Sample-Efficient Omniprediction for Proper Losses

Isaac Gibbs*† Ryan J. Tibshirani*

Abstract

We study the problem of constructing probabilistic predictions that lead to accurate decisions when employed by downstream users to inform actions. Given a single decision maker, developing an optimal predictor is equivalent to minimizing a proper loss function corresponding to the negative utility of that individual. For multiple decision makers, our problem can be viewed as a variant of omniprediction in which the goal is to develop a single predictor which simultaneously minimizes multiple losses. Existing algorithms for achieving omniprediction broadly fall into two categories: first, boosting methods, which optimize auxiliary targets such as multicalibration and obtain omniprediction as a corollary, and second, adversarial two-player game based approaches, which estimate and respond to the worst-case loss in an online fashion. We give lower bounds which demonstrate that multicalibration is a strictly more difficult problem than omniprediction and hence the first approach must incur suboptimal sample complexity. For the latter approach, we discuss how these ideas can be used to obtain a sample-efficient algorithm for our problem through an online-to-batch conversion. This conversion has the downside of returning a complex, randomized predictor. We therefore improve on this method by designing a more direct nonrandomized algorithm that exploits structural elements of the set of proper losses.

1 Introduction

A standard method for fitting a predictive model is to minimize a single loss function measuring its accuracy. Commonly, this framework is employed under the implicit assumption that accurate predictions are sufficient to guide the decisions of downstream users. While this may hold true in some examples, in general, predictive accuracy does not preclude the possibility that the trained model fails on the most decision-critical examples. For example, neural network classifiers trained using empirical risk minimization have been observed to be miscalibrated and thus cannot be relied upon to accurately measure outcome uncertainty [Guo et al., 2017].

In response to this, a growing body of literature has focused on designing predictors that simultaneously satisfy multiple performance criteria. Rather than solely targeting low empirical loss, multiaccuracy instead requires the predictor to be unbiased over a collection of reweightings of the feature space [Hébert-Johnson et al., 2018, Kim et al., 2019]. In applications, these reweightings typically include subgroup indicators and thus multiaccuracy can ensure that the predictor remains unbiased across sensitive subpopulations. This is strengthened by multicalibration, which further requires the same unbiasedness criteria to hold conditional on the specific prediction that was issued [Hébert-Johnson et al., 2018]. Another line of work on distributional robustness looks to construct predictors that are simultaneously accurate across a variety of covariate shifts or subpopulations [Mansour et al., 2008, Blum et al., 2017, Mohri et al., 2019, Rothblum and Yona, 2021, Duchi et al., 2023].

In this paper, we will focus on constructing predictors that provide simultaneously optimal performance when applied by multiple downstream users to inform decisions. More formally, consider a decision-making task with covariates X and a binary outcome $Y \in \{0, 1\}$. Let $\hat{p}(X)$ denote an estimate of $\mathbb{P}(Y = 1 \mid X)$, the conditional probability that Y is equal to one given X, and consider a setting in which a downstream user will use $\hat{p}(X)$ to choose an action $a \in \mathcal{A}$. Let $u(a, y)$ be a utility function that characterizes the user's benefit

*Department of Statistics, University of California, Berkeley.
†Email: igibbs@berkeley.edu.

from the action a under true outcome y. A natural decision-making procedure is to treat the prediction as though it were perfectly accurate and select an action

$$a(\hat{p}(X); u) \in \operatorname*{argmax}_{a \in \mathcal{A}} \mathbb{E}_{Y' \sim \mathrm{Ber}(\hat{p}(X))}[u(a, Y')], \tag{1.1}$$

that maximizes the expected utility under $Y' \sim \mathrm{Ber}(\hat{p}(X))$. Our goal is to construct predictors that lead to good decisions when applied in this manner by *any* downstream user, i.e., to construct predictors that lead to good performance in (1.1) when applied to arbitrary utility functions.

A motivation for this framework comes from settings in which a single centralized entity with access to data and statistical expertise must issue predictions that are useful to a diverse array of end users. This type of interaction is common in domains such as weather and epidemiological forecasting, where government organizations regularly broadcast predictions that are utilized by the general public. Alternatively, one may consider technologies such as language or vision models which are frequently treated as black-boxes by their users. In these settings, the estimated probability $\hat{p}(X)$ could indicate the likelihood that the text or image output by the model contains an error and the user may use this information to decide whether to trust the model or seek out additional assistance.

Without any further restrictions, obtaining optimal decisions in (1.1) is as difficult as learning the true conditional probability function, $p^*(X) = \mathbb{P}(Y = 1 \mid X)$. Indeed, as we will show in Section 2, the reduction in expected utility that is suffered by taking action $a(\hat{p}(X); u)$ instead of the optimal action $a(p^*(X); u)$ is comparable, in the worst-case (over all utility functions), to the L_1 distance between $\hat{p}(X)$ and $p^*(X)$. By standard results in nonparametric estimation, this problem quickly becomes intractable when X is of even moderate dimension (e.g., see Stone [1982], Devroye et al. [1996], Györfi et al. [2002]). As a result, instead of asking for optimal decisions overall, we will judge $\hat{p}(X)$ by comparing its performance against the best in a restricted class of predictors $\mathcal{F}$. More formally, we aim to minimize

$$\sup_{u} \sup_{f \in \mathcal{F}} \mathbb{E}[u(a(f(X); u), Y)] - \mathbb{E}[u(a(\hat{p}(X); u), Y)], \tag{1.2}$$

where the first supremum is over all bounded utility functions* and the expectations are taken over the test point (X, Y). Importantly, unlike the standard setup in nonparametric estimation, we place no smoothness assumptions or other restrictions on the distribution of the data. Additionally, note that in this objective the comparator in $\mathcal{F}$ is allowed to depend on the utility function. On the other hand, the prediction $\hat{p}(X)$ that we construct must be universal to all decision-making problems.

By reformulating (1.2) slightly, our problem can be seen as a special case of a more general framework known as omniprediction [Gopalan et al., 2022]. Stated simply, omniprediction is the task of constructing predictors which minimize multiple loss functions simultaneously. Given a class of losses $\mathcal{L}$ and competitor functions $\mathcal{F}$, we aim to produce $\hat{p}(X)$ to minimize

$$\sup_{\ell \in \mathcal{L}} \sup_{f \in \mathcal{F}} \mathbb{E}[\ell(\hat{p}(X), Y)] - \mathbb{E}[\ell(f(X), Y)]. \tag{1.3}$$

To connect this to our current setting, let $\ell^u(\hat{p}(X), Y) = -u(a(\hat{p}(X); u), Y)$ denote the loss induced by utility function u. We thus obtain the equivalence

$$\mathbb{E}[u(a(f(X); u), Y)] - \mathbb{E}[u(a(\hat{p}(X); u), Y)] = \mathbb{E}[\ell^u(\hat{p}(X), Y)] - \mathbb{E}[\ell^u(f(X), Y)].$$

Further, it is easy to check that $p \in \operatorname{argmin}_{q \in [0,1]} \mathbb{E}_{Y \sim \mathrm{Ber}(p)}[\ell^u(q, Y)]$. Our problem can then be equivalently formulated as bounding the omniprediction error (1.3) with $\mathcal{L}$ taken to be the set of bounded loss functions that are minimized by predicting the true probabilities. In the probabilistic forecasting literature, losses with this last property are referred to as *proper* [Gneiting and Raftery, 2007].

Following the initial proposal by Gopalan et al. [2022], numerous authors have examined algorithms for achieving omniprediction. These methods can be broadly categorized into two groups. The first are boosting

*And, by extension, all possible action spaces.

algorithms [Gopalan et al., 2022, 2023b,a, Globus-Harris et al., 2023, Kim and Perdomo, 2023, Gopalan et al., 2024]. These contributions begin by noting that in order to have low omniprediction error it is sufficient for $\hat{p}(X)$ to satisfy a corresponding set of multiaccuracy, calibration, and/or multicalibration criteria. Then, a predictor that satisfies these criteria is constructed in an iterative fashion by identifying and correcting any criteria which are not currently met. The second class of methods is based on algorithms for what are known as two-player games [Garg et al., 2024, Noarov et al., 2025, Okoroafor et al., 2025, Lu et al., 2025]. Here, the omniprediction problem is framed as a game in which one player constructs a mixture loss that serves as a proxy for the supremum in (1.3) and the second player constructs the predictor as a best response to this loss. By drawing on tools from the online learning literature, these two players can be designed to guarantee that the predictors returned by the second player satisfy an online form of omniprediction. As shown in Okoroafor et al. [2025] and Lu et al. [2025], standard online-to-batch conversion methods can then be used to obtain a predictor with low error on i.i.d. data.

As an aside, we note that a third approach for omniprediction that does not directly use the two-player game set-up, but draws on closely related tools from the online learning literature, is developed in Dwork et al. [2024]. Their method is designed specifically for cases in which compositions of the loss and comparator functions can be efficiently embedded in a kernel class. In general, this embedding leads to suboptimal rates for the problems we are interested in and thus we will not focus on this method in detail.

The remainder of this paper is devoted to studying the sample efficiency of algorithms for omniprediction with respect to the class of proper loss functions. We begin in Section 2 by giving a precise characterization of the omniprediction error when no restrictions are placed on the comparator class. We show that in this case omniprediction is equivalent to L_1 estimation of $p^*(X)$, and thus suffers from poor (nonparametric) learning rates. Section 3 considers the performance of boosting methods under the setting in which $\mathcal{F}$ has finite VC dimension $\mathrm{VC}(\mathcal{F}) < \infty$. We establish that for a sample of size n the sufficient conditions of multicalibration and calibrated multiaccuracy can be achieved together at a rate no better than $\sqrt{\mathrm{VC}(\mathcal{F})/n} + n^{-2/5}$. This is strictly worse than the error bound of $\tilde{O}(\sqrt{\mathrm{VC}(\mathcal{F})/n})$ obtained by two-player game based methods [Okoroafor et al., 2025], where recall the notation $\tilde{O}(\cdot)$ hides polylogarithmic factors in $\mathrm{VC}(\mathcal{F})$ and n. Therefore, existing boosting methods that target these criteria must be suboptimal.

It is interesting to note that the error rate obtained by two-player game based methods for omniprediction is (up to polylog terms) identical to the optimal learning rate for standard risk minimization of a single loss function. A classical result in learning theory shows that the best possible error rate for binary classification with respect to the 0-1 loss is $\sqrt{\mathrm{VC}(\mathcal{F})/n}$ (e.g., see Theorem 14.5 in Devroye et al. [1996]). Since the 0-1 loss is proper, this lower bound also applies to our present omniprediction problem. In what follows, we refer to $\sqrt{\mathrm{VC}(\mathcal{F})/n}$ as the optimal rate for omniprediction and we say that any method that achieves this rate up to polylogarithmic factors is sample-efficient.

Sections 4 and 5 give our presentation of such sample-efficient algorithms for omniprediction. Section 4 presents a general reduction of the omniprediction problem into the comparatively simpler task of ensembling a finite set of predictors over a small collection of loss functions. Here, we draw heavily on the work of Savage [1971] and Ehm et al. [2016] which demonstrates that all proper losses can be decomposed as mixtures over a class of weighted 0-1 losses. Section 5 then presents two methods. In Section 5.1, we discuss two-player game based algorithms and derive a new variant of such methods which is simpler to compute. Like all two-player game based methods, this procedure achieves sample efficiency, but does so at the cost of producing a complex randomized predictor. To overcome this shortcoming, in Section 5.2 we present a new algorithm that more directly exploits structural properties of the set of proper loss functions to obtain a nonrandomized predictor that yields the same optimal error rate. This partially answers an open question of Okoroafor et al. [2025] who raised the problem of constructing nonrandomized omnipredictors which achieve sample efficiency.

Empirical evaluation of the aforementioned methods on both simulated examples and a sales forecasting dataset is given in Section 6. As expected, we find that boosting methods give suboptimal performance when compared to the other approaches. On the other hand, methods based on two-player games and our direct ensembling approach realize similar error rates in practice.

While our methods are designed for the binary prediction problem, they can be readily extended to other targets. In Section 7, we describe a result of Steinwart et al. [2014] that provides a general characterization

of proper losses for other point prediction targets for continuous outcomes such as (conditional) means or quantiles. By comparing this result to the binary case, we find that our methods can be applied to construct point predictors that are simultaneously accurate over all proper losses for a given one-dimensional target (a single mean or quantile). Estimation of multivariate targets is considerably more challenging and provides an interesting open direction for future work.

Notation. In what follows, we let $\{(X_i, Y_i)\}_{i=1}^n \subseteq \mathcal{X} \times \{0,1\}$ denote an i.i.d. training sample. We use (X, Y) to denote a test sample taken independently from the same distribution and $p^*(X) = \mathbb{P}(Y = 1 \mid X)$ to denote the true conditional probability function. Throughout, we will work with the class

$$\mathcal{L}_0 = \left\{ \ell : [0,1] \times \{0,1\} \to [0,1] : p \in \operatorname*{argmin}_{q \in [0,1]} \mathbb{E}_{Y' \sim \mathrm{Ber}(p)}[\ell(q, Y')], \text{ for all } p \in [0,1] \right\},$$

of bounded, proper loss functions. Our goal is to use $\{(X_i, Y_i)\}_{i=1}^n$ to construct a predictor $\hat{p}(X)$ with low omniprediction error (1.3) specialized to $\mathcal{L}_0$, i.e., a low value of

$$\mathrm{OP}(\hat{p}; \mathcal{L}_0, \mathcal{F}) = \sup_{\ell \in \mathcal{L}_0} \sup_{f \in \mathcal{F}} \mathbb{E}[\ell(\hat{p}(X), Y)] - \mathbb{E}[\ell(f(X), Y)], \tag{1.4}$$

where as usual, the expectations are taken over the test point (X, Y).

2 Comparison to nonparametric estimation

To begin understanding the omniprediction problem, it is useful to first consider how (1.4) behaves when $\mathcal{F}$ is allowed to include all possible competitor functions. First, as a sanity check, let us verify that $p^*(X)$ does indeed achieve the minimum possible omniprediction error in this case. Indeed, for any proper loss ℓ and for any predictor $p(X)$,

$$\mathbb{E}[\ell(p^*(X), Y)] = \mathbb{E}[\mathbb{E}[\ell(p^*(X), Y) \mid X]] \leq \mathbb{E}[\mathbb{E}[\ell(p(X), Y) \mid X]] = \mathbb{E}[\ell(p(X), Y)],$$

where the inequality follows from the definition of propriety. Furthermore, since the loss ℓ^u induced by an arbitrary utility function u is proper, the true conditional probability $p^*(X)$ is the optimal predictor for any decision-making problem,

$$\mathbb{E}[u(a(p^*(X); u), Y)] \geq \mathbb{E}[u(a(p(X); u), Y)],$$

where the inequality again follows by conditioning on X and applying propriety.

As $\hat{p}(X)$ moves away from $p^*(X)$ it will no longer give optimal performance over all proper losses. This is quantified in the following proposition which shows that for a general predictor, the maximum performance gap relative to $p^*(X)$ scales with the L_1 distance. As $p^*(X)$ is always the optimal predictor, this proposition can be interpreted as giving bounds on the omniprediction error in the case where no restrictions are placed on $\mathcal{F}$. The proof of this result is deferred to Appendix A, as will be the default for proofs in this paper.

Proposition 1. *For any predictor $p : \mathcal{X} \to [0,1]$,*

$$\frac{1}{210}\mathbb{E}[|p(X) - p^*(X)|]^2 \leq \sup_{\ell \in \mathcal{L}_0} \mathbb{E}[\ell(p(X), Y)] - \mathbb{E}[\ell(p^*(X), Y)] \leq 2\mathbb{E}[|p(X) - p^*(X)|].$$

It is well-known that without parametric assumptions, L_1 estimation of $p^*(X)$ suffers from a strong curse of dimensionality. For instance, if X is uniformly distributed on $[-1,1]^d$, and $p^*(X)$ can be any Lipschitz continuous function (with say, a Lipschitz constant of at most 1) then we have $\mathbb{E}[|\hat{p}(X) - p^*(X)|] \geq \Omega(n^{-1/(d+2)})$, where the expectation is taken over X and the training samples $\{(X_i, Y_i)\}_{i=1}^n$ used to fit $\hat{p}(X)$ [Stone, 1982]. In omniprediction, we compare to predictors in a restricted class $\mathcal{F}$, which allows us to circumvent the curse of dimensionality and recover more tractable rates. Furthermore, we note that this is not the same as simply targeting the projection of $p^*(X)$ onto $\mathcal{F}$. Such a projection will be loss-dependent, whereas omniprediction requires high accuracy against all losses in $\mathcal{L}_0$ simultaneously.

3 Omniprediction via multicalibration or calibrated multiaccuracy

Starting with Gopalan et al. [2022], various authors have considered algorithms for obtaining omniprediction via the stronger notions of multicalibration and calibrated multiaccuracy, such as Gopalan et al. [2023b,a], Globus-Harris et al. [2023], Kim and Perdomo [2023], Gopalan et al. [2024]. To define these targets formally, let $\mathcal{G}$ denote a class of functions mapping $\mathcal{X}$ to $\mathbb{R}$ and $p : \mathcal{X} \to [0,1]$ denote a prediction of $p^*(X)$. We say that $p(X)$ is multicalibrated with respect to $\mathcal{G}$ if

$$\mathbb{E}[g(X)(Y - p(X)) \mid p(X)] \overset{\text{as}}{=} 0, \text{ for all } g \in \mathcal{G}.$$

We say that $p(X)$ is calibrated if $\mathbb{E}[Y \mid p(X)] \overset{\text{as}}{=} p(X)$ and multiaccurate if

$$\mathbb{E}[g(X)(Y - p(X))] = 0, \text{ for all } g \in \mathcal{G}.$$

We use the term calibrated multiaccuracy to refer to predictors that are both calibrated and multiaccurate. In essence, multiaccuracy requires the predictor to be unbiased under all reweightings of the feature space by functions in $\mathcal{G}$, whereas calibration requires that the true and predicted frequencies of $Y = 1$ match over all instances where we make the same prediction. Multicalibration goes further by combining these conditions into a single statement. As a sanity check, one can verify that the true conditional probability $p^*(X)$ satisfies all three of these conditions.

Of course, our estimated predictor will never be exactly calibrated or multiaccurate, and to measure its discrepancy from these targets, we define multicalibration, multiaccuracy, and expected calibration errors by

$$\begin{aligned}
\mathrm{MC}(p; \mathcal{G}) &= \sup_{g \in \mathcal{G}} \mathbb{E}\big[\big|\mathbb{E}[g(X)(Y - p(X)) \mid p(X)]\big|\big], \\
\mathrm{MA}(p; \mathcal{G}) &= \sup_{g \in \mathcal{G}} \big|\mathbb{E}[g(X)(Y - p(X))]\big|, \\
\mathrm{ECE}(p) &= \mathbb{E}\big[\big|p(X) - \mathbb{E}[Y \mid p(X)]\big|\big],
\end{aligned}$$

respectively. It is easy to verify that if the constant function $x \mapsto 1$ is in $\mathcal{G}$, then the multicalibration error of a predictor upper bounds both its multiaccuracy and expected calibration errors.

To connect these definitions to omniprediction, we will need to study a specific choice of the class $\mathcal{G}$. Let $\partial\mathcal{L}_0 = \{p \mapsto \ell(p,1) - \ell(p,0) : \ell \in \mathcal{L}_0\}$ denote the set of discrete derivatives of proper losses (with respect to theor second argument), and $\partial\mathcal{L}_0 \circ \mathcal{F} = \{x \mapsto \ell(f(x),1) - \ell(f(x),0) : \ell \in \mathcal{L}_0, f \in \mathcal{F}\}$ denote the composition of these functions with the comparator class $\mathcal{F}$. Below we recall a known bound on omniprediction error.

