\mathbf{y}}$  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.

### 2.1.1 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:

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

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

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

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  $\tilde{p}$  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.

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

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  $\mathcal{L}_{\text{cvx}}$ , it does not imply loss OI even for the  $\ell_4$  loss.

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

The [10] proof of omniprediction was tailored specifically to multicalibration and the specific class of convex loss functions  $\mathcal{L}_{\text{cvx}}$ . 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  $\mathcal{C}$ .

### 2.1.2 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  $\tilde{p}(\mathbf{x})$ . It is known from the work of [6] that when distinguishers receive simultaneous access to  $c(\mathbf{x})$  and  $\tilde{p}(\mathbf{x})$ , 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  $\tilde{p}(\mathbf{x})$  *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*.

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

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

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

Subtracting (7) from (6), we obtain (5), albeit with a slightly larger error parameter. In other words, if  $\tilde{p}$  satisfies both hypothesis OI and decision OI, then  $\tilde{p}$  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  $\alpha$ -calibrated if  $\mathbf{E}[\mathbf{y} | \tilde{p}(\mathbf{x}) = v] \approx_\alpha v$ . Using a more nuanced notion called weighted calibration from [11], we can get an exact characterization of decision OI (see Theorem 17).



## 60:8 Loss Minimization Through the Lens of Outcome Indistinguishability

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  $\tilde{p}$  is  $(\mathcal{C}, \alpha)$ -multiaccurate if for every  $c \in \mathcal{C}$ , the correlation between  $c$  and  $\mathbf{y}^* - \tilde{p}(\mathbf{x})$  is at most  $\alpha$ . Formally, we require

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

For a loss function  $\ell$ , 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  $\tilde{p}$  is  $(\mathcal{L}, \mathcal{C}, \varepsilon)$ -hypothesis OI iff it is  $(\partial\mathcal{L} \circ \mathcal{C}, \varepsilon)$ -multiaccurate. Thus, if  $\tilde{p}$  is  $\varepsilon$ -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.

### 2.1.3 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

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

That is,  $\text{level}(\mathcal{C})$  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  $\text{level}(\mathcal{C})$  stems from the fact that  $\partial\ell \circ c$  belongs to this class, hence  $\text{level}(\mathcal{C})$ -multiaccuracy suffices for Hypothesis-OI over all loss functions.

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

### 2.1.4 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  $\ell$  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  $\text{Int}(\mathcal{C}, \alpha)$  which are the indicators of the event that  $c(x)$  lies in a certain interval  $I \subset [-1, 1]$  of width  $\alpha$ , over all  $c \in \mathcal{C}$  and intervals  $I$ .

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



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

## 2.2 Loss OI in GLMs

GLMs are an important class of models from statistics that generalize linear and logistic regression [26, 1]. On a technical level, GLMs are constructed using the following recipe:

1. We start with an arbitrary monotone increasing transfer function  $g' : \mathbb{R} \rightarrow \mathbb{R}$ , so that its integral  $g(t)$  is convex.
2. We define its matching loss  $\ell_g(y, t) = g(t) - yt$  which is a convex function of  $t$ .
3. We look for the model  $h \in \mathcal{H}$  which minimizes  $\mathbf{E}[\ell_g(y, h(x))]$  where the hypothesis class  $\mathcal{H}$  is taken to be linear combinations over some base class  $\mathcal{C}$ . This gives rise to a convex optimization problem that can be solved efficiently [1, 28].

When we take  $g'(t) = t$ , this recipe gives linear regression with the squared loss. When  $g'(t) = \sigma(t)$  is the sigmoid, we get logistic regression. The class of losses  $\mathcal{L}_{\text{GLM}}$  that arise in this manner are convex. Thus, by the results of [10],  $\mathcal{C}$ -multicalibration suffices to obtain omniprediction for  $\mathcal{L}_{\text{GLM}}$ .

Our first result on GLMs shows that the class  $\partial\mathcal{L}_{\text{GLM}} \circ \mathcal{C} = \mathcal{C}$ . This holds for the simple reason that every loss  $\ell_g \in \mathcal{L}_{\text{GLM}}$  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.

