Multigroup Fairness and Omniprediction: Separations and Equivalences

Sílvia Casacuberta[*] Stanford University

Parikshit Gopalan Apple

Varun Kanade[†] University of Oxford

Omer Reingold[‡] Stanford University

Konstantinos Stavropoulos[§] Institute for Advanced Study

Pranay Tankala[¶] Harvard University

Abstract

Omniprediction [1] is a powerful learning guarantee that requires a predictor to be competitive with the best hypothesis from a benchmark class, not just for a single loss function, but simultaneously across all loss functions in a prespecified family. Currently, learning algorithms for achieving omniprediction rely on notions of *multigroup fairness*, such as multiaccuracy and multicalibration [2] or calibrated multiaccuracy [3]. In this work, we ask whether this reliance is necessary: *Does omniprediction require some form of multigroup fairness?* We answer this question in the negative. While prior works have shown that various multigroup fairness notions imply omniprediction [1, 3, 4], we rule out even a weak converse. Specifically, we show that omniprediction for proper losses does not even require accuracy in expectation, a much weaker notion than calibration or multiaccuracy. We complement our negative answer for omniprediction with an affirmative answer for *loss outcome indistinguishability* (loss OI) [3], a related but stronger learning guarantee. Loss OI, which implies omniprediction, requires the predicted label distribution to be indistinguishable from the true label distribution by a certain class of tests depending on the loss functions and the benchmark class. Prior work showed how to achieve loss OI from a combination of calibration and multiaccuracy [3]. We prove the converse, establishing that loss OI is in fact equivalent to a form of calibrated multiaccuracy.

1 Introduction

Consider the setting of binary classification, where labeled data points $(\mathbf{x}, \mathbf{y})$ are sampled from an unknown distribution over a domain $\mathcal{X} \times \{0, 1\}$, and our goal is to learn a *predictor*, which is a function $p : \mathcal{X} \to [0, 1]$ that estimates the probability that the label $\mathbf{y}$ will be 1 given features $\mathbf{x} \in \mathcal{X}$. In this work, we investigate the relationship between the notion of *omniprediction* introduced by [1] and its variants [3, 5], and the various *multigroup fairness* criteria satisfied by the predictor p.

[*]Supported by a Knight-Hennessy Scholarship, the Simons Foundation Collaboration on the Theory of Algorithmic Fairness, and the Simons Foundation investigators award 17351.

[†]This research has been supported in part by EPSRC grant EP/Y028872/1, Mathematical Foundations of Intelligence: An "Erlangen Programme" for AI.

[‡]Supported by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness, and the Simons Foundation investigators award 17351.

[§]This work was done in part while the author was an intern at Apple and while at The University of Texas at Austin. The author was supported in part by the NSF AI Institute for Foundations of Machine Learning (IFML) and, during the 2026–27 academic year, by the Institute for Advanced Study and an anonymous donor.

[¶]Work done during an internship at Apple, as well as with support from Simons Foundation Grant 733782 and Cooperative Agreement CB20ADR0160001 with the United States Census Bureau.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Omniprediction is a powerful new paradigm for loss minimization, first proposed by [1]. Specifically, an $(\mathcal{L}, \mathcal{C})$-*omnipredictor* is a predictor that performs as well as any hypothesis from a baseline class $\mathcal{C}$, for any loss function coming from a family $\mathcal{L}$. The first construction for such a "loss-oblivious" learning notion was achieved through a surprising connection to *multigroup fairness* notions; specifically, through the notion of $\mathcal{C}$-*multicalibration*, which first originated in the algorithmic fairness literature [2].

$\mathcal{C}$-*multicalibration* requires the predictor p to be *calibrated* conditioned on the event $\mathbf{x} \in S$, for all S in a prespecified collection $\mathcal{C} \subseteq 2^{\mathcal{X}}$ of protected subsets of the domain. The weaker notion of $\mathcal{C}$-*multiaccuracy* from [2] requires that the expectation of the predictor matches the expectation of the label, conditioned on membership in $S \in \mathcal{C}$. An intermediate notion is $\mathcal{C}$-*calibrated multiaccuracy*, introduced in [3], which requires both global calibration (without conditioning) and $\mathcal{C}$-multiaccuracy. These notions fall under the umbrella of *multigroup fairness*, a viewpoint that attempts to bridge the gap between *group fairness* and *individual fairness* [6] by allowing the collection of protected groups $\mathcal{C}$ in the domain to be numerous and overlapping.

Although originally motivated by fairness considerations in prediction, multicalibration and multiaccuracy can be viewed as extensions of two classical notions in forecasting: *calibration* ($\mathbf{E}[\mathbf{y}|p(\mathbf{x})] = p(\mathbf{x})$, [7]) and *accuracy in expectation* ($\mathbf{E}[\mathbf{y}] = \mathbf{E}[p(\mathbf{x})]$), respectively. $\mathcal{C}$-multiaccuracy is also equivalent to the notion of $\mathcal{C}$-*computational indistinguishability* in the pseudorandomness and complexity theory literature, where constructions of multiaccurate predictors originally appeared [8]. An active body of research has drawn connections between these notions and topics in learning theory and theoretical computer science, including strong guarantees for learnability [1, 3, 4], pseudorandomness [9, 10] and computational complexity [11, 12].

Multiaccuracy and multicalibration are motivated by fairness, and their definitions do not reference loss functions. But [1] showed that these multigroup fairness notions imply surprisingly strong and robust loss minimization guarantees: namely, $\mathcal{C}$-multicalibration (appropriately generalized to real-valued function classes $\mathcal{C}$) yields omniprediction for all convex, Lipschitz loss functions.[6] This initiated an exciting line of research, which derived various flavors of omniprediction from different multigroup fairness notions [3, 5, 13–16, 4, 12]. Even more recent state-of-the-art constructions of omnipredictors go through some form of multigroup fairness: either multicalibration [1, 14] or calibrated multiaccuracy [16, 4, 15]. This naturally raises the following fundamental question:

Is (some form of) multigroup fairness necessary for omniprediction?
What about calibration or accuracy in expectation?

These were considered in the work of [5], who call it a glaring question in the theory of omniprediction. They showed that multicalibration is both necessary and sufficient for a stronger version of omniprediction called *swap omniprediction*.[7] However, the question of whether multigroup fairness is required for the standard notion of omniprediction raised in that work has remained open.

A strengthening of omniprediction is the notion of $(\mathcal{L}, \mathcal{C})$-*loss outcome indistinguishability (OI)*, introduced by [3]. Intuitively, given a loss function $\ell \in \mathcal{L}$ and hypothesis $c \in \mathcal{C}$, a predictor p satisfies loss OI if the excess loss that p suffers compared to the hypothesis c is roughly unchanged on replacing the real outcomes $\mathbf{y}$ with *modeled* or *simulated* outcomes $\mathbf{y}_p|\mathbf{x} \sim \mathrm{Ber}(p(\mathbf{x}))$. The name "outcome indistinguishability" is derived from the more general OI framework introduced by [9]. It was shown by [3] that calibrated multiaccuracy implies loss OI, which in turn implies omniprediction.

The motivation for studying loss OI is that it can be tailored to specific classes of loss functions, in contrast with known results on omniprediction (e.g., [1] which applies to all Lipschitz, convex losses), and this can give more efficient algorithms (see the discussion in [3]). Indeed, the best known algorithms for omniprediction for specific loss families [4, 15] use the loss OI formulation as their starting point. For instance, the weaker condition of *proper* calibrated multiaccuracy is crucially used by [4] to achieve loss OI for proper losses with optimal rate. However, as in the case of omniprediction, it was unclear from previous work whether either calibration or multiaccuracy is *necessary* for loss OI. The work of [5] showed an equivalence between a stronger variant called *swap* loss OI, and a suitably defined notion of multicalibration.

[6] This requires extending the notion of multicalibration to allow for real-valued functions c rather than indicators of subsets.

[7] Technically, they show an equivalence with *swap multicalibration*, but this is essentially the same as standard multicalibration up to polynomial changes in the error parameters.

1.1 Results Overview

In this work, we show a sharp contrast between the notions of omniprediction and loss OI: while omniprediction does *not* require any of calibration, multiaccuracy, or multicalibration, loss OI is equivalent to a form of calibrated multiaccuracy.

Omniprediction does not require accuracy in expectation (Section 3). Formally, given a hypothesis class $\mathcal{C}$, a family $\mathcal{L}$ of loss functions, and $\varepsilon > 0$, an $(\mathcal{L}, \mathcal{C}, \varepsilon)$-*omnipredictor* is a predictor p such that for every $\ell \in \mathcal{L}$,

$$\mathbf{E}[\ell(\mathbf{y}, k_\ell \circ p(\mathbf{x}))] \leq \min_{c \in \mathcal{C}} \mathbf{E}[\ell(\mathbf{y}, c(\mathbf{x}))] + \varepsilon,$$

where $k_\ell(v) = \arg\min_{t \in \mathcal{A}} \ell(v, t)$ is the optimal action function for each loss function $\ell : [0,1] \times \mathcal{A} \to \mathbb{R}$. Recall that a predictor is *accurate in expectation* if $\mathbf{E}[\mathbf{y}] = \mathbf{E}[p(\mathbf{x})]$. The *expected accuracy error* (EAE) of p is naturally defined as $\mathsf{EAE}(p) = |\mathbf{E}[\mathbf{y}] - \mathbf{E}[p(\mathbf{x})]|$.