Theorem 1 (Corollary of Lemma 12, Proposition 13, and Theorem 17 in Gopalan et al. [2023a])**.** *For any predictor $p : \mathcal{X} \to [0,1]$,*

$$\mathrm{OP}(p; \mathcal{L}_0, \mathcal{F}) \le \mathrm{MA}(p; \partial\mathcal{L}_0 \circ \mathcal{F}) + \mathrm{ECE}(p) \le 2\mathrm{MC}(p; \partial\mathcal{L}_0 \circ \mathcal{F} \cup \{x \mapsto 1\}).$$

Despite the extensive study of calibrated multiaccuracy as a vehicle for omniprediction, little seems to be known about the relative difficulty of these two problems beyond Theorem 1. As we will now argue, the former is strictly more difficult and necessarily incurs a greater sample complexity. The underlying reason for this stems from two simple observations. First, in order to construct a predictor $\hat{p}(X)$ with low calibration error we must restrict the range of its outputs. In particular, to verify $|\mathbb{E}[Y \mid p(X) = q] - q|$ is small we need to have many points X_i for which $p(X_i) = q$, and this is only possible when $p(X)$ takes on a small number of distinct values. On the other hand, even for very simple function classes, all (approximately) multiaccurate predictors must have sufficient complexity to capture the correlations between $p^*(X)$ and $g(X)$. These two considerations create a natural tension between calibration and multiaccuracy resulting in the following.

Proposition 2. *Suppose $\mathcal{X} = \mathbb{R}$ and let $\mathcal{G} = \{x \mapsto x\}$ denote the singleton function class containing just the identity. Then for a universal constant $c > 0$,*

$$\inf_{\hat{p}} \sup_{P_{XY}} \mathbb{E}\big[\max\big\{\mathrm{MA}(\hat{p}; \mathcal{G}), \mathrm{ECE}(\hat{p})\big\}\big] \ge cn^{-2/5},$$

where the expectation is taken over training samples $\{(X_i, Y_i)\}_{i=1}^n \overset{\text{iid}}{\sim} P_{XY}$ used to fit $\hat{p} : \mathcal{X} \to [0,1]$.

Proposition 2 evaluates calibrated multiaccuracy over a simple singleton function class. To connect this choice of $\mathcal{G}$ with the compositional class $\partial\mathcal{L}_0 \circ \mathcal{F}$ appearing in Theorem 1, one may simply note that by taking $\mathcal{F} = \mathcal{G} = \{x \mapsto x\}$ and considering the squared loss we have $2x - 1 = (x-1)^2 - x^2 \in \mathcal{L}_0 \circ \mathcal{F}$. Using this fact, one may argue that in this case Proposition 2 goes through with $\mathcal{G}$ replaced by $\partial\mathcal{L}_0 \circ \mathcal{F}$, hence providing a lower bound on the difficulty of calibrated multiaccuracy when it is used to ensure omniprediction.

After quantifying the difficulty of calibration and multiaccuracy in combination, we will now also give a lower bound on the difficulty of obtaining multiaccuracy alone. Notably, (up to polylogarithmic factors) this lower bound matches the upper bound previously derived in Okoroafor et al. [2025].

Proposition 3. *Let $\mathcal{G}$ denote a set of functions of finite VC dimension which take values in $\{-1, 1\}$. Then for a universal constant $c > 0$,*

$$\inf_{\hat{p}} \sup_{P_{XY}} \mathbb{E}[\mathrm{MA}(p; \mathcal{G})] \geq c\sqrt{\frac{\mathrm{VC}(\mathcal{G})}{n}},$$

where again the expectation is taken over training samples $\{(X_i, Y_i)\}_{i=1}^n \overset{\text{iid}}{\sim} P_{XY}$ used to fit $\hat{p} : \mathcal{X} \to [0, 1]$.

Once again, by choosing $\mathcal{F}$ appropriately, we can connect Proposition 3 to omniprediction. For instance, note that the standard 0-1 loss $\ell(p, y) = \mathbb{1}\{p \leq 1/2, y = 1\} + \mathbb{1}\{p > 1/2, y = 0\}$ is proper. If the functions in $\mathcal{F}$ output values in $\{0, 1\}$, their composition with the discrete derivative of ℓ can be written as

$$\ell(f(x), 1) - \ell(f(x), 0) = \begin{cases} -1, & f(x) = 1, \\ +1, & f(x) = 0. \end{cases}$$

and the lower bound of Proposition 3 holds with $\mathcal{G}$ replaced by $\mathcal{L}_0 \circ \mathcal{F}$ and $\mathrm{VC}(\mathcal{G})$ replaced by $\mathrm{VC}(\mathcal{F})$.

More generally, by combining the above two results, we see that for $\mathcal{F}$ of finite VC dimension calibrated multiaccuracy cannot be obtained at a rate better than $\sqrt{\mathrm{VC}(\mathcal{F})/n} + n^{-2/5}$. As we will see shortly, this is strictly worse than the optimal rate of $\sqrt{\mathrm{VC}(\mathcal{F})/n}$ (up to polylogarithmic factors) for omniprediction. Thus, methods targeting calibrated multiaccuracy and multicalibration cannot possibly produce optimal algorithms for the omniprediction problem.

To round out our discussion, we finish this section by giving a new algorithm for calibrated multiaccuracy which obtains an error bound of $O_{\mathbb{P}}(\sqrt{\mathrm{VC}(\mathcal{F})/n} + n^{-1/3})$. This rate is almost identical to our lower bound, which has a slightly larger exponent on the second term, and improves on previous methods for this problem as well as for multicalibration, which typically incur sample complexities of order $(\mathrm{VC}(\mathcal{F})/n)^{1/k}$ where $k \geq 4$ (e.g., Gopalan et al. [2023a], Globus-Harris et al. [2023], Okoroafor et al. [2025]). Unfortunately, the algorithm we present is not computationally tractable due to the fact that it requires iterating over all functions in $\mathcal{G}$. Hence, our goal in presenting this result is not to give a new practical method for calibrated multiaccuracy, but instead to help delineate the best rates one can expect for this problem. We leave it as an open problem to close the gap between the upper bound provided by this method and our lower bounds. Finally, we note that while we state the next result for finite function classes, it can be readily extended to infinite classes by taking an appropriate cover.

Proposition 4. *Let $\mathcal{G}$ be a finite class of functions which take values in $[-1, 1]$. Then, given i.i.d. training samples $\{(X_i, Y_i)\}_{i=1}^n \subseteq \mathcal{X} \times \{0, 1\}$, there is an algorithm which inputs these samples and outputs a predictor $\hat{p}(X)$ such that*

$$\max\left\{\mathrm{MA}(\hat{p}; \mathcal{G}), \mathrm{ECE}(\hat{p})\right\} \leq O_{\mathbb{P}}\left(\sqrt{\frac{\log(|\mathcal{G}|)}{n}} + \frac{1}{n^{1/3}}\right).$$

At a high-level, our method for achieving calibrated multiaccuracy uses a similar construction to two-player game based algorithms for omniprediction. Namely, it enumerates multiaccuracy and calibration as a list of multiple objectives for $\hat{p}(X)$ and best-responds to mixtures of these objectives in an online fashion. The next section gives a discussion of methods of this type for omniprediction. Therefore, to avoid repetition, we defer a detailed description of our method for calibrated multiaccuracy (which provides the result in Proposition 4) to Appendix B.

4 Reduction to finite ensembling

In the following section, we will present two methods for obtaining omniprediction at optimal rates. Both of these algorithms will be based on a simplification of the omniprediction problem that replaces the general set of all proper losses with a discrete collection. This allows us to reduce omniprediction to an ensembling task over a finite set of competitors. Structured characterizations of certain classes of proper loss functions have a long history in the literature dating back to the foundational work of Savage [1971]. In what follows, we will draw in particular on Ehm et al. [2016].

To begin simplifying the problem, we will first restrict the omniprediction task to the set of losses which are left-continuous in the prediction. This simplification is not critical and in practice we believe it will have little effect on the performance of the predictors. For instance, for a finite action space the decision-making function

$$a(p;u) \in \operatorname*{argmax}_{a\in\mathcal{A}} \mathbb{E}_{Y'\sim\mathrm{Ber}(p)}[u(a,Y')] = \operatorname*{argmax}_{a\in\mathcal{A}} p(u(a,1)-u(a,0)) - u(a,0),$$

is an argmax over a finite collection of functions linear in p. In particular, this implies that as a function of its first argument $a(p;u)$ is piecewise constant with discontinuities at the values of p for which there are multiple optimal actions. To break ties at these points, we may define $a(p;u) = \lim_{q\uparrow p} \operatorname{argmax}_{a\in\mathcal{A}} \mathbb{E}_{Y'\sim\mathrm{Ber}(q)}[u(a,Y')]$ as the limiting action over as q approaches p from below. One can verify that under this choice the induced loss $\ell^u(p,y) = -u(a(p;u),y)$ is left-continuous. In general, we believe that the restriction of left-continuity is benign and captures most practical settings. A brief discussion on potential avenues for extending our results to non-left-continuous losses is given in Appendix C.

In addition to this continuity requirement, we will also restrict ourselves to losses that satisfy $\ell(0,0) = \ell(1,1) = 0$. This restriction has no material impact on our results, since given an arbitrary proper loss ℓ one may always substitute it with the translated loss $\tilde{\ell}(p,y) = \ell(p,y) - \ell(y,y)$ without changing the omniprediction error. In what follows, we use $\mathcal{L}_{\mathrm{lc}}$ to denote the set of losses satisfying the above restrictions.

Now, our main tool for simplifying $\mathcal{L}_{\mathrm{lc}}$ will be a representation for members of this class as mixtures of weighted 0-1 losses. More precisely, for any $\theta \in [0,1]$ let ℓ_θ denote the weighted 0-1 loss given by

$$\ell_\theta(p,y) = \theta\mathbb{1}\{p > \theta, y = 0\} + (1-\theta)\mathbb{1}\{p \le \theta, y = 1\}.$$

To develop intuition, one can interpret the cases $p > \theta$ and $p \le \theta$ as corresponding to predictions that $y = 1$ and $y = 0$, respectively. The values θ and $1-\theta$ then determine the relative weights given to errors in each of these predictions.

It is easy to verify that ℓ_θ is proper since for any $p \in [0,1]$,

$$\mathbb{E}_{Y'\sim\mathrm{Ber}(p)}[\ell_\theta(q,Y')] = \theta(1-p)\mathbb{1}\{q > \theta\} + p(1-\theta)\mathbb{1}\{q \le \theta\}, \tag{4.1}$$

and thus the minimizers of the loss are given by

$$\operatorname*{argmin}_{q\in[0,1]} \mathbb{E}_{Y'\sim\mathrm{Ber}(p)}[\ell_\theta(q,Y')] = \begin{cases} [0,\theta), & p < \theta, \\ (\theta,1], & p > \theta, \\ [0,1], & p = \theta. \end{cases}$$

In particular, we see that p is always a minimizer.

The key fact that we will use to simplify the omniprediction problem is the following decomposition of Ehm et al. [2016] which shows that any element of $\mathcal{L}_{\mathrm{lc}}$ can be obtained as a mixture of weighted 0-1 losses ℓ_θ, $\theta \in [0,1]$.

Theorem 2 (Theorem 1 of Ehm et al. [2016])**.** *For any $\ell \in \mathcal{L}_{\mathrm{lc}}$ there exists a nonnegative measure μ on $[0,1]$ such that $\mu([0,1]) \le 2$ and*

$$\ell(p,y) = \int_0^1 \ell_\theta(p,y)\, d\mu(\theta), \text{ for all } p \in [0,1] \text{ and } y \in \{0,1\}.$$

Applying Theorem 2 to the omniprediction problem we have the equalities,

$$\begin{aligned}\mathrm{OP}(\hat{p}; \mathcal{L}_{\mathrm{lc}}, \mathcal{F}) &= \sup_{\ell \in \mathcal{L}_{\mathrm{lc}}, f \in \mathcal{F}} \mathbb{E}[\ell(\hat{p}(X), Y)] - \mathbb{E}[\ell(f(X), Y)] \\ &= \sup_{\mu, f \in \mathcal{F}} \int_0^1 \mathbb{E}[\ell_\theta(\hat{p}(X), Y) - \ell_\theta(f(X), Y)]\, d\mu(\theta) \\ &= 2 \sup_{\theta \in [0,1], f \in \mathcal{F}} \mathbb{E}[\ell_\theta(\hat{p}(X), Y)] - \mathbb{E}[\ell_\theta(f(X), Y)].\end{aligned}$$

In particular, we find that the omniprediction error equals twice the maximum possible error over all weighted 0-1 losses. To complete our simplification, we will show that it is sufficient to approximate this last quantity by θ falling in a discrete set.

Fix $m \in \mathbb{N}$. Given an arbitrary $\theta \in [0, 1]$ our goal will be to round it to the grid $\{\frac{i}{m} - \frac{1}{2m} : i \in \{1, \ldots, m\}\}$. For ease of notation in what follows, let $\theta_i = \frac{i}{m} - \frac{1}{2m}$. Our first step will be to restrict our predictor $\hat{p}(X)$ to lie on the grid $\{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\}$. This restriction is completely innocuous and will be guaranteed by all of the algorithms developed in the subsequent sections. Second, we will assume that the function class $\mathcal{F}$ is closed under all constant translations. The issue we need to avoid is the case in which there is a predictor $f_\theta \in \mathcal{F}$ that is optimal under a loss ℓ_θ but its performance cannot be (approximately) replicated under the rounded loss ℓ_{θ_i} for θ_i taken to be the value on the grid closest to θ. This edge case is ruled out by the assumption of closedness of $\mathcal{F}$ under constant translations (however, this particular condition is not critical, and it can be replaced by a variety of other sufficient conditions).

Under the above restrictions, we have the following simplification of the omniprediction error.

Lemma 1. *Suppose that $\mathcal{F}$ is closed under constant addition. Then for any predictor $p : \mathcal{X} \to \{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\}$,*

$$\sup_{\theta \in [0,1], f \in \mathcal{F}} \mathbb{E}[\ell_\theta(p(X), Y)] - \mathbb{E}[\ell_\theta(f(X), Y)] \leq \sup_{i \in \{1, \ldots, m\}, f \in \mathcal{F}} \mathbb{E}[\ell_{\theta_i}(p(X), Y)] - \mathbb{E}[\ell_{\theta_i}(f(X), Y)] + \frac{1}{m}.$$

Using this simplification, we will split our methods for constructing a predictor $\hat{p}(X)$ into two steps. In the first step, we find predictors $\{\hat{f}_{\theta_i}\}_{i=1}^m$ which empirically minimize the losses $\{\ell_{\theta_i}\}_{i=1}^m$, respectively. If $\mathcal{F}$ is a class of finite VC dimension and we define $\hat{f}_{\theta_i}$ to be the empirical risk minimizer of ℓ_{θ_i} over a sample of size n, then standard arguments (e.g., see Theorem 6.8 of Shalev-Shwartz and Ben-David [2014]) guarantee

$$\sup_{i \in \{1, \ldots, m\}, f \in \mathcal{F}} \mathbb{E}[\ell_{\theta_i}(\hat{f}_{\theta_i}(X), Y)] - \mathbb{E}[\ell_{\theta_i}(f(X), Y)] \leq O_{\mathbb{P}}\left(\sqrt{\frac{\mathrm{VC}(\mathcal{F}) \log(m)}{n}}\right). \tag{4.2}$$

Then, in the second step we will ensemble $\{\hat{f}_{\theta_i}\}_{i=1}^m$ into a single predictor $\hat{p}(X)$ minimizing

$$\sup_{i \in \{1, \ldots, m\}} \mathbb{E}[\ell_{\theta_i}(\hat{p}(X), Y)] - \mathbb{E}[\ell_{\theta_i}(\hat{f}_{\theta_i}(X), Y)]. \tag{4.3}$$

The remainder of this article will be focused on methods for performing the second step (4.3). For simplicity in what follows, we will assume that $\{\hat{f}_{\theta_i}\}_{i=1}^m$ are fixed in advance and the entire dataset $\{(X_i, Y_i)\}_{i=1}^n$ is used for ensembling. In practice, and in the examples we consider later, we will split the data into two parts: one for fitting the base predictors $\{\hat{f}_{\theta_i}\}_{i=1}^m$ and the other for ensembling them to derive the omnipredictor.

5 Sample-efficient methods for omniprediction

5.1 Method based on two-player games

We now present our first sample-efficient algorithm for omniprediction, based on a formulation of omniprediction as a two-player game in which one player maintains a mixture over the omniprediction objectives and the other player responds with a predictor which performs well on that mixture. To formalize this, let $q = (q_i)_{i=1}^m$

denote a probability distribution over $\{\theta_i\}_{i=1}^m$ where q_i denotes the probability of observing θ_i. Consider the mixture over omniprediction objectives given by

$$\ell(p,(x,y);q) = \sum_{i=1}^m q_i(\ell_{\theta_i}(p,y) - \ell(\hat{f}_{\theta_i}(x),y)).$$

The goal of the first player in the game will be to maximize the mixture loss in expectation, i.e., construct a mixture such that

$$\mathbb{E}[\ell(\hat{p}(X),(X,Y);q)] \approx \max_{i\in\{1,\dots,m\}} \mathbb{E}[\ell_{\theta_i}(\hat{p}(X),Y)] - \mathbb{E}[\ell(\hat{f}_{\theta_i}(X),Y)]. \tag{5.1}$$

The goal of the second player is to learn $\hat{p}(X)$ that minimizes the expected mixture loss. Under (5.1), this is equivalent to the ensembling step (4.3) (minimizing the worst-case excess loss against a base predictor), and by the results of the last section, is sufficient to guarantee that $\hat{p}(X)$ has small omniprediction error.

In our algorithm, the two players will execute on these objectives in an online fashion. To guarantee (5.1), the first player will use the well-known hedge algorithm, which learns q using online mirror descent (or more accurately, mirror ascent) over the probability simplex [Vovk, 1990, Littlestone and Warmuth, 1994, Freund and Schapire, 1997]. To respond to q, the second player will solve a min-max program that protects against the unknown distribution of $Y \mid X$. More precisely, if we let Δ_m denote the set of probability distributions over $\{0, \frac{1}{m}, \frac{2}{m}, \dots, 1\}$, then the second player will form its (randomized) prediction at x by solving

$$\min_{P\in\Delta_m} \max_{p_y\in[0,1]} \mathbb{E}_{Y'\sim\mathrm{Ber}(p_y),p\sim P}[\ell(p,(x,Y');q)]. \tag{5.2}$$

A critical observation underlying the success of this algorithm is the following bound.

Lemma 2. *For any $x \in \mathcal{X}$, the program in (5.2) has optimal value at most zero.*

Proof. The optimization problem (5.2) is bilinear in P and p_y, and thus by von Neumann's min-max theorem [von Neumann et al., 1944] we may swap the order of minimization and maximization to obtain

$$\min_{P\in\Delta_m} \max_{p_y\in[0,1]} \mathbb{E}_{Y'\sim\mathrm{Ber}(p_y),p\sim P}[\ell(p,(x,Y');q)] = \max_{p_y\in[0,1]} \min_{P\in\Delta_m} \mathbb{E}_{Y'\sim\mathrm{Ber}(p_y),p\sim P}[\ell(p,(x,Y');q)]. \tag{5.3}$$

Since each of the losses $\{\ell_{\theta_i}\}_{i=1}^m$ are proper, we additionally have that for each i,

$$\mathbb{E}_{Y'\sim\mathrm{Ber}(p_y)}[\ell_{\theta_i}(p_y,Y')] - \mathbb{E}_{Y'\sim\mathrm{Ber}(p_y)}[\ell_{\theta_i}(\hat{f}_{\theta_i}(x),Y')] \le 0,$$

thus $\mathbb{E}_{Y'\sim\mathrm{Ber}(p_y)}[\ell(p_y,(x,Y');q)] \le 0$. Moreover, it is easy to check that the value of $\ell_{\theta_i}(p_y,Y')$ is unchanged when p_y is rounded to its nearest value on the grid $\{0, \frac{1}{m}, \frac{2}{m}, \dots, 1\}$ (where ties are broken by rounding down). Setting P to be the distribution that puts all its mass on this rounded value in the inner minimization on the right-hand side in (5.3) gives the desired result. □

After only minor transformations, the optimization problem (5.2) can be written as a linear program with m variables and two constraints corresponding to the values $y \in \{0,1\}$. It can can then be solved by calling any standard convex solver. However, in the implementation of our two-played game based omniprediction algorithm, we will need to solve (5.2) repeatedly, and hence even a slight inefficiency will be magnified when the solver is called many times. Fortunately, by exploiting the structure of weighted 0-1 losses we can derive a more direct characterization of the solution, allowing us to solve (5.2) in $O(m)$ time.