► **Theorem 5 (Informal).** *If  $\tilde{p}$  is  $(\mathcal{C}, \alpha)$ -multiaccurate and calibrated, then it is  $(\mathcal{L}_{\text{GLM}}, \mathcal{C}, O(\alpha))$ -Loss OI.*

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 which shows a reverse connection between multiaccuracy and GLM optimality with  $\ell_1$ -regularization. We state the result informally here.

► **Proposition 6 (Informal).** *For any GLM loss and  $\alpha > 0$ , the optimizer of the  $\ell_1$ -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  $\ell_1$ -regularized GLM learner, like Lasso [33] 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 [34]. 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 (Informal).** *Let  $g'$  be strictly monotonically increasing, let  $f$  be the Legendre dual of  $g$ , and let  $D_f$  be the corresponding Bregman divergence. A predictor  $\tilde{p}$  is  $(\ell_g, \mathcal{H}, \alpha)$ -Loss OI if and only if the following approximate Pythagorean theorem holds approximately.*

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

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  $\tilde{p}$  plus the distance between  $\tilde{p}$  and the predictor from  $\mathcal{H}$ . In other words, if a predictor  $\tilde{p}$  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 class  $\mathcal{C}$ , we define the following classes of predictors.

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

Then we have  $\text{MA}(\alpha) \supseteq \text{calMA}(\alpha) \supseteq \text{MC}(\alpha)$ . We compare the complexity of computing a predictor in each of these classes given access to a  $(\rho, \sigma)$ -weak learner for  $\mathcal{C}$  [3, 19, 17]. 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 in [9] 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 [14] algorithm as analyzed in Lemma 7.6. For calMA, we use our analysis of Algorithm 7.2 in Theorem 17. For MC, we use the analysis of the algorithm from [10, Section 9], which is derived from the boosting by branching programs algorithm by [24].

- For  $\text{MA}(\alpha)$ , the number of calls made by the algorithm of [14] is bounded by  $O(1/\sigma^2)$ .
- For  $\text{calMA}(\alpha)$ , the number of calls made by Algorithm 7.2 bounded by  $O(1/\sigma^2)$ .
- For  $\text{MC}(\alpha)$ , the number of calls made by the algorithm of [10] 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 [11].

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 [18, 15] 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 [12]. For a collection of common losses (including 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 [10] which introduced omnipredictors, and the outcome indistinguishability framework of [6]. 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 [23] to the “performative” prediction setting, where predictions can influence the distribution over outcomes.

Rothblum and Yona [29] 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 [11] 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  $\{MC_d\}$  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 [1, 28]. 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 [28, 26]. The set of convex losses  $\mathcal{L}_{\text{GLM}}$  derived from GLMs are also referred to as matching losses in the literature [2].

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 [10]) hold only for the realizable setting: the data are generated so that  $\mathbf{E}[\mathbf{y}^*|\mathbf{x}] = g'(h(\mathbf{x}))$  for some  $h \in \text{Lin}(\mathcal{C})$ , both  $g'$  and  $h$  are unknown. The first algorithm to give guarantees in this scenario was [16], 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 [18, 15], it is a proper learning algorithm where the output is of the form  $u \circ \tilde{h}$ , where  $\tilde{h} \in \text{Lin}(\mathcal{C})$  and  $u$  is monotone.

Both our work and the work of [10] depart from these works in that they do not require the realizability assumption. We give a single predictor  $\tilde{p}$ , with the guarantee that for any convex  $f$  (with Legendre dual  $g$ ),  $D_f(p^*, \tilde{p})$  is comparable to  $D_f(p^*, g' \circ h)$  for any  $h \in \text{Lin}(\mathcal{C})$ . Under the realizability assumption, for any strongly convex function  $f$  bounding  $D_f$  implies

## 60:12 Loss Minimization Through the Lens of Outcome Indistinguishability

a squared loss bound [15, 20], 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 [32, 8] 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  $\ell_1$  and the squared loss. In these settings, a polynomial dependence on the accuracy parameter  $\varepsilon$  is not possible.