While most known constructions of omnipredictors require multigroup fairness notions [1, 14, 16, 4, 15], we show that the converse is not true; namely, we show that omniprediction for proper losses does not require accuracy in expectation. We require that the class $\mathcal{C}$ and marginal distribution $\mathcal{D}_\mathcal{X}$ over $\mathcal{X}$ are such that there exists a Boolean function f which is strongly hard on average for $\mathcal{C}$, meaning that for every $c \in \mathcal{C}$,

$$\left| \mathop{\mathbf{E}}_{\mathcal{D}_\mathcal{X}} \left[c(\mathbf{x}) \Big(f(\mathbf{x}) - \frac{1}{2} \Big) \right] \right| \leq \varepsilon. \tag{1}$$

Theorem 1.1. *Fix $\varepsilon > 0$ and a distribution $\mathcal{D}_\mathcal{X}$ over $\mathcal{X}$. Let $\mathcal{C}$ be any hypothesis class that admits a strong hard-on-average function f with respect to $\mathcal{D}_\mathcal{X}$, as in Equation* (1)*. Consider the distribution $\mathcal{D} = (\mathbf{x}, f(\mathbf{x}))$ for $\mathbf{x} \sim \mathcal{D}_\mathcal{X}$. There exists a predictor $p : \mathcal{X} \to [0,1]$ such that (1) p is an $(\mathcal{L}, \mathcal{C}, O(\varepsilon))$-omnipredictor for the class $\mathcal{L}$ of all convex, Lipschitz losses, and (2) $\mathsf{EAE}(p) \geq 1/4$.*

As we observe in Remark 3.6, the result can also be extended to proper losses using the recent result of [4]. Since accuracy in expectation is an especially mild condition, implied by any one of calibration, multiaccuracy, or multicalibration, this result implies that *none* of these conditions are required for omniprediction. This also shows that neither calibration nor multiaccuracy is necessary for agnostic learning, although together they are sufficient, as shown in [12]. The work of [5] asked if there is *a characterization of standard omniprediction in terms of multigroup fairness notions*; our work answers this in the negative.

The theorem statement above does not address the efficiency of computing the predictor p, but it is essentially identical to that of computing the hard function f. If f is efficiently computable by some class (it would need to be more expressive than $\mathcal{C}$ itself), then so is p.

The key conceptual idea in the proof is to study the closure properties of the set of omnipredictors. We show that they are closed under a between-ness operation, which implies that the set of omnipredictors is quite large. We then show that such a large set must contain points that significantly violate even simple conditions like calibration and accuracy in expectation. The proof involves three steps.

1. First we prove that if q is an omnipredictor, then any predictor r that lies between q and the Bayes optimal predictor $p^* = \mathbf{E}[\mathbf{y}|x]$, point-wise for every $x \in \mathcal{X}$, is also an omnipredictor (Theorem 3.2). Intuitively, moving a predictor point-wise towards the Bayes optimal p^* can only further reduce its loss, and so the omniprediction guarantee satisfied by q is preserved.
2. Second, we apply this observation starting from an omnipredictor q satisfying $\mathbf{E}[|p^*(\mathbf{x}) - q(\mathbf{x})|] \geq \alpha$ for some large α. By decomposing the ℓ_1 loss between p^* and q into the regions $\{x : p^*(x) \geq q(x)\}$ and $\{x : q(x) > p^*(x)\}$, we construct a predictor r that lies between q and p^*, and hence is also an omnipredictor by Theorem 3.2, but which has $\mathsf{EAE}(r) \geq \alpha/2$ (Lemma 3.3). We remark that this holds even in the case where q (and hence r) are perfect omnipredictors (i.e., with $\varepsilon = 0$).
3. Third, we show that for any hypothesis class $\mathcal{C}$ that admits a strong hard-on-average Boolean function, there exists an omnipredictor q satisfying $\mathbf{E}[|p^*(\mathbf{x}) - q(\mathbf{x})|] = 1/2$, and hence by Lemma 3.3 an omnipredictor r satisfying $\mathsf{EAE}(r) \geq 1/4$ (Theorem 3.5).

Loss OI requires calibrated multiaccuracy (Section 4). We show that the situation is very different for the stronger notion of loss OI (which implies omniprediction [3]). Indeed, we prove that, in contrast to omniprediction, both calibration and multiaccuracy are necessary as long as $\mathcal{L}$ contains some simple and natural loss functions. This result establishes a satisfying converse to a key result of [3], which showed that calibration and multiaccuracy are together sufficient for loss OI.

Formally, given a predictor p, we associate a simulated label distribution $\mathcal{D}(p)$ on $(\mathbf{x}, \mathbf{y}_p) \in \mathcal{X} \times \{0,1\}$, where $\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}$ and $\mathbf{y}_p|\mathbf{x} \sim \mathrm{Ber}(p(\mathbf{x}))$. For $\ell \in \mathcal{L}$ and $c \in \mathcal{C}$, we define the distinguisher $u_{\ell,c} : \mathcal{X} \times \{0,1\} \times [0,1] \to \mathbb{R}$ as $u_{\ell,c}(x, y, v) = \ell(y, c(x)) - \ell(y, k_\ell(v))$. We can now formally state the definitions of loss OI, decision OI, and hypothesis OI [3]:

Definition 1.2. *We say that the predictor $p : \mathcal{X} \to [0,1]$ satisfies:*

- $(\mathcal{L}, \mathcal{C}, \alpha)$-loss OI *if for every $\ell \in \mathcal{L}$ and $c \in \mathcal{C}$,*
$$\left|\mathbf{E}\big[u_{\ell,c}(\mathbf{x}, \mathbf{y}, p(\mathbf{x}))\big] - \mathbf{E}\big[u_{\ell,c}(\mathbf{x}, \mathbf{y}_p, p(\mathbf{x}))\big]\right| \leq \alpha.$$
- $(\mathcal{L}, \beta)$-decision OI *if for every $\ell \in \mathcal{L}$,*
$$\left|\mathbf{E}\left[\ell(\mathbf{y}, k_\ell(p(\mathbf{x})))\right] - \mathbf{E}\left[\ell(\mathbf{y}_p, k_\ell(p(\mathbf{x})))\right]\right| \leq \beta.$$
- $(\mathcal{L}, \mathcal{C}, \gamma)$-hypothesis OI *if for every $\ell \in \mathcal{L}$ and every $c \in \mathcal{C}$,*
$$\left|\mathbf{E}\left[\ell(\mathbf{y}, c(\mathbf{x}))\right] - \mathbf{E}\left[\ell(\mathbf{y}_p, c(\mathbf{x}))\right]\right| \leq \gamma.$$

The work of [3] shows that decision OI corresponds to a form of calibration, whereas hypothesis OI corresponds to a form of multiaccuracy. They also show that decision and hypothesis OI together (i.e., a form of calibrated multiaccuracy) suffice for loss OI.

In this paper we show a converse to this statement. The only mild condition that we require on the loss function is for it to have a *neutral action*: an action $t \in \mathcal{A}$ where the loss is the same whether y is 0 or 1. Formally, if $\partial\ell(t) = \ell(1,t) - \ell(0,t)$, then a neutral action is one where $\partial\ell(t) = 0$. For instance, $t = 1/2$ is a neutral action for the squared loss and the cross-entropy loss. In general, a simple shift-based transformation can transform an arbitrary loss into one with a neutral action, while essentially preserving its behavior (see Section 4).

Theorem 1.3. *Given ℓ with a neutral action and $\mathcal{A} \subseteq \mathcal{C}$,[8] if a predictor satisfies $(\ell, \mathcal{C}, \alpha)$-loss OI, then it also satisfies both (ℓ, α)-decision OI and $(\ell, \mathcal{C}, 2\alpha)$-hypothesis OI.*

The idea behind our proof is the following. The key property ensured by a neutral action t is that the loss is independent of the (distribution of the) label and consequently $\mathbf{E}[\ell(\mathbf{y}, t)] = \mathbf{E}[\ell(\mathbf{y}_p, t)]$. Under the assumption that p satisfies $(\ell, \mathcal{C}, \alpha)$-loss OI and that $t \in \mathcal{C}$, this implies that the inequality $\left|\mathbf{E}\big[u_{\ell,c}(\mathbf{x}, \mathbf{y}, p(\mathbf{x}))\big] - \mathbf{E}\big[u_{\ell,c}(\mathbf{x}, \mathbf{y}_p, p(\mathbf{x}))\big]\right| \leq \alpha$ is in particular satisfied with the distinguisher $u_{\ell,t}$ (i.e., for the neutral action t). Using the fact that $\mathbf{E}[\ell(\mathbf{y}, t)] = \mathbf{E}[\ell(\mathbf{y}_p, t)]$, we can turn the loss OI guarantee for $u_{\ell,t}$ into a decision OI guarantee. From here we derive the hypothesis OI guarantee for every $c \in \mathcal{C}$ by applying a triangle inequality and also using the general loss OI guarantee. As we shall discuss in Section 4, the assumption of a neutral action is not strictly necessary and can easily be relaxed or replaced with other, similarly mild assumptions.