Lemma 3. *Fix any $m \in \mathbb{N}$, $x \in \mathcal{X}$, and probability distribution q. Define the optimal values*

$$\theta^* = \max\left\{\theta \in \left\{0, \frac{1}{m}, \frac{2}{m}, \dots, 1\right\} : \sum_{i=1}^m q_i \mathbb{1}\{\theta \le \theta_i\} \ge \sum_{i=1}^m q_i \mathbb{1}\{\hat{f}_{\theta_i}(x) \le \theta_i\}\right\},$$

$$\rho^* = \frac{\sum_{i=1}^m q_i \mathbb{1}\{\theta^* \le \theta_i\} - \sum_{i=1}^m q_i \mathbb{1}\{\hat{f}_{\theta_i}(x) \le \theta_i\}}{q_{m\theta^*+1}},$$

with the caveat that $\rho^* = 0$ if $\theta^* = 1$. Then, $P^* = (1-\rho^*)\delta_{\theta^*} + \rho^*\delta_{\theta^*+1/m}$ solves (5.2).

Algorithm 1: Two-player game based omniprediction

Input: training samples $\{(X_i, Y_i)\}_{i=1}^n$, base predictors $\{\hat{f}_{\theta_i}\}_{i=1}^m$, learning rate $\eta > 0$

initialize $q_i(1) = \frac{1}{m}$, for all $i \in \{1, \ldots, m\}$;

for $t = 1, \ldots, n$ **do**

  $\hat{P}_t(x) \in \operatorname{argmin}_{P \in \Delta_m} \max_{p_y \in [0,1]} \mathbb{E}_{Y' \sim \mathrm{Ber}(p_y), p \sim P}[\ell(p, (x, Y'); q(t))]$, for all $x \in \mathcal{X}$;

  $\tilde{q}_i(t+1) = q_i(t) \exp(\eta(\mathbb{E}_{p \sim \hat{P}_t(X_t)}[\ell_{\theta_i}(p, Y_t)] - \ell_{\theta_i}(\hat{f}_{\theta_i}(X_t), Y_t)))$, for all $i \in \{1, \ldots, m\}$;

  $q_i(t+1) = \frac{\tilde{q}_i(t+1)}{\sum_{j=1}^m \tilde{q}_j(t+1)}$, for all $i \in \{1, \ldots, m\}$;

return $\hat{P} = \frac{1}{n} \sum_{t=1}^n \hat{P}_t$

Algorithm 1 now gives a complete description of our two-player game based method for omniprediction. As stated in Theorem 3 below, this obtains (up to polylog factors) the optimal omniprediction error rate of $\sqrt{\mathrm{VC}(\mathcal{F})/n}$. The proof of this theorem is provided in Appendix E.1. The main idea is to combine Lemma 2 with a regret bound for $q(t)$ that formalizes (5.1) and guarantees that the learned mixture losses are a good proxy for the omniprediction objective. These two results are sufficient to control the online omniprediction error. Generalization to new test samples is then obtained through a standard online-to-batch conversion and the Azuma-Hoeffding inequality.

Theorem 3. *Let $\mathcal{F}$ be a function class with finite VC dimension and assume that the base predictors $\{\hat{f}_{\theta_i}\}_{i=1}^m$ satisfy (4.2). Then, setting $m = \Theta(\sqrt{\log(n)/n})$ and $\eta = \Theta(\sqrt{n/\log(m)})$, Algorithm 1 produces a distribution $\hat{P}(X)$ such that the randomized predictor $\hat{p}(X) \sim \hat{P}(X)$ has omniprediction error*

$$\mathrm{OP}(\hat{p}; \mathcal{L}_{\mathrm{lc}}, \mathcal{F}) \leq \tilde{O}_{\mathbb{P}}\left(\sqrt{\frac{\mathrm{VC}(\mathcal{F})}{n}}\right).$$

As discussed in the introduction, we are not the first to propose a method for omniprediction based on two-player games. Garg et al. [2024], Okoroafor et al. [2025] both develop two-player game based algorithms that achieve an online omniprediction (up to polylog terms) at the rate $\sqrt{\mathrm{VC}(\mathcal{F})/n}$. As noted by Okoroafor et al. [2025], applying an online-to-batch conversion to these procedures then gives an offline omniprediction method with the same error rate. The main contribution of Algorithm 1 relative to these methods is that it is simpler to compute and implement. This stems from the fact that we have offloaded the optimization over $\mathcal{F}$ to a separate step where we construct the base predictors $\{\hat{f}_{\theta_i}\}_{i=1}^m$. In contrast, the methods of Garg et al. [2024] and Okoroafor et al. [2025] must perform substantial additional computation to handle the entire set of competitors in $\mathcal{F}$ at every iteration of the online algorithm. Nonetheless, Algorithm 1 is similar to existing approaches. In the next subsection, we go farther and develop a direct ensembling approach which achieves sample efficiency without randomization.

5.2 Direct ensembling

We develop a new omniprediction algorithm that more directly exploits the structure of weighted 0-1 losses. The predictor produced by Algorithm 1 is randomized, and to compute the distribution of its prediction at a given test point X, we must solve a sequence of $2n$ linear programs (minimally, we require $\hat{P}_t(X_t)$ and $\hat{P}_t(X)$ for each t). These issues are not unique to Algorithm 1, and other two-player game based algorithms share similar shortcomings. Recently, Okoroafor et al. [2025] raised the question of whether it is possible to achieve sample-efficient omniprediction without randomization. Here, we answer this question in the affirmative for proper losses.

5.2.1 Warm-up: ensembling two predictors

To motivate our method, it is useful to consider the simplest case in which we have just two base predictors, $\hat{f}_{\theta_h}$ and $\hat{f}_{\theta_l}$, associated with parameters $\theta_h > \theta_l$. Recall that for weighted 0-1 losses there are effectively only

two predictions: namely, given parameter θ we may either output the prediction $\hat{p}(X) > \theta$ or $\hat{p}(X) \le \theta$. The first (respectively, second) prediction is optimal when $p^*(X) \ge \theta$ (respectively, $p^*(X) \le \theta$). Extending this to the pair $\hat{f}_{\theta_h}(X)$ and $\hat{f}_{\theta_l}(X)$ we find that there are four possible cases:

$$\text{i. } \{\hat{f}_{\theta_h}(X) > \theta_h,\ \hat{f}_{\theta_l}(X) > \theta_l\}, \quad \text{ii. } \{\hat{f}_{\theta_h}(X) \le \theta_h,\ \hat{f}_{\theta_l}(X) \le \theta_l\},$$
$$\text{iii. } \{\hat{f}_{\theta_h}(X) \le \theta_h,\ \hat{f}_{\theta_l}(X) > \theta_l\}, \quad \text{iv. } \{\hat{f}_{\theta_h}(X) > \theta_h,\ \hat{f}_{\theta_l}(X) \le \theta_l\}.$$

In the first three cases, the predictions of $\hat{f}_{\theta_h}(X)$ and $\hat{f}_{\theta_l}(X)$ are consistent with each other, and to obtain a small omniprediction error we may simply define $\hat{p}(X)$ to agree with both of them. In particular, in case i we can set $\hat{p}(X) > \theta_h > \theta_l$, in case ii we can set $\hat{p}(X) \le \theta_l < \theta_h$ and in case iii we can set $\theta_l < \hat{p}(X) \le \theta_h$. On the other hand, in case iv the predictions of $\hat{f}_{\theta_h}(X)$ and $\hat{f}_{\theta_l}(X)$ are contradictory. To resolve this disagreement, we can examine the data and set

$$\hat{p}(X) \in \begin{cases} (\theta_h, 1], & \text{if } \hat{\mathbb{E}}_n(Y \mid \hat{f}_{\theta_h}(X) > \theta_h, \hat{f}_{\theta_l}(X) \le \theta_l) > \theta_h, \\ (\theta_l, \theta_h], & \text{if } \hat{\mathbb{E}}_n(Y \mid \hat{f}_{\theta_h}(X) > \theta_h, \hat{f}_{\theta_l}(X) \le \theta_l) \in (\theta_l, \theta_h], \\ [0, \theta_l], & \text{if } \hat{\mathbb{E}}_n(Y \mid \hat{f}_{\theta_h}(X) > \theta_h, \hat{f}_{\theta_l}(X) \le \theta_l) \le \theta_l, \end{cases}$$

where $\hat{\mathbb{E}}_n$ denotes the empirical expectation operator over $\{(X_i, Y_i)\}_{i=1}^n$. As the following lemma shows, this definition produces low omniprediction error.

Lemma 4. *Fix any $\theta_h > \theta_l$ and predictors $\hat{f}_{\theta_h}$ and $\hat{f}_{\theta_l}$. Let $\hat{p}(X)$ be defined as above. Then,*

$$\max_{\theta \in \{\theta_l, \theta_h\}} \mathbb{E}[\ell_\theta(\hat{p}(X), Y)] - \mathbb{E}[\ell_\theta(\hat{f}_\theta(X), Y)] \le O_{\mathbb{P}}\left(\sqrt{\frac{1}{n}}\right).$$

Proof. For simplicity, we will only consider the case $\theta = \theta_l$. The case $\theta = \theta_h$ is similar. Let $E = \{\hat{f}_{\theta_h}(X) > \theta_h, \hat{f}_{\theta_l}(X) \le \theta_l\}$ denote the event where the predictors disagree. By construction, we have

$$\begin{aligned} \mathbb{E}[\ell_{\theta_l}(\hat{p}(X), Y)] - \mathbb{E}[\ell_{\theta_l}(\hat{f}_{\theta_l}(X), Y)] &= \mathbb{E}[(\ell_{\theta_l}(\hat{p}(X), Y) - \ell_{\theta_l}(\hat{f}_{\theta_l}(X), Y))\mathbb{1}\{E\}] \\ &= \mathbb{E}[(\ell_{\theta_l}(1, Y) - \ell_{\theta_l}(0, Y))\mathbb{1}\{E\}]\mathbb{1}\{\hat{\mathbb{E}}_n[Y \mid E] > \theta_l\} \\ &= \mathbb{E}[(\ell_{\theta_l}(1, Y) - \ell_{\theta_l}(0, Y))\mathbb{1}\{E\}]\mathbb{1}\big\{\hat{\mathbb{E}}_n[(\ell_{\theta_l}(1, Y) - \ell_{\theta_l}(0, Y))\mathbb{1}\{E\}] < 0\big\}, \end{aligned}$$

where the last equality follows from the definition of ℓ_{θ_l}. This last quantity can be bounded using Hoeffding's inequality. □

5.2.2 General case: ensembling m predictors

Extending Lemma 4 beyond two predictors requires considerable care. The m predictors $\{\hat{f}_{\theta_i}\}_{i=1}^m$ can make a total of 2^m combinations of predictions (corresponding to each predictor $\hat{f}_{\theta_i}$ lying on either side of θ_i), the vast majority of which contain some disagreements. Notably, we cannot obtain accurate estimates of the true probability of $Y = 1$ under all of these combinations simultaneously. As a result, instead of evaluating these events individually, we will devise an iterative scheme in which the predictors are ensembled in groups.

The main primitive in these iterations is a merge algorithm that takes as input two predictors $\hat{p}_h(X)$ and $\hat{p}_l(X)$ which are designed to give low error on losses ℓ_θ for $\theta \in \Theta_h$ and $\theta \in \Theta_l$, respectively. The sets (Θ_h, Θ_l) are constructed so that $\theta_h > \theta_l$ for all $\theta_h \in \Theta_h$ and $\theta_l \in \Theta_l$. The output of the merge method will be a single predictor, $\hat{p}_m(X)$ that obtains loss comparable to $\hat{p}_h(X)$ on all parameters $\theta_h \in \Theta_h$ and comparable to $\hat{p}_l(X)$ on all parameters $\theta_l \in \Theta_l$.

This merge procedure resolves disagreements between $\hat{p}_h(X)$ and $\hat{p}_l(X)$. This is done using the following iterative scheme. We begin by simply positing that $\hat{p}_h(X)$ is a good predictor, and hence set $\hat{p}_m(X) = \hat{p}_h(X)$. This will guarantee that $\hat{p}_m(X)$ has good performance on Θ_h, but it leaves open the possibility that it fails on (a subset of) Θ_l. To address this, we iterate through the elements $\theta_l \in \Theta_l$ in descending order and examine the empirical expectation $\hat{\mathbb{E}}_n[(\ell_{\theta_l}(1, Y) - \ell_{\theta_l}(0, Y))\mathbb{1}\{X \in E\}]$, where $E = \{x : \hat{p}_h(x) > \min \Theta_h,\ \hat{p}_l(x) \le \theta_l\}$

is the event where the predictors disagree. If this expectation is negative then it means that predicting a high value gives a low loss and thus $\hat{p}_m(X)$ will give good performance on ℓ_{θ_l}. On the other hand, if it is positive then we need to predict a small value. To account for this, we modify our predictor such that $\hat{p}_m(x) = \hat{p}_l(x)$ for all $x \in E$. Notably, due to the hierarchical structure of the weighted 0-1 loss family, this modification will maintain that $\hat{p}_m(X)$ is a good predictor on all previously considered parameters $\theta \in \Theta_l$ with $\theta > \theta_l$. To see this, note that for any such θ,

$$\begin{aligned}\hat{\mathbb{E}}_n[(\ell_\theta(1,Y) - \ell_\theta(0,Y)) \mid X \in E] &= \theta - \hat{\mathbb{P}}_n(Y = 1 \mid X \in E) \\ &\geq \theta_l - \hat{\mathbb{P}}_n(Y = 1 \mid X \in E) \\ &= \hat{\mathbb{E}}_n[(\ell_{\theta_l}(1,Y) - \ell_{\theta_l}(0,Y)) \mid X \in E] > 0.\end{aligned}$$

However, it may now give poor performance on some losses in Θ_h; this is corrected by performing a similar set of iterations over the parameters in Θ_h. Eventually, after repeating this entire process many times we will have evaluated all parameters in Θ_h and Θ_l and certified the performance of $\hat{p}_m(X)$ on each of them.

Algorithm 2: Merge

Input: training samples $\{(X_i, Y_i)\}_{i=1}^n$, predictors $\hat{p}_l$, $\hat{p}_h$, parameter sets Θ_l, Θ_h, hyperparameter $\epsilon \geq 0$

$\hat{p}_m = \hat{p}_h$;
$\theta_h = \min \Theta_h$;
$\theta_l = \max \Theta_l$;
dir = low;
while $\theta_l \neq -\infty$, $\theta_h \neq \infty$ **do**
  $E = \{x : \hat{p}_h(x) > \theta_l, \hat{p}_l(x) \leq \theta_l\}$;
  if *dir* = *low* **then**
    if $\hat{\mathbb{E}}_n[(\ell_{\theta_l}(1,Y) - \ell_{\theta_l}(0,Y))\mathbb{1}\{X \in E\}] < -\epsilon$ **then**
      $\hat{p}_m(x) = \hat{p}_l(x)$, for all $x \in E$;
      $\theta_h = \min\{\theta \in \Theta_h : \theta > \theta_h\}$;
      dir = high;
    else
      $\theta_l = \max\{\theta \in \Theta_l : \theta < \theta_l\}$;
  else
    // Do a symmetric set of iterations through Θ_h in which we alter the value
    // of $\hat{p}_m(X)$ if $\hat{\mathbb{E}}_n[(\ell_{\theta_l}(1,Y) - \ell_{\theta_l}(0,Y))\mathbb{1}\{X \in E\}] > \epsilon$.
return $\hat{p}_m$

Algorithm 2 gives a summary of the merge method, a more detailed description of which can be found in Appendix E.2. In total, this algorithm will evaluate each element of Θ_h and Θ_l at most once and thus will be guaranteed to run in at most $|\Theta_h| + |\Theta_l|$ iterations. In addition to the description given above, Algorithm 2 contains one additional hyperparameter ϵ, which gives a buffer on the improvement in the loss that must be observed before swapping $\hat{p}_m(X)$ between $\hat{p}_h(X)$ and $\hat{p}_l(X)$. In our theoretical results, correct specification of this hyperparameter is used to mitigate the sensitivity of $\hat{p}_m(X)$ to noise, and ensure its generalization to new data. The approach we take here is partially inspired by Deng and Hsu [2024], who use a similar buffer hyperparameter in a different context. On the other hand, in our experiments we find that the choise of the hyperparameter ϵ not crucial and the lowest omniprediction error is achieved when $\epsilon = 0$. As a result, we will not place a heavy emphasis on the choice of ϵ.

Lemma 5 states our formal guarantee on the omniprediction error of the merge procedure. In this lemma we assume that the values of $\hat{p}_h(X)$ and $\hat{p}_l(X)$ are restricted to $(\max \Theta_l, 1]$ and $[0, \min \Theta_h)$, respectively. The idea here is that $\hat{p}_h(X)$ (respectively $\hat{p}_l(X)$) only gives information about the parameters in Θ_h (respectively Θ_l) and does not give any signal about Θ_l (respectively Θ_h). In our applications of the merge procedure this assumption will be guaranteed to hold by construction.

Lemma 5. *Let Θ_h, Θ_l be finite subsets of $[0,1]$, with $\min\Theta_h > \max\Theta_l$, and assume that $\hat{p}_l(X) \in [0, \min\Theta_h)$ and $\hat{p}_h(X) \in (\max\Theta_l, 1]$. Then, setting $\epsilon = \Theta(\sqrt{\log(|\Theta_h| + |\Theta_l|)/n})$, Algorithm 2 returns a predictor $\hat{p}_m(X)$ such that*

$$\max_{a\in\{h,l\}} \max_{\theta\in\Theta_a} \mathbb{E}[\ell_\theta(\hat{p}_m(X), Y)] - \mathbb{E}[\ell_\theta(\hat{p}_a(X), Y)] \leq O_{\mathbb{P}}\left(\sqrt{\frac{\log(|\Theta_h|) + \log(|\Theta_l|)}{n}}\right).$$

With this merge procedure in hand, ensembling the full collection of base predictors $\{\hat{f}_{\theta_i}\}_{i=1}^m$ is relatively straightforward. Namely, we simply apply the merge procedure repeatedly, joining together predictors with adjacent parameters until we are left with a single function. Concretely, assume that $m = 2^k$ is a power of 2. Then, we will proceed in k rounds, where in each round adjacent predictors are paired up and then merged (e.g., in round 1 we merge the pairs $(\hat{f}_{\theta_1}, \hat{f}_{\theta_2}), \ldots, (\hat{f}_{\theta_{m-1}}, \hat{f}_{\theta_m})$). In order to guarantee the generalization of this method, each of these k rounds will use fresh data. This is specified on line 3 of Algorithm 3, where we use the notation $\text{Split}(\{(X_i, Y_i)\}_{i=1}^n)$ to denote a division of the training dataset into $\log_2(m)$ equally-sized folds. Here, data splitting ensures that the empirical expectations that appear in the merge procedure stay uniformly close to their population counterparts. In practice, we find that this is unnecessary and all of the data can be used at every round without issue.