Bregman divergences and Pythagorean theorems for them are studied in information geometry [27, 5], 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 [10, Lemma 8.4]. This guarantee was subsequently shown to hold even with degree-2 multicalibration [11, 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<sup>3</sup>  $\mathbf{y}^* \in \{0, 1\}$ . 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.  $\text{Ber}(p)$  denotes the Bernoulli distribution on  $\{0, 1\}$  with parameter  $p$ . For a real valued function  $f : \mathcal{T} \rightarrow \mathbb{R}$ , let  $\|f\|_{\infty} = \max_{\mathcal{T}} |f(x)|$ . For a family of such functions  $\mathcal{F}$ , let  $\|\mathcal{F}\|_{\infty} = \max_{f \in \mathcal{F}} \|f\|_{\infty}$ .

**Predictors.** A predictor is a function  $\tilde{p} : \mathcal{X} \rightarrow [0, 1]$  be a predictor, where  $\tilde{p}(x)$  is interpreted as an estimate of the label being 1, conditioned on  $x$ . For a predictor  $\tilde{p} : \mathcal{X} \rightarrow [0, 1]$ , we define the distribution  $(\mathbf{x}, \tilde{\mathbf{y}}) \sim \mathcal{D}(\tilde{p})$  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  $\mathbf{x}$ ,  $\tilde{\mathbf{y}} \sim \text{Ber}(\tilde{p}(\mathbf{x}))$  so that

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

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  $\tilde{p}(\mathbf{x}) = v$ , the expected outcome is close to  $v$ .

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

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

► **Definition 1** (ECE and Approximate calibration). *We define the expected calibration error (ECE) of a predictor  $\tilde{p}$  as*

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

For  $\alpha \geq 0$ , a predictor  $\tilde{p} : \mathcal{X} \rightarrow [0, 1]$  is  $\alpha$ -calibrated if  $\text{ECE}(\tilde{p}) \leq \alpha$ .

<sup>3</sup> 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.

A predictor  $\tilde{p}$  is perfectly calibrated if  $\alpha = 0$ , so that  $\mathbf{E}_{\mathcal{D}}[\mathbf{y}^*[\tilde{p}(\mathbf{x}) = v]] = v$ . 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 [7, 14]. For such predictors, one can check for  $\alpha$ -calibration given black-box access to  $\tilde{p}$  in time  $\text{poly}(1/\alpha, 1/\delta)$ , using labeled samples.

Following [11], we will allow for weighted notions of calibration, parametrized by a family of weight functions  $\mathcal{W} = \{w : [0, 1] \rightarrow \mathbb{R}\}$ . Intuitively, we think of a weight function as highlighting predictions belonging to certain regions of  $[0, 1]$ .

► **Definition 2.** Let  $\mathcal{W} = \{w : [0, 1] \rightarrow \mathbb{R}\}$  be a family of weight functions. For a predictor  $\tilde{p} : \mathcal{X} \rightarrow [0, 1]$  we define

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

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

► **Lemma 3.**

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

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

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

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

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  $\tilde{p}$  is  $\alpha$ -smoothly calibrated if  $\text{CE}(\mathcal{W}_1, \tilde{p}) \leq \alpha$ .

**Loss functions and decision functions.** A loss function is a function  $\ell : \{0, 1\} \times \mathbb{R} \rightarrow \mathbb{R}$ . For instance, we define the squared loss by  $\ell_2(y, t) = \|y - t\|_2^2$  and the  $\ell_p$  loss by  $\ell_p(y, t) = \|y - t\|_p^p$ . We define  $b_\ell$ , the Lipschitz constant of  $\ell$ , to be the smallest constant so that  $|\ell(y, t_1) - \ell(y, t_2)| \leq b_\ell |t_1 - t_2|$ . We let  $\text{Lip}_b$  denote the set of all  $b$ -Lipschitz functions. We say that a loss  $\ell$  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  $\ell$  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  $\ell$  via linearity so that the first argument can take values in  $[0, 1]$ . We define