Summary. Our results demonstrate an important distinction between the notions of omniprediction and loss OI: the former does not enforce any form of multigroup fairness guarantees (or even accuracy in expectation), while the latter requires a form of calibrated multiaccuracy. This resolves an open question in the literature [5]. Our findings suggest that although most algorithmic results on omniprediction go through the "loss OI route", they might be sacrificing efficiency, since we now know that omnipredictors are "more plentiful" than predictors satisfying loss OI.

1.2 Related Work

Our work continues a line of investigation into connections between multigroup fairness and various notions of learning. Much of this work aims to derive learning guarantees from multigroup fairness

[8] We think of each $t \in \mathcal{A}$ as the constant function that always takes the value t.

notions, including calibration [1, 3, 17, 10, 5, 13–16, 4, 12]. Papers which investigate the reverse connection include [5, 12]. Alternative algorithms for omniprediction, which are based on 2-player game methods rather than boosting algorithms, do achieve optimal error bounds for omniprediction (up to polynomial factors) through online-to-batch conversions [18, 19].

The work of [5] presents tight connections between the swap analogues of omniprediction and loss OI on the one hand and swap multicalibration for families $\mathcal{C}'$ derived from $\mathcal{L}$ and $\mathcal{C}$ on the other hand. From this equivalence they conclude that standard omniprediction, which they show is weaker than swap omniprediction, is unlikely to require multicalibration, which they show is equivalent to swap multicalibration, using a rather indirect chain of reasoning. However, this does not rule out the necessity of some weaker multigroup notion, or even calibration.

Multigroup fairness notions have several connections to complexity theory and pseudorandomness. An indistinguishability perspective on these notions was presented in [9], which motivated the definition of loss OI. Connections to hardcore set construction in complexity theory and questions in pseudorandomness were first noted by [8], and expanded upon in [11, 12, 20]. [10] connect multigroup fairness notions of varying strengths to graph regularity lemmas in combinatorics; see also [21] for more connections to complexity theory. In recent work, Gibbs and Tibshirani [18] proved that when calibration error is measured via expected calibration error (ECE), both multicalibration and calibrated multiaccuracy are strictly harder problems than omniprediction or loss OI in terms of sample complexity. Their result demonstrates a tension between the desiderata of low ECE and multiaccuracy. Their result relies critically on ECE being the calibration measure: while they frame their result as showing that *omniprediction (or rather loss OI) is easier to achieve than multicalibration or calibrated multiaccuracy*, this is only true if calibration error is measured using ECE. Indeed, the result of [4] shows how one can obtain loss OI with optimal sample complexity using predictors that satisfy multiaccuracy and a weaker form of calibration they call *proper calibration*, which captures decision OI for proper losses. Our result shows that proper calibration and multiaccuracy are both necessary to obtain loss OI for proper losses.

2 Preliminaries

We will use boldface (e.g. $\mathbf{x}$) to denote random variables. Given a distribution $\mathcal{D}$ over pairs $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \{0, 1\}$, we denote the marginal distribution of $\mathbf{x}$ by $\mathcal{D}_{\mathcal{X}}$. A *predictor* is a function $p : \mathcal{X} \to [0, 1]$, a *hypothesis* is a function $c : \mathcal{X} \to [-1, 1]$, and a *loss* is a function $\ell : \{0, 1\} \times \mathcal{A} \to \mathbb{R}$ for a set of *actions* $\mathcal{A}$. The *Bayes optimal predictor* is given by $p^*(x) = \mathbf{E}[\mathbf{y}|x]$. The reason to have both predictors and hypotheses is that certain notions like calibration are traditionally defined for $[0, 1]$-valued functions, whereas others like correlation are more natural for $[-1, 1]$-valued functions. To every predictor p, we associate a simulated label distribution $\mathcal{D}(p)$ on $(\mathbf{x}, \mathbf{y}_p) \in \mathcal{X} \times \{0, 1\}$ where $\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}$ and $\mathbf{y}_p|\mathbf{x} \sim \mathrm{Ber}(p(\mathbf{x}))$. In other words, $\mathcal{D}(p)$ is a distribution whose Bayes optimal predictor is p. By this notation, $\mathcal{D} = \mathcal{D}(p^*)$; we prefer the former for simplicity.

Multigroup fairness notions. We now define *multiaccuracy*, *multicalibration*, and (global, weight-restricted) *calibration*:

Definition 2.1 ([2, 22, 23])**.** *Fix a hypothesis class $\mathcal{C}$ of hypotheses $c : \mathcal{X} \to [-1, 1]$ and a class $\mathcal{W}$ of* weight functions *$w : [0, 1] \to [-1, 1]$. We say a predictor p satisfies* $(\mathcal{C}, \delta)$-multiaccuracy *(with respect to a distribution $\mathcal{D}$ over $(\mathbf{x}, \mathbf{y})$ pairs) if*

$$\max_{c \in \mathcal{C}} \Big| \mathbf{E}\big[c(\mathbf{x})(\mathbf{y} - p(\mathbf{x}))\big] \Big| \leq \delta.$$

We say that p satisfies $(\mathcal{C}, \delta)$-multicalibration *if*

$$\max_{c \in \mathcal{C}} \mathbf{E} \Big| \mathbf{E}\big[c(\mathbf{x})(\mathbf{y} - p(\mathbf{x})) \mid p(\mathbf{x})\big] \Big| \leq \delta.$$

We say that p satisfies $(\mathcal{W}, \delta)$-weight restricted calibration *if*

$$\max_{w \in \mathcal{W}} \Big| \mathbf{E}\big[w(p(\mathbf{x}))(\mathbf{y} - p(\mathbf{x}))\big] \Big| \leq \delta.$$

Assume that $\mathcal{C}$ and $\mathcal{W}$ are closed under complement and contain the constant function 1. We define the *expected accuracy error* (EAE) and the *expected calibration error* (ECE) of a predictor as:

$$\mathsf{EAE}(p) = \Big| \mathbf{E}[\mathbf{y}] - \mathbf{E}[p(\mathbf{x})] \Big|, \quad \mathsf{ECE}(p) = \mathbf{E} \Big| \mathbf{E}[\mathbf{y}|p(\mathbf{x})] - p(\mathbf{x}) \Big|.$$

Observe that $\mathsf{EAE}(p) \leq \mathsf{ECE}(p)$, $(\mathcal{C}, \delta)$-multicalibration implies $\mathsf{ECE}(p) \leq \delta$, and either one of $(\mathcal{C}, \delta)$-multiaccuracy or $(\mathcal{W}, \delta)$-weight restricted calibration implies $\mathsf{EAE}(p) \leq \delta$ (since $1 \in \mathcal{C}, \mathcal{W}$).

Loss functions. In a loss minimization task, on input $\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}$, we take an action $t \in \mathcal{A}$ (typically based on a prediction $p(\mathbf{x})$ of the label), and suffer loss $\ell(\mathbf{y}, t)$ for a loss function $\ell : \{0,1\} \times \mathcal{A} \to \mathbb{R}$. We allow for abstract action spaces $\mathcal{A}$ which might be discrete or continuous. Typically, we pick the optimal action from a hypothesis class of functions $\mathcal{C} = \{c : \mathcal{X} \to \mathcal{A}\}$. This results in the loss minimization problem

$$\min_{c \in \mathcal{C}} \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\ell(\mathbf{y}, c(\mathbf{x}))], \tag{2}$$

which we shall refer to as *agnostically learning $\mathcal{C}$ with the loss function ℓ*.

We extend ℓ linearly to a function $\ell : [0,1] \times \mathcal{A} \to \mathbb{R}$ by defining

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

An important class of loss functions is the class of *proper losses*:

Definition 2.2. *We say that a loss ℓ is* proper *if $\mathcal{A} = [0,1]$ and $\ell(p, q) \geq \ell(p, p)$ for all $q \in [0,1]$.*

In other words, if $\mathbf{y} \sim \mathrm{Ber}(p)$, then the expected loss is minimized by the action $q = p$.

Recall that given a loss function $\ell : [0,1] \times \mathcal{A} \to \mathbb{R}$, the optimal action function $k_\ell : [0,1] \to \mathcal{A}$ is given by $k_\ell(v) = \arg\min_{t \in \mathcal{A}} \ell(v, t)$. [24] reduce the case of arbitrary loss functions to proper losses, by showing how we can post-process an arbitrary loss function in order to make it proper:

Lemma 2.3 ([24])**.** *Given a loss function $\ell : [0,1] \times \mathcal{A} \to \mathbb{R}$, the loss function $\ell^{\mathrm{pr}} : [0,1] \times [0,1] \to \mathbb{R}$ defined as $\ell^{\mathrm{pr}}(p, q) = \ell(p, k_\ell(q))$ is a proper loss function.*

We will use the well-known characterization of proper loss functions in terms of Bregman divergences [25, 26], which we define below.