Algorithm 3: Direct ensembling scheme for omniprediction

Input: training samples $\{(X_i, Y_i)\}_{i=1}^n$, base predictors $\{\hat{f}_{\theta_i}\}_{i=1}^m$, hyperparameter $\epsilon \geq 0$

1 $\hat{p}_{1,i} = \hat{f}_{\theta_i}$, for all $i \in \{1, \ldots, m\}$;
2 $\Theta_{1,i} = \{\theta_i\}$, for all $i \in \{1, \ldots, m\}$; `// $\hat{p}_{t,i}$ is designed to be "optimal" on $\Theta_{t,i}$`
3 $D_1, \ldots, D_{\log_2(m)} = \text{Split}(\{(X_i, Y_i)\}_{i=1}^n)$; `// Split the data into equally-sized parts`
4 **for** $t = 1, \ldots, \log_2(m)$ **do**
5 **for** $i = 1, \ldots, \frac{m}{2^t}$ **do**
6 $\hat{p}_{t+1,i} = \text{Merge}(D_t, \hat{p}_{t,2i-1}, \hat{p}_{t,2i}, \Theta_{t,2i-1}, \Theta_{t,2i}, \epsilon)$;
7 $\Theta_{t+1,i} = \Theta_{t,2i-1} \cup \Theta_{t,2i}$.
8 **return** $\hat{p} = \hat{p}_{\log_2(m),1}$

Algorithm 3 states our method formally. In this algorithm, and in what follows, we will assume that $\hat{f}_{\theta_i}$ takes values in $\{\theta_i - \frac{1}{2m}, \theta_i + \frac{1}{2m}\}$. This is always possible since given an arbitrary predictor $\tilde{f}_{\theta_i}$ with good performance under ℓ_{θ_i} we may always equivalently recode its predictions as

$$\hat{f}_{\theta_i}(X) = \left(\theta_i - \frac{1}{2m}\right)\mathbb{1}\{\tilde{f}_{\theta_i}(X) \leq \theta_i\} + \left(\theta_i + \frac{1}{2m}\right)\mathbb{1}\{\tilde{f}_{\theta_i}(X) > \theta_i\},$$

since we only need $\hat{f}_{\theta_i}(X)$ to provide information on whether $p^*(X)$ lies above or below θ_i. The next result establishes that this direct ensembling method achieves the optimal omniprediction error rate (up to polylog terms).

Theorem 4. *Let $\mathcal{F}$ be a function class with finite VC dimension and assume that the base predictors $\{\hat{f}_{\theta_i}\}_{i=1}^m$ satisfy (4.2). Then, setting $m = \Theta(2^{\lfloor\log_2(\sqrt{n})\rfloor})$ and $\epsilon = \Theta(\sqrt{\log(n)/n})$, Algorithm 3 returns a predictor $\hat{p}(X)$ with omniprediction error*

$$\text{OP}(\hat{p}; \mathcal{L}_{\text{lc}}, \mathcal{F}) \leq \tilde{O}_{\mathbb{P}}\left(\sqrt{\frac{\text{VC}(\mathcal{F})}{n}}\right).$$

6 Empirical comparisons

We now turn our attention to a set of empirical comparisons. Following the discussion in the earlier sections, we will evaluate three methods for omniprediction:

- **CalMA:** Our first method is the calibrated multiaccuracy scheme proposed in Algorithm 2 of Gopalan et al. [2023a]. This is a boosting method that iteratively updates $\hat{p}(X)$ by alternating between improving its multiaccuracy error and improving its calibration error. We will implement this algorithm so that it targets multiaccuracy with respect to the class $\mathcal{G} = \{x \mapsto \ell(\hat{f}_{\theta_i}(x), 1) - \ell(\hat{f}_{\theta_i}(x), 0) : i \in \{1, \ldots, m\}\}$. A straightforward consequence of Theorem 1 shows that this (combined with calibration) is sufficient to give low omniprediction error.

  The calibrated multiaccuracy procedure of Gopalan et al. [2023a] has a hyperparameter α, that specifies the target omniprediction error. The theory presented in that work suggests that this parameter should be chosen to be of order $\alpha = \Theta((\log(m)/n)^{-1/4} + n^{-1/10})$. In practice, we find that this is needlessly pessimistic and will prefer to take $\alpha = c\sqrt{\log(m)/n}$ for some constant c that we vary.

  Additionally, the theory for this method requires extensive data splitting in order to ensure that fresh samples are available for each of up to $O(1/\alpha^2)$ iterations of the algorithm. For the sample sizes we consider, this would give us only a handful of data points at each iteration with which to improve the multiaccuracy and calibration error. As this is clearly impractical, we do not perform any data splitting and simply use all available data at every step. As we will see shortly, this does not appear to be an issue and the algorithm gives reasonable empirical performance.

- **Two-player:** Our second algorithm is the two-player game based procedure given in Algorithm 1. We implement this method with hyperparameter $\eta = c\sqrt{\log(m)/n}$ for varying levels of c.

- **Direct ensembling:** Our third method is the direct ensembling scheme given in Algorithm 3. Similar to the previous methods, we implement this method with hyperparameter $\epsilon = c\sqrt{\log(m)/n}$ for varying levels of c. Additionally, as above, we do not utilize data splitting. We find that although our theoretical results require fresh data for every round of merging, in practice this method offers robust performance when all the available data is used at each step.

All three methods are implemented with the same value of m and the same set of base predictors $\{\hat{f}_{\theta_i}\}_{i=1}^m$. The exact procedure for obtaining these quantities varies for each experiment and is specified in the relevant subsections below.

6.1 Simulated example

For our first example, we consider a simple simulated dataset which illustrates the core ensembling problem. Define $\mathcal{F} = \{x \mapsto \beta_0 + \beta_1 x : \beta_0, \beta_1 \in \mathbb{R}\}$ to be the class of linear predictors on $\mathbb{R}$. Take X to be supported on $\{0.05, 0.45, 0.85\}$, with distribution $\mathbb{P}(X = 0.05) = 0.1$, $\mathbb{P}(X = 0.45) = 0.6$, and $\mathbb{P}(X = 0.85) = 0.3$; then let $Y \in \{0, 1\}$ be sampled according to $\mathbb{P}(Y \mid X = 0.05) = 0.3$, $\mathbb{P}(Y \mid X = 0.45) = 0.9$, and $\mathbb{P}(Y \mid X = 0.85) = 0.4$. By design, this distribution for (X, Y) has the property that the linear predictor $f^*_\theta \in \mathcal{F}$, optimal under loss ℓ_θ, creates inconsistent predictions as θ varies. For example, at $\theta = 0.35$ and $X = 0.05$, the optimal predictor outputs $f^*_{0.35}(0.05) \leq 0.35$, while at $\theta = 0.75$ it predicts $f^*_{0.75}(0.05) > 0.75$. This inconsistency in the optimal predictions is illustrated in the left panel of Figure 1, which plots the conditional distribution of Y given X alongside these optima.

The rightmost two panels of Figure 1 inspect the performance of the three main omniprediction methods over different sample sizes n and settings of m. To simplify our initial comparisons, the results in this figure show only a single hyperparameter setting for each method which was found to give good performance (with details given in the figure caption). Dots display empirical estimates of the average omniprediction error,

$$\mathbb{E}_{\{(X_i,Y_i)\}_{i=1}^n}\left[\max_{i\in\{1,\ldots,m\}} \mathbb{E}_{(X,Y)}[\ell_{\theta_i}(\hat{p}(X), Y)] - \mathbb{E}_{(X,Y)}[\ell_{\theta_i}(\hat{f}_{\theta_i}(X), Y)]\right],$$

over multiple draws of the training dataset $\{(X_i, Y_i)\}_{i=1}^n$; error bars show empirical estimates of the standard deviation of this error. The center panel shows results for a fixed value of $m = 16$ while the right panel gives results for $m = 2^{\lfloor \log_2(\sqrt{n}) \rfloor}$ increasing with the sample size. In both cases, each base predictor $\hat{f}_{\theta_i}$ is obtained

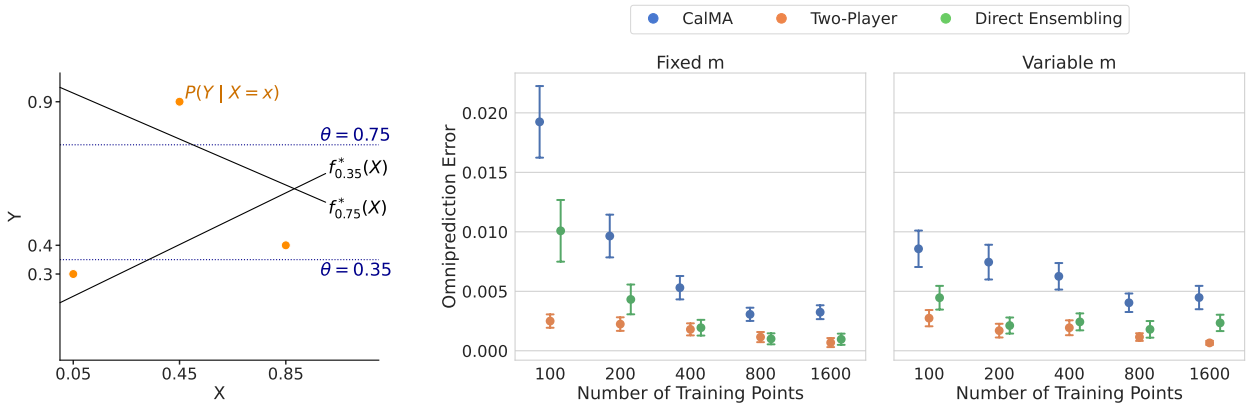

Figure 1: Illustration of the core ensembling problem for our simulated example (left panel) and realized average omniprediction error of the calibrated multiaccuracy (blue), two-player game based (orange), and direct ensembling (green) methods for various sample sizes with $m = 16$ fixed (center panel) or chosen variably as $m = 2^{\lfloor \log_2(\sqrt{n}) \rfloor}$ (right panel), for a simulated dataset. Dots and error bars display the means and standard errors obtained by evaluating the omniprediction error over 2000 test points for a repeated 40 draws of the training dataset. Hyperparameters for the calibrated multiaccuracy, two-player, and direct ensembling methods are set according to $c = 0.5$, $c = 32$, and $c = 0$, respectively.

by empirical minimization of the the loss ℓ_{θ_i} over an independent dataset of size 500 (this minimization can be recast as a mixed integer program).

The figure shows that the method based on calibrated multiaccuracy realizes the highest omniprediction error across all sample sizes and settings of m. Further, the two-player game based algorithm performs better than the direct ensembling method at smaller sample sizes, but the two exhibit similar performance at larger values of n. An advantage of the direct ensembling method is that it offers simplified hyperparameter tuning. Figure 2 displays the results for the three methods as the scaling constant c varies. We find that the direct ensembling method always performs best with $\epsilon = 0$. On the other hand, to obtain optimal performance with the two-player game based approach we must choose an intermediate value of η. In practice, selecting such a value may be challenging and could require additional data splitting.

6.2 Sales forecasting

Our second experiment compares the three omniprediction methods on a retail sales forecasting dataset taken from the M5 forecasting challenge [Makridakis et al., 2022]. In this challenge, competitors were tasked with constructing quantile forecasts of the daily sales of various items at ten different Walmart stores over a 28-day period. We transform this task to a binary prediction problem in which the goal is to estimate the probability that at least one unit of an item is sold at a given store on a given day. To do this, we use linear interpolation to convert the quantile forecasts given by the competitors into estimates of the full cumulative distribution function of the sales. We then set our function class $\mathcal{F}$ to be the corresponding forecasts of the probability that at least one sale is made. Details of this procedure are given in Appendix F. In total, the M5 dataset contains quantile forecasts from the top 50 participants in the competition, but to obtain a sufficient sample size for our experiments, we restrict our attention to the 43 forecasters who issued predictions for at least 10,000 product-store pairs on day 7.

We evaluate the omniprediction methods in three steps. First, to obtain $\{\hat{f}_{\theta_i}\}_{i=1}^m$ we randomly select 500 product-store pairs from the day 7 data. Then, for each $i \in \{1, \ldots, m\}$ we set $\hat{f}_{\theta_i}$ to be the element of $\mathcal{F}$ that minimizes the empirical loss ℓ_{θ_i}, over these 500 samples. With these initial predictors in hand, we then run the three omniprediction methods on a randomly chosen subset of the data from day 14. Finally, all methods are evaluated on the data from day 21.

Figure 3 shows the results of this experiment for various sample sizes n and settings of m. Similar to the previous subsection, the left panel shows results for a fixed value of $m = 16$ while the right panel gives results for $m = 2^{\lfloor \log_2(\sqrt{n}) \rfloor}$ increasing with n. Also, we display the best performing hyperparameter for each method.

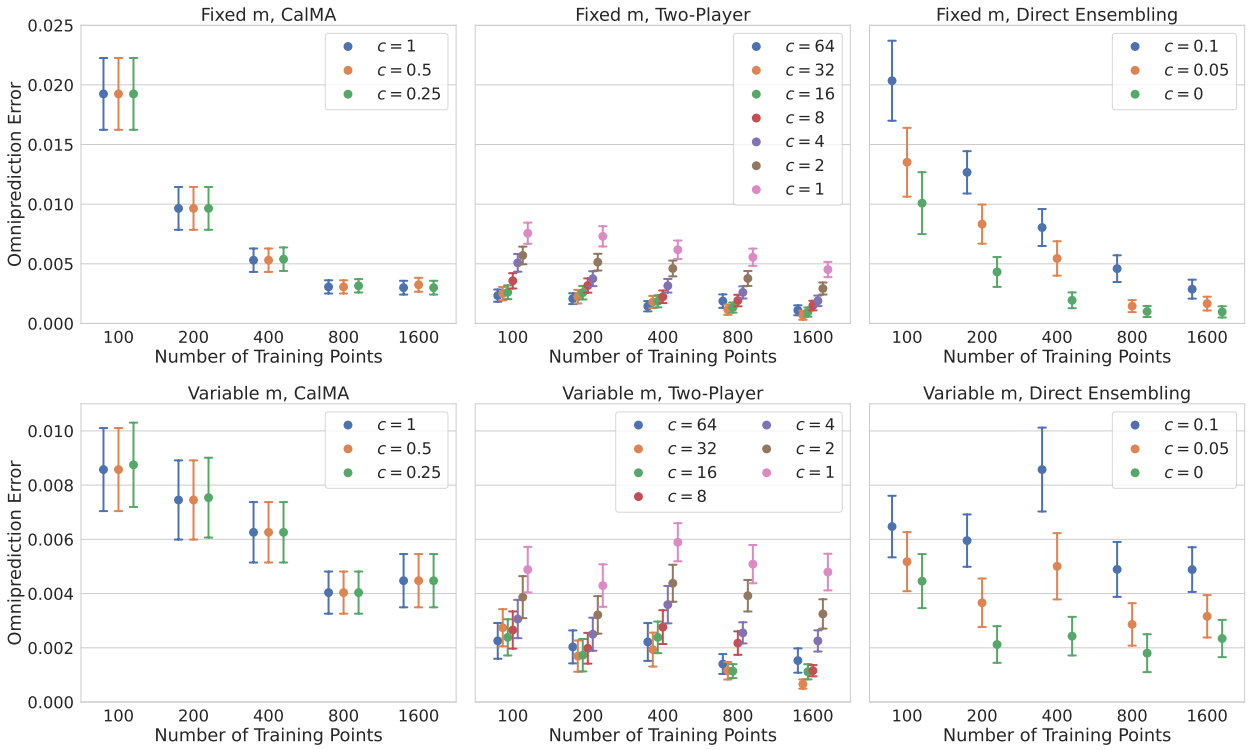

Figure 2: Omniprediction error of the calibrated multiaccuracy (left panels), two-player game based (center panels), and direct ensembling (right panels) methods across various sample sizes with $m = 16$ fixed (top row) or chosen variably as $m = 2^{\lfloor \log_2(\sqrt{n}) \rfloor}$ (bottom row) as the scaling constant c varies, for a simulated dataset. Dots and error bars show the means and standard errors obtained by evaluating the omniprediction error over 2000 test points for 40 draws of the training dataset.

Corresponding results for other parameter choices are given in Figure 4 in the appendix. In addition to the three omniprediction methods discussed above, this figure also displays results for the best performing base model, i.e., the predictor

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \max_{i \in \{1,\ldots,m\}} \frac{1}{n} \sum_{i=1}^{n} \ell_{\theta_i}(f_{\theta_i}(X_i), Y_i) - \frac{1}{n} \sum_{i=1}^{n} \ell_{\theta_i}(\hat{f}_{\theta_i}(X_i), Y_i),$$

that minimizes the empirical omniprediction error on the day 14 data.

As in the simulated example, the calibrated multiaccuracy method once again realizes the largest errors. Notably, this method is even outperformed by the best base model which offers no omniprediction guarantee. The two-player game based method again performs the best for small sample sizes and the direct ensembling method begins to close the gap at larger sample sizes. The two-player game based method has surprisingly strong performance for even the smallest sample sizes, with an omniprediction error of nearly zero for $n = 25$ (and varying m). This is likely due to the fact that even before observing any training samples the two-player game based approach forms an initial baseline ensemble of the available predictors (recall Lemma 3). In this example, this baseline performs well and thus the method does not require significant training data. On the other hand, the direct ensembling procedure requires additional training samples to effectively learn how to resolve disagreements between base predictors.

7 Discussion

This article studied three algorithmic frameworks for constructing predictors with low omniprediction error over the class of bounded proper losses (subject to minor continuity conditions). Our theoretical and empirical

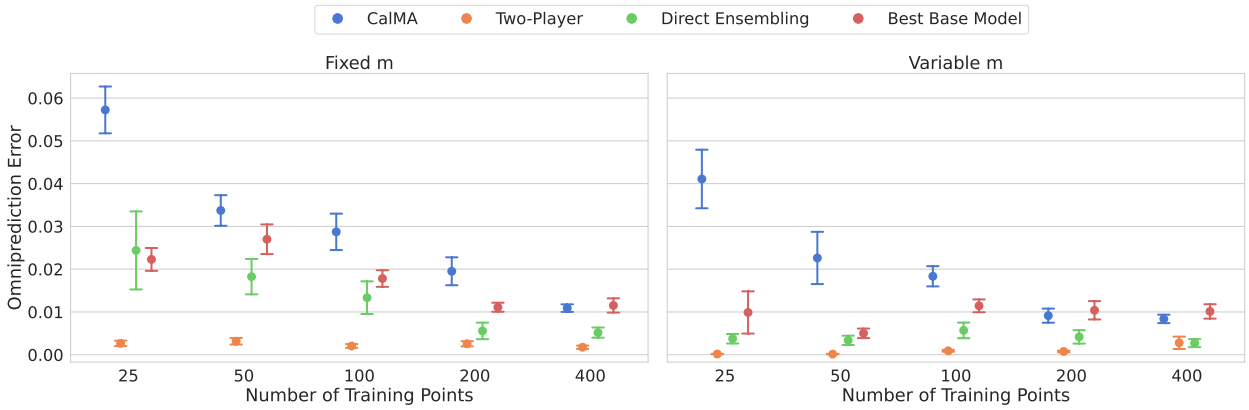

Figure 3: Realized average omniprediction error of the calibrated multiaccuracy (blue), two-player game based (orange), direct ensembling (green) methods, as well as the error of the best base model (red) across various sample sizes with $m = 16$ fixed (left panel) or chosen variably as $m = 2^{\lfloor \log_2(\sqrt{n}) \rfloor}$ (right panel), for the M5 sales forecasting dataset. Dots and error bars display the means and standard errors obtained by evaluating the omniprediction error over 2000 test points for a repeated 20 draws of the training dataset. Hyperparameters for the calibrated multiaccuracy, two-player, and direct ensembling procedures are set using $c = 0.5$, $c = 32$, and $c = 0$, respectively.

results show that methods based on calibrated multiaccuracy lead to larger error rates than those based on two-player games and a new direct ensembling approach, which takes advantage of the hierarchical structure of weighted 0-1 losses for classification. The latter two methods provide similar theoretical guarantees, with the two-player game based methods offering better empirical performance at smaller sample sizes.