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

A decision function is a function  $k : [0, 1] \rightarrow \mathbb{R}$ . 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  $\ell$ , we define the Bayes-optimal decision function  $k_\ell : [0, 1] \rightarrow \mathbb{R}$  by

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

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

## 60:14 Loss Minimization Through the Lens of Outcome Indistinguishability

**Hypotheses.** A bounded hypothesis class is a family of functions  $\mathcal{C} \subseteq \{c : \mathcal{X} \rightarrow [-1, 1]\}$ . 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  $\text{Lin}(\mathcal{C}, B)$  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  $h \in \text{Lin}(\mathcal{C}, B)$ . We will consider loss minimization problems with the hypothesis class  $\mathcal{H} = \text{Lin}(\mathcal{C}, B)$  (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 [14], *multicalibration* and related notions have seen application beyond fair prediction in recent years. Intuitively, multicalibration requires that the predictions of  $\tilde{p}$  appear calibrated even when we restrict our attention to structured subpopulations. [14] 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 [14, 22], which requires that predictions appear accurate in expectation (unbiased) over each  $c \in \mathcal{C}$ .

► **Definition 4.** Let  $\mathcal{C} = \{c : \mathcal{X} \rightarrow [-1, 1]\}$  be a family of hypotheses and  $\alpha \geq 0$ . We say that the predictor  $\tilde{p} : \mathcal{X} \rightarrow [0, 1]$  is  $(\mathcal{C}, \alpha)$ -multiaccurate if for every  $c \in \mathcal{C}$  it holds that

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

Multicalibration strengthens both calibration and multiaccuracy, requiring approximate calibration over each  $c \in \mathcal{C}$ . We adapt the definitions in [14, 10] to our notion of approximate calibration.

► **Definition 5.** Let  $\mathcal{C} = \{c : \mathcal{X} \rightarrow [-1, 1]\}$  be a family of hypotheses and  $\alpha \geq 0$ . We say that the predictor  $\tilde{p} : \mathcal{X} \rightarrow [0, 1]$  is  $(\mathcal{C}, \alpha)$ -multicalibrated if for every  $c \in \mathcal{C}$  it holds that

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

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  $\alpha$ -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

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

We will generally reserve the term  $(\mathcal{C}, \alpha)$ -multiaccuracy to denote a bounded hypothesis class  $\mathcal{C}$  where  $\text{MAE}(\mathcal{C}, \tilde{p}) \leq \alpha$ . The hypothesis classes  $\mathcal{H}$  most relevant to us are of the form  $\mathcal{H} = \text{Lin}(\mathcal{C}, B)$ . 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 6.** If the predictor  $\tilde{p}$  is  $(\mathcal{C}, \alpha)$ -multiaccurate, then for  $B \geq 1$  and  $\mathcal{H} = \text{Lin}(\mathcal{C}, B)$  we have

$$\text{MAE}(\mathcal{H}, \tilde{p}) \leq B\alpha.$$

**Omnipredictors.** The notion of omniprediction introduced by [10] 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 7.** We say that the predictor  $\tilde{p} : \mathcal{X} \rightarrow [0, 1]$  is an  $(\mathcal{L}, \mathcal{H}, \delta)$ -omnipredictor if for every loss  $\ell \in \mathcal{L}$  and hypothesis  $h \in \mathcal{H}$ ,

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

**Outcome Indistinguishability.** Outcome indistinguishability introduced by [6] provides an elegant framework for reasoning about the quality predictions made by a predictor  $\tilde{p}$ , by measuring their ability to fool statistical tests when nature’s labels  $\mathbf{y}^*$  and replaced by simulated labels  $\tilde{\mathbf{y}}$ . The notion is parameterized by a class of algorithms  $\mathcal{A} \subseteq \{a : \mathcal{X} \times \{0, 1\} \times [0, 1] \rightarrow [-1, 1]\}$ , whose goal is to “distinguish” Nature’s distribution and the modeled distribution.