Definition 2.4. *Given a concave function $h : [0,1] \to \mathbb{R}$, the* associated Bregman divergence *$D_h : [0,1] \times [0,1] \to \mathbb{R}^+$ is given by*

$$D_h(p, q) = h(q) + (p - q)h'(q) - h(p), \tag{3}$$

where h' is the derivative of h in the case that h happens to be differentiable, or an arbitrary superderivative of h otherwise.

Theorem 2.5 ([25, 26])**.** *If $\ell^{\mathrm{pr}} : [0,1] \times [0,1] \to \mathbb{R}$ is a proper loss, then there exists a concave function h such that*

$$\ell^{\mathrm{pr}}(p, q) = D_h(p, q) + h(p) = h(q) + (p - q)h'(q),$$

where h' is the derivative of h in the case that h happens to be differentiable, or a superderivative of h otherwise. Moreover, h is given by the formula $h(p) = \ell^{\mathrm{pr}}(p, p)$, and $h'(p) = \ell^{\mathrm{pr}}(1, p) - \ell^{\mathrm{pr}}(0, p)$ is a superderivative.

The key property of $\ell^{\mathrm{pr}}(p, q)$ is that it is linear in p, so it can be viewed as the linear extension of $\ell^{\mathrm{pr}}(y, q) = h(q) + (y - q)h'(q)$. If the loss functions we consider are bounded to lie in the range $[0,1]$, then it follows that $h'(q) = \ell^{\mathrm{pr}}(1, q) - \ell^{\mathrm{pr}}(0, q) \in [-1, 1]$.

Due to the space constraints, we defer all proofs to the appendix.

3 Omniprediction and Accuracy in Expectation

In this section, we state and prove our result showing that omniprediction does not even require the very mild notion of accuracy in expectation.

Informally, a predictor p is an $(\mathcal{L}, \mathcal{C})$-*omnipredictor* if for every loss $\ell \in \mathcal{L}$, we would achieve better loss by not taking actions prescribed by any $c \in \mathcal{C}$ but instead taking the *optimal action according to* p, which we now define. Specifically, given a loss function $\ell : [0,1] \times \mathcal{A} \to \mathbb{R}$, the *optimal action* function $k_\ell : [0,1] \to \mathcal{A}$ is defined as $k_\ell(v) = \arg\min_{t \in \mathcal{A}} \ell(v, t)$.[9]

[9] If there are multiple actions that minimize ℓ, we pick one arbitrarily.

We remark that the function k_ℓ is loss-specific, but is typically simple for standard loss functions. An omnipredictor allows for simultaneous loss minimization across a large class of loss functions, without considering each loss separately:

Definition 3.1 ([1]). *For a family $\mathcal{L}$ of loss functions, a hypothesis class $\mathcal{C}$ and $\varepsilon > 0$, an* $(\mathcal{L}, \mathcal{C}, \varepsilon)$-omnipredictor *is a predictor p such that for every $\ell \in \mathcal{L}$, we have*

$$\mathbf{E}[\ell(\mathbf{y}, k_\ell \circ p(\mathbf{x}))] \leq \min_{c \in \mathcal{C}} \mathbf{E}[\ell(\mathbf{y}, c(\mathbf{x}))] + \varepsilon \tag{4}$$

The right hand side is the loss incurred by agnostically learning $\mathcal{C}$ with loss ℓ up to an additive factor of ε, the optimal solution to which will typically depend on the choice of ℓ. The power of omniprediction comes from the fact that the predictor p is learned from the data in one shot. Its predictions can be used to minimize any loss from $\mathcal{L}$ relative to $\mathcal{C}$ without any further learning. The existence of omnipredictors follows from observing that p^* is an omnipredictor for all $(\mathcal{L}, \mathcal{C})$. The key challenge with omnipredictors is efficient learning. [1] first showed how to construct omnipredictors using a weak agnostic learning oracle for $\mathcal{C}$.

Let $\mathcal{P} = \{p : \mathcal{X} \to [0, 1]\}$ denote the space of all predictors, and let $\mathrm{OP}_{\mathcal{L},\mathcal{C},\varepsilon} \subseteq \mathcal{P}$ denote the set of all $(\mathcal{L}, \mathcal{C}, \varepsilon)$-omnipredictors. In this section, we explore closure properties of $\mathrm{OP}_{\mathcal{L},\mathcal{C},\varepsilon}$ under two simple operations. We can also consider the case of *perfect* omniprediction for the case when $\varepsilon = 0$; we will denote $\mathrm{OP}_{\mathcal{L},\mathcal{C},0}$ by $\mathrm{OP}_{\mathcal{L},\mathcal{C}}$.

Convex combinations. Given $p, q \in \mathcal{P}$ and $\lambda \in [0, 1]$, define the predictor $r = \lambda p + (1 - \lambda)q \in \mathcal{P}$ by $r(x) = \lambda p(x) + (1 - \lambda)q(x)$. Let $\mathrm{Conv}(p, q)$ be the set of all such convex combinations of p, q.

Between-ness. Given $p, q \in \mathcal{P}$, we define $\mathrm{Btwn}(p, q)$ to be the set of predictors $r \in \mathcal{P}$ such that for every $x \in \mathcal{X}$, $\min(p(x), q(x)) \leq r(x) \leq \max(p(x), q(x))$.

We say that such an r *lies between* p and q. Alternately, we can view each $r \in \mathrm{Btwn}(p, q)$ as being obtained by choosing $\lambda(x) \in [0, 1]$ independently for every $x \in \mathcal{X}$, and setting $r(x) = \lambda(x)p(x) + (1 - \lambda(x))q(x)$. Note that the set $\mathrm{Btwn}(p, q)$ is convex and $\mathrm{Conv}(p, q) \subseteq \mathrm{Btwn}(p, q)$. Observe that $\mathrm{Btwn}(p, q)$ may be $|\mathcal{X}|$-dimensional, e.g. if $p(x) \neq q(x)$ for every $x \in \mathcal{X}$, as opposed to $\mathrm{Conv}(p, q)$ which is one dimensional, so $\mathrm{Btwn}(p, q)$ is a much larger set.

We first show that if q is an omnipredictor, then any r which lies between q and the Bayes optimal predictor p^* is also an omnipredictor. The main idea is to show that starting from an omnipredictor and moving towards the Bayes optimal predictor only reduces loss, hence it preserves the property of omniprediction. The proof uses a result of [24] to reduce to the case of proper losses, and a characterization of proper losses in terms of Bregman divergences of concave functions [25, 26]. Formally:

Theorem 3.2. *For any classes $\mathcal{L}$ and $\mathcal{C}$, if $q \in \mathrm{OP}_{\mathcal{L},\mathcal{C},\varepsilon}$, then $\mathrm{Btwn}(p^*, q) \subseteq \mathrm{OP}_{\mathcal{L},\mathcal{C},\varepsilon}$.*

An important consequence is that omnipredictors do not even need to be accurate in expectation. Note that accuracy in expectation is a very weak requirement which is implied by both calibration and multiaccuracy. This holds even in the case of perfect omniprediction with $\varepsilon = 0$.

Lemma 3.3. *Assume there exists $q \in \mathrm{OP}_{\mathcal{L},\mathcal{C}}$ such that $\mathbf{E}[|p^*(\mathbf{x}) - q(\mathbf{x})|] \geq \alpha$. Then there exists $r \in \mathrm{OP}_{\mathcal{L},\mathcal{C}}$ such that $\mathsf{EAE}(r) \geq \alpha/2$.*

For simple classes, we can explicitly construct examples of a p^* and an omnipredictor for which accuracy in expectation does not hold. We outline such a construction for the case of coordinate functions when the ground truth p^* is the parity function.

- $\mathcal{D}^*$ is the distribution where $\mathbf{x}$ is uniform on $\{0, 1\}^k$ and $\mathbf{y} = \chi(\mathbf{x})$ is the parity of the bits.
- Let $\mathcal{L}$ be all convex, Lipschitz losses and $\mathcal{C} = \{x \mapsto x_i \mid i \in [k]\}$ be all coordinate functions. (We could alternatively consider parities, conjunctions, or disjunctions on strictly fewer than k bits.)
- The constant predictor $q(x) = 1/2$ is $(\mathcal{C}, 0)$-multicalibrated. So it is an $(\mathcal{L}, \mathcal{C}, 0)$-omnipredictor by the main result of [1].
- The predictor r defined as $r(x) = 1/2$ when $\chi(x) = 0$ and $r(x) = 1$ when $\chi(x) = 1$ is also an $(\mathcal{L}, \mathcal{C}, 0)$-omnipredictor since it lies between q and p^*. But $\mathbf{E}[r(\mathbf{x})] = 3/4$ whereas $\mathbf{E}[\mathbf{y}] = 1/2$, so $\mathsf{EAE}(r) = 1/4$.