7.1 Extensions to other prediction targets

In this paper, we have chosen to focus on binary classification in which the goal is to estimate the conditional probability function, $\mathbb{P}(Y = 1 \mid X)$. Perhaps surprisingly, the algorithms and theory we have developed are not unique to this problem and can be extended to handle a large variety of estimation targets. To formalize this, let T denote a map that takes in a distribution P on $\mathcal{Y}$ and returns an estimation target $T(P)$ of interest. In the previous sections, we studied $\mathcal{Y} = \{0, 1\}$ and $T(P) = \mathbb{P}_P(Y = 1)$. More generally, one may consider prediction tasks such as estimating the mean, $T(P) = \mathbb{E}_P[Y]$ or τ-quantile, $T(P) = \inf\{z : \mathbb{P}_P(Y \leq z) \geq \tau\}$ with $\mathcal{Y} = \mathbb{R}$. We say that T is an elicitable property of P if there exists at least one loss function which is minimized at $T(P)$, i.e., there exists ℓ such that for all P,

$$T(P) \in \operatorname*{argmin}_t \mathbb{E}_P[\ell(t, Y)].$$

It is worth noting that while some popular prediction targets such as means and quantiles are elicitable, not every property of a distribution can be obtained this way. A notable example is the conditional value-at-risk which is well-known to be nonelicitable [Gneiting, 2011].

Restricting now to elicitable properties, the goal is to design predictors that estimate $T(P_{Y|X})$ well under all possible losses for T. As above, we say that ℓ is a proper loss[†] for T if $T(P) \in \operatorname{argmin}_t \mathbb{E}_P[\ell(t, Y)]$ for all P and strictly proper if $T(P)$ is the unique minimizer. Recall, the key technical tool that allowed us to handle arbitrary proper losses in binary prediction was Theorem 2, which gave a decomposition of proper losses as mixtures of a one-dimensional family of weighted 0-1 losses. To extend our results beyond binary prediction, we can leverage the following result from Steinwart et al. [2014], which demonstrates the existence of similar decompositions for other targets. This result requires that T be strictly locally nonconstant: informally, this means that slight changes in P can shift $T(P)$ up or down. A more precise definition of this property is given as Definition 4 in Steinwart et al. [2014].

[†]Some authors call loss functions satisfying this condition consistent losses, while reserving the term proper for loss functions of entire distributions, not just functionals.

Proposition 5 (Variant of Corollary 9 of Steinwart et al. [2014]). *Let $(\mathcal{Y}, \mathcal{A}, \mu)$ be a separable finite measure space, let $\mathcal{P}$ be a set of μ-absolutely continuous distributions on $\mathcal{Y}$ and let $T : \mathcal{P} \to \mathbb{R}$ be continuous, elicitable, and strictly locally nonconstant, for which* $\mathrm{Image}(T) \subseteq \mathbb{R}$ *is an interval. Then, there is a measurable function $V : \mathrm{Image}(T) \times \mathcal{Y} \to \mathbb{R}$ that identifies T, i.e., a function V with the property that for all $t \in \mathrm{int}(\mathrm{Image}(T))$,*

$$\mathbb{E}_{Y\sim P}[V(t, Y)] = 0 \iff t = T(P) \quad \text{and} \quad \mathbb{E}_{Y\sim P}[V(t, Y)] > 0 \iff t > T(P).$$

Moreover, all strictly proper losses ℓ for T that are locally-Lipschitz in their first argument can be written as

$$\ell(t, y) = \int_{-\infty}^{\infty} V(\theta, y)\mathbb{1}\{t \le \theta\}w(\theta)\, d\theta + \kappa(y), \text{ for all } t \in \mathbb{R} \text{ and } \mu\text{-almost all } y \in \mathcal{Y}, \tag{7.1}$$

for some functions $w : \mathbb{R} \to [0, \infty)$ and $\kappa : \mathcal{Y} \to \mathbb{R}$ that depend on ℓ.

A key feature of Proposition 5 is the identification function V; common examples include $V(t, y) = t - y$, which identifies the mean, and $V(t, y) = \mathbb{1}\{y \le t\} - \tau$, which identifies the τ-quantile. The perhaps surprising implication of this proposition is that any (appropriately smooth) proper loss for the mean or τ-quantile can be written as a mixture over such identification functions.

With Proposition 5 in hand, omniprediction algorithms for other point prediction targets can be developed by replacing the weighted 0-1 losses underlying our methods with the threshold loss $\ell_\theta^T(t, y) = V(\theta, y)\mathbb{1}\{t \le \theta\}$. Similar to the binary case, the loss $\ell_\theta^T(t, y)$ is proper and effectively considers only two predictions, depending on whether t falls above or below θ. By replacing all instances of ℓ_θ with ℓ_θ^T in the previous sections, we can adapt Algorithms 1 or 3 to construct predictors $\hat{t}(X)$ satisfying the corresponding omniprediction guarantee

$$\sup_{\ell, f\in\mathcal{F}} \mathbb{E}[\ell(\hat{t}(X), Y)] - \mathbb{E}[\ell(f(X), Y)] \le \tilde{O}_{\mathbb{P}}\left(\sqrt{\frac{\mathrm{VC}(\mathcal{F})}{n}}\right),$$

where the supremum is over all proper losses for T satisfying appropriate regularity conditions. Making this statement precise requires some minor additional technical assumptions to ensure that the weight function w is appropriately bounded and the parameters θ can be discretized. We do not pursue this here.

A more challenging task is to extend our results beyond point prediction problems. For instance, given a multiclass outcome $Y \in \{1, \ldots, k\}$, one may attempt to construct estimates of the entire vector of conditional probabilities $(\mathbb{P}(Y = 1 \mid X), \ldots, \mathbb{P}(Y = k \mid X))$. However, the class of proper losses in this problem setting is significantly more complex. While in binary prediction we were able to decompose proper losses in terms of a one-dimensional family, Kleinberg et al. [2023] showed that the space of proper losses in the multiclass setting is fundamentally more complex, and it is impossible to construct a finite-dimensional family of losses that produce a similar decomposition. Determining whether efficient omniprediction algorithms exist in this setting is an interesting open problem for future work.

Acknowledgments

This work was supported by the Office of Naval Research, ONR grant N00014-20-1-2787. The authors thank Sivaraman Balakrishnan for helpful discussions.

References

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19 (3):357–367, 1967.

Avrim Blum, Nika Haghtalab, Ariel D. Procaccia, and Mingda Qiao. Collaborative PAC learning. In *Advances in Neural Information Processing Systems*, 2017.

Samuel Deng and Daniel Hsu. Multi-group learning for hierarchical groups. In *Proceedings of the International Conference on Machine Learning*, 2024.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses for latent covariate mixtures. *Operations Research*, 71(2):649–664, 2023.

Cynthia Dwork, Chris Hays, Nicole Immorlica, Juan C. Perdomo, and Pranay Tankala. From fairness to infinity: Outcome-indistinguishable (omni)prediction in evolving graphs. *arXiv preprint*, 2024. arXiv:2411.17582.

Werner Ehm, Tilmann Gneiting, Alexander Jordan, and Fabian Krüger. Of quantiles and expectiles: Consistent scoring functions, choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B*, 78(3):505–562, 2016.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Sumegha Garg, Christopher Jung, Omer Reingold, and Aaron Roth. Oracle efficient online multicalibration and omniprediction. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2024.

Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, and Jessica Sorrell. Multicalibration as boosting for regression. In *Proceedings of the International Conference on Machine Learning*, 2023.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

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

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *Innovations in Theoretical Computer Science Conference*, 2022.

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 Conference*, 2023a.

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

Parikshit Gopalan, Princewill Okoroafor, Prasad Raghavendra, Abhishek Sherry, and Mihir Singhal. Omnipredictors for regression and the approximate rank of convex functions. In *Proceedings of the Conference on Learning Theory*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, 2017.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

Elad Hazan. Introduction to online convex optimization. *arXiv preprint*, 2019. arXiv:1909.05207.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the International Conference on Machine Learning*, 2018.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Michael P. Kim and Juan C. Perdomo. Making decisions under outcome performativity. In *Innovations in Theoretical Computer Science Conference*, 2023.

Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

Robert Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. In *Proceedings of the Conference on Learning Theory*, 2023.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

Jiuyao Lu, Aaron Roth, and Mirah Shi. Sample efficient omniprediction and downstream swap regret for non-linear losses. *arXiv preprint*, 2025. arXiv:2502.12564.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364, 2022.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems*, 2008.

Pascal Massart. *Concentration Inequalities and Model Selection*. Springer, 2007.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the International Conference on Machine Learning*, 2019.

Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional prediction for sequential decision making. In *Proceedings of the International Conference on Machine Learning*, 2025.

Princewill Okoroafor, Robert Kleinberg, and Michael P. Kim. Near-optimal algorithms for omniprediction. *arXiv preprint*, 2025. arXiv:2501.17205.

Guy N. Rothblum and Gal Yona. Multi-group agnostic PAC learnability. In *Proceedings of the International Conference on Machine Learning*, 2021.

Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Ingo Steinwart, Chloé Pasin, Robert Williamson, and Siyu Zhang. Elicitation and identification of properties. In *Proceedings of the Conference on Learning Theory*, 2014.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10 (4):1040–1053, 1982.

John von Neumann, Oskar Morgenstern, and Ariel Rubinstein. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Volodimir Vovk. Aggregating strategies. In *Proceedings of the Workshop on Computational Learning Theory*, 1990.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*. Springer, 1997.

A Proofs for Section 2

In this section, we prove Proposition 1, which connects omniprediction error with unrestricted competitors to L_1 estimation error.

Proof of Proposition 1. To get the upper bound, fix any bounded, proper loss $\ell \in \mathcal{L}_0$. Then,

$$\begin{aligned}
\mathbb{E}[\ell(p(X),Y)] - \mathbb{E}[\ell(p^*(X),Y)] &= \mathbb{E}[\ell(p(X),Y) - \ell(p^*(X),Y)] - \mathbb{E}_{Y'|X\sim p(X)}[\ell(p(X),Y') - \ell(p^*(X),Y')] \\
&\quad + \mathbb{E}_{Y'|X\sim p(X)}[\ell(p(X),Y') - \ell(p^*(X),Y')] \\
&\le \mathbb{E}[\ell(p(X),Y) - \ell(p^*(X),Y)] - \mathbb{E}_{Y'|X\sim p(X)}[\ell(p(X),Y') - \ell(p^*(X),Y')] \\
&= \mathbb{E}[(p^*(X) - p(X))(\ell(p(X),1) - \ell(p(X),0) - \ell(p^*(X),1) + \ell(p^*(X),0))] \\
&\le 2\mathbb{E}[|p(X) - p^*(X)|],
\end{aligned}$$

where the first inequality uses the fact that ℓ is proper to bound the second term by 0.

For the lower bound, let $m \in \mathbb{N}$ be an positive integer to be specified shortly. Then,

$$\begin{aligned}
\mathbb{E}[|p(X) - p^*(X)|] &= 2\mathbb{E}\Big[\Big|p^*(X) - \frac{p(X)+p^*(X)}{2}\Big|\Big] \\
&\le \frac{2}{m} + 2\mathbb{E}\Big[\Big|p^*(X) - \frac{p(X)+p^*(X)}{2}\Big|\mathbb{1}\Big\{|p^*(X) - p(X)| > \frac{2}{m}\Big\}\Big] \\
&= \frac{2}{m} + \sum_{i=0}^{m} 2\mathbb{E}\Big[\Big|p^*(X) - \frac{p(X)+p^*(X)}{2}\Big|\mathbb{1}\Big\{|p^*(X) - p(X)| > \frac{2}{m}\Big\}\mathbb{1}\Big\{\Big\lfloor m\frac{p(X)+p^*(X)}{2}\Big\rfloor = i\Big\}\Big] \\
&\le \frac{4}{m} + \sum_{i=0}^{m} 2\mathbb{E}\Big[\Big|p^*(X) - \frac{i}{m}\Big|\mathbb{1}\Big\{|p^*(X) - p(X)| > \frac{2}{m}\Big\}\mathbb{1}\Big\{\Big\lfloor m\frac{p(X)+p^*(X)}{2}\Big\rfloor = i\Big\}\Big] \\
&\le \frac{4}{m} + \sum_{i=0}^{m} 2\mathbb{E}\Big[\Big|p^*(X) - \frac{i}{m}\Big|\mathbb{1}\Big\{p(X) \le \frac{i}{m} < p^*(X) \text{ or } p^*(X) \le \frac{i}{m} < p(X)\Big\}\Big] \\
&= \frac{4}{m} + \sum_{i=0}^{m} 2(\mathbb{E}[\ell_{i/m}(p(X),Y)] - \mathbb{E}[\ell_{i/m}(p^*(X),Y)]),
\end{aligned}$$

where we recall that $\ell_{i/m}$ denotes the proper loss function given by

$$\ell_{i/m}(p,y) = \frac{i}{m}\mathbb{1}\Big\{p > \frac{i}{m}, y = 0\Big\} + \Big(1 - \frac{i}{m}\Big)\mathbb{1}\Big\{p \le \frac{i}{m}, y = 1\Big\}.$$

So, rearranging we find that

$$\sup_{\ell\in\mathcal{L}_0} \mathbb{E}[\ell(p(X),Y)] - \mathbb{E}[\ell(p^*(X),Y)] \ge \frac{\mathbb{E}[|p(X) - p^*(X)|]}{2(m+1)} - \frac{4}{2m(m+1)}.$$

Finally, setting $m = \lfloor 7\mathbb{E}[|p(X) - p^*(X)|]^{-1}\rfloor - 1$ gives

$$\begin{aligned}
&\frac{\mathbb{E}[|p(X) - p^*(X)|]}{m+1} - \frac{4}{m(m+1)} \\
&\qquad \ge \frac{\mathbb{E}[|p(X) - p^*(X)|]^2}{7} - \frac{4}{(7\mathbb{E}[|p(X) - p^*(X)|]^{-1} - 2)(7\mathbb{E}[|p(X) - p^*(X)|]^{-1} - 1)} \\
&\qquad \ge \frac{\mathbb{E}[|p(X) - p^*(X)|]^2}{7} - \frac{4\mathbb{E}[|p(X) - p^*(X)|]^2}{30} \\
&\qquad = \frac{\mathbb{E}[|p(X) - p^*(X)|]^2}{105},
\end{aligned}$$

where to get the second inequality we have used the fact that $\mathbb{E}[|p(X) - p^*(X)|] \le 1$. □

B Proofs for Section 3

In this section, we prove Propositions 2, 3 and 4 which give lower and upper bounds on the minimax rate of calibrated multiaccuracy. We begin with the lower bound for calibrated multiaccuracy in Proposition 2.

Proof of Proposition 2. We will prove this result using Fano's method [Yu, 1997]. Let $k \in \mathbb{N}$ be a large value that we will specify shortly and set X_i to be uniformly distributed on $\{\frac{1}{k}, \frac{2}{k}, \dots, 1\}$. By the Varshamov–Gilbert lemma (e.g., see Lemma 4.7 of Massart [2007]), we know there is a collection of vectors $V \subseteq \{0,1\}^k$ such that $|V| \geq \exp(k/4)$ and for all $v, v' \in V$ with $v \neq v'$, $\|v - v'\|_1 \geq k/8$. Our goal will be to apply Fano's inequality to the set of distributions given by $p^*(X) = p_v(X) = \frac{1}{4} + \frac{X}{2} + \delta v_X$ for $v \in V$ and some appropriately small value $\delta > 0$. The idea here is that in order to be multiaccurate the predictor $\hat{p}(X)$ must correctly capture the linear component of $p_v(X)$ present in the term $\frac{X}{2}$. Then, the only way for $\hat{p}(X)$ to additionally be calibrated is if it accurately determines the value of v_x for most $x \in \{\frac{1}{k}, \frac{2}{k}, \dots, 1\}$. This latter problem is difficult and it suffers a worst-case estimation rate of $\Omega(n^{-2/5})$.

To formalize this, we first lower bound the ability of the predictor to hedge between two sign vectors. In particular, fix $v, v' \in V$ with $v \neq v'$. Then, we will lower bound

$$\inf_p \max_{p^* \in \{p_v, p_{v'}\}} \max\left\{\mathbb{E}_{p^*}[X(Y - p(X))], \mathbb{E}\left[\left|p(X) - \mathbb{E}[p^*(X) \mid p(X)]\right|\right]\right\},$$

where the infimum is taken over all functions $p : \{\frac{1}{k}, \frac{2}{k}, \dots, 1\} \to [0,1]$ and the notation $\mathbb{E}_{p^*}$ is used to denote the distribution in which $X \sim \text{Unif}(\{\frac{1}{k}, \frac{2}{k}, \dots, 1\})$ and $Y \mid X \sim \text{Ber}(p^*(X))$.