► **Definition 8 (Outcome Indistinguishability).** A predictor  $\tilde{p} : \mathcal{X} \rightarrow [0, 1]$  is  $(\mathcal{A}, \varepsilon)$ -outcome indistinguishable if for every  $a \in \mathcal{A}$ ,

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

In fact, [6] 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 8 corresponds to “sample-access OI” where the distinguisher receives access to  $\mathbf{x}$ ,  $\tilde{p}(\mathbf{x})$ , and outcomes sampled either from  $\mathbf{y}^* \sim \text{Ber}(p^*(\mathbf{x}))$  or  $\tilde{\mathbf{y}} \sim \text{Ber}(\tilde{p}(\mathbf{x}))$ .

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  $\tilde{p}(\mathbf{x})$ , and simply has access to either  $(\mathbf{x}^*, \mathbf{y}^*) \sim \mathcal{D}$  or  $(\mathbf{x}, \tilde{\mathbf{y}}) \sim \mathcal{D}(\tilde{p})$ . Sample-access OI and No-access OI are in tight correspondence with multicalibration and multiaccuracy, respectively [6]. Another interesting special case of sample-access OI is when we are given access to  $\tilde{p}(\mathbf{x})$  but not to the point  $\mathbf{x}$ . Here, the goal is to distinguish between  $(\mathbf{y}^*, \tilde{p}(\mathbf{x})) \sim \mathcal{D}$  and  $(\tilde{\mathbf{y}}, \tilde{p}(\mathbf{x})) \sim \mathcal{D}(\tilde{p})$ . 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  $\tilde{p}$  with regard to distinguishers that are derived from a loss function  $\ell$ . 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  $k_\ell$  based on the predictor  $\tilde{p}$ .

► **Definition 9 (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  $u_{\ell, c} : \{0, 1\} \times [0, 1] \times \mathcal{X} \rightarrow \mathbb{R}$  by

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

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

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



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  $\tilde{p}$  and computing the expected loss cannot distinguish between  $\mathbf{y}^*$  and  $\tilde{\mathbf{y}}$ .

► **Definition 10** (Decision OI). *Let  $\mathcal{L}$  be a family of loss functions, and  $\varepsilon > 0$ . We say that predictor  $\tilde{p}$  is  $(\mathcal{L}, \varepsilon)$ -decision-OI if for every  $\ell \in \mathcal{L}$  it holds that*

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

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 11** (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  $\tilde{p}$  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*

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

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

► **Lemma 12** (Decomposition lemma). *If the predictor  $\tilde{p}: \mathcal{X} \rightarrow [0, 1]$  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

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

Hence by the triangle inequality,

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

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 13** (Formal Restatement of Proposition 1). *If the predictor  $\tilde{p}: \mathcal{X} \rightarrow [0, 1]$  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

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

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

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

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

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

Plugging this into Equation (10) gives

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

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

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

hence  $\tilde{p}$  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  $\ell_4$  loss. We prove an even stronger statement, that multicalibration does not imply loss-OI. This statement is stronger because of the result of [10] that multicalibration implies omniprediction for a broad class of convex loss functions. We define the  $\ell_p$  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 14.** *There exist a distribution  $\mathcal{D}$ , a class  $\mathcal{C}$  and a predictor  $\tilde{p}$  such that*
  - $\tilde{p}$  is  $(\mathcal{C}, 0)$ -multicalibrated, so it is an  $(L_p, \mathcal{C}, 0)$ -omnipredictor.
  - $\tilde{p}$  is not  $(\{\ell_4\}, \mathcal{C}, \varepsilon)$ -loss OI for any  $\varepsilon < 4/9$ .

The proof which is given in Appendix 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 15.** *Given a loss  $\ell : \{0, 1\} \times \mathbb{R} \rightarrow \mathbb{R}$ , define the function  $\partial\ell : \mathbb{R} \rightarrow \mathbb{R}$  as*

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

The following lemma justifies the analogy to partial derivatives.

- ▶ **Lemma 16.** *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 12, this gives a decomposition of loss OI as a calibration condition and a multiaccuracy condition.