The question then is whether there always exist omnipredictors q for which $\mathbf{E}[|p^*(\mathbf{x}) - q(\mathbf{x})|]$ is large. Under mild complexity-theoretic assumptions on the hypothesis class $\mathcal{C}$ and the marginal distribution $\mathcal{D}_{\mathcal{X}}$, we show that such omnipredictors always exist. Our assumption asserts the existence of a Boolean function which is extremely hard for $\mathcal{C}$ under the distribution $\mathcal{D}_{\mathcal{X}}$.

Definition 3.4. *For a distribution $\mathcal{D}_{\mathcal{X}}$ on $\mathcal{X}$, a hypothesis class $\mathcal{C}$ and $\varepsilon > 0$, we say that the Boolean function $f : \mathcal{X} \to \{0,1\}$ is* $(1/2 - \varepsilon)$-hard *for $\mathcal{C}$ under $\mathcal{D}_{\mathcal{X}}$ if for every $c \in \mathcal{C}$,*

$$\left| \mathop{\mathbf{E}}_{\mathcal{D}_{\mathcal{X}}} \left[c(\mathbf{x}) \Big(f(\mathbf{x}) - \frac{1}{2} \Big) \right] \right| \leq \varepsilon. \tag{5}$$

To see why this implies average-case hardness, note that for Boolean functions c, and a balanced function f, this definition implies that

$$\Pr_{\mathcal{D}_{\mathcal{X}}}[c(x) = f(x)] \leq 1/2 + 2\varepsilon.$$

Our result applies to the class $\mathcal{L}_c$ of convex, Lipschitz loss functions. These are loss functions where $\mathcal{A} = \mathbb{R}$, hence $\ell : \{0,1\} \times \mathbb{R} \to \mathbb{R}$, and $\ell(y,t)$ is a convex, Lipschitz function of t for $y \in \{0,1\}$. The main result of [1] shows that if p is $\mathcal{C}$-multicalibrated, then it is an $(\mathcal{L}_c, \mathcal{C})$-omnipredictor. We use this to show the following:

Theorem 3.5. *Let $\mathcal{C}$ be a hypothesis class of functions with range $[-1,1]$, $\mathcal{D}_{\mathcal{X}}$ be a distribution over $\mathcal{X}$, and let $\mathcal{L}$ be the class of all convex, Lipschitz losses.*[10] *Suppose that there exists a Boolean function $p^* : \mathcal{X} \to \{0,1\}$ and $\varepsilon > 0$ such that p^* is $(1/2 - \varepsilon)$-hard for $\mathcal{C}$ under $\mathcal{D}_{\mathcal{X}}$. Then, for the distribution $\mathcal{D} = \mathcal{D}(p^*)$, there exists a predictor $r : \mathcal{X} \to [0,1]$ such that r is an $(\mathcal{L}, \mathcal{C}, O(\varepsilon))$-omnipredictor and $\mathsf{EAE}(r) \geq \frac{1}{4}$.*

Remark 3.6. *The result of Theorem 3.5 extends beyond the class $\mathcal{L}_c$ of convex, Lipschitz loss functions. The recent work by [4] shows that if q is $(\mathrm{Th}(\mathcal{C}), \varepsilon)$-multicalibrated, then it is an $(\mathcal{L}_{BV}, \mathcal{C}, O(\varepsilon))$-omnipredictor for the class $\mathcal{L}_{BV}$ of bounded variation losses, which include all convex, all Lipschitz, and all proper losses, where $\mathrm{Th}(\mathcal{C})$ denotes the concept class consisting of thresholds applied to $\mathcal{C}$. Applying this result to the proof of Theorem 3.5, we similarly obtain a predictor r that is an $(\mathcal{L}_{BV}, \mathcal{C}, O(\varepsilon))$-omnipredictor and $\mathsf{EAE}(r) \geq \frac{1}{4}$.*

Another consequence of this construction is that neither calibration nor multiaccuracy is necessary for strong agnostic learning (although together they are sufficient [3, 12]). The proof of the result below follows directly from the above construction by considering the ℓ_1 loss. This result can be made more general as in the case of Theorem 3.5.

Corollary 3.7. *Under the assumptions of Theorem 3.5, there exists a predictor $p : \mathcal{X} \to [0,1]$ and a post-processing function $\kappa : [0,1] \to [-1,1]$ such that*

$$\mathbf{E}[(2\mathbf{y} - 1)\kappa(p(\mathbf{x}))] \geq \max_{c \in \mathcal{C}} \mathbf{E}[(2\mathbf{y} - 1)c(\mathbf{x})] - O(\varepsilon),$$

but $\mathsf{EAE}(p) \geq 1/4$.

For the ℓ_1 loss, the optimal post-processing function corresponds to $\kappa(v) = 2 \cdot \mathbb{1}[v \geq 1/2] - 1$.

On the existence of average-case hard functions. For what classes $\mathcal{C}$ and distributions $\mathcal{D}_{\mathcal{X}}$ can we expect there to be a Boolean function f that satisfies Definition 3.4? Clearly we need the $\mathcal{D}_{\mathcal{X}}$ distribution to have sufficient min-entropy: else even a constant function will give a sufficiently good approximation to f. So let us assume this to be the case. We will consider the standard complexity theory setting where $\mathcal{X} = \{0,1\}^n$ and $\mathcal{D}_{\mathcal{X}}$ is the uniform distribution; analogous statements hold for other domains and (high-entropy) distributions, but are tedious to formalize.

We now focus on the class $\mathcal{C}$. The answer depends on whether we want the omnipredictor to be efficiently computable or not. The definition of an omnipredictor itself does not require it to be efficiently computable. Yet, as noted in [1], the definition is most interesting when the predictor is not only efficiently computable but also learnable; this typically means that it lies in a more powerful class than $\mathcal{C}$. Hence, we will consider both settings: the information-theoretic setting where the omnipredictor is not required to be computationally efficient, and the setting where it is.

[10] By *convex*, we mean that both functions $\ell(0,\cdot)$ and $\ell(1,\cdot)$ are convex, for each $\ell \in \mathcal{L}$. As in [1] we can relax this to include all (B,ε)-nice functions and drop the global Lipschitz requirement.

Consider the information-theoretic setting where we do not require p^* or p to be efficiently computable. As long as functions in the class $\mathcal{C}$ have bounded description lengths, or small covering numbers (in the continuous setting), a simple counting argument implies that there exists a p^* which is average-case hard for $\mathcal{C}$. This gives an omnipredictor p satisfying the guarantees of Theorem 3.5.

The more challenging setting is where we require a Boolean p^* that is hard for $\mathcal{C}$ under $\mathcal{D}_{\mathcal{X}}$, but is efficiently computable by a more powerful class. This requires making unproven but plausible computational assumptions about $\mathcal{C}$. Similar assumptions are used in the seminal work of Nisan and Wigderson [27] to construct pseudorandom generators that can be used to fool statistical tests modeled by functions from $\mathcal{C}$. Such assumptions are stronger than worst-case hardness assumptions, but weaker than the average-case hardness assumptions on which cryptography is based; see Impagliazzo's well-known Five Worlds survey [28].

4 Loss OI and Calibrated Multiaccuracy

In this section, we present our result characterizing loss OI [3] as equivalent in power to calibrated multiaccuracy. Loss OI is an instantiation of the general notion of sample-access *outcome indistinguishability* (OI) from [9], which concerns the ability of a prespecified class of *distinguisher* functions to tell apart genuine data points of the form $(\mathbf{x}, \mathbf{y}, p(\mathbf{x}))$ from modeled ones of the form $(\mathbf{x}, \mathbf{y}_p, p(\mathbf{x}))$, where $\mathbf{y}_p|\mathbf{x} \sim \mathrm{Ber}(p(\mathbf{x}))$. It was shown in [3] that loss OI, while stronger than omniprediction, is nevertheless implied by calibrated multiaccuracy. The present result can be viewed as a converse to that prior result, fully characterizing loss OI as *equivalent* to calibrated multiaccuracy. This is in contrast to our result in the previous section, which showed that the weaker notion of omniprediction does not even require accuracy in expectation. We start with the definition of loss outcome indistinguishability:

Definition 4.1 (Loss Outcome Indistinguishability [3])**.** *For $\ell \in \mathcal{L}$ and $c \in \mathcal{C}$, we define the distinguisher $u_{\ell,c} : \mathcal{X} \times \{0,1\} \times [0,1] \to \mathbb{R}$ as $u_{\ell,c}(x, y, v) = \ell(y, c(x)) - \ell(y, k_\ell(v))$. We say the predictor $p : \mathcal{X} \to [0,1]$ is $(\mathcal{L}, \mathcal{C}, \alpha)$*-loss OI *if for every $\ell \in \mathcal{L}$ and $c \in \mathcal{C}$,*

$$\Big|\mathbf{E}\big[u_{\ell,c}(\mathbf{x}, \mathbf{y}, p(\mathbf{x}))\big] - \mathbf{E}\big[u_{\ell,c}(\mathbf{x}, \mathbf{y}_p, p(\mathbf{x}))\big]\Big| \le \alpha. \tag{6}$$