Fix any $p : \{\frac{1}{k}, \frac{2}{k}, \dots, 1\} \to [0,1]$. Let $p_1, \dots, p_r$ denote the distinct values in the support of $p(X)$ and for $i \in \{1, \dots, r\}$ let $G_i = \{x \in \{\frac{1}{k}, \frac{2}{k}, \dots, 1\} : p(x) = p_i\}$. For ease of notation, define

$$\begin{aligned}
\text{ECE}_{\max}(p; v, v') &= \max_{p^* \in \{p_v, p_{v'}\}} \left|\mathbb{E}\left[\left|p(X) - \mathbb{E}[p^*(X) \mid p(X)]\right|\right]\right| \\
&= \max_{\tilde{v} \in \{v, v'\}} \sum_{i=1}^r \frac{|G_i|}{k} \left| \frac{1}{|G_i|} \sum_{x \in G_i} \frac{1}{4} + \frac{x}{2} + \delta \tilde{v}_x - p_i \right|,
\end{aligned}$$

as the maximum calibration error, and note that

$$\begin{aligned}
\sum_{i=1}^r \frac{|G_i|}{k} \left| \frac{1}{|G_i|} \sum_{x \in G_i} (v_x - v'_x) \right| &\leq \frac{1}{\delta} \sum_{i=1}^r \frac{|G_i|}{k} \left(\left| \frac{1}{|G_i|} \sum_{x \in G_i} \frac{1}{4} + \frac{x}{2} + \delta v_x - p_i \right| + \left| \frac{1}{|G_i|} \sum_{x \in G_i} \frac{1}{4} + \frac{x}{2} + \delta v'_x - p_i \right| \right) \\
&\leq \frac{2\text{ECE}_{\max}(p; v, v')}{\delta}.
\end{aligned}$$

In particular, applying this bound alongside our assumptions on V gives

$$\begin{aligned}
\frac{k}{8} \leq \|v - v'\|_1 &\leq \sum_{i=1}^r \mathbb{1}\{G_i = 1\} |v_i - v'_i| + \sum_{i=1}^r |G_i| \mathbb{1}\{G_i > 1\} \\
&\leq \sum_{i=1}^r |G_i| \left| \frac{1}{|G_i|} \sum_{x \in G_i} (v_x - v'_x) \right| + \sum_{i=1}^r |G_i| \mathbb{1}\{G_i > 1\} \\
&\leq \frac{2\text{ECE}_{\max}(p; v, v')}{\delta} + \sum_{i=1}^r |G_i| \mathbb{1}\{G_i > 1\},
\end{aligned}$$

and rearranging we have that

$$\sum_{i=1}^r |G_i| \mathbb{1}\{G_i > 1\} \geq \frac{k}{8} - \frac{2\text{ECE}_{\max}(p; v, v')}{\delta}.$$

On the other hand, by considering the multiaccuracy error with $g(x) = x$ we find that

$$\begin{aligned}
&\mathbb{E}_{p_v}[X(Y - p(X))] \\
&\quad \geq \sum_{i=1}^{r} \frac{|G_i|}{k} \left(\frac{1}{|G_i|} \sum_{x \in G_i} x\left(\frac{1}{4} + \frac{x}{2} + \delta v_x\right) - \frac{1}{|G_i|} \sum_{x \in G_i} x\left(\frac{1}{4} + \frac{1}{|G_i|} \sum_{x \in G_i} \frac{x}{2} + \delta v_x\right) \right) - \mathrm{ECE}_{\max}(p; v, v') \\
&\quad \geq \sum_{i=1}^{r} \frac{|G_i|}{2k} \left(\frac{1}{|G_i|} \sum_{x \in G_i} x^2 - \left(\frac{1}{|G_i|} \sum_{x \in G_i} x \right)^2 \right) - \delta - \mathrm{ECE}_{\max}(p; v, v') \\
&\quad = \sum_{i=1}^{r} \frac{|G_i|}{4k} \frac{1}{|G_i|^2} \sum_{x, x' \in G_i} (x - x')^2 - \delta - \mathrm{ECE}_{\max}(p; v, v') \\
&\quad \geq \sum_{i=1}^{r} \frac{|G_i|}{4k} \Big(1 - \frac{1}{|G_i|}\Big) \frac{1}{k^2} - \delta - \mathrm{ECE}_{\max}(p; v, v') \\
&\quad \geq \frac{1}{8k^3} \sum_{i=1}^{r} |G_i| \mathbb{1}\{|G_i| > 1\} - \delta - \mathrm{ECE}_{\max}(p; v, v') \\
&\quad \geq \frac{1}{8k^3} \Big(\frac{k}{8} - \frac{2\mathrm{ECE}_{\max}(p; v, v')}{\delta} \Big) - \delta - \mathrm{ECE}_{\max}(p; v, v'),
\end{aligned}$$

and rearranging the first and last inequalities gives

$$\mathbb{E}_{p_v}[X(Y - p(X))] + \mathrm{ECE}_{\max}(p; v, v') + \frac{1}{4k^3} \frac{\mathrm{ECE}_{\max}(p; v, v')}{\delta} \geq \frac{1}{64k^2} - \delta.$$

Finally, setting $\delta = 1/(128k^2)$ we find that

$$\inf_{p} \max_{p^* \in \{p_v, p_{v'}\}} \max\big\{ \mathbb{E}_{p^*}[X(Y - p(X))], \mathbb{E}\big[\big|p(X) - \mathbb{E}[p^*(X) \mid p(X)]\big|\big] \big\} \geq \frac{k}{k + 32} \frac{1}{128k^2}.$$

With this inequality in hand, the proof of our desired result follows from a straightforward application of Fano's inequality (e.g., as stated in Lemma 3 of Yu [1997]). Let $\hat{p} : \mathcal{X} \to [0, 1]$ be an arbitrary estimator, and define an associated classifier by

$$\hat{v} \in \operatorname*{argmin}_{v \in V} \max\big\{ |\mathbb{E}_{p_v}[X(Y - \hat{p}(X))]|, \mathbb{E}\big[\big|p(X) - \mathbb{E}[p_v(X) \mid \hat{p}(X)]\big|\big] \big\},$$

where both here and in what follows the expectations are taken with respect to (X, Y), with the estimator $\hat{p}$ (which is a random function of the training data) held fixed. By our previous calculations, we have that for any $v^* \in V$,

$$\max\big\{ \mathbb{E}_{p_{v^*}}[X(Y - \hat{p}(X))], \mathbb{E}\big[\big|p(X) - \mathbb{E}[p_{v^*}(X) \mid \hat{p}(X)]\big|\big] \big\} \geq \frac{k}{k + 32} \frac{1}{128k^2} \mathbb{1}\{\hat{v} \neq v^*\},$$

and thus,

$$\begin{aligned}
&\max_{v^* \in V} \mathbb{E}_{\{(X_i, Y_i)\}_{i=1}^n \overset{\text{iid}}{\sim} p_{v^*}} \Big[\max\big\{ \mathbb{E}_{p_{v^*}}[X(Y - \hat{p}(X))], \mathbb{E}\big[\big|\hat{p}(X) - \mathbb{E}[p_{v^*}(X) \mid \hat{p}(X)]\big|\big] \big\} \Big] \\
&\quad \geq \mathbb{E}_{v^* \sim \mathrm{Unif}(V)} \mathbb{E}_{\{(X_i, Y_i)\}_{i=1}^n \overset{\text{iid}}{\sim} p_{v^*}} \Big[\max\big\{ \mathbb{E}_{p_{v^*}}[X(Y - \hat{p}(X))], \mathbb{E}\big[\big|\hat{p}(X) - \mathbb{E}[p_{v^*}(X) \mid \hat{p}(X)]\big|\big] \big\} \Big] \\
&\quad \geq \frac{k}{k + 16} \frac{1}{128k^2} \mathbb{P}_{v^* \sim \mathrm{Unif}(V), \{(X_i, Y_i)\}_{i=1}^n \overset{\text{iid}}{\sim} p_v} (\hat{v} \neq v^*) \\
&\quad \geq \frac{k}{k + 16} \frac{1}{128k^2} \left(1 - \frac{\frac{1}{|V|^2} \sum_{v, v' \in V} n D_{\mathrm{KL}}(p_v || p_{v'}) + \log(2)}{\log(|V|)} \right),
\end{aligned}$$

where $D_{\mathrm{KL}}(p_v||p_{v'})$ denotes the KL-divergence between the distribution of (X,Y) under p_v and $p_{v'}$. Now by a direct calculation,

$$\begin{aligned}
D_{\mathrm{KL}}(p_v||p_{v'}) &= \mathbb{E}_X[p_v(X)\log\left(\frac{p_v(X)}{p_{v'}(X)}\right) + (1-p_v(X))\log\left(\frac{1-p_v(X)}{1-p_{v'}(X)}\right)\Big] \\
&\leq \mathbb{E}_X\left[p_v(X)\left(\frac{p_v(X)}{p_{v'}(X)}-1\right) + (1-p_v(X))\left(\frac{1-p_v(X)}{1-p_{v'}(X)}-1\right)\right] \\
&= \mathbb{E}_X\left[\frac{(p_v(X)-p_{v'})^2}{p_{v'}(X)(1-p_{v'}(X))}\right] \\
&\leq \frac{64}{7}\delta^2,
\end{aligned}$$

where the last inequality holds for $\delta \leq 1/8$. Plugging this into the previous expression gives a lower bound of

$$\frac{k}{k+32}\frac{1}{128k^2}\left(1-\frac{n\frac{64}{7}\delta^2+\log(2)}{k/4}\right) = \frac{k}{k+32}\frac{1}{128k^2}\left(1-\frac{n\frac{64}{7}128^{-2}k^{-4}+\log(2)}{k/4}\right).$$

The desired result follows immediately by taking $k = C\lceil n^{-1/5}\rceil$ for an appropriately chosen constant C. □

We next give a proof of our lower bound for multiaccuracy given in Proposition 3.

Proof of Proposition 3. Abbreviate $d = \mathrm{VC}(\mathcal{G})$. Once again, we will use Fano's method. By definition of the VC dimension, we may find $x_1,\ldots,x_d$ such that for all $v \in \{-1,1\}^d$ there exists $g_v \in \mathcal{G}$ with $g_v(x_i) = v_i$ for all $i \in \{1,\ldots,d\}$. Consider distributions on (X,Y) given by $X \sim \mathrm{Unif}(x_1,\ldots,x_d)$ and $Y \mid X \sim \mathrm{Ber}(\frac{1+\delta g_v(X)}{2})$ for some small $\delta > 0$ that we will specify shortly. Denote the expectation over this distribution by $\mathbb{E}_v$. Let V be as in the proof of Proposition 2. For any $v \neq v'$ with $v, v' \in V$ and $p : \{x_1,\ldots,x_d\} \to [0,1]$ we have that

$$\begin{aligned}
&\max_{v^* \in \{v,v'\}} \sup_{g\in\mathcal{G}} \mathbb{E}_{v^*}[g(X)(Y-p(X))] \\
&\quad= \max_{v^* \in \{v,v'\}} \sup_{g\in\mathcal{G}} \mathbb{E}_X\left[g(X)\left(\frac{1+\delta g_{v^*}(X)}{2}-p(X)\right)\right] \\
&\quad\geq \frac{1}{2}\sup_{g\in\mathcal{G}} \mathbb{E}_X\left[g(X)\left(\frac{1+\delta g_v(X)}{2}-p(X)\right)\right] + \frac{1}{2}\sup_{g\in\mathcal{G}} \mathbb{E}_X\left[g(X)\left(\frac{1+\delta g_{v'}(X)}{2}-p(X)\right)\right] \\
&\quad\geq \frac{1}{2}\sup_{g\in\mathcal{G}} \mathbb{E}_X\left[g(X)\left(\frac{1+\delta g_v(X)}{2}-\frac{1+\delta g_{v'}(X)}{2}\right)\right] \\
&\quad= \frac{\delta}{4}\mathbb{E}_X[|g_v(X)-g_{v'}(X)|] = \frac{1}{4}\delta\frac{\|v-v'\|_1}{d} \geq \frac{\delta}{32}.
\end{aligned}$$

Proceeding as in the proof of Proposition 2, we obtain the lower bound,

$$\min_{\hat{p}} \sup_{P_{XY}} \sup_{g\in\mathcal{G}} \mathbb{E}\Big[g(X)(Y-\hat{p}(X))\Big] \geq \frac{\delta}{32}\left(1-\frac{n\frac{64}{7}\delta^2+\log(2)}{d\log(2)}\right).$$

The expectation is taken over training samples $\{(X_i,Y_i)\}_{i=1}^n$ used to fit $\hat{p}$ and the test sample (X,Y), all i.i.d. from P_{XY}. Setting $\delta = C\sqrt{d/n}$ for a sufficiently small constant $C > 0$ gives the result. □

We turn to the proof of Proposition 4, and present our algorithm for obtaining calibrated multiaccuracy. This will follow a similar structure to the two-player game based algorithms for omniprediction introduced in Section 5.1. Namely, we expand the calibration and multiaccuracy criteria as a set of objectives and use a multiplicative weights algorithm to obtain useful mixtures of these targets.

Fix a hyperparameter $m \in \mathbb{N}$. We will learn a predictor that returns randomized outputs in $\{\frac{1}{m},\frac{2}{m},\ldots,1\}$. Let $\mathcal{G}_m = \{g : \{\frac{1}{m},\frac{2}{m},\ldots,1\} \to \{-1,1\}\}$ denote the set of sign functions on $\{\frac{1}{m},\frac{2}{m},\ldots,1\}$. Let Δ_m denote the

space of probability distributions on $\{\frac{1}{m}, \frac{2}{m}, \ldots, 1\}$, and note that for any randomized predictor $P : \mathcal{X} \to \Delta_m$ the expected calibration error can be written as

$$\mathbb{E}_{p(X)\sim P(X)}[|p(X) - \mathbb{E}[Y \mid p(X)]|] = \sup_{g\in\mathcal{G}_m} \mathbb{E}_{p(X)\sim P(X)}[g(p(X))(Y - p(X))].$$

Thus, to guarantee calibration it is sufficient to guarantee that our predictor gives multiaccurate predictions with respect to $\mathcal{G}_m$. Combining this with the original multiaccuracy target class $\mathcal{G}$ gives us the necessary set of objectives for a two-player game based algorithm. A formal description is given in Algorithm 4. As stated in Proposition 6, this algorithm obtains calibrated multiaccuracy at the rate $\sqrt{\log(|\mathcal{G}|)/n} + n^{-1/3}$, and this proves the claim in Proposition 4.

Algorithm 4: Two-player game based calibrated multiaccuracy

Input: training samples $\{(X_i, Y_i)\}_{i=1}^n$, finite function class $\mathcal{G}$, learning rate $\eta > 0$

$\mathcal{G}_\pm = \mathcal{G} \cup \{-g : g \in \mathcal{G}\}$;

$q_g(1) = \frac{1}{|\mathcal{G}_\pm \cup \mathcal{G}_m|}$, for all $g \in \mathcal{G}_\pm \cup \mathcal{G}_m$;

for $t = 1, \ldots, n$ **do**

  $\hat{P}_t(x) \in \operatorname{argmin}_{P\in\Delta_m} \max_{p_y\in[0,1]} \sum_{g\in\mathcal{G}_\pm} q_g(t)\mathbb{E}_{p\sim P}[g(X)(p_y - p)] + \sum_{g\in\mathcal{G}_m} q_g(t)\mathbb{E}_{p\sim P}[g(p)(p_y - p)]$, for all $x \in \mathcal{X}$;

  $\tilde{q}_g(t+1) = \tilde{q}_g(t)\exp(\eta\mathbb{E}_{p\sim\hat{P}_t(X_t)}[g(X_t)(Y_t - p)])$, for all $g \in \mathcal{G}_\pm$;

  $\tilde{q}_g(t+1) = \tilde{q}_g(t)\exp(\eta\mathbb{E}_{p\sim\hat{P}_t(X_t)}[g(p)(Y_t - p)])$, for all $g \in \mathcal{G}_m$;

  $q_g(t+1) = \frac{\tilde{q}_g(t+1)}{\sum_{g'\in\mathcal{G}_\pm\cup\mathcal{G}_m}\tilde{q}_{g'}(t+1)}$, for all $g \in \mathcal{G}_\pm \cup \mathcal{G}_m$;

return $\hat{P} = \frac{1}{n}\sum_{t=1}^n \hat{P}_t$

Proposition 6. *Setting $m = \lceil n^{1/3}\rceil$ and $\eta = \sqrt{(\log(|\mathcal{G}|) + m)/n}$, Algorithm 4 produces a distribution $\hat{P}(X)$ such that the randomized predictor $\hat{p}(X) \sim \hat{P}(X)$ has calibrated multiaccuracy error*

$$\max\{\mathrm{MA}(\hat{p}; \mathcal{G}), \mathrm{ECE}(\hat{p})\} \le O_{\mathbb{P}}\left(\sqrt{\frac{\log(|\mathcal{G}|)}{n}} + \frac{1}{n^{1/3}}\right).$$

Proof. In what follows, all expectations are taken with respect to (X, Y) and a random draw from $\hat{P}$ (or its constituents $\hat{P}_t$). In particular, the training samples are treated as fixed. Fix any $g \in \mathcal{G}$. By definition,

$$\mathbb{E}[g(X)(Y - \hat{p}(X))] = \frac{1}{n}\sum_{t=1}^n \mathbb{E}_{\hat{p}(X)\sim\hat{P}_t(X)}[g(X)(Y - \hat{p}(X))].$$

Now, by the Azuma-Hoeffding inequality (Theorem 5 below) we may guarantee that for any $c > 0$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{t=1}^n \mathbb{E}_{\hat{p}(X)\sim\hat{P}_t(X)}[g(X)(Y - \hat{p}(X))] - \frac{1}{n}\sum_{t=1}^n \mathbb{E}_{p\sim\hat{P}_t(X_t)}[g(X_t)(Y_t - p)]\right| \ge c\sqrt{\frac{\log(|\mathcal{G}|)}{n}}\right) \le 2\exp\left(-\frac{c^2}{8}\right).$$

Applying this to the previous expression, we find that

$$\mathbb{E}[g(X)(Y - \hat{p}(X))] \le \frac{1}{n}\sum_{t=1}^n \mathbb{E}_{p\sim\hat{P}_t(X_t)}[g(X_t)(Y_t - p)] + O_{\mathbb{P}}\left(\sqrt{\frac{\log(|\mathcal{G}|)}{n}}\right).$$

The updates for q_g given in Algorithm 4 are exactly the updates for the hedge method [Vovk, 1990, Littlestone and Warmuth, 1994, Freund and Schapire, 1997]. By known regret bounds for this algorithm (see Theorem 6 below), we have the inequality

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{p\sim\hat{P}_t(X_t)}[g(X_t)(Y_t-p)] \le \frac{1}{n}\sum_{t=1}^{n}\sum_{g'\in\mathcal{G}_\pm\cup\mathcal{G}_m} q_{g'}(t)\mathbb{E}_{p\sim\hat{P}_t(X_t)}[g'(X_t)(Y_t-p)] + O\left(\sqrt{\frac{\log(|\mathcal{G}|)+m}{n}}\right).$$

Finally, by definition of $\hat{P}_i(X_i)$ and von Neumann's minimax theorem [von Neumann et al., 1944],

$$\begin{aligned}
&\sum_{t=1}^{n}\sum_{g'\in\mathcal{G}_\pm\cup\mathcal{G}_m} q_{g'}(t)\mathbb{E}_{p\sim\hat{P}_t(X_t)}[g'(X_t)(Y_t-p)] \\
&\qquad\le \min_{P\in\Delta_m}\max_{p_y\in[0,1]}\sum_{g'\in\mathcal{G}_\pm} q_{g'}(t)\mathbb{E}_{p\sim P}[g'(X_t)(p_y-p)] + \sum_{g'\in\mathcal{G}_m} q_{g'}(t)\mathbb{E}_{p\sim P}[g'(p)(p_y-p)] \\
&\qquad= \max_{p_y\in[0,1]}\min_{P\in\Delta_m}\sum_{g'\in\mathcal{G}_\pm} q_{g'}(t)\mathbb{E}_{p\sim P}[g'(X_t)(p_y-p)] + \sum_{g'\in\mathcal{G}_m} q_{g'}(t)\mathbb{E}_{p\sim P}[g'(p)(p_y-p)] \le \frac{1}{m},
\end{aligned}$$

where to get the last inequality one may simply set P to give probability one to the element of $\{\frac{1}{m},\frac{2}{m},\dots,1\}$ that is closest to p_y. Combining all of the previous steps, we arrive at the bound

$$\sup_{g\in\mathcal{G}}\mathbb{E}_{\hat{p}(X)\sim\hat{P}(X)}[g(X)(Y-\hat{p}(X))] \le O_{\mathbb{P}}\left(\sqrt{\frac{\log(|\mathcal{G}|)}{n}}\right) + O\left(\sqrt{\frac{\log(|\mathcal{G}|)+m}{n}}\right) + \frac{1}{m} = O_{\mathbb{P}}\left(\sqrt{\frac{\log(|\mathcal{G}|)}{n}} + \frac{1}{n^{1/3}}\right),$$

by our choice of $m = \lceil n^{1/3}\rceil$. A bound on the multiaccuracy follows by applying the same argument to $-g$. Finally, a bound on the expected calibration error follows by applying the preceding argument to $\mathcal{G}_m$. □

C Extensions of Theorem 2 beyond left-continuity

While we will not pursue this in detail, it should be possible to extend the decomposition result in Theorem 2 beyond left-continuous losses. To motivate this, let us first consider the discontinuity point of ℓ_θ. From (4.1), we see that when the true underlying probability is equal to θ all predictions have the same expected loss. As a result, one can modify the value of the loss substantially at $p=\theta$ without affecting its propriety. Indeed, one can verify with some additional calculation that the family of losses

$$\ell_{\theta,\beta} = \begin{cases} \theta, & \text{if } p>\theta \text{ and } y=0, \\ 1-\theta, & \text{if } p<\theta \text{ and } y=1, \\ \theta(1-\theta)+\beta(y-\theta), & \text{if } p=\theta, \end{cases}$$

is proper for all $\theta\in[0,1]$ and $\beta\in[-\theta,1-\theta]$. By varying the second parameter β, one can encode a variety of jump discontinuities in $\ell_{\theta,\beta}$. While not a complete proof, the calculations in Kleinberg et al. [2023] suggest that these jumps are sufficient to capture all possible discontinuities in proper losses and thus enable an extension of Theorem 2 to a decomposition of arbitrary proper losses in terms of mixtures over the two-parameter class $\{\ell_{\theta,\beta}:\theta\in[0,1],\beta\in[-\theta,1-\theta]\}$. We do not believe that this extra layer of complexity has a large impact on practical results for omniprediction and hence we have chosen to omit these details and restrict ourselves to left-continuous losses.