► **Theorem 17.** *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  $\tilde{p}: \mathcal{X} \rightarrow [0, 1]$  is  $(\mathcal{L}, \mathcal{C}, \varepsilon_1)$ -hypothesis-OI where  $\varepsilon_1 = \text{MAE}(\mathcal{C}', \tilde{p})$ .*
2. *Define the family of weight functions  $\mathcal{W}' = \{\partial\ell \circ k_\ell\}_{\ell \in \mathcal{L}}$ . The predictor  $\tilde{p}: \mathcal{X} \rightarrow [0, 1]$  is  $(\mathcal{L}, \varepsilon_2)$ -decision-OI where  $\varepsilon_2 = \text{CE}(\mathcal{W}', \tilde{p})$ .*

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

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

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

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

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  $t = k_\ell(\tilde{p}(x))$ ,

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

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

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

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  $\text{MAE}(\mathcal{C}', \tilde{p})$  is larger than  $\varepsilon'$ , then there exist a  $c, \ell$  pair that distinguishes between  $\mathbf{y}^*$  and  $\tilde{\mathbf{y}}$  with advantage  $\varepsilon'$ .

---

## References

- 1 Alan Agresti. *Foundations of Linear and Generalized Linear Models*. Wiley, 2015.
- 2 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.
- 3 Shai Ben-David, Philip M. Long, and Yishay Mansour. Agnostic boosting. In *14th Annual Conference on Computational Learning Theory, COLT, 2001*.
- 4 Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- 5 Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006. URL: <http://www.elementsofinformationtheory.com/>.
- 6 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. arXiv: 2011.13426.
- 7 Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

- 8 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. URL: <http://proceedings.mlr.press/v65/goel17a.html>.
- 9 Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. *arXiv preprint arXiv:2210.08649*, 2022.
- 10 Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Ominipredictors. In *Innovations in Theoretical Computer Science (ITCS'2022)*, 2022. [arXiv: 2109.05389](https://arxiv.org/abs/2109.05389).
- 11 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.
- 12 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.
- 13 Moritz Hardt and Benjamin Recht. Patterns, predictions, and actions: A story about machine learning. *arXiv preprint*, 2021. [arXiv:2102.05242](https://arxiv.org/abs/2102.05242).
- 14 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*.
- 15 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.
- 16 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. doi:10.1007/978-3-540-27819-1\_34.
- 17 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. doi:10.1145/1374376.1374466.
- 18 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.
- 19 Adam Tauman Kalai and Rocco A. Servedio. Boosting in the presence of noise. *Journal of Computer and System Sciences*, 71(3):266–290, 2005.
- 20 Varun Kanade. Computational learning theory. learning real-valued functions, michaelmas term 2018. URL: <https://www.cs.ox.ac.uk/people/varun.kanade/teaching/CLT-MT2018/>.
- 21 Michael J Kearns and Umesh Vazirani. *An introduction to computational learning theory*. MIT press, 1994.
- 22 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.
- 23 Michael P. Kim and Juan C. Perdomo. Making decisions under outcome performativity. *arXiv preprint arXiv:2210.01745*, 2022.
- 24 Yishay Mansour and David McAllester. Boosting using branching programs. *Journal of Computer and System Sciences*, 64(1):103–112, 2002.
- 25 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.
- 26 P. McCullagh and J. A. Nelder. *Generalized Linear Models (2nd ed.)*. Chapman and Hall, 1989.

## 60:20 Loss Minimization Through the Lens of Outcome Indistinguishability

- 27 Frank Nielsen. An elementary introduction to information geometry. *CoRR*, abs/1808.08271, 2018. [arXiv:1808.08271](https://arxiv.org/abs/1808.08271).
- 28 Philippe Rigollet. Statistics for applications, lecture notes. lecture 10: Generalized linear models, fall 2016.
- 29 Guy N Rothblum and Gal Yona. Multi-group agnostic pac learnability. In *International Conference on Machine Learning*, pages 9107–9115. PMLR, 2021.
- 30 Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- 31 Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.
- 32 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. [doi:10.1137/100806126](https://doi.org/10.1137/100806126).
- 33 Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
- 34 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.