To gain intuition for the distinguisher $u_{\ell,c}$ and the notion of loss OI, we make three key observations. First, we note that the test $u_{\ell,c}$ compares—for each labeled instance (x, y) and prediction value v—the difference in loss suffered by two actions: the action $c(x)$ suggested by the hypothesis c and the action $k_\ell(v)$, which is the optimal action when $\mathbf{y} \sim \mathrm{Ber}(v)$. Next, we note $u_{\ell,c}$ is always nonnegative in expectation when fed prediction values v generated by the predictor p and simulated labels $\mathbf{y}_p|\mathbf{x} \sim \mathrm{Ber}(p(\mathbf{x}))$. To see this, observe that by the definition of k_ℓ,

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

This is because $\mathbf{y}_p|\mathbf{x} \sim \mathrm{Ber}(p(\mathbf{x}))$, so $k_\ell \circ p(\mathbf{x})$ incurs the lowest loss among all possible actions. Finally, we observe that if a predictor p satisfies (perfect) loss OI, then it too incurs less loss than c, because we can switch $\mathbf{y}_p$ and $\mathbf{y}$ to get

$$\mathop{\mathbf{E}}_{\mathcal{D}}\big[\ell(\mathbf{y}, c(\mathbf{x}))\big] - \mathop{\mathbf{E}}_{\mathcal{D}}\big[\ell(\mathbf{y}, k_\ell(p(\mathbf{x})))\big] = \mathbf{E}\big[u_{\ell,c}(\mathbf{x}, \mathbf{y}, p(\mathbf{x}))\big] = \mathbf{E}\big[u_{\ell,c}(\mathbf{x}, \mathbf{y}_p, p(\mathbf{x}))\big] \ge 0. \tag{7}$$

This is precisely the condition required for omniprediction.

Having established the relationship between loss OI and omniprediction, we now recall the two simpler notions of *decision OI* and *hypothesis OI*, also introduced in [3], which together are sufficient for loss OI.

Definition 4.2 (Decision OI)**.** *The predictor $p : \mathcal{X} \to [0,1]$ satisfies $(\mathcal{L}, \beta)$*-decision OI *if for every $\ell \in \mathcal{L}$ it holds that*

$$\Big|\mathbf{E}\left[\ell(\mathbf{y}, k_\ell(p(\mathbf{x})))\right] - \mathbf{E}\left[\ell(\mathbf{y}_p, k_\ell(p(\mathbf{x})))\right]\Big| \le \beta. \tag{8}$$

Definition 4.3 (Hypothesis OI)**.** *The predictor $p : \mathcal{X} \to [0,1]$ satisfies $(\mathcal{L}, \mathcal{C}, \gamma)$*-hypothesis OI *if for every $\ell \in \mathcal{L}$ and every $c \in \mathcal{C}$, it holds that*

$$\Big|\mathbf{E}\left[\ell(\mathbf{y}, c(\mathbf{x}))\right] - \mathbf{E}\left[\ell(\mathbf{y}_p, c(\mathbf{x}))\right]\Big| \le \gamma. \tag{9}$$

They show a decomposition lemma which shows that decision OI and hypothesis OI together imply loss OI:

Lemma 4.4 (Lemma 4.4 in [3])**.** *If a predictor satisfies $(\mathcal{L}, \beta)$-decision OI and $(\mathcal{L}, \mathcal{C}, \gamma)$-hypothesis OI, then it satisfies $(\mathcal{L}, \mathcal{C}, \beta + \gamma)$-loss OI.*

In the same paper, they show that decision OI is a form of weight-restricted calibration (as defined in [23]), and hypothesis OI is multiaccuracy for a class of hypotheses derived from $\mathcal{L}$ and $\mathcal{C}$. To state these prior results, let $\partial\ell(t) = \ell(1, t) - \ell(0, t)$ denote the discrete partial derivative of ℓ with respect to the label.

Lemma 4.5 (Characterizing decision and hypothesis OI [3])**.**

- *A predictor p satisfies (ℓ, β)-decision OI if and only if p satisfies weight-restricted calibration with error at most β with respect to the weight function $\partial\ell \circ k_\ell$.*
- *A predictor p satisfies $(\ell, \mathcal{C}, \gamma)$-hypothesis OI if and only if p satisfies multiaccuracy with error at most γ with respect to the function class $\partial\ell \circ \mathcal{C}$.*

The main result of this section is essentially a converse to Lemma 4.4. We show that under a very mild condition on the loss function ℓ and class $\mathcal{C}$, any predictor satisfying $(\ell, \mathcal{C}, \alpha)$-loss OI must *also* satisfy both $(\ell, O(\alpha))$-decision OI and $(\ell, \mathcal{C}, O(\alpha))$-hypothesis OI. Combined with Lemma 4.5, which says that decision OI is a form of calibration and hypothesis OI is a form of multiaccuracy, this establishes the desired equivalence between loss OI and calibrated multiaccuracy. We prove our equivalence for loss functions ℓ that have a *neutral* action:

Definition 4.6 (Neutrality)**.** *We say that an action $t \in \mathcal{A}$ is* neutral *for a loss function $\ell : \{0, 1\} \times \mathcal{A} \to \mathbb{R}$ if $\partial\ell(t) = 0$.*

Many natural loss functions have neutral actions. For example, $t = 1/2$ is a neutral action for the squared loss and cross-entropy loss, and $t = 0$ is neutral for the logistic loss. Although there do exist loss functions without neutral actions, such losses can always be easily tweaked to correct this anomaly. For example, the loss function $\ell(y, p) = y$ is proper, but the ranges of $\ell(1, p)$ and $\ell(0, p)$ are disjoint, so there cannot exist a neutral action. In this example, the additive dependence on y was the cause of the problem. The following lemma shows that the presence of such a term is the only possible obstruction to having a neutral action, and that this obstruction can be bypassed by subtracting away the term:

Lemma 4.7. *Fix a loss function $\ell_0 : \{0, 1\} \times \mathcal{A} \to \mathbb{R}$ and an action $t_0 \in \mathcal{A}$. Then t_0 is neutral for the loss function $\ell(y, t) = \ell_0(y, t) - \partial\ell_0(t_0)y$. Moreover, if $\mathcal{A} = [0, 1]$ and ℓ_0 is proper, then so is ℓ.*

The last mild assumption that we need concerns the class $\mathcal{C}$, not the loss ℓ. We require $\mathcal{C}$ to contain all the *constant* action functions $c(x) = t$ for each fixed action $t \in \mathcal{A}$. We denote this by $\mathcal{A} \subseteq \mathcal{C}$. With this notation in hand, we are ready to state the main result of this section:

Theorem 4.8. *Suppose the loss function $\ell : \{0, 1\} \times \mathcal{A} \to \mathbb{R}$ has a neutral action and $\mathcal{A} \subseteq \mathcal{C}$. Then, a predictor p satisfies $(\ell, \mathcal{C}, O(\alpha))$-loss OI if and only if it satisfies both $(\partial\ell \circ k_\ell, O(\alpha))$-weight restricted calibration and $(\partial\ell \circ \mathcal{C}, O(\alpha))$-multiaccuracy.*

Once again, we emphasize that the neutral action assumption is a mild one, and is essentially only necessary to rule out the trivial loss function $\ell(y, p) = y$ that does not depend on the prediction p, as well as rescalings and small perturbations thereof (i.e. functions where $\partial\ell$ is close to a constant function). We could, for example, replace the neutral action assumption with the similarly mild requirements that $|\mathrm{range}(\partial\ell)| \geq \Omega(1)$ and $\mathrm{range}(\ell) \subseteq [-1, 1]$, which also rules out these bad cases.

Theorem 4.8 extends the previously known Lemma 4.4, which was a one directional implication. Bearing in mind the connections between decision OI and calibration and between hypothesis OI and multiaccuracy (Lemma 4.5), we see that the new direction of Theorem 4.8 follows immediately from the following lemma regarding losses with neutral actions.

Lemma 4.9. *Assume that the loss function $\ell : \{0, 1\} \times \mathcal{A} \to \mathbb{R}$ has a neutral action and that $\mathcal{A} \subseteq \mathcal{C}$. If a predictor p satisfies $(\ell, \mathcal{C}, \alpha)$-loss OI, then it also satisfies both (ℓ, α)-decision OI and $(\ell, \mathcal{C}, 2\alpha)$-hypothesis OI.*

References

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

[2] 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)*, pages 1939–1948. PMLR, 2018.

[3] Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. In *Innovations in Theoretical Computer Science (ITCS)*, 2023.

[4] Princewill Okoroafor, Robert Kleinberg, and Michael P. Kim. Near-optimal algorithms for omniprediction. In *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1595–1609. IEEE, 2025.

[5] Parikshit Gopalan, Michael Kim, and Omer Reingold. Swap agnostic learning, or characterizing omniprediction via multicalibration. *Advances in Neural Information Processing Systems*, 36, 2023.