D Proofs for Section 4

In this section, we prove Lemma 1, which bounds the discretization error for omniprediction with respect to weighted 0-1 losses.

Proof of Lemma 1. Fix any $\theta \in [0,1]$ and $\epsilon > 0$. Let $f_{\theta,\epsilon}$ be such that

$$\sup_{f \in \mathcal{F}} \mathbb{E}[\ell_\theta(p(X), Y)] - \mathbb{E}[\ell_\theta(f(X), Y)] \leq \mathbb{E}[\ell_\theta(p(X), Y) - \mathbb{E}[\ell_\theta(f_{\theta,\epsilon}(X), Y)] + \epsilon.$$

Let θ_i denote the value on the grid $\{\frac{i}{m} - \frac{1}{2m} : i \in \{1, \ldots, m\}\}$ closest to θ, subject to the extra specification that in the case of ties we always round up. By our assumption of the support of $p(X)$ we have that

$$|\mathbb{E}[\ell_\theta(p(X), Y) - \ell_{\theta_i}(p(X), Y)]| = |\mathbb{E}[(\theta - \theta_i)\mathbb{1}\{Y = 0, p(X) > \theta\} + (\theta_i - \theta)\mathbb{1}\{Y = 1, p(X) \leq \theta\}]| \leq \frac{1}{2m}.$$

Similarly, we also have

$$\begin{aligned}&|\mathbb{E}[\ell_\theta(f_{\theta,\epsilon}(X), Y) - \ell_{\theta_i}(f_{\theta,\epsilon}(X) - \theta + \theta_i, Y)]| \\ &\qquad = |\mathbb{E}[(\theta - \theta_i)\mathbb{1}\{Y = 0, f_{\theta,\epsilon}(X) > \theta\} + (\theta_i - \theta)\mathbb{1}\{Y = 1, f_{\theta,\epsilon}(X) \leq \theta\}]| \leq \frac{1}{2m}.\end{aligned}$$

So, putting these two facts together we find that

$$\mathbb{E}[\ell_\theta(p(X), Y) - \ell_\theta(f_{\theta,\epsilon}(X), Y)] \leq \sup_{f \in \mathcal{F}} \mathbb{E}[\ell_{\theta_i}(p(X), Y) - \ell_{\theta_i}(f(X), Y)] + \frac{1}{m},$$

and sending $\epsilon \to 0$ gives the desired result. □

E Proofs for Section 5

E.1 Proofs for Section 5.1

In this section, we prove Lemma 3 and Theorem 3, which provide our theory for the two-player game based omniprediction method. We begin with the explicit solution to the min-max game in Lemma 3.

Proof of Lemma 3. Consider the distribution $P^* = (1 - \rho^*)\delta_{\theta^*} + \rho^*\delta_{\theta^*+i/m}$, where θ^* and ρ^* are as defined in the lemma. For ease of notation, let $q_{m+1} = 1$, and define $p_y^* = \min\{\theta^* + \frac{1}{2m}, 1\}$. To prove P^* is optimal it is sufficient to prove that the pair (P^*, p_y^*) is a saddle point of the min-max program. To see this, observe that for any (P, p_y) the optimization objective can be written as

$$\begin{aligned}O(P, p_y) &:= \mathbb{E}_{p \sim P, Y' \sim \mathrm{Ber}(p_y)}\left[\sum_{i=1}^m q_i(\ell_{\theta_i}(p, Y') - \ell_{\theta_i}(\hat{f}_{\theta_i}(x), Y'))\right] \\ &= \mathbb{E}_{p \sim P, Y' \sim \mathrm{Ber}(p_y)}\Bigg[\sum_{i=1}^m q_i\Big(\theta_i\mathbb{1}\{p > \theta_i, Y' = 0\} + (1 - \theta_i)\mathbb{1}\{p \leq \theta_i, Y' = 1\} \\ &\qquad - \theta_i\mathbb{1}\{\hat{f}_{\theta_i}(x) > \theta_i, Y' = 0\} - (1 - \theta_i)\mathbb{1}\{\hat{f}_{\theta_i}(x) \leq \theta_i, Y' = 1\}\Big)\Bigg] \\ &= \mathbb{E}_{p \sim P}\Bigg[\sum_{i=1}^m q_i\Big(\theta_i(1 - p_y)\mathbb{1}\{p > \theta_i\} + (1 - \theta_i)p_y\mathbb{1}\{p \leq \theta_i\} \\ &\qquad - \theta_i(1 - p_y)\mathbb{1}\{\hat{f}_{\theta_i}(x) > \theta_i\} - (1 - \theta_i)p_y\mathbb{1}\{\hat{f}_{\theta_i}(x) \leq \theta_i\}\Big)\Bigg] \\ &= \mathbb{E}_{p \sim P}\left[\sum_{i=1}^m q_i(p_y - \theta_i)(\mathbb{1}\{p \leq \theta_i\} - \mathbb{1}\{\hat{f}_{\theta_i}(x) \leq \theta_i\})\right].\end{aligned}$$

Now, plugging in our choice of P^* gives an objective value of

$$\begin{aligned} O(P^*, p_y) &= \sum_{i=1}^{m} q_i p_y (\mathbb{1}\{\theta^* \le \theta_i\} - \mathbb{1}\{\hat{f}_{\theta_i}(x) \le \theta_i\}) - \rho^* p_y q_{m\theta^*+1} \\ &\qquad - \mathbb{E}_{p\sim P}\Big[\sum_{i=1}^{m} q_i \theta_i (\mathbb{1}\{p \le \theta_i\} - \mathbb{1}\{\hat{f}_{\theta_i}(x) \le \theta_i\})\Big] \\ &= -\mathbb{E}_{p\sim P}\Big[\sum_{i=1}^{m} q_i \theta_i (\mathbb{1}\{p \le \theta_i\} - \mathbb{1}\{\hat{f}_{\theta_i}(x) \le \theta_i\})\Big], \end{aligned}$$

where the second equality follows from our choice of ρ^*. Since this last expression does not depend on p_y, we must have that $O(P^*, p_y^*) = \max_{p_y \in [0,1]} O(P^*, p_y)$.

On the other hand, because the losses $\{\ell_{\theta_i}\}_{i=1}^m$ are proper we know that at $p_y = p_y^*$, the objective $O(P, p_y^*)$ is minimized by setting $P = \delta_{p_y^*}$. Moreover, it is easy to check that for all $i \in \{1, \dots, m\}$,

$$\mathbb{E}_{Y'\sim \mathrm{Ber}(p_y^*)}[\ell_{\theta_i}(p_y^*, Y')] = \mathbb{E}_{Y'\sim \mathrm{Ber}(p_y^*)}[\ell_{\theta_i}(\theta^*, Y')] = \mathbb{E}_{Y'\sim \mathrm{Ber}(p_y^*)}[\ell_{\theta_i}(\theta^* + 1/m, Y')].$$

In particular, this implies that $O(P^*, p_y^*) = O(\delta_{p_y^*}, p_y^*)$, hence $O(P^*, p_y^*) = \min_{P\in\Delta_m} O(P, p_y^*)$, as desired. □

Proof of Theorem 3. In what follows, all expectations are taken with respect to (X, Y) and a random draw from $\hat{P}$ (or its constituents $\hat{P}_t$). In particular, the training samples are treated as fixed throughout. By the results of Section 4, it is sufficient to bound (4.3). Fix any $i \in \{1, \dots, m\}$. By definition of $\hat{P}$, we have that

$$\mathbb{E}_{\hat{p}(X)\sim\hat{P}(X)}\big[\ell_{\theta_i}(\hat{p}(X), Y) - \ell_{\theta_i}(\hat{f}_{\theta_i}(X), Y)\big] = \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{\hat{p}(X)\sim\hat{P}_t(X)}\big[\ell_{\theta_i}(\hat{p}(X), Y) - \ell_{\theta_i}(\hat{f}_{\theta_i}(X), Y)\big].$$

Now, consider the martingale

$$M_n(i) = \sum_{t=1}^{n} \Big(\mathbb{E}_{p\sim\hat{P}_t(X_t)}\big[\ell_{\theta_i}(p, Y_t) - \ell_{\theta_i}(\hat{f}_{\theta_i}(X_t), Y_t)\big] - \mathbb{E}_{\hat{p}(X)\sim\hat{P}_t(X)}\big[\ell_{\theta_i}(\hat{p}(X), Y) - \ell_{\theta_i}(\hat{f}_{\theta_i}(X), Y)\big]\Big).$$

By the Azuma-Hoeffding inequality (Theorem 5 below),

$$\max_{i\in\{1,\dots,m\}} |M_n(i)|/n \le O_{\mathbb{P}}(\sqrt{\log(m)/n}),$$

and so, in particular,

$$\mathbb{E}_{\hat{p}(X)\sim\hat{P}(X)}\big[\ell_{\theta_i}(\hat{p}(X), Y) - \ell_{\theta_i}(\hat{f}_{\theta_i}(X), Y)\big] \le \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{p\sim\hat{P}_t(X_t)}\big[\ell_{\theta_i}(p, Y_t) - \ell_{\theta_i}(\hat{f}_{\theta_i}(X_t), Y_t)\big] + O_{\mathbb{P}}\left(\sqrt{\frac{\log(m)}{n}}\right).$$

By regret bounds for the hedge algorithm (Theorem 6 below) the first term above is bounded by

$$\frac{1}{n}\sum_{t=1}^{n}\sum_{j=1}^{m} q_j(s)\mathbb{E}_{p\sim\hat{P}_t(X_t)}\big[\ell_{\theta_j}(p, Y_t) - \ell_{\theta_j}(\hat{f}_{\theta_j}(X_t), Y_t)\big]) + 4\eta + \frac{\log(m)}{n\eta}.$$

By Lemma 2 we know that the first term above is nonpositive. Putting the above inequalities together,

$$\mathbb{E}_{\hat{p}(X)\sim\hat{P}(X)}\big[\ell_{\theta_i}(\hat{p}(X), Y) - \ell_{\theta_i}(\hat{f}_{\theta_i}(X), Y)\big] \le O_{\mathbb{P}}\left(\sqrt{\frac{\log(m)}{n}}\right) + \eta + \frac{\log(m)}{n\eta},$$

and plugging in our choices of η and m gives the desired result. □

E.2 Proofs for Section 5.2

In this section, we prove Lemma 5 and Theorem 4. We begin by stating a more detailed version of our merge algorithm which defines a number of additional quantities that will be useful in the proof. Most crucially, we use $A_{h,t}$ and $A_{l,t}$ (evolving over iterations t) to denote the sets for which $\hat{p}_m(x) = \hat{p}_h(x)$ and $\hat{p}_m(x) = \hat{p}_l(x)$, respectively. We also use $\{\theta^s_{h,0}, \ldots, \theta^s_{h,k_h}\}$ and $\{\theta^s_{l,0}, \ldots, \theta^s_{l,k_l}\}$ to denote the sets where the algorithm switches direction (i.e., swaps from examining parameters in Θ_h to examining parameters in Θ_l and vice versa).

Algorithm 5: Detailed merge procedure

Input: training samples $\{(X_i, Y_i)\}_{i=1}^n$, predictors $\hat{p}_l$, $\hat{p}_h$, parameter sets Θ_l, Θ_h, hyperparameter $\epsilon \geq 0$

$\theta^s_{l,0} = \theta^s_{h,0} = \theta_{h,0} = -1$;
$\theta_{l,0} = 1$;
$k_l = k_h = 0$;
$t = 1$;
$A_{l,1} = \emptyset$;
$A_{h,1} = \mathcal{X}$;
$\theta_{l,1} = \max \Theta_l$;
$\theta_{h,1} = \min \Theta_h$;
$\mathrm{dir}(1) = \mathrm{low}$;
while $\theta_l \neq -\infty$, $\theta_h \neq \infty$ **do**
  $E = \mathbb{1}\{x : \hat{p}_h(x) > \theta_{h,t}, \hat{p}_l(x) \leq \theta_{l,t}\}$;
  if $\mathrm{dir}(t) = \mathrm{low}$ **then**
    if $\hat{\mathbb{E}}_n[(\ell_{\theta_{l,t}}(0, Y) - \ell_{\theta_{l,t}}(1, Y))\mathbb{1}\{X \in E\}] < -\epsilon$ **then**
      $A_{l,t+1} = A_{l,t} \cup E$;
      $A_{h,t+1} = A_{h,t} \setminus E$;
      $\theta_{h,t+1} = \min\{\theta \in \Theta_h : \theta > \theta_{h,t}\}$;
      $\mathrm{dir}(t+1) = \mathrm{high}$;
      $k_l = k_l + 1$;
      $\theta^s_{l,k_l} = \theta_{l,t}$;
    else
      $\theta_{l,t+1} = \max\{\theta \in \Theta_l : \theta < \theta_{l,t}\}$;
      $\mathrm{dir}(t+1) = \mathrm{low}$;
  else
    if $\hat{\mathbb{E}}_n[(\ell_{\theta_{h,t}}(1, Y) - \ell_{\theta_{h,t}}(0, Y))\mathbb{1}\{X \in E\}] < -\epsilon$ **then**
      $A_{h,t+1} = A_{h,t} \cup E$;
      $A_{l,t+1} = A_{l,t} \setminus E$;
      $\theta_{l,t+1} = \max\{\theta \in \Theta_l : \theta < \theta_{l,t}\}$;
      $\mathrm{dir}(t+1) = \mathrm{low}$;
      $k_h = k_h + 1$;
      $\theta^s_{h,k_h} = \theta_{h,t}$;
    else
      $\theta_{h,t+1} = \min\{\theta \in \Theta_h : \theta > \theta_{h,t}\}$;
      $\mathrm{dir}(t+1) = \mathrm{high}$;
  $t = t + 1$;
return $\hat{p}_m(X) = \hat{p}_l(X)\mathbb{1}\{X \in A_l\} + \hat{p}_h(X)\mathbb{1}\{X \in A_h\}$

Finally, we let $c_{h,t} = |\{s < t : \mathrm{dir}(s) = \mathrm{high}, \mathrm{dir}(s+1) = \mathrm{low}\}|$ denote the number of times the direction switches from high to low before time t, and oppositely, $c_{\ell,t} = |\{s < t : \mathrm{dir}(s) = \mathrm{low}, \mathrm{dir}(s+1) = \mathrm{high}\}|$. We will now prove Lemma 5 using a sequence of smaller results. Our first lemma characterizes the structure of

the sets $A_{h,t}$ and $A_{l,t}$.

Lemma 6. *Let Θ_h, Θ_l be finite subsets of $[0,1]$, with $\min \Theta_h > \max \Theta_l$, and let $\hat{p}_h$ and $\hat{p}_l$ be predictors which values in $[0,1]$. Then, for each time t for which $\mathrm{dir}(t) = \mathrm{high}$,*

$$\begin{aligned} A_{h,t} &= \bigcup_{i=1}^{c_{h,t}} \{x : \theta^s_{h,i-1} < \hat{p}_h(x) \le \theta^s_{h,i},\, \hat{p}_l(x) > \theta^s_{\ell,i}\} \cup \{x : \hat{p}_h(x) > \theta^s_{h,c_{h,t}},\, \hat{p}_l(x) > \theta^s_{\ell,c_{\ell,t}}\}, \\ A_{l,t} &= \bigcup_{i=1}^{c_{\ell,t}} \{x : \theta^s_{\ell,i} < \hat{p}_l(x) \le \theta^s_{\ell,i-1},\, \hat{p}_h(x) \le \theta^s_{h,i-1}\} \cup \{x : \hat{p}_l(x) \le \theta^s_{\ell,c_{\ell,t}}\}. \end{aligned} \tag{E.1}$$

Moreover, for each timestep t on which $\mathrm{dir}(t) = \mathrm{low}$,

$$\begin{aligned} A_{h,t} &= \bigcup_{i=1}^{c_{h,t}} \{x : \theta^s_{h,i-1} < \hat{p}_h(x) \le \theta^s_{h,i},\, \hat{p}_l(x) > \theta^s_{\ell,i}\} \cup \{x : \hat{p}_h(x) > \theta^s_{h,c_{h,t}}\}, \\ A_{l,t} &= \bigcup_{i=1}^{c_{\ell,t}} \{x : \theta^s_{\ell,i} < \hat{p}_l(x) \le \theta^s_{\ell,i-1},\, \hat{p}_h(x) \le \theta^s_{h,i-1}\} \cup \{x : \hat{p}_h(x) \le \theta^s_{h,c_{h,t}},\, \hat{p}_l(x) \le \theta^s_{\ell,c_{\ell,t}}\}. \end{aligned} \tag{E.2}$$

Proof. We proceed by induction on t. The base case of $t = 0$ is immediate. For the induction step, suppose that the result holds at time t and for simplicity that $\mathrm{dir}(t) = \mathrm{low}$ (the case where $\mathrm{dir}(t) = \mathrm{high}$ is identical). If $\mathrm{dir}(t+1) = \mathrm{dir}(t) = \mathrm{low}$ there is nothing to prove. Suppose $\mathrm{dir}(t+1) = \mathrm{high}$. Then,

$$\begin{aligned} A_{h,t+1} &= A_{h,t} \setminus \{x : \hat{p}_h(x) > \theta_{h,t},\, \hat{p}_l(x) \le \theta_{l,t}\} \\ &= \bigcup_{i=1}^{c_{h,t}} \{x : \theta^s_{h,i-1} < \hat{p}_h(x) \le \theta^s_{h,i},\, \hat{p}_l(x) > \theta^s_{\ell,i}\} \\ &\qquad \cup \{x : \hat{p}_h(x) > \theta^s_{h,c_{h,t}}\} \setminus \{x : \hat{p}_h(x) > \theta_{h,t},\, \hat{p}_l(x) \le \theta_{l,t}\}. \end{aligned}$$

By definition $c_{h,t+1} = c_{h,t}$, $\theta^s_{c_{h,t}} = \theta_{h,t}$, $c_{\ell,t+1} = c_{\ell,t} + 1$, and $\theta^s_{\ell,c_{\ell,t+1}} = \theta_{l,t}$. The above can be rewritten as

$$\bigcup_{i=1}^{c_{h,t+1}} \{x : \theta^s_{h,i-1} < \hat{p}_h(x) \le \theta^s_{h,i},\, \hat{p}_l(x) > \theta^s_{\ell,i}\} \cup \{x : \hat{p}_h(x) > \theta^s_{h,c_{h,t+1}},\, \hat{p}_l(x) > \theta^s_{\ell,c_{\ell,t+1}}\},$$

as desired. Moreover, note that by construction $c_{\ell,t+1} = c_{h,t} + 1$. So, we also have that

$$\begin{aligned} A_{l,t+1} &= A_{l,t} \cup \{x : \hat{p}_h(x) > \theta_{h,t},\, \hat{p}_l(x) \le \theta_{l,t}\} \\ &= \bigcup_{i=1}^{c_{\ell,t}} \{x : \theta^s_{\ell,i} < \hat{p}_l(x) \le \theta^s_{\ell,i-1},\, \hat{p}_h(x) \le \theta^s_{h,i-1}\} \\ &\qquad \cup \{x : \hat{p}_h(x) \le \theta^s_{h,c_{h,t}}, \hat{p}_l(x) \le \theta^s_{\ell,c_{\ell,t}}\} \cup \{x : \hat{p}_h(x) > \theta_{h,t},\, \hat{p}_l(x) \le \theta_{l,t}\} \\ &= \bigcup_{i=1}^{c_{\ell,t}} \{x : \theta^s_{\ell,i} < \hat{p}_l(x) \le \theta^s_{\ell,i-1},\, \hat{p}_h(x) \le \theta^s_{h,i-1}\} \\ &\qquad \cup \{x : \hat{p}_h(x) \le \theta^s_{h,c_{h,t}},\, \hat{p}_l(x) \le \theta^s_{\ell,c_{\ell,t}}\} \cup \{x : \hat{p}_h(x) > \theta^s_{h,c_{h,t}},\, \hat{p}_l(x) \le \theta^s_{\ell,c_{\ell,t+1}}\} \\ &= \bigcup_{i=1}^{c_{\ell,t+1}} \{x : \theta^s_{\ell,i} < \hat{p}_l(x) \le \theta^s_{\ell,i-1},\, \hat{p}_h(x) \le \theta^s_{h,i-1}\} \cup \{x : \hat{p}_l(x) \le \theta^s_{\ell,c_{\ell,t+1}}\}. \end{aligned}$$

□

Our next lemma upper bounds the loss of the ensembled predictor computed by the merge procedure at each iteration of the algorithm.