[6] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science conference (ITCS)*, pages 214–226. ACM, 2012.

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

[8] Luca Trevisan, Madhur Tulsiani, and Salil P. Vadhan. Regularity, boosting, and efficiently simulating every high-entropy distribution. In *Proceedings of the 24th Annual IEEE Conference on Computational Complexity (CCC)*, pages 126–136. IEEE Computer Society, 2009.

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

[10] Cynthia Dwork, Daniel Lee, Huijia Lin, and Pranay Tankala. From pseudorandomness to multi-group fairness and back. In *The Thirty-Sixth Annual Conference on Learning Theory (COLT)*, volume 195 of *Proceedings of Machine Learning Research*, pages 3566–3614. PMLR, 2023.

[11] Sílvia Casacuberta, Cynthia Dwork, and Salil Vadhan. Complexity-theoretic implications of multicalibration. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, pages 1071–1082, 2024.

[12] Sílvia Casacuberta, Parikshit Gopalan, Varun Kanade, and Omer Reingold. How global calibration strengthens multiaccuracy. In *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1198–1227. IEEE, 2025.

[13] Parikshit Gopalan, Princewill Okoroafor, Prasad Raghavendra, Abhishek Shetty, and Mihir Singhal. Omnipredictors for regression and the approximate rank of convex functions. In *The Thirty Seventh Annual Conference on Learning Theory (COLT)*, pages 2027–2070. PMLR, 2024.

[14] Sumegha Garg, Christopher Jung, Omer Reingold, and Aaron Roth. Oracle efficient online multicalibration and omniprediction. In David P. Woodruff, editor, *Proceedings of the 2024 ACM-SIAM Symposium on Discrete Algorithms, SODA 2024, Alexandria, VA, USA, January 7-10, 2024*, pages 2725–2792. SIAM, 2024.

[15] Lunjia Hu, Kevin Tian, and Chutong Yang. Omnipredicting single-index models with multi-index models. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing (STOC)*, pages 1762–1773. ACM, 2025.

[16] Cynthia Dwork, Chris Hays, Nicole Immorlica, Juan C. Perdomo, and Pranay Tankala. From fairness to infinity: Outcome-indistinguishable (omni)prediction in evolving graphs. In *The Thirty Eighth Annual Conference on Learning Theory (COLT)*, volume 291 of *Proceedings of Machine Learning Research*, pages 1564–1637. PMLR, 2025.

[17] Aravind Gollakota, Parikshit Gopalan, Adam Klivans, and Konstantinos Stavropoulos. Agnostically learning single-index models using omnipredictors. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:14685–14704, 2023.

[18] Isaac Gibbs and Ryan J. Tibshirani. Sample-efficient omniprediction for proper losses. In Steve Hanneke and Tor Lattimore, editors, *The Thirty Ninth Annual Conference on Learning Theory (COLT)*, volume 336 of *Proceedings of Machine Learning Research*, pages 2679–2719. PMLR, 2026. URL `https://proceedings.mlr.press/v336/gibbs26a.html`.

[19] Sivaraman Balakrishnan, Nika Haghtalab, Daniel Hsu, Brian W Lee, and Eric Zhao. Panprediction: Optimal predictions for any downstream task and loss. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

[20] Lunjia Hu and Salil P. Vadhan. Generalized and unified equivalences between hardness and pseudoentropy. In *Theory of Cryptography - 23rd International Conference, TCC 2025*, volume 16271 of *Lecture Notes in Computer Science*, pages 258–288. Springer, 2025.

[21] Cynthia Dwork and Pranay Tankala. Efficient testing implies structured symmetry. In *2026 IEEE 67th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2026.

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

[23] 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.

[24] Bobby Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory (COLT)*, volume 195, pages 5143–5145. PMLR, 12–15 Jul 2023.

[25] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[26] Frank Nielsen. Legendre transformation and information geometry. Technical Report CIG-MEMO2, September 2010.

[27] Noam Nisan and Avi Wigderson. Hardness vs randomness. *Journal of Computer and System Sciences*, 49(2):149–167, 1994.

[28] Russell Impagliazzo. A personal view of average-case complexity. In *Proceedings of the Tenth Annual Structure in Complexity Theory Conference, Minneapolis, Minnesota, USA, June 19-22, 1995*, pages 134–147. IEEE Computer Society, 1995.

A Proofs for Section 3

Theorem 3.2. *For any classes $\mathcal{L}$ and $\mathcal{C}$, if $q \in \mathrm{OP}_{\mathcal{L},\mathcal{C},\varepsilon}$, then $\mathrm{Btwn}(p^*, q) \subseteq \mathrm{OP}_{\mathcal{L},\mathcal{C},\varepsilon}$.*

Proof of Theorem 3.2. Let $\ell \in \mathcal{L}$ be a loss function, and let $q \in \mathrm{OP}_{\mathcal{L},\mathcal{C},\varepsilon}$, so that Equation (4) holds. Let $r \in \mathrm{Btwn}(p^*, q)$. Our goal is to show that

$$\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\ell(\mathbf{y}, k_\ell(r(\mathbf{x})))] \leq \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\ell(\mathbf{y}, k_\ell(q(\mathbf{x})))]. \tag{10}$$

Since $q \in \mathrm{OP}_{\mathcal{L},\mathcal{C},\varepsilon}$, this will imply that $r \in \mathrm{OP}_{\mathcal{L},\mathcal{C},\varepsilon}$, which proves the result.

By Lemma 2.3, there exists a proper loss ℓ^{pr} so that Equation (10) is equivalent to

$$\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\ell^{\mathrm{pr}}(\mathbf{y}, r(\mathbf{x}))] \leq \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\ell^{\mathrm{pr}}(\mathbf{y}, q(\mathbf{x}))]. \tag{11}$$

We will in fact show that Equation (11) holds even conditioned on $\mathbf{x}$:

$$\mathop{\mathbf{E}}_{\mathbf{y}|\mathbf{x}}[\ell^{\mathrm{pr}}(\mathbf{y}, r(\mathbf{x}))] = \ell^{\mathrm{pr}}(p^*(\mathbf{x}), r(\mathbf{x})) \leq \ell^{\mathrm{pr}}(p^*(\mathbf{x}), q(\mathbf{x})) = \mathop{\mathbf{E}}_{\mathbf{y}|\mathbf{x}}[\ell^{\mathrm{pr}}(\mathbf{y}, q(\mathbf{x}))]. \tag{12}$$

Consider $\mathbf{x}$ such that $q(\mathbf{x}) \leq p^*(\mathbf{x})$. The case $q(\mathbf{x}) \geq p^*(\mathbf{x})$ is handled similarly. Since the inequality holds pointwise for every x, we will drop the x and show that for $q \leq r \leq p^*$ all values in $[0, 1]$,

$$\ell^{\mathrm{pr}}(p^*, q) \geq \ell^{\mathrm{pr}}(p^*, r). \tag{13}$$

By Theorem 2.5, there exists a concave function h (depending on ℓ^{pr}) so that Equation (13) is equivalent to

$$h(q) + (p^* - q)h'(q) \geq h(r) + (p^* - r)h'(r). \tag{14}$$

Geometrically, this says that the supergradient approximation to $h(p^*)$ decreases (and hence improves) as we move from q towards p^*. To prove Equation 14, we note that

$$\begin{aligned} h(q) + (p^* - q)h'(q) &= h(q) + (r - q)h'(q) + (p^* - r)h'(q) \\ &\geq h(r) + (p^* - r)h'(r), \end{aligned}$$

where we use $h(q) + (r - q)h'(q) \geq h(r)$ by concavity, and $h'(q) \geq h'(r)$ since $q \leq r$ and by concavity, the derivative h' is decreasing. □

Lemma 3.3. *Assume there exists $q \in \mathrm{OP}_{\mathcal{L},\mathcal{C}}$ such that $\mathbf{E}[|p^*(\mathbf{x}) - q(\mathbf{x})|] \geq \alpha$. Then there exists $r \in \mathrm{OP}_{\mathcal{L},\mathcal{C}}$ such that $\mathsf{EAE}(r) \geq \alpha/2$.*

Proof of Lemma 3.3. Note that

$$\mathbf{E}[|p^*(\mathbf{x}) - q(\mathbf{x})|] = \mathbf{E}[\mathbf{1}[p^*(\mathbf{x}) \geq q(\mathbf{x})](p^*(\mathbf{x}) - q(\mathbf{x}))] + \mathbf{E}[\mathbf{1}[q(\mathbf{x}) > p^*(\mathbf{x})](q(\mathbf{x}) - p^*(\mathbf{x}))]. \tag{15}$$

Thus, at least one of the two terms on the RHS of Eq. (15) is at least $\alpha/2$; without loss of generality, assume that $\mathbf{E}[\mathbf{1}[q(\mathbf{x}) > p^*(\mathbf{x})](q(\mathbf{x}) - p^*(\mathbf{x}))] \geq \alpha/2$.

Let us define the predictor r as

$$r(x) = \begin{cases} p^*(x) \text{ if } q(x) \leq p^*(x) \\ q(x) \text{ if } q(x) > p^*(x). \end{cases} \tag{16}$$

It follows that $r \in \mathrm{Btwn}(p^*, q) \subseteq \mathrm{OP}_{\mathcal{L},\mathcal{C}}$. But we have

$$\mathsf{EAE}(r) = \Big|\mathbf{E}[r(\mathbf{x}) - \mathbf{y}]\Big| = \Big|\mathbf{E}[r(\mathbf{x}) - p^*(\mathbf{x})]\Big| = \Big|\mathbf{E}[\mathbf{1}[q(\mathbf{x}) > p^*(\mathbf{x})](q(\mathbf{x}) - p^*(\mathbf{x}))]\Big| \geq \alpha/2.$$

□

Theorem 3.5. *Let $\mathcal{C}$ be a hypothesis class of functions with range $[-1,1]$, $\mathcal{D}_{\mathcal{X}}$ be a distribution over $\mathcal{X}$, and let $\mathcal{L}$ be the class of all convex, Lipschitz losses.*[11] *Suppose that there exists a Boolean function $p^* : \mathcal{X} \to \{0,1\}$ and $\varepsilon > 0$ such that p^* is $(1/2-\varepsilon)$-hard for $\mathcal{C}$ under $\mathcal{D}_{\mathcal{X}}$. Then, for the distribution $\mathcal{D} = \mathcal{D}(p^*)$, there exists a predictor $r : \mathcal{X} \to [0,1]$ such that r is an $(\mathcal{L}, \mathcal{C}, O(\varepsilon))$-omnipredictor and $\mathsf{EAE}(r) \geq \frac{1}{4}$.*

Proof of Theorem 3.5. Consider the function $q(x) = 1/2$. For $\mathcal{D} = \mathcal{D}(p^*)$, it is straightforward to see that q is $(\mathcal{C}, \varepsilon)$-multicalibrated. Since p^* is $(1/2-\varepsilon)$-hard for $\mathcal{C}$ under $\mathcal{D}_{\mathcal{X}}$, for each $c \in \mathcal{C}$,

$$|\,\mathbf{E}[c(\mathbf{x})(\mathbf{y} - q(\mathbf{x}))]| = \left|\mathbf{E}\left[c(\mathbf{x})\Big(p^*(\mathbf{x}) - \frac{1}{2}\Big)\right]\right| \leq \varepsilon.$$

Then the main result of [1] implies that q is an $(\mathcal{L}, \mathcal{C}, O(\varepsilon))$-omnipredictor where $\mathcal{L}$ is the class of convex, Lipschitz losses; here the hidden constant in the O-notation depends only on the Lipschitz constant of the class of loss functions.

Since the range of p^* is $\{0,1\}$,

$$\mathbf{E}\left[|q(\mathbf{x}) - p^*(\mathbf{x})|\right] = \frac{1}{2},$$

and the conclusion follows by applying Lemma 3.3. □

B Proofs for Section 4

Lemma 4.7. *Fix a loss function $\ell_0 : \{0,1\} \times \mathcal{A} \to \mathbb{R}$ and an action $t_0 \in \mathcal{A}$. Then t_0 is neutral for the loss function $\ell(y,t) = \ell_0(y,t) - \partial\ell_0(t_0)y$. Moreover, if $\mathcal{A} = [0,1]$ and ℓ_0 is proper, then so is ℓ.*

Proof of Lemma 4.7. The fact that t_0 is neutral for ℓ follows immediately from the definitions of neutrality and ℓ. Next, recall that ℓ_0 is proper if and only if there exists a concave function h and a superderivative h' of h such that $\ell_0(y,p) = h(p) + h'(p)(y-p)$. If $h(p)$ is concave, then so is the function $h(p) + c \cdot p$ for any constant c, of which $h'(p) + c$ is a valid superderivative. Thus, the following loss function is also proper:

$$\big(h(p) + c \cdot p\big) + \big(h'(p) + c\big)(y - p).$$

When $c = -\partial\ell_0(t_0)$, the expression above exactly coincides with the function ℓ of interest. □

Lemma 4.9. *Assume that the loss function $\ell : \{0,1\} \times \mathcal{A} \to \mathbb{R}$ has a neutral action and that $\mathcal{A} \subseteq \mathcal{C}$. If a predictor p satisfies $(\ell, \mathcal{C}, \alpha)$-loss OI, then it also satisfies both (ℓ, α)-decision OI and $(\ell, \mathcal{C}, 2\alpha)$-hypothesis OI.*

Proof of Lemma 4.9. Let us consider a neutral action t for ℓ. Since $t \in \mathcal{C}$, the loss OI condition on p implies

$$\Big|\mathbf{E}[u_{\ell,t}(\mathbf{x}, \mathbf{y}, p(\mathbf{x}))] - \mathbf{E}[u_{\ell,t}(\mathbf{x}, \mathbf{y}_p, p(\mathbf{x}))]\Big| \leq \alpha. \tag{17}$$

But since t is neutral, the distribution on labels is irrelevant. Hence we have

$$\mathbf{E}[\ell(\mathbf{y}, t)] = \mathbf{E}[\ell(\mathbf{y}_p, t)] = \ell(0,t) = \ell(1,t).$$

Plugging in the definition of $u_{\ell,t}$, namely $\ell(y,t) - \ell(y, k_\ell(v))$ and using the fact that $\mathbf{E}[\ell(\mathbf{y},t)] = \mathbf{E}[\ell(\mathbf{y}_p,t)]$, the left side of Equation (17) equals

$$\begin{aligned}&\Big|\mathbf{E}[\ell(\mathbf{y}, k_\ell(p(\mathbf{x}))) - \ell(\mathbf{y},t)] - \mathbf{E}[(\ell(\mathbf{y}_p, k_\ell(p(\mathbf{x})))) - \ell(\mathbf{y}_p,t)]\Big| \\ &= \Big|\mathbf{E}[\ell(\mathbf{y}, k_\ell(p(\mathbf{x})))] - \mathbf{E}[\ell(\mathbf{y}_p, k_\ell(p(\mathbf{x})))]\Big|, \end{aligned} \tag{18}$$

which implies the claim about decision OI. For the claim about hypothesis OI, observe that

$$\begin{aligned}\mathbf{E}[\ell(\mathbf{y}, c(\mathbf{x}))] &= \mathbf{E}[u_{\ell,c}(\mathbf{x}, \mathbf{y}, p(\mathbf{x}))] + \mathbf{E}[\ell(\mathbf{y}, k_\ell(p(\mathbf{x})))] \\ \mathbf{E}[\ell(\mathbf{y}_p, c(\mathbf{x}))] &= \mathbf{E}[u_{\ell,c}(\mathbf{x}, \mathbf{y}_p, p(\mathbf{x}))] + \mathbf{E}[\ell(\mathbf{y}_p, k_\ell(p(\mathbf{x})))].\end{aligned}$$

[11] By *convex*, we mean that both functions $\ell(0,\cdot)$ and $\ell(1,\cdot)$ are convex, for each $\ell \in \mathcal{L}$. As in [1] we can relax this to include all (B,ε)-nice functions and drop the global Lipschitz requirement.

For every $c \in \mathcal{C}$, each expectation on the right side changes only by α in going from $\mathbf{y}$ to $\mathbf{y}_p$, by loss OI applied to $u_{\ell,c}$ (for the term $\mathbf{E}[u_{\ell,c}(\mathbf{x}, \mathbf{y}, p(\mathbf{x}))]$) and by decision OI applied to ℓ (for the term $\mathbf{E}[\ell(\mathbf{y}, k_\ell(p(\mathbf{x})))]$), respectively. Therefore, by the triangle inequality, the expectations on the right side can change by at most 2α, and so $\left|\mathbf{E}[\ell(\mathbf{y}, c(\mathbf{x}))] - \mathbf{E}[\ell(\mathbf{y}_p, c(\mathbf{x}))]\right| \leq 2\alpha$ for every $c \in \mathcal{C}$. This proves the claim about hypothesis OI. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We rigorously prove all of the mathematical results that we claim.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We spell out all of the assumptions that we make in our theorem statements.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide rigorous proofs for all of our theorems and mathematical claims.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: All of our results are theoretical and we do not present empirical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: All of our results are theoretical and we do not present empirical results (nor code or data).

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: All of our results are theoretical and we do not present any experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: All of our results are theoretical and we do not present any experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: All of our results are theoretical and we do not present any experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read the Code of ethics and confirm that our research conducted for this paper conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work continues the established line of research on multigroup fairness, where the primary goal is to ensure that algorithmic predictions are unbiased/calibrated among many subgroups of the domain. Hence our theoretical results on this framework contribute to more positive societal impacts of computation.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: All of our results are theoretical and we do not release any data or models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: All of our results are theoretical and we do not present or use any assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: All of our results are theoretical and we do not present any experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: All of our results are theoretical and we do not conduct research with crowdsourcing experiments or with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: All of our results are theoretical and we do not conduct research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: We have not used LLMs in this research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.