Lemma 7. *Let Θ_h, Θ_l be finite subsets of $[0,1]$, with $\min\Theta_h > \max\Theta_l$, and assume that $\hat{p}_l(X) \in [0, \min\Theta_h)$ and $\hat{p}_h(X) \in (\max\Theta_l, 1]$. For each t, let*

$$\hat{p}_{m,t}(x) = \hat{p}_l(x)\mathbb{1}\{x \in A_{l,t}\} + \hat{p}_h(x)\mathbb{1}\{x \in A_{h,t}\}.$$

Fix $\epsilon > 0$ and suppose that,

$$\max_{\theta_h\in\Theta_h, \theta_l\in\Theta_l, \theta\in\{\theta_h,\theta_l\}} \left|(\hat{\mathbb{E}}_n - \mathbb{E})[(\ell_\theta(1,Y) - \ell_\theta(0,Y))\mathbb{1}\{\hat{p}_h(X) > \theta_h,\, \hat{p}_l(X) \le \theta_l\}]\right| \le \epsilon.$$

Then, for all t such that $\mathrm{dir}(t) = \text{high}$ we have

$$\max_{\theta\in\Theta_h:\theta<\theta_{h,t}} \mathbb{E}[\ell_\theta(\hat{p}_{m,t}(X),Y) - \ell_\theta(\hat{p}_h(X),Y)] \le 2\epsilon \;\; \textit{and} \;\; \max_{\theta\in\Theta_l} \mathbb{E}[\ell_\theta(\hat{p}_{m,t}(X),Y) - \ell_\theta(\hat{p}_l(X),Y)] \le 2\epsilon.$$

Similarly, for all t such that $\mathrm{dir}(t) = \text{low}$ we have

$$\max_{\theta\in\Theta_h} \mathbb{E}[\ell_\theta(\hat{p}_{m,t}(X),Y) - \ell_\theta(\hat{p}_h(X),Y)] \le 2\epsilon \;\; \textit{and} \;\; \max_{\theta\in\Theta_l:\theta>\theta_{l,t}} \mathbb{E}[\ell_\theta(\hat{p}_{m,t}(X),Y) - \ell_\theta(\hat{p}_l(X),Y)] \le 2\epsilon,$$

where the expectations above are taken over the randomness in (X,Y) with the predictors held fixed.

Proof. We prove this by induction. The base case of $t = 0$ is immediate. For the inductive step, suppose the result holds at timestep t. Assume for simplicity that $\mathrm{dir}(t) = \text{high}$ (the case $\mathrm{dir}(t) = \text{low}$ is identical). There are two cases.

Case 1: dir$(t+1)$ = high. In this case the predictor does not change. Thus, to obtain the desired result we just need to show that

$$\mathbb{E}[\ell_{\theta_{h,t}}(\hat{p}_{m,t}(X),Y) - \ell_{\theta_{h,t}}(\hat{p}_h(X),Y)] \le 2\epsilon.$$

By Lemma 6, we have

$$\begin{aligned}
&\mathbb{E}[\ell_{\theta_{h,t}}(\hat{p}_{m,t}(X),Y) - \ell_{\theta_{h,t}}(\hat{p}_h(X),Y)] \\
&\qquad = \mathbb{E}[(\ell_{\theta_{h,t}}(0,Y) - \ell_{\theta_{h,t}}(1,Y))\mathbb{1}\{X \in A_{l,t},\, \hat{p}_h(X) > \theta_{h,t},\, \hat{p}_l(X) \le \theta_{h,t}\}] \\
&\qquad = \mathbb{E}[(\ell_{\theta_{h,t}}(0,Y) - \ell_{\theta_{h,t}}(1,Y))\mathbb{1}\{\hat{p}_l(X) \le \theta^s_{c_{\ell,t}},\, \hat{p}_h(X) > \theta_{h,t}\}].
\end{aligned}$$

Now, by construction, $\theta^s_{c_{\ell,t}} = \theta_{l,t}$. So, the above is quantity is exactly equal to

$$\begin{aligned}
&\mathbb{E}[(\ell_{\theta_{h,t}}(0,Y) - \ell_{\theta_{h,t}}(1,Y))\mathbb{1}\{\hat{p}_l(X) \le \theta_{l,t},\, \hat{p}_h(X) > \theta_{h,t}\}] \\
&\qquad = (\mathbb{E} - \hat{\mathbb{E}}_n)[(\ell_{\theta_{h,t}}(0,Y) - \ell_{\theta_{h,t}}(1,Y))\mathbb{1}\{\hat{p}_l(X) \le \theta_{l,t},\, \hat{p}_h(X) > \theta_{h,t}\}] \\
&\qquad\qquad + \hat{\mathbb{E}}_n[(\ell_{\theta_{h,t}}(0,Y) - \ell_{\theta_{h,t}}(1,Y))\mathbb{1}\{\hat{p}_l(X) \le \theta_{l,t},\, \hat{p}_h(X) > \theta_{h,t}\}] \\
&\qquad \le 2\epsilon,
\end{aligned}$$

where to obtain the last line we recall that $\mathrm{dir}(t) = \mathrm{dir}(t+1) = \text{high}$ and thus the empirical expectation in the second term must be at most ϵ.

Case 2: dir$(t+1)$ = low. In this case, by construction, in order to have $\mathrm{dir}(t) = \text{high}$ and $\mathrm{dir}(t+1) = \text{low}$ we must have that

$$\hat{\mathbb{E}}_n[(\ell_{\theta_{h,t}}(1,Y) - \ell_{\theta_{h,t}}(0,Y))\mathbb{1}\{\hat{p}_l(X) \le \theta_{l,t},\, \hat{p}_h(X) > \theta_{h,t}\}] < -\epsilon.$$

Notably, it follows immediately that

$$\hat{\mathbb{E}}_n[(\ell_\theta(1,Y) - \ell_\theta(0,Y))\mathbb{1}\{\hat{p}_l(X) \le \theta_{l,t},\, \hat{p}_h(X) > \theta_{h,t}\}] < -\epsilon, \;\; \text{for all } \theta \le \theta_{h,t}.$$

We will use this fact multiple times in the calculations that follow.

We consider a series of subcases. First, consider the case where $\theta \in \{\theta' \in \Theta_l : \theta' \geq \theta_{l,t}\}$. By the induction hypothesis,

$$\begin{aligned}
\mathbb{E}[\ell_\theta(\hat{p}_{m,t+1}(X), Y) - \ell_\theta(\hat{p}_l(X), Y)] &\leq \mathbb{E}[\ell_\theta(\hat{p}_{m,t+1}(X), Y) - \ell_\theta(\hat{p}_{m,t}(X), Y)] + 2\epsilon \\
&= \mathbb{E}[(\ell_\theta(1, Y) - \ell_\theta(0, Y))\mathbb{1}\{\hat{p}_h(X) > \theta_{h,t},\, \hat{p}_l(X) \leq \theta_{\ell,t}\}] + 2\epsilon \\
&\leq (\mathbb{E} - \hat{\mathbb{E}}_n)[(\ell_\theta(1, Y) - \ell_\theta(0, Y))\mathbb{1}\{\hat{p}_h(X) > \theta_{h,t},\, \hat{p}_l(X) \leq \theta_{\ell,t}\}] \\
&\qquad + \hat{\mathbb{E}}_n[(\ell_\theta(1, Y) - \ell_\theta(0, Y))\mathbb{1}\{\hat{p}_h(X) > \theta_{h,t},\, \hat{p}_l(X) \leq \theta_{\ell,t}\}] + 2\epsilon \\
&\leq \epsilon - \epsilon + 2\epsilon \\
&= 2\epsilon.
\end{aligned}$$

On the other hand, for $\theta \geq \theta_{h,t}$ we have that $\hat{p}_{m,t+1}(x) > \theta \iff \hat{p}_h(x) > \theta$ (recalling Lemma 6 and the fact that $\theta^s_{h,c_{h,t}} = \theta_{h,t}$) and thus,

$$\mathbb{E}[\ell_\theta(\hat{p}_{m,t+1}(X), Y) - \ell_\theta(\hat{p}_h(X), Y)] = 0.$$

Finally, for $\theta \in \{\theta' \in \Theta_h : \theta' < \theta_{h,t}\}$ we have

$$\begin{aligned}
\mathbb{E}[\ell_\theta(\hat{p}_{m,t+1}(X), Y) - \ell_\theta(\hat{p}_h(X), Y)] &\leq \mathbb{E}[\ell_\theta(\hat{p}_{m,t+1}(X), Y) - \ell_\theta(\hat{p}_{m,t}(X), Y)] + 2\epsilon \\
&= \mathbb{E}[(\ell_\theta(1, Y) - \ell_\theta(0, Y))\mathbb{1}\{\hat{p}_h(X) > \theta_{h,t},\, \hat{p}_l(X) \leq \theta_{\ell,t}\}] + 2\epsilon \\
&\leq 2\epsilon,
\end{aligned}$$

as above. □

We are now ready to prove Lemma 5 which follows as an almost immediate corollary of Lemma 7.

Proof of Lemma 5. By Hoeffding's inequality we have that

$$\max_{\theta_h \in \Theta_h, \theta_l \in \Theta_l, \theta \in \{\theta_h, \theta_l\}} \left|(\hat{\mathbb{E}}_n - \mathbb{E})[(\ell_\theta(1, Y) - \ell_\theta(0, Y))\mathbb{1}\{\hat{p}_h(X) > \theta_h,\, \hat{p}_l(X) \leq \theta_l]\right| = O_{\mathbb{P}}\left(\sqrt{\frac{\log(|\Theta_h| \cdot |\Theta_l|)}{n}}\right).$$

Plugging this fact into the statement of Lemma 7 and taking t to be the last timestep of Algorithm 5 proves the desired result. □

With the above lemmas in hand the proof of Theorem 4 is immediate.

Proof of Theorem 4. This result follows from combining Lemma 5 with the results of Section 4 then adding up the cumulative error over all $\log_2(m)$ rounds of Algorithm 3. □

F Additional details for the sales forecasting example

For our sales forecasting example in Section 6.2 we need to compute the forecasted probability of observing a nonzero number of sales given a predicted set of quantiles. Formally, let $Y_c \in \mathbb{R}$ denote the number of sales of an item on a given day at a given Walmart location. For probability levels $0 < \tau_1 < \cdots < \tau_k < 1$, denote the corresponding quantile estimates by $\hat{q}^{\tau_1} \leq \cdots \leq \hat{q}^{\tau_k}$. Note that we can estimate the cumulative distribution function of Y_c via linear interpolation: for $x \in \mathbb{R}$,

$$\hat{\mathbb{P}}(Y_c \leq x) = \begin{cases} 1, & x \geq \hat{q}^{\tau_k}, \\ 0, & x < \hat{q}^{\tau_1}, \\ \tau_{i-1} + \frac{\tau_i - \tau_{i-1}}{\hat{q}^{\tau_i} - \hat{q}^{\tau_{i-1}}}(x - \hat{q}^{\tau_{i-1}}), & \hat{q}^{\tau_{i-1}} \leq x < \hat{q}^{\tau_i}. \end{cases}$$

We conclude this section with Figure 4 which displays the results of our sales forecasting experiments for varying hyperparameter values.

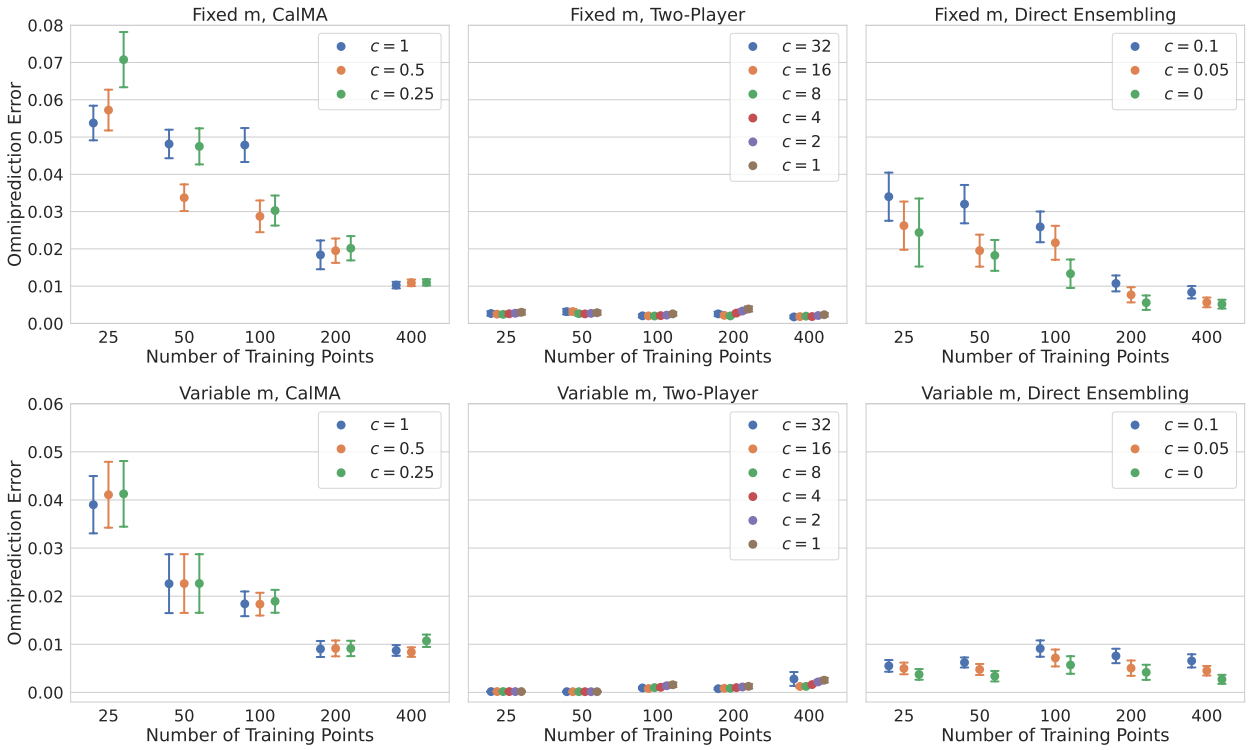

Figure 4: Omniprediction error of the calibrated multiaccuracy (left panels), two-player game based (center panels), and direct ensembling (right panels) methods across various sample sizes with $m = 16$ fixed (top row) or chosen variably as $m = 2^{\lfloor \log_2(\sqrt{n}) \rfloor}$ (bottom row) as the scaling constant c varies, for the M5 sales forecasting dataset. Dots and error bars show the means and standard errors obtained by evaluating the omniprediction error over 2000 test points for 20 draws of the training dataset.

G Proofs for Section 7

In this section, we prove Proposition 5.

Proof of Proposition 5. The statement given in Proposition 5 is a slight variation on Corollary 9 of Steinwart et al. [2014]. In particular, we assume that the losses under consideration are strictly proper, while Steinwart et al. [2014] instead assumes that the losses are order sensitive. To be precise, they restrict to losses ℓ such that for all distributions $P \in \mathcal{P}$ and all $t_1, t_2 \in \text{Image}(T)$ such that either $t_2 < t_1 < T(P)$ or $T(P) < t_1 < t_2$,

$$\mathbb{E}_P[\ell(t_1, Y)] < \mathbb{E}_P[\ell(t_2, Y)].$$

We show here that this latter condition is implied by strict propriety.

To this end, let ℓ be a strictly proper loss for T and fix any $t_1, t_2 \in \text{Image}(T)$ such that $t_2 < t_1 < T(P)$ or $T(P) < t_1 < t_2$. Let P_1 and P_2 be such that $T(P_1) = t_1$ and $T(P_2) = t_2$. By the continuity of T, there exists $\lambda \in (0, 1)$ such that $T(\lambda P_2 + (1 - \lambda)P) = T(P_1)$. Moreover, since ℓ is strictly proper we must have that

$$\begin{aligned}
\lambda \mathbb{E}_{P_2}[\ell(t_1, Y)] + (1 - \lambda)\mathbb{E}_P[\ell(t_1, Y)] &= \mathbb{E}_{\lambda P_2 + (1-\lambda)P}[\ell(t_1, Y)] \\
&< \mathbb{E}_{\lambda P_2 + (1-\lambda)P}[\ell(t_2, Y)] \\
&= \lambda \mathbb{E}_{P_2}[\ell(t_2, Y)] + (1 - \lambda)\mathbb{E}_P[\ell(t_2, Y)],
\end{aligned}$$

and so in particular,

$$(1 - \lambda)(\mathbb{E}_P[\ell(t_2, Y)] - \mathbb{E}_P[\ell(t_1, Y)]) > \lambda(\mathbb{E}_{P_2}[\ell(t_1, Y)] - \mathbb{E}_{P_2}[\ell(t_2, Y)]) > 0,$$

as desired. □

H Auxiliary results

In this section, we state a few results from past work that were used in the proofs from the previous sections. We begin by recalling the well-known Azuma-Hoeffding inequality [Hoeffding, 1963, Azuma, 1967].

Theorem 5 (As stated in Theorem 9.7 of Hazan [2019])**.** *Let $\{X_t\}_{t=1}^T$ be a martingale with bounded differences $\mathbb{P}(|X_t - X_{t-1}| \leq B) = 1$, for all $2 \leq t \leq T$. Then, for all $c \in \mathbb{R}$,*

$$\mathbb{P}(|X_T - \mathbb{E}[X_T]| \geq c) \leq 2 \exp\left(- \frac{c^2}{2B^2 T} \right).$$

We next recall the regret bound for the hedge algorithm from the online learning literature [Vovk, 1990, Littlestone and Warmuth, 1994, Freund and Schapire, 1997].

Theorem 6 (As stated in Theorem 1.5 of Hazan [2019])**.** *Consider an online learning problem with m experts receiving bounded losses $\{\ell_{t,i}\}_{1 \leq i \leq m, 1 \leq t \leq T}$ with $\sup_{1 \leq i \leq m, 1 \leq t \leq T} \ell_{t,i} \leq B$. Suppose that at time t we make the same prediction as expert i with probability*

$$q_{t,i} = \frac{\exp(-\eta \sum_{s<t} \ell_{s,i})}{\sum_{j=1}^m \exp(-\eta \sum_{s<t} \ell_{s,j})},$$

for some $\eta > 0$. Then,

$$\sum_{t=1}^T \mathbb{E}_{I \sim q_t}[\ell_{t,I}] \leq \min_{1 \leq i \leq m} \sum_{t=1}^T \ell_{t,i} + \eta T B^2 + \frac{\log(m)}{\eta}.